What to Expect When You Are Expecting on the Grassmannian

Citation for published version:
Eftekhar, A, Balzano, L & Wakin, MB 2016 'What to Expect When You Are Expecting on the Grassmannian'.

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Early version, also known as pre-print

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
What to Expect
When You Are Expecting on the Grassmannian

Armin Eftekhari, Laura Balzano, 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 find 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 orthonormal basis \( S \in \mathbb{R}^{n \times r} \). 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, for example, system identification [4], [5] where data commonly suffers from erasures. Similar applications arise in imaging, computer vision, and communications [6]–[10], to name a few.

More concretely, for an integer \( T \), let \( \{q_t\}_{t=1}^T \subset \mathbb{R}^r \) be independent copies of a random vector \( q \in \mathbb{R}^r \). At time \( t \in [1 : T] := \{1, 2, \ldots, T\} \), we observe each entry of \( s_t := S \cdot q_t \in \mathbb{S} \) 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 GROUSE [11], as well as a generalization of the classic power method [12]. Both algorithms partition the incoming measurements into non-overlapping blocks and iteratively update their estimate of the true subspace \( S \) with each incoming measurement block.\(^1\)

This paper does not offer a more efficient algorithm for subspace identification but rather intends to enhance our understanding of this subject by answering a basic question regarding averaging zero-filled data on the Grassmannian. We next state the problem 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 \( S \), where \( \{S_k\}_{k=1}^K \subset \mathbb{R}^{n \times b} \) and the covariance matrices \( \{Q_k\}_{k=1}^K \subset \mathbb{R}^{b \times b} \) 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 \( Y_k \subset \mathbb{R}^{n \times b} \) 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 \( S_k = \text{span}(Y_k) \), and recall that \( Y_k \) best approximates \( \text{span}(S) \) 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 \( \mathcal{G}(n, r) \), 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 \( Y_k \) as an estimator of 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 [13]: If \( \{\theta_i(A, B)\}_{i=1}^r \) are the principal angles between \( r \)-dimensional

\(^1\)Strictly speaking, GROUSE uses blocks of size one, updating its estimate of \( S \) with each new measurement vector. However, the extension to larger blocks is straightforward. In fact, the authors are preparing a manuscript that introduces SNIPE, a new algorithm for subspace identification which might be considered a generalization of GROUSE to large blocks, and with stronger supporting theory.
subspaces $A, B \in \mathbb{G}(n, r)$, their distance is

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

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 $\mathbb{G}(n, r)$ as the subspace(s) to which the expected squared distance is minimized [15], [16]. More specifically, a Fréchet expectation $F \in \mathbb{G}(n, r)$ of random subspace $Y_k$ is a minimizer of the program

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

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 it from a Fréchet expectation $F$ from $S$? We answer these questions in the rest of this section.

Let us continue with a toy example with $n = 2, r = 1$. For a very small $\epsilon \ll 1$, we set

$$S = \left[ \sqrt{1 - \epsilon^2}, \epsilon \right]^*,$$

so that $S = \text{span}(S)$ is nearly aligned with the first canonical vector $e_1 = [1, 0]^*$. Suppose that every entry of each incoming vector is independently observed with probability of $1/2$. Therefore, every $y_i$ is either parallel to $e_1$, or parallel to $e_2 = [0, 1]^*$, or parallel to $S$, or degenerate ($y_i = 0$), each with probability of $1/4$. 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 $Y_i$, in expectation and if rank$(\mathbb{E}[Q_k]) = r$, spans the true subspace $S$, namely span$(\mathbb{E}[Y_k]) = S$. In dealing with partial samples, an important property of $S$ proves to be its coherence, defined as

$$\mu(S) := \max_{i \in [1:n]} \|S[i, :]\|_2, \quad (3)$$

where we use MATLAB’s matrix notation to specify the rows of $S$ [11], [17]. One can verify that $\mu(S)$ is independent of the choice of orthobasis $S$ in (4), 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

$$\text{r columns of the standard Fourier matrix in } \mathbb{C}^n, \mu(S) = 1.$$\footnote{Principal angles between subspaces generalize the notion of angle between lines. See [14] for more details.}

