Journal of Machine Learning Research 20 (2019) 1-62 Submitted 12/16; Revised 2/19; Published 2/19

Streaming Principal Component Analysis From Incomplete Data

Armin Eftekhari ARMIN.EFTEKHARI@EPFL.CH
Institute of Electrical Engineering
École Polytechnique Fédérale de Lausanne
Lausanne, VD 1015, Switzerland

Gregory Ongie GONGIE@UCHICAGO.EDU
Department of Statistics
University of Chicago
Chicago, IL 60637, USA

Laura Balzano GIRASOLE@UMICH.EDU
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109, USA

Michael B. Wakin MWAKIN@MINES.EDU
Department of Electrical Engineering
Colorado School of Mines
Golden, CO 80401, USA

Editor: Tong Zhang

Abstract

Linear subspace models are pervasive in computational sciences and particularly used for large datasets which are often incomplete due to privacy issues or sampling constraints. Therefore, a critical problem is developing an efficient algorithm for detecting low-dimensional linear structure from incomplete data efficiently, in terms of both computational complexity and storage.

In this paper we propose a streaming subspace estimation algorithm called Subspace Navigation via Interpolation from Partial Entries (SNIPE) that efficiently processes blocks of incomplete data to estimate the underlying subspace model. In every iteration, SNIPE finds the subspace that best fits the new data block but remains close to the previous estimate. We show that SNIPE is a streaming solver for the underlying nonconvex matrix completion problem, that it converges globally to a stationary point of this program regardless of initialization, and that the convergence is locally linear with high probability. We also find that SNIPE shows state-of-the-art performance in our numerical simulations.

Keywords: Principal component analysis, Subspace identification, Matrix completion, Streaming algorithms, Nonconvex optimization, Global convergence

1. Introduction

Linear models are the backbone of computational science, and Principal Component Analysis (PCA) in particular is an indispensable tool for detecting linear structure in collected data (van Overschee and de Moor, 2012; Ardekani et al., 1999; Krim and Viberg, 1996; Tong

Figure 1: The sequence of generic vectors $\{s_t\}_{t=1}^T$ drawn from an unknown r-dimensional subspace $\mathcal{S}$ in panel (a) is only partially observed on random index sets $\{\omega_t\}_{t=1}^T$. That is, we only have access to incomplete data vectors $\{y_t\}_{t=1}^T$ in panel (b), where the white entries are missing. Our objective is to estimate the subspace $\mathcal{S}$ from the incomplete data vectors, when limited storage and processing resources are available. See Section 1 for the detailed setup.

and Perreau, 1998). Principal components of a dataset are used, for example, to perform linear dimensionality reduction, which is in turn at the heart of classification, regression and other learning tasks that often suffer from the "curse of dimensionality", where having a small number of training samples in relation to the number of features typically leads to overfitting (Hastie et al., 2013).

In this work, we are particularly interested in applying PCA to data that is presented sequentially to a user, with limited processing time available for each item. Moreover, due to hardware limitations, we assume the user can only store small amounts of data. Finally, we also consider the possibility that the incoming data is incomplete, either due to physical sampling constraints, or deliberately subsampled to facilitate faster processing times or to address privacy concerns.

As one example, consider monitoring network traffic over time, where acquiring complete network measurements at fine time-scales is impractical and subsampling is necessary (Lakhina et al., 2004; Gershenfeld et al., 2010). As another example, suppose we have a network of cheap, battery-powered sensors that must relay summary statistics of their measurements, say their principal components, to a central node on a daily basis. Each sensor cannot store or process all its daily measurements locally, nor does it have the power to relay all the raw data to the central node. Moreover, many measurements are not reliable and can be treated as missing. It is in this and similar contexts that we hope to develop a *streaming* algorithm for PCA from incomplete data.

More formally, we consider the following problem: Let $\mathcal{S}$ be an r-dimensional subspace with orthonormal basis $S \in \mathbb{R}^{n\times r}$. For an integer T, let the coefficient vectors $\{q_t\}_{t=1}^T \subset \mathbb{R}^r$ be independent copies of a random vector $q \in \mathbb{R}^r$ with bounded expectation, namely, $\mathbb{E}\|q\|_2 < \infty$. Consider the sequence of vectors $\{Sq_t\}_{t=1}^T \subset \mathcal{S}$ and set $s_t := Sq_t$ for short. At each time $t \in [1:T] := \{1, 2, \cdots, T\}$, we observe each entry of s_t independently with a probability of p and collect the observed entries in $y_t \in \mathbb{R}^n$. Formally, we let $\omega_t \subseteq [1:n]$ be the random index set over which s_t is observed and write this measurement process as $y_t = P_{\omega_t} \cdot s_t$, where $P_{\omega_t} \in \mathbb{R}^{n\times n}$ is the projection onto the coordinate set ω_t, namely, it equals one on its diagonal entries corresponding to the index set ω_t and is zero elsewhere.

Our objective in this paper is to design a streaming algorithm to identify the subspace $\mathcal{S}$ from the incomplete data $\{y_t\}_{t=1}^T$ supported on the index sets $\{\omega_t\}_{t=1}^T$. Put differently, our objective is to design a streaming algorithm to compute leading r principal components of the full (but hidden) data matrix $[s_1\, s_2 \cdots s_T]$ from the incomplete observations $[y_1\, y_2\, \cdots y_T]$, see Figure 1. By the Eckart-Young-Mirsky Theorem, this task is equivalent to computing leading r left singular vectors of the full data matrix from its partial observations (Eckart and Young, 1936; Mirsky, 1966).

Assuming that $r = \dim(\mathcal{S})$ is known *a priori* (or estimated from data by other means), we present the SNIPE algorithm for this task in Section 2. SNIPE is designed based on the *principle of least-change* and, in every iteration, finds the subspace that best fits the new data block but remains close to the previous estimate. SNIPE requires $O(n)$ bits of memory and performs $O(n)$ flops in every iteration, which is optimal in its dependence on the ambient dimension n. As discussed in Section 3, SNIPE has a natural interpretation as a streaming algorithm for low-rank matrix completion (Davenport and Romberg, 2016).

Section 4 discusses the global and local convergence of SNIPE. In particular, the local convergence rate is linear near the true subspace, namely, the estimation error of SNIPE reduces by a factor of $1-cp$ in every iteration, for a certain factor c and with high probability. This local convergence guarantee for SNIPE is a key technical contribution of this paper which is absent in its close competitors, see Section 5.

Even though we limit ourselves to the "noise-free" case of $y_t = P_{\omega_t} s_t$ in this paper, SNIPE can also be applied (after minimal changes) when $y_t = P_{\omega_t}(s_t + n_t)$, where we might think of $n_t \in \mathbb{R}^n$ as measurement noise from a signal processing viewpoint. Alternatively from a statistical viewpoint, n_t represents the tail of the covariance matrix of the generative model, from which $\{s_t\}_t$ are drawn. Moreover, SNIPE can be easily adapted to the *dynamic* case where the underlying subspace $\mathcal{S} = \mathcal{S}(t)$ changes over time. We leave the convergence analysis of SNIPE under a noisy time-variant model to a future work. Similarly, entries of incoming vectors are observed uniformly at random in our theoretical analysis but SNIPE also applies to any incomplete data.

A review of prior art is presented in Section 5, and the performance of SNIPE and rival algorithms are examined numerically in Section 6, where we find that SNIPE shows the state-of-the-art performance. Technical proofs appear in Section 7 and in the appendices, with Appendix A (Toolbox) collecting some of the frequently-used mathematical tools. Finally, Appendix L offers an alternative initialization for SNIPE.

2. SNIPE

In this section, we present Subspace Navigation via Interpolation from Partial Entries (SNIPE), a streaming algorithm for subspace identification from incomplete data, received sequentially.

Let us first introduce some additional notation. Recall that we denote the incoming sequence of incomplete vectors by $\{y_t\}_{t=1}^T \subset \mathbb{R}^n$, which are supported on index sets $\{\omega_t\}_{t=1}^T \subseteq [1:n]$. For a block size $b \geq r$, we concatenate every b consecutive vectors into a data block, thereby partitioning the incoming data into $K = T/b$ non-overlapping blocks $\{Y_k\}_{k=1}^K$, where $Y_k \in \mathbb{R}^{n\times b}$ for every k. We assume for convenience that K is an integer. We also often take $b = O(r)$ to maximize the efficiency of SNIPE, as discussed below.

At a high level, SNIPE processes the first incomplete block Y_1 to produce an estimate $\widehat{\mathcal{S}}_1$ of the true subspace $\mathcal{S}$. This estimate is then iteratively updated after receiving each of the new incomplete blocks $\{Y_k\}_{k=2}^K$, thereby producing a sequence of estimates $\{\widehat{\mathcal{S}}_k\}_{k=2}^K$, see Figure 2. Every $\widehat{\mathcal{S}}_k$ is an r-dimensional subspace of $\mathbb{R}^n$ with orthonormal basis $\widehat{S}_k \in \mathbb{R}^{n\times r}$; the particular choice of orthonormal basis is inconsequential throughout the paper.

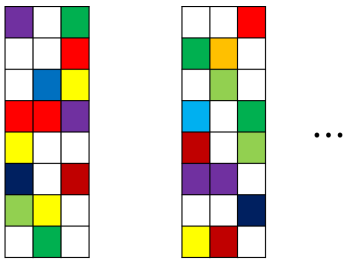


Figure 2: SNIPE concatenates every b incoming vectors into a block and iteratively updates its estimate of the true subspace $\mathcal{S}$ after receiving each new block. That is, SNIPE updates its estimate of $\mathcal{S}$, from $\widehat{\mathcal{S}}_{k-1}$ to $\widehat{\mathcal{S}}_k$, after receiving the incomplete data block $Y_k \in \mathbb{R}^{n\times b}$, see Section 2 for the details.

More concretely, SNIPE sets $\widehat{\mathcal{S}}_1$ to be the span of leading r left singular vectors of Y_1, namely, the left singular vectors corresponding to largest r singular values of Y_1, with ties broken arbitrarily. Then, at iteration $k \in [2 : K]$ and given the previous estimate $\widehat{\mathcal{S}}_{k-1} = \text{span}(\widehat{S}_{k-1})$, SNIPE processes the columns of the kth incomplete block Y_k one by one and forms the matrix

$$R_k = \left[\cdots \quad y_t + P_{\omega_t^C} \widehat{S}_{k-1} \left(\widehat{S}_{k-1}^* P_{\omega_t} \widehat{S}_{k-1} + \lambda I_r \right)^\dagger \widehat{S}_{k-1}^* y_t \quad \cdots \right] \in \mathbb{R}^{n\times b}, \qquad t \in [(k-1)b + 1 : kb], \tag{1}$$

where $\dagger$ denotes the pseudo-inverse and $\lambda \geq 0$ is a parameter. Above, $P_{\omega_t^C} = I_n - P_{\omega_t} \in \mathbb{R}^{n\times n}$ projects a vector onto the complement of the index set ω_t. The motivation for the particular choice of R_k above will become clear in Section 3. SNIPE then updates its estimate by setting $\widehat{\mathcal{S}}_k$ to be the span of leading r left singular vectors of R_k. Algorithm 1 summarizes these steps. Note that Algorithm 1 rejects ill-conditioned updates in Step 3 for the convenience of analysis and that similar reject options have precedence in the literature (Balzano and Wright, 2015). We however found implementing this reject option to be unnecessary in numerical simulations.

Remark 1 [Computational complexity of SNIPE] We measure the *per-iteration* algorithmic complexity of SNIPE by calculating the average number of floating-point operations (flops) performed on an incoming vector. Every iteration of SNIPE involves finding leading r left singular vectors of an $n \times b$ matrix. Assuming that $b = O(r)$, this could be done with $O(nr^2)$ flops. At the kth iteration with $k \geq 2$, SNIPE also requires finding the pseudo-inverse of $P_{\omega_j} \widehat{S}_{k-1} \in \mathbb{R}^{n\times r}$ for each incoming vector which costs $O(nr^2)$ flops. Therefore the overall computational complexity of SNIPE is $O(nr^2)$ flops per vector. As further discussed in Section 5, this matches the complexity of other algorithms for streaming PCA even though here the received data is highly incomplete.

Remark 2 [Storage requirements of SNIPE] We measure the storage required by SNIPE by calculating the number of memory elements stored by SNIPE at any given instant. At the kth iteration, SNIPE must store the current estimate $\widehat{S}_{k-1} \in \mathbb{R}^{n\times r}$ (if available) and the new incomplete block $Y_k \in \mathbb{R}^{n\times b}$. Assuming that $b = O(r)$, this translates into $O(nr) + O(pnr) = O(nr)$ memory elements. SNIPE therefore requires $O(nr)$ bits of storage, which is optimal up to a constant factor.

Algorithm 1 SNIPE for streaming PCA from incomplete data

Input:

- Dimension r,
- Received data $\{y_t\}_{t=1}^T \subset \mathbb{R}^n$ supported on index sets $\{\omega_t\}_{t=1}^T \subseteq [1:n]$, presented sequentially in $K = T/b$ blocks of size $b \geq r$,
- Tuning parameter $\lambda \geq 0$,
- Reject thresholds $\sigma_{\min}, \tau \geq 0$.

Output:

- r-dimensional subspace $\widehat{\mathcal{S}}_K$.

Body:

- Form $Y_1 \in \mathbb{R}^{n\times b}$ by concatenating the first b received vectors $\{y_t\}_{t=1}^b$. Let $\widehat{\mathcal{S}}_1$, with orthonormal basis $\widehat{S}_1 \in \mathbb{R}^{n\times r}$, be the span of leading r left singular vectors of Y_1, namely, those corresponding to r largest singular values. Ties are broken arbitrarily.
- For $k \in [2:K]$, repeat:
  1. Set $R_k \leftarrow \{\}$.
  2. For $t \in [(k-1)b+1 : kb]$, repeat
     - Set

$$R_k \leftarrow \begin{bmatrix} R_k & y_t + P_{\omega_t^C}\widehat{S}_{k-1}\left(\widehat{S}_{k-1}^* P_{\omega_t}\widehat{S}_{k-1} + \lambda I_r\right)^\dagger \widehat{S}_{k-1}^* y_t \end{bmatrix},$$

       where $P_{\omega_t} \in \mathbb{R}^{n\times n}$ equals one on its diagonal entries corresponding to the index set ω_t, and is zero elsewhere. Likewise, $P_{\omega_t^C}$ projects a vector onto the complement of the index set ω_t.
  3. If $\sigma_r(R_k) < \sigma_{\min}$ or $\sigma_r(R_k) \leq (1+\tau)\cdot\sigma_{r+1}(R_k)$, then set $\widehat{\mathcal{S}}_k \leftarrow \widehat{\mathcal{S}}_{k-1}$. Otherwise, let $\widehat{\mathcal{S}}_k$, with orthonormal basis $\widehat{S}_k \in \mathbb{R}^{n\times r}$, be the span of leading r left singular vectors of R_k. Ties are broken arbitrarily. Here, $\sigma_i(R_k)$ is the ith largest singular value of R_k.
- Return $\widehat{\mathcal{S}}_K$.

3. Interpretation of SNIPE

SNIPE has a natural interpretation as a streaming algorithm for low-rank matrix completion, which we now discuss. First let us enrich our notation. Recall the incomplete data blocks $\{Y_k\}_{k=1}^K \subset \mathbb{R}^{n\times b}$ and let the random index set $\Omega_k \subseteq [1:n]\times[1:b]$ be the support of Y_k for every k. We write that $Y_k = P_{\Omega_k}(S_k)$ for every k, where the complete (but hidden) data block $S_k \in \mathbb{R}^{n\times b}$ is formed by concatenating $\{s_t\}_{t=(k-1)b+1}^{kb}$. Here, $P_{\Omega_k}(S_k)$ retains only the entries of S_k on the index set Ω_k, setting the rest to zero. By design, $s_t = Sq_t$ for every t and we may therefore write that $S_k = S\cdot Q_k$ for the coefficient matrix $Q_k \in \mathbb{R}^{r\times b}$ formed by concatenating $\{q_t\}_{t=(k-1)b+1}^{kb}$. To summarize, $\{S_k\}_{k=1}^K$ is formed by partitioning $\{s_t\}_{t=1}^T$ into K blocks. Likewise, $\{Q_k, Y_k, \Omega_k\}_{k=1}^K$ are formed by partitioning $\{q_t, y_t, \omega_t\}_{t=1}^T$, respectively.

With this introduction, let us form $Y \in \mathbb{R}^{n\times T}$ by concatenating the incomplete blocks $\{Y_k\}_{k=1}^K \subset \mathbb{R}^{n\times b}$, supported on the index sets $\Omega \subseteq [1:n]\times[1:T]$. To find the true subspace $\mathcal{S}$, one might consider solving

$$\begin{cases} \min\limits_{X,\mathcal{U}} \ \|P_{\mathcal{U}^\perp} X\|_F^2 + \lambda \|P_{\Omega^C}(X)\|_F^2 \\ P_\Omega(X) = Y, \end{cases} \tag{2}$$

where the minimization is over a matrix $X \in \mathbb{R}^{n\times T}$ and r-dimensional subspace $\mathcal{U} \subset \mathbb{R}^n$. Above, $P_{\mathcal{U}^\perp} \in \mathbb{R}^{n\times n}$ is the orthogonal projection onto the orthogonal complement of subspace $\mathcal{U}$ and $P_\Omega(X)$ retains only the entries of X on the index set Ω, setting the rest to zero. Note that Program (2) encourages its solution(s) to be low-rank while matching the observations Y on the index set Ω. The term $\lambda\|P_{\Omega^C}(X)\|_F^2$ for $\lambda \geq 0$ is the *Tikhonov regularizer* that controls the energy of solution(s) on the complement of index set Ω.

With complete data, namely, when $\Omega = [1:n]\times[1:T]$, Program (2) reduces to PCA, as it returns $X = Y$ and searches for an r-dimensional subspace that captures most of the energy of Y. That is, Program (2) reduces to $\min_{\mathcal{U}} \|P_{\mathcal{U}^\perp} Y\|_F^2$ when $\Omega = [1:n]\times[1:T]$, solution of which is the span of leading r left singular vectors of Y in light of the Eckart-Young-Mirsky Theorem (Eckart and Young, 1936; Mirsky, 1966). In this sense then, Program (2) performs PCA from incomplete data. Note crucially that Program (2) is a nonconvex problem because the Grassmannian $\mathbb{G}(n,r)$, the set of all r-dimensional subspaces in $\mathbb{R}^n$, is a nonconvex set.[1] However, given a fixed subspace $\mathcal{U} \in \mathbb{G}(n,r)$, Program (2) reduces to the simple least-squares program

$$\begin{cases} \min\limits_{X} \ \|P_{\mathcal{U}^\perp} X\|_F^2 + \lambda \|P_{\Omega^C}(X)\|_F^2 \\ P_\Omega(X) = Y, \end{cases} \tag{3}$$

where the minimization is over $X \in \mathbb{R}^{n\times T}$. If in addition λ is positive, then Program (3) is strongly convex and has a unique minimizer. Given a fixed feasible $X \in \mathbb{R}^{n\times T}$, Program (2) has the same minimizers as

$$\min_{\mathcal{U}\in\mathbb{G}(n,r)} \|P_{\mathcal{U}^\perp} X\|_F^2. \tag{4}$$

1. The Grassmannian can be embedded in $\mathbb{R}^{n\times n}$ via the map that takes $\mathcal{U} \in \mathbb{G}(n,r)$ to the corresponding orthogonal projection $P_{\mathcal{U}} \in \mathbb{R}^{n\times n}$. The resulting submanifold of $\mathbb{R}^{n\times n}$ is a nonconvex set.

That is, for a fixed feasible X, Program (2) simply performs PCA on X. We might also view Program (2) from a matrix completion perspective. More specifically, let

$$\rho_r^2(X) = \sum_{i \geq r+1} \sigma_i^2(X), \tag{5}$$

be the *residual* of X, namely, the energy of its trailing singular values $\sigma_{r+1}(X) \geq \sigma_{r+2}(X) \geq \cdots$. Like the popular nuclear norm $\|X\|_* = \sum_{i \geq 1} \sigma_i(X)$ in (Davenport and Romberg, 2016), the residual $\rho_r(X)$ gauges the rank of X. In particular, $\rho_r(X) = 0$ if and only if $\text{rank}(X) \leq r$. Unlike the nuclear norm, however, the residual is still a nonconvex function of X. We now rewrite Program (2) as

$$\begin{aligned} \begin{cases} \min\limits_{X,\mathcal{U}} \ \|P_{\mathcal{U}^\perp} X\|_F^2 + \lambda \|P_{\Omega^C}(X)\|_F^2 \\ P_\Omega(X) = Y \end{cases} &= \begin{cases} \min\limits_{X} \min\limits_{\mathcal{U} \in \mathbb{G}(n,r)} \ \|P_{\mathcal{U}^\perp} X\|_F^2 + \lambda \|P_{\Omega^C}(X)\|_F^2 \\ P_\Omega(X) = Y \end{cases} \\ &= \begin{cases} \min\limits_{X} \ \rho_r^2(X) + \lambda \|P_{\Omega^C}(X)\|_F^2 \\ P_\Omega(X) = Y. \end{cases} \end{aligned} \tag{6}$$

That is, if we ignore the regularization term $\lambda \|P_{\Omega^C}(X)\|_F^2$, Program (2) searches for a matrix with the least residual, as a proxy for least rank, that matches the observations Y. In this sense then, Program (2) is a "relaxation" of the low-rank matrix completion problem. Several other formulations for the matrix completion problem are reviewed in (Davenport and Romberg, 2016; Eftekhari et al., 2018b,a). We can also rewrite Program (2) in terms of its data blocks by considering the equivalent program

$$\begin{cases} \min \ \sum_{k=1}^K \|P_{\mathcal{U}}^\perp X_k\|_F^2 + \lambda \|P_{\Omega_k^C}(X_k)\|_F^2 \\ P_{\Omega_k}(X_k) = Y_k \qquad k \in [1:K], \end{cases} \tag{7}$$

where the minimization is over matrices $\{X_k\}_{k=1}^K \subset \mathbb{R}^{n \times b}$ and subspace $\mathcal{U} \in \mathbb{G}(n,r)$. Let us additionally introduce a number of auxiliary variables into Program (7) by considering the equivalent program

$$\begin{cases} \min \ \sum_{k=1}^K \|P_{\mathcal{U}_k}^\perp X_k\|_F^2 + \lambda \|P_{\Omega_k^C}(X_k)\|_F^2 \\ P_{\Omega_k}(X_k) = Y_k \qquad k \in [1:K] \\ \mathcal{U}_1 = \mathcal{U}_2 = \cdots = \mathcal{U}_K, \end{cases} \tag{8}$$

where the minimization is over matrices $\{X_k\}_{k=1}^K \subset \mathbb{R}^{n \times b}$ and subspaces $\{\mathcal{U}_k\}_{k=1}^K \subset \mathbb{G}(n,r)$. Indeed, Programs (2,7,8) are all equivalent and all nonconvex. Now consider the following approximate solver for Program (8) that alternatively solves for matrices and subspaces:

- Setting $X_1 = Y_1$ in Program (8), we minimize $\|P_{\mathcal{U}_1}^\perp Y_1\|_F^2$ over $\mathcal{U}_1 \in \mathbb{G}(n,r)$ and, by the Eckart-Young-Mirsky Theorem, find a minimizer to be the span of leading r left singular vectors of Y_1, which coincides with $\widehat{\mathcal{S}}_1$ in SNIPE.
- For $k \in [2:K]$, repeat:

– Setting $\mathcal{U}_k = \widehat{\mathcal{S}}_{k-1}$ in Program (8), we solve

$$\begin{cases} \min_{X_k} \ \|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} X_k\|_F^2 + \lambda \|P_{\Omega_k^C}(X_k)\|_F^2 \\ P_{\Omega_k}(X_k) = Y_k, \end{cases} \tag{9}$$

over matrix $X_k \in \mathbb{R}^{n\times b}$. We verify in Appendix B that the minimizer of Program (9) coincides with R_k in SNIPE, see (1).

– If $\sigma_r(R_k) < \sigma_{\min}$ or $\sigma_r(R_k) \leq (1+\tau)\sigma_{r+1}(R_k)$, then no update is made, namely, we set $\widehat{\mathcal{S}}_k = \widehat{\mathcal{S}}_{k-1}$. Otherwise, setting $X_k = R_k$ in Program (8), we solve $\min \|P_{\mathcal{U}_k}^{\perp} R_k\|_F^2$ over $\mathcal{U}_k \in \mathbb{G}(n,r)$ to find $\widehat{\mathcal{S}}_k$. That is, by the Eckart-Young-Mirsky Theorem again, $\widehat{\mathcal{S}}_k$ is the span of leading r left singular vectors of R_k. The output of this step matches $\widehat{\mathcal{S}}_k$ produced in SNIPE.

To summarize, following the above procedure produces $\{R_k\}_{k=1}^K$ and $\{\widehat{\mathcal{S}}_k\}_{k=1}^K$ in SNIPE. In other words, we might think of SNIPE as an approximate solver for Program (2), namely, SNIPE is a streaming algorithm for low-rank matrix completion. In fact, the output of SNIPE always converges to a stationary point of Program (2) in the sense described in Section 4.

Another insight about the choice of R_k in (1) is as follows. Let us set $\lambda = 0$ for simplicity. At the beginning of the kth iteration of SNIPE with $k \geq 2$, the available estimate of the true subspace is $\widehat{\mathcal{S}}_{k-1}$ with orthonormal basis $\widehat{S}_{k-1}$. Given a new incomplete vector $y \in \mathbb{R}^n$, supported on the index set $\omega \subseteq [1:n]$, $z = \widehat{S}_{k-1}(P_\omega \widehat{S}_{k-1})^\dagger y$ best approximates y in $\widehat{\mathcal{S}}_{k-1}$ in ℓ_2 sense. In order to agree with the measurements, we minimally adjust this to $y + P_{\omega^C} z$, where P_{ω^C} projects onto the complement of index set ω. This indeed matches the expression for the columns of R_k in SNIPE. We note that this type of least-change strategy has been successfully used in the development of quasi-Newton methods for optimization (Nocedal and Wright, 2006, Chapter 6).

4. Performance of SNIPE

To measure the performance of SNIPE—whose output is a subspace—we naturally use principal angles as an error metric. More specifically, recall that $\mathcal{S}$ and $\widehat{\mathcal{S}}_K$ denote the true subspace and the output of SNIPE, respectively. Then the ith largest singular value of $P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_K}$ equals $\sin(\theta_i(\mathcal{S}, \widehat{\mathcal{S}}_K))$, where

$$\theta_1(\mathcal{S}, \widehat{\mathcal{S}}_K) \geq \theta_2(\mathcal{S}, \widehat{\mathcal{S}}_K) \geq \cdots \geq \theta_r(\mathcal{S}, \widehat{\mathcal{S}}_K)$$

denote the principal angles between the two r-dimensional subspaces $\mathcal{S}, \widehat{\mathcal{S}}_K$ (Golub and Van Loan, 2013). The estimation error of SNIPE is then

$$d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_K) := \sqrt{\frac{1}{r}\sum_{i=1}^{r} \sin^2\left(\theta_i(\mathcal{S}, \widehat{\mathcal{S}}_K)\right)} = \frac{\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_K}\|_F}{\sqrt{r}}, \tag{10}$$

which also naturally induces a metric topology on the Grassmannian $\mathbb{G}(n,r)$.[2] Note also that we will always reserve calligraphic letters for subspaces and capital letters for their orthonormal bases, for example subspace $\mathcal{S}$ and its orthonormal basis S.

Our first result loosely speaking states that a subsequence of SNIPE converges to a stationary point of the nonconvex Program (2) as T goes to infinity, see Section 7.1 for the proof.

Theorem 3 [Global convergence] *Consider an r-dimensional subspace $\mathcal{S}$ with orthonormal basis $S \in \mathbb{R}^{n\times r}$. For an integer T, let the coefficient vectors $\{q_t\}_{t=1}^T \subset \mathbb{R}^r$ be independent copies of a random vector $q \in \mathbb{R}^r$ with bounded expectation, namely,, $\mathbb{E}\|q\|_2 < \infty$. For every $t \in [1:T]$, we observe each coordinate of $s_t = Sq_t \in \mathcal{S}$ independently with a probability of p and collect the observations in $y_t \in \mathbb{R}^n$, supported on a random index set $\omega_t \subseteq [1:n]$. Fix positive λ, block size $b \geq r$, positive reject thresholds $\sigma_{\min}, \tau$, and consider the output sequence of* SNIPE *in Algorithm 1, namely, $\{(R_k, \widehat{\mathcal{S}}_k)\}_k$. Also by partitioning $\{q_t, s_t, y_t, \omega_t\}_t$, form the coefficient blocks $\{Q_k\}_k \subset \mathbb{R}^{r\times b}$, data blocks $\{S_k\}_k \subset \mathbb{R}^{n\times b}$, and incomplete data blocks $\{Y_k\}_k \subset \mathbb{R}^{n\times b}$ supported on index sets $\{\Omega_k\}_k \subseteq [1:n]\times[1:b]$, as described in Section 3.*

For every integer l, there exists an integer k_l, for which the following asymptotic statement is almost surely true as $T \to \infty$. Consider the restriction of Program (2) to iteration k_l, namely,

$$\begin{cases} \min_{X,\mathcal{U}} \ \ \|P_{\mathcal{U}^\perp} X\|_F^2 + \lambda \|P_{\Omega_{k_l}^C}(X)\|_F^2 \\ P_{\Omega_{k_l}}(X) = Y_{k_l}, \end{cases} \tag{11}$$

where the minimization is over matrix $X \in \mathbb{R}^{n\times b}$ and r-dimensional subspace $\mathcal{U}$. Then there exists $R \in \mathbb{R}^{n\times b}$ and r-dimensional subspace $\widehat{\mathcal{S}}$ such that

- *$\widehat{S}$ is the span of leading r left singular vectors of R,*
- *$(R, \widehat{\mathcal{S}})$ is a stationary pair of Program (11), namely, it satisfies the first-order optimality conditions of Program (11),*
- $\lim_{l\to\infty} \|R_{k_l} - R\|_F = 0$,
- $\lim_{l\to\infty} d_{\mathbb{G}}(\widehat{\mathcal{S}}_{k_l}, \mathcal{S}) = 0$.

