LOCAL CONVERGENCE OF AN ALGORITHM FOR SUBSPACE IDENTIFICATION FROM PARTIAL DATA

LAURA BALZANO* AND STEPHEN J. WRIGHT†

Abstract. GROUSE (Grassmannian Rank-One Update Subspace Estimation) is an iterative algorithm for identifying a linear subspace of $\mathbb{R}^n$ from data consisting of partial observations of random vectors from that subspace. This paper examines local convergence properties of GROUSE, under assumptions on the randomness of the observed vectors, the randomness of the subset of elements observed at each iteration, and incoherence of the subspace with the coordinate directions. Convergence at an expected linear rate is demonstrated under certain assumptions. The case in which the full random vector is revealed at each iteration allows for much simpler analysis, and is also described. GROUSE is related to incremental SVD methods and to gradient projection algorithms in optimization.

Key words. Subspace Identification, Optimization

1. Introduction.

We seek to identify an unknown subspace $\mathcal{S}$ of dimension d in $\mathbb{R}^n$, described by an $n \times d$ matrix $\bar{U}$ whose orthonormal columns span $\mathcal{S}$. Our data consist of a sequence of vectors v_t of the form

$$v_t = \bar{U} s_t, \tag{1.1}$$

where $s_t \in \mathbb{R}^d$ is a random vector whose elements are independent and identically distributed (i.i.d.) in $\mathcal{N}(0,1)$. Critically, we observe only a subset $\Omega_t \subset \{1,2,\ldots,n\}$ of the components of v_t.

GROUSE [2, 3] (Grassmannian Rank-One Update Subspace Estimation) is an algorithm that generates a sequence $\{U_t\}_{t=0,1,\ldots}$ of $n \times d$ matrices with orthonormal columns with the goal that $R(U_t) \to \mathcal{S}$ (where $R(\cdot)$ denotes range). Partial observation of the vector v_t is used to update U_t to U_{t+1}. We present GROUSE (slightly modified from earlier descriptions) as Algorithm 1.

1.1. Applications of Subspace Identification.

Subspace identification problems arise in a great variety of applications. They are the simplest form of the more general class of problems in which we seek to identify a low-dimensional manifold in a high-dimensional ambient space from a sequence of incomplete observations. Subspace identification finds applications in medical [1] and hyperspectral [14] imaging, communications [19], source localization and target tracking in radar and sonar [12], computer vision for object tracking [8], and in control for system identification [21, 20], where one is interested in estimating the range space of the observability matrix of a system. Subspaces have also been used to represent images of a single scene under varying illuminations [6] and to model origin-destination flows in a computer network [13]. Environmental monitoring of soil and crop conditions [10], water contamination [16], and seismological activity [22] can all be summarized efficiently by low-dimensional subspace representations.

1.2. GROUSE.

Each iteration of the GROUSE algorithm (Algorithm 1) essentially performs a gradient projection step onto the Grassmannian manifold of subspaces of dimension d, based on the latest partially observed sample $[v_t]_{\Omega_t}$ of the

*Department of Electrical Engineering and Computer Science, University of Michigan. girasole@umich.edu

†Department of Computer Sciences, University of Wisconsin-Madison. swright@cs.wisc.edu

random vector $v_t \in \mathcal{S}$. In this description, we use $[U]_{\Omega_t}$ to denote the row submatrix of the $n \times d$ matrix U corresponding to the index set $\Omega_t \subset \{1, 2, \ldots, n\}$. Similarly, $[z]_{\Omega_t}$ denotes the subvector of elements of $z \in \mathbb{R}^n$ corresponding to elements of Ω_t. We use $|\Omega_t|$ to denote cardinality of the set Ω_t and Ω_t^c to denote the complement $\{1, 2, \ldots, n\} \setminus \Omega_t$.

Algorithm 1 GROUSE: Partial Data

Given U_0, an $n \times d$ matrix with orthonormal columns, with $0 < d < n$;
Set $t := 1$;
repeat
  Draw a random subset $\Omega_t \subset \{1, 2, \ldots, n\}$ and and observe $[v_t]_{\Omega_t}$ where $v_t \in \mathcal{S}$;
  if the eigenvalues of $[U_t]_{\Omega_t}^T [U_t]_{\Omega_t}$ lie in the range $[0.5|\Omega_t|/n, 1.5|\Omega_t|/n]$ **then**
    Define $w_t := \arg\min_w \|[U_t]_{\Omega_t} w - [v_t]_{\Omega_t}\|_2^2$;
    Define $p_t := U_t w_t$; $[r_t]_{\Omega_t} := [v_t]_{\Omega_t} - [p_t]_{\Omega_t}$; $[r_t]_{\Omega_t^c} := 0$; $\sigma_t := \|r_t\| \, \|p_t\|$;
    Choose $\eta_t > 0$ and set

$$U_{t+1} := U_t + \left[(\cos(\sigma_t \eta_t) - 1) \frac{p_t}{\|p_t\|} + \sin(\sigma_t \eta_t) \frac{r_t}{\|r_t\|} \right] \frac{w_t^T}{\|w_t\|}. \tag{1.2}$$

  end if
  $t := t + 1$;
until termination

This description in Algorithm 1 differs from that of [3] only in that the following condition is required for the eigenvalues of $[U_t]_{\Omega_t}^T [U_t]_{\Omega_t}$:

$$\lambda_i([U_t]_{\Omega_t}^T [U_t]_{\Omega_t}) \in \left[.5 \frac{|\Omega_t|}{n}, 1.5 \frac{|\Omega_t|}{n} \right], \quad i = 1, 2, \ldots, d, \tag{1.3}$$

where $\lambda_i(\cdot)$ denotes the ith eigenvalue (in decreasing order). A consequence is that

$$\|([U_t]_{\Omega_t}^T [U_t]_{\Omega_t})^{-1}\| \leq \frac{2n}{|\Omega_t|}. \tag{1.4}$$

As we see later in Theorem 2.6, this condition ensures that the sample Ω_t is such that $[v_t]_{\Omega_t}$ captures useful information about $\mathcal{S}$; if it is not satisfied, the weight vector w_t may not accurately reflect how the latest observation $[v_t]_{\Omega_t}$ is explained by the current basis vectors (the columns of $[U_t]_{\Omega_t}$). Since we need to factor the matrix $[U_t]_{\Omega_t}$ in order to calculate w_t, and since we have typically that $d \ll n$, the marginal cost of determining or estimating the singular values of $[U_t]_{\Omega_t}$ and checking the condition (1.4) is not excessive. We show in our analysis that the condition (1.3) is satisfied at most iterations.

We note several elementary facts about the vector quantities that appear in GROUSE. Let $P_{R(\cdot)}$ denote the projection operator onto the range, and $P_{N(\cdot)}$ denote the projection onto the nullspace of a matrix. Since

$$[p_t]_{\Omega_t} = P_{R([U_t]_{\Omega_t})}([v_t]_{\Omega_t}), \quad [r_t]_{\Omega_t} = P_{N([U_t]_{\Omega_t}^T)}([v_t]_{\Omega_t}),$$

we have that

$$p_t^T r_t = [p_t]_{\Omega_t}^T [r_t]_{\Omega_t} = 0 \tag{1.5}$$

and

$$\|p_t + r_t\|^2 = \|p_t\|^2 + \|r_t\|^2. \tag{1.6}$$

By orthonormality of the columns of U_t, we also have that

$$\|p_t\| = \|w_t\|. \tag{1.7}$$

1.3. GROUSE in Context. The derivation of GROUSE as a stochastic gradient algorithm on the Grassmannian manifold can be found in [3], along with a discussion of its relationship to matrix completion. In this subsection, we discuss several other aspects of GROUSE's convergence behavior, focusing on the regime in which the iterates U_t are close to identifying the correct subspace $\mathcal{S}$, so that $\|r_t\| \ll \|p_t\|$. We assume that the steplength η_t is chosen to satisfy

$$\sin \sigma_t \eta_t = \frac{\|r_t\|}{\|p_t\|}. \tag{1.8}$$

Since $1 - \cos \sigma_t \eta_t = O(\|r_t\|^2/\|p_t\|^2)$, by multiplying both sides of (1.2) by w_t, and using (1.7), we have that

$$U_{t+1} w_t = U_t w_t + \frac{\|r_t\|}{\|p_t\|} \frac{r_t}{\|r_t\|} \frac{w_t^T w_t}{\|w_t\|} + O\left(\frac{\|r_t\|^2}{\|p_t\|^2} \right) = p_t + r_t + O\left(\frac{\|r_t\|^2}{\|p_t\|^2} \right).$$

It follows from the definition of r_t that

$$[U_{t+1} w_t]_{\Omega_t} \approx [p_t + r_t]_{\Omega_t} = [v_t]_{\Omega_t}, \tag{1.9a}$$

$$[U_{t+1} w_t]_{\Omega_t^c} \approx [p_t + r_t]_{\Omega_t^c} = [U_t w_t]_{\Omega_t^c}, \tag{1.9b}$$

where $\Omega_t^c := \{1, 2, \dots, n\} \setminus \Omega_t$. Moreover, in any direction z orthogonal to w_t, we have $U_{t+1} z = U_t z$. Thus, the update (1.2) has the effect of (approximately) matching the newly revealed information $[v_t]_{\Omega_t}$ along the direction w_t, while leaving the values of $U_t w_t$ almost unchanged in the non-revealed components Ω_t^c, and making no change at all in the remaining $(d-1)$-dimensional subspace $\{z \,|\, w_t^T z = 0\}$. In this sense, (1.2) is a "least-change" update, leaving the current iterate U_t undisturbed as far as possible, but making just enough of a change to match the new information. The least-change strategy is key to the development of quasi-Newton methods for optimization [15, Chapter 6], in which low-rank, least-change updates are made to approximate Hessian matrices, to match the curvature information gained in each step.

The relationship of GROUSE to gradient projection becomes clearer when we define the following measure of inconsistency between $R(U_t)$ and $\mathcal{S}$, based on the information revealed in $[v_t]_{\Omega_t}$:

$$\mathcal{E}(U_t) := \min_w \|[U_t]_{\Omega_t} w - [v_t]_{\Omega_t}\|_2^2.$$

It can be shown that

$$\frac{d\mathcal{E}}{dU_t} = -2 r_t w_t^T.$$

With the choice (1.8) of η_t, we have from (1.2) that

$$U_{t+1} = U_t + \frac{1}{\|p_t\|^2} r_t w_t^T + O\left(\frac{\|r_t\|^2}{\|p_t\|^2} \right),$$

so that the GROUSE step is a step in the negative gradient direction for $\mathcal{E}$, projected onto the space of $n \times d$ matrices with orthonormal columns.

GROUSE is related too to an incremental singular value decomposition (ISVD) approach that maintains an approximation U_t with orthonormal columns, and iterates in the following way. First, the new random vector v_t is appended to U_t to form an $n \times (d+1)$ matrix, with missing elements of v_t "imputed" from the current estimate U_t and the weight vector w_t obtained as in GROUSE. Second, the SVD of this expanded matrix is computed, and the first d columns of its left factor (an $n \times (d+1)$ matrix with orthonormal columns) are taken as the new iterate U_{t+1}. (The final column is discarded.) It is shown in [5] that for a certain choice of steplength parameter η_t in GROUSE, the ISVD and GROUSE algorithms are equivalent.

In our analysis below, we use the following generalization of (1.8) for the choice of η_t:

$$\sin \sigma_t \eta_t = \alpha_t \frac{\|r_t\|}{\|p_t\|}, \tag{1.10}$$

where $\alpha_t \in (0, 2)$ is a user-defined "fudge factor." We show that the best asymptotic results are obtained by setting $\alpha_t \equiv 1$.

1.4. Summary of Results. Our main result is expected local linear convergence of the sequence of subspaces $\{R(U_t)\}$ to $\mathcal{S}$. This section outlines the assumptions needed to prove our result and discusses their relevance to computational experience.

We recall the assumption that the observation vector v_t has the form $\bar{U} s_t$ (1.1), with the elements of s_t being i.i.d. normal with zero mean and identical variance. We assume too that the set Ω_t of observed elements of v_t is chosen independently at each iteration.

The discrepancy between the d-dimensional subspaces $R(U_t)$ and $\mathcal{S}$ is measured in terms of the *d principal angles* between these subspaces, $\phi_i(U_t, \bar{U})$ [18, Chapter 5], which are defined by

$$\cos \phi_i(U_t, \bar{U}) = \sigma_i(\bar{U}^T U_t), \quad i = 1, 2, \dots, d, \tag{1.11}$$

where $\sigma_i(\bar{U}^T U_t)$, $i = 1, 2, \dots, d$ are the singular values of $U_t^T \bar{U}$. The quantity ϵ_t defined by

$$\epsilon_t := \sum_{i=1}^{d} \sin^2 \phi_i(\bar{U}, U_t) = \sum_{i=1}^{d} (1 - \sigma_i^2(\bar{U}^T U_t)) = d - \|\bar{U}^T U_t\|_F^2 \tag{1.12}$$

is central to our analysis. We show that for small ϵ_t, we have

$$\epsilon_{t+1} \approx \epsilon_t - \frac{\|r_t\|^2}{\|w_t\|^2}, \tag{1.13}$$

and that the expected value of the decrease $\|r_t\|^2 / \|w_t\|^2$ is bounded below by a small multiple of ϵ_t, provided that the eigenvalue check (1.3) is satisfied. (Higher-order terms complicate the analysis considerably.)

