What to Expect When You Are Expecting on the Grassmannian

Armin Eftekhar, Laura Balzano, and Michael B. Wakin

Abstract—Consider an incoming sequence of vectors, all belonging to an unknown subspace $S$, and each with many missing entries. In order to estimate $S$, it is common to partition the data into blocks and iteratively update the estimate of $S$ with each new incoming measurement block.

In this paper, we investigate a rather basic question: Is it possible to identify $S$ by averaging the column span of the partially observed incoming measurement blocks on the Grassmannian?

We show that in general the span of the incoming blocks is in fact a biased estimator of $S$ when data suffers from erasures, and we find an upper bound for this bias. We reach this conclusion by examining the defining optimization program for the Fréchet expectation on the Grassmannian, and with the aid of a sharp perturbation bound and standard large deviation results.

I. PROBLEM STATEMENT

Consider an $r$-dimensional subspace $S$ with orthonbasis $S \in \mathbb{R}^{n \times r}$, where $n > r$. We wish to identify $S$ from incomplete data, received sequentially, using only limited memory [1]–[3]. Streaming subspace identification from incomplete data finds application in system identification [4], [5] where data commonly suffers from erasures, or in monitoring network traffic [6], where collecting complete data is infeasible. Such applications arise also in imaging, computer vision, and communications [7]–[11], to name a few.

More concretely, for an integer $T$, let $\{q_t\}_{t=1}^T \subset \mathbb{R}^n$ be independent copies of a random vector $q \in \mathbb{R}^r$. At time $t \in \{1 : T\} := \{1, 2, \cdots , T\}$, we observe each entry of $s_t := S \cdot q_t \in S$ with a probability of $p \in (0, 1)$, and we collect the measurements in $y_t \in \mathbb{R}^n$, setting the unobserved entries to zero. To reiterate, our objective is to identify the subspace $S$ from the measurement vectors $\{y_t\}_{t=1}^T$. Throughout, we assume that $r = \text{dim}(S)$ is known a priori or estimated from data by other means.

The literature of modern signal processing offers a number of efficient algorithms to solve this problem, including (the new) SNIPE [12], GROUSE [13], as well as a generalization of the classic power method [14]. These algorithms partition the incoming measurements into non-overlapping blocks and iteratively update their estimate of the true subspace $S$ with each incoming measurement block.

This paper intends to enhance our understanding of this subject by answering a basic, but hopefully interesting, question about averaging zero-filled data on the Grassmannian manifold of $r$-dimensional subspaces. Estimation on the Grassmannian is a problem of general interest [15]. Meanwhile, zero-filling is a common step with missing data and has been shown to have reasonable estimation properties in some cases [16] or can be used as algorithmic initialization [17], [18].

While a byproduct of our work is an algorithm for subspace identification (that computes the mentioned average), more efficient techniques for subspace identification exist in the literature [12]–[14] that utilize the information from previous blocks instead of zero-filling the measurement blocks. We also note that block-based subspace averaging algorithms have been used successfully with fully observed $(p = 1)$ data [19]–[21]. With these remarks, let us now state this question in detail.

For an integer $b \geq r$, suppose we partition the incoming measurements $\{y_t\}_{t=1}^T$ into (non-overlapping) blocks of size $b$, which we denote by $\{Y_k\}_{k=1}^K \subset \mathbb{R}^{n \times b}$, assuming that the number of blocks $K = T/b$ is an integer for simplicity. Each measurement block $Y_k$ is a partially-observed copy of $S_Y = S \cdot Q_k$, where $\{Q_k\}_{k=1}^K \subset \mathbb{R}^{r \times b}$ and the coefficient matrices $\{Q_k\}_{k=1}^K \subset \mathbb{R}^{r \times b}$ are obtained by partitioning $\{s_t\}_{t=1}^T$ and $\{q_t\}_{t=1}^T$ into blocks of size $b$, respectively.

Each measurement block $Y_k$ provides a simple, if not accurate, estimate of the underlying subspace $S$. Indeed, let $Y_k \in \mathbb{R}^{n \times b}$ be a rank-$r$ truncation of $Y_k$, obtained by truncating all but the largest $r$ singular values of $Y_k$. Consider the $r$-dimensional subspace $Y_k = \text{span}(Y_k)$, and recall that $Y_k$ best approximates $\text{span}(Y_k)$ among all $r$-dimensional subspaces. We may consider $Y_k$ as an estimate of $S$. In particular, $Y_k = S$ when there is no erasure $(p = 1)$ and $Q_k$ is rank-$r$.

By construction, $\{Y_k\}_{k=1}^K$ are independent and identically distributed random subspaces on the Grassmannian $\mathbb{G}(n, r)$, the manifold of all $r$-dimensional subspaces of $\mathbb{R}^n$. It is therefore natural to consider the “average” of the subspaces $\{Y_k\}_{k=1}^K$ as an estimate of $S$. (As we will see in Section II, some care must be taken in defining this average. We will also point out that, under mild conditions, this average can be updated in a streaming fashion, making the scheme suitable for memory-limited scenarios.)