Remark 4 [Discussion of Theorem 3] Theorem 3 roughly speaking states that there is a subsequence of SNIPE that converges to a stationary point of Program (2), which was the program designed in Section 3 for PCA from incomplete data or, from a different perspective, for low-rank matrix completion. Theorem 3 is however silent about the nature of this stationary point, whether it is a local or global minimizer/maximizer, or a saddle point. To some extent, this question is addressed below in Proposition 6.

More generally, we have been able to show that this stationary point is in fact rank-r. When the block size of SNIPE is sufficiently large, namely when $b = \Omega(n)$, we can further establish that the limit point of SNIPE is indeed a global minimizer of Program (2)

2. Another possible error metric is simply the largest principal angle $\theta_1(\mathcal{S}, \widehat{\mathcal{S}}_K)$. The two metrics are very closely related: $\theta_1(\mathcal{S}, \widehat{\mathcal{S}}_K)/\sqrt{r} \leq d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_K) \leq \theta_1(\mathcal{S}, \widehat{\mathcal{S}}_K)$. However, we find that $\theta_1(\mathcal{S}, \widehat{\mathcal{S}}_K)$ is not amenable for analysis of our problem, as opposed to $d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_K)$.

and moreover SNIPE recovers the true subspace, namely, $\lim_{l\to\infty} d(\widehat{\mathcal{S}}_{k_l}, \mathcal{S}) = 0$, with high probability and under certain standard conditions on the *coherence* of the true subspace $\mathcal{S}$ and sampling probability p. We have not included these results here because SNIPE is intended as a streaming algorithm and we are therefore more interested in the setting where $b = O(r)$, see Remarks 1 and 2 about the implementation of SNIPE. It is not currently clear to us when SNIPE converges in general but, as suggested by Proposition 6 below, if SNIPE converges, it does indeed converge to the true subspace $\mathcal{S}$.

Remark 5 [Technical point about Theorem 3] Note that Theorem 3 is proved for positive (but possibly arbitrarily small) $\sigma_{\min}$ and τ. In particular, an update $(R_k, \widehat{\mathcal{S}}_k)$ is rejected in Algorithm 1 if

$$\frac{\sigma_r(R_k)}{\sigma_{r+1}(R_k)} \leq 1 + \tau, \tag{12}$$

for an (otherwise arbitrary) positive τ and whenever the ratio is well-defined. Here, $\sigma_i(R_k)$ is the ith largest singular value of R_k. This is merely a technical nuance, limited to the rth singular value gap, that helps prevent the the output subspace $\widehat{\mathcal{S}}_k$ from oscillating in the limit. Likewise, Theorem 3 does not address the case $\lambda = 0$ in Program (11), even though λ can be made arbitrarily small in Theorem 3. This is again for technical convenience and in fact the numerical simulations in Section 6 are all conducted with $\lambda = 0$.

Our second result establishes that, if SNIPE converges, then it converges to the true subspace $\mathcal{S}$, see Section 7.2 for the proof.

Proposition 6 [Convergence] *Consider the setup in the first paragraph of Theorem 3. Suppose that r independent copies of random coefficient vector $q \in \mathbb{R}^r$ almost surely form a basis for $\mathbb{R}^r$.*[3] *Suppose also that the output of* SNIPE *converges to an r-dimensional subspace $\widehat{\mathcal{S}}$, namely,*

$$\lim_{k\to\infty} d_{\mathbb{G}}(\widehat{\mathcal{S}}_k, \widehat{\mathcal{S}}) = 0. \tag{13}$$

Then almost surely it must hold that $\widehat{\mathcal{S}} = \mathcal{S}$.

Remark 7 [Discussion of Proposition 6] Proposition 6 does not specify the conditions under which SNIPE converges. Indeed, if the sampling probability p is too small, namely, if very few of the entries of incoming vectors are observed, then SNIPE might not converge at all as the numerical evidence suggests, see also Remark 4. However, if SNIPE converges, then it converges to the true subspace $\mathcal{S}$. The local rate of convergence is specified below.

The concept of *coherence* is critical in specifying the local convergence rate, since we consider entrywise subsampling. The coherence of an r-dimensional subspace $\mathcal{S}$ with orthonormal basis $S \in \mathbb{R}^{n\times r}$ is defined as

$$\eta(\mathcal{S}) := \frac{n}{r} \max \|S[i,:]\|_2^2, \tag{14}$$

3. For example, this requirement is met if entries of q are independent Gaussian random variables with zero-mean and unit variance.

where $S[i,:]$ is the ith row of S. It is easy to verify that $\eta(\mathcal{S})$ is independent of the choice of orthonormal basis S and that

$$1 \le \eta(\mathcal{S}) \le \frac{n}{r}. \tag{15}$$

It is also common to say that $\mathcal{S}$ is *coherent* (*incoherent*) when $\eta(\mathcal{S})$ is large (small). Loosely speaking, when $\mathcal{S}$ is coherent, its orthonormal basis S is "spiky." An example is when $\mathcal{S}$ is the span of a column-subset of the identity matrix. In contrast, when $\mathcal{S}$ is incoherent, entries of S tend to be "diffuse." Not surprisingly, identifying a coherent subspace from subsampled data may require many more samples (Balzano and Wright, 2015; Mitliagkas et al., 2014; Chen, 2015).

We will also use $\lesssim$ and $\gtrsim$ below to suppress (most of) the universal constants. Moreover, throughout C represents a universal constant, the value of which is subject to change in every appearance.

Our next results specify the local convergence rate of SNIPE. Indeed, the convergence speed near the true subspace $\mathcal{S}$ is linear as detailed in Theorems 8 and 10, and proved in Section 7.3. In particular, Theorem 8 states that, when sufficiently small, the expected estimation error of SNIPE reduces by a factor of $1 - p/32$ in every iteration.

Theorem 8 [Locally linear convergence of SNIPE **in expectation]** *Consider the setup in the first paragraph of Theorem 3. Fix a positive tuning parameter α, iteration $k \in [2:K]$, and let $\mathfrak{E}_{k-1}$ be the event where*

$$\frac{1}{\sqrt{nb}} \gtrsim p \gtrsim \log n \log^2\left(p\sqrt{r} d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1})\right) \frac{\eta(\mathcal{S}) r}{n}, \tag{16}$$

$$d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}) \log\left(\frac{16}{p\sqrt{r} d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1})}\right) \lesssim \frac{p^{\frac{3}{2}}}{\sqrt{\log n}}, \tag{17}$$

$$\|Q_k\| \le \frac{\sigma_{\min}}{\sqrt{1 - p/4}}, \tag{18}$$

where $\sigma_{\min}$ is the reject threshold in SNIPE. *Let also $\mathcal{A}_k$ be the event where the kth iteration of* SNIPE *is not rejected (see Step 3 of Algorithm 1) and let $1_{\mathcal{A}_k}$ be the indicator for this event, taking one if the event happens and zero otherwise. Then it holds that*

$$\mathbb{E}\left[1_{\mathcal{A}_k} \cdot d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_k) \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}\right] \le \left(1 - \frac{p}{32}\right) d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}). \tag{19}$$

Remark 9 [Discussion of Theorem 8] When the sampling probability p is large enough and SNIPE is near the true subspace $\mathcal{S}$, Theorem 8 states that the expected estimation error of SNIPE reduces by a factor of $1 - p/32$, if the iterate of SNIPE is not rejected. Note that: ① The lower bound on the sampling probability p in (16) matches the one in the low-rank matrix completion literature up to a logarithmic factor (Davenport and Romberg, 2016). Indeed, SNIPE can be interpreted as a streaming matrix completion algorithm as discussed in Section 3. The upper bound on p in (16) is merely for technical convenience and a tidier presentation in the most interesting regime for p. Indeed, since we often take $b = O(r) \ll n$, one might loosely read (16) as

$$\frac{1}{\sqrt{nr}} \gtrsim p \gtrsim \frac{\eta(\mathcal{S}) r}{n}, \tag{20}$$

in which the upper bound hardly poses a restriction even for moderately large data dimension n, as it forces $r = O(n^{\frac{1}{3}})$. ② Ignoring the logarithmic factors for simplicity, we may read (17) as $d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}) \lesssim p^{3/2}$, which "activates" (19). In other words, the basin of attraction of the true subspace $\mathcal{S}$ as a (possibly local) minimizer of the (nonconvex) Program (2) has a radius of $O(p^{3/2})$. ③ The indicator $1_{\mathcal{A}_k}$ in (19) removes the rejected iterates and similar conditions implicitly exist in the analysis of other streaming PCA algorithms (Balzano and Wright, 2015).

Note that Theorem 8 *cannot* tell us what the local convergence rate of SNIPE is, even in expectation. Indeed, the expected reduction in the estimation error of SNIPE, specified in (19), is not enough to activate (17) for the next iteration (namely, with k instead of $k-1$). That is, we cannot apply Theorem 8 iteratively and find the expected convergence rate of SNIPE. A key technical contribution of this paper is specifying the local behaviour of SNIPE below. With high probability, the estimation error does not *increase* by much in every iteration near the true subspace. However, only in some of these iterations does the error reduce. Overall, on a long enough interval, the estimation error of SNIPE near the true subspace indeed reduces substantially and with high probability as detailed in Theorem 10 and proved in Section 7.3. Performance guarantees for stochastic algorithms on long intervals is not uncommon, see for example (Cartis and Scheinberg, 2017).

Theorem 10 [Locally linear convergence of SNIPE**]** *Consider the setup in the first paragraph of Theorem 3. Suppose that the output $\widehat{\mathcal{S}}_{K_0}$ of* SNIPE *at iteration $K_0 \in [2:K]$ satisfies*

$$d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{K_0}) \lesssim \frac{e^{-\frac{Cp^3nb}{\widetilde{\eta}}} p^{\frac{7}{2}} nb}{\widetilde{\eta} \log b \log\left(C\widetilde{\eta}n\right) \log^2(K-K_0)}, \tag{21}$$

and that

$$K - K_0 \gtrsim \frac{\widetilde{\eta} \log b \log(K-K_0)}{p^2 nb}. \tag{22}$$

Above, C is a universal constant that might change in every appearance. Then it holds that

$$\prod_{k=K_0+1}^{K} 1_{\mathcal{A}_k} \cdot d(\mathcal{S}, \widehat{\mathcal{S}}_K) \lesssim \left(1 - \frac{Cp^3nb}{\widetilde{\eta} \log b \log(K-K_0)}\right)^{K-K_0} d(\mathcal{S}, \widehat{\mathcal{S}}_{K_0}), \tag{23}$$

except with a probability of at most

$$b^{-C\log(K-K_0)} + \sum_{k=K_0+1}^{K} \Pr\left[\|Q_k\| > \left(1 + \frac{Cp^3nb}{\widetilde{\eta}\log b}\right)\sigma_{\min}\right] \tag{24}$$

and provided that

$$\frac{1}{\sqrt{nb}} \gtrsim p \gtrsim \log^2 b \log n \frac{\eta(\mathcal{S}) r}{n}. \tag{25}$$

Above, $\sigma_{\min}$ is the reject threshold of SNIPE *and*

$$\widetilde{\eta} = \max_{k \in [K_0:K]} \widetilde{\eta}_k, \tag{26}$$

$$\widetilde{\eta}_k := nb \cdot \frac{\|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k\|_{\infty}^2}{\|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k\|_F^2}. \tag{27}$$

Remark 11 [Discussion of Theorem 10] Loosely speaking, Theorem 10 states that, with high probability, the estimation error of SNIPE over $O(\widetilde{\eta} n)$ iterations reduces linearly (i.e., exponentially fast) when SNIPE is near the true subspace and the sampling probability p is large enough. Most of the remarks about Theorem 8 are also valid here. Let us also point out that the dependence on the coefficient matrix Q_k in (24) is mild but necessary. As an example, consider the case where the coefficient vectors $\{q_t\}_t$ are standard random Gaussian vectors so that the coefficient matrices $\{Q_k\}_k$ are standard random Gaussian matrices, namely, populated with independent zero-mean Gaussian random variables with unit variance. Then by taking

$$\sigma_{\min} = C\sqrt{b} / \left(1 + \frac{Cp^3 nb}{\widetilde{\eta} \log b}\right),$$

we find that

$$\Pr\left[\|Q_k\| > \left(1 + \frac{Cp^3 nb}{\widetilde{\eta} \log b}\right) \sigma_{\min}\right] = \Pr\left[\|Q_k\| \gtrsim \sqrt{b}\right] \le e^{-Cb} \tag{28}$$

and consequently the failure probability in (24) becomes $b^{-C \log(K-K_0)} + (K-K_0) e^{-Cb}$, which can be made arbitrarily small by modestly increasing the block size b. For the reader's convenience, Appendix K collects the relevant spectral properties of a standard random Gaussian matrix. The dependence on $\|Q_k\|$ in Theorem 8 is *not* an artifact of our proof techniques. Indeed, when $\|Q_k\| \gg 1$, it is likely that certain directions in $\mathcal{S}$ are over represented which will skew the estimate of SNIPE in their favor.

Remark 12 [Coherence factor $\widetilde{\eta}$] A key quantity in Theorem 10 is the new "coherence" factor $\widetilde{\eta}$ which is absent in the expected behavior of SNIPE in Theorem 8. Somewhat similar to the coherence $\eta(\cdot)$ in (14), $\widetilde{\eta}_k$ measures how "spiky" the matrix $P_{\mathcal{S}_{k-1}^{\perp}} S_k \in \mathbb{R}^{n \times b}$ is. In fact, one may easily verify that

$$\widetilde{\eta}_k \le \operatorname{rank}(P_{\mathcal{S}_{k-1}^{\perp}} S_k) \cdot \nu(P_{\mathcal{S}_{k-1}^{\perp}} S_k)^2 \cdot \eta(\operatorname{span}(P_{\mathcal{S}_{k-1}^{\perp}} S_k)) \cdot \eta(\operatorname{span}(S_k^* P_{\mathcal{S}_{k-1}^{\perp}})), \tag{29}$$

where $\nu(P_{\mathcal{S}_{k-1}^{\perp}} S_k)$ is the condition number of $P_{\mathcal{S}_{k-1}^{\perp}} S_k$, namely, the ratio of largest to smallest nonzero singular values. The number of iterations needed to see a reduction in estimation error in (22) and the convergence rate of SNIPE in (23) both prefer $\widetilde{\eta}$ to be small, namely, prefer that $\{P_{\mathcal{S}_{k-1}^{\perp}} S_k\}_k$ are all incoherent as measured by $\widetilde{\eta}_k$.

When SNIPE is close enough to true subspace $\mathcal{S}$ as required in (21), one would expect that iterates of SNIPE would be nearly as coherent as $\mathcal{S}$ itself in the sense that $\eta(\widehat{\mathcal{S}}_k) \approx \eta(\mathcal{S})$. This intuition is indeed correct and also utilized in our analysis. However, even when $d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1})$ is small and $\eta(\widehat{\mathcal{S}}_{k-1}), \eta(\mathcal{S})$ are both small, $\widetilde{\eta}_k$ might be very large, namely, $P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k$ might be a spiky matrix. Indeed, when $b = r$, $(\widehat{S}_{k-1}^{\perp})^* S_k$ is approximately the (horizontal) tangent at $\widehat{\mathcal{S}}_{k-1}$ to the geodesic on the Grassmannian $\mathbb{G}(n, r)$ that connects $\widehat{\mathcal{S}}_{k-1}$

to $\mathcal{S}$. Even though both $\widehat{\mathcal{S}}_{k-1}$ and $\mathcal{S}$ are incoherent subspaces, namely, $\eta(\widehat{\mathcal{S}}_{k-1}), \eta(\mathcal{S})$ are both small, the tangent direction connecting the two is not necessarily incoherent. Despite the dependence of our results on $\widetilde{\eta}$, it is entirely possible that SNIPE with high probability approaches the true subspace $\mathcal{S}$ from an incoherent tangent direction, of which there are many. Such a result has remained beyond our reach. In fact, similar questions arise in matrix completion. Iterative Hard Thresholding (IHT) is a powerful algorithm for matrix completion with excellent empirical performance (Tanner and Wei, 2013), the convergence rate of which has remained unknown, to the best of our knowledge. With $\{M_k\}_k$ denoting the iterates of IHT, it is not difficult to see that if the differences $\{M_k - M_{k-1}\}_k$ are incoherent matrices (i.e., not spiky), then the linear convergence rate of IHT would follow from rather standard arguments.

5. Related Work

In this paper, we presented SNIPE for streaming PCA from incomplete data and, from a different perspective, SNIPE might be considered as a streaming matrix completion algorithm, see Section 3. In other words, SNIPE is a "subspace tracking" algorithm that identifies the linear structure of data as it arrives. Note also that t in our framework need not correspond to time, see Figure 1. For example, only a small portion of a large data matrix Y can be stored in the fast access memory of the processing unit, which could instead use SNIPE to fetch and process the data in small chunks and iteratively update the principal components. Moreover, SNIPE can be easily adapted to the dynamic case where the distribution of data changes over time. In dynamic subspace tracking for example, the (hidden) data vector s_t is drawn from the subspace $\mathcal{S}(t) \in \mathbb{G}(n, r)$ that varies with time. Likewise, it is easy to slightly modify SNIPE to handle noisy observations or equivalently to the case where $\{s_t\}_t$ are generated from a distribution with possibly full-rank covariance matrix. We leave investigating both of these directions to a future work.

Among several algorithms that have been proposed for tracking low-dimensional structure in a dataset from partially observed streaming data (Mitliagkas et al., 2014; Chi et al., 2013; Mardani et al., 2015; Xie et al., 2013; Eftekhari et al., 2017), SNIPE might be most closely related to GROUSE (Balzano et al., 2010; Balzano and Wright, 2013). GROUSE performs streaming PCA from incomplete data using stochastic gradient projection on the Grassmannian, updating its estimate of the true subspace with each new incomplete vector. Both GROUSE and SNIPE were designed based on the principle of least-change, discussed in Section 3. In fact, when GROUSE is sufficiently close to the true subspace and with a specific choice of its step length, both algorithms have nearly identical updates, see (Balzano and Wright, 2015, Equation 1.9). A weaker analogue of Theorem 8 for GROUSE was recently established in (Balzano and Wright, 2015). More specifically, (Balzano and Wright, 2015) stipulates that, if the current estimate $\widehat{\mathcal{S}}_k$ is sufficiently close to the true subspace $\mathcal{S}$, then $\widehat{\mathcal{S}}_{k+1}$ will be even closer to $\mathcal{S}$ in expectation. Such a result however *cannot* tell us what the local convergence rate of SNIPE is, even in expectation, see the discussion right before Theorem 10 above. In this sense, a key technical contribution of our work is establishing the local linear convergence of SNIPE, see Theorem 10, which is missing from its close competitor GROUSE. In fact, the global convergence guarantees listed in Theorem 3

and Proposition 6 are also unique to SNIPE; such theoretical guarantees are not available for GROUSE.

It might be interesting to add that our proposed update in SNIPE was inspired by that of GROUSE when we found zero-filled updates were unreliable (Eftekhari et al., 2017). However, GROUSE was derived as a purely streaming algorithm, and it therefore is not designed to leverage common low-rank structure that may be revealed when a block of vectors is processed at once. Therefore, for each block SNIPE often achieves a more significant reduction in error than is possible with GROUSE.

Lastly, both SNIPE and GROUSE have a computational complexity of $O(nr^2)$ flops per incoming vector, see Remark 1. Also, SNIPE and GROUSE both require $O(nr)$ memory elements of storage, see Remark 2. With complete data, namely, when no entries are missing, a close relative of both SNIPE and GROUSE are incremental SVD algorithms, a class of algorithms that efficiently compute the SVD of streaming data (Bunch and Nielsen, 1978; Balsubramani et al., 2013; Oja and Karhunen, 1985; Watanabe and Pakvasa, 1973; Balsubramani et al., 2013; Brand, 2002).

A streaming PCA algorithm might also be interpreted as a stochastic algorithm for PCA (Arora et al., 2012). Stochastic projected gradient ascent in this context is closely related to the classical power method. In particular, the algorithm in (Mitliagkas et al., 2014) extends the power method to handle missing data, in part by improving the main result of (Lounici, 2014). With high probability, this algorithm converges globally and linearly to the true subspace and, most notably, succeeds for arbitrarily small sampling probability p, if the scope of the algorithm T is large enough. Additionally, this algorithm too has a computational complexity of $O(nr^2)$ operations per vector and a storage requirement of $O(nr)$ memory elements. In practice, SNIPE substantially outperforms the power method, as we will see in Section 6. A disadvantage of the power method is that it updates its estimate of the true subspace with every $O(n)$ incoming vectors; the waiting time might be prohibitively long if n is large. In contrast, SNIPE frequently updates its estimate with every $b = O(r)$ incoming vectors. As we will see in Section 6, SNIPE substantially outperforms the power method in practice. Let us add that POPCA (Gonen et al., 2016) is closely related to the power method, for which the authors provide lower bounds on the achievable sample complexity. However, POPCA has substantially greater memory demand than SNIPE, since it maintains an estimate of the possibly dense $n \times n$ sample covariance matrix of incoming data.

The PETRELS algorithm (Chi et al., 2013) operates on one column at a time (rather than blocks) and global convergence for PETRELS, namely, convergence to a stationary point of the underlying nonconvex program, is known. Designed for streaming matrix completion, the algorithm in (Mardani et al., 2015) also operates on one column at a time and asymptotic onvergence to the true subspace is established, see Propositions 2 and 3 therein. This framework is also extended to tensors. MOUSSE in (Xie et al., 2013) tracks a union of subspaces rather than just one; SNIPE would function more like an ingredient of this algorithm. Asymptotic consistency of MOUSSE is also established there. The theoretical guarantees of SNIPE are more comprehensive in the sense that we also offer local convergence rate for SNIPE, see Theorems 8 and 10. ReProcs, introduced in (Lois and Vaswani, 2015), tracks a slowly changing subspace when initialized sufficiently close.

In the next section, we compare the performance of several of these algorithms in practice and find that SNIPE competes empirically with state-of-the-art algorithms.

Even though we consider uniform random sampling of the entries of incoming vectors, SNIPE can be applied to any incomplete data. For example, instead of uniform sampling analyzed here, one can perhaps sample the entries of every incoming vector based on their estimated importance. More specifically, in iteration k, one might observe each entry of the incoming vector with a probability proportional to the *leverage score* of the corresponding row of the current estimate $\widehat{\mathcal{S}}_{k-1}$. In batch or offline matrix completion, using the idea of *leveraged sampling* (as opposed to uniform sampling) alleviates the dependence on the coherence factor $\eta(\mathcal{S})$ in (16) (Eftekhari et al., 2018a; Chen, 2015). While interesting, we have not pursued this direction in the current work.

6. Simulations

This section consists of two parts: first, we empirically study the dependence of SNIPE on various parameters, and second we compare SNIPE with existing algorithms for streaming subspace estimation with missing data. In all simulations, we consider an r-dimensional subspace $\mathcal{S} \subset \mathbb{R}^n$ and a sequence of generic vectors $\{s_t\}_{t=1}^T \subset \mathcal{S}$. Each entry of these vectors is observed with probability $p \in (0, 1]$ and collected in vectors $\{y_t\}_{t=1}^T \subset \mathbb{R}^n$. Our objective is to estimate $\mathcal{S}$ from $\{y_t\}$, as described in Section 1.

Sampling probability We first set $n = 100$, $r = 5$, and let $\mathcal{S}$ be a generic r-dimensional subspace, namely, the span of an $n \times r$ standard random Gaussian matrix. For various values of probability p, we run SNIPE with block size $b = 2r = 10$ and scope of $T = 500r = 2500$, recording the average estimation error $d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_K)$ over 50 trials, see (10). The average error versus probability is plotted in Figure 3a.

Subspace dimension With the same setting as the previous paragraph, we now set $p = 3r/n = 0.15$ and vary the subspace dimension r, block size $b = 2r$, and scope $T = 500r$. The average error versus subspace dimension is plotted in Figure 3b.

Ambient dimension This time, we set $r = 5$, $p = 3r/n$, $b = 2r$, $T = 500r$, and vary the ambient dimension n. In other words, we vary n while keeping the number of samples per vector fixed at about $pn = 3r$. The average error versus ambient dimension is plotted in Figure 3c. Observe that the performance of SNIPE steadily degrades as n increases. This is in agreement with Theorem 8 by substituting $p = 3r/n$ there, which states that the error reduces by a factor of $1 - Cr/n$, in expectation. A similar behavior is observed for our close competitor, namely, GROUSE (Balzano and Wright, 2015).

Block size Next we set $n = 100$, $r = 5$, $p = 3r/n$, $T = 500r$, and vary the block size b. The average error versus block size in both cases is depicted in Figure 3d. From Step 3 of Algorithm 1, a block size of $b \geq r$ is necessary for the success of SNIPE and qualitatively speaking larger values of b lead to better stability in face of missing data, which might explain the poor performance of SNIPE for very small values of b. However, as b increases, the number of blocks $K = T/b$ reduces because the scope T is held fixed. As the estimation error of SNIPE scales like $(1 - cp)^{-K}$ in Theorem 10 for a certain factor c, the performance

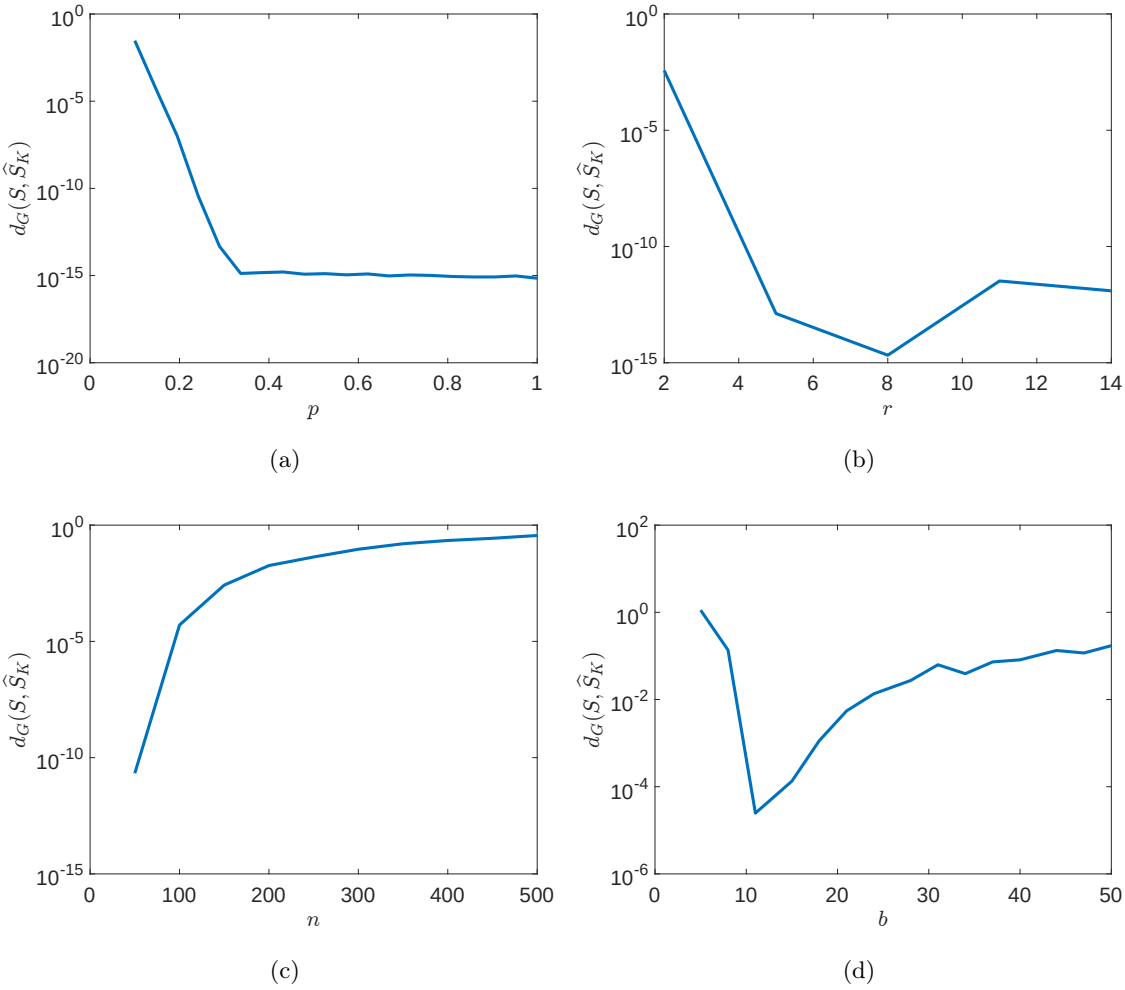


Figure 3: Performance of SNIPE as (a) sampling probability p, (b) subspace dimension r, (c) ambient dimension n, (d) block size b vary. $\widehat{\mathcal{S}}_K$ is the output of SNIPE and $d_G(\mathcal{S}, \widehat{\mathcal{S}}_K)$ is its distance to the true subspace $\mathcal{S}$, which generated the input of SNIPE. See Section 6 for details and note that each panel is generated with a different and random subspace $\mathcal{S}$.

suffers in Figure 3d. It appears that the choice of $b = 2r$ in SNIPE guarantees the best empirical performance.