A critical assumption, made precise below, is *incoherence* of the subspace $\mathcal{S}$ with respect to the coordinate directions. Concepts of incoherence have been well studied in the context of compressed sensing (see for example [7]). If $\mathcal{S}$ were to align closely with one or two principal axes, then observation subsets Ω_t that did not include the corresponding index would be missing important information about $\mathcal{S}$. We would need to choose larger sample sets Ω_t (of size $|\Omega_t|$ related to n), or to take many more iterations, in order to have a good chance of capturing the components of v_t that align with $\mathcal{S}$.

Our analysis requires another kind of incoherence too. We assume that the error in U_t revealed by the observation vector — the part of v_t that is *not* explained by the current iterate U_t — is usually incoherent with respect to the coordinate directions. (Our computations indicate that such is the case.) This incoherence measure is denoted by $\mu(x_t)$, where $x_t := (I - U_t U_t^T)v_t$, and our assumption on this quantity is spelled out in Lemma 2.9.

High-probability results play a key role in the analysis. Our lower bound on the quantity $\|r_t\|^2/\|w_t\|^2$ in (1.13), for instance, is not proved to hold at every iteration but only at a substantial majority of iterations. In fact, it is possible that $\epsilon_{t+1} > \epsilon_t$ for some t; the sequence $\{\epsilon_t\}$ may not decrease monotonically.

We state at the outset that the expected linear convergence behavior is proved to hold in only a limited regime, that is, the main theorem requires ϵ_t to be quite small and each $|\Omega_t|$ to be on the order of $d(\log d)(\log^2 n)$ in order for the claimed linear rate to be observed. This requirement on observations is only $\log d$ greater than what is required for batch matrix completion algorithms [17]. The linear convergence rate observed in computational experiments is, roughly speaking, a factor of $(1 - Xq/(nd))$ per iteration, where q is a lower bound on $|\Omega_t|$ and X is some number not too much less than one. We see in Section 4 that this rate appears to hold in a much wider regime than the analysis would strictly predict, both for much smaller $|\Omega_t|$ and for much larger ϵ_t. In fact, the same "gap" between theory and practice of local convergence is seen in many optimization algorithms. We point out too that the mismatch largely disappears in the full-data case, where $\Omega_t = \{1, 2, \dots, n\}$ for all t. In this case, the theoretical restrictions on ϵ_t are mild, incoherence is irrelevant, and the predicted convergence behavior matches closely the computational observations.

1.5. Outline. Section 2 contains the proof of our claim of expected linear convergence. This long section is broken into subsections, with a "roadmap" given at the start. Section 3 analyzes the full-data case in which $\Omega_t \equiv \{1, 2, \dots, n\}$. Many of the complications of the general case vanish here, but the specialized analysis holds some interest and convergence still occurs only in an expected sense, because of the random nature of the observation vectors v_t.

Notation. As noted earlier, we use $N(\cdot)$ to denote the null space (kernel) of a matrix and $P_{\mathcal{T}}$ to denote projection onto a subspace $\mathcal{T}$.

The notation $\|\cdot\|$ (without subscript) on either vector or matrix indicates $\|\cdot\|_2$. Recall that the Frobenius norm is related to $\|\cdot\|_2$ by the following inequalities:

$$\|A\|_2 \le \|A\|_F \le \sqrt{r}\|A\|_2,$$

where r is the rank of A. We note too that the norms $\|\cdot\|_2$ and $\|\cdot\|_F$ are invariant under orthogonal transformations of the matrix argument.

We drop the subscripts t frequently during the paper, when it causes no confusion to do so, and reminding the reader of this practice where appropriate.

2. Expected Linear Convergence. We develop the local convergence results for GROUSE in this section. The analysis is surprisingly technical for such a simple method, so we break the exposition into relatively short subsections. We give a brief outline of our proof strategy here.

Subsection 2.1 obtains a lower-bounding expression for the improvement in the measure ϵ_t (1.12) made over a single step. This bound involves three different quantities, and the rest of the paper focuses on controlling each of them. Subsection 2.2 shows that the Frobenius-norm difference between U_t and $\bar{U}$ can be bounded above and below by multiples of ϵ_t. Subsection 2.3 examines some consequences of the fact that only a subset Ω_t of the elements of v_t is revealed at each iteration. This subsection introduces an assumed lower bound q on the cardinality of Ω_t, and obtains bounds on $\|r_t\|$ and $\|p_t\|$ (and their ratio) in terms of the norm of the vector s_t from (1.1).

Subsection 2.4 examines a particular term $(\bar{U}^T p_t)^T(\bar{U}^T r_t)$ that appears in the lower-bounding expression for $\epsilon_t - \epsilon_{t+1}$ obtained earlier in Subsection 2.1, deriving bounds for this quantity in terms of ϵ_t, $\|p_t\|$, and $\|r_t\|$. These bounds are used in Subsection 2.5 to make the results of Subsection 2.1 more precise.

Subsection 2.6 defines the concept of coherence used in this paper, and uses a measure concentration result to show that the eigenvalue condition (1.3) is satisfied on most iterations. Subsection 2.7 proves a high-probability bound for the ratio $\|r_t\|^2/\|p_t\|^2$, which is the dominant term in the error improvement $\epsilon_t - \epsilon_{t+1}$. This bound is given in terms of the angle θ_t that is the angle between $R(U_t)$ and $\mathcal{S}$ that is revealed by the (full) random observation vector v_t. Subsection 2.8 shows that the expected value of $\sin^2\theta_t$ is ϵ_t/d. Finally, Subsection 2.9 puts the pieces together, proving expected linear convergence rate by combining bounds for the "good" iterations with those for the "anomalous" iterations, where the latter category includes those for which the update is skipped because condition (1.3) fails to hold.

2.1. A Bound for $\epsilon_t - \epsilon_{t+1}$. In this subsection, we obtain an expression for $\epsilon_{t+1} - \epsilon_t$, where ϵ_t is the quantity defined in (1.12). We deal mostly with the case in which a step is actually taken by the algorithm, that is, condition (1.3) holds. (If such is not the case, we have trivially that $\epsilon_{t+1} = \epsilon_t$.) We start by defining the $d \times d$ orthogonal matrix W_t as

$$W_t := \left[\frac{w_t}{\|w_t\|} \mid Z_t\right], \tag{2.1}$$

where Z_t is a $d \times (d-1)$ matrix with orthonormal columns whose columns span $N(w_t^T)$. It is clear that the first column of $U_t W_t$ is

$$\frac{U_t w_t}{\|w_t\|} = \frac{p_t}{\|p_t\|}.$$

Let us now write the update formula (1.2) as follows

$$U_{t+1} := U_t + \left[\frac{y_t}{\|y_t\|} - \frac{p_t}{\|p_t\|}\right]\frac{w_t^T}{\|w_t\|}, \tag{2.2}$$

$$\text{where } \frac{y_t}{\|y_t\|} := \cos(\sigma_t\eta_t)\frac{p_t}{\|p_t\|} + \sin(\sigma_t\eta_t)\frac{r_t}{\|r_t\|}. \tag{2.3}$$

By using a trigonometric identity together with (1.5), we can see that the right-hand side of (2.3) has unit norm. From (2.2), we have

$$\begin{aligned} U_{t+1}W_t &= U_t W_t + \left[\frac{y_t}{\|y_t\|} - \frac{p_t}{\|p_t\|}\right]\frac{w_t^T}{\|w_t\|}W_t \\ &= U_t W_t + \left[\frac{y_t}{\|y_t\|} - \frac{p_t}{\|p_t\|}\right]\begin{bmatrix}1 & 0 & 0 & \ldots & 0\end{bmatrix}, \end{aligned}$$

where y_t is defined in (2.3). Thus, the update has the effect of replacing the first column $p_t/\|p_t\|$ of $U_t W_t$ by $y_t/\|y_t\|$, and leaving the other columns unchanged. Recalling that the Frobenius norm is invariant under orthogonal transformations, using (1.12) and (2.3), and dropping the subscript t freely on scalars and vectors, we obtain

$$
\begin{aligned}
\epsilon_t - \epsilon_{t+1} &= \|\bar{U}^T U_{t+1}\|_F^2 - \|\bar{U}^T U_t\|_F^2 \\
&= \|\bar{U}^T U_{t+1} W_t\|_F^2 - \|\bar{U}^T U_t W_t\|_F^2 \\
&= \left\| \cos(\sigma\eta) \frac{\bar{U}^T p}{\|p\|} + \sin(\sigma\eta) \frac{\bar{U}^T r}{\|r\|} \right\|_2^2 - \left\| \frac{\bar{U}^T p}{\|p\|} \right\|_2^2 \\
&= (\cos^2(\sigma\eta) - 1) \frac{\|\bar{U}^T p\|^2}{\|p\|^2} + 2\cos(\sigma\eta)\sin(\sigma\eta) \frac{(\bar{U}^T p)^T (\bar{U}^T r)}{\|p\|\|r\|} + \sin^2(\sigma\eta) \frac{\|\bar{U}^T r\|^2}{\|r\|^2} \\
&= \sin^2(\sigma\eta) \left(\frac{\|\bar{U}^T r\|^2}{\|r\|^2} - \frac{\|\bar{U}^T p\|^2}{\|p\|^2} \right) + \sin(2\sigma\eta) \frac{(\bar{U}^T p)^T (\bar{U}^T r)}{\|p\|\|r\|} \\
&\geq -\sin^2(\sigma\eta) + \sin(2\sigma\eta) \frac{(\bar{U}^T p)^T (\bar{U}^T r)}{\|p\|\|r\|},
\end{aligned} \tag{2.4}
$$

where the final inequality follows from $\|\bar{U}^T p\| \leq \|p\|$ (since the columns of $\bar{U}$ are orthonormal) and $\|\bar{U}^T r\|^2/\|r\|^2 \geq 0$. Choosing η_t so that (1.10) is satisfied, we have from $\sin^2(\sigma\eta) \in [0,1]$ and for any scalar β that

$$\beta\sqrt{1 - \sin^2(\sigma\eta)} \geq \beta - |\beta| \sin^2(\sigma\eta),$$

and thus by substituting (1.10), we have

$$
\begin{aligned}
\sin(2\sigma\eta)\beta &= 2\sin(\sigma\eta)\beta\sqrt{1 - \sin^2(\sigma\eta)} \\
&\geq 2\sin(\sigma\eta)\beta - 2\sin^3(\sigma\eta)|\beta| = 2\alpha\frac{\|r\|}{\|p\|}\beta - 2\alpha^3\frac{\|r\|^3}{\|p\|^3}|\beta|.
\end{aligned}
$$

By substituting into (2.4), we obtain

$$\epsilon_t - \epsilon_{t+1} \geq -\alpha^2 \frac{\|r\|^2}{\|p\|^2} + 2\alpha\frac{\|r\|}{\|p\|}\frac{(\bar{U}^T p)^T(\bar{U}^T r)}{\|p\|\|r\|} - 2\alpha^3 \frac{\|r\|^2}{\|p\|^2} \frac{|(\bar{U}^T p)^T(\bar{U}^T r)|}{\|p\|^2}. \tag{2.5}$$

We will return to formula (2.5) in Section 2.4. To preview: we will show that

$$(\bar{U}^T p)^T(\bar{U}^T r) \approx \|r\|^2, \tag{2.6}$$

and that the final term on the right-hand side is higher-order. Thus, we can deduce that the right-hand side of (2.5) is approximately

$$\alpha(2-\alpha)\frac{\|r\|^2}{\|p\|^2},$$

and hence that the approximate maximal improvement $\epsilon_t - \epsilon_{t+1}$ is obtained by setting $\alpha = 1$, as claimed earlier.

2.2. Relating U_t to $\bar{U}$. We state here a fundamental result about the relationship between U_t, $\bar{U}$, and the quantity ϵ_t defined in (1.12). After an orthogonal transformation, the squared-Frobenius-norm difference between U_t and $\bar{U}$ is of the same order as ϵ_t.