With this introduction, the present work answers the following question: What is the bias of the expectation of each $Y_k$ with respect to the true subspace $S$? In the next section, we formalize this question and find an upper bound for the bias, with our main result summarized in Theorem 1 below.
II. EXPECTATION ON GRASSMANNIAN

Consider the following metric on the Grassmannian $\mathcal{G}(n, r)$ [22]: If \(\{\theta_i(A, B)\}_{i=1}^r\) are the principal angles between \(r\)-dimensional subspaces \(A, B \in \mathcal{G}(n, r)\), their distance is

\[
d_G(A, B) = \sqrt{\frac{1}{r} \sum_{i=1}^r \theta_i(A, B)^2}.
\] (1)

For example, the distance between two one-dimensional subspaces (namely, two lines) is the smaller angle that they make. We can now define the Fréchet expectation of \(Y_k\) on \(\mathcal{G}(n, r)\) as the subspace(s) to which the expected squared distance is minimized [24], [25]. More specifically, a Fréchet expectation \(E \in \mathcal{G}(n, r)\) of random subspace \(Y_k\) is a minimizer of the program

\[
F = \arg \min_{S \in \mathcal{G}(n, r)} \mathbb{E} \left[ d_G \left( S', Y_k \right)^2 \right],
\] (2)

where the expectation is with respect to the coefficient matrix \(Q_k \in \mathbb{R}^{r \times b}\) and the support of \(Y_k\).

Is \(Y_k\) an unbiased estimator of the true subspace \(S\)? If not, how far is a Fréchet expectation \(F\) from \(S\)? We answer these questions in the rest of this section. Note also that \(F\) may be computed by empirically averaging the incoming sequence \(\{Y_t\}\) on the Grassmannian, as explained later in Remark 5.

Let us continue with a toy example with \(n = 2, r = 1\). For a very small \(\epsilon \ll 1\), we set \(S = \begin{bmatrix} \sqrt{1 - \epsilon^2} & \epsilon \end{bmatrix}^T\), so that \(S = \text{span}(S)\) is nearly aligned with the first canonical vector \(e_1 = [1, 0]^T\). Suppose that every entry of each incoming vector is independently observed with probability \(1/2\). Therefore, every \(y_t\) is either parallel to \(e_1\), or parallel to \(e_2 = [0, 1]^T\), or parallel to \(S\), or degenerate \((y_t = 0)\), each with probability \(1/4\). With block size \(b = 1\) and after ignoring the degenerate inputs, it follows that either \(Y_k = \text{span}(e_1)\), or \(Y_k = \text{span}(e_2)\), or \(Y_k = S\), each with probability \(1/3\). A short calculation reveals that the minimizer of Program (2), namely the Fréchet expectation of \(Y_k\), is unique in this case and makes an angle of about \(\pi/6\) with \(S\). That is, the Fréchet expectation of \(Y_k\) is a biased estimator of the true subspace \(S\), in general.

Perhaps this bias is somewhat unexpected, especially since each measurement block \(Y_k\) satisfies \(\mathbb{E}[Y_k] = pS \cdot \mathbb{E}[Q_k]\), and also because with probability one, \(Q_k\) will be rank-\(r\), in which case \(\text{span}(pS \cdot Q_k) = \text{span}(S) = S\). Interestingly, results in a recent paper [21] suggest that this bias disappears when averaging the span of fully observed \((p = 1)\) data blocks.

In dealing with partial samples, an important property of \(S\) proves to be its coherence, defined as

\[
\mu(S) := \frac{n}{r} \max_{i \in [1:n]} \| S[i, :]\|_2^2
\] (3)

where we use MATLAB’s matrix notation to specify the rows of \(S\) [13], [26]. One can verify that \(\mu(S)\) is independent of the choice of orthobasis \(S\) in (3), and that \(\mu(S) \in [1, n/r]\). For example, when \(S\) consists of \(r\) columns of the \(n \times n\) identity matrix, \(\mu(S) = n/r\). In contrast, when \(S\) comprises \(r\) columns of the standard Fourier matrix in \(\mathbb{C}^n\), \(\mu(S) = 1\). Moreover, introducing a second quantity,

\[
\nu(S) := \frac{n}{r} \left\| \begin{array}{cc}
\|S[1,:]\|_2 & \cdots \\
\|S[n,:]\|_2
\end{array} \right\|^2,
\] (4)

will presently enable us to more tightly control the bias of the expectation of \(Y_k\). Above, \(S^\perp \in \mathbb{R}^{n \times (n-r)}\) is an orthobasis for the orthogonal complement of the subspace \(S\). Note that \(\nu(S)\) too is independent of the choice of orthobasis in (4) and that

\[
\nu(S) \leq \mu(S).
\] (5)

In the examples above, \(\nu(S) = 0\) when \(S\) spans \(r\) columns of the identity matrix and \(\nu(S) = 1\) when \(S\) spans \(r\) columns of the Fourier matrix. As another example, \(\nu(S)\) is large when \(S = \text{span}(S)\) and the only nonzero entries of \(S\) are \(S[1,1] = S[2,1] = S[3,2] = S[4,2] = 1/\sqrt{2}\).