Coherence Lastly, we set $n = 300$, $r = 10$, $p = 3r/n$, $b = 2r$, and $T = 500r$. We then test the performance of SNIPE as the coherence of $\mathcal{S}$ varies, see (14). To that end, let $\mathcal{S} \subset \mathbb{R}^n$ be a generic subspace with orthonormal basis $S \in \mathbb{R}^{n \times r}$. In particular S is obtained by orthogonalizing the columns of a standard $n \times n$ random Gaussian matrix. Then, the average coherence of $\mathcal{S}$ over 50 trials was $3.3334 \ll n/r$ and the average estimation error of SNIPE was $2.795 \cdot 10^{-5}$. On the other hand, let $D \in \mathbb{R}^{n \times n}$ be a diagonal matrix with entries $D[i,i] = i^{-1}$ and consider $S' = DS$. Unlike $\mathcal{S}$, the new subspace $\mathcal{S}' := \text{span}(S')$ is

	$p = 0.15$		$p = 0.30$		$p = 0.45$		$p = 0.60$		$p = 0.75$	
	$d_{\mathbb{G}} < 10^{-3}$	$d_{\mathbb{G}} < 10^{-7}$	$d_{\mathbb{G}} < 10^{-3}$	$d_{\mathbb{G}} < 10^{-7}$	$d_{\mathbb{G}} < 10^{-3}$	$d_{\mathbb{G}} < 10^{-7}$	$d_{\mathbb{G}} < 10^{-3}$	$d_{\mathbb{G}} < 10^{-7}$	$d_{\mathbb{G}} < 10^{-3}$	$d_{\mathbb{G}} < 10^{-7}$
GROUSE	**878.0** (76.1)	**1852.4** (85.2)	**294.9** (20.6)	**646.1** (26.8)	181.6 (13.8)	391.2 (18.0)	130.2 (10.3)	277.2 (12.8)	105.1 (11.7)	213.5 (12.8)
PETRELS	1689.0 (1394.1)	2853.7 (916.9)	421.8 (31.3)	1100.5 (64.2)	262.1 (25.3)	802.0 (31.1)	181.8 (20.2)	671.4 (22.8)	133.3 (21.3)	599.1 (24.0)
SNIPE	1815.7 (137.9)	3946.9 (182.4)	537.4 (39.9)	1236.4 (62.5)	282.4 (25.8)	649.6 (41.5)	171.4 (17.2)	391.4 (25.4)	105.5 (9.1)	241.6 (15.6)
SNIPE-overlap	1588.3 (183.8)	3319.5 (232.9)	318.7 (27.1)	704.6 (36.8)	**131.8** (11.6)	**296.4** (17.4)	**71.3** (5.7)	**155.6** (9.9)	**44.2** (4.2)	**91.8** (6.8)

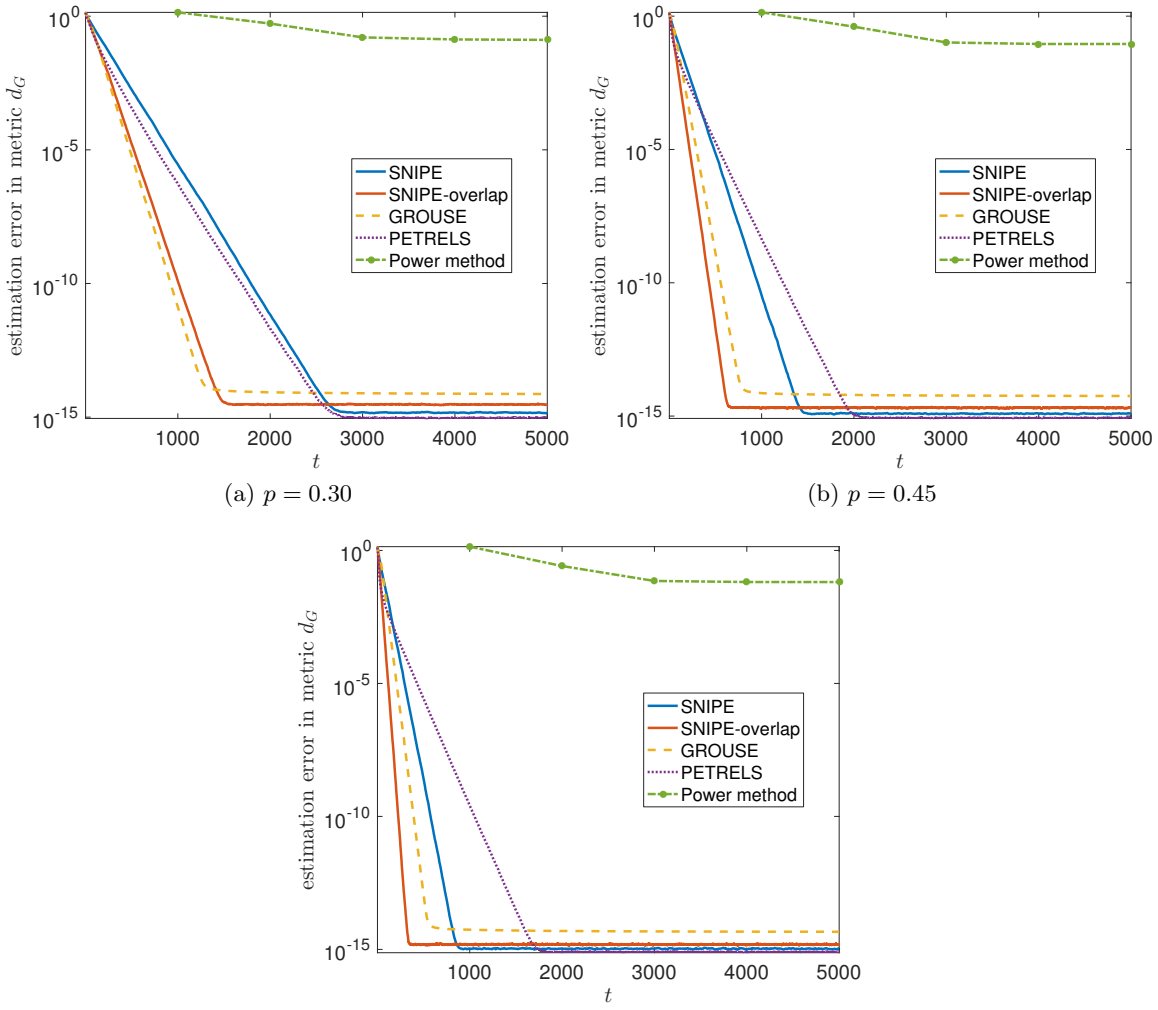
Table 1: Average number of revealed data columns needed to reach the indicated subspace recovery error for various sampling probabilities p over 100 random trials. Standard deviations are given in parenthesis.

typically coherent since the energy of its orthonormal basis S' is mostly concentrated along its first few rows. This time, the average coherence of $\mathcal{S}'$ over 50 trials was $19.1773 \approx n/r$ and the average estimation error of SNIPE was substantially worse at 0.4286.

Comparisons Next we empirically compare SNIPE with GROUSE (Balzano et al., 2010; Zhang and Balzano, 2016), PETRELS (Chi et al., 2013), and the modified power method in (Mitliagkas et al., 2014). In addition to the version of SNIPE given in Algorithm 1, we also include comparisons with a simple variant of SNIPE, which we call SNIPE-overlap. Unlike SNIPE which processes disjoint blocks, SNIPE-overlap processes all overlapping blocks of data. More precisely, for a block size b, SNIPE-overlap first processes data columns $t = 1, 2, ..., b$, followed by columns $t = 2, 3, ..., b+1$, and so on, whereas regular SNIPE processes columns $t = 1, 2, ..., b$ followed by $t = b+1, b+2, ..., 2b$, etc. The theory developed in this paper does not hold for SNIPE-overlap because of lack of statistical independence between iterations, but we include the algorithm in the comparisons since it represents a minor modifications of the SNIPE-overlap framework and appears to have some empirical benefits, as detailed below.

In these experiments, we set $n = 100$, $r = 5$, $T = 5000$, and take $\mathcal{S} \subset \mathbb{R}^n$ to be a generic r-dimensional subspace and simulate noiseless data samples as before. In Figure 4 we compare the algorithms for three values of sampling probability p, which shows the average over 100 trials of the estimation error of algorithms (with respect to the metric $d_{\mathbb{G}}$) relative to the number of revealed data columns. For SNIPE, we used the block size of $b = 2r$. Having tried to get the best performance from GROUSE, we used the "greedy" step-size as proposed in (Zhang and Balzano, 2016). For (Mitliagkas et al., 2014), we set the block size as $b = 1000$ which was found empirically to yield the lowest subspace error after $T = 5000$ iterations.

In Table 1 we also compare the average number of revealed columns needed to reach a given error tolerance for each algorithm (as measured by error metric $d_{\mathbb{G}}$) for various values of the sampling probability p. We omit the modified power method from the results since it was unable to reach the given error tolerances in all cases. For the medium/high sampling rates $p = 0.45, 0.60, 0.75$, SNIPE-overlap is fastest to converge, while regular SNIPE is competitive with GROUSE and PETRELS. For the lower sampling rates $p = 0.15, 0.30$ we find GROUSE yields the fastest convergence, although SNIPE-overlap is also competitive with GROUSE for $p = 0.30$.



(a) $p = 0.30$ (b) $p = 0.45$

(c) $p = 0.60$

Figure 4: Average subspace estimation error versus number of revealed data columns at the specified sampling probability p, see Section 6 for details.

7. Theory

In this section, we prove the technical results presented in Section 4. A short word on notation is in order first. We will frequently use MATLAB's matrix notation so that, for example, $A[i, j]$ is the $[i, j]$th entry of A, and the row-vector $A[i, :]$ corresponds to the ith row of A. By $\{\epsilon_i\}_i \overset{\text{ind.}}{\sim} \text{Bernoulli}(p)$, we mean that $\{\epsilon_i\}_i$ are independent Bernoulli random variables taking one with probability of p and zero otherwise. Throughout, $E_{i,j}$ stands for the $[i, j]$th canonical matrix so that $E_{i,j}[i, j] = 1$ is its only nonzero entry. The size of $E_{i,j}$ may be inferred from the context. As usual, $\|\cdot\|$ and $\|\cdot\|_F$ stand for the spectral and Frobenius norms. In addition, $\|A\|_\infty$ and $\|A\|_{2\to\infty}$ return the largest entry of a matrix A (in magnitude) and the largest ℓ_2 norm of the rows of A, respectively. Singular values of

a matrix A are denoted by $\sigma_1(A) \geq \sigma_2(A) \geq \cdots$. For purely aesthetic reasons, we will occasionally use the notation $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$.

7.1. Convergence of SNIPE to a Stationary Point (Proof of Theorem 3)

Consider Program (2), namely,

$$\begin{cases} \min \;\; f_\Omega\left(X, \mathcal{U}\right) := \|P_{\mathcal{U}}^{\perp} X\|_F^2 + \lambda \|P_{\Omega^C}(X)\|_F^2, \\ P_\Omega(X) = Y, \end{cases} \tag{30}$$

where the minimization is over matrix $X \in \mathbb{R}^{n\times T}$ and subspace $\mathcal{U} \in \mathbb{G}(n,r)$. Before proceeding with the rest of the proof, let us for future reference record the partial derivatives of f_Ω below. Consider a small perturbation to X in the form of $X + \Delta$, where $\Delta \in \mathbb{R}^{n\times T}$. Let $U \in \mathbb{R}^{n\times r}$ and $U^\perp$ be orthonormal bases for $\mathcal{U}$ and its orthogonal complement $\mathcal{U}^\perp$, respectively. Consider also a small perturbation to U in the form of $U + U^\perp \Delta' \in \mathbb{R}^{n\times r}$, where $\Delta' \in \mathbb{R}^{(n-r)\times r}$. The perturbation to $f_\Omega(X,\mathcal{U}) = f_\Omega(X,U)$ can be written as

$$\begin{aligned} f_\Omega\left(X+\Delta, U + U^\perp \Delta'\right) = f_\Omega\left(X,\mathcal{U}\right) &+ \langle \Delta, \partial_X f_\Omega\left(X,\mathcal{U}\right)\rangle \\ &+ \left\langle \Delta', \partial_{\mathcal{U}} f_\Omega\left(X,\mathcal{U}\right)\right\rangle + o(\|\Delta\|_F) + o(\|\Delta'\|_F), \end{aligned} \tag{31}$$

where $o(\cdot)$ is the standard little-o notation. The partial derivatives of f_Ω are listed below and derived in Appendix D.

Lemma 13 *For f_Ω in Program (30), the first-order partial derivatives at $(X,\mathcal{U}) \in \mathbb{R}^{n\times T} \times \mathbb{G}(n,r)$ are*

$$\begin{aligned} \partial_X f_\Omega\left(X,\mathcal{U}\right) &= 2P_{\mathcal{U}^\perp} X + 2\lambda P_{\Omega^C}(X) \in \mathbb{R}^{n\times T}, \\ \partial_{\mathcal{U}} f_\Omega\left(X,\mathcal{U}\right) &= -2(U^\perp)^* X X^* U \in \mathbb{R}^{(n-r)\times r}, \end{aligned} \tag{32}$$

where $U \in \mathbb{R}^{n\times r}$ and $U^\perp \in \mathbb{R}^{n\times(n-r)}$ are orthonormal bases for $\mathcal{U}$ and its orthogonal complement, respectively.

Recall that $Q_k \in \mathbb{R}^{r\times b}$ is a random matrix with bounded expectation, namely, $\mathbb{E}\|\mathcal{Q}_k\|_F < \infty$. As $K \to \infty$, $\{Q_k\}_{k=1}^K$ therefore has a bounded subsequence. To keep the notation simple and without any loss of generality, we assume that in fact the sequence $\{Q_k\}_k$ is itself bounded. As $K \to \infty$ and for an integer l, we can always find an interval of length l over which the same index set and nearly the same coefficient matrix repeats. More specifically, consider an index set $\widehat{\Omega} \subseteq [1:n] \times [1:b]$ and a matrix $\widehat{Q} \in \mathbb{R}^{r\times b}$ in the support of the distributions from which $\{\Omega_k\}_{k=1}^K$ and $\{Q_k\}_{k=1}^K$ are drawn. For every integer l, as a result of the second Borel-Cantelli lemma (Durrett, 2010, pg. 64), almost surely there exists a contiguous interval

$$\kappa_l := [k_l - l + 1 : k_l], \tag{33}$$

such that

$$\Omega_k = \widehat{\Omega}, \qquad k \in \kappa_l, \tag{34}$$

$$\max_{k\in\kappa_l} \|Q_k - \widehat{Q}\|_F \leq \frac{1}{l}. \tag{35}$$

As $l \to \infty$, the measurements corresponding to the interval κ_l converge. To be specific, let $\widehat{Y} := P_{\widehat{\Omega}}(S\widehat{Q})$ and note that

$$\begin{aligned}
\lim_{l\to\infty} \max_{k\in\kappa_l} \ \|Y_k - \widehat{Y}\|_F &= \lim_{l\to\infty} \max_{k\in\kappa_l} \ \|P_{\Omega_k}(SQ_k) - \widehat{Y}\|_F \\
&= \lim_{l\to\infty} \max_{k\in\kappa_l} \ \|P_{\widehat{\Omega}}(SQ_k) - \widehat{Y}\|_F \\
&= \lim_{l\to\infty} \max_{k\in\kappa_l} \ \|P_{\widehat{\Omega}}(S(Q_k - \widehat{Q}))\|_F \\
&\le \lim_{l\to\infty} \max_{k\in\kappa_l} \ \|Q_k - \widehat{Q}\|_F \\
&= 0. \qquad (36)
\end{aligned}$$

The above observation encourages us to exchange (Ω_k, Y_k) with $(\widehat{\Omega}, \widehat{Y})$ on the interval κ_l. Let us therefore study the program

$$\begin{cases} \min \ \ f_{\widehat{\Omega}}(X, \mathcal{U}), \\ P_{\widehat{\Omega}}(X) = \widehat{Y}, \end{cases} \qquad (37)$$

where the minimization is over all matrices $X \in \mathbb{R}^{n\times b}$ and subspaces $\mathcal{U} \in \mathbb{G}(n, r)$. From a technical viewpoint, it is in fact more convenient to relax the equality constraint above as

$$\begin{cases} \min \ \ f_{\widehat{\Omega}}(X, \mathcal{U}), \\ \|P_{\widehat{\Omega}}(X) - \widehat{Y}\|_F \le \epsilon, \end{cases} \qquad (38)$$

for $\epsilon > 0$. We fix ϵ for now. Let us next use alternative minimization to solve Program (38). More specifically, recall (33) and consider the initialization $\widehat{\mathcal{S}}_{k_l-l,\epsilon} := \widehat{\mathcal{S}}_{k_l-l}$, where $\widehat{\mathcal{S}}_{k_l-l}$ is the output of SNIPE at iteration $k_l - l$, see Algorithm 1. For every $k \in \kappa_l$, consider the program

$$\begin{cases} \min \ \ f_{\widehat{\Omega}}(X, \widehat{\mathcal{S}}_{k-1,\epsilon}), \\ \|P_{\widehat{\Omega}}(X) - \widehat{Y}\|_F \le \epsilon, \end{cases} \qquad (39)$$

and let $R_{k,\epsilon}$ be a minimizer of Program (39). We then update the subspace by solving

$$\min_{\mathcal{U}\in\mathbb{G}(n,r)} \ f_{\widehat{\Omega}}(R_{k,\epsilon}, \mathcal{U}), \qquad (40)$$

and setting $\widehat{\mathcal{S}}_{k,\epsilon}$ to be a minimizer of Program (40). Recalling the definition of $f_{\widehat{\Omega}}$ in Program (30) and in light of the Eckart-Young-Mirsky Theorem, Program (40) can be solved by computing top r left singular vectors of $R_{k,\epsilon}$ (Eckart and Young, 1936; Mirsky, 1966). For future reference, note that the optimality and hence stationarity of $\widehat{\mathcal{S}}_{k,\epsilon}$ in Program (40) dictates that

$$\partial_{\mathcal{U}} f_{\widehat{\Omega}}(R_{k,\epsilon}, \widehat{\mathcal{S}}_{k,\epsilon}) = 0, \qquad k \in \kappa_l, \qquad (41)$$

where $\partial_{\mathcal{U}} f$ was specified in Lemma 13. From the above construction of the sequence $\{(R_{k,\epsilon}, \widehat{\mathcal{S}}_{k,\epsilon})\}_{k\in\kappa_l}$, we also observe that

$$0 \le f_{\widehat{\Omega}}(R_{k,\epsilon}, \widehat{\mathcal{S}}_{k,\epsilon}) \le f_{\widehat{\Omega}}(R_{k,\epsilon}, \widehat{\mathcal{S}}_{k-1,\epsilon}) \le f_{\widehat{\Omega}}(R_{k-1,\epsilon}, \widehat{\mathcal{S}}_{k-1,\epsilon}), \qquad (42)$$

for every $k \in [k_l - l + 2 : k_l] \subset \kappa_l$, see (33). That is, $\{f_{\widehat{\Omega}}(R_{k,\epsilon}, \widehat{\mathcal{S}}_{k,\epsilon})\}_{k\in\kappa_l}$ is a nonincreasing and nonnegative sequence. It therefore holds that

$$\lim_{l\to\infty} \left| f_{\widehat{\Omega}}(R_{k_l-1,\epsilon}, \widehat{\mathcal{S}}_{k_l-1,\epsilon}) - f_{\widehat{\Omega}}(R_{k_l,\epsilon}, \widehat{\mathcal{S}}_{k_l-1,\epsilon}) \right| = 0. \tag{43}$$

By the feasibility of $R_{k_l-1,\epsilon}$ in Program (39) and by the continuity of $f_{\widehat{\Omega}}(X, \mathcal{U})$ in X, we conclude in light of (43) that $R_{k_l-1,\epsilon}$ too is a minimizer (and hence also a stationary point) of Program (39) and in the limit of $l \to \infty$. We therefore find by writing the stationarity conditions of Program (39) at $R_{k_l,\epsilon}$ that

$$\|P_{\widehat{\Omega}}(R_{k_l,\epsilon}) - \widehat{Y}\|_F \leq \epsilon, \tag{44}$$

$$\lim_{l\to\infty} \left\| \partial_X f_{\widehat{\Omega}}(R_{k_l,\epsilon}, \widehat{\mathcal{S}}_{k_l,\epsilon}) + \lambda_{k_l,\epsilon}(P_{\widehat{\Omega}}(R_{k_l,\epsilon}) - \widehat{Y}) \right\|_F = 0. \tag{45}$$

for nonnegative $\lambda_{k_l,\epsilon}$. Recalling the definition of $f_{\widehat{\Omega}}$ and that $\lambda > 0$ by assumption, we observe that Program (39) is strongly convex in $P_{\widehat{\Omega}^C}(X)$ and consequently any pair of minimizers of Program (39) must agree on the index set $\widehat{\Omega}^C$. Optimality of $R_{k_l,\epsilon}$ and limit optimality of $R_{k_l-1,\epsilon}$ in Program (39) therefore imply that

$$\lim_{l\to\infty} \left\| P_{\widehat{\Omega}^C}(R_{k_l-1,\epsilon} - R_{k_l,\epsilon}) \right\|_F = 0. \tag{46}$$

On the index set $\widehat{\Omega}$, on the other hand, the feasibility of both $R_{k_l-1,\epsilon}$ and $R_{k_l,\epsilon}$ in Program (39) implies that

$$\begin{aligned} \|P_{\widehat{\Omega}}(R_{k_l-1,\epsilon} - R_{k_l,\epsilon})\|_F &\leq \|P_{\widehat{\Omega}}(R_{k_l-1,\epsilon}) - \widehat{Y}\|_F + \|P_{\widehat{\Omega}}(R_{k_l,\epsilon}) - \widehat{Y}\|_F \qquad \text{(triangle inequality)} \\ &\leq 2\epsilon. \end{aligned} \tag{47}$$

Combining (46) and (47) yields that

$$\begin{aligned} &\lim_{l\to\infty} \|R_{k_l-1,\epsilon} - R_{k_l,\epsilon}\|_F \\ &\leq \lim_{l\to\infty} \|P_{\widehat{\Omega}}(R_{k_l-1,\epsilon} - R_{k_l,\epsilon})\|_F + \lim_{l\to\infty} \|P_{\widehat{\Omega}^C}(R_{k_l-1,\epsilon} - R_{k_l,\epsilon})\|_F \qquad \text{(triangle inequality)} \\ &\leq 2\epsilon, \qquad \text{(see (46,47))} \end{aligned} \tag{48}$$

In light of (39), $\{R_{k_l,\epsilon}\}_l$ is bounded and consequently has a convergent subsequence. Without loss of generality and to simplify the notation, we assume that $\{R_{k_l,\epsilon}\}_l$ is itself convergent, namely, that there exists $R_\epsilon \in \mathbb{R}^{n\times b}$ for which

$$\lim_{l\to\infty} \|R_{k_l,\epsilon} - R_\epsilon\|_F \leq 2\epsilon. \tag{49}$$

Let us now send ϵ to zero in (49) to obtain that

$$\lim_{\epsilon\to 0} \lim_{l\to\infty} \|R_{k_l,\epsilon} - R_\epsilon\|_F \leq \lim_{\epsilon\to 0} 2\epsilon = 0. \tag{50}$$

We next show that it is possible to essentially change the order of limits above and also conclude that $(R, \widehat{\mathcal{S}})$ coincides with the output of SNIPE in limit. The following result is proved in Appendix E.

Lemma 14 *With the setup above, there exist a sequence* $\{\epsilon_i\}_i$ *with* $\lim_{i\to\infty}\epsilon_i = 0$ *and a matrix* $R \in \mathbb{R}^{n\times b}$ *such that*

$$\lim_{l,i\to\infty} \|R_{k_l,\epsilon_i} - R\|_F = \lim_{l\to\infty}\lim_{i\to\infty} \|R_{k_l,\epsilon_i} - R\|_F = \lim_{i\to\infty}\lim_{l\to\infty} \|R_{k_l,\epsilon_i} - R\|_F = 0. \tag{51}$$

Moreover, suppose that the output of SNIPE *in every iteration has a spectral gap in the sense that there exists* $\tau > 0$ *such that*

$$\frac{\sigma_r(R_k)}{\sigma_{r+1}(R_k)} \geq 1 + \tau, \tag{52}$$

for every k. *Let* $\widehat{\mathcal{S}}_{k_l,\epsilon_i}$ *and* $\widehat{\mathcal{S}}$ *be the span of top* r *left singular vectors of* R_{k_l,ϵ_i} *and* R, *respectively. Then it holds that*

$$\lim_{l,i\to\infty} d_{\mathbb{G}}(\widehat{\mathcal{S}}_{k_l,\epsilon_i}, \mathcal{S}) = 0. \tag{53}$$

Lastly, in the limit of $l \to \infty$, SNIPE *produces* $(R, \widehat{\mathcal{S}})$ *in every iteration, namely,*

$$\lim_{l\to\infty} \|R_{k_l} - R\|_F = \lim_{l\to\infty} d_{\mathbb{G}}(\widehat{\mathcal{S}}_{k_l}, \widehat{\mathcal{S}}) = 0, \tag{54}$$

where $(R_{k_l}, \widehat{\mathcal{S}}_{k_l})$ *is the output of* SNIPE *in iteration* k_l, *see Algorithm 1.*

In fact, the pair $(R, \widehat{S})$ from Lemma 14 is stationary in limit in the sense described next and proved in Appendix F.

Lemma 15 *The pair* $(R, \widehat{S})$ *in Lemma 14 is a stationary point of the program*

$$\begin{cases} \min \;\; f_{\Omega_{k_l}}(X, \mathcal{U}), \\ P_{\Omega_{k_l}}(X) = Y_{k_l}, \end{cases} \tag{55}$$

as $l \to \infty$. *The minimization above is over all matrices* $X \in \mathbb{R}^{n\times b}$ *and subspaces* $\mathcal{U} \in \mathbb{G}(n, r)$. *More specifically, it holds that*

$$\lim_{l\to\infty} \|\partial_{\mathcal{U}} f_{\Omega_{k_l}}(R, \widehat{\mathcal{S}})\|_F = 0, \tag{56}$$

$$\lim_{l\to\infty} \|P_{\Omega_{k_l}}(R) - Y_{k_l}\|_F = 0, \tag{57}$$

$$\lim_{l\to\infty} \|P_{\Omega_{k_l}^C}(\partial_X f_{\Omega_{k_l}}(R, \widehat{\mathcal{S}}))\|_F = 0. \tag{58}$$

In words, Lemmas 14 and 15 together imply that the output of SNIPE in limit is a stationary point of Program (55). This completes the proof of Theorem 3.