Recalling the definition (1.11) of the principal angles $\phi_i(U_t, \bar{U})$ between the subspaces spanned by the columns of U_t and the columns of $\bar{U}$, we define

$$\Sigma_t := \text{diag}\,(\sin\phi_i(\bar{U}, U_t)), \quad \Gamma_t := \text{diag}\,(\cos\phi_i(\bar{U}, U_t)). \tag{2.7}$$

Recalling (1.12) and using the definitions (1.11) and (2.7), we have

$$\|\Sigma_t\|_F^2 = \sum_{i=1}^{d} \sin^2\phi_i(\bar{U}, U_t) = \epsilon_t, \tag{2.8a}$$

$$\|\Gamma_t\|_F^2 = \sum_{i=1}^{d} \sigma_i(\bar{U}^T U_t) = \sum_{i=1}^{d} \cos^2\phi_i(\bar{U}, U_t) = d - \epsilon_t, \tag{2.8b}$$

$$\|\bar{U}\bar{U}^T - U_t U_t^T\|_F^2 = 2d - 2\|\bar{U}^T U_t\|_F^2 = 2\epsilon_t. \tag{2.8c}$$

We have the following lemma.

LEMMA 2.1. *Let ϵ_t be as in (1.12) and suppose $n \geq 2d$. Then there is an orthogonal matrix $V_t \in \mathbb{R}^{d\times d}$ such that*

$$\epsilon_t \leq \|\bar{U}V_t - U_t\|_F^2 \leq 2\epsilon_t,$$

and thus $\|\bar{U}^T U_t - V_t\|_F^2 \leq 2\epsilon_t$.

Proof. The proof uses [18, Theorem 5.2]. There are unitary matrices Q_t, $\bar{Y}$, and Y_t such that

$$Q_t\bar{U}\bar{Y} := \begin{matrix} d \\ d \\ n-2d \end{matrix} \overset{d}{\begin{pmatrix} I \\ 0 \\ 0 \end{pmatrix}}, \quad Q_t U_t Y_t := \begin{matrix} d \\ d \\ n-2d \end{matrix} \overset{d}{\begin{pmatrix} \Gamma_t \\ \Sigma_t \\ 0 \end{pmatrix}}, \tag{2.9}$$

where Γ_t and Σ_t are as defined in (2.7). Defining the orthogonal matrix $V_t := \bar{Y}Y_t^T$ we have that

$$\begin{aligned} \bar{U}V_t &= Q_t^T \begin{pmatrix} I \\ 0 \\ 0 \end{pmatrix} \bar{Y}^T\bar{Y}Y_t^T \\ &= Q_t^T \begin{pmatrix} I \\ 0 \\ 0 \end{pmatrix} Y_t^T \\ &= Q_t^T \begin{pmatrix} \Gamma_t \\ \Sigma_t \\ 0 \end{pmatrix} Y_t^T + Q_t^T \begin{pmatrix} I - \Gamma_t \\ -\Sigma_t \\ 0 \end{pmatrix} Y_t^T \\ &= U_t + Q_t^T \begin{pmatrix} I - \Gamma_t \\ -\Sigma_t \\ 0 \end{pmatrix} Y_t^T. \end{aligned}$$

Therefore, using the abbreviated notation $\phi_i := \phi_i(\bar{U}, U_t)$, together with orthogonality of Q_t and Y_t, we have

$$\|\bar{U}V_t - U_t\|_F^2 = \|I - \Gamma_t\|_F^2 + \|\Sigma_t\|_F^2 = \sum_{i=1}^{d} [(1 - \cos\phi_i)^2 + \sin^2\phi_i].$$

By dropping the cosine part of each summation term, we obtain from (1.12) that $\|\bar{U}V_t - U_t\|_F^2 \geq \sum_{i=1}^d \sin^2\phi_i = \epsilon_t$, proving the lower bound. For the upper bound, we have

$$\begin{aligned}\|U_t^T\bar{U} - V_t^T\|_F^2 &= \sum_{i=1}^d [(1-\cos\phi_i)^2 + \sin^2\phi_i] \\ &= \sum_{i=1}^d [2 - 2\cos\phi_i] \leq \sum_{i=1}^d [2 - 2\cos^2\phi_i] = 2\sum_{i=1}^d \sin^2\phi_i = 2\epsilon_t,\end{aligned}$$

as required. The final claim is an immediate consequence of this upper bound. ∎

2.3. Consequences of Sampling. In this subsection we investigate some of the issues raised by observing the subspace vector v_t only on a sample set $\Omega_t \subset \{1, 2, \ldots, n\}$, seeing how some of the identities and bounds of Sections 2.1 and 2.2 are affected. We state a lower bound on the cardinality of Ω_t and an upper bound on ϵ_t that give sufficient conditions for these looser bounds to hold. These bounds are vital to the analysis of later subsections.

We start with a simple result about the relationship between $[\bar{U}]_{\Omega_t}$ and $[U_t]_{\Omega_t}$, based on Lemma 2.1.

LEMMA 2.2. *Let V_t be the matrix from Lemma 2.1. Then* $\|[\bar{U}]_{\Omega_t} - [U_t]_{\Omega_t}V_t^T\|_F^2 = \|[\bar{U}]_{\Omega_t}V_t - [U_t]_{\Omega_t}\|_F^2 \leq 2\epsilon_t$.

Proof. We have

$$\|[\bar{U}]_{\Omega_t}V_t - [U_t]_{\Omega_t}\|_F^2 \leq \|\bar{U}V_t - U_t\|_F^2 \leq 2\epsilon_t,$$

where the last inequality follows from Lemma 2.1. ∎

We now introduce some simplified notation for important quantities in our analysis, and state the representations of the key vectors w_t, p_t, w_t, and v_t in terms of this notation. We also make use of the vector s_t defined in (1.1). As in other parts of the paper, we drop the subscript t freely on vector quantities.

$$B := \bar{U}_{\Omega_t}, \tag{2.10a}$$
$$C := [U_t]_{\Omega_t}, \tag{2.10b}$$
$$P_{N(C^T)} = (I - C(C^TC)^{-1}C^T), \tag{2.10c}$$
$$P_{N(B^T)} = (I - B(B^TB)^{-1}B^T), \tag{2.10d}$$
$$[v_t]_{\Omega_t} = Bs, \tag{2.10e}$$
$$w = (C^TC)^{-1}C^TBs, \tag{2.10f}$$
$$p = U_tw = U_t(C^TC)^{-1}C^TBs, \tag{2.10g}$$
$$[p_t]_{\Omega_t} = P_{R(C)}Bs, \tag{2.10h}$$
$$[r_t]_{\Omega_t} = Bs - [p_t]_{\Omega_t} = (I - C(C^TC)^{-1}C^T)Bs = P_{N(C^T)}Bs. \tag{2.10i}$$

The notation B and C from (2.10a) and (2.10b) is used for simplicity in this subsection and the next. The reader will note that we have used both $(B^TB)^{-1}$ and $(C^TC)^{-1}$ freely. This property requires that our assumption (1.3) holds for U_t. It requires a similar property for $[\bar{U}]_{\Omega_t}$, something we simply assume for now, but prove later as a consequence of incoherence; see Theorem 2.6.

Note that we have from (2.10h) and (2.10i) that

$$P_{N(C^T)}[r_t]_{\Omega_t} = P_{N(C^T)}(Bs_t - Cw_t) = P_{N(C^T)}Bs_t. \tag{2.11}$$

For the remainder of the paper, we make the following assumptions on the size of the sample set Ω_t and the size of ϵ_t:

$$|\Omega_t| \geq q, \tag{2.12}$$

$$\epsilon_t \leq \frac{1}{128}\frac{q^2}{n^2 d}. \tag{2.13}$$

In later subsections, we will derive conditions on q that facilitate the convergence results. For now, we have the following estimates on vectors of interest.

LEMMA 2.3. *Suppose that (1.3) holds and that (2.12) and (2.13) are satisfied. Then we have*

$$\|r_t\| \leq \sqrt{2\epsilon_t}\|s_t\|, \tag{2.14}$$

$$\|p_t\| \in \left[\frac{3}{4}\|s_t\|, \frac{5}{4}\|s_t\|\right], \tag{2.15}$$

$$\frac{\|r_t\|^2}{\|p_t\|^2} \leq \frac{32}{9}\epsilon_t \tag{2.16}$$

Proof. We have from (2.10i), $[r_t]_{\Omega_t^c} = 0$, the fact that $P_{N(C^T)}C = 0$, and Lemma 2.2 that

$$\|r_t\| = \|[r_t]_{\Omega_t}\| = \|P_{N(C^T)}Bs_t\| = \\ \|P_{N(C^T)}(B - CV^T)s_t\| \leq \|B - CV^T\|\|s_t\| \leq \sqrt{2\epsilon_t}\|s_t\|,$$

proving (2.14).

We prove the lower bound in (2.15) (the upper bound is similar). We have from $\|C\| \leq \|U_t\| = 1$, Lemma 2.2, (1.4), (2.12), and (2.13), that

$$\begin{aligned}
\|p_t\| = \|w_t\| &= \|(C^TC)^{-1}C^TBs_t\| \\
&= \|(C^TC)^{-1}C^TBV(V^Ts_t)\| \\
&\geq \|(C^TC)^{-1}C^TC(V^Ts_t)\| - \|(C^TC)^{-1}C^T(BV - C)(V^Ts_t)\| \\
&\geq \|s_t\| - \|(C^TC)^{-1}\|\|C\|\|BV - C\|\|s_t\| \\
&\geq \|s_t\| - \frac{2n}{|\Omega_t|}\sqrt{2\epsilon_t}\|s_t\| \\
&\geq \|s_t\| - \frac{2n}{q}\sqrt{2\epsilon_t}\|s_t\| \\
&\geq \|s_t\| - \frac{2n}{q}\frac{1}{8}\frac{q}{n\sqrt{d}}\|s_t\| \\
&\geq \|s_t\| - \frac{1}{4}\|s_t\| = \frac{3}{4}\|s_t\|.
\end{aligned}$$

The final bound (2.16) follows immediately from the preceding two results.

□

2.4. Estimating $(\bar{U}^Tp)^T(\bar{U}^Tr)$. We return now to the key quantity $(\bar{U}^Tp)^T(\bar{U}r)$ that appears in (2.5), with the goal of establishing a precise form of the estimate (2.6). Throughout this section, we assume that the conditions (2.12) and (2.13) are satisfied.

We start by noting from (2.11) that

$$\bar{U}^T r = B^T [r_t]_{\Omega_t} = B^T P_{N(C^T)} B s_t = B^T P_{N(C^T)} [r_t]_{\Omega_t}$$
$$\bar{U}^T p = \bar{U}^T U_t w_t,$$

and therefore

$$(\bar{U}^T r)^T (\bar{U}^T p) = [r_t]_{\Omega_t}^T P_{N(C^T)} B \bar{U}^T U_t w. \tag{2.17}$$

By replacing $\bar{U}^T U_t$ with $V + \bar{U}^T U_t - V$ (where $V = V_t$ is the orthogonal matrix from Lemma 2.1) and manipulating, we obtain

$$\begin{aligned}
P_{N(C^T)} B \bar{U}^T U_t w &= P_{N(C^T)} B V w + P_{N(C^T)} B (\bar{U}^T U_t - V) w \\
&= P_{N(C^T)} B V (C^T C)^{-1} C^T B s + P_{N(C^T)} (BV - C) V^T (\bar{U}^T U_t - V) w \\
&= P_{N(C^T)} (BV - C)(C^T C)^{-1} C^T (C + BV - C) V^T s + \\
&\quad\ P_{N(C^T)} (BV - C) V^T (\bar{U}^T U_t - V) w \\
&= P_{N(C^T)} (BV - C)(C^T C)^{-1} C^T C V^T s + \\
&\quad\ P_{N(C^T)} (BV - C)(C^T C)^{-1} C^T (BV - C) V^T s + \\
&\quad\ P_{N(C^T)} (BV - C) V^T (\bar{U}^T U_t - V) w \\
&= P_{N(C^T)} (B - CV^T) s + P_{N(C^T)} (B - CV^T) z \\
&= [r_t]_{\Omega_t} + P_{N(C^T)} (B - CV^T) z,
\end{aligned} \tag{2.18}$$

where the final equality comes from (2.11), and we define

$$z := V (C^T C)^{-1} C^T (BV - C) V^T s + (\bar{U}^T U_t - V)(C^T C)^{-1} C^T B s.$$

From (1.3), we have

$$\|C\| \le \sqrt{\frac{3|\Omega_t|}{2n}}.$$

By using this bound, together with (1.4), $\|B\| \le \|\bar{U}\| = 1$, Lemma 2.1, and (2.12), we obtain

$$\begin{aligned}
\|z\| &\le \|(C^T C)^{-1}\|_2 \|C\|_2 \|BV - C\|_2 \|s\|_2 + \|\bar{U}^T U_t - V\|_2 \|(C^T C)^{-1}\|_2 \|C\|_2 \|B\|_2 \|s\|_2 \\
&\le \frac{2n}{|\Omega_t|} \sqrt{\frac{3|\Omega_t|}{2n}} \sqrt{2\epsilon_t} \|s\|_2 + \sqrt{2\epsilon_t} \frac{2n}{|\Omega_t|} \sqrt{\frac{3|\Omega_t|}{2n}} \|s\|_2 \\
&= 4\sqrt{3} \sqrt{\frac{n}{|\Omega_t|}} \sqrt{\epsilon_t} \|s\|_2 \le 4\sqrt{3} \sqrt{\frac{n}{q}} \sqrt{\epsilon_t} \|s\|_2.
\end{aligned}$$