We are now in position to state the main result of this paper, proved in Section III. In a nutshell, this result states that the estimation bias of the Fréchet expectation is bounded by a factor of \(\sqrt{(1 + \frac{n}{r}) \mathbb{E}[\nu(S)]}\). Here and elsewhere, \(a \vee b = \max(a, b)\).

Theorem 1. (Bias of Fréchet expectation) Consider a subspace \(S \in \mathcal{G}(n, r)\). Consider also a random vector \(q \in \mathbb{R}^r\) and construct \(Q \in \mathbb{R}^{r \times b}\) by concatenating \(b\) independent copies of \(q\). Let \(\kappa(Q)\) be the condition number of \(Q\), and set \(Q = \text{span}(Q^*) \in \mathcal{G}(n, r)\). As described in Section I, construct also the random subspace \(Y_k \in \mathcal{G}(n, r)\) and its Fréchet expectation \(F \in \mathcal{G}(n, r)\). (The distribution of \(Y_k\) and hence its expectation are independent of \(k\).) Then, for any \(\alpha, \tilde{\kappa}, \tilde{\mu} Q \geq 1\), it holds that

\[
d_G(F, S)^2 \leq \alpha^2 \kappa^2 \left(1 + \frac{n}{b}\right) \frac{r \nu(S \vee \tilde{\mu} Q) \log(n \vee b)}{p n} + e^{-a} + \Pr[\kappa(Q) > \tilde{\kappa}] + \Pr[\mu(Q) > \tilde{\mu} Q],
\] (6)

provided that

\[
p \geq \alpha^2 \kappa^2 \left(1 + \frac{n}{b}\right) \frac{r \mu(S \vee \tilde{\mu} Q) \log(n \vee b)}{n}
\] (7)

holds. Here, the notation \(\lesssim\) suppresses any universal factors for simplicity.

A few remarks are in order.

Remark 1. (Coefficients) Recall that, at time \(t\), we partially observe \(S \cdot q_t\), where \(q_t\) is an independent copy of a random vector \(q \in \mathbb{R}^r\). The bound in (6) depends on the properties of the random matrix \(Q \in \mathbb{R}^{r \times b}\), formed by concatenating \(b\) independent copies of \(q\). This dependence is indeed present though often mild in practice. Indeed, if the energy of \(Q\) is concentrated, say, along its first row, then \(SQ\) hardly contains any information about, say, the second column of \(S\). In
practice, however, it is common to assume that \(q\) is a standard random Gaussian vector, in which case, \(Q\) becomes a standard random Gaussian matrix. Then, basic arguments in random matrix theory predict that

\[
\mu(Q) \lesssim \log b, \quad \kappa(Q) \lesssim \left| \frac{\sqrt{b} + \sqrt{r}}{\sqrt{b} - \sqrt{r}} \right|,
\]

with overwhelming probability [27]. In particular, \(Q\) is well-conditioned when the block size \(b\) is sufficiently large.

**Remark 2. (Coherence)** The bound on the bias in (6) depends on one of the coherence factors of \(S\), namely \(\nu(S)\) (see (4)). This dependence suggests the estimation bias is small when \(\nu(S)\) is small. In particular, for both of the earlier examples (column-subset of identity matrix and standard Fourier matrix), recall that \(\nu\) is small. We point out that the role of this coherence seems inherent to the problem. Indeed, in the example with large \(\nu(S)\) after (5), rarely are both the first and second rows observed (and, thus, nonzero).

**Remark 3. (Block size)** The bound in (6) also depends on the block size \(b\) suggesting that, to minimize the bias, the block size should ideally be comparable to \(n\), namely \(b = O(n)\). This dependence on block size was anticipated. Indeed, it is well-understood that estimating the rank can be expected to increasingly suffer as \(b\) decreases.

**Remark 4. (Measurements)** The bound on the estimation bias in (6) reduces as \(p\) increases, namely as the number of measurements collected from each incoming vector increases. In particular, \(p = 1\) means no erasure and \(Y_k = S\), if \(Q\) is almost surely full-rank. Moreover, the bound on bias is proportional to \(1/\sqrt{p}\), decreasing as \(p\) increases.