7.2. Convergence of SNIPE (Proof of Proposition 6)

In iteration k of SNIPE, we partially observe the data block $SQ_k \in \mathbb{R}^{n\times b}$ on a random index set $\Omega_k \subset [1:n]\times[1:b]$, where $Q_k \in \mathbb{R}^{r\times b}$ is a random coefficient matrix. We collect the observations in $Y_k = P_{\Omega_k}(SQ_k) \in \mathbb{R}^{n\times b}$, see Sections 2 and 3 for the detailed setup. Note that R_k in (1) can be written as

$$R_k = Y_k + P_{\Omega_k^C}(\widehat{S}_{k-1}Q'_k) = P_{\Omega_k}(SQ_k) + P_{\Omega_k^C}(\widehat{S}_{k-1}Q'_k), \tag{59}$$

where

$$Q'_k := \begin{bmatrix} \cdots & \left(\widehat{S}_{k-1}^* P_{\omega_t}\widehat{S}_{k-1} + \lambda I_r\right)^\dagger y_t & \cdots \end{bmatrix} \in \mathbb{R}^{r\times b}. \tag{60}$$

By (13), there exists $Q''_k \in \mathbb{R}^{r\times b}$ and

$$R'_k := P_{\Omega_k}(SQ_k) + P_{\Omega_k^C}(\widehat{S}Q''_k), \tag{61}$$

such that

$$\lim_{k\to\infty} \|R_k - R'_k\|_F = 0. \tag{62}$$

In (61) above, $\widehat{S} \in \mathbb{R}^{n\times r}$ is an orthonormal basis for the subspace $\widehat{\mathcal{S}}$. In Algorithm 1, the rank-r truncated SVD of R_k spans $\widehat{\mathcal{S}}_k \in \mathbb{G}(n,r)$, namely, the output of SNIPE in iteration k. Let also $\widehat{\mathcal{S}}'_k \in \mathbb{G}(n,r)$ denote the span of rank-r truncated SVD of R'_k. The existence of the reject option in Algorithm 1 with positive τ implies that R_k has a spectral gap and therefore $\widehat{\mathcal{S}}_k$ is uniquely defined. Combining this with (62), we find that $\widehat{\mathcal{S}}'_k$ too is uniquely defined in the limit of $k\to\infty$. Therefore another consequence of (62) is that

$$\lim_{k\to\infty} d_{\mathbb{G}}(\widehat{\mathcal{S}}_k, \widehat{\mathcal{S}}'_k) = 0. \tag{63}$$

Then we have that

$$\lim_{k\to\infty} d_{\mathbb{G}}(\widehat{\mathcal{S}}'_k, \widehat{\mathcal{S}}) \le \lim_{k\to\infty} d_{\mathbb{G}}(\widehat{\mathcal{S}}'_k, \widehat{\mathcal{S}}_k) + \lim_{k\to\infty} d_{\mathbb{G}}(\widehat{\mathcal{S}}_k, \widehat{\mathcal{S}}) \le 0 + 0, \qquad \text{(see (63,13))} \tag{64}$$

namely, $\widehat{\mathcal{S}}'_k$ converges to $\widehat{\mathcal{S}}$ in the limit too. Let us now rewrite R'_k as

$$R'_k = P_{\Omega_k}(SQ_k - \widehat{S}Q''_k) + \widehat{S}Q''_k, \qquad \text{(see (61))} \tag{65}$$

which, together with (64), implies that

$$\lim_{k\to\infty} \|\widehat{S}^* P_{\Omega_k}(SQ_k - \widehat{S}Q''_k)\|_F = 0. \tag{66}$$

We can rewrite the above limit in terms of the data vectors (rather than data blocks) to obtain that

$$\lim_{t\to\infty} \|\widehat{S}^* P_{\omega_t}(Sq_t - \widehat{S}q''_t)\|_F = 0, \tag{67}$$

where $\{q_t, q''_t\}_t$ form the columns of the blocks $\{Q_k, Q''_k\}_k$, and the index sets $\{\omega_t\}_t \subseteq [1:n]$ form $\{\Omega_k\}_k$. There almost surely exists a subsequence $\{t_i\}_i$ over which $\omega_{t_i} = \{1\}$, namely, there is a subsequence where we only observe the first entry of the incoming data vector. Consider a vector $q^1 \in \mathbb{R}^r$ in the support of the distribution from which $\{q_t\}_t$

are drawn. Then there also exists a subsequence of $\{t_i\}_i$, denoted by $\{t_{i_j}\}_{i_j}$, such that $\lim_{j\to\infty} \|q_{t_{i_j}} - q^1\|_2 = 0$. Restricted to the subsequence $\{t_{i_j}\}_j$, (67) reads as

$$0 = \lim_{j\to\infty} \|\widehat{S}^* P_{\omega_{t_{i_j}}}(Sq_{t_{i_j}} - \widehat{S}q''_{t_{i_j}})\|_F = \lim_{j\to\infty} \|\widehat{S}^* P_{\{1\}}(Sq_{t_{i_j}} - \widehat{S}q''_{t_{i_j}})\|_F = \|\widehat{S}^* P_{\{1\}}(Sq^1 - \widehat{S}q''^1)\|_F, \tag{68}$$

where we set $q''^1 := \lim_{j\to\infty} q''_{t_{i_j}}$; the limit exists by (60). Likewise, we can show that

$$v^l := P_{\{l\}}(Sq^l - \widehat{S}q''^l) \in \mathcal{S}^\perp, \qquad l \in [1:n], \tag{69}$$

where $\{q''^l\}_l$ are defined similarly. Because $\dim(\mathcal{S}^\perp) = n - r$, at most $n - r$ of the vectors $\{v^l\}_{l=1}^n$ are linearly independent. Because the supports of $\{v^l\}_l$ are disjoint, it follows that there are at most $n - r$ of the vectors $\{v^l\}_{l=1}^n$ are nonzero. Put differently, there exists an index set $I \subset [1:n]$ of size at least r such that

$$Sq^l = \widehat{S}q''^l, \qquad l \in I. \tag{70}$$

Almost surely, $\{q^l\}_{l\in I} \subset \mathbb{R}^r$ form a basis for $\mathbb{R}^r$, and therefore $\mathcal{S} \subseteq \widehat{\mathcal{S}}$. Because $\widehat{\mathcal{S}} \in \mathbb{G}(n,r)$ by assumption, it follows that $\widehat{\mathcal{S}} = \mathcal{S}$, which completes the proof of Proposition 6.

7.3. Locally Linear Convergence of SNIPE (Proof of Theorems 8 and 10)

At iteration $k \in [2:K]$, SNIPE uses the current estimate $\widehat{\mathcal{S}}_{k-1}$ and the new incomplete block Y_k to produce a new estimate $\widehat{\mathcal{S}}_k$ of the true subspace $\mathcal{S}$. The main challenge here is to compare the new and old principal angles with $\mathcal{S}$, namely, compare $d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_k)$ and $d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1})$. Lemma 16 below, proved in Appendix G, loosely speaking states that $d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_k)$ reduces by a factor of $1 - O(p)$ in expectation in every iteration, when $d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_k) \lesssim p^{\frac{5}{2}}$ and ignoring all other parameters in this qualitative discussion. In other words, when sufficiently small, the estimation error of SNIPE reduces in every iteration, but in expectation. The actual behavior of SNIPE is more nuanced. Indeed, Lemma 16 below also adds that the estimation error $d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_k)$ in fact contracts in *some* iterations by a factor of $1 - Cp$, namely,

$$d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_k) \lesssim (1 - Cp) \cdot d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}),$$

provided that $d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}) \lesssim p^{\frac{5}{2}}$. That is, when sufficiently small, the estimation error of SNIPE reduces in some but not all iterations. In the rest of iterations, the error does not increase by much, namely,

$$d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_k) \approx d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}),$$

with high probability and provided that $d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}) \lesssim p^{\frac{5}{2}}$.

Lemma 16 *Fix $k \in [2:K]$, $\alpha, \nu \geq 1$, and $c > 0$. Let $\mathfrak{E}_{k-1}$ be the event where*

$$p \gtrsim \alpha^2 \log^2 b \log n \frac{\eta(\widehat{\mathcal{S}}_{k-1}) r}{n}, \tag{71}$$

$$d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}) \lesssim \frac{p^{\frac{7}{2}} nb}{\alpha c \log b \sqrt{r \log n}}, \tag{72}$$

and let $\mathfrak{E}'_k$ *be the event where* $\|Q_k\| \le \nu \cdot \sigma_{\min}$*, where* $\sigma_{\min}$ *is the reject threshold in* SNIPE*, see Algorithm 1. Then it holds that*

$$\mathbb{E}\left[d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_k) \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}'_k\right] \le \nu \left(1 - \frac{p}{2} + \frac{p^3 nb}{c}\right) d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}) + \frac{b^{-C\alpha}}{\sqrt{r}}. \tag{73}$$

Moreover, conditioned on $\widehat{\mathcal{S}}_{k-1}$ *and the event* $\mathfrak{E}_{k-1} \cap \mathfrak{E}'_k$*, it holds that*

$$d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_k) \le \nu \left(1 + \frac{p^3 nb}{c}\right) d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}), \tag{74}$$

except with a probability of at most $b^{-C\alpha}$*. Lastly, a stronger bound holds conditioned on* $\widehat{\mathcal{S}}_{k-1}$ *and the event* $\mathfrak{E}_{k-1} \cap \mathfrak{E}'_k$*, namely,*

$$d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_k) \le \nu \left(1 - \frac{p}{4} + \frac{p^3 nb}{c}\right) d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}) \tag{75}$$

with a probability of at least

$$\phi_k(\alpha) := 1 - \exp\left(-\frac{C_1 p^2 nb}{\widetilde{\eta}_{k-1}}\right) - b^{-C\alpha}, \tag{76}$$

where

$$\widetilde{\eta}_k = \widetilde{\eta}(P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k) := nb \cdot \frac{\|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k\|_\infty^2}{\|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k\|_F^2}. \tag{77}$$

Let us now use Lemma 16 to complete the proofs of Theorems 8 and 10.

7.3.1. Proof of Theorem 8

With the choice of $c = 4p^2 nb$ and $\nu = 1/\sqrt{1 - p/4}$, (73) reads as

$$\begin{aligned}
\mathbb{E}\left[d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_k) \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}'_k\right] &\le \nu \left(1 - \frac{p}{2} + \frac{p^3 nb}{c}\right) d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}) + \frac{b^{-C\alpha}}{\sqrt{r}} \qquad \text{(see (73))} \\
&= \sqrt{1 - \frac{p}{4}} d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}) + \frac{b^{-C\alpha}}{\sqrt{r}} \\
&\le \left(1 - \frac{p}{8}\right) d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}) + \frac{b^{-C\alpha}}{\sqrt{r}}.
\end{aligned} \tag{78}$$

With the choice of

$$\alpha = -\frac{C \log\left(p\sqrt{r} d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1})/16\right)}{\log b}, \tag{79}$$

for an appropriate constant C above, the bound in (78) simplifies to

$$\begin{aligned}
\mathbb{E}\left[d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_k) \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}'_k\right] &\le \left(1 - \frac{p}{8}\right) d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}) + \frac{b^{-C\alpha}}{\sqrt{r}} \qquad \text{(see (78))} \\
&\le \left(1 - \frac{p}{8}\right) d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}) + \frac{p}{16} d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1})
\end{aligned}$$

$$\leq \left(1 - \frac{p}{16}\right) d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}). \tag{80}$$

Lastly we remove the conditioning on $\mathfrak{E}'_k$ above. Using the law of total expectation, we write that

$$\begin{aligned}
&\mathbb{E}\left[d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_k) \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}\right] \\
&= \mathbb{E}\left[d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_k) \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}'_k\right] \cdot \Pr[\mathfrak{E}'_k] + \mathbb{E}\left[d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_k) \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}'^C_k\right] \cdot \Pr[\mathfrak{E}'^C_k] \\
&\leq \mathbb{E}\left[d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_k) \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}'_k\right] + \Pr[\mathfrak{E}'^C_k] \qquad \text{(see (10))} \\
&\leq \left(1 - \frac{p}{16}\right) d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}) + \Pr[\mathfrak{E}'^C_k] \qquad \text{(see (80))} \\
&\leq \left(1 - \frac{p}{32}\right) d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}),
\end{aligned} \tag{81}$$

where the last line holds if

$$\Pr[\mathfrak{E}'^C_k] \leq \frac{p}{32} d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}).$$

With the choice of c, ν, α above, let us also rewrite the event $\mathfrak{E}_{k-1}$ in Lemma 16. First, we rewrite (72) as

$$\begin{aligned}
d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1}) &\lesssim \frac{p^{\frac{7}{2}} nb}{\alpha c \log b \sqrt{r \log n}} = \frac{C p^{\frac{3}{2}}}{\alpha \log b \sqrt{r \log n}} \\
&= -\frac{C p^{\frac{3}{2}}}{\log\left(p\sqrt{r} d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1})/16\right) \sqrt{r \log n}}. \qquad \text{(see (79))}
\end{aligned} \tag{82}$$

Second, we replace the coherence $\eta(\widehat{\mathcal{S}}_{k-1})$ in (71) with the simpler quantity $\eta(\mathcal{S})$. We can do so thanks to Lemma 20 which roughly speaking states that a pair of subspaces $\mathcal{A}$ and $\mathcal{B}$ with a small principal angle have similar coherences, namely, $\theta_1(\mathcal{A}, \mathcal{B}) \approx 0 \implies \eta(\mathcal{A}) \approx \eta(\mathcal{B})$. More concretely, note that

$$\begin{aligned}
\sqrt{\eta(\widehat{\mathcal{S}}_{k-1})} &\leq \sqrt{\eta\left(\mathcal{S}\right)} + d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{k-1})\sqrt{n} \qquad \text{(see Lemma 20)} \\
&\leq \sqrt{\eta(\mathcal{S})} + C p^{\frac{3}{2}} \sqrt{\frac{n}{r \log n}} \qquad \text{(see (82))} \\
&\leq \sqrt{\eta(\mathcal{S})} + 1 \qquad \left(\text{if } p \lesssim \frac{1}{\sqrt{nb}}\right) \\
&\leq 2\sqrt{\eta(\mathcal{S})}. \qquad \text{(see (15))}
\end{aligned} \tag{83}$$

This completes the proof of Theorem 8.

7.3.2. Proof of Theorem 10

For $K_0 \in [1 : K]$, we condition on $\widehat{\mathcal{S}}_{K_0}$. For positive c to be set later, suppose that

$$d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{K_0}) \lesssim \frac{e^{-\frac{C p^3 nb}{\tilde{\eta}}} p^{\frac{7}{2}} nb}{\alpha c \log b \sqrt{\log n}}. \tag{84}$$

In particular, (84) implies that the error at iteration K_0 is small enough to activate Lemma 16, see (72). For $\nu \geq 1$ to be set later, we condition for now on the event

$$\mathfrak{E}' := \cap_{k=K_0+1}^{K} \mathfrak{E}'_k, \tag{85}$$

where the event $\mathfrak{E}'_k$ was defined in Lemma 16. Suppose also that (71) holds for every $k \in [K_0+1:K]$, namely,

$$p \gtrsim \max_{k\in[K_0:K-1]} \eta(\widehat{\mathcal{S}}_k) \cdot \frac{r\log^2 b \log n}{n}, \tag{86}$$

which will next allow us to apply Lemma 16 repeatedly to all iterations in the interval $[K_0+1:K]$. With the success probability $\phi_k(\alpha)$ defined in Lemma 16, let us also define

$$\phi(\alpha) := \min_{k\in[K_0+1:K]} \phi_k(\alpha) = 1 - \exp\left(-\frac{C_1 p^2 nb}{\widetilde{\eta}}\right) - b^{-C\alpha}, \qquad \widetilde{\eta} := \max_{k\in[K_0+1:K]} \widetilde{\eta}_k \geq 1, \tag{87}$$

where the inequality above follows because $\widetilde{\eta}_k \geq 1$ for every k, see (77). We now partition $[K_0+1:K]$ into (non-overlapping) intervals $\{I_i\}_i$, each with the length

$$l = \frac{C_2 \log b \log(K-K_0)}{\phi(\alpha)}, \tag{88}$$

except possibly the last interval which might be shorter. Consider one of these intervals, say I_1. Then by Lemma 16 and the union bound, (74) holds for every iteration $k \in I_1$ except with a probability of at most $l \cdot b^{-C\alpha}$ because the length of I_1 is l. That is, the estimation error does not increase by much in every iteration in the interval I_1. In some of these iterations, the error in fact reduces. More specifically, (75) holds in iteration k with a probability of at least $\phi_k(\alpha)$, see (76). While $b^{-C\alpha}$ in (76) can be made arbitrary small by increasing the tuning parameter α, this of course would not necessarily make $\phi_k(\alpha)$ arbitrary close to one. That is, there is a sizable chance that the estimation error does not contract in iteration k. However, (75) holds at least in one iteration in the interval I_1 except with a probability of at most

$$(1-\phi(\alpha))^l \leq e^{-\phi(\alpha)l} = b^{-C_2\log(K-K_0)}. \qquad (1+a\leq e^a)$$

Therefore, except with a probabilty of at most $lb^{-C\alpha} + b^{-C_2\log(K-K_0)}$, (75) holds at least once and (74) holds for all iterations in the interval I_1. It immediately follows that

$$\begin{aligned}
&\frac{d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{K_0+l})}{d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{K_0})} \\
&\leq \nu\left(1 - \frac{p}{4} + \frac{p^3 nb}{c}\right) \cdot \nu^{l-1}\left(1 + \frac{p^3 nb}{c}\right)^{l-1} \qquad \text{(see (74,75))} \\
&\leq \nu^l \exp\left(-\frac{p}{4} + \frac{p^3 np}{c} + (l-1)\frac{p^3 nb}{c}\right) \qquad (1+a\leq e^a) \\
&= \nu^l \exp\left(-\frac{p}{4} + \frac{lp^3 nb}{c}\right),
\end{aligned} \tag{89}$$

except with a probability of at most

$$lb^{-C\alpha} + b^{-C_2 \log(K-K_0)}. \tag{90}$$

In particular, suppose that

$$c \geq 4lp^2nb, \tag{91}$$

so that the exponent in the last line of (89) is negative. Let

$$i_{\max} := \left\lfloor \frac{K-K_0}{l} \right\rfloor,$$

and note that

$$\frac{d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_K)}{d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{K_0})} = \prod_{i=1}^{i_{\max}} \frac{d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{K_0+il})}{d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{K_0+(i-1)l})} \cdot \frac{d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_K)}{d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{K_0+i_{\max}l})}. \tag{92}$$

By applying the bound in (89) to all intervals $\{I_i\}_i$, we then conclude that

$$\begin{aligned}
&\frac{d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_K)}{d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{K_0})} \\
&= \prod_{i=1}^{i_{\max}} \frac{d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{K_0+il})}{d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{K_0+(i-1)l})} \cdot \frac{d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_K)}{d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{K_0+i_{\max}l})} \qquad \text{(see (92))} \\
&\leq \left(\nu^l \exp\left(-\frac{p}{4} + \frac{lp^3nb}{c}\right)\right)^{\left\lfloor \frac{K-K_0}{l} \right\rfloor} \cdot \nu^l \left(1 + \frac{p^3nb}{c}\right)^l \qquad \text{(see (89,74))} \\
&\leq \left(\nu^l \exp\left(-\frac{p}{4} + \frac{lp^3nb}{c}\right)\right)^{\left\lfloor \frac{K-K_0}{l} \right\rfloor} \cdot \nu^l \exp\left(\frac{lp^3nb}{c}\right) \qquad (1+a \leq e^a) \\
&\leq \nu^{K-K_0} \exp\left(\left\lfloor \frac{K-K_0}{2l} \right\rfloor \left(-\frac{p}{4} + \frac{lp^3nb}{c}\right)\right) \cdot \exp\left(\frac{lp^3nb}{c}\right) \\
&\leq \nu^{K-K_0} \exp\left(\frac{K-K_0}{2l} \left(-\frac{p}{4} + \frac{lp^3nb}{c}\right)\right) \cdot \exp\left(\frac{lp^3nb}{c}\right) \quad \text{(if } K-K_0 \geq l \text{ and (91) holds)} \\
&\leq \nu^{K-K_0} \exp\left(\frac{K-K_0}{2l} \left(-\frac{p}{4} + \frac{3lp^3nb}{c}\right)\right), \qquad \text{(if } K-K_0 \geq l\text{)}
\end{aligned} \tag{93}$$

except with a probability of at most

$$\begin{aligned}
&\left\lceil \frac{K-K_0}{l} \right\rceil \left(lb^{-C\alpha} + b^{-C_2 \log(K-K_0)}\right) \\
&\quad \leq \frac{2(K-K_0)}{l} \left(lb^{-C\alpha} + b^{-C_2 \log(K-K_0)}\right) \qquad \text{(if } K-K_0 \geq l\text{)} \\
&\quad \leq (K-K_0)e^{-C\alpha} + b^{-CC_2 \log(K-K_0)},
\end{aligned} \tag{94}$$

which follows from an application of the union bound to the failure probability in (90). With the choice of $\alpha = \alpha' \log b \log(K-K_0)$ with sufficiently large α', the failure probability in (94) simplifies to

$$(K-K_0)e^{-C\alpha} + b^{-C\log(K-K_0)} = (K-K_0) \cdot (K-K_0)^{-C\alpha' \log b} + b^{-C\log(K-K_0)}$$

$$\begin{aligned}
&\leq (K-K_0)^{-C\alpha' \log b} + b^{-C\log(K-K_0)} \\
&= b^{-C\alpha' \log(K-K_0)} + b^{-C\log(K-K_0)} \\
&\leq b^{-C\log(K-K_0)}. \qquad (95)
\end{aligned}$$

The next step involves elementary bookkeeping to upper-bound the last line of (93). Suppose that

$$\alpha \gtrsim \frac{\log\left(\frac{C\widetilde{\eta}}{p^2 nb}\right)}{\log b}, \qquad (96)$$

$$p \lesssim \frac{1}{\sqrt{nb}}. \qquad (97)$$

Using (96) and (97) with appropriate constants replacing $\gtrsim$ and $\lesssim$ above, we may verify that

$$b^{-C\alpha} \leq \frac{C_1 p^2 nb}{4\widetilde{\eta}}, \qquad \text{(see (96))} \qquad (98)$$

$$\frac{C_1 p^2 nb}{\widetilde{\eta}} \lesssim 2, \qquad \text{((97) and } \widetilde{\eta} \geq 1) \qquad (99)$$

$$\begin{aligned}
\phi(\alpha) &= 1 - \exp\left(-\frac{C_1 p^2 nb}{\widetilde{\eta}}\right) - b^{-C\alpha} \qquad \text{(see (87))} \\
&\geq \frac{1}{2} \cdot \frac{C_1 p^2 nb}{\widetilde{\eta}} - b^{-C\alpha} \qquad \left(\text{(99) and } e^{-a} \leq 1 - \frac{a}{2} \text{ for } a \lesssim 2\right) \\
&\geq \frac{C_1 p^2 nb}{4\widetilde{\eta}}, \qquad \text{(see (98))} \qquad (100)
\end{aligned}$$

$$\begin{aligned}
l &= \frac{C_2 \log b \log(K-K_0)}{\phi(\alpha)} \qquad \text{(see (88))} \\
&\leq \frac{4C_2 \widetilde{\eta} \log b \log(K-K_0)}{C_1 p^2 nb}. \qquad \text{(see (100))} \qquad (101)
\end{aligned}$$

Now with the choice of

$$c = \frac{96 C_2}{C_1} \widetilde{\eta} \log b \log(K-K_0), \qquad (102)$$

we may verify that

$$-\frac{p}{4} + \frac{3lp^3 nb}{c} \leq -\frac{p}{8}, \qquad \text{(see (101,102))} \qquad (103)$$

and, revisiting (93), we find that

$$\begin{aligned}
&\frac{d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_K)}{d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_{K_0})} \\
&\leq \nu^{K-K_0} \exp\left(\frac{K-K_0}{2l}\left(-\frac{p}{4} + \frac{3lp^3 nb}{c}\right)\right) \qquad \text{(see (93))}
\end{aligned}$$

$$
\begin{aligned}
&\le \nu^{K-K_0} \exp\left(-\frac{(K-K_0)p}{16l}\right) \qquad \text{(see (103))}\\
&= \left(\nu \exp\left(-\frac{p}{16l}\right)\right)^{K-K_0}\\
&\le \left(\nu \exp\left(-\frac{Cp^3nb}{\widetilde{\eta}\log b \log(K-K_0)}\right)\right)^{K-K_0} \qquad \text{(see (101))}\\
&\le \nu^{K-K_0}\left(1-\frac{Cp^3nb}{\widetilde{\eta}\log b \log(K-K_0)}\right)^{K-K_0} \qquad \text{((97) and } \widetilde{\eta}\ge 1)\\
&\le \left(1-\frac{Cp^3nb}{\widetilde{\eta}\log b \log(K-K_0)}\right)^{K-K_0}, \qquad (104)
\end{aligned}
$$

where we set

$$\nu = 1+\frac{Cp^3nb}{\widetilde{\eta}\log b\log(K-K_0)}, \qquad (105)$$

for an appropriate choice of constant C. To reiterate, conditioned on the event $\mathfrak{E}'$ in (85), (104) is valid provided that (84,86,96,97) hold and except with the probability of at most $b^{-C\log(K-K_0)}$, see (95). In particular, to better interpret (86), we next replace the coherence $\eta(\widehat{\mathcal{S}}_k)$ therein with the simpler quantity $\eta(\mathcal{S})$. We can do so thanks to Lemma 20 which roughly speaking states that a pair of subspaces $\mathcal{A}$ and $\mathcal{B}$ with a small principal angle have similar coherences, namely, $\theta_1(\mathcal{A},\mathcal{B})\approx 0 \implies \eta(\mathcal{A})\approx\eta(\mathcal{B})$. More concretely, Lemma 20 implies that

$$\sqrt{\eta(\widehat{\mathcal{S}}_k)} \le \sqrt{\eta\left(\mathcal{S}\right)} + d_{\mathbb{G}}(\mathcal{S},\widehat{\mathcal{S}}_k)\sqrt{n}. \qquad (106)$$

for every k. To bound the distance in the last line above, we observe that (104) holds also after replacing K with any $k\in[K_0+l:K]$, implying in particular that

$$d_{\mathbb{G}}(\mathcal{S},\widehat{\mathcal{S}}_k)\le d_{\mathbb{G}}(\mathcal{S},\widehat{\mathcal{S}}_{K_0}), \qquad k\in[K_0+l:K]. \qquad (107)$$

When $k\in[K_0+1:K_0+l-1]$ however, we cannot guarantee that the error reduces and all we can say is that that the error does not increase by much. That is, for every $k\in[K_0+1:K_0+l-1]$, we have that

$$
\begin{aligned}
d_{\mathbb{G}}(\mathcal{S},\widehat{\mathcal{S}}_k) &\le \nu^{k-K_0}\left(1+\frac{p^3nb}{c}\right)^{k-K_0} d_{\mathbb{G}}(\mathcal{S},\widehat{\mathcal{S}}_{K_0}) \qquad \text{(see (74))}\\
&\le \nu^{l}\left(1+\frac{p^3nb}{c}\right)^{l} d_{\mathbb{G}}(\mathcal{S},\widehat{\mathcal{S}}_{K_0})\\
&= \nu^{l}\left(1+\frac{Cp^3nb}{\widetilde{\eta}\log b\log(K-K_0)}\right)^{l} d_{\mathbb{G}}(\mathcal{S},\widehat{\mathcal{S}}_{K_0})\\
&\le \left(1+\frac{Cp^3nb}{\widetilde{\eta}\log b\log(K-K_0)}\right)^{2l} d_{\mathbb{G}}(\mathcal{S},\widehat{\mathcal{S}}_{K_0}), \qquad \text{(see (105))} \qquad (108)
\end{aligned}
$$

with an appropriate choice of C in (105). We continue by writing that

$$d_{\mathbb{G}}(\mathcal{S},\widehat{\mathcal{S}}_k)\le\left(1+\frac{Cp^3nb}{\widetilde{\eta}\log b\log(K-K_0)}\right)^{2l} d_{\mathbb{G}}(\mathcal{S},\widehat{\mathcal{S}}_{K_0}) \qquad \text{(see (108))}$$

$$
\begin{aligned}
&\leq \exp\left(\frac{Cp^3nb}{\widetilde{\eta}\log b\log(K-K_0)}\cdot l\right) d_{\mathbb{G}}(\mathcal{S},\widehat{\mathcal{S}}_{K_0}) \qquad (1+a\leq e^a)\\
&\leq e^{Cp} d_{\mathbb{G}}(\mathcal{S},\widehat{\mathcal{S}}_{K_0}) \qquad (\text{see (101)})\\
&\lesssim d_{\mathbb{G}}(\mathcal{S},\widehat{\mathcal{S}}_{K_0}). \qquad (p\leq 1)
\end{aligned} \tag{109}
$$

Combining (107) and (109), we arrive at

$$
d_{\mathbb{G}}(\mathcal{S},\widehat{\mathcal{S}}_k)\lesssim d_{\mathbb{G}}(\mathcal{S},\widehat{\mathcal{S}}_{K_0}), \tag{110}
$$

for every $k\in[K_0+1:K]$, provided that (84,86,96,97) hold and except with a probability of at most $b^{-C\log(K-K_0)}$, see (95). Substituting the above bound into (106) yields that

$$
\begin{aligned}
\sqrt{\eta(\widehat{\mathcal{S}}_k)} &\leq \sqrt{\eta(\mathcal{S})}+d_{\mathbb{G}}(\mathcal{S},\widehat{\mathcal{S}}_k)\sqrt{n} \qquad (\text{see (106)})\\
&\leq \sqrt{\eta(\mathcal{S})}+Cd_{\mathbb{G}}(\mathcal{S},\widehat{\mathcal{S}}_{K_0})\sqrt{n} \qquad (\text{see (110)})\\
&\leq \sqrt{\eta(\mathcal{S})}+\frac{Cp^{\frac{7}{2}}n^{\frac{3}{2}}b}{\alpha c\log b\sqrt{\log n}} \qquad (\text{see (84)})\\
&\leq \sqrt{\eta(\mathcal{S})}+\frac{Cp^{\frac{7}{2}}n^{\frac{3}{2}}b}{\alpha\widetilde{\eta}\log^2 b\sqrt{\log n}\log(K-K_0)} \qquad (\text{see (102)})\\
&\leq \sqrt{\eta(\mathcal{S})}+1. \qquad ((97)\text{ and }\widetilde{\eta}\geq 1)\\
&\leq 2\sqrt{\eta(\mathcal{S})}. \qquad (\eta(\mathcal{S})\geq 1)
\end{aligned} \tag{111}
$$

Plugging back the bound above into (86) yields that

$$
p\gtrsim \log^2 b\log n\frac{\eta(\mathcal{S})r}{n}. \tag{112}
$$

To summarize, conditioned on the event $\mathfrak{E}'$, we have established that (104) is valid under (84,96,97,112) and except with a probability of at most $b^{-C\log(K-K_0)}$. The event $\mathfrak{E}'=\cap_{k=K_0+1}^{K}\mathfrak{E}'_k$ itself holds except with a probability of at most $\sum_{k=K_0+1}^{K}\Pr[\mathfrak{E}'^C_k]$ by the union bound. With an application of the law of total probability, (104) is therefore valid except with a probability of at most $b^{-C\log(K-K_0)}+\sum_{k=K_0+1}^{K}\Pr[\mathfrak{E}'^C_k]$. This completes the proof of Theorem 10.