By combining this bound with (2.18) and (2.17), we obtain

$$\begin{aligned}
(\bar{U}^T r)^T (\bar{U}^T p) &\ge \|r\|^2 - \|[r_t]_{\Omega_t}^T P_{N(C^T)} (B - CV^T) z\| \\
&\ge \|r\|^2 - \|r\| \|B - CV^T\| \|z\| \\
&\ge \|r\|^2 - \|r\| \sqrt{2\epsilon_t} 4\sqrt{3} \sqrt{\frac{n}{q}} \sqrt{\epsilon_t} \|s\|_2 \\
&\ge \|r\|^2 - 8\sqrt{3} \sqrt{\frac{n}{q}} \epsilon_t^{3/2} \|s\|^2,
\end{aligned} \tag{2.19}$$

where for the final inequality we used (2.14). We apply a similar argument to obtain an upper bound on $(\bar{U}^T r)^T(\bar{U}^T p)$, leading to a bound on the absolute value:

$$|(\bar{U}^T r)^T(\bar{U}^T p)| \le \|r\|^2 + 8\sqrt{3}\sqrt{\frac{n}{q}}\epsilon_t^{3/2}\|s\|^2. \tag{2.20}$$

From the bound (2.14), we have $\|s_t\|^2/\|p_t\|^2 \le (4/3)^2$, and thus from (2.19) we obtain

$$\frac{(\bar{U}^T p)^T(\bar{U}^T r)}{\|p\|^2} \ge \frac{\|r\|^2}{\|p\|^2} - 8\sqrt{3}\sqrt{\frac{n}{q}}\epsilon_t^{3/2}\frac{16}{9} = \frac{\|r\|^2}{\|p\|^2} - \frac{128\sqrt{3}}{9}\sqrt{\frac{n}{q}}\epsilon_t^{3/2}. \tag{2.21}$$

Likewise, from (2.20), we have

$$\frac{|(\bar{U}^T p)^T(\bar{U}^T r)|}{\|p\|^2} \le \frac{\|r\|^2}{\|p\|^2} + \frac{128\sqrt{3}}{9}\sqrt{\frac{n}{q}}\epsilon_t^{3/2}. \tag{2.22}$$

Using (2.16) together with the bound (2.13) on ϵ_t, we obtain from (2.22) that

$$\begin{aligned}
\frac{|(\bar{U}^T p)^T(\bar{U}^T r)|}{\|p\|^2} &\le \frac{32}{9}\epsilon_t + \frac{128\sqrt{3}}{9}\sqrt{\frac{n}{q}}\frac{1}{8\sqrt{2}}\frac{q}{n\sqrt{d}}\epsilon_t \\
&= \frac{32}{9}\epsilon_t + \frac{8\sqrt{6}}{9}\sqrt{\frac{q}{nd}}\epsilon_t \\
&\le \frac{64}{9}\epsilon_t,
\end{aligned} \tag{2.23}$$

where the last inequality follows from $q \le n$ and $d \ge 1$.

2.5. Bounding ϵ_{t+1}. We return now to the inequality (2.5), using the bounds from the previous subsection to refine our upper bound on ϵ_{t+1}. By rearranging (2.5) and substituting (2.21), (2.22), (2.16), and (2.23), we obtain

$$\begin{aligned}
\epsilon_{t+1} &\le \epsilon_t + \alpha^2\frac{\|r\|^2}{\|p\|^2} - 2\alpha\frac{(\bar{U}^T p)^T(\bar{U}^T r)}{\|p\|^2} + 2\alpha^3\frac{\|r\|^2}{\|p\|^2}\frac{|(\bar{U}^T p)^T(\bar{U}^T r)|}{\|p\|^2} \\
&\le \epsilon_t + \alpha^2\frac{\|r\|^2}{\|p\|^2} - 2\alpha\frac{\|r\|^2}{\|p\|^2} + \frac{256\sqrt{3}}{9}\alpha\sqrt{\frac{n}{q}}\epsilon_t^{3/2} + 2\alpha^3\left(\frac{32}{9}\epsilon_t\right)\left(\frac{64}{9}\epsilon_t\right).
\end{aligned}$$

By bounding $\epsilon_t^{1/2}$ using (2.13) in the last term, we obtain

$$\begin{aligned}
\epsilon_{t+1} &\le \epsilon_t - \alpha(2-\alpha)\frac{\|r\|^2}{\|p\|^2} + \frac{256\sqrt{3}}{9}\alpha\sqrt{\frac{n}{q}}\epsilon_t^{3/2} + \alpha^3\frac{64}{9\sqrt{2}}\frac{q}{n\sqrt{d}}\epsilon_t^{3/2} \\
&\le \epsilon_t - \alpha(2-\alpha)\frac{\|r\|^2}{\|p\|^2} + \left(\frac{256\sqrt{3}}{9}\alpha + \frac{64}{9\sqrt{2}}\alpha^3\right)\sqrt{\frac{n}{q}}\epsilon_t^{3/2},
\end{aligned}$$

where we used $q \le n$ and $d \ge 1$ in the second inequality. It is evident from this expression that $\alpha_t \equiv 1$ is a good choice for the steplength "fudge factor" in (1.10). By fixing $\alpha_t = 1$ and simplifying the numerical constants in the expression above, we obtain

$$\epsilon_{t+1} \le \epsilon_t - \frac{\|r\|^2}{\|p\|^2} + 55\sqrt{\frac{n}{q}}\epsilon_t^{3/2}. \tag{2.24}$$

We proceed in Subsection 2.7 to develop a high-probability lower bound on $\|r\|^2/\|p\|^2$, showing that this term is usually at least a small positive multiple of ϵ_t. Before doing this, however, it is necessary to discuss the incoherence assumptions and their consequences.

2.6. Incoherence and its Consequences. It is essential to our convergence results that the subspace $\mathcal{S}$ to be identified is *incoherent* with the coordinate directions, that is, the projection of each coordinate unit vector onto $\mathcal{S}$ should not be too long. This assumption is needed to ensure that the partially sampled observation vectors have sufficient expected information content. We make these concept precise in this subsection.

DEFINITION 2.4. *Given a matrix U of dimension $n \times d$ with orthonormal columns, defining the subspace $\mathcal{T} := R(U)$, the* coherence *of $\mathcal{T}$ is*

$$\mu(\mathcal{T}) := \frac{n}{d} \max_{i=1,2,\dots,n} \|P_{\mathcal{T}} e_i\|_2^2 ,$$

where e_i is the ith unit vector in $\mathbb{R}^n$. Note that $1 \le \mu(\mathcal{T}) \le n/d$. Since $P_{\mathcal{T}} = UU^T$, we have (with a slight change of notation)

$$\mu(U) = \frac{n}{d} \max_{i=1,2,\dots,n} \|U_{i\cdot}\|_2^2,$$

where $U_{i\cdot}$ denotes the ith row of U. As a special case of this definition, we have for a vector $x \in \mathbb{R}^n$ that

$$\mu(x) = n \frac{\|x\|_\infty^2}{\|x\|_2^2}.$$

We have the following result that relates the coherence of $R(U_t)$ to that of $R(\bar{U})$, for small values of ϵ_t.

LEMMA 2.5. *Suppose that $|\Omega_t| \ge q$, as in (2.12) and that*

$$\epsilon_t \le \frac{d}{16n} \mu(\bar{U}). \tag{2.25}$$

Then

$$\mu(U_t) \le \mu(\bar{U}) + 4\sqrt{\frac{n}{d}} \epsilon_t^{1/2} \mu(\bar{U})^{1/2} \le 2\mu(\bar{U}).$$

Proof. We have by Lemma 2.1 that

$$\|(U_t)_{i\cdot}\|_2 \le \|\bar{U}_{i\cdot}\|_2 + \|\bar{U}_{i\cdot} V_t - (U_t)_{i\cdot|}\|_2 \le \|\bar{U}_{i\cdot}\|_2 + \sqrt{2\epsilon_t}, \quad i = 1, 2, \dots, d.$$

By squaring both sides of this inequality and multiplying by n/d, we obtain

$$\frac{n}{d} \|(U_t)_{i\cdot}\|_2^2 \le \frac{n}{d} \|\bar{U}_{i\cdot}\|_2^2 + 2^{3/2} \epsilon_t^{1/2} \frac{n}{d} \|\bar{U}_{i\cdot}\|_2 + 2\frac{n}{d} \epsilon_t.$$

By taking the maxima of both sides over $i = 1, 2, \dots, d$, we have from Definition 2.4 that

$$\mu(U_t) \le \mu(\bar{U}) + 2^{3/2} \sqrt{\frac{n}{d}} \epsilon_t^{1/2} \mu(\bar{U})^{1/2} + 2\frac{n}{d} \epsilon_t.$$

By substituting the bound (2.25) into this expression, we obtain

$$\mu(U_t) \le \mu(\bar{U}) + 2^{3/2} \sqrt{\frac{n}{d}} \frac{1}{4} \sqrt{\frac{d}{n}} \mu(\bar{U}) + 2\frac{n}{d} \frac{d}{16n} \mu(\bar{U}) \le 2\mu(\bar{U}),$$

as required. □

We use coherence to analyze the key condition (1.3) that is used in the algorithm to check acceptability of the sample Ω_t. We show in the following result that the singular values of $U_\Omega^T U_\Omega$ are all approximately $|\Omega|/n$, under an assumption that relates the size of Ω_t to the coherence of U. The proof of this result appears in Appendix A.

THEOREM 2.6. *Given an $n \times d$ matrix U with orthonormal columns and a parameter $\delta > 0$, let $\Omega \subset \{1, 2, \ldots, n\}$ be chosen uniformly with replacement at random such that*

$$|\Omega| > \frac{8}{3} d\mu(U) \log\left(\frac{2d}{\delta}\right).$$

Then with probability at least $1 - \delta$, the eigenvalues of $U_\Omega^T U_\Omega$ lie in an interval

$$\lambda_i \left(U_\Omega^T U_\Omega\right) \in \left[\frac{|\Omega|}{n}(1-\gamma), \frac{|\Omega|}{n}(1+\gamma)\right], \quad i = 1, 2, \ldots, d,$$

where

$$\gamma := \sqrt{\frac{8d\mu(U)}{3|\Omega|} \log\left(\frac{2d}{\delta}\right)}. \tag{2.26}$$

We conclude this subsection with the following result, which quantifies the probability that the condition (1.3) is satisfied. We provide a specific choice for the lower bound q on sample size that is excessive for current purposes, but useful in later subsections.

COROLLARY 2.7. *Suppose that ϵ_t satisfies the bounds (2.13) and (2.25). Then condition (1.3) is satisfied with probability at least $1 - \delta$ if $|\Omega_t| \geq q$ and q satisfies*

$$\delta \geq 2d \exp\left(\frac{-3q}{64d\mu(\bar{U})}\right). \tag{2.27}$$

In particular, (2.27) is satisfied provided that the following conditions hold:

$$\delta = .1, \quad q \geq C_1 (\log n)^2 d\mu(\bar{U}) \log(20d), \quad C_1 \geq \frac{64}{3}. \tag{2.28}$$

Proof. Note first that (2.27) is equivalent to

$$\log\left(\frac{2d}{\delta}\right) \leq \frac{3q}{64d\mu(\bar{U})}. \tag{2.29}$$

Since (2.25) is assumed to hold, we can apply Lemma 2.5 to obtain

$$|\Omega_t| \geq q \geq \frac{64}{3} d\mu(\bar{U}) \log\left(\frac{2d}{\delta}\right) \geq \frac{32}{3} d\mu(U_t) \log\left(\frac{2d}{\delta}\right),$$

so that the condition of Theorem 2.6 is satisfied, with a factor of 4 to spare. By applying this theorem, we obtain

$$\gamma^2 = \frac{8d\mu(U_t)}{3|\Omega_t|} \log\left(\frac{2d}{\delta}\right) \leq \frac{16d\mu(\bar{U})}{3q} \log\left(\frac{2d}{\delta}\right) \leq \frac{1}{4}, \tag{2.30}$$

with the final inequality following from (2.27). Thus (1.3) is satisfied with probability at least $1-\delta$.

We now verify that (2.28) implies (2.27). From the inequality for q in (2.28), and the assumption that $C_1 \geq 64/3$, we have

$$q \geq C_1 (\log n)^2 d\mu(\bar{U}) \log(20d) \geq \frac{64}{3} d\mu(\bar{U}) \log(20d).$$

We obtain by rearranging this expression that

$$.1 \geq 2d \exp\left(\frac{-3q}{64 d\mu(\bar{U})}\right),$$

so that the second condition in (2.27) holds for the specific values of δ and q. Thus, we have shown that the values of δ and q in (2.28) satisfy (2.27), so that (1.3) is satisfied with probability at least .9 for these values of δ and q. ◻

2.7. A High-Probability Lower Bound on $\|r_t\|^2/\|p_t\|^2$. Theorem 2.6 can be used to derive a high-probability result for a lower bound on the quantity $\|r_t\|^2/\|p_t\|^2$, which is the key part of the the error decrease expression (2.24) and is therefore critical to our analysis.