Moreover, introducing a second quantity,

$$\nu(S) := \frac{n}{r} \left[ \begin{array}{c} \|S[1, :]\|_2 \\ \vdots \\ \|S[n, :]\|_2 \\ \end{array} \right] \left[ \begin{array}{c} S[1, :] \\ \vdots \\ S[n, :] \end{array} \right]^2, \quad (4)$$

will presently enable us to more tightly control the bias of the expectation of $Y_k$. Above, $S^2 \subseteq \mathbb{R}^{n \times (n-r)}$ 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

$$\nu(S) \leq \mu(S). \quad (5)$$

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.$^4$\footnote{Such a minimizer exists by Weierstrass’s theorem since the objective function of Program (2) is continuous and $\mathbb{G}(n, r)$ is compact. We also remark that, alternatively, one might define the Fréchet expectation only if there exists a unique minimizer to Program (2). This alternative definition will not be used here, as it does not fit the nature of our analysis. Also, we have discarded from this expectation any matrices for which rank$(Y_k) < r$. Such matrices can arise, with low probability, when too few samples are collected from $S_k$.}

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 $\sqrt{\left(1 + \frac{\nu}{\mu} \right) \frac{\nu}{\mu} \frac{n}{r}}$. Here and elsewhere, $a \lor b = \max(a, b)$.

**Theorem 1. (Bias of Fréchet expectation)** Consider a subspace $S \in \mathbb{G}(n, r)$. 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 $Q = \text{span}(Q^*) \in \mathbb{G}(n, r)$. As described in Section I, construct also the random subspace $Y_k \in \mathbb{G}(n, r)$ and its Fréchet expectation $F \in \mathbb{G}(n, r)$. (The distribution of $Y_k$ and hence its expectation are independent of $k$.) Then, for any $\alpha, \beta, \mu > 0$, it holds that

$$d_G(F, S)^2 \lesssim \alpha^2 \kappa^2 \left( 1+\frac{n}{b} \right) \frac{r \nu(S) \lor \mu(Q) \log(n \lor b)}{\sqrt{\beta}} + e^{-\alpha} + \text{Pr}[\kappa(Q) > \kappa] + \text{Pr}[\mu(Q) > \mu] \cdot \frac{1}{2}, \quad (6)$$

provided that the right-hand side is $O(1)$. 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 often mild in practice. For example, 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

$$\mu(Q) \lesssim \log b, \quad \kappa(Q) \lesssim \frac{\sqrt{b} + \sqrt{r}}{b - \sqrt{r}} \quad (7)$$

with overwhelming probability [18]. 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 the two coherence factors of $S$, namely $\mu(S)$ and $\nu(S)$ (see $^4$As another example, $\nu(S)$ is large when $S = \text{span}(S)$ and the only nonzero entries of $S$ are $S[1, 1] = S[2, 1] = S[3, 2] = S[4, 2] = 1/\sqrt{2}$.}
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.

**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-$r$ covariance matrix of a random vector $X \in \mathbb{R}^n$ requires $O(n)$ samples in the presence of noise [19].

**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 for computing Fréchet expectation exist in the literature of computer vision and machine learning; the recent works [20], [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 $A \#_r B$ be a point on the geodesic (itself a subspace in $\mathbb{G}(n, r)$) such that $d_G(A, A \#_r B) = \rho \cdot d_G(A, B)$. For example, $A \#_{1/2} B$ is half way between $A$ and $B$. (The explicit expression for $A \#_r B$ is given in [20].) Suppose also, for simplicity, that $(Y_k)_k$ belongs to a geodesic ball on the Grassmannian with radius smaller than $\frac{\pi}{4}$. Then, starting with $F_1 = Y_1$, the recursion $F_k = F_{k-1} \#_{1/k} Y_k$ converges, 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 of $p = 3r/n$ 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 Footnote 4. In the first two cases, $\nu(S)$ is small (see (5) and (7)), 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

$$P_p(A) := \sum_{i=1}^n \sum_{j=1}^n \epsilon_{i,j} A[i, j] \cdot E_{i,j},$$

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 $\mathbb{R}^{n \times b}$.