Acknowledgments

AE would like to thank Anand Vidyashankar and Chris Williams for separately pointing out the possibility of a statistical interpretation of SNIPE, as discussed at the end of Section 3. AE is also extremely grateful to Raphael Hauser for his helpful insights. Lastly, the authors would like to acknowledge and thank Dehui Yang for his involvement in the early phases of this project. For this project, AE was supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1 and partially by the Turing Seed Funding grant SF019. GO and LB were supported by ARO Grant W911NF-14-1-0634. LB was also supported by DARPA grant 16-43-D3M-FP-037. MBW was partially supported by NSF grant CCF-1409258 and NSF CAREER grant CCF-1149225.

Appendix A. Toolbox

This section collects a number of standard results for the reader's convenience. We begin with the following large-deviation bounds that are repeatedly used in the rest of the appendices (Gross, 2011; Tropp, 2012). Throughout, C is a universal constant the value of which might change in every appearance.

Lemma 17 [Hoeffding inequality] *Let $\{z_i\}_i$ be a finite sequence of zero-mean independent random variables and assume that almost surely every z_i belongs to a compact interval of length l_i on the real line. Then, for positive α and except with a probability of at most $e^{-C\alpha^2/\sum_i l_i^2}$, it holds that $\sum_i z_i \le \alpha$.*

Lemma 18 [Matrix Bernstein inequality for spectral norm] *Let $\{Z_i\}_i \subset \mathbb{R}^{n\times b}$ be a finite sequence of zero-mean independent random matrices, and set*

$$\beta := \max_i \|Z_i\|,$$

$$\sigma^2 := \left\|\sum_i \mathbb{E}\left[Z_i^* Z_i\right]\right\| \vee \left\|\sum_i \mathbb{E}\left[Z_i Z_i^*\right]\right\|.$$

Then, for $\alpha \ge 1$ and except with a probability of at most $e^{-C\alpha}$, it holds that

$$\left\|\sum_i Z_i\right\| \lesssim \alpha \cdot \max\left(\log(n\vee b)\cdot\beta, \sqrt{\log(n\vee b)}\cdot\sigma\right).$$

For two r-dimensional subspaces $\mathcal{A}$ and $\mathcal{B}$ with principal angles $\theta_1(\mathcal{A},\mathcal{B}) \ge \theta_2(\mathcal{A},\mathcal{B}) \ge \cdots \ge \theta_r(\mathcal{A},\mathcal{B})$, recall the following useful identities about the principal angles between them:

$$\sin\left(\theta_1\left(\mathcal{A},\mathcal{B}\right)\right) = \|P_{\mathcal{A}^\perp} P_{\mathcal{B}}\| = \|P_{\mathcal{A}} - P_{\mathcal{B}}\|, \tag{113}$$

$$\sqrt{\sum_{i=1}^r \sin^2\left(\theta_i\left(\mathcal{A},\mathcal{B}\right)\right)} = \|P_{\mathcal{A}^\perp} P_B\|_F = \frac{1}{\sqrt{2}}\|P_{\mathcal{A}} - P_{\mathcal{B}}\|_F. \tag{114}$$

Note also the following perturbation bound that is slightly stronger than the standard ones, but proved similarly nonetheless (Wedin, 1972).

Lemma 19 [Perturbation bound] *Fix a rank-r matrix A and let $\mathcal{A} = \mathrm{span}(A)$. For matrix B, let B_r be a rank-r truncation of B obtained via SVD and set $\mathcal{B}_r = \mathrm{span}(B_r)$. Then, it holds that*

$$\|P_{\mathcal{A}} - P_{\mathcal{B}_r}\| = \|P_{\mathcal{A}^\perp} P_{\mathcal{B}_r}\| \le \frac{\|P_{\mathcal{A}^\perp} B\|}{\sigma_r(B)} \le \frac{\|B-A\|}{\sigma_r(B)} \le \frac{\|B-A\|}{\sigma_r(A) - \|B-A\|},$$

$$\frac{1}{\sqrt{2}}\|P_{\mathcal{A}} - P_{\mathcal{B}_r}\|_F = \|P_{\mathcal{A}^\perp} P_{\mathcal{B}_r}\|_F \le \frac{\|P_{\mathcal{A}^\perp} B\|_F}{\sigma_r(B)} \le \frac{\|B-A\|_F}{\sigma_r(B)},$$

where $\sigma_r(A)$ is the r largest singular value of A.

Proof Let $B_{r^+} := B - B_r$ denote the residual and note that

$$\begin{aligned}
\|P_{\mathcal{A}^\perp} P_{\mathcal{B}_r}\| &= \left\| P_{\mathcal{A}^\perp} B_r B_r^\dagger \right\| \qquad (\mathcal{B}_r = \mathrm{span}(B_r)) \\
&= \left\| P_{\mathcal{A}^\perp} \left(B - B_{r^+}\right) B_r^\dagger \right\| \qquad (B = B_r + B_{r^+}) \\
&= \left\| P_{\mathcal{A}^\perp} B B_r^\dagger \right\| \qquad \left(\mathrm{span}\left(B_{r^+}^*\right) \perp \mathrm{span}\left(B_r^*\right) = \mathrm{span}\left(B_r^\dagger\right)\right) \\
&\leq \|P_{\mathcal{A}^\perp} B\| \cdot \|B_r^\dagger\| \\
&= \frac{\|P_{\mathcal{A}^\perp} B\|}{\sigma_r\left(B_r\right)} \\
&= \frac{\|P_{\mathcal{A}^\perp} (B - A)\|}{\sigma_r\left(B_r\right)} \\
&\leq \frac{\|B - A\|}{\sigma_r\left(B_r\right)} \\
&\leq \frac{\|B - A\|}{\sigma_r\left(A\right) - \|B - A\|}. \qquad \text{(Weyl's inequality)}
\end{aligned}$$

The proof is identical for the claim with the Frobenius norm and is therefore omitted. ■

Lastly, let us record what happens to the coherence of a subspace under a small perturbation, see (14).

Lemma 20 [Coherence under perturbation] *Let $\mathcal{A}, \mathcal{B}$ be two r-dimensional subspaces in $\mathbb{R}^n$, and let $d_{\mathbb{G}}(\mathcal{A}, \mathcal{B})$ denote their distance, see (10). Then their coherences are related as*

$$\sqrt{\eta\left(\mathcal{B}\right)} \leq \sqrt{\eta\left(\mathcal{A}\right)} + d_{\mathbb{G}}(\mathcal{A}, \mathcal{B})\sqrt{n}.$$

Proof Let $\theta_i = \theta_i(\mathcal{A}, \mathcal{B})$ be the shorthand for the ith principal angle between the subspaces $\mathcal{A}$ and $\mathcal{B}$. It is well-known (Golub and Van Loan, 2013) that there exist orthonormal bases $A, B \in \mathbb{R}^{n \times r}$ for the subspaces $\mathcal{A}$ and $\mathcal{B}$, respectively, such that

$$A^* B = \mathrm{diag}\left(\begin{bmatrix} \cos\theta_1 & \cos\theta_2 & \cdots & \cos\theta_r \end{bmatrix}\right) =: \Gamma \in \mathbb{R}^{r \times r}, \tag{115}$$

where $\mathrm{diag}(a)$ is the diagonal matrix formed from vector a. There also exists $A' \in \mathbb{R}^{n \times r}$ with orthonormal columns such that

$$\left(A'\right)^* B = \mathrm{diag}\left(\begin{bmatrix} \sin\theta_1 & \sin\theta_2 & \cdots & \sin\theta_r \end{bmatrix}\right) =: \Sigma \in \mathbb{R}^{r \times r}, \qquad \left(A'\right)^* A = 0, \tag{116}$$

and, moreover,

$$\mathrm{span}\left(\begin{bmatrix} A & B \end{bmatrix}\right) = \mathrm{span}\left(\begin{bmatrix} A & A' \end{bmatrix}\right). \tag{117}$$

With $\mathcal{A}' = \mathrm{span}(A')$, it follows that

$$B = P_{\mathcal{A}} B + P_{\mathcal{A}'} B = A A^* B + A'(A')^* B = A\Gamma + A'\Sigma. \qquad \text{(see (115) and (116))} \tag{118}$$

Consequently,

$$\sqrt{\eta\left(\mathcal{B}\right)} = \sqrt{\frac{n}{r}} \max_i \|B[i,:]\|_2 \qquad \text{(see (14))}$$

$$
\begin{aligned}
&\leq \sqrt{\frac{n}{r}} \max_i \|A[i,:] \cdot \Gamma\|_2 + \sqrt{\frac{n}{r}} \max_i \left\|A'[i,:] \cdot \Sigma\right\|_2 \quad \text{((118) and triangle inequality)} \\
&\leq \sqrt{\frac{n}{r}} \max_i \|A[i,:]\|_2 \, \|\Gamma\| + \sqrt{\frac{n}{r}} \max_i \left\|A'[i,:]\right\|_2 \|\Sigma\| \\
&= \sqrt{\eta\left(\mathcal{A}\right)} \, \|\Gamma\| + \sqrt{\eta(\mathcal{A}')} \, \|\Sigma\| \qquad \text{(see (14))} \\
&\leq \sqrt{\eta\left(\mathcal{A}\right)} \, \|\Gamma\| + \sqrt{\frac{n}{r}} \, \|\Sigma\| \qquad \left(\eta(\mathcal{A}') \leq \frac{n}{r}\right) \\
&\leq \sqrt{\eta\left(\mathcal{A}\right)} + \sqrt{\frac{n}{r}} \sin\theta_1 \qquad \text{(see (115) and (116))} \\
&= \sqrt{\eta\left(\mathcal{A}\right)} + \sqrt{\frac{n}{r}} \, \|P_{\mathcal{A}^\perp} P_{\mathcal{B}}\| \qquad \text{(see (113))} \\
&\leq \sqrt{\eta\left(\mathcal{A}\right)} + \sqrt{\frac{n}{r}} \, \|P_{\mathcal{A}^\perp} P_{\mathcal{B}}\|_F \\
&= \sqrt{\eta\left(\mathcal{A}\right)} + d_{\mathbb{G}}(\mathcal{A},\mathcal{B})\sqrt{n}, \qquad \text{(see (10))}
\end{aligned} \tag{119}
$$

which completes the proof of Lemma 20. ■

Appendix B. Supplement to Section 3

In this section, we verify that

$$
R_k = \begin{cases} \arg\min \ \ \|P_{\widehat{\mathcal{S}}_{k-1}^\perp} X_k\|_F^2 + \lambda \|P_{\Omega_k^C}(X_k)\|_F^2 \\ P_{\Omega_k}\left(X_k\right) = Y_k, \end{cases} \tag{120}
$$

when $k \geq 2$. The optimization above is over $X_k \in \mathbb{R}^{n\times b}$. First note that Program (120) is separable and equivalent to the following b programs:

$$
R_k[:,j] = \begin{cases} \arg\min \ \ \|P_{\widehat{\mathcal{S}}_{k-1}^\perp} x\|_2^2 + \lambda \|P_{\omega_t^C} x\|_F^2 \\ P_{\omega_t} \cdot x = y_t, \end{cases} \qquad t = (k-1)b + j, \ \ j \in [1:b]. \tag{121}
$$

Above, $R_k[:,j] \in \mathbb{R}^n$ is the jth column of R_k in MATLAB's matrix notation and the optimization is over $x \in \mathbb{R}^n$. To solve the jth program in (121), we make the change of variables $x = y_t + W_{\omega_t^C} \cdot z$. Here, $z \in \mathbb{R}^{n-m}$ and $W_{\omega_t^C} \in \{0,1\}^{n\times(n-m)}$ is defined naturally so that $P_{\omega_t^C} = W_{\omega_t^C} W_{\omega_t^C}^*$. We now rewrite (121) as the following b unconstrained programs:

$$
\begin{aligned}
z_j &:= \arg\min \ \ \|P_{\widehat{\mathcal{S}}_{k-1}^\perp} y_t + P_{\widehat{\mathcal{S}}_{k-1}^\perp} W_{\omega_t^C} z\|_2^2 + \lambda \|P_{\omega_t^C} W_{\omega_t^C} z\|_F^2 \\
&= \arg\min \ \ \|(\widehat{S}_{k-1}^\perp)^* y_t + (\widehat{S}_{k-1}^\perp)^* W_{\omega_t^C} z\|_2^2 + \lambda \|z\|_F^2, \qquad t = (k-1)b + j, \ \ j \in [1:b].
\end{aligned} \tag{122}
$$

Above, $\widehat{S}_{k-1}^\perp \in \mathbb{R}^{n\times(n-r)}$ is an orthonormal basis for the subspace $\widehat{\mathcal{S}}_{k-1}^\perp \in \mathbb{G}(n, n-r)$ and in particular $P_{\widehat{\mathcal{S}}_{k-1}^\perp} = \widehat{S}_{k-1}^\perp (\widehat{S}_{k-1}^\perp)^*$. The optimization above is over $z \in \mathbb{R}^{n-m}$. Note that

$$
z_j = -\left(W_{\omega_t^C}^* P_{\widehat{\mathcal{S}}_{k-1}^\perp} W_{\omega_t^C} + \lambda I_{n-m}\right)^{-1} W_{\omega_t^C}^* P_{\widehat{\mathcal{S}}_{k-1}^\perp} y_t, \qquad j \in [1:b], \tag{123}
$$

are solutions of the least squares programs in (122) when m is large enough. For fixed j, we simplify the expression for z_j as follows:

$$\begin{aligned}
z_j &= -\left(\lambda I_{n-m} + W_{\omega_t^C}^* P_{\widehat{\mathcal{S}}_{k-1}^\perp} W_{\omega_t^C}\right)^{-1} W_{\omega_t^C}^* P_{\widehat{\mathcal{S}}_{k-1}^\perp} y_t \\
&= -\left((1+\lambda) I_{n-m} - W_{\omega_t^C}^* P_{\widehat{\mathcal{S}}_{k-1}} W_{\omega_t^C}\right)^{-1} W_{\omega_t^C}^* P_{\widehat{\mathcal{S}}_{k-1}^\perp} y_t \qquad \left(P_{\widehat{\mathcal{S}}_{k-1}^\perp} = I_n - P_{\widehat{\mathcal{S}}_{k-1}}\right) \\
&= \frac{-1}{1+\lambda}\left(I_{n-m} + W_{\omega_t^C}^* \widehat{S}_{k-1}\left((1+\lambda) I_r - \widehat{S}_{k-1}^* P_{\omega_t^C} \widehat{S}_{k-1}\right)^{-1} \widehat{S}_{k-1}^* W_{\omega_t^C}\right) W_{\omega_t^C}^* P_{\widehat{\mathcal{S}}_{k-1}^\perp} y_t \\
&\qquad \text{(inversion lemma)} \\
&= \frac{-1}{1+\lambda}\left(I_{n-m} + W_{\omega_t^C}^* \widehat{S}_{k-1}\left((1+\lambda) I_r - \widehat{S}_{k-1}^* P_{\omega_t^C} \widehat{S}_{k-1}\right)^{-1} \widehat{S}_{k-1}^* W_{\omega_t^C}\right) W_{\omega_t^C}^* P_{\widehat{\mathcal{S}}_{k-1}^\perp} P_{\omega_t} y_t \\
&\qquad (y_t = P_{\omega_t} y_t) \\
&= \frac{1}{1+\lambda}\left(I_{n-m} + W_{\omega_t^C}^* \widehat{S}_{k-1}\left((1+\lambda) I_r - \widehat{S}_{k-1}^* P_{\omega_t^C} \widehat{S}_{k-1}\right)^{-1} \widehat{S}_{k-1}^* W_{\omega_t^C}\right) W_{\omega_t^C}^* P_{\widehat{\mathcal{S}}_{k-1}} y_t \\
&\qquad \left(W_{\omega_t^C}^* P_{\omega_t} = 0\right) \\
&= \frac{W_{\omega_t^C}^* P_{\widehat{\mathcal{S}}_{k-1}} y_t}{1+\lambda} + \frac{W_{\omega_t^C}^* \widehat{S}_{k-1}}{1+\lambda}\left((1+\lambda) I_r - \widehat{S}_{k-1}^* P_{\omega_t^C} \widehat{S}_{k-1}\right)^{-1} \widehat{S}_{k-1}^* P_{\omega_t^C} P_{\widehat{\mathcal{S}}_{k-1}} y_t \\
&\qquad \left(P_{\omega_t^C} = W_{\omega_t^C} W_{\omega_t^C}^*\right) \\
&= \frac{W_{\omega_t^C}^* P_{\widehat{\mathcal{S}}_{k-1}} y_t}{1+\lambda} \\
&\quad + \frac{W_{\omega_t^C}^* \widehat{S}_{k-1}}{1+\lambda}\left((1+\lambda) I_r - \widehat{S}_{k-1}^* P_{\omega_t^C} \widehat{S}_{k-1}\right)^{-1}\left(\widehat{S}_{k-1}^* P_{\omega_t^C} \widehat{S}_{k-1} - (1+\lambda) I_r\right) \widehat{S}_{k-1}^* y_t \\
&\quad + W_{\omega_t^C}^* \widehat{S}_{k-1}\left((1+\lambda) I_r - \widehat{S}_{k-1}^* P_{\omega_t^C} \widehat{S}_{k-1}\right)^{-1} \widehat{S}_{k-1}^* y_t \\
&= W_{\omega_t^C}^* \widehat{S}_{k-1}\left((1+\lambda) I_r - \widehat{S}_{k-1}^* P_{\omega_t^C} \widehat{S}_{k-1}\right)^{-1} \widehat{S}_{k-1}^* y_t \\
&= W_{\omega_t^C}^* \widehat{S}_{k-1}\left(\lambda I_r + \widehat{S}_{k-1}^* P_{\omega_t} \widehat{S}_{k-1}\right)^{-1} \widehat{S}_{k-1}^* y_t, \qquad (124)
\end{aligned}$$

which means that

$$y_t + W_{\omega_t^C} z_j = y_t + P_{\omega_t^C} \widehat{S}_{k-1}\left(\lambda I_r + \widehat{S}_{k-1}^* P_{\omega_t} \widehat{S}_{k-1}\right)^{-1} \widehat{S}_{k-1}^* y_t, \qquad (125)$$

is a solution of the jth program in (121) which indeed matches the jth column of R_k defined in (1).

Appendix C. Proof of Proposition 24

Let us form the blocks $Q_1 \in \mathbb{R}^{b_1 \times r}$, $S_1 \in \mathbb{R}^{n \times b_1}$, and $\Omega_1 \subseteq [1:n] \times [1:b_1]$ as usual, see Section 3. As in that section, we also write the measurement process as $Y_1 = P_{\Omega_1}(S_1)$, where $P_{\Omega_1}(\cdot)$ projects onto the index set Ω_1. Let us fix Q_1 for now. Also let $Y_{1,r} \in \mathbb{R}^{n \times b_1}$ be a

rank-r truncation of Y_1 obtained via SVD. SNIPE then sets $\widehat{\mathcal{S}}_1 = \text{span}(Y_{1,r})$. Our objective here is to control $\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_1}\|_F$. Since

$$\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_1}\|_F \le \sqrt{r}\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_1}\|, \qquad \left(\widehat{\mathcal{S}}_1 \in \mathbb{G}(n,r)\right) \tag{126}$$

it suffices to bound the spectral norm. Conditioned on Q_1, it is easy to verify that $\mathbb{E}[Y_1] = \mathbb{E}[P_{\Omega_1}(S_1)] = p \cdot S_1$, suggesting that we might consider Y_1 as a perturbed copy of $p \cdot S_1$ and perhaps consider $\widehat{\mathcal{S}}_1 = \text{span}(Y_{1,r})$ as a perturbation of $\mathcal{S} = \text{span}(p \cdot S_1)$. Indeed, the perturbation bound in Lemma 19 dictates that

$$\begin{aligned}
\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_1}\| &\le \frac{\|Y_1 - pS_1\|}{p \cdot \sigma_r(S_1) - \|Y_1 - pS_1\|} \\
&= \frac{\|Y_1 - pS_1\|}{p \cdot \sigma_r(Q_1) - \|Y_1 - pS_1\|} \qquad (S_1 = SQ_1,\ S^*S = I_r) \\
&\le \frac{2}{p} \cdot \frac{\|Y_1 - pS_1\|}{\sigma_r(Q_1)}. \qquad \left(\text{if } \|Y_1 - pS_{Q_1}\| \le \frac{p}{2} \cdot \sigma_r(Q_1)\right)
\end{aligned} \tag{127}$$

It remains to bound the norm in the last line above. To that end, we study the concentration of Y_1 about its expectation by writing that

$$Y_1 - pS_1 = P_{\Omega_1}(S_1) - pS_1 = \sum_{i,j} (\epsilon_{i,j} - p) S_1[i,j] \cdot E_{i,j} =: \sum_{i,j} Z_{i,j}, \tag{128}$$

where $\{\epsilon_{i,j}\}_{i,j} \overset{\text{ind.}}{\sim} \text{Bernoulli}(p)$ and $E_{i,j} \in \mathbb{R}^{n \times b_1}$ is the $[i,j]$th canonical matrix. Additionally, $\{Z_{i,j}\}_{i,j}$ are independent zero-mean random matrices. In order to appeal to the matrix Bernstein inequality (Lemma 18), we compute the β and σ parameters below, starting with β:

$$\begin{aligned}
\|Z_{i,j}\| &= \|(\epsilon_{i,j} - p) S_1[i,j] \cdot E_{i,j}\| \\
&= |(\epsilon_{i,j} - p) S_1[i,j]| \qquad (\|E_{i,j}\| = 1) \\
&\le |S_1[i,j]| \qquad (\epsilon_{i,j} \in \{0,1\}) \\
&\le \|S_1\|_\infty \\
&\le \|S\|_{2\to\infty} \|Q_1\|_{2\to\infty} \qquad (S_1 = SQ_1,\ \|AB^*\|_\infty \le \|A\|_{2\to\infty}\|B\|_{2\to\infty}) \\
&\le \sqrt{\frac{\eta(\mathcal{S})\, r}{n}} \cdot \sqrt{\frac{\eta(\mathcal{Q}_1)\, r}{b_1}} \cdot \|Q_1\| \qquad (\text{see (14)}) \\
&=: \beta.
\end{aligned} \tag{129}$$

Above, $\|A\|_\infty$ and $\|A\|_{2\to\infty}$ return the largest entry of A in magnitude and the largest ℓ_2 norm of the rows of matrix A, respectively. As for σ, we write that

$$\begin{aligned}
\left\| \mathbb{E}\left[\sum_{i,j} Z_{i,j} Z_{i,j}^* \right] \right\| &= \left\| \sum_{i,j} \mathbb{E}\left[(\epsilon_{i,j} - p)^2 \right] S_1[i,j]^2 \cdot E_{i,i} \right\| \\
&= \left\| \sum_{i,j} p(1-p) S_1[i,j]^2 \cdot E_{i,i} \right\| \qquad (\epsilon_{i,j} \sim \text{Bernoulli}(p))
\end{aligned}$$

$$
\begin{aligned}
&\leq p \left\| \sum_{i,j} S_1[i,j]^2 \cdot E_{i,j} \right\| \\
&= p \left\| \sum_{i} \|S_1[i,:]\|_2^2 \cdot E_{i,i} \right\| \\
&= p \max_i \|S_1[i,:]\|_2^2 \\
&= p \, \|S_1\|_{2\to\infty}^2 \\
&\leq p \, \|S\|_{2\to\infty}^2 \cdot \|Q_1\|^2 \qquad (S_1 = SQ_1,\ \|AB\|_{2\to\infty} \leq \|A\|_{2\to\infty}\|B\|) \\
&= p \cdot \frac{\eta(\mathcal{S})\, r}{n} \cdot \|Q_1\|^2. \qquad \text{(see (14))}
\end{aligned} \tag{130}
$$

In a similar fashion, we find that

$$
\begin{aligned}
\left\| \mathbb{E}\left[\sum_{i,j} Z_{i,j}^* Z_{i,j} \right] \right\| &\leq p \left\| \sum_j \|S_1[:,j]\|_2^2 \, E_{j,j} \right\| \\
&= p \, \|S_1^*\|_{2\to\infty}^2 \\
&\leq p \cdot \|S\|^2 \cdot \|Q_1\|_{2\to\infty}^2 \qquad (S_1 = SQ_1,\ \|AB\|_{2\to\infty} \leq \|A\|_{2\to\infty}\|B\|) \\
&\leq p \cdot \frac{\eta(\mathcal{Q}_1)\, r}{b_1} \cdot \|Q_1\|^2, \qquad (\|S\| = 1, \text{ see (14)})
\end{aligned} \tag{131}
$$

and eventually

$$
\begin{aligned}
\sigma^2 &= \left\| \mathbb{E}\left[\sum_{i,j} Z_{i,j}^* Z_{i,j} \right] \right\| \vee \left\| \mathbb{E}\left[\sum_{i,j} Z_{i,j} Z_{i,j}^* \right] \right\| \\
&\leq \frac{pr}{n} \left(1 \vee \frac{n}{b_1} \right) \left(\eta(\mathcal{S}) \vee \eta(\mathcal{Q}_1) \right) \|Q_1\|^2. \qquad \text{(see (130) and (131))}
\end{aligned} \tag{132}
$$

Lastly,

$$
\begin{aligned}
&\max\left(\log(n \vee b_1) \cdot \beta, \sqrt{\log(n \vee b_1)} \cdot \sigma \right) \\
&\lesssim \max\left(\log(n \vee b_1) \cdot \frac{r}{n}, \sqrt{\log(n \vee b_1)} \cdot \sqrt{\frac{pr}{n}} \right) \sqrt{1 \vee \frac{n}{b_1}} \cdot \sqrt{\eta(\mathcal{S}) \vee \eta(\mathcal{Q}_1)} \cdot \|Q_1\| \\
&\qquad \text{(see (129) and (132))} \\
&\leq \sqrt{\log(n \vee b_1)} \cdot \sqrt{\frac{pr}{n}} \sqrt{1 \vee \frac{n}{b_1}} \cdot \sqrt{\eta(\mathcal{S}) \vee \eta(\mathcal{Q}_1)} \cdot \|Q_1\|. \qquad \left(\text{if } p \geq \frac{\log(n \vee b_1)\, r}{n} \right)
\end{aligned} \tag{133}
$$

The Bernstein inequality now dictates that

$$
\|Y_1 - pS_1\| = \left\| \sum_{i,j} Z_{i,j} \right\| \qquad \text{(see (128))}
$$

$$\lesssim \alpha \max\left(\log(n \vee b_1) \cdot \beta, \sqrt{\log(n \vee b_1} \cdot \sigma\right) \qquad \text{(see Lemma 18)}$$
$$\lesssim \alpha \sqrt{\log(n \vee b_1)} \cdot \sqrt{\frac{rp}{n}} \sqrt{1 \vee \frac{n}{b_1}} \cdot \sqrt{\eta\left(\mathcal{S}\right) \vee \eta\left(\mathcal{Q}_1\right)} \cdot \|Q_1\|, \qquad \text{(see (133))} \tag{134}$$

except with a probability of at most $e^{-\alpha}$. In particular, suppose that

$$p \gtrsim \alpha^2 \nu(Q_1)^2 \left(1 \vee \frac{n}{b_1}\right) \frac{\left(\eta\left(\mathcal{S}\right) \vee \eta\left(\mathcal{Q}_1\right)\right) r \log(n \vee b_1)}{n}, \qquad \left(\nu(Q_1) = \frac{\|Q_1\|}{\sigma_r(Q_1)}\right) \tag{135}$$

so that (127) holds. Then, by substituting (134) back into (127) and then applying (126), we find that

$$\begin{aligned}
\frac{\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_1}\|_F}{\sqrt{r}} &\leq \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_1}\| \qquad \text{(see (126))} \\
&\leq \frac{2}{p} \cdot \frac{\|Y_1 - pS_1\|}{\sigma_r(Q_1)} \qquad \text{(see (127))} \\
&\lesssim \alpha \sqrt{\log(n \vee b_1) \cdot \frac{r}{pn} \left(1 \vee \frac{n}{b_1}\right) \left(\eta\left(\mathcal{S}\right) \vee \eta\left(\mathcal{Q}_1\right)\right)} \cdot \nu\left(Q_1\right) \qquad \text{(see (134))} \\
&=: \delta_1\left(\nu(Q_1), \eta(Q_1)\right),
\end{aligned} \tag{136}$$

except with a probability of at most $e^{-\alpha}$ and for fixed Q_1. In order to remove the conditioning on Q_1, fix $\nu \geq 1$, $1 \leq \eta_1 \leq \frac{b_1}{r}$, and recall the following inequality for events $\mathcal{A}$ and $\mathcal{B}$:

$$\Pr\left[\mathcal{A}\right] = \Pr\left[\mathcal{A}|\mathcal{B}\right] \cdot \Pr\left[\mathcal{B}\right] + \Pr\left[\mathcal{A}|\mathcal{B}^C\right] \cdot \Pr\left[\mathcal{B}^C\right] \leq \Pr\left[\mathcal{A}|\mathcal{B}\right] + \Pr\left[\mathcal{B}^C\right]. \tag{137}$$

Set $\mathcal{Q}_1 = \operatorname{span}(Q_1^*)$ and let $\mathcal{E}$ be the event where both $\nu(Q_1) \leq \nu$ and $\eta(\mathcal{Q}_1) \leq \eta_1$. Thanks to the inequality above, we find that

$$\begin{aligned}
&\Pr\left[\frac{\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_1}\|_F}{\sqrt{r}} \gtrsim \delta_1\left(\nu, \eta_1\right)\right] \\
&\leq \Pr\left[\frac{\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_1}\|_F}{\sqrt{r}} \gtrsim \delta\left(\nu, \eta_1\right) \mid \mathcal{E}\right] + \Pr\left[\mathcal{E}^C\right] \qquad \text{(see (137))} \\
&\leq e^{-\alpha} + \Pr\left[\nu(Q_1) > \nu\right] + \Pr\left[\eta(Q_1) > \eta_1\right], \qquad \text{(see (136))}
\end{aligned} \tag{138}$$

which completes the proof of Proposition 24.