We have the following result, which is the main result of [4] and is proved there.

LEMMA 2.8. *Let $\delta > 0$ be given, and suppose that*

$$|\Omega_t| > \frac{8}{3} d\mu(U_t) \log\left(\frac{2d}{\delta}\right). \tag{2.31}$$

Then with probability at least $1-3\delta$, we have

$$\|[v_t]_{\Omega_t} - [p_t]_{\Omega_t}\|_2^2 \geq \left(\frac{|\Omega_t|(1-\xi_t) - d\mu(U_t)\frac{(1+\beta_t)^2}{1-\gamma_t}}{n}\right) \|v_t - U_t U_t^T v_t\|_2^2, \tag{2.32}$$

where we define $x_t := v_t - U_t U_t^T v_t$, set γ_t as in (2.26), and define

$$\xi_t := \sqrt{\frac{2\mu(x_t)^2}{|\Omega_t|} \log\left(\frac{1}{\delta}\right)}, \quad \beta_t := \sqrt{2\mu(x_t) \log\left(\frac{1}{\delta}\right)}. \tag{2.33}$$

We focus now on the factor in parentheses in (2.32), proposing conditions on $\mu(x_t)$ and q under which it can be bounded below. The conditions on $\mu(x_t)$ are meant to be "realistic" in the sense that this quantity is observed to vary like $\log n$ in practice, so the upper bounds are designed to be a (possibly large) multiple of this quantity.

LEMMA 2.9. *Suppose that $\delta = .1$. Suppose that on some iteration t, we have that $|\Omega_t| \geq q$, where q and C_1 satisfy the bounds (2.28) Suppose that ϵ_t satisfies the bounds (2.13) and (2.25) for this value of q, and that $\mu(x_t)$ satisfies the following two upper bounds*

$$\mu(x_t) \leq \log n \left[\frac{.045}{\log 10} C_1 d\mu(\bar{U}) \log(20d)\right]^{1/2} \tag{2.34a}$$

$$\mu(x_t) \leq (\log n)^2 \left[\frac{.05}{8 \log 10} C_1 \log(20d)\right], \tag{2.34b}$$

where $x_t = v_t - U_t U_t^T v_t$ as in Lemma 2.8. Then we have

$$|\Omega_t|(1-\xi_t) - d\mu(U_t)\frac{(1+\beta_t)^2}{1-\gamma_t} \geq \frac{q}{2}, \tag{2.35}$$

where ξ_t, β_t, and γ_t are as defined in Lemma 2.8.

Proof. We show first that $\xi_t \leq .3$, where ξ_t is defined in (2.33). Since $|\Omega_t| \geq q$ and $\delta = .1$, we have

$$\begin{aligned}
\xi_t^2 &= \frac{2\mu(x_t)^2}{|\Omega_t|}\log\frac{1}{\delta} \\
&\leq \frac{2\mu(x_t)^2}{q}\log 10 \\
&\leq \frac{1}{q}\left[2(\log n)^2\frac{(.045)C_1}{\log 10}\log(20d)d\mu(\bar{U})\right]\log 10 \quad \text{by (2.34a)} \\
&\leq \frac{(.09)q}{q} \quad \text{by (2.28)} \\
&= .09,
\end{aligned}$$

establishing the claim.

We show next that the last term on the left-hand side of (2.35) is bounded by $.2q$. The first step is to verify that $\gamma_t \leq .5$ for γ_t defined in (2.26). When $\delta = .1$ and q and C_1 satisfy the bounds (2.28) (as we assume here), we have

$$\gamma_t^2 = \frac{8d\mu(U_t)}{3|\Omega_t|}\log\left(\frac{2d}{\delta}\right) \leq \frac{16d\mu(\bar{U})}{3|\Omega_t|}\log(20d) \leq \frac{C_1 d\mu(\bar{U})}{4q}\log(20d) \leq \frac{1}{4},$$

where we used Lemma 2.5 in the first inequality and (2.28) for the remaining inequalities.

We obtain next a bound on $(1+\beta_t)^2$. From the definition of β_t and the fact that $\mu(x_t) \geq 1$, we have

$$\beta_t = \sqrt{2\mu(x_t)\log\frac{1}{\delta}} = \sqrt{2\mu(x_t)\log 10} \geq 1,$$

so that

$$\begin{aligned}
(1+\beta_t)^2 &\leq (2\beta_t)^2 \\
&\leq 8\mu(x_t)\log 10 \\
&\leq (\log n)^2(.05)C_1\log(20d) \quad \text{by (2.34b)} \\
&\leq (.05)\frac{q}{d\mu(\bar{U})} \quad \text{by (2.28).}
\end{aligned}$$

By using these last two bounds in conjunction with Lemma 2.5, we obtain

$$d\mu(U_t)\frac{(1+\beta_t)^2}{1-\gamma_t} \leq 2d\mu(\bar{U})\frac{(.05)\frac{q}{d\mu(\bar{U})}}{.5} \leq \frac{.1}{.5}q = .2q.$$

The result (2.35) follows by combining this bound with $1-\xi_t \geq .7$ (proved earlier) and $|\Omega_t| \geq q$ (assumed). ☐

We now derive a high-probability lower bound on $\|r_t\|^2/\|p_t\|^2$.

LEMMA 2.10. *Suppose that $|\Omega_t| \geq q$ for all t, where q and C_1 satisfy the bounds (2.28). Suppose that ϵ_t satisfies the bounds (2.13) and (2.25). Suppose that there is a quantity $\bar{\delta} \in (0, .6)$ such that the bounds (2.34) are satisfied by $x_t = v_t - U_t U_t^T v_t$ with probability at least $1 - \bar{\delta}$. Then with probability at least $.6 - \bar{\delta}$, we have that*

$$\frac{\|r_t\|^2}{\|p_t\|^2} \geq (.32)\frac{q}{n}\sin^2\theta_t,$$

where θ_t is the angle between v_t and $R(U_t)$.

Proof. Since we assume (2.28), and thus that $\delta = .1$, we have from Corollary 2.7 that the check on the eigenvalues of $[U_t]_{\Omega_t}^T [U_t]_{\Omega_t}$ in (1.3) is satisfied with probability at least $1 - \delta = 1 - .1 = .9$. From Lemma 2.8, we have that (2.32) holds with probability at least $1 - 3\delta = 1 - .3 = .7$. We have assumed further that (2.34) holds with probability $1 - \bar{\delta}$. Thus, from the union bound, we have under our assumptions that the bounds (1.3), (2.32), and (2.34) all hold with probability at least $.6 - \bar{\delta}$. Since, in particular, the conditions (1.3), (2.12), and (2.13) are satisfied under this scenario, we have from Lemma 2.3 that

$$\|p_t\| \leq \frac{5}{4}\|s_t\| = \frac{5}{4}\|v_t\|.$$

By using this bound together with the definitions of $[r_t]_{\Omega_t}$ and $[r_t]_{\Omega_t^c}$ in Algorithm 1 and Lemmas 2.8 and 2.9, we obtain

$$\frac{\|r_t\|^2}{\|p_t\|^2} \geq \frac{16}{25}\frac{\|[r_t]_{\Omega_t}\|^2}{\|s_t\|^2} = \frac{16}{25}\frac{\|[v_t]_{\Omega_t} - [p_t]_{\Omega_t}\|^2}{\|v_t\|^2} \geq \frac{16}{25}\frac{q}{2n}\frac{\|v_t - U_t U_t^T v_t\|_2^2}{\|v_t\|^2}.$$

Using orthonormality of the columns of U_t and the definition of $\cos\theta_t$, we obtain

$$\frac{\|v_t - U_t U_t^T v_t\|_2^2}{\|v_t\|^2} = \frac{\|v_t\|^2 - v_t^T U_t U_t^T v_t}{\|v_t\|^2} = 1 - \frac{[v_t^T(U_t U_t^T v_t)]^2}{\|v_t\|^2\|U_t U_t^T v_t\|^2} = 1 - \cos^2\theta_t = \sin^2\theta_t.$$

We complete the proof by combining the last two expressions. ☐

2.8. Expectation for the Angle Captured by v_t. Here we obtain an expected value for the quantity $\sin^2\theta_t$, where θ_t is the angle between the (full) random sample vector and the subspace $R(U_t)$. Noting that $v_t = \bar{U}s_t$, where s_t is random, we have

$$\cos^2\theta_t = \frac{[(\bar{U}s_t)^T(U_t U_t^T \bar{U}s_t)]^2}{\|\bar{U}s_t\|^2\|U_t U_t^T \bar{U}s_t\|^2} = \frac{s_t^T \bar{U}^T U_t U_t^T \bar{U}s_t}{\|s_t\|^2}. \tag{2.36}$$

We start with two elementary technical results.

LEMMA 2.11. *Let $w \in \mathbb{R}^d$ be a random vector whose components w_i, $i = 1, 2, \ldots, d$ are independent and identically distributed. Then*

$$E_w\left(\frac{w_i^2}{\sum_{j=1}^d w_j^2}\right) = \frac{1}{d}. \tag{2.37}$$

Proof. By the additive property of expectation, we have

$$1 = E\left(\frac{\sum_{i=1}^d w_i^2}{\sum_{j=1}^d w_j^2}\right) = \sum_{i=1}^d E\left(\frac{w_i^2}{\sum_{j=1}^d w_j^2}\right) = dE\left(\frac{w_i^2}{\sum_{j=1}^d w_j^2}\right), \quad i = 1, 2, \ldots, d,$$

since each of the w_i is identically distributed. □

LEMMA 2.12. *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, $i = 1, 2, \ldots, d$ are all i.i.d. $\mathcal{N}(0, 1)$. Then*

$$E\left(\frac{w^T Q w}{w^T w}\right) = \frac{1}{d}\operatorname{trace} Q.$$

Proof.

$$\begin{aligned} E\left(\frac{w^T Q w}{w^T w}\right) &= \sum_{i\neq j} E\left(\frac{w_i w_j Q_{ij}}{\|w\|^2}\right) + \sum_{i=1}^{n} E\left(\frac{w_i^2 Q_{ii}}{\|w\|^2}\right) \\ &= \sum_{i=1}^{n} Q_{ii} E\left(\frac{w_i^2}{\|w\|^2}\right) = \frac{1}{d}\operatorname{trace} Q, \end{aligned}$$

where the second equality follows from Lemma 2.11 and the fact that $E(w_i w_j/\|w\|^2) = 0$ for $i \neq j$. □

The main result of this subsection follows.

LEMMA 2.13. *Suppose that $s_t \in \mathbb{R}^d$ is a random vector whose components are i.i.d. $\mathcal{N}(0, 1)$. Then*

$$E(\sin^2\theta_t) = \epsilon_t/d,$$

where ϵ_t is defined in (1.12).

Proof. From (2.36), Lemma 2.12, and (1.12) we have

$$E(\cos^2\theta_t) = \frac{1}{d}\operatorname{trace}(\bar{U}^T U_t U_t^T \bar{U}) = \frac{1}{d}\|U_t^T \bar{U}\|_F^2 = \frac{1}{d}(d - \epsilon_t) = 1 - \frac{\epsilon_t}{d},$$

giving the result. □

2.9. Expected Linear Decrease. We now put the pieces of theory derived in the previous subsections together, to demonstrate the expected decrease in ϵ_t over a single iteration.

THEOREM 2.14. *Suppose that $|\Omega_t| \geq q$ for all t, where q and C_1 satisfy the bounds (2.28). Suppose that ϵ_t satisfies the bounds (2.13) and (2.25). Suppose that there is a quantity $\bar{\delta} \in (0, .6)$ such that the bounds (2.34) are satisfied by $x_t = v_t - U_t U_t^T v_t$ with probability at least $1 - \bar{\delta}$. Suppose that at each iteration, s_t in (1.1) is a random vector whose components are i.i.d. $\mathcal{N}(0, 1)$. Then*

$$E[\epsilon_{t+1} \mid \epsilon_t] \leq \epsilon_t - (.32)(.6 - \bar{\delta})\frac{q}{nd}\epsilon_t + 55\sqrt{\frac{n}{q}}\epsilon_t^{3/2}. \tag{2.38}$$

Proof. Under the given assumptions, we have from (2.24) and Lemma 2.10 that

$$\epsilon_{t+1} \leq \epsilon_t - .32\frac{q}{n}\sin^2\theta_t + 55\sqrt{\frac{n}{q}}\epsilon_t^{3/2}, \quad \text{with probability at least } .6 - \bar{\delta},$$

while

$$\epsilon_{t+1} \leq \epsilon_t + 55\sqrt{\frac{n}{q}}\epsilon_t^{3/2}, \quad \text{otherwise.}$$

("Otherwise" includes iterations on which no step is taken because condition (1.3) fails to hold; we have $\epsilon_{t+1} = \epsilon_t$ on these iterations.) We obtain the proof by combining these two results and using Lemma 2.13. □

COROLLARY 2.15. *Suppose that the conditions of Theorem 2.14 hold and that in addition, ϵ_t satisfies the following bound:*

$$\epsilon_t \le (8 \times 10^{-6})(.6 - \bar{\delta})^2 \frac{q^3}{n^3 d^2}. \tag{2.39}$$

We then have

$$E[\epsilon_{t+1} \,|\, \epsilon_t] \le \left(1 - (.16)(.6 - \bar{\delta}) \frac{q}{nd}\right) \epsilon_t. \tag{2.40}$$

Proof. Given the bound (2.39), we have

$$55\sqrt{\frac{n}{q}}\epsilon_t^{1/2} \le 55\sqrt{\frac{n}{q}}(.0029)(.6 - \bar{\delta})\frac{q^{3/2}}{n^{3/2}d} \le (.16)(.6 - \bar{\delta})\frac{q}{nd}.$$

We obtain the result by combining this bound with (2.38). □

This result indicates that the rate of decrease in error metric ϵ_t is more rapid for higher values of the sampling ratio q/n, and becomes slower as subspace dimension d increases. The expected decrease in (2.40) is consistent with the factor $(1 - 1/d)$ that we prove in the next section for the full-data case ($q = n$), modulo the factor $(.16)(.6 - \bar{\delta})$. The appearance of the latter factor is of course due to the uncertainty caused by sampling.