**Figure 1:** The numerical example described in Remark 5.

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 $P_p(A)$. We set $Y = P_p(SQ)$ and let $Y_r \in \mathbb{R}^{n \times b}$ be a rank-$r$ truncation of $Y$, obtained via singular value decomposition (SVD). We also let $Y = \text{span}(Y_r)$. Note that $Y$ is a random subspace on the Grassmannian $\mathcal{G}(n, r)$.

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

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

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 $S', Y \in \mathcal{G}(n, r)$. It is well-known that $\{ \sin(\theta_i(S', Y)) \}_{i=1}^r$ are in fact the singular values of $P_{S'} P_Y$, and that

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

where $\arcsin(\cdot)$, applied to a matrix, acts only on the singular values, leaving the singular vectors intact [14]. The geodesic distance $d_G(S', Y)$ is tightly controlled as follows:

$$d_G(S', Y)^2 = \frac{1}{r} \| \arcsin{(P_{S' \perp} P_Y)} \|^2_F \quad (\text{see } (10))$$

$$\lesssim \frac{1}{r} \| P_{S' \perp} P_Y \|^2_F \quad (| \arcsin| a | \leq \pi |a|/2)$$

$$\leq \| P_{S' \perp} P_Y \|^2_F \quad \text{ (rank } (P_{S' \perp} P_Y) \leq r) \quad (11)$$

$$d_G(S', Y)^2 = \frac{1}{r} \sum_{i=1}^r \theta_i(S', Y)^2 \quad (\text{see } (10))$$

$$\geq \frac{1}{r} \sum_{i=1}^r \sin^2(\theta_i(S', Y)) \quad (|a| \geq |\sin a|)$$

$$= \frac{1}{r} \| P_{S' \perp} P_Y \|^2_F.$$  

(12)
In turn, (11) and (12) allow us to tightly control \( f(S') \) for arbitrary \( S' \in \mathbb{R}^{r \times b} \):

\[
\begin{align*}
\mathbb{E} \left[ \| P_{S'} \cdot P_Y \| \right] &\leq \mathbb{E} \left[ \| P_{S'} \cdot P_Y \| ^2 \right] \quad \text{(see (13))} \\
\mathbb{E} \left[ \| P_{S'} \cdot P_Y \| \right] &\geq \mathbb{E} \left[ \| P_{S'} \cdot P_Y \| ^2 \right] \quad \text{(see (12))}
\end{align*}
\]

(14)

Substituting the bound above back into (16), we find that

\[
\begin{align*}
f(S) &\leq \mathbb{E} \left[ \| P_{S} \cdot P_Y \| ^2 \right] + \mathbb{P} \left[ \mathcal{E} \right] \quad \text{(see (16))} \\
&\leq \Delta^2 + e^{-a} + \mathbb{P} \left[ \mathcal{E} \right].
\end{align*}
\]

(18)

In words, when \( p \) is sufficiently large, \( f(S) \) is small. Let us next find a lower bound on \( f(\cdot) \) far from \( S \): For an arbitrary subspace \( S' \in \mathbb{R}^{r \times b} \), we note that

\[
\begin{align*}
f(S') &\geq \mathbb{E} \left[ \| P_{S'} \cdot P_Y \| ^2 \right] \quad \text{(see (14))} \\
&\geq \mathbb{E} \left[ \| P_{S'} \cdot P_Y \| ^2 \right] \quad \text{(Jensen’s ineq.)} \\
&\geq \mathbb{E} \left[ \| P_{S'} \cdot P_Y \| ^2 \right] \quad \text{(triangle ineq.)} \\
&\geq \mathbb{E} \left[ \| P_{S'} \cdot P_Y \| ^2 \right] \quad \text{(text)} \\
&\geq \mathbb{E} \left[ \| P_{S'} \cdot P_Y \| ^2 \right] \quad \text{(Jensen’s ineq.)} \\
&\geq \mathbb{E} \left[ \| P_{S'} \cdot P_Y \| ^2 \right] \quad \text{(text)} \\
&\geq \frac{1}{2} \mathbb{E} \left[ \| P_{S'} \cdot P_Y \| ^2 \right] \quad \text{(rank \( P_S \cdot P_Y \leq r \))}
\end{align*}
\]