**Remark 5. (Implementation)** Efficient algorithms with convergence guarantees for computing Fréchet expectation exist in the literature of computer vision and machine learning; the recent works [19]–[21] suit us best here. For subspaces \(A, B \in \mathbb{G}(n, r)\), consider a geodesic connecting \(A\) and \(B\), namely a curve of shortest length connecting \(A\) and \(B\) (with respect to the canonical metric on the Grassmannian). Let \(A \#_p B\) be a point on the geodesic (itself a subspace in \(\mathbb{G}(n, r)\)) such that \(d_G(A, A \#_p B) = \rho \cdot d_G(A, B)\). For example, \(A \#_{1/2} B\) is half way between \(A\) and \(B\). The explicit expression for \(A \#_p B\) is given in [19]. Suppose also, for simplicity, that \(\{Y_k\}_k\) belongs to a geodesic ball on the Grassmannian with radius smaller than \(\pi/4\). Then, starting with \(F_1 = Y_1\), the recursion \(F_k = F_{k-1} \#_{1/k} Y_k\) converges linearly [21], in probability, to the Fréchet expectation \(F\), if it is unique. This recursion might be considered as a “running average” on the Grassmannian.

Let us consider an example with \(n = 50, r = 2\), setting \(q \in \mathbb{R}^r\) to be the standard random Gaussian vector. Entries of incoming vectors are observed with a probability \(p = 3r/n\) and measurements are partitioned into blocks of size \(b = 5r\). Figure 1 plots the geodesic distance \(d_G(F_k, S)\) versus \(k\) in three cases (with \(\{F_k\}_k\) defined above): first, when \(S\) is the span of a column subset of the identity matrix and, second, when \(S\) is a generic subspace (say, the span of a standard random Gaussian matrix), and third, when \(S\) is as described in the example with large \(\nu(S)\) after (5). In the first two cases, \(\nu(S)\) is small (see (5) and (8)), predicting a relatively small estimation bias. In the third case, however, \(\nu(S)\) and consequently the bias are large. This is indeed corroborated by Figure 1.

### III. Proof of Theorem 1

Let us first simplify the notation and introduce some helpful details. For \(A \in \mathbb{R}^{n \times b}\), set

\[
\mathcal{P}_p(A) := \sum_{i=1}^n \sum_{j=1}^b \epsilon_{i,j} A[i,j] \cdot E_{i,j},
\]

where \(\{\epsilon_{i,j}\}_{i,j}\) is a sequence of independent Bernoulli random variables, taking one with probability \(p\) and zero otherwise. Also, \(E_{i,j} \in \mathbb{R}^{n \times b}\) is the \([i,j]\)-th canonical matrix in \(\mathbb{R}^{n \times b}\), i.e., \(E_{i,j}[i,j] = 1\) is the only nonzero entry of \(E_{i,j}\). Let \(\Omega \subset [1 : n] \times [1 : b]\) be the random index set corresponding to the support of \(\mathcal{P}_p(A)\). We set \(Y = \mathcal{P}_p(SQ)\) and let \(Y_r\) be a rank-\(r\) truncation of \(Y\), obtained via singular value decomposition (SVD). We also let \(S = \text{span}(Y_r)\). Note that \(S\) is a rank subspace on the Grassmannian \(\mathbb{G}(n, r)\).

We wish to calculate how far the true subspace \(S\) is from Fréchet expectation(s) of \(Y\), defined as solution(s) of the program

\[
\min_{S' \in \mathbb{G}(n,r)} f(S') \quad f(S') := \mathbb{E} \left[ d_G(S', Y)^2 \right],
\]

where the expectation is with respect to the coefficient matrix \(Q\) and the support \(\Omega\). Let \(\{\theta_i(S', Y)\}_{i=1}^r\) denote the principal angles between the two subspaces \(S', Y \in \mathbb{G}(n, r)\). It is well-known that \(\{\sin(\theta_i(S', Y))\}_{i=1}^r\) are in fact the singular values of \(P_{S', Y}\), where \(P_A\) denotes the orthogonal projection matrix onto a subspace \(A\). Moreover,

\[
d_G(S', Y)^2 = \frac{1}{r} \sum_{i=1}^r \theta_i(S', Y)^2 = \frac{1}{r} \| \arcsin(P_{S' \perp Y}) \|_F^2,
\]

Figure 1: The numerical example described in Remark 5.
where \( \arcsin(\cdot) \), applied to a matrix, acts only on the singular values, leaving the singular vectors intact [23]. The geodesic distance \( d_G(S', Y) \) is tightly controlled as follows:

\[
d_G(S', Y)^2 = \frac{1}{r} \left\| \arcsin(P_{S'\perp} P_Y) \right\|_F^2 \quad \text{(see (11))}
\leq \frac{1}{r} \left\| P_{S'\perp} P_Y \right\|_F^2 \quad \text{(|| arcsin a \| \leq |a|/2)}
\leq \| P_{S'\perp} P_Y \|^2, \quad \text{(rank } P_{S'\perp} P_Y \text{) \leq r)} \quad (12)
\]

\[
d_G(S', Y)^2 = \frac{1}{r} \sum_{i=1}^{r} \theta_i(S', Y)^2 \quad \text{(see (11))}
\geq \frac{1}{r} \sum_{i=1}^{r} \sin^2(\theta_i(S', Y)) \quad \text{(|a| \geq |\sin a|)}
= \frac{1}{r} \| P_{S'\perp} P_Y \|_F^2 \quad . \quad (13)
\]