3. The Full-Data Case: $q = n$.

When a random full vector $v_t \in \mathcal{S}$ is available at each iteration of GROUSE (that is, $\Omega_t \equiv \{1, 2, \dots, n\}$), the algorithm and its analysis simplify considerably, as we show in this section. The expected decrease factor in ϵ_t at each iteration is asymptotically $(1 - 1/d)$.

While the ISVD algorithm is preferred for this no-noise, full-data case, we note that gradient algorithms are more flexible than algorithms based explicitly on linear algebra when additional constraints or regularizers are present, such as a sparsity regularizer on the data fit or factor weights. It may be possible to build on the full-data analysis presented in this section to obtain a convergence guarantee for a Grassmannian gradient-descent algorithm on such regularized problems.

Algorithm 2 shown below is the specialization of Algorithm 1 to the full-data case. Since $(U_t)_{\Omega_t}^T (U_t)_{\Omega_t} = U_t^T U_t = I$ for all i, the eigenvalue check (1.3) is no longer needed. The definitions of certain quantities above are simplified in the full-data case, as we demonstrate here (with the introduction of notation $A_t := U_t^T \bar{U}$):

$$v_t = \bar{U} s_t, \tag{3.2a}$$

$$A_t := U_t^T \bar{U}, \tag{3.2b}$$

$$w_t = U_t^T v_t = U_t^T \bar{U} s_t = A_t s_t, \tag{3.2c}$$

$$p_t = U_t w_t, \tag{3.2d}$$

$$r_t = v_t - U_t w_t = (I - U_t U_t^T) \bar{U} s_t. \tag{3.2e}$$

We continue to use θ_t to denote the angle between $\mathcal{S}$ and $R(U_t)$ that is exposed by the update vector v_t. We have

$$\cos\theta_t = \frac{\|w_t\|}{\|v_t\|} = \frac{\|p_t\|}{\|v_t\|}, \quad \sin\theta_t = \frac{\sqrt{\|v_t\|^2 - \|w_t\|^2}}{\|v_t\|} = \frac{\|r_t\|}{\|v_t\|}. \tag{3.3}$$

Algorithm 2 GROUSE: Full Data

Given U_0 and $n \times d$ matrix with orthonormal columns, with $0 < d < n$;
for $t = 0, 1, 2, \dots$ **do**
  Take $v_t \in \mathcal{S}$;
  Define $w_t := \arg\min_w \|U_t w - v_t\|_2^2 = U_t^T v_t$;
  Define $p_t := U_t w_t$; $r_t := v_t - p_t$; $\sigma_t := \|r_t\| \, \|p_t\|$;
  Choose $\eta_t > 0$ and set

$$U_{t+1} := U_t + \left[(\cos(\sigma_t \eta_t) - 1) \frac{p_t}{\|p_t\|} + \sin(\sigma_t \eta_t) \frac{r_t}{\|r_t\|} \right] \frac{w_t^T}{\|w_t\|}. \tag{3.1}$$

end for

Thus

$$\sigma_t = \|r_t\| \|p_t\| = \|v_t\|^2 \sin\theta_t \cos\theta_t = \frac{1}{2} \|v_t\|^2 \sin 2\theta_t. \tag{3.4}$$

By using A_t defined in (3.2b) we have from (1.12) that

$$\epsilon_t = d - \|A_t\|_F^2 = d - \operatorname{trace}(A_t A_t^T). \tag{3.5}$$

Our first result provides an exact expression for the relationship between ϵ_{t+1} and ϵ_t. It also motivates an "optimal" choice for η_t consistent with the one discussed in Subsection 1.3. The proof of this result is quite technical, involving various trigonometric identities and elementary linear algebra manipulations, so we relegate it to Appendix B.

LEMMA 3.1. *We have for all t that*

$$\epsilon_t - \epsilon_{t+1} = \frac{\sin(\sigma_t \eta_t) \sin(2\theta_t - \sigma_t \eta_t)}{\sin^2 \theta_t} \left(1 - \frac{w_t^T A_t A_t^T w_t}{w_t^T w_t} \right). \tag{3.6}$$

Moreover, the right-hand side is nonnegative for $\sigma_t \eta_t \in (0, 2\theta_t)$, and zero if $v_t \in R(U_t) = \mathcal{S}_t$ or $v_t \perp \mathcal{S}_t$ (that is, $\theta_t = 0$ or $\theta_t = \pi/2$).

The expression (3.6) immediately suggests the following choice for η_t:

$$\eta_t := \frac{\theta_t}{\sigma_t} = \frac{2\theta_t}{\|v_t\|^2 \sin 2\theta_t}, \tag{3.7}$$

for which $\sin \sigma_t \eta_t = \|r_t\| / \|v_t\|$. (In the regime $\|r_t\| \ll \|p_t\|$, this choice is similar to (1.8) made for the general case, because of (1.6).) Given (3.7), (3.6) simplifies to

$$\epsilon_t - \epsilon_{t+1} = \left(1 - \frac{w_t^T A_t A_t^T w_t}{w_t^T w_t} \right) \tag{3.8}$$

We now proceed with an expected convergence analysis for the choice of η_t in (3.7), for which the convergence bound is (3.8).

THEOREM 3.2. *Suppose that in Algorithm 2, $v_t = \bar{U} s_t$, where the components of s_t are chosen i.i.d. from $\mathcal{N}(0, 1)$, and that η_t chosen as in (3.7) for each t. Suppose too that $\epsilon_t \le \bar{\epsilon}$ for some $\bar{\epsilon} \in (0, 1/3)$. Then*

$$E\left[\epsilon_{t+1} \,|\, \epsilon_t\right] \le \left(1 - \frac{1 - 3\epsilon_t}{d} \right) \epsilon_t. \tag{3.9}$$

Proof. From Lemma 2.1, using the $d \times d$ orthogonal matrices Y_t and $\bar{Y}$ defined in (2.9), we have that

$$A_t = U_t^T \bar{U} = Y_t \Gamma_t \bar{Y}^T,$$

so that

$$A_t^T A_t = \bar{Y} \Gamma_t^2 \bar{Y}^T, \quad A_t^T A_t A_t^T A_t = \bar{Y} \Gamma_t^4 \bar{Y}^T.$$

Thus for the critical term in (3.8), using $w_t = U_t^T \bar{U} s_t = A_t s_t$, dropping the subscript t freely, and recalling the definition (2.7) of Γ_t, we can write

$$\frac{w^T A A^T w}{w^T w} = \frac{s^T A^T A A^T A s}{s^T A^T A s} = \frac{\tilde{s}^T \Gamma^4 \tilde{s}}{\tilde{s}^T \Gamma^2 \tilde{s}} = \frac{\sum_{i=1}^d \tilde{s}_i^2 \cos^4 \phi_i}{\sum_{i=1}^d \tilde{s}_i^2 \cos^2 \phi_i}, \tag{3.10}$$

where $\tilde{s} = \bar{Y}^T s$ and $\phi_i = \phi_i(\bar{U}, U_t)$. Since the components of s are chosen i.i.d. from $\mathcal{N}(0,1)$, the components of $\tilde{s}$ are also i.i.d. from $\mathcal{N}(0,1)$.

We make two useful observations before proceeding. First, from the definition of ϵ_t in (1.12), we have

$$0 \le \frac{\sum_{i=1}^d \tilde{s}_i^2 \sin^2 \phi_i}{\sum_{i=1}^d \tilde{s}_i^2} \le \max_{i=1,2,\ldots,d} \sin^2 \phi_i \le \epsilon_t. \tag{3.11}$$

Second, for any scalar u with $0 \le u \le \bar{\epsilon}$, for any $\bar{\epsilon} \in [0, 1/2)$, we have

$$\frac{1}{1-u} = 1 + \frac{u}{1-u} \le 1 + 2u.$$

Returning to (3.10), dropping the indices on the summation terms for clarity, introducing the notation

$$\psi_t := \frac{\sum \tilde{s}_i^2 \sin^2 \phi_i}{\sum \tilde{s}_i^2}, \tag{3.12}$$

and noting from (3.11) that $\psi_t \in [0, \epsilon_t]$, we have

$$\begin{aligned}
\frac{\sum \tilde{s}_i^2 \cos^4 \phi_i}{\sum \tilde{s}_i^2 \cos^2 \phi_i} &= \frac{\sum \tilde{s}_i^2 [1 - 2\sin^2 \phi_i + \sin^4 \phi_i]}{\sum \tilde{s}_i^2 (1 - \sin^2 \phi_i)} \\
&\le \frac{\sum \tilde{s}_i^2 - (2 - \epsilon_t) \sum \tilde{s}_i^2 \sin^2 \phi_i}{\sum \tilde{s}_i^2 - \sum \tilde{s}_i^2 \sin^2 \phi_i} && \text{from (3.11)} \\
&= \frac{1 - (2 - \epsilon_t)\psi_t}{1 - \psi_t} && \text{from (3.12)} \\
&\le [1 - (2 - \epsilon_t)\psi_t] \left[1 + \frac{1}{1 - \epsilon_t} \psi_t\right] \\
&= 1 - \left[2 - \epsilon_t - \frac{1}{1 - \epsilon_t}\right] \psi_t - \frac{2 - \epsilon_t}{1 - \epsilon_t} \psi_t^2 \\
&\le 1 - \left[2 - \epsilon_t - \frac{1}{1 - \epsilon_t}\right] \psi_t \\
&\le 1 - [2 - \epsilon_t - (1 + 2\epsilon_t)] \psi_t \\
&= 1 - (1 - 3\epsilon_t)\psi_t.
\end{aligned}$$

We have too from Lemma 2.12 that

$$E(\psi_t) = \frac{1}{d}\sum_{i=1}^{d} \sin^2 \phi_i = \frac{\epsilon_t}{d},$$

where the expectation is taken over s_t. Assembling these results, we have that

$$E\left(1 - \frac{w_t^T A_t A_t^T w_t}{w_t^T w_t}\right) = E\left(1 - \frac{\sum \tilde{s}_i^2 \cos^4 \phi_i}{\sum \tilde{s}_i^2 \cos^2 \phi_i}\right) \geq (1 - 3\epsilon_t)E(\psi_t) = (1 - 3\epsilon_t)\frac{\epsilon_t}{d}.$$

The result follows by taking the conditional expectation of both sides in (3.8), and using the bound just derived. □

This result shows that the sequence $\{\epsilon_t\}$ converges linearly in expectation with an asymptotic rate of $(1 - 1/d)$. This rate allows for some interesting observations. First, if $d = 1$, it suggests convergence in a single step — as indeed we would expect, as the full vector $v_t \in \mathcal{S}$ would in this case reveal the solution in one step. More generally, we have from the bound

$$(1 - 1/d)^d \leq \frac{1}{e}$$

that a decrease factor of about e can be expected over each set of d consecutive iterations. By comparison, the same amount of information — d full vectors randomly drawn from $\mathcal{S}$ — is sufficient to reveal the subspace completely (with probability 1). We could obtain an orthonormal basis by assembling these d vectors into a $n \times d$ matrix and performing a singular value decomposition (SVD). (Of course, extension of an SVD-based approach to the case of partial data is not straightforward.)

4. Computational Results.

We present some computational results on random problems to illustrate the convergence behavior of GROUSE in both the partial-data and full-data cases.