(19)

for an absolute constant \( C_1 > 0 \). Above, we used the inequality \( a \cdot b \geq \frac{a^2}{2} - b^2 \) for scalars \( a,b \) and the fact that \( \| P_S \cdot P_Y \| ^2 = 2 \| P_S \cdot P_Y \| ^2 \). To summarize in words, \( f(S) \) is small for large \( p \) because of (18). Moreover, thanks to (19), we know that \( f(S') \) is large for any subspace \( S' \) far from \( S \). Therefore, any minimizer of \( f(\cdot) \) in (9) (namely, any Fréchet expectation \( F \)) must be close to the true subspace \( S \). More formally,

\[
\begin{align*}
\| P_{S'} \cdot P_Y \| ^2 &\geq \frac{1}{r} \sum_{i=1}^{r} \sin^2 (\theta_i (F,S)) = \frac{\| P_{F} \cdot P_Y \| ^2}{r} \leq \Delta^2 + e^{-a} + \mathbb{P} \left[ \mathcal{E} \right],
\end{align*}
\]

(21)

which completes the proof of Theorem 1 after noting that \( \mathbb{P} \left[ \mathcal{E} \right] \leq \mathbb{P} \left[ \kappa(Q) > \bar{\kappa} \right] + \mathbb{P} \left[ \mu(Q) > \bar{\mu} \right] \), and using the fact that \( |a| \leq \pi |\sin(a)| / 2 \) when \( |a| \leq \pi / 2 \).

IV. PROOF OF LEMMA 1

Fix \( Q \in \mathbb{R}^{r \times b} \) for now and assume \( Q \) is rank-\( r \). Consider the measurement matrix \( Y = P_p(SQ) \in \mathbb{R}^{n \times b} \) and let \( Y_r \in \mathbb{R}^{n \times b} \) be a rank-\( r \) truncation of \( Y \), obtained via SVD, and set

\[
\begin{align*}
\mathbb{E} \left[ \| P_{S'} \cdot P_Y \| ^2 \right] &\geq \mathbb{E} \left[ \| P_{S'} \cdot P_Y \| ^2 \right] \quad \text{(see Lemma 1)} \\
&\leq \Delta^2 + e^{-a}
\end{align*}
\]

(17)
Note that

\[ Y = \text{span}(Y_r). \] Let also \( Y_{r+} := Y - Y_r \) denote the residual. Note that

\[
\|P_{S^\perp} P_{Y_r}\| = \left\| P_{S^\perp} (Y - Y_{r+}) Y_r \right\| \\
= \left\| P_{S^\perp} Y_r \right\| \quad (Y_r Y_{r+}^* = 0) \\
\leq \left\| P_{S^\perp} Y_r \right\| \cdot \left\| Y_r \right\| \\
= \left\| P_{S^\perp} Y \right\| \quad \sigma_r(Y_r) \\
\leq \frac{\left\| P_{S^\perp} Y \right\|}{\sigma_r(pSQ) - \left\| Y_r - pSQ \right\|} \quad \text{(Weyl's inequality)} \\
= \frac{p \cdot \sigma_r(Q) - \left\| Y_r - pSQ \right\|}{\left\| P_{S^\perp} Y \right\|} \quad \text{(\(S^* S = I_r\))} \\
\leq \frac{p\|Q\|^2 \left\| \sum_i |S[i, :]|_2^2 \cdot P_{S^\perp} E_{i,i} P_{S^\perp} \right\|}{\left\| P_{S^\perp} Y \right\|} \quad \text{(22)}
\]

which is slightly sharper than the standard perturbation bound [22, 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

\[
P_{S^\perp} Y = P_{S^\perp} P_{p} (SQ) \\
= P_{S^\perp} \sum_{i,j} \epsilon_{i,j} \cdot (SQ)[i, j] \cdot E_{i,i} \quad \text{(see (8))} \\
= \sum_{i,j} Z_{i,j}, \quad \text{(23)}
\]