Appendix D. Proof of Lemma 13

We conveniently define the orthonormal basis

$$B = \begin{bmatrix} U & U^\perp \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Then the perturbation from $\mathcal{U}$ can be written more compactly as

$$U + U^\perp \Delta' = B \begin{bmatrix} I_r \\ \Delta' \end{bmatrix} \in \mathbb{R}^{n \times r}.$$

In particular, orthogonal projection onto $\text{span}(U + U^\perp \Delta')$ is

$$\begin{aligned}(U + U^\perp\Delta')(U + U^\perp\Delta')^\dagger &= B \begin{bmatrix} I_r \\ \Delta' \end{bmatrix} \cdot \begin{bmatrix} I_r \\ \Delta' \end{bmatrix}^\dagger B^* \\ &= B \begin{bmatrix} I_r \\ \Delta' \end{bmatrix} \left(I_r + \Delta'^*\Delta'\right)^{-1} \begin{bmatrix} I_r & \Delta'^* \end{bmatrix} O^* \\ &= B \begin{bmatrix} I_r & \Delta'^* \\ \Delta' & 0_{n-r} \end{bmatrix} B^* + o(\|\Delta'\|_F), \end{aligned} \tag{139}$$

where $0_{n-r} \in \mathbb{R}^{(n-r)\times(n-r)}$ is the matrix of zeros of size $n - r$. From (30), note also that

$$f_\Omega(X, \mathcal{U}) = \langle P_{\mathcal{U}^\perp}, XX^* \rangle + \lambda \|P_{\Omega^C}(X)\|_F^2 = \left\langle I_n - UU^\dagger, XX^* \right\rangle + \lambda \|P_{\Omega^C}(X)\|_F^2.$$

We can now write that

$$\begin{aligned} &f_\Omega\left(X + \Delta, U + U^\perp\Delta'\right) \\ &= \left\langle I_n - B \begin{bmatrix} I_r & \Delta'^* \\ \Delta' & 0_{n-r} \end{bmatrix} B^*, (X+\Delta)(X+\Delta)^* \right\rangle + 2\lambda P_{\Omega^C}(X) + o(\|\Delta\|_F) + o(\|\Delta'\|_F) \\ &\qquad \text{(see (139))} \\ &= \left\langle B \begin{bmatrix} 0_r & -\Delta'^* \\ -\Delta' & I_{n-r} \end{bmatrix} B^*, (X+\Delta)(X+\Delta)^* \right\rangle + 2\lambda P_{\Omega^C}(X) + o(\|\Delta\|_F) + o(\|\Delta'\|_F) \\ &= \left\langle \begin{bmatrix} 0_r & -\Delta'^* \\ -\Delta' & I_{n-r} \end{bmatrix}, B^*\left(XX^* + \Delta X^* + X\Delta^*\right)B \right\rangle + 2\lambda P_{\Omega^C}(X) + o(\|\Delta\|_F) + o(\|\Delta'\|_F) \\ &= \left\langle \begin{bmatrix} 0_r & 0_{r\times(n-r)} \\ 0_{(n-r)\times r} & I_{n-r} \end{bmatrix}, B^*XX^*B \right\rangle \\ &\quad + \left\langle \begin{bmatrix} 0_r & 0_{r\times(n-r)} \\ 0_{(n-r)\times r} & I_{n-r} \end{bmatrix}, B^*\Delta X^*B + B^*X\Delta^*B \right\rangle \\ &\quad - \left\langle \begin{bmatrix} 0_r & \Delta'^* \\ \Delta' & 0_{n-r} \end{bmatrix}, B^*XX^*B \right\rangle + 2\lambda P_{\Omega^C}(X) + o(\|\Delta\|_F) + o(\|\Delta'\|_F). \end{aligned} \tag{140}$$

We can further simplify the above expansion as

$$\begin{aligned} &f_\Omega\left(X + \Delta, U + U^\perp\Delta'\right) \\ &= f_\Omega(X, \mathcal{U}) + 2\langle \Delta, P_{\mathcal{U}^\perp} X \rangle - 2\left\langle \Delta', (U^\perp)^* XX^* U \right\rangle + 2\lambda P_{\Omega^C}(X) + o(\|\Delta\|_F) + o(\|\Delta'\|_F). \end{aligned} \tag{141}$$

Therefore the partial derivatives of f are

$$\begin{aligned} \partial_X f_\Omega(X, \mathcal{U}) &= 2P_{\mathcal{U}^\perp} X + 2\lambda P_{\Omega^C}(X) \in \mathbb{R}^{n\times T}, \\ \partial_{\mathcal{U}} f_\Omega(X, \mathcal{U}) &= -2(U^\perp)^* XX^* U \in \mathbb{R}^{(n-r)\times r}, \end{aligned} \tag{142}$$

which completes the proof of Lemma 13.

Appendix E. Proof of Lemma 14

By definition in (39), $\{R_{\epsilon'}\}_{\epsilon' \leq \epsilon}$ is a bounded set, see also Program (30) for the definition of $f_{\widehat{\Omega}}$. Therefore there exist a subsequence $\{R_{\epsilon_i}\}_i$ and a matrix $R \in \mathbb{R}^{n \times b}$ such that

$$\lim_{i \to \infty} \epsilon_i = 0, \tag{143}$$

$$\lim_{i \to \infty} \|R_{\epsilon_i} - R\|_F = 0. \tag{144}$$

That is, $\{R_{\epsilon_i}\}_i$ converges to R. On the other hand, for every $\delta \geq 0$, (49) implies that there exists an integer l_δ that depneds on δ and

$$\|R_{k_l,\epsilon} - R_\epsilon\|_F \leq 2\epsilon l' + \delta, \qquad l \geq l_\delta. \tag{145}$$

Restricted to the sequence $\{\epsilon_i\}_i$ above, (145) reads as

$$\|R_{k_l,\epsilon_i} - R_{\epsilon_i}\|_F \leq 2\epsilon_i l' + \delta, \qquad l \geq l_\delta, \tag{146}$$

which, in the limit of $i \to \infty$, yields that

$$\lim_{i \to \infty} \|R_{k_l,\epsilon_i} - R_{\epsilon_i}\|_F \leq \lim_{i \to \infty} 2\epsilon_i + \delta = \delta, \qquad l \geq l_\delta. \tag{147}$$

We used (143) to obtain the identity above. An immediate consequence of (147) is that

$$\lim_{l \to \infty} \lim_{i \to \infty} \|R_{k_l,\epsilon_i} - R_{\epsilon_i}\|_F = 0. \tag{148}$$

Invoking (144), it then follows that

$$\begin{aligned} \lim_{l \to \infty} \lim_{i \to \infty} \|R_{k_l,\epsilon_i} - R\|_F &= \lim_{l \to \infty} \lim_{i \to \infty} \|R_{k_l,\epsilon_i} - R_{\epsilon_i}\|_F \qquad \text{(see (144))} \\ &= 0. \qquad \text{(see (148))} \end{aligned} \tag{149}$$

Exchanging the order of limits above yields that

$$\begin{aligned} &\lim_{i \to \infty} \lim_{l \to \infty} \|R_{k_l,\epsilon_i} - R\|_F \\ &\leq \lim_{i \to \infty} \lim_{l \to \infty} \|R_{k_l,\epsilon_i} - R_{\epsilon_i}\|_F + \lim_{i \to \infty} \|R_{\epsilon_i} - R\|_F \qquad \text{(triangle inequality)} \\ &= \lim_{i \to \infty} \lim_{l \to \infty} \|R_{k_l,\epsilon_i} - R_{\epsilon_i}\|_F \qquad \text{(see (144))} \\ &= 0. \qquad \text{(see (50))} \end{aligned} \tag{150}$$

Therefore, (149) and (150) together state that R_{k,ϵ_i} converges to R as $l, i \to \infty$, namely,

$$\lim_{l,i \to \infty} \|R_{k_l,\epsilon_i} - R\|_F = 0, \tag{151}$$

thereby proving the first claim in Lemma 14, see (51). In order to prove the claim about subspaces in Lemma 14, we proceed as follows. Recall the output of SNIPE, namely, $\{(R_k, \widehat{\mathcal{S}}_k)\}_k$ constructed in Algorithm 1. In light of Section 3, R_k is also the unique minimizer of Program

(9). Recall that $\kappa_l = [k_l - l + 1 : k_l]$ from (33). Recall also the construction of sequence $\{(R_{k,\epsilon}, \widehat{\mathcal{S}}_{k,\epsilon})\}_{k\in\kappa_l}$ in the beginning of Section 7.1 and note that both procedures are initialized identically at the beginning of interval κ_l, namely, $\widehat{\mathcal{S}}_{k_l-l,\epsilon} = \widehat{\mathcal{S}}_{k_l-1}$. Therefore, for fixed l, observe that[4]

$$\lim_{\epsilon\to 0} \|R_{k,\epsilon} - R_k\|_F = 0, \qquad k \in \kappa_l, \tag{152}$$

which, when restricted to $\{\epsilon_i\}_i$, reads as

$$\lim_{i\to\infty} \|R_{k,\epsilon_i} - R_k\|_F = 0, \qquad k \in \kappa_l. \tag{153}$$

By design, every R_k has a spectral gap in the sense that there exists $\tau > 0$ such that

$$\frac{\sigma_r(R_k)}{\sigma_{r+1}(R_k)} \geq 1 + \tau, \tag{154}$$

for every k. Recall that $\widehat{\mathcal{S}}_k$ is the span of top r left singular vectors of R_k. An immediate consequence of (154) is that $\widehat{\mathcal{S}}_k$ is uniquely defined, namely, there are no ties in the spectrum of R_k. By (153), there are no ties in the spectrum of R_{k,ϵ_i} as well for sufficiently large i, namely,

$$\lim_{i\to\infty} \frac{\sigma_r(R_{k,\epsilon_i})}{\sigma_{r+1}(R_{k,\epsilon_i})} \geq 1 + \tau, \qquad k \in \kappa_l. \tag{155}$$

By sending l to infinity above, we find that

$$\begin{aligned} 1 + \tau &\leq \lim_{l\to\infty} \lim_{i\to\infty} \frac{\sigma_r(R_{k_l,\epsilon_i})}{\sigma_{r+1}(R_{k_l,\epsilon_i})} \qquad \text{(see (155))} \\ &= \lim_{l,i\to\infty} \frac{\sigma_r(R_{k_l,\epsilon_i})}{\sigma_{r+1}(R_{k_l,\epsilon_i})} \qquad \text{(see (151))} \\ &= \frac{\sigma_r(R)}{\sigma_{r+1}(R)}. \qquad \text{(see (151))} \end{aligned} \tag{156}$$

Recall that $\widehat{\mathcal{S}}_{k,\epsilon_i}$ is by definition the span of leading r left singular vectors of R_{k,ϵ_i}. Likewise, let $\widehat{\mathcal{S}}$ be the span of leading r left singular vectors of R. An immediate consequence of the second line of (156) is that $\widehat{\mathcal{S}}_{k_l,\epsilon_i}$ is uniquely defined, namely, no ties in the spectrum of R_{k_l,ϵ_i} in the limit of $l, i \to \infty$. The third line of (156) similarly implies that $\widehat{\mathcal{S}}$ is uniquely defined. Given the uniqueness of these subspaces, another implication of (151) is that

$$\lim_{l,i\to\infty} d_{\mathbb{G}}(\widehat{\mathcal{S}}_{k_l,\epsilon_i}, \widehat{\mathcal{S}}) = 0, \tag{157}$$

where $d_{\mathbb{G}}$ is the metric on Grassmannian defined in (10). Lastly we show that SNIPE in the limit produces copies of $(R, \widehat{\mathcal{S}})$ on the interval $\kappa_{l,l'}$. This is done by simply noting that

$$\lim_{l\to\infty} \|R_{k_l} - R\|_F = \lim_{l\to\infty} \lim_{i\to\infty} \|R_{k_l,\epsilon_i} - R\|_F \qquad \text{(see (153))}$$

4. To verify (152), note that for every feasible X_ϵ in Program (38), $X = Y + P_{\widehat{\Omega}^C}(X_\epsilon)$ is feasible for Program (37). Moreover, as $\epsilon \to 0$, $\|X_\epsilon - X\|_F \to 0$ and consequently, by continuity, $|f_{\widehat{\Omega}}(X_\epsilon, \mathcal{U}) - f_{\widehat{\Omega}}(X, \mathcal{U})| \to 0$. On the other hand, by definition, R_k is the unique minimizer of Program (37), namely, $f_{\widehat{\Omega}}(R_k, \mathcal{U}) < f_{\widehat{\Omega}}(X, \mathcal{U})$ for any $X \neq R_k$ feasible for Program (37). For sufficiently small ϵ, it follows that $f_{\widehat{\Omega}}(R_k, \mathcal{U}) < f_{\widehat{\Omega}}(X_\epsilon, \mathcal{U})$ for any X_ϵ that is feasible for Program (38) and $\liminf_{\epsilon\to 0} \|X_\epsilon - R_k\| > 0$. That is, the unique minimizer of Program (38) approaches R_k in the limit, namely, $\lim_{\epsilon\to 0} \|R_{k,\epsilon} - R_k\|_F = 0$.

$$\begin{aligned} &= \lim_{l,i\to\infty} \|R_{k_l,\epsilon_i} - R\|_F \qquad \text{(see (151))} \\ &= 0, \qquad \text{(see (151))} \end{aligned} \tag{158}$$

$$\begin{aligned} \lim_{l\to\infty} d_{\mathbb{G}}(\widehat{\mathcal{S}}_{k_l}, \widehat{\mathcal{S}}) &= \lim_{l\to\infty} \lim_{i\to\infty} d_{\mathbb{G}}(\widehat{\mathcal{S}}_{k_l,\epsilon_i}, \widehat{\mathcal{S}}) \qquad \text{(see (153))} \\ &= \lim_{l,i\to\infty} \lim_{i\to\infty} d_{\mathbb{G}}(\widehat{\mathcal{S}}_{k_l,\epsilon_i}, \widehat{\mathcal{S}}) \qquad \text{(see (157))} \\ &= 0, \qquad \text{(see (157))} \end{aligned} \tag{159}$$

which completes the proof of Lemma 14.

Appendix F. Proof of Lemma 15

By restricting (41) to $\{\epsilon_i\}_i$, we find that

$$\begin{aligned} 0 &= \partial_{\mathcal{U}} f_{\widehat{\Omega}}(R_{k_l,\epsilon_i}, \widehat{\mathcal{S}}_{k,\epsilon_i}) \qquad \text{(see (41))} \\ &= \partial_{\mathcal{U}} f_{\Omega_{k_l}}(R_{k_l,\epsilon_i}, \widehat{\mathcal{S}}_{k,\epsilon_i}), \qquad \text{(see (34))} \end{aligned} \tag{160}$$

for every l, i. By the joint continuity of $\partial_{\mathcal{U}} f_{\Omega_{k_l}}$ in Lemma 13, it follows that

$$\begin{aligned} 0 &= \lim_{l,i\to\infty} \left\| \partial_{\mathcal{U}} f_{\Omega_{k_l}}(R_{k_l,\epsilon_i}, \widehat{\mathcal{S}}_{k_l,\epsilon_i}) \right\|_F \\ &= \lim_{l\to\infty} \left\| \partial_{\mathcal{U}} f_{\Omega_{k_l}}(R, \widehat{\mathcal{S}}) \right\|_F, \qquad \text{(see Lemma 14)} \end{aligned} \tag{161}$$

which establishes (56). To establish (57), we restrict (44) to $\{\epsilon_i\}_i$ and then send i, l to infinity to find that

$$\begin{aligned} 0 &= \lim_{i\to\infty} \lim_{l\to\infty} \|P_{\widehat{\Omega}}(R_{k_l,\epsilon_i}) - \widehat{Y}\|_F \qquad \text{(see (44))} \\ &= \lim_{i\to\infty} \lim_{l\to\infty} \|P_{\Omega_k}(R_{k_l,\epsilon_i}) - Y_{k_l}\|_F \qquad \text{(see (34,36))} \\ &= \lim_{l\to\infty} \|P_{\Omega_{k_l}}(R) - Y_{k_l}\|_F. \qquad \text{(see Lemma 14)} \end{aligned} \tag{162}$$

To establish (58), we restrict (45) to $\{\epsilon_i\}_i$ and then send i to infinity to find that

$$\begin{aligned} 0 &= \lim_{i\to\infty} \lim_{l\to\infty} \left\| \partial_X f_{\widehat{\Omega}}(R_{k_l,\epsilon_i}, \widehat{\mathcal{S}}_{k_l,\epsilon_i}) + \lambda_{k_l,\epsilon_i}(P_{\widehat{\Omega}}(R_{k_l,\epsilon_i}) - \widehat{Y}) \right\|_F^2 \\ &= \lim_{i\to\infty} \lim_{l\to\infty} \left\| P_{\widehat{\Omega}^C} \left(\partial_X f_{\widehat{\Omega}}(R_{k_l,\epsilon_i}, \widehat{\mathcal{S}}_{k_l,\epsilon_i}) \right) \right\|_F^2 \\ &\qquad + \lim_{i\to\infty} \lim_{l\to\infty} \left\| P_{\widehat{\Omega}} \left(\partial_X f_{\widehat{\Omega}}(R_{k_l,\epsilon}, \widehat{\mathcal{S}}_{k_l,\epsilon_i}) + \lambda_{k,\epsilon_i}(R_{k_l,\epsilon_i} - \widehat{Y}) \right) \right\|_F^2 \\ &\geq \lim_{i\to\infty} \lim_{l\to\infty} \left\| P_{\widehat{\Omega}^C} \left(\partial_X f_{\widehat{\Omega}}(R_{k_l,\epsilon_i}, \widehat{\mathcal{S}}_{k_l,\epsilon_i}) \right) \right\|_F^2 \\ &= \lim_{l\to\infty} \left\| P_{\Omega_{k_l}^C} \left(\partial_X f_{\Omega_{k_l}}(R, \widehat{\mathcal{S}}) \right) \right\|_F^2. \qquad \text{(see (34) and Lemma 14)} \end{aligned} \tag{163}$$

This completes the proof of Lemma 15.

Appendix G. Proof of Lemma 16

Recall that by construction in Section 3, the rows of the coefficient matrix $Q_k \in \mathbb{R}^{b\times r}$ are independent copies of the random vector $q \in \mathbb{R}^r$. Setting $S_k = S \cdot Q_k^* \in \mathbb{R}^{n\times b}$, we observe in iteration k each entry of the data block S_k independently with a probability of p, collect the observed entries in $Y_k \in \mathbb{R}^{n\times b}$, supported on the index set $\Omega_k \subseteq [1:n]\times[1:b]$. We write this as $Y_k = P_{\Omega_k}(S_k)$, where the linear operator P_{Ω_k} retains the entries on the index set Ω_k and sets the rest to zero. Recall also that Q_k is obtained by concatenating the coefficient vectors $\{q_t\}_{t=(k-1)b+1}^{kb} \subset \mathbb{R}^r$. To unburden the notation, we enumerate these vectors as $\{q_j\}_{j=1}^b$. Likewise, we use the indexing $\{s_j, y_j, \omega_j\}_{j=1}^b$ for the data vectors, incomplete data vectors, and their supports, respectively.

Given the new incomplete block Y_k at iteration k, we update our estimate of the true subspace $\mathcal{S}$ from the old $\widehat{\mathcal{S}}_{k-1}$ as follows. We calculate the random matrix

$$\mathbb{R}^{n\times b} \ni R_k := Y_k + P_{\Omega_k^C}\left(O\left(Y_k\right)\right), \tag{164}$$

where the linear operator $P_{\Omega_k^C}$ projects onto the complement of index set Ω_k, and

$$O(Y_k) = \left[\cdots \widehat{S}_{k-1}(P_{\omega_j}\widehat{S}_{k-1})^\dagger y_j \cdots\right] \in \mathbb{R}^{n\times b}. \tag{165}$$

Above, $\widehat{S}_{k-1} \in \mathbb{R}^{n\times r}$ is an orthonormal basis for the r-dimensional subspace $\widehat{\mathcal{S}}_{k-1}$. If $\sigma_r(R_k) < \sigma_{\min}$, reject this iteration, see Algorithm 1. Otherwise, let $R_{k,r}$ denote a rank-r truncation of R_k obtained via SVD. Then our updated estimate is $\widehat{\mathcal{S}}_k = \mathrm{span}(R_{k,r})$.

We condition on the subspace $\widehat{\mathcal{S}}_{k-1}$ and the coefficient matrix Q_k for now. To control the estimation error $d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_k)$, our strategy is to treat R_k as a perturbed copy of $S_k = SQ_k^*$ and in turn treat $\widehat{\mathcal{S}}_k = \mathrm{span}(R_{k,r})$ as a perturbed copy of $\mathcal{S} = \mathrm{span}(S_k)$. Indeed, an application of the perturbation bound in Lemma 19 yields that

$$\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_k}\|_F \le \frac{\|P_{\mathcal{S}^\perp} R_k\|_F}{\sigma_r\left(R_k\right)}. \tag{166}$$

To control the numerator above, we begin with some preparation. First, recalling the definition of $O(Y_k)$ from (165), we observe that

$$\begin{aligned} O\left(Y_k\right) &= O\left(P_{\Omega_k}\left(S_k\right)\right) \qquad (Y_k = P_{\Omega_k}\left(S_k\right)) \\ &= O(P_{\Omega_k}(P_{\widehat{\mathcal{S}}_{k-1}} S_k)) + O(P_{\Omega_k}(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k)) \qquad \text{(linearity of } O\text{)} \\ &= P_{\widehat{\mathcal{S}}_{k-1}} S_k + O(P_{\Omega_k}(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k)). \qquad \text{(see (165))} \end{aligned} \tag{167}$$

The above decomposition allows us to rewrite R in (164) as

$$\begin{aligned} R_k &= Y_k + P_{\Omega_k^C}\left(O\left(Y_k\right)\right) \qquad \text{(see (164))} \\ &= P_{\Omega_k}(S_k) + P_{\Omega_k^C}\left(O\left(Y_k\right)\right) \qquad (Y_k = P_{\Omega_k}(S_k)) \\ &= P_{\Omega_k}\left(S_k\right) + P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}_{k-1}} S_k) + [P_{\Omega_k^C} \circ O \circ P_{\Omega_k}](P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) \\ &\qquad \text{(see (167), } f\circ g(x) := f(g(x))) \end{aligned}$$

$$
\begin{aligned}
&= S_k - P_{\Omega_k^C}(S_k) + P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}_{k-1}} S_k) + [P_{\Omega_k^C} \circ O \circ P_{\Omega_k}](P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) \\
&= S_k - P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) + [P_{\Omega_k^C} \circ O \circ P_{\Omega_k}](P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) \\
&= S_k - P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) - [P_{\Omega_k} \circ O \circ P_{\Omega_k}]\left(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k\right) + [O \circ P_{\Omega_k}]\left(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k\right). \qquad (168)
\end{aligned}
$$

Since $P_{\mathcal{S}^\perp} S_k = P_{\mathcal{S}^\perp} S Q_k^* = 0$, it immediately follows that

$$
P_{\mathcal{S}^\perp} R_k = -P_{\mathcal{S}^\perp} \cdot P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) - P_{\mathcal{S}^\perp} \cdot [P_{\Omega_k} \circ O \circ P_{\Omega_k}](P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) + P_{\mathcal{S}^\perp} \cdot [O \circ P_{\Omega_k}]\left(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k\right).
$$

In particular, with an application of the triangle inequality above, we find that

$$
\begin{aligned}
&\|P_{\mathcal{S}^\perp} R_k\|_F \\
&\le \left\| P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) + [P_{\Omega_k} \circ O \circ P_{\Omega_k}](P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) \right\|_F + \left\| P_{\mathcal{S}^\perp} \cdot [O \circ P_{\Omega_k}]\left(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k\right) \right\|_F . \qquad (169)
\end{aligned}
$$

We proceed by controlling each norm on the right-hand side above using the next two technical lemmas, proved in Appendices H and I, respectively.

Lemma 21 *It holds that*

$$
\begin{aligned}
&\mathbb{E}\left[\left\| P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) + [P_{\Omega_k} \circ O \circ P_{\Omega_k}]\left(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k\right) \right\|_F \mid \widehat{\mathcal{S}}_{k-1}, Q_k \right] \\
&\le \sqrt{1-p} \cdot \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F \cdot \|Q_k\|. \qquad (170)
\end{aligned}
$$

For fixed $\widehat{\mathcal{S}}_{k-1}$ and Q_k, we also have

$$
\left\| P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) + [P_{\Omega_k} \circ O \circ P_{\Omega_k}]\left(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k\right) \right\|_F \le \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F \cdot \|Q_k\|. \qquad (171)
$$

and the stronger bound

$$
\left\| P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) + [P_{\Omega_k} \circ O \circ P_{\Omega_k}]\left(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k\right) \right\|_F \le \sqrt{1-p/2} \cdot \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F \cdot \|Q_k\|, \qquad (172)
$$

except with a probability of at most

$$
\exp\left(-\frac{C_1 p^2 n b}{\widetilde{\eta}_k} \right), \qquad (173)
$$

where

$$
\widetilde{\eta}_k = \widetilde{\eta}(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) = nb \cdot \frac{\|P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k\|_\infty^2}{\|P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k\|_F^2}. \qquad (174)
$$

Lemma 22 *For fixed $\widehat{\mathcal{S}}_{k-1}, Q_k$ and $\alpha \ge 1$, it holds that*

$$
\left\| P_{\mathcal{S}^\perp} \cdot [O \circ P_{\Omega_k}]\left(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k\right) \right\|_F \lesssim \alpha \log b \sqrt{\frac{\log n}{p}} \cdot \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\| \cdot \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F \cdot \|Q_k\|, \qquad (175)
$$

except with a probability of at most $b^{-C\alpha}$ and provided that $p \gtrsim \alpha^2 \log^2 b \log n \cdot \eta(\widehat{\mathcal{S}}_{k-1}) r/n$.