For the full-data case, we implemented Algorithm 2 in Matlab on a problem for which the $n \times d$ subspace $\mathcal{S}$ was chosen randomly, as the range space of an $n \times d$ matrix whose elements are i.i.d. in $\mathcal{N}(0, 1)$. We used a random starting matrix U_0 whose columns were orthonormalized. Figure 4.1 shows results for $n = 10000$ and $d = 4$, $d = 6$. The straight line in this semilog plot (t vs $\log \epsilon_t$) represents the predicted asymptotic convergence rate $(1 - 1/d)$, while the irregular line represents the actual error. There is a close correspondence between these results and the predictions of Theorem 3.2. On early iterations, when ϵ_t is large, convergence is slower than the asymptotic rate, as predicted by the presence of the factor $1 - 3\epsilon_t$ in the expression (3.9). On later iterations, this factor approaches 1, and the asymptotic rate emerges — the curve of actual errors becomes parallel to the straight line.

For the general case, we chose various values of the dimensions n and d and the sampling cardinality q, and ran a number of trials that were constructed in the following manner. The target space $\mathcal{S}$ was defined to be the range space of an $n \times d$ matrix T whose entries were chosen i.i.d. from $\mathcal{N}(0, 1)$, and $\bar{U}$ was obtained by orthonormalizing the columns of T. To obtain a starting matrix U_0, we added to T an $n \times d$ matrix whose elements were chosen i.i.d. from $\mathcal{N}(0, 1/4)$, and orthonormalized the resulting matrix. We then generated vectors from $\bar{U}$ using Gaussian vectors s_t, and updated U_t using the GROUSE algorithm. In the computational experiments we did not check whether condition (1.3) was satisfied, and instead took every step. When $|\Omega|$ was sufficiently large, in alignment with the theory, this bound was almost

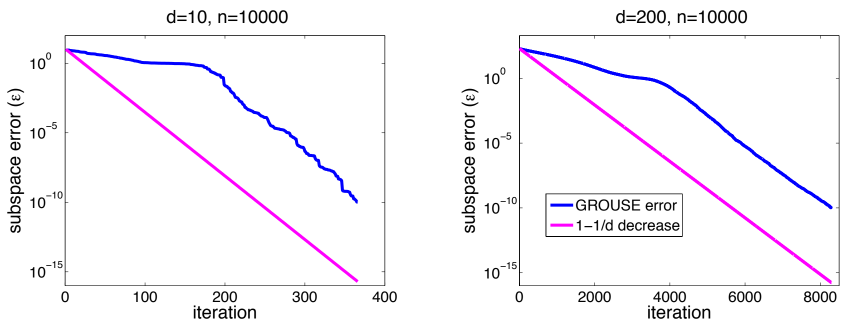

FIG. 4.1. *Illustration of convergence on Full-Data Case for $n = 10000$ with $d = 10$ (left) and $d = 200$ (right).*

always satisfied. As a specific example, for $n = 10,000$ and $|\Omega| = d \log d \log n$, out of 1000 trials, the bounds of (1.3) were satisfied 98.5% of the time for $d = 10$ and 100% of the time for $d = 100$.

After running each trial for a large enough number of iterations N to establish an asymptotic convergence rate, we computed the value X to satisfy the following expression:

$$\epsilon_N = \epsilon_0 \left(1 - X \frac{q}{nd}\right)^N , \tag{4.1}$$

By comparing with (2.40), we see that X absorbs the factor $(.16)(.6 - \bar{\delta})$ that is independent of n, d, and q, and that arises because of the errors introduced by sampling.

The values of X for various values of n, d, and q are shown in Figure 4.2. For all q larger than some modest multiple of d, X is not too far from 1, showing that the actual convergence rate is not too much different from $(1 - q/(nd))$ and that indeed the analysis is somewhat conservative.

5. Conclusion. We have analyzed the GROUSE algorithm to find that near a solution, GROUSE decreases subspace error at a linear rate, in expectation. Our estimate of the linear rate depends on the problem dimensions and the number of entries observed per vector, and matches well with our computational observations.

We believe that there are deep connections between our analysis and recent important work on randomized linear algebra (see, for example, [11]). Often this work seeks to find a low-rank approximation to a matrix or an approximation to its leading eigenspace, and randomized column or row sampling is used to make algorithms more efficient on large matrices. Our work randomly samples matrix entries instead. A deeper understanding of how these approaches are related is an interesting area for future investigation. GROUSE's connections to the ISVD algorithm, which are explored further in [5], may be a step in this direction.

Several other questions concerning GROUSE's convergence behavior warrant further investigation. We have observed empirically that convergence to a solution occurs from any random starting point, given a sufficiently large number of observed elements. This motivates us to pursue a better mathematical understanding of the global convergence properties. Moreover, we observe convergence even for coherent subspaces, and would like to understand why. Also of interest is the behavior of GROUSE in the case of noisy observations. A diminishing step size is needed here,

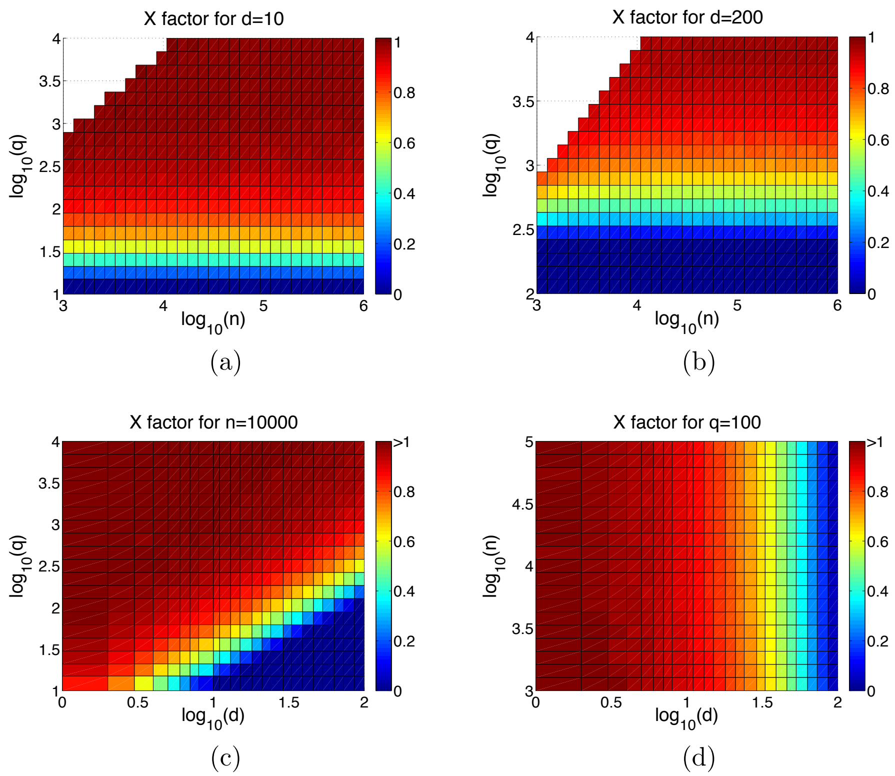

FIG. 4.2. *Observed convergence factor X for various values of d, n, and q. X is computed using (4.1), using $N = 500$ iterations of Algorithm 1 per trial. (a) $d = 10$, showing X averaged over ten trials for each n and q. (b) $d = 200$, with X plotted over one trial for each n and q. (c) $n = 10000$, with X plotted for d and q, averaged over ten trials. (d) $q = 100$, with X plotted for n and d, averaged over ten trials. In (a) and (b), the darkest red takes value 1; in (c) and (d), the darkest red is from 1 to 1.4. The plots all clearly show a phase transition in X, occurring when q is some modest multiple of d, but with X otherwise independent of q, n, and d, as suggested by the analysis.*

leading to slower convergence. We would like a better analytical understanding of this case under different noise models of interest.

Acknowledgments. We are grateful to two referees for helpful and constructive comments on the original version of this manuscript.

Appendix A. Proof of Theorem 2.6.

We start with a key result on matrix concentration.

THEOREM A.1 (Noncommutative Bernstein Inequality [9, 17]). *Let $X_1, \ldots, X_m$ be independent zero-mean square $d \times d$ random matrices. Suppose*

$$\rho_k^2 := \max\{\|\mathbb{E}[X_k X_k^T]\|_2, \|\mathbb{E}[X_k^T X_k]\|_2\}$$

and $\|X_k\|_2 \le M$ almost surely for all k. Then for any $\tau > 0$,

$$\mathbb{P}\left[\left\|\sum_{k=1}^m X_k\right\|_2 > \tau\right] \le 2d \exp\left(\frac{-\tau^2/2}{\sum_{k=1}^m \rho_k^2 + M\tau/3}\right).$$

We proceed with the proof of Theorem 2.6.

Proof. We start by defining the notation

$$u_k := U^T_{\Omega(k)\cdot} \in \mathbb{R}^d,$$

that is, u_k is the transpose of the row of the row of U that corresponds to the kth element of Ω. We thus define

$$X_k := u_k u_k^T - \frac{1}{n} I_d,$$

where I_d is the $d \times d$ identity matrix. Because of orthonormality of the columns of U, this random variable has zero mean.

To apply Theorem A.1, we must compute the values of ρ_k and M that correspond to this definition of X_k. Since $\Omega(k)$ is chosen uniformly with replacement, the X_k are distributed identically for all k, and ρ_k is independent of k (and can thus be denoted by ρ).

Using the fact that

$$\|A - B\|_2 \le \max\{\|A\|_2, \|B\|_2\} \text{ for positive semidefinite matrices } A \text{ and } B, \quad \text{(A.1)}$$

and recalling that $\|u_k\|_2^2 = \|U_{\Omega(k)\cdot}\|_2^2 \le d\mu(U)/n$, we have

$$\left\| u_k u_k^T - \frac{1}{n} I_d \right\|_2 \le \max\left\{ \frac{d\mu(U)}{n}, \frac{1}{n} \right\}.$$

Thus we can define $M := d\mu(U)/n$. For ρ, we note by symmetry of X_k that

$$\begin{aligned} \rho^2 = \left\|\mathbb{E}\left[X_k^2\right]\right\|_2 &= \left\| \mathbb{E}\left[u_k u_k^T u_k u_k^T - \frac{2}{n} u_k u_k^T + \frac{1}{n^2} I_d \right] \right\|_2 \\ &= \left\| \mathbb{E}\left[u_k (u_k^T u_k) u_k^T \right] - \frac{1}{n^2} I_d \right\|_2, \end{aligned} \quad \text{(A.2)}$$

where the last step follows from linearity of expectation, and $E(u_k u_k^T) = (1/n) I_d$.

For the next step, we define S to be the $n \times n$ diagonal matrix with diagonal elements $\|U_{i\cdot}\|_2^2$, $i = 1, 2, \dots, n$. We thus have

$$\|E[u_k(u_k^T u_k)u_k^T]\|_2 = \left\| \frac{1}{n} U^T S U \right\| \le \frac{1}{n} \|U\|_2^2 \|S\|_2 = \frac{1}{n} \frac{d\mu(U)}{n} = \frac{d\mu(U)}{n^2}.$$

Using (A.1), we have from (A.2) that

$$\rho^2 \le \max\left(\left\| \mathbb{E}\left[u_k (u_k^T u_k) u_k^T \right] \right\|, \frac{1}{n^2} \right) \le \max\left(\frac{d\mu(U)}{n^2}, \frac{1}{n^2} \right) = \frac{d\mu(U)}{n^2},$$

since $d\mu(U) \ge d \ge 1$.

We now apply Theorem A.1. First, we restrict τ to be such that $M\tau \le |\Omega|\rho^2$ to simplify the denominator of the exponent. We obtain

$$2d \exp\left(\frac{-\tau^2/2}{|\Omega|\rho^2 + M\tau/3} \right) \le 2d \exp\left(\frac{-\tau^2/2}{\frac{4}{3}|\Omega| \frac{d\mu(U)}{n^2}} \right),$$

and thus

$$\mathbb{P}\left[\left\|\sum_{k\in\Omega}\left(u_k u_k^T - \frac{1}{n}I_d\right)\right\| > \tau\right] \le 2d\exp\left(\frac{-3n^2\tau^2}{8|\Omega|d\mu(U)}\right).$$

Now take $\tau = \gamma|\Omega|/n$ with γ defined in the statement of the lemma. Since $\gamma < 1$ by assumption, $M\tau \le |\Omega|\rho^2$ holds and we have

$$\mathbb{P}\left[\left\|\sum_{k\in\Omega}\left(u_k u_k^T - \frac{1}{n}I_d\right)\right\|_2 \le \frac{|\Omega|}{n}\gamma\right] \ge 1-\delta. \tag{A.3}$$

We have, by symmetry of $\sum_{k\in\Omega} u_k u_k^T$ and the fact that

$$\lambda_i\left(\sum_{k\in\Omega} u_k u_k^T - \frac{|\Omega|}{n}I\right) = \lambda_i\left(\sum_{k\in\Omega} u_k u_k^T\right) - \frac{|\Omega|}{n},$$

that

$$\begin{aligned}\left\|\sum_{k\in\Omega}\left(u_k u_k^T - \frac{1}{n}I_d\right)\right\|_2 &= \left\|\left(\sum_{k\in\Omega} u_k u_k^T\right) - \frac{|\Omega|}{n}I_d\right\|_2 \\ &= \max_{i=1,2,\dots,n}\left|\lambda_i\left(\sum_{k\in\Omega} u_k u_k^T\right) - \frac{|\Omega|}{n}\right|,\end{aligned}$$

From (A.3), we have with probability $1-\delta$ that

$$\lambda_i\left(\sum_{k\in\Omega} u_k u_k^T\right) \in \left[(1-\gamma)\frac{|\Omega|}{n}, (1+\gamma)\frac{|\Omega|}{n}\right] \quad \text{for all } i=1,2,\dots,n,$$

completing the proof. □

Appendix B. Proof of Lemma 3.1.