In turn, (12) and (13) allow us to tightly control \( f(S') \) for arbitrary \( S' \in G(n, r) \):

\[
f(S') = \frac{1}{r} E \left[ d_G(S', Y)^2 \right] \quad \text{(see (10))}
\leq E \left[ \| P_{S'\perp} P_Y \|^2 \right], \quad \text{(see (12))} \quad (14)
\]

\[
f(S') \geq \frac{1}{r} E \left[ \| P_{S'\perp} P_Y \|^2 \right] \quad . \quad (15)
\]

Of particular interest to us is evaluating \( f(S) \) which, as (14) and (15) suggest, requires an estimate of the principal angle \( \sin(\theta_1(S, Y)) = \| P_{S\perp} P_Y \| \). The following result is proved in Section IV with the aid of a sharp perturbation bound, as well as basic large deviation results.

**Lemma 1.** Fix \( Q \in \mathbb{R}^r \times b \) with rank \( r \) and let \( \kappa(Q) \) be its condition number. Also set \( Q = \text{span}(Q^*) \). For \( \alpha \geq 1 \) and except with a probability of at most \( e^{-\alpha} \), it holds that

\[
\| P_{S\perp} P_Y \|
\leq \alpha \cdot \kappa(Q) \left( 1 + \sqrt{\frac{n}{b}} \right) \sqrt{\frac{r (\nu(S) \vee \mu(Q)) \log (n + b)}{pn}},
\]

provided that

\[
p \geq \alpha^2 \kappa(Q)^2 \left( 1 + \frac{n}{b} \right) \frac{r (\mu(S) \vee \mu(Q)) \log (n + b)}{n}.
\]

Lemma 1 readily translates into a bound on \( f(S) \): For \( \alpha, \bar{\kappa}, \bar{\mu}_Q \geq 1 \), suppose that

\[
p \geq \alpha^2 \bar{\kappa}^2 \left( 1 + \frac{n}{b} \right) \frac{r (\mu(S) \vee \bar{\mu}_Q) \log (n + b)}{n},
\]

and conveniently let \( \mathcal{E} \) denote the event where both \( \kappa(Q) \leq \bar{\kappa}_Q \) and \( \mu(Q) \leq \bar{\mu}_Q \) hold. Also set

\[
\Delta := \alpha \cdot \kappa(Q) \left( 1 + \sqrt{\frac{n}{b}} \right) \sqrt{\frac{r (\nu(S) \vee \mu(Q)) \log (n + b)}{pn}}.
\]

and let \( \mathcal{E}' \) be the event where \( \| P_{S\perp} P_Y \|^2 \leq \Delta^2 \). Then, note that

\[
f(S) \leq E \left[ \| P_{S\perp} P_Y \|^2 \right] \quad \text{(seen (14))}
= E \left[ \| P_{S\perp} P_Y \|^2 | \mathcal{E}' \right] \cdot \Pr[\mathcal{E}]
+ E \left[ \| P_{S\perp} P_Y \|^2 | \mathcal{E}^C \right] \cdot \Pr[\mathcal{E}^C]
\leq E \left[ \| P_{S\perp} P_Y \|^2 | \mathcal{E}' \right] + \Pr[\mathcal{E}^C], \quad (17)
\]

where \( \mathcal{E}^C \) is the complement of the event \( \mathcal{E} \). To bound the expectation in the last line above, let \( \mathcal{E}' \) be the event where \( \| P_{S\perp} P_Y \|^2 \leq \Delta^2 \) and note that

\[
E \left[ \| P_{S\perp} P_Y \|^2 | \mathcal{E}' \right] = E Q \left[ \| P_{S\perp} P_Y \|^2 | \mathcal{E}' \right]
= E Q \left[ \| P_{S\perp} P_Y \|^2 | \mathcal{E}^C \right] \Pr[\mathcal{E}^C]
+ E Q \left[ \| P_{S\perp} P_Y \|^2 | \mathcal{E}^C \right] \Pr[\mathcal{E}^C] \cdot |E|
\leq E Q \left[ \| P_{S\perp} P_Y \|^2 | \mathcal{E}^C \right] \Pr[\mathcal{E}^C] \cdot |E|
\leq \Delta^2 + e^{-\alpha} |E| \quad \text{(see Lemma 1)} \quad (18)
\]