We next use Lemmas 21 and 22 to derive two bounds for the numerator of (166), the weaker bound holds with high probability but the stronger bound holds with only some probability. More specifically, substituting (171) and (175) into (169) yields that

$$\begin{aligned}
&\|P_{\mathcal{S}^\perp} R_k\|_F \\
&\leq \left\|P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) + [P_{\Omega_k} \circ O \circ P_{\Omega_k}]\,(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k)\right\|_F + \left\|P_{\mathcal{S}^\perp} \cdot [O \circ P_{\Omega_k}]\,(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k)\right\|_F \\
&\qquad \text{(see (169))} \\
&\leq \left(1 + C\alpha \log b \sqrt{\frac{\log n}{p}} \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|\right) \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F \|Q_k\|, \qquad (176)
\end{aligned}$$

except with a probability of at most $b^{-C\alpha}$ and provided that $p \gtrsim \alpha^2 \log^2 b \log n \cdot \eta(\widehat{\mathcal{S}}_{k-1}) r/n$. For positive c to be set later, let us further assume that

$$\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\| \leq \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F \lesssim \frac{p^{\frac{7}{2}} nb}{c\alpha \log b \sqrt{\log n}}. \qquad (177)$$

With an appropriate constant replacing $\lesssim$ above, (176) simplifies to

$$\|P_{\mathcal{S}^\perp} R_k\|_F \lesssim \left(1 + \frac{p^3 nb}{c}\right) \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F \|Q_k\|. \qquad (178)$$

A stronger bound is obtained by substituting (172, 175) into (169), namely,

$$\begin{aligned}
&\|P_{\mathcal{S}^\perp} R_k\|_F \\
&\leq \left\|P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) + [P_{\Omega_k} \circ O \circ P_{\Omega_k}](P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k)\right\|_F + \left\|P_{\mathcal{S}^\perp} \cdot [O \circ P_{\Omega_k}](P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k)\right\|_F \\
&\qquad \text{(see (169))} \\
&\leq \left(\sqrt{1 - \frac{p}{2}} + \frac{p^3 nb}{c}\right) \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F \|Q_k\| \qquad \text{(see (172,175,177))} \\
&\leq \left(1 - \frac{p}{4} + \frac{p^3 nb}{c}\right) \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F \|Q_k\|, \qquad (179)
\end{aligned}$$

provided that $p \gtrsim \alpha^2 \log^2 b \log n \cdot \eta(\widehat{\mathcal{S}}_{k-1}) r/n$ and except with a probability of at most

$$\exp\left(-\frac{C_1 p^2 nb}{\widetilde{\eta}_k}\right) + b^{-C\alpha}.$$

Note that (178) and (179) offer two alternative bounds for the numerator in the last line (166), which we will next use to complete the proof of Lemma 16.

Fix the subspace $\widehat{\mathcal{S}}_{k-1}$ for now. Let $\mathfrak{E}_{k-1}$ be the event where $p \gtrsim \alpha^2 \log^2 b \log n \cdot \eta(\widehat{\mathcal{S}}_{k-1}) r/n$ and (177) holds. For $\nu \geq 1$ to be set later, let $\mathfrak{E}_k'$ be the event where

$$\|Q_k\| \leq \nu \cdot \sigma_{\min}. \qquad (180)$$

Conditioned on the event $\mathfrak{E}_{k-1} \cap \mathfrak{E}_k'$, we write that

$$\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_k}\|_F \leq \frac{\|P_{\mathcal{S}^\perp} R_k\|_F}{\sigma_r(R_k)} \qquad \text{(see (166))}$$

$$\leq \frac{\|P_{\mathcal{S}^\perp} R_k\|_F}{\sigma_{\min}}$$
$$\lesssim \nu \left(1 + \frac{p^3 nb}{c}\right) \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F, \qquad \text{(see (178))} \tag{181}$$

except with a probability of at most $b^{-C\alpha}$. A stronger bound is obtained from (179), namely,

$$\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_k}\|_F \leq \frac{\|P_{\mathcal{S}^\perp} R_k\|_F}{\sigma_r(R_k)} \qquad \text{(see (166))}$$
$$\leq \nu \left(1 - \frac{p}{4} + \frac{p^3 nb}{c}\right) \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F, \qquad \text{(see (179))} \tag{182}$$

conditioned on the event $\mathfrak{E}_{k-1} \cap \mathfrak{E}'_k$ and except with a probability of at most

$$\exp\left(-\frac{C_1 p^2 nb}{\widetilde{\eta}_k}\right) + b^{-C\alpha}. \tag{183}$$

This completes the proof of the probabilistic claims in Lemma 16, namely, (74) and (75). To complete the proof of Lemma 16, we next derive a bound for the conditional expectation of $\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_k}\|_F$. Let $\mathfrak{E}''_k$ be the event where $p \gtrsim \alpha^2 \log^2 b \log n \cdot \eta(\widehat{\mathcal{S}}_{k-1}) r/n$ and

$$\left\| P_{\mathcal{S}^\perp} \cdot [O \circ P_\Omega] \left(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k\right) \right\|_F \lesssim \frac{p^3 nb}{c} \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F \|Q_k\|. \tag{184}$$

In light of Lemma 22, we have that

$$\Pr[\mathfrak{E}''_k | \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_k, \mathfrak{E}'_k] \geq 1 - b^{-C\alpha}. \tag{185}$$

Using the law of total expectation, we now write that

$$\begin{aligned}
&\mathbb{E}\left[\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_k}\|_F \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}'_k\right] \\
&= \mathbb{E}\left[\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_k}\|_F \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}'_k, \mathfrak{E}''_k\right] \cdot \Pr[\mathfrak{E}''_k | \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}'_k] \\
&\quad + \mathbb{E}\left[\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_k}\|_F \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}'_k, \mathfrak{E}''^C_k\right] \cdot \Pr[\mathfrak{E}''^C_k | \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}'_k] \\
&\leq \mathbb{E}\left[\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_k}\|_F \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}'_k, \mathfrak{E}''_k\right] + \sqrt{r} \Pr[\mathfrak{E}''^C_k | \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}'_k] \qquad \left(\widehat{\mathcal{S}}_k \in \mathbb{G}(n,r)\right) \\
&\leq \mathbb{E}\left[\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_k}\|_F \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}'_k, \mathfrak{E}''_k\right] + \sqrt{r} b^{-C\alpha} \qquad \text{(see (185))} \\
&\leq \mathbb{E}\left[\frac{\|P_{\mathcal{S}^\perp} R_k\|_F}{\sigma_r(R_k)} \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}'_k, \mathfrak{E}''_k\right] + b^{-C\alpha} \qquad \text{((166) and } b \geq r) \\
&\leq \sigma_{\min}^{-1} \mathbb{E}\left[\|P_{\mathcal{S}^\perp} R_k\|_F \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}'_k, \mathfrak{E}''_k\right] + b^{-C\alpha}.
\end{aligned} \tag{186}$$

We next bound the remaining expectation above by writing that

$$\begin{aligned}
&\mathbb{E}\left[\|P_{\mathcal{S}^\perp} R_k\|_F | \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}'_k, \mathfrak{E}''_k\right] \\
&\leq \mathbb{E}\left[\left\| P_{\Omega_k^C} (P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) + [P_{\Omega_k} \circ O \circ P_{\Omega_k}] (P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) \right\|_F \right.
\end{aligned}$$

$$
\begin{aligned}
&\quad + \left\| P_{\mathcal{S}^\perp} \cdot [O \circ P_\Omega] \left(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k \right) \right\|_F \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}_k', \mathfrak{E}_k'' \Big] \qquad \text{(see (169))} \\
&\leq \mathbb{E}\left[\left\| P_{\Omega_k^C} (P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) + [P_{\Omega_k} \circ O \circ P_{\Omega_k}] \left(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k \right) \right\|_F \right. \\
&\quad \left. + \frac{p^3 nb}{c} \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F \|Q_k\| \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}_k', \mathfrak{E}_k'' \right] \qquad \text{(see (184))} \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\| P_{\Omega_k^C} (P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) + [P_{\Omega_k} \circ O \circ P_{\Omega_k}] \left(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k \right) \right\|_F | \widehat{\mathcal{S}}_{k-1}, Q_k \right] \right. \\
&\quad \left. + \frac{p^3 nb}{c} \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F \|Q_k\| | \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}_k', \mathfrak{E}_k'' \right] \\
&\leq \mathbb{E}\left[\sqrt{1-p} \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F \|Q_k\| + \frac{p^3 nb}{c} \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F \|Q_k\| | \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}_k', \mathfrak{E}_k'' \right] \\
&\qquad \text{(see (170))} \\
&= \mathbb{E}\left[\nu \cdot \sigma_{\min} \left(\sqrt{1-p} + \frac{p^3 nb}{c} \right) \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F | \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}_k', \mathfrak{E}_k'' \right] \qquad \text{(see (180))} \\
&= \nu \cdot \sigma_{\min} \left(\sqrt{1-p} + \frac{p^3 nb}{c} \right) \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F \\
&\leq \nu \cdot \sigma_{\min} \left(1 - \frac{p}{2} + \frac{p^3 nb}{c} \right) \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F. \qquad (187)
\end{aligned}
$$

Plugging the bound above back into (186) yields that

$$
\begin{aligned}
&\mathbb{E}\left[\|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_k}\|_F | \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}_k' \right] \\
&\leq \sigma_{\min}^{-1} \mathbb{E}\left[\|P_{\mathcal{S}^\perp} R_k\|_F \,|\, \widehat{\mathcal{S}}_{k-1}, \mathfrak{E}_{k-1}, \mathfrak{E}_k', \mathfrak{E}_k'' \right] + b^{-C\alpha} \qquad \text{(see (186))} \\
&= \nu \left(1 - \frac{p}{2} + \frac{p^3 nb}{c} \right) \|P_{\mathcal{S}^\perp} P_{\widehat{\mathcal{S}}_{k-1}}\|_F + b^{-C\alpha}, \qquad (188)
\end{aligned}
$$

which proves (73) and completes the proof of Lemma 16.

Appendix H. Proof of Lemma 21

Throughout, $\widehat{\mathcal{S}}_{k-1}, Q_k$ is fixed. Recalling the definition of operator $O(\cdot)$ from (165), we write that

$$
\begin{aligned}
&[P_{\Omega_k} \circ O \circ P_{\Omega_k}] \left(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k \right) \\
&= \left[\begin{array}{cccc} \cdots & \left(P_{\omega_j} \widehat{S}_{k-1} \right) \left(P_{\omega_j} \widehat{S}_{k-1} \right) \cdot P_{\omega_j} P_{\widehat{\mathcal{S}}_{k-1}^\perp} S q_j & \cdots \end{array} \right] \qquad \text{(see (165))} \\
&= \left[\begin{array}{cccc} \cdots & \left(P_{\omega_j} \widehat{S}_{k-1} \right) \left(P_{\omega_j} \widehat{S}_{k-1} \right) \cdot P_{\widehat{\mathcal{S}}_{k-1}^\perp} S q_j & \cdots \end{array} \right] \\
&= \left[\begin{array}{cccc} \cdots & P_{\widehat{\mathcal{S}}_{k-1,j}} \cdot P_{\widehat{\mathcal{S}}_{k-1}^\perp} S q_j & \cdots \end{array} \right]. \qquad \left(\widehat{\mathcal{S}}_{k-1,j} := \operatorname{span}(P_{\omega_j} \widehat{S}_{k-1}) \right) \qquad (189)
\end{aligned}
$$

Let also $\widehat{\mathcal{S}}_{k-1,j}^C := \operatorname{span}(P_{\omega_j^C} \widehat{S}_{k-1})$. Note that $\widehat{S}_{k-1} = P_{\omega_j} \widehat{S}_{k-1} + P_{\omega_j^C} \widehat{S}_{k-1}$ and that $(P_{\omega_j} \widehat{S}_{k-1})^* (P_{\omega_j^C} \widehat{S}_{k-1}) = 0$. Consequently, $\widehat{\mathcal{S}}_{k-1,j} \perp \widehat{\mathcal{S}}_{k-1,j}^C$ and then $P_{\widehat{\mathcal{S}}_{k-1}} = P_{\widehat{\mathcal{S}}_{k-1,j}} +$

$P_{\widehat{\mathcal{S}}^C_{k-1,j}}$. Put differently, $\widehat{\mathcal{S}}_{k-1} = \widehat{\mathcal{S}}_{k-1,j} \oplus \widehat{\mathcal{S}}^C_{k-1,j}$. Using this decomposition, we simplify (189) as

$$\begin{aligned}
&[P_{\Omega_k} \circ O \circ P_{\Omega_k}] \left(P_{\widehat{\mathcal{S}}^\perp_{k-1}} S_k\right) \\
&= \begin{bmatrix} \cdots & P_{\widehat{\mathcal{S}}_{k-1,j}} \cdot P_{\widehat{\mathcal{S}}^\perp_{k-1}} Sq_j & \cdots \end{bmatrix} \qquad \text{(see (189))} \\
&= \begin{bmatrix} \cdots & (P_{\widehat{\mathcal{S}}_{k-1}} - P_{\widehat{\mathcal{S}}^C_{k-1,j}}) \cdot P_{\widehat{\mathcal{S}}^\perp_{k-1}} Sq_j & \cdots \end{bmatrix} \qquad \left(\widehat{\mathcal{S}}_{k-1} = \widehat{\mathcal{S}}_{k-1,j} \oplus \widehat{\mathcal{S}}^C_{k-1,j}\right) \\
&= -\begin{bmatrix} \cdots & P_{\widehat{\mathcal{S}}^C_{k-1,j}} \cdot P_{\widehat{\mathcal{S}}^\perp_{k-1}} Sq_j & \cdots \end{bmatrix}. \qquad (190)
\end{aligned}$$

It immediately follows that

$$\begin{aligned}
&P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}^\perp_{k-1}} S_k) + [P_{\Omega_k} \circ O \circ P_{\Omega_k}] \left(P_{\widehat{\mathcal{S}}^\perp_{k-1}} S_k\right) \\
&= \begin{bmatrix} \cdots & P_{\omega_j^C} \cdot P_{\widehat{\mathcal{S}}^\perp_{k-1}} Sq_j - P_{\widehat{\mathcal{S}}^C_{k-1,j}} \cdot P_{\widehat{\mathcal{S}}^\perp_{k-1}} Sq_j & \cdots \end{bmatrix} \\
&= \begin{bmatrix} \cdots & P_{\omega_j^C}(I_n - P_{\widehat{\mathcal{S}}^C_{k-1,j}}) P_{\omega_j^C} \cdot P_{\widehat{\mathcal{S}}^\perp_{k-1}} Sq_j & \cdots \end{bmatrix}, \qquad \left(\widehat{\mathcal{S}}^C_{k-1,j} = \operatorname{span}(P_{\omega_j^C} \widehat{S}_{k-1})\right) \quad (191)
\end{aligned}$$

and consequently

$$\begin{aligned}
&\left\| P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}^\perp_{k-1}} S_k) + [P_{\Omega_k} \circ O \circ P_{\Omega_k}] \left(P_{\widehat{\mathcal{S}}^\perp_{k-1}} S_k\right) \right\|_F^2 \\
&= \sum_{j=1}^{b} \left\| P_{\omega_j^C}(I_n - P_{\widehat{\mathcal{S}}^C_{k-1,j}}) P_{\omega_j^C} \cdot P_{\widehat{\mathcal{S}}^\perp_{k-1}} Sq_j \right\|_2^2 \qquad \text{(see (191))} \\
&\le \sum_{j=1}^{b} \left\| P_{\omega_j^C} \cdot P_{\widehat{\mathcal{S}}^\perp_{k-1}} Sq_j \right\|_2^2 \\
&= \left\| P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}^\perp_{k-1}} SQ_k) \right\|_F^2 \\
&= \left\| P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}^\perp_{k-1}} S_k) \right\|_F^2. \qquad (S_k = SQ_k) \qquad (192)
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathbb{E}\left[\left\| P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}^\perp_{k-1}} S_k) + [P_{\Omega_k} \circ O \circ P_{\Omega_k}] \left(P_{\widehat{\mathcal{S}}^\perp_{k-1}} S_k\right) \right\|_F \mid \widehat{\mathcal{S}}_{k-1}, Q_k \right] \\
&\le \sqrt{\mathbb{E}\left[\left\| P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}^\perp_{k-1}} S_k) + [P_{\Omega_k} \circ O \circ P_{\Omega_k}] \left(P_{\mathcal{S}^\perp_{k-1}} S_k\right) \right\|_F^2 \mid \widehat{\mathcal{S}}_{k-1}, Q_k \right]} \qquad \text{(Jensen's inequality)} \\
&\le \sqrt{\mathbb{E}\left[\left\| P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}^\perp_{k-1}} S_k) \right\|_F^2 \mid \widehat{\mathcal{S}}_{k-1}, Q_k \right]} \qquad \text{(see (192))} \\
&= \sqrt{\mathbb{E}\left[\sum_{i,j} (1 - \epsilon_{i,j}) \left| (P_{\widehat{\mathcal{S}}^\perp_{k-1}} S_k)[i,j] \right|^2 \mid \widehat{\mathcal{S}}_{k-1}, Q_k \right]} \\
&= \sqrt{1-p} \| P_{\widehat{\mathcal{S}}^\perp_{k-1}} S_k \|_F
\end{aligned}$$

$$\leq \sqrt{1-p}\|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S\|_F \|Q_k\| \qquad (S_k = SQ_k^*, \qquad \|AB\|_F \leq \|A\|_F \cdot \|B\|) \tag{193}$$

where $\{\epsilon_{i,j}\}_{i,j} \overset{\text{ind.}}{\sim} \text{Bernoulli}(p)$. We thus proved the first claim in Lemma 21, namely, (170). Then note that

$$\begin{aligned}
&\left\| P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k) + [P_{\Omega_k} \circ O \circ P_{\Omega_k}] \left(P_{\mathcal{S}_{k-1}^{\perp}} S_k\right) \right\|_F^2 \\
&\leq \left\| P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k) \right\|_F^2 \qquad \text{(see (192))} \\
&\leq \|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k\|_F^2 \\
&\leq \|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S\|_F^2 \|Q_k\|^2, \qquad (S_k = SQ_k)
\end{aligned} \tag{194}$$

which proves the second claim, namely, (171). In fact, with some probability, a stronger bound can be derived by controlling the deviation from the expectation in (193) using the Hoeffding inequality (Lemma 17). With $\alpha = \frac{p}{2}\|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k\|_F^2$ in Lemma 17 and recalling that $\widehat{\mathcal{S}}_{k-1}, Q_k$ are fixed for now, we find that

$$\begin{aligned}
\left\| P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k) \right\|_F^2 &\leq \mathbb{E}\left[\left\| P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k) \right\|_F^2 \mid \widehat{\mathcal{S}}_{k-1}, Q_k \right] + \alpha \\
&= (1 - p/2)\|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k\|_F^2 \qquad \text{(see (193))} \\
&= (1 - p/2)\|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} SQ_k\|_F^2 \qquad (S_k = SQ_k) \\
&\leq (1 - p/2)\|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S\|_F^2 \, \|Q_k\|^2 ,
\end{aligned} \tag{195}$$

except with a probability of at most

$$\begin{aligned}
\exp\left(-\frac{C_1 p^2 \|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k\|_F^4}{\sum_{i,j} |(P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k)[i,j]|^4} \right) &\leq \exp\left(-\frac{C_1 p^2 \|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k\|_F^4}{\|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k\|_\infty^2 \cdot \|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k\|_F^2} \right) \\
&= \exp\left(-\frac{C_1 p^2 \|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k\|_F^2}{\|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k\|_\infty^2} \right) \\
&=: \exp\left(\frac{-C_1 p^2 nb}{\widetilde{\eta}(P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k)} \right),
\end{aligned} \tag{196}$$

where $\|A\|_\infty$ returns the largest entry of A in magnitude. Substituting (195) back into (194) yields that

$$\begin{aligned}
&\left\| P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k) + [P_{\Omega_k} \circ O \circ P_{\Omega_k}] \left(P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k\right) \right\|_F \\
&\leq \left\| P_{\Omega_k^C}(P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S_k) \right\|_F \qquad \text{(see (194))} \\
&\leq \sqrt{1 - p/2} \cdot \|P_{\widehat{\mathcal{S}}_{k-1}^{\perp}} S\|_F \|Q_k\|,
\end{aligned} \tag{197}$$

which proves the last claim in Lemma 21, namely, (172).

Appendix I. Proof of Lemma 22

We fix $\widehat{\mathcal{S}}_{k-1}, Q_k$ throughout. We begin by bounding the target quantity as

$$\begin{aligned}
&\left\|P_{\mathcal{S}^\perp}\cdot[O\circ P_{\Omega_k}]\left(P_{\widehat{\mathcal{S}}_{k-1}^\perp}S_k\right)\right\|_F \\
&= \left\|P_{\mathcal{S}^\perp}P_{\widehat{\mathcal{S}}_{k-1}}\cdot[O\circ P_{\Omega_k}]\left(P_{\widehat{\mathcal{S}}_{k-1}^\perp}S_k\right)\right\|_F \qquad \text{(see (165))} \\
&\le \|P_{\mathcal{S}^\perp}P_{\widehat{\mathcal{S}}_{k-1}}\|\cdot\left\|[O\circ P_{\Omega_k}]\left(P_{\widehat{\mathcal{S}}_{k-1}^\perp}S_k\right)\right\|_F \qquad (\|AB\|_F\le\|A\|\cdot\|B\|_F)\,. \qquad (198)
\end{aligned}$$

We bound the random norm in the last line above in Appendix J.

Lemma 23 *For $\alpha\ge 1$ and except with a probability of at most $b^{-C\alpha}$, it holds that*

$$\left\|[O\circ P_{\Omega_k}]\left(P_{\widehat{\mathcal{S}}_{k-1}^\perp}S_k\right)\right\|_F \lesssim \alpha\log b\sqrt{\frac{\log n}{p}}\|P_{\mathcal{S}^\perp}P_{\widehat{\mathcal{S}}_{k-1}}\|_F\cdot\|Q_k\|, \qquad (199)$$

provided that $p\gtrsim\alpha^2\log^2 b\log n\cdot\eta(\widehat{\mathcal{S}}_{k-1})r/n$.

In light of the above lemma, we conclude that

$$\begin{aligned}
&\left\|P_{\mathcal{S}^\perp}\cdot[O\circ P_{\Omega_k}]\left(P_{\widehat{\mathcal{S}}_{k-1}^\perp}S_k\right)\right\|_F \\
&\le \|P_{\mathcal{S}^\perp}P_{\widehat{\mathcal{S}}_{k-1}}\|\cdot\left\|[O\circ P_{\Omega_k}]\left(P_{\widehat{\mathcal{S}}_{k-1}^\perp}S_k\right)\right\|_F \qquad \text{(see (198))} \\
&\lesssim \alpha\log b\sqrt{\frac{\log n}{p}}\|P_{\mathcal{S}^\perp}P_{\widehat{\mathcal{S}}_{k-1}}\|\cdot\|P_{\mathcal{S}^\perp}P_{\widehat{\mathcal{S}}_{k-1}}\|_F\cdot\|Q_k\|, \qquad \text{(see (199))} \qquad (200)
\end{aligned}$$

except with a probability of at most $b^{-C\alpha}$ and provided that $p\gtrsim\alpha^2\log^2 b\log n\cdot\eta(\widehat{\mathcal{S}}_{k-1})r/n$. This completes the proof of Lemma 22.

Appendix J. Proof of Lemma 23

Using the definition of operator O in (165), we write that

$$\begin{aligned}
&\left\|[O\circ P_{\Omega_k}]\left(P_{\widehat{\mathcal{S}}_{k-1}^\perp}S_k\right)\right\|_F^2 \\
&= \sum_{j=1}^{b}\left\|\widehat{S}_{k-1}(P_{\omega_j}\widehat{S}_{k-1})^\dagger P_{\omega_j}\cdot P_{\widehat{\mathcal{S}}_{k-1}^\perp}Sq_j\right\|_2^2 \qquad (\text{(165) and } S_k=SQ_k) \\
&= \sum_{j=1}^{b}\left\|\widehat{S}_{k-1}(P_{\omega_j}\widehat{S}_{k-1})^\dagger P_{\widehat{\mathcal{S}}_{k-1}^\perp}Sq_j\right\|_2^2 \\
&= \sum_{j=1}^{b}\left\|(P_{\omega_j}\widehat{S}_{k-1})^\dagger P_{\widehat{\mathcal{S}}_{k-1}^\perp}Sq_j\right\|_2^2. \qquad \left(\widehat{S}_{k-1}^*\widehat{S}_{k-1}=I_r\right) \qquad (201)
\end{aligned}$$

For fixed $j\in[1:b]$, consider the summand in the last line above:

$$\left\|(P_{\omega_j}\widehat{S}_{k-1})^\dagger P_{\widehat{\mathcal{S}}_{k-1}^\perp}S\cdot q_j\right\|_2 \le \left\|(P_{\omega_j}\widehat{S}_{k-1})^\dagger P_{\widehat{\mathcal{S}}_{k-1}^\perp}\right\|\cdot\|P_{\widehat{\mathcal{S}}_{k-1}^\perp}Sq_j\|_2$$

$$= \left\| (P_{\omega_j} \widehat{S}_{k-1})^\dagger \widehat{S}_{k-1}^\perp \right\| \cdot \|P_{\widehat{\mathcal{S}}_{k-1}^\perp} S q_j\|_2 \qquad \left(P_{\widehat{\mathcal{S}}_{k-1}^\perp} = \widehat{S}_{k-1}^\perp (\widehat{S}_{k-1}^\perp)^* \right)$$
$$=: \|\widehat{Z}_j\| \cdot \|P_{\widehat{\mathcal{S}}_{k-1}^\perp} S q_j\|_2. \tag{202}$$

Above, $\widehat{S}_{k-1}^\perp$ is as usual an orthonormal basis for the subspace $\widehat{\mathcal{S}}_{k-1}^\perp$. We can now revisit (201) and write that

$$\begin{aligned}
\left\| O(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k) \right\|_F^2 &= \sum_{j=1}^b \left\| (P_{\omega_j} \widehat{S}_{k-1})^\dagger P_{\widehat{\mathcal{S}}_{k-1}^\perp} S q_j) \right\|_2^2 \qquad \text{(see (201))} \\
&\leq \max_j \|\widehat{Z}_j\|^2 \cdot \sum_{j=1}^b \|P_{\widehat{\mathcal{S}}_{k-1}^\perp} S q_j\|_2^2 \qquad \text{(see (202))} \\
&= \max_j \|\widehat{Z}_j\|^2 \cdot \|P_{\widehat{\mathcal{S}}_{k-1}^\perp} S Q_k\|_F^2 \\
&\leq \max_j \|\widehat{Z}_j\|^2 \cdot \|P_{\widehat{\mathcal{S}}_{k-1}^\perp} S\|_F^2 \|Q_k\|^2. \qquad (\|AB\|_F \leq \|A\|_F \cdot \|B\|) \qquad (203)
\end{aligned}$$

It remains to control the maximum in the last line above. We first focus on controlling $\|\widehat{Z}_j\|$ for fixed $j \in [1:b]$. Observe that $\widehat{Z}_j$ is a solution of the least-squares problem

$$\min_{Z \in \mathbb{R}^{n \times (n-r)}} \left\| \widehat{S}_{k-1}^\perp - (P_{\omega_j} \widehat{S}_{k-1}) Z \right\|_F^2,$$

and therefore satisfies the *normal equation*

$$(P_{\omega_j} \widehat{S}_{k-1})^* \left((P_{\omega_j} \widehat{S}_{k-1}) \widehat{Z}_j - \widehat{S}_{k-1}^\perp \right) = 0,$$

which is itself equivalent to

$$(\widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1}) \widehat{Z}_j = \widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1}^\perp. \qquad \left(P_{\omega_j}^2 = P_{\omega_j} \right) \tag{204}$$

In fact, since

$$\mathbb{E}\left[\widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1}^\perp \right] = p \cdot \widehat{S}_{k-1}^* \widehat{S}_{k-1}^\perp = 0,$$
$$\mathbb{E}\left[\widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1} \right] = p \cdot I_r, \qquad \left(\widehat{S}_{k-1}^* \widehat{S}_{k-1} = I_r \right)$$

we can rewrite (204) as

$$\left(\widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1} - \mathbb{E}\left[\widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1} \right] \right) \widehat{Z}_j + p \cdot \widehat{Z}_j = \widehat{S}_{k-1}^T P_{\omega_j} \widehat{S}_{k-1}^\perp - \mathbb{E}\left[\widehat{S}_{k-1}^T P_{\omega_j} \widehat{S}_{k-1}^\perp \right].$$

An application of the triangle inequality above immediately implies that

$$p\|\widehat{Z}_j\| \leq \left\| \widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1} - \mathbb{E}\left[\widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1} \right] \right\| \cdot \|\widehat{Z}_j\| + \left\| \widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1}^\perp - \mathbb{E}\left[\widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1}^\perp \right] \right\|. \tag{205}$$

To control $\|\widehat{Z}_j\|$, we therefore need to derive large devation bounds for the two remaining norms on the right-hand side above. For the first spectral norm, we write that

$$\left\|\widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1} - \mathbb{E}\left[\widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1}\right]\right\| = \left\|\sum_i (\epsilon_i - p) \cdot \widehat{S}_{k-1}^* E_{i,i} \widehat{S}_{k-1}\right\| =: \left\|\sum_i A_i\right\|, \quad (206)$$

where $\{\epsilon_i\}_i \overset{\text{ind.}}{\sim} \text{Bernoulli}(p)$ and $E_{i,i} \in \mathbb{R}^{n\times n}$ is the $[i,i]$th canonical matrix. Furthermore, $\{A_i\}_i \subset \mathbb{R}^{r\times r}$ above are independent and zero-mean random matrices. To apply the Bernstein inequality (Lemma 18), we first compute the parameter β as

$$\begin{aligned}
\|A_i\| &= \left\|(\epsilon_i - p) \cdot S_{k-1}^* E_{i,i} S_{k-1}\right\| \qquad \text{(see (206))} \\
&\le \|\widehat{S}_{k-1}^* E_{i,i} \widehat{S}_{k-1}\| \qquad (\epsilon_i \in \{0,1\}) \\
&= \|\widehat{S}_{k-1}[i,:]\|_2^2 \\
&\le \frac{\eta(\widehat{\mathcal{S}}_{k-1}) r}{n} =: \beta. \qquad \text{(see (14))}
\end{aligned} \quad (207)$$