We drop the subscript "t" throughout the proof and use A_+ in place of A_{t+1}. From (3.1), and using the definitions (3.2), we have

$$\begin{aligned}A_+^T &= \bar{U}^T U_+ \\ &= \bar{U}^T U + \left\{(\cos(\sigma\eta)-1)\frac{\bar{U}^T U U^T \bar{U} s}{\|w\|} + \sin(\sigma\eta)\frac{(I-\bar{U}^T U U^T \bar{U})s}{\|r\|}\right\}\frac{s^T\bar{U}^T U}{\|w\|} \\ &= \left\{I + (\cos(\sigma\eta)-1)\frac{A^T A s s^T}{\|w\|^2} + \sin(\sigma\eta)\frac{(I-A^T A)ss^T}{\|r\|\|w\|}\right\}A^T = HA^T,\end{aligned}$$

where the matrix H is defined in an obvious way. Thus

$$\|A_+\|_F^2 = \mathrm{trace}(A_+A_+^T) = \mathrm{trace}(AH^T HA^T).$$

Focusing initially on $H^T H$ we obtain

$$
\begin{aligned}
H^T H = I &+ (\cos(\sigma\eta) - 1)^2 \frac{ss^T A^T A A^T A s s^T}{\|w\|^4} \\
&+ (\cos(\sigma\eta) - 1) \frac{ss^T A^T A + A^T A s s^T}{\|w\|^2} \\
&+ \sin(\sigma\eta) \frac{2ss^T - ss^T A^T A - A^T A s s^T}{\|r\|\|w\|} \\
&+ 2\sin(\sigma\eta)(\cos(\sigma\eta) - 1) \frac{ss^T A^T A s s^T - ss^T A^T A A^T A s s^T}{\|r\|\|w\|^3} \\
&+ \sin^2(\sigma\eta) \frac{s(s^T s - 2s^T A^T A s + s^T A^T A A^T A s) s^T}{\|r\|^2 \|w\|^2}.
\end{aligned}
$$

It follows immediately that

$$
\begin{aligned}
A_+ A_+^T = AA^T &+ (\cos(\sigma\eta) - 1)^2 \frac{Ass^T A^T A A^T A s s^T A^T}{\|w\|^4} \\
&+ (\cos(\sigma\eta) - 1) \frac{Ass^T A^T A A^T + A A^T A s s^T A^T}{\|w\|^2} \\
&+ \sin(\sigma\eta) \frac{2Ass^T A^T - Ass^T A^T A A^T - A A^T A s s^T A^T}{\|r\|\|w\|} \\
&+ 2\sin(\sigma\eta)(\cos(\sigma\eta) - 1) \frac{Ass^T A^T A s s^T A^T - Ass^T A^T A A^T A s s^T A^T}{\|r\|\|w\|^3} \\
&+ \sin^2(\sigma\eta) \frac{As(s^T s - 2s^T A^T A s + s^T A^T A A^T A s) s^T A^T}{\|r\|^2 \|w\|^2}.
\end{aligned}
$$

We now use repeatedly the fact that $\operatorname{trace} ab^T = a^T b$ to deduce that

$$
\begin{aligned}
\operatorname{trace}(A_+ A_+^T) = \operatorname{trace}(AA^T) &+ (\cos(\sigma\eta) - 1)^2 \frac{(s^T A^T A s) s^T A^T A A^T A s}{\|w\|^4} \\
&+ (\cos(\sigma\eta) - 1) \frac{2s^T A^T A A^T A s}{\|w\|^2} \\
&+ \sin(\sigma\eta) \frac{2s^T A^T A s - 2s^T A^T A A^T A s}{\|r\|\|w\|} \\
&+ 2\sin(\sigma\eta)(\cos(\sigma\eta) - 1) \frac{(s^T A^T A s)^2 - (s^T A^T A A^T A s)(s^T A^T A s)}{\|r\|\|w\|^3} \\
&+ \sin^2(\sigma\eta) \frac{\|s\|^2 s^T A^T A s - 2(s^T A^T A s)^2 + (s^T A^T A A^T A s)(s^T A^T A s)}{\|r\|^2 \|w\|^2}.
\end{aligned}
$$

Now using $w = As$ (and hence $s^T A^T As = \|w\|^2$), we have

$$
\begin{aligned}
\text{trace}(A_+ A_+^T) = \text{trace}(AA^T) &+ (\cos(\sigma\eta) - 1)^2 \frac{s^T A^T AA^T As}{\|w\|^2} \\
&+ (\cos(\sigma\eta) - 1) \frac{2s^T A^T AA^T As}{\|w\|^2} \\
&+ 2\sin(\sigma\eta) \frac{\|w\|^2 - s^T A^T AA^T As}{\|r\|\|w\|} \\
&+ 2\sin(\sigma\eta)(\cos(\sigma\eta) - 1) \frac{\|w\|^2 - s^T A^T AA^T As}{\|r\|\|w\|} \\
&+ \sin^2(\sigma\eta) \frac{\|s\|^2 - 2\|w\|^2 + (s^T A^T AA^T As)}{\|r\|^2},
\end{aligned}
$$

For the second and third terms on the right-hand side, we use the identity

$$
(\cos(\sigma\eta) - 1)^2 + 2(\cos(\sigma\eta) - 1) = \cos^2(\sigma\eta) - 1 = -\sin^2(\eta\sigma),
$$

allowing us to combine these terms with the final $\sin^2(\sigma\eta)$ term. Using also the identity $\|r\|^2 = \|s\|^2 - \|w\|^2$, we obtain for the combination of these three terms that

$$
\begin{aligned}
&\sin^2(\sigma\eta) \left[1 - \frac{\|w\|^2}{\|r\|^2} + s^T A^T AA^T As \left(\frac{1}{\|r\|^2} - \frac{1}{\|w\|^2} \right) \right] \\
&\quad = \sin^2(\sigma\eta) \left(1 - \frac{\|w\|^2}{\|r\|^2} \right) \left(1 - \frac{s^T A^T AA^T As}{\|w\|^2} \right).
\end{aligned}
$$

We can also combine the third and fourth terms in the right-hand side above to yield a combined quantity

$$
2\sin(\sigma\eta)\cos(\sigma\eta) \frac{\|w\|}{\|r\|} \left(1 - \frac{s^T A^T AA^T As}{\|w\|^2} \right).
$$

By substituting these two compressed terms into the expression above, we obtain

$$
\begin{aligned}
&\text{trace}(A_+ A_+^T) = \text{trace}(AA^T) \\
&\quad + \sin(\sigma\eta) \left(1 - \frac{s^T A^T AA^T As}{\|w\|^2} \right) \left[\left(1 - \frac{\|w\|^2}{\|r\|^2} \right) \sin(\sigma\eta) + 2\cos(\sigma\eta) \frac{\|w\|}{\|r\|} \right].
\end{aligned}
$$

We now use the relations (3.3) to deduce that

$$
\frac{\|w\|}{\|r\|} = \frac{\cos\theta}{\sin\theta}, \quad 1 - \frac{\|w\|^2}{\|r\|^2} = -\frac{\cos(2\theta)}{\sin^2\theta},
$$

and thus the increment $\text{trace}(A_+ A_+^T) - \text{trace}(AA^T)$ becomes

$$
\begin{aligned}
&\sin(\sigma\eta) \left(1 - \frac{s^T A^T AA^T As}{\|w\|^2} \right) \left[-\frac{\cos(2\theta)}{\sin^2\theta} \sin(\sigma\eta) + 2\cos(\sigma\eta) \frac{\cos\theta}{\sin\theta} \right] \\
&\quad = \frac{\sin(\sigma\eta)\sin(2\theta - \sigma\eta)}{\sin^2\theta} \left(1 - \frac{s^T A^T AA^T As}{\|w\|^2} \right).
\end{aligned}
$$

The result (3.6) follows by substituting $w = As$ and (3.4).

Nonnegativity of the right-hand side follows from $\theta_t \geq 0$, $2\theta_t - \sigma_t \eta_t \geq 0$, and $\|A_t^T w_t\| \leq \|\bar{U}^T U_t\| \|w_t\| \leq \|w_t\|$.

To prove that the right-hand side of (3.6) is zero when $v_t \in \mathcal{S}$ or $v_t \perp \mathcal{S}$, we take the former case first. Here, there exists $\hat{s}_t \in \mathbb{R}^d$ such that

$$v_t = \bar{U} s_t = U_t \hat{s}_t.$$

Thus

$$w_t = A_t s_t = U_t^T \bar{U} s_t = U_t^T U_t \hat{s}_t = \hat{s}_t,$$

so that $\|v_t\| = \|w_t\|$ and thus $\theta_t = 0$, from (3.3). This implies that the right-hand side of (3.6) is zero. When $v_t \perp \mathcal{S}_t$, we have $w_t = U_t^T v_t = 0$ and so $\theta_t = \pi/2$ and $\sigma_t = 0$, implying again that the right-hand side of (3.6) is zero.

REFERENCES

[1] B. A. Ardekani, J. Kershaw, K. Kashikura, and I. Kanno, *Activation detection in functional MRI using subspace modeling and maximum likelihood estimation*, IEEE Transactions on Medical Imaging, 18 (1999).

[2] L. Balzano, *Handling Missing Data in High-Dimensional Subspace Modeling*, PhD thesis, University of Wisconsin-Madison, May 2012.

[3] L. Balzano, R. Nowak, and B. Recht, *Online identification and tracking of subspaces from highly incomplete information*, in Proceedings of Allerton, September 2010. Available at `http://arxiv.org/abs/1006.4046`.

[4] L. Balzano, B. Recht, and R. Nowak, *High-dimensional matched subspace detection when data are missing*, in Proceedings of the International Symposium on Information Theory, June 2010.

[5] L. Balzano and S. J. Wright, *On GROUSE and incremental SVD*, in Proceedings of the 5th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2013, pp. 1–4.

[6] R. Basri and D. Jacobs, *Lambertian reflectance and linear subspaces*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 25 (2003), pp. 218–233.

[7] E. Candès and J. Romberg, *Sparsity and incoherence in compressive sampling*, Inverse Problems, 23 (2007), p. 969.

[8] J. P. Costeira and T. Kanade, *A multibody factorization method for independently moving objects*, International Journal of Computer Vision, 29 (1998).

[9] D. Gross, *Recovering low-rank matrices from few coefficients in any basis*, IEEE Transactions on Information Theory, 57 (2011), pp. 1548 –1566.

[10] J. Gupchup, R. Burns, A. Terzis, and A. Szalay, *Model-based event detection in wireless sensor networks*, in Proceedings of the Workshop on Data Sharing and Interoperability (DSI), 2007.

[11] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM review, 53 (2011), pp. 217–288.

[12] H. Krim and M. Viberg, *Two decades of array signal processing research: the parametric approach*, Signal Processing Magazine, IEEE, 13 (1996), pp. 67–94.

[13] A. Lakhina, M. Crovella, and C. Diot, *Diagnosing network-wide traffic anomalies*, in Proceedings of SIGCOMM, 2004.

[14] D. Manolakis and G. Shaw, *Detection algorithms for hyperspectral imaging applications*, Signal Processing Magazine, IEEE, 19 (2002), pp. 29–43.

[15] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, second ed., 2006.

[16] S. Papadimitriou, J. Sun, and C. Faloutsos, *Streaming pattern discovery in multiple time-series*, in Proceedings of VLDB Conference, 2005.

[17] B. Recht, *A simpler approach to matrix completion*, Journal of Machine Learning Research, 12 (2011), pp. 3413–3430.

[18] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*, Computer Science and Scientific Computing, Academic Press, New York, 1990.

[19] L. Tong and S. Perreau, *Multichannel blind identification: From subspace to maximum likelihood methods*, Proceedings of the IEEE, 86 (1998), pp. 1951–1968.
[20] P. van Overschee and B. de Moor, *Subspace Identification for Linear Systems*, Kluwer Academic Publishers, Norwell, Massachusetts, 1996.
[21] L. Vandenberghe, *Convex optimization techniques in system identification*, in Proceedings of the IFAC Symposium on System Identification, July 2012.
[22] G. S. Wagner and T. J. Owens, *Signal detection using multi-channel seismic data*, Bulletin of the Seismological Society of America, 86 (1996), pp. 221–231.