Substituting the bound above back into (17), we find that

\[
f(S)
\leq \Delta^2 + e^{-\alpha} + \Pr[\mathcal{E}^C]. \quad \text{see (17)}
\]

(19)

In words, when \( p \) is sufficiently large, \( f(S) \) is small. Let us next find a lower bound on \( f(S) \) far from \( S \): For an arbitrary subspace \( S' \in G(n, r) \), we note that

\[
f(S') \geq \frac{1}{r} E \left[ \| P_{S'\perp} P_Y \|^2 \right] \quad \text{(seen (15))}
\geq \frac{1}{r} \left( \| P_{S'\perp} P_S \|^2 - \| P_{S'\perp} (P_S - E[P_Y]) \|^2 \right) \quad \text{(triangle inequ.)}
\geq \frac{1}{2r} \left( \| P_{S'\perp} P_S \|^2 - \| P_{S'\perp} (P_S - E[P_Y]) \|^2 \right)
\geq \frac{1}{2r} \left( \| P_{S'\perp} P_S \|^2 - \frac{2}{r} E \left[ \| P_{S\perp} P_Y \|^2 \right] \right)
\geq \frac{1}{2r} \left( \| P_{S'\perp} P_S \|^2 - \frac{2}{r} E \left[ \| P_{S\perp} P_Y \|^2 \right] \right)
\geq \frac{1}{2r} \left( \| P_{S'\perp} P_S \|^2 - 2E \left[ \| P_{S\perp} P_Y \|^2 \right] \right) \quad \text{(rank } P_{S\perp} P_Y \text{) \leq r)}
\geq \frac{1}{2r} \left( \| P_{S'\perp} P_S \|^2 - 2C_1 \left( \Delta^2 + e^{-\alpha} + \Pr[\mathcal{E}^C] \right) \right) \quad \text{(see (19))}
\]

for an absolute constant \( C_1 > 0 \). Above, we used the inequality \( (a - b)^2 \geq \frac{a^2}{2} - b^2 \) for scalars \( a, b \) and the fact that \( \| P_S - P_Y \|^2 = \| P_{S\perp} P_Y \|^2 \). To summarize in words, \( f(S) \) is small for large \( p \) because of (19). Moreover, thanks to (20), we know that \( f(S') \) is large for any subspace \( S' \) far
from $S$. Therefore, any minimizer of $f(\cdot)$ in (10) (namely, any Fréchet expectation $F$) must be close to the true subspace $S$. More formally,

$$\|P_{S^r} P_{S^r}\|_F^2 \geq \Delta^2 + \Pr[\mathcal{E}^C] \implies f(S') > f(S),$$

(21)

and, consequently,

$$\frac{1}{r} \sum_{i=1}^r \sin^2(\theta_i(F, S)) = \frac{\|P_{E^r} P_{E^r}\|_F^2}{r} \leq \Delta^2 + e^{-\alpha} + \Pr[\mathcal{E}^C],$$

(22)

which completes the proof of Theorem 1 after noting that $\Pr[\mathcal{E}^C] \leq \Pr[s(Q) > \ell] + \Pr[|\mu(Q) | > \mu_Q]$, and using the fact that $|a| \leq \pi |\sin(a)|/2$ when $|a| \leq \pi/2$.

REFERENCES

[1] M.K. Warmuth and D. Kuzmin. Randomized online pca algorithms with regret bounds that are logarithmic in the dimension. Journal of Machine Learning Research, 9(Oct):2287–2320, 2008.

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

[3] W. Yang and H. Xu. Streaming sparse principal component analysis. In Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pages 494–503, 2015.

[4] P. van Overschee and B.L. de Moor. Subspace identifiability: Theory, implementation, applications. Springer US, 2012.

[5] Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. SIAM Journal on Matrix Analysis and Applications, 31(3):1235–1256, 2009.

[6] 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.

[7] 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.

[8] D. Manolakis and G. Shaw. Detection algorithms for hyperspectral imaging applications. IEEE Signal Processing Magazine, 19(1):29–43, 2002.

[9] J.P. Costeira and T. Kanade. A multibody factorization method for independently moving objects. International Journal of Computer Vision, 29(3):159–179, 1998.

[10] H. Krim and M. Viberg. Two decades of array signal processing: the parametric approach. IEEE Signal processing magazine, 13(4):67–94, 1996.

[11] L. Tong and S. Perreau. Multichannel blind identification: From subspace to maximum likelihood methods. PROCEEDINGS-IEEE, 86:1951–1968, 1998.

[12] A. Efekhari, L. Balzano, D. Yang, and M. B. Wakin. SNPE for memory-limited PCA from incomplete data. arXiv preprint arXiv:1612.00904, 2016.

[13] 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.