To compute the weak variance σ, we write that

$$\begin{aligned}
\left\|\mathbb{E}\left[\sum_i A_i^2\right]\right\| &= \left\|\sum_i \mathbb{E}\left[(\epsilon_i - p)^2\right] (\widehat{S}_{k-1}^* E_{i,i} \widehat{S}_{k-1})^2\right\| \qquad \text{(see (206))} \\
&= \left\|\sum_i p(1-p)(\widehat{S}_{k-1}^* E_{i,i} \widehat{S}_{k-1})^2\right\| \qquad (\epsilon_i \sim \text{Bernoulli}(p)) \\
&\le p \left\|\sum_i (\widehat{S}_{k-1}^* E_{i,i} \widehat{S}_{k-1})^2\right\| \\
&= p \left\|\sum_i \widehat{S}_{k-1}^* E_{i,i} \widehat{S}_{k-1} \widehat{S}_{k-1}^* E_{i,i} \widehat{S}_{k-1}\right\| \\
&= p \left\|\sum_i \|\widehat{S}_{k-1}[i,:]\|_2^2 \cdot \widehat{S}_{k-1}^* E_{i,i} \widehat{S}_{k-1}\right\| \\
&\le p \cdot \max_i \|\widehat{S}_{k-1}[i,:]\|_2^2 \cdot \left\|\sum_i \widehat{S}_{k-1}^* E_{i,i} \widehat{S}_{k-1}\right\| \\
&= p \left\|\sum_i \widehat{S}_{k-1}^* E_{i,i} \widehat{S}_{k-1}\right\| \\
&= p \cdot \frac{\eta(\widehat{\mathcal{S}}_{k-1}) r}{n} \cdot \left\|\sum_i \widehat{S}_{k-1}^* E_{i,i} \widehat{S}_{k-1}\right\| \qquad \text{(see (14))} \\
&= p \cdot \frac{\eta(\widehat{\mathcal{S}}_{k-1}) r}{n} \qquad \left(\sum_i E_{i,i} = I_n,\ \widehat{S}_{k-1}^* \widehat{S}_{k-1} = I_r\right) \\
&=: \sigma^2.
\end{aligned} \quad (208)$$

It also follows that

$$\begin{aligned}
&\max\left(\log r\cdot\beta, \sqrt{\log r}\cdot\sigma\right)\\
&= \max\left(\frac{\log r\cdot\eta(\widehat{\mathcal{S}}_{k-1})r}{n}, \sqrt{\frac{\log r\cdot p\cdot\eta(\widehat{\mathcal{S}}_{k-1})r}{n}}\right) \qquad \text{(see (207) and (208))}\\
&\leq \sqrt{\frac{\log r\cdot p\cdot\eta(\widehat{\mathcal{S}}_{k-1})r}{n}}. \qquad \left(\text{if } p\geq\frac{\log r\cdot\eta(\widehat{\mathcal{S}}_{k-1})r}{n}\right) \qquad (209)
\end{aligned}$$

As a result, for $\alpha\geq 1$ and except with a probability of at most $e^{-C\alpha}$, it holds that

$$\begin{aligned}
\left\|\widehat{S}_{k-1}^* P_{\omega_j}\widehat{S}_{k-1} - \mathbb{E}\left[\widehat{S}_{k-1}^* P_{\omega_j}\widehat{S}_{k-1}\right]\right\| &\lesssim \alpha\max\left(\log r\cdot\beta, \sqrt{\log r}\cdot\sigma\right) \qquad \text{(see Lemma 18)}\\
&\leq \alpha\sqrt{\frac{\log r\cdot p\cdot\eta(\widehat{\mathcal{S}}_{k-1})r}{n}}. \qquad (210)
\end{aligned}$$

On the other hand, in order to apply the Bernstein inequality to the second spectral norm in (205), we write that

$$\left\|\widehat{S}_{k-1}^* P_{\omega_j}\widehat{S}_{k-1}^\perp - \mathbb{E}\left[\widehat{S}_{k-1}^* P_{\omega_j}\widehat{S}_{k-1}^\perp\right]\right\| = \left\|\sum_i (\epsilon_i - p)\,\widehat{S}_{k-1}^* E_{i,i}\widehat{S}_{k-1}^\perp\right\| =: \left\|\sum_i A_i\right\|, \qquad (211)$$

where $\{\epsilon_i\}_i \overset{\text{ind.}}{\sim} \text{Bernoulli}(p)$, $E_{i,i}\in\mathbb{R}^{n\times n}$ is the ith canonical matrix, and $\{A_i\}_i\subset\mathbb{R}^{r\times(n-r)}$ are zero-mean and independent random matrices. To compute the parameter β here, we write that

$$\begin{aligned}
\|A_i\| &= \left\|(\epsilon_i - p)\,\widehat{S}_{k-1}^* E_{i,i}\widehat{S}_{k-1}^\perp\right\| \qquad \text{(see (211))}\\
&\leq \|\widehat{S}_{k-1}^* E_{i,i}\widehat{S}_{k-1}^\perp\| \qquad (\epsilon_i\in\{0,1\})\\
&\leq \|\widehat{S}_{k-1}^* E_{i,i}\| \qquad \left((\widehat{S}_{k-1}^\perp)^*\widehat{S}_{k-1}^\perp = I_{n-r}\right)\\
&= \|\widehat{S}_{k-1}[i,:]\|_2\\
&\leq \sqrt{\frac{\eta(\widehat{\mathcal{S}}_{k-1})r}{n}} =: \beta. \qquad \text{(see (14))} \qquad (212)
\end{aligned}$$

To compute the weak variance σ, we notice that

$$\begin{aligned}
\left\|\mathbb{E}\left[\sum_i A_i A_i^*\right]\right\| &= \left\|\sum_i \mathbb{E}\left[(\epsilon_i - p)^2\right]\widehat{S}_{k-1}^* E_{i,i}\widehat{S}_{k-1}^\perp(\widehat{S}_{k-1}^\perp)^* E_{i,i}\widehat{S}_{k-1}\right\| \qquad \text{(see (211))}\\
&= \left\|\sum_i p(1-p)\cdot\widehat{S}_{k-1}^* E_{i,i}\widehat{S}_{k-1}^\perp(\widehat{S}_{k-1}^\perp)^* E_{i,i}\widehat{S}_{k-1}\right\| \qquad (\epsilon_i\sim\text{Benoulli}(p))\\
&\leq p\left\|\sum_i \widehat{S}_{k-1}^* E_{i,i}\widehat{S}_{k-1}^\perp(\widehat{S}_{k-1}^\perp)^* E_{i,i}\widehat{S}_{k-1}\right\|
\end{aligned}$$

$$\begin{aligned}
&\leq p \left\| \sum_i \widehat{S}_{k-1}^* E_{i,i} E_{i,i} \widehat{S}_{k-1} \right\| \qquad \left(\widehat{S}_{k-1}^\perp (\widehat{S}_{k-1}^\perp)^* \preceq I_n \right) \\
&= p \left\| \sum_i \widehat{S}_{k-1}^* E_{i,i} \widehat{S}_{k-1} \right\| \\
&= p. \qquad \left(\sum_i E_{i,i} = I_n,\ \widehat{S}_{k-1}^* \widehat{S}_{k-1} = I_r \right)
\end{aligned} \tag{213}$$

In a similar fashion, we find that

$$\begin{aligned}
\left\| \mathbb{E}\left[\sum_i A_i^* A_i \right] \right\| &\leq p \left\| \sum_i (\widehat{S}_{k-1}^\perp)^* E_{i,i} \widehat{S}_{k-1} \widehat{S}_{k-1}^* E_{i,i} \widehat{S}_{k-1}^\perp \right\| \\
&= p \left\| \sum_i \|\widehat{S}_{k-1}[i,:]\|_2^2 \cdot (\widehat{S}_{k-1}^\perp)^* E_{i,i} \widehat{S}_{k-1}^\perp \right\| \\
&\leq p \cdot \max_i \|\widehat{S}_{k-1}[i,:]\|_2^2 \cdot \left\| \sum_i (\widehat{S}_{k-1}^\perp)^* E_{i,i} \widehat{S}_{k-1}^\perp \right\| \\
&= p \cdot \max_i \|\widehat{S}_{k-1}[i,:]\|_2^2 \qquad \left(\sum_i E_{i,i} = I_n,\ (\widehat{S}_{k-1}^\perp)^* \widehat{S}_{k-1}^\perp = I_{n-r} \right) \\
&= p \cdot \frac{\eta(\widehat{\mathcal{S}}_{k-1}) r}{n}, \qquad \text{(see (14))}
\end{aligned} \tag{214}$$

and finally

$$\begin{aligned}
\sigma &= \max \left(\left\| \mathbb{E} \sum_i A_i A_i^* \right\|, \left\| \mathbb{E} \sum_i A_i^* A_i \right\| \right) \\
&= \max \left(\sqrt{p}, \sqrt{p} \cdot \sqrt{\frac{\eta(\widehat{\mathcal{S}}_{k-1}) r}{n}} \right) \qquad \text{(see (213) and (214))} \\
&= \sqrt{p}. \qquad \left(\eta(\widehat{\mathcal{S}}_{k-1}) \leq \frac{n}{r} \right)
\end{aligned} \tag{215}$$

We now compute

$$\begin{aligned}
\max \left(\log n \cdot \beta, \sqrt{\log n} \cdot \sigma \right) &= \max \left(\log n \sqrt{\frac{\eta(\widehat{\mathcal{S}}_{k-1}}{n}}, \sqrt{\log n \cdot p} \right) \qquad \text{(see (212) and (215))} \\
&= \sqrt{\log n \cdot p}. \qquad \left(\text{if } p \geq \frac{\log n \cdot \eta(\widehat{\mathcal{S}}_{k-1}) r}{n} \right)
\end{aligned} \tag{216}$$

Therefore, for $\alpha \geq 1$ and except with a probability of at most $e^{-C\alpha}$, it holds that

$$\left\| \widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1}^\perp - \mathbb{E}\left[\widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1}^\perp \right] \right\| \lesssim \alpha \max \left(\log n \cdot \beta, \sqrt{\log n} \cdot \sigma \right) \qquad \text{(see Lemma 18)}$$

$$= \alpha\sqrt{\log n \cdot p}. \tag{217}$$

Overall, by substituting the large deviation bounds (210) and (217) into (205), we find that

$$\begin{aligned} p\|\widehat{Z}_j\| &\le \left\|\widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1} - \mathbb{E}\left[\widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1}\right]\right\| \cdot \|\widehat{Z}_j\| \\ &\quad + \left\|\widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1}^\perp - \mathbb{E}\left[\widehat{S}_{k-1}^* P_{\omega_j} \widehat{S}_{k-1}^\perp\right]\right\| \qquad \text{(see (205))} \\ &\lesssim \alpha\sqrt{\frac{\log r \cdot p \cdot \eta(\widehat{\mathcal{S}}_{k-1})r}{n}} \cdot \|\widehat{Z}_j\| + \alpha\sqrt{\log n \cdot p}, \qquad \text{(see (210) and (217))} \end{aligned}$$

except with a probability of at most $e^{-C\alpha}$ and under (209) and (216). It immediately follows that

$$\begin{aligned} \|\widehat{Z}_j\| &\lesssim \frac{\alpha\sqrt{\frac{\log n}{p}}}{1 - \sqrt{\frac{\alpha^2 \log r \cdot \eta(\widehat{\mathcal{S}}_{k-1})r}{pn}}} \qquad \text{(see the next line)} \\ &\lesssim \alpha\sqrt{\frac{\log n}{p}}, \qquad \left(\text{if } \frac{\alpha^2 \log r \cdot \eta(\widehat{\mathcal{S}}_{k-1})r}{pn} \lesssim 1\right) \end{aligned} \tag{218}$$

except with a probability of at most $e^{-C\alpha}$. In light of (209) and (216), we assume that $p \gtrsim \alpha^2 \log n \cdot \eta(\widehat{\mathcal{S}}_{k-1})r/n$. Then using the union bound and with the choice of $\alpha = \alpha' \log b$, it follows that

$$\max_{j\in[1:b]} \|\widehat{Z}_j\| \lesssim \alpha' \log b \sqrt{\frac{\log n}{p}},$$

provided that $p \gtrsim \alpha'^2 \log^2 b \cdot \log n \cdot \eta(\widehat{\mathcal{S}}_{k-1})r/n$ and except with a probability of at most $be^{-C\alpha' \log b} = b^{-C\alpha'}$. Invoking (203), we finally conclude that

$$\begin{aligned} \left\|O(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k)\right\|_F &\le \max_j \|\widehat{Z}_j\| \cdot \|P_{\widehat{\mathcal{S}}_{k-1}^\perp} P_{\mathcal{S}}\|_F \|Q_k\| \qquad \text{(see (203))} \\ &\lesssim \alpha' \log b \sqrt{\frac{\log n}{p}} \cdot \|P_{\widehat{\mathcal{S}}_{k-1}^\perp} P_{\mathcal{S}}\|_F \|Q_k\|. \end{aligned}$$

A bound in expectation also easily follows: Let δ denote the factor of δ' in last line above. Then we have that

$$\begin{aligned} \mathbb{E}\left\|P_{\mathcal{S}^\perp} \cdot [O \circ P_{\Omega_k}]\,(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k)\right\|_F &= \delta \int_0^\infty \Pr\left[\left\|P_{\mathcal{S}^\perp} \cdot [O \circ P_{\Omega}]\,(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k)\right\|_F > \alpha'\delta\right] d\alpha' \\ &\le \delta\left(1 + \int_1^\infty \Pr\left[\left\|P_{\mathcal{S}^\perp} \cdot [O \circ P_{\Omega}]\,(P_{\widehat{\mathcal{S}}_{k-1}^\perp} S_k)\right\|_F > \alpha'\delta\right] d\alpha'\right) \\ &\le \delta\left(1 + \int_1^\infty b^{-\alpha'}\, d\alpha'\right) \\ &\le 2\delta \\ &= 2\log b \sqrt{\frac{\log n}{p}} \cdot \|P_{\widehat{\mathcal{S}}_{k-1}^\perp} P_{\mathcal{S}}\|_F \|Q_k\|. \end{aligned} \tag{219}$$

This completes the proof of Lemma 23.

Appendix K. Properties of a Standard Random Gaussian Matrix

As a supplement to Remark 25, we show here that a standard random Gaussian matrix $G \in \mathbb{R}^{b\times r}$ is well-conditioned and incoherent when b is sufficiently large. From (Vershynin, 2012a, Corollary 5.35) and for fixed $\alpha \geq 1$, recall that

$$\sqrt{b} - C_3\alpha\sqrt{r} \leq \sigma_r(G) \leq \sigma_1(G) \leq \sqrt{b} + C_3\alpha\sqrt{r}, \tag{220}$$

except with a probability of at most $e^{-\alpha^2 r}$. It follows that

$$\nu(G) = \frac{\sigma_1(G)}{\sigma_r(G)} \leq \frac{\sqrt{b} + C_3\alpha\sqrt{r}}{\sqrt{b} - C_3\alpha\sqrt{r}}, \tag{221}$$

which can be made close to one by choosing $b \gtrsim \alpha^2 r$.

For the coherence, note that $G(G^*G)^{-\frac{1}{2}} \in \mathbb{R}^{b\times r}$ is an orthonormal basis for $\mathrm{span}(G)$. Using the definition of coherence in (14), we then write that

$$\begin{aligned}
\eta\left(\mathrm{span}(G)\right) &= \frac{b}{r}\max_{i\in[1:b]}\left\|G[i,:]\left(G^*G\right)^{-\frac{1}{2}}\right\|_2^2 \qquad \text{(see (14))} \\
&\leq \frac{b}{r}\max_i \|G[i,:]\|_2^2 \cdot \|\left(G^*G\right)^{-\frac{1}{2}}\|^2 \\
&= \frac{b}{r}\max_i \|G[i,:]\|_2^2 \cdot \left(\sigma_r(G)\right)^{-2} \\
&\leq \frac{b}{r}\max_i \|G[i,:]\|_2^2 \cdot \left(\sqrt{b} - C_3\alpha\sqrt{r}\right)^{-2} \qquad \text{(see (220))} \\
&\leq \frac{b}{r}\max_i \|G[i,:]\|_2^2 \cdot \left(\frac{b}{2} - C_3^2\alpha^2 r\right)^{-1} \qquad \left((a-b)^2 \geq \frac{a^2}{2} - b^2\right) \\
&\lesssim \frac{b}{r}\max_i \|G[i,:]\|_2^2 \cdot \left(b - C_4\alpha^2 r\right)^{-1},
\end{aligned} \tag{222}$$

except with a probability of at most $e^{-\alpha^2 r}$. For fixed i, $\|G[i,:]\|_2^2$ is a chi-squared random variable with r degrees of freedom so that

$$\Pr\left[\|G[i,:]\|_2^2 \gtrsim \beta \cdot r\right] \leq e^{-\beta}, \tag{223}$$

for $\beta \geq 1$. An application of the union bound and the choice of $\beta = C\alpha\log b$ then leads us to

$$\Pr\left[\max_{i\in[1:b]} \|G[i,:]\|_2^2 \gtrsim \alpha\log b \cdot r\right] \leq b \cdot b^{-C\alpha} = b^{-C\alpha}. \tag{224}$$

Substituting the bound above back into (222) yields that

$$\begin{aligned}
\eta\left(\mathrm{span}(G)\right) &\lesssim \frac{b}{r}\max_i \|G[i,:]\|_2^2 \cdot \left(b - C_4 r\right)^{-1} \qquad \text{(see (222))} \\
&\lesssim \frac{b}{r} \cdot r\alpha\log b \cdot \left(b - C_4 r\right)^{-1} \qquad \text{(see (224))} \\
&\lesssim \frac{\alpha b\log b}{b - C_4\alpha^2 r},
\end{aligned} \tag{225}$$

except with a probability of at most $e^{-\alpha r} + b^{-C\alpha}$. In particular, when $b \geq 2C_4\alpha^2 r$, we find that $\eta(\mathrm{span}(G)) \lesssim \alpha\log b$ except with a probability of at most $e^{-\alpha r} + b^{-C\alpha}$.

Appendix L. Alternative Initialization

SNIPE in Algorithm 1 is initialized by truncating the SVD of the first incomplete block $Y_1 \in \mathbb{R}^{n\times b}$, where we often take $b = O(r)$ to keep the computational and storage requirements of SNIPE minimal, see Remarks 1 and 2. Put differently, with the notation of Section 3, even though our end goal is to compute rank-r truncated SVD of the full (but hidden) data block $S_1 \in \mathbb{R}^{n\times b}$, SNIPE is initialized with truncated SVD of the incomplete (but available) block $Y_1 = P_{\Omega_1}(S_1) \in \mathbb{R}^{n\times b}$, which fills in the missing entries with zeros. Indeed, when the first incomplete block Y_1 arrives, there is no prior knowledge to leverage and zero-filling the missing entries in Y_1 is a sensible strategy. In contrast, for the rest of blocks $k \geq 2$, SNIPE uses its previous estimate $\widehat{\mathcal{S}}_{k-1}$ to fill out the erased entries in Y_k before updating its estimate to $\widehat{\mathcal{S}}_k$, see Algorithm 1.

One might instead initialize SNIPE with a larger block. More specifically, suppose that we change the first block size to $b_1 \geq b$ while keeping the rest of the blocks at the same size b. Then we set $\widehat{\mathcal{S}}_1$ to be the span of leading r left singular vectors of the first incomplete block $Y_1 \in \mathbb{R}^{n\times b_1}$, while the rest of steps in Algorithm 1 do not change. As the size of the first block b_1 increases, $\widehat{\mathcal{S}}_1$ increasingly better approximates the true subspace $\mathcal{S}$. Indeed, one might consider $Y_1 = P_{\Omega_1}(S_1) = S_1 + (P_{\Omega_1}(S_1) - S_1)$ as a "noisy" copy of S_1, where the noise is due to the erasures. Roughly speaking then, as b_1 increases, the energy of the "signal" part, namely, $\|S_1\|_F$, grows faster than and eventually dominates the energy of the random noise $\|P_{\Omega_1}(S_1) - S_1\|_F$. This intuition is made precise by the following result which loosely speaking states that

$$d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_1) \lesssim \sqrt{\frac{r}{pn}},$$

when $b_1 = \Omega(n)$. This result is proved in Appendix C with the aid of standard large deviation bounds.

Proposition 24 *Consider an r-dimensional subspace $\mathcal{S}$ with orthonormal basis $S \in \mathbb{R}^{n\times r}$. For an integer $b_1 \geq r$, let the coefficient vectors $\{q_t\}_{t=1}^{b_1} \subset \mathbb{R}^r$ be independent copies of a random vector $q \in \mathbb{R}^r$. For every $t \in [1 : b_1]$, we observe each coordinate of $s_t = Sq_t \in \mathcal{S}$ independently with a probability of p and collect the observed entries in $y_t \in \mathbb{R}^n$, supported on the random index set $\omega_t \subseteq [1 : n]$. We set $Q_1 = [q_1\, q_2 \cdots q_{b_1}] \in \mathbb{R}^{r\times b_1}$ and $Y_1 = [y_1\, y_2 \cdots y_{b_1}] \in \mathbb{R}^{n\times b_1}$ for short. Let also $\widehat{\mathcal{S}}_1$ be the span of leading r left singular vectors of Y_1.*

Then, for fixed $\alpha, \nu \geq 1$ and $1 \leq \eta \leq b_1/r$, it holds that

$$d_{\mathbb{G}}(\mathcal{S}, \widehat{\mathcal{S}}_1) \lesssim \alpha \cdot \nu \sqrt{\left(1 \vee \frac{n}{b_1}\right) \frac{(\eta \vee \eta(\mathcal{S}))\, r \log(n \vee b_1)}{pn}}, \tag{226}$$

except with a probability of at most $e^{-\alpha} + \Pr[\nu(Q_1) > \nu] + \Pr[\eta(\mathcal{Q}_1) > \eta]$. Above, $\nu(Q_1)$ is the condition number of Q_1, $\eta(\mathcal{Q}_1)$ is the coherence of $\mathcal{Q}_1 = \text{span}(Q_1^*)$ (see (14)), and $a \vee b := \max\{a, b\}$.

Remark 25 [Discussion of Proposition 24] As (226) suggests, for $\widehat{\mathcal{S}}_1$ to be close to $\mathcal{S}_1$, the first block should be a wide matrix, namely, $b_1 = O(n)$. This dependence on the block size was anticipated. Indeed, it is well-understood that one needs $O(n)$ samples in order

for the sample covariance matrix to closely approximate the covariance matrix of a random vector in $\mathbb{R}^n$ (Vershynin, 2012b). As an example, consider the case where the coefficient vectors $\{q_t\}_{t=1}^{b_1}$ are standard random Gaussian vectors and so $Q_1 \in \mathbb{R}^{n\times b_1}$ is a standard random Gaussian matrix, namely, populated with zero-mean independent Gaussian random variables with unit variance. Then both probabilities above are small when b_1 is sufficiently large. More specifically, we show in Appendix K that

$$\Pr[\nu(Q_1) > \nu] \le \exp\left(-C\frac{\nu-1}{\nu+1}\right), \qquad \text{when } b \gtrsim r, \tag{227}$$

$$\Pr[\eta(\mathcal{Q}_1) > \eta] \le \exp(-C\eta/\log b), \qquad \text{when } b\log^2 b \gtrsim \eta^2 r, \tag{228}$$

for a Gaussian coefficient matrix Q_1.

It is also worth noting that initializing SNIPE with a large first block can be done *without* increasing the storage requirement or computational complexity of SNIPE, namely, replacing the block size $b = O(r)$ in first step of Algorithm 1 with $b_1 = O(n)$ can be done without losing the streaming nature of SNIPE. More specifically, with the alternative initialization, naively computing the truncated SVD of the first block requires $O(b_1 n) = O(n^2)$ bits of storage and $O(rb_1 n) = O(rn^2)$ flops. These requirements can be significantly lowered by implementing a state-of-the-art streaming PCA algorithm such as the "power method" in (Mitliagkas et al., 2013). This suggests a two-phase algorithm. In the first phase, the power method is applied to the incoming data, where the missing entries are filled with zeros. This phase produces the estimate $\widehat{\mathcal{S}}_1$ in SNIPE, which serves as an initialization for the second phase in which the main loop of SNIPE is applied to the incoming blocks, producing the estimates $\{\widehat{\mathcal{S}}_k\}_{k\ge 2}$. If b_1 is sufficiently large, the first phase brings us within the basin of attraction of the true subspace $\mathcal{S}$ and activates the locally linear convergence of SNIPE to $\mathcal{S}$ in the second phase, see Theorems 8 and 10.

References

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(2):101–114, 1999.

R. Arora, A. Cotter, K. Livescu, and N. Srebro. Stochastic optimization for PCA and PLS. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 861–868. IEEE, 2012.

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

L. Balzano and S. J Wright. On GROUSE and incremental SVD. In *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 1–4. IEEE, 2013.

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

L. Balzano, R. Nowak, and B. Recht. Online identification and tracking of subspaces from highly incomplete information. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 704–711. IEEE, 2010.

M. Brand. Incremental singular value decomposition of uncertain data with missing values. In *European Conference on Computer Vision*, pages 707–720. Springer, 2002.

J. R. Bunch and C. P. Nielsen. Updating the singular value decomposition. *Numerische Mathematik*, 31(2):111–129, 1978.

C. Cartis and K. Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming*, pages 1–39, 2017.

Y. Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.

Y. Chi, Y. C. Eldar, and R. Calderbank. PETRELS: Parallel subspace estimation and tracking by recursive least squares from partial observations. *IEEE Transactions on Signal Processing*, 61(23):5947–5959, 2013.

M. A. Davenport and J. Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

R. Durrett. *Probability: Theory and examples*. Cambridge university press, 2010.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936. doi: 10.1007/BF02288367.

A. Eftekhari, L. Balzano, and M. B. Wakin. What to expect when you are expecting on the Grassmannian. *IEEE Signal Processing Letters*, 24(6):872–876, 2017.

A. Eftekhari, M. B. Wakin, and R. A. Ward. MC^2: A two-phase algorithm for leveraged matrix completion. *Information and Inference: A Journal of the IMA*, 7(3):581–604, 2018a.

A. Eftekhari, D. Yang, and M. B. Wakin. Weighted matrix completion and recovery with prior subspace information. *IEEE Transactions on Information Theory*, 64(6):4044–4071, 2018b.

N. Gershenfeld, S. Samouhos, and B. Nordman. Intelligent infrastructure for energy efficiency. *Science*, 327(5969):1086–1088, 2010.

G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2013.

A. Gonen, D. Rosenbaum, Y. C. Eldar, and S. Shalev-Shwartz. Subspace learning with partial information. *Journal of Machine Learning Research*, 17(52):1–21, 2016.

D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer New York, 2013.

H. Krim and M. Viberg. Two decades of array signal processing research: The parametric approach. *IEEE Signal Processing Magazine*, 13(4):67–94, 1996.

A. Lakhina, M. Crovella, and C. Diot. Diagnosing network-wide traffic anomalies. In *ACM SIGCOMM Computer Communication Review*, volume 34, pages 219–230. ACM, 2004.

B. Lois and N. Vaswani. Online matrix completion and online robust PCA. In *IEEE International Symposium on Information Theory (ISIT)*, pages 1826–1830. IEEE, 2015.

K. Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.

M. Mardani, G. Mateos, and G. B. Giannakis. Subspace learning and imputation for streaming big data matrices and tensors. *IEEE Transactions on Signal Processing*, 63(10):2663–2677, 2015.

L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *Quart. J. Math. Oxford*, pages 1156–1159, 1966.

I. Mitliagkas, C. Caramanis, and P. Jain. Memory limited, streaming PCA. In *Advances in Neural Information Processing Systems*, pages 2886–2894, 2013.

I. Mitliagkas, C. Caramanis, and P. Jain. Streaming PCA with many missing entries. *Preprint*, 2014.

J. Nocedal and S. J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2006.

E. Oja and J. Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, 106(1):69–84, 1985.

J. Tanner and K. Wei. Normalized iterative hard thresholding for matrix completion. *SIAM Journal on Scientific Computing*, 35(5):S104–S125, 2013.

L. Tong and S. Perreau. Multichannel blind identification: From subspace to maximum likelihood methods. *Proceedings of IEEE*, 86:1951–1968, 1998.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

P. van Overschee and B. L. de Moor. *Subspace identification for linear systems: Theory, implementation, applications*. Springer US, 2012.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 95–110. Cambridge University Press, 2012a.

R. Vershynin. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012b.

S. Watanabe and N. Pakvasa. Subspace method of pattern recognition. In *Proc. 1st. IJCPR*, pages 25–32, 1973.

P. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

Y. Xie, J. Huang, and R. Willett. Change-point detection for high-dimensional time series with missing data. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):12–27, 2013.

D. Zhang and L. Balzano. Global convergence of a grassmannian gradient descent algorithm for subspace estimation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, page 1460Ð1468, 2016.