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

[15] Y. Hong, X. Yang, R. Kwitt, M. Stynner, and M. Niethammer. Regression uncertainty on the Grassmannian. In Proceedings of the AISTATS Conference, 2017.

[16] S. Chatterjee. Matrix estimation by universal singular value thresholding. The Annals of Statistics, 43(1):177–214, 2015.

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

[18] R.S. Ganti, L. Balzano, and R. Willett. Matrix completion under monotonic single index models. In Advances in Neural Information Processing Systems, pages 1873–1881, 2015.

[19] R. Chakraborty and B.C. Venuri. Recursive Fréchet mean computation on the Grassmannian and its applications to computer vision. In Proceedings of the IEEE International Conference on Computer Vision, pages 4229–4237, 2015.

[20] H. Salehian, R. Chakraborty, E. Ofori, D. Vaillancourt, and B.C. Venuri. An efficient recursive estimator of the Fréchet mean on a hypersphere with applications to medical image analysis. Mathematical Foundations of Computational Anatomy, 2015.

[21] R. Chakraborty, S. Hauberg, and B. C. Venuri. Intrinsic Grassmann averages for online linear and robust subspace learning. arXiv preprint arXiv:1702.00105, 2017.

[22] P-A Abilis. Robert Mahony, and Rodolphe Sepulchre. Riemannian geometry of grassmann manifolds with a view on algorithmic computation. Acta Applicandae Mathematica, 80(2):199–220, 2004.

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

[24] M. Fréchet. Les élménts aléleatoires de nature quelconque dans un espace distancié. Annales de l’institut Henri Poincaré, 10:215–310, 1948.

[25] H. Karcher. Riemannian center of mass and mollifier smoothing. Journal of the American Statistical Association, 99, pages 505–511, 2004.

[26] Y. Chen. Incoherence-optimal matrix completion. IEEE Transactions on Information Theory, 57(6):2990–2993, 2011.

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

[28] I.M. Johnstone and A.Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. Journal of the American Statistical Association, 2012.

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

[30] J.A. Tropp. User-friendly tail bounds for sums of random matrices. arXiv preprint arXiv:1011.3027, 2010.

IV. PROOF OF LEMMA 1

Fix $Q \in \mathbb{R}^{n \times r}$ for now and assume $Q$ is rank-$r$. Consider the measurement matrix $Y = P_{S^r}(SQ)$ in $\mathbb{R}^{n \times r}$ and let $Y_r \in \mathbb{R}^{n \times k}$ be a rank-$r$ truncation of $Y$, obtained via SVD, and set $Y \in \mathbb{R}^{n \times r} = \text{span}(Y_r)$. Let also $Y_\rightarrow := Y - Y_r$ denote the residual. Note that

$$\|P_{S^r}Y\| = \|P_{S^r}Y_\rightarrow + Y_r\| = \|P_{S^r}(Y - Y_\rightarrow) + Y_r\| \leq \|P_{S^r}Y_\rightarrow\| \sigma_r(Y_r) \leq \frac{\|P_{S^r}Y_\rightarrow\|}{\sigma_r(pSQ) - \|Y_r - pSQ\|} \leq \frac{\|P_{S^r}Y\|}{\sigma_r(\mu(Q)) - \|Y_r - pSQ\|},$$

(23)

which is slightly sharper than the standard perturbation bound [29, Theorem 3], and the difference is consequential in our problem. We next control both norms in the last line above. Beginning with the numerator, we write that

$$P_{S^r}Y = P_{S^r}P_{S}(SQ) = P_{S^r} \sum_{i,j}^r \epsilon_{i,j} \cdot (SQ)[i,j] \cdot E_{i,j} \quad (\text{see } (9))$$

$$= \sum_{i,j}^r Z_{i,j},$$

(24)
where \( \{ Z_{i,j} \}_{i,j} \subset \mathbb{R}^{n \times b} \) are independent zero-mean random matrices. In order to appeal to the matrix Bernstein inequality [30], some preparation is required:

\[
\nu(S) = \frac{n}{r} \left\| S \right\|_2^2 = \frac{n}{r} \left( \sum_{i,j} \left| S_{i,j} \right|^2 \right)^{1/2} \cdot \left\| S^\perp \right\|_2 \cdot \left\| S \right\|_2 \quad \text{(see (4))}
\]

\[
\max_{i,j} \| Z_{i,j} \| = \max_{i,j} \| \epsilon_{i,j} \cdot (SQ)[i,j] \cdot P_{S^\perp} E_{i,j} \| \quad \text{(see (24))}
\]

\[
\leq \max_{i,j} \| (SQ)[i,j] \cdot P_{S^\perp} E_{i,j} \| \quad (\epsilon_{i,j} \in \{0, 1\})
\]

\[
\leq \max_{i,j} \| S[i,:] \|_2 \cdot \| Q[:,j] \|_2 \cdot \| P_{S^\perp} E_{i,j} \|
\]

\[
\leq \left\| Q \right\| \sqrt{\frac{r \nu(Q)}{b}} \max_{i,j} \| S[i,:] \|_2 \cdot \| P_{S^\perp} E_{i,j} \| \quad \text{(see (3))}
\]

\[
\leq \left\| Q \right\| \sqrt{\frac{r \nu(Q)}{b}} \max_{i,j} \| S[i,:] \|_2 \cdot \| S^\perp[i,:] \|_2 \leq \left\| Q \right\| \sqrt{\frac{r \nu(S)}{b}} =: \beta, \quad \text{(see (25))}
\]

\[
\left\| \mathbb{E}_\Omega \left[ \sum_{i,j} Z_{i,j} Z_{i,j}^* \right] \right\| \leq \sum_{i,j} \mathbb{E}_\Omega \left[ \epsilon_{i,j}^2 \cdot (SQ)[i,j] \right] \cdot \| P_{S^\perp} E_{i,j} P_{S^\perp} \| \quad (\epsilon_{i,j} \in \{0, 1\})
\]

\[
= p \sum_{i,j} \left| (SQ)[i,j] \right|^2 \cdot E_{i,j} P_{S^\perp} \quad (\epsilon_{i,j} \sim \text{Bernoulli}(p))
\]

\[
= p \sum_{i,j} \left\| S[i,:] \right\|_2 \cdot \| P_{S^\perp} E_{i,j} P_{S^\perp} \| \leq p \| Q \|^2 \left( 1 + \sqrt{\frac{n}{b}} \cdot \sqrt{\frac{r \nu(S) \cup \mu(Q)}{n}} \right)
\]

(see (27) and (28))

(26)

(28)

In fact, using a slightly different argument, \( p \) in (29) can be replaced with \( p(1 - p) \). However, since \( p \) is typically small, this does not lead to a substantial improvement in final results and is therefore ignored here.) We are now in position to apply the matrix Bernstein inequality [30]: For \( \alpha \geq 1 \) and except with a probability of at most \( e^{-\alpha} \), it holds that

\[
\| P_{S^\perp} Y \| \leq \sum_{i,j} \mathbb{E}_\Omega \left[ Z_{i,j} \right] \quad (\text{see (24))}
\]

\[
\leq \alpha \cdot \max \left( \log(n \cup b) \cdot \beta, \sqrt{\log(n \cup b)} \cdot \sigma \right)
\]

\[
\leq \alpha \| Q \| \left( 1 + \sqrt{\frac{n}{b}} \cdot \max \left( \log(n \cup b) \cdot \sqrt{\frac{r \nu(S) \cup \mu(Q)}{n}} \right) \right)
\]

\[
= \alpha \| Q \| \left( 1 + \sqrt{\frac{n}{b}} \cdot \sqrt{\log(n \cup b)} \cdot \sqrt{\frac{r \nu(S) \cup \mu(Q)}{n}} \right)
\]

(see (9))

(30)
when $p$ is sufficiently large. On the other hand, note that $E_\Omega[\mathcal{P}_p(SQ)] = pSQ$. In fact, another application of the matrix Bernstein’s inequality proves that

$$\| \mathcal{P}_p(SQ) - pSQ \|
\leq C\alpha \|Q\| \left(1 \vee \sqrt{\frac{n}{\kappa}}\right) \cdot \sqrt{\log(n \vee b)} \cdot \sqrt{\frac{pr(\mu(S) \vee \mu(Q))}{n}},$$

(31)

for some constant $C$.

In particular, with $\kappa(Q) = \|Q\|/\sigma_r(Q)$ denoting the condition number of $Q$ and after taking

$$p \geq C^2\alpha^2\kappa(Q)^2 \left(1 \vee \frac{n}{\kappa}\right) \frac{r(\mu(S) \vee \nu(Q)) \log(n \vee b)}{n},$$

(32)

we find that

$$\| \mathcal{P}_p(SQ) - pSQ \| \leq \frac{p \cdot \sigma_r(Q)}{2}.$$

(33)

Plugging (30) and (33) back into (23) yields that

$$\| \mathbb{P}_{S^\perp} \mathbb{P}_Y \|
\leq \frac{\| \mathbb{P}_{S^\perp} Y \|}{\| \mathbb{P}_{S^\perp} Y\| - \| Y_r - pSQ \|} \quad \text{(see (23))}
\leq \frac{\| \mathbb{P}_{S^\perp} Y \|}{\| \mathbb{P}_{S^\perp} Y\|/2} \quad \text{(see 33)}
\lesssim \alpha \cdot \kappa(Q) \left(1 \vee \sqrt{\frac{n}{\kappa}}\right) \frac{r(\nu(S) \vee \mu(Q)) \log(n \vee b)}{pn},$$

provided that $p$ satisfies (32). The last line above uses (30). This completes the proof of Lemma 1.