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 [23], some preparation is required:

\[
\nu(S) = \frac{n}{r} \left\| \left[ \left\| S[1, :]\right\|_2 \right. \right. \\
\left. \left. \quad \quad \vdots \right. \right. \left. \quad \left. \| S[n, :]\right\|_2 \right]\| \cdot S^\perp |^{2} \quad \text{(see (4))} \\
= \frac{n}{r} \max_i \left\| S[i, :]\right\|_2^2 \cdot P_{S^\perp} E_{i,i} P_{S^\perp} \| \\
\leq \frac{n}{r} \max_i \left\| S[i, :]\right\|_2^2 \cdot P_{S^\perp} E_{i,i} P_{S^\perp} \| \\
= \frac{n}{r} \max_i \left\| S[i, :]\right\|_2^2 \cdot \left\| S^\perp[i, :]\right\|_2^2 \quad \text{(24)}
\]

\[
\max_{i,j} \|Z_{i,j}\| \\
\leq \max_{i,j} \|\epsilon_{i,j} \cdot (SQ)[i, j] \cdot P_{S^\perp} E_{i,i}\| \quad \text{(see (23))} \\
\leq \max_{i,j} \|(SQ)[i, j] \cdot P_{S^\perp} E_{i,i}\| \quad \epsilon_{i,j} \in \{0, 1\} \\
\leq \max_{i,j} \|S[i, :]\|_2 \cdot \|Q[\cdot, j]\|_2 \cdot \|P_{S^\perp} E_{i,i}\| \\
\leq \|Q\| \frac{r \mu(Q)}{b} \max_{i,j} \|S[i, :]\|_2 \cdot \|P_{S^\perp} E_{i,i}\| \quad \text{(see (3))} \\
= \|Q\| \frac{r \mu(Q)}{b} \max_{i,j} \|S[i, :]\|_2 \cdot \|S^\perp[i, :]\|_2 \\
\leq \|Q\| \frac{r \mu(Q)}{b} \sqrt{\frac{r \nu(S)}{n}} =: \beta, \quad \text{(see (24))} \quad \text{(25)}
\]

\[
\sigma^2 := \left\| E_{i,j} \sum_{i,j} Z_{i,j} Z_{i,j}^* \right\| \vee \left\| E_{i,j} \sum_{i,j} Z_{i,j}^* Z_{i,j} \right\| \\
\leq \|Q\|^2 \left( 1 + \frac{n}{b} \right) \cdot \frac{r \nu(S) \vee \mu(Q)}{n} \quad \text{(26)} \quad \text{and (27)}.
\]

(In fact, using a slightly different argument, \( p \) in (28) 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
Plugging (29) and (32) back into (22) yields that
\[\|P_{3 \perp P}Y\| \leq \frac{\|P_{S^2}Y\|}{p \cdot \sigma_r(Q)} \] (see (22))
\[\lesssim \alpha \cdot \kappa(Q) \left(1 + \sqrt{\frac{n}{b}}\right) \cdot \sqrt{\log(n \vee b)} \cdot \sqrt{p} \cdot \sqrt{n} \cdot \sqrt{\mu(S) \vee \mu(Q)}\] (Bernstein’s ineq.)
for some constant \(E\).

In particular, with \(\kappa(Q) = \|Q\|/\sigma_r(Q)\) denoting the condition number of \(Q\) and after taking
\[p \geq C^2 \alpha^2 \kappa^2(Q) \left(1 + \sqrt{\frac{n}{b}}\right) r \left(\mu(S) \vee \mu(Q)\right) \log(n \vee b)\]
we find that
\[\|P_{p}(SQ) - pSQ\| \leq \frac{\|P_{p}(SQ) - pSQ\|}{p \cdot \sigma_r(Q)} \leq \frac{\|P_{p}(SQ) - pSQ\|}{p \cdot \sigma_r(Q)/2}\]
(32)

Plugging (29) and (32) back into (22) yields that
\[\|P_{3 \perp P}Y\| \leq \frac{\|P_{S^2}Y\|}{p \cdot \sigma_r(Q)} - \|Y_{r} - pSQ\|\] (see (22))
\[\lesssim \alpha \cdot \kappa(Q) \left(1 + \sqrt{\frac{n}{b}}\right) \cdot \sqrt{\log(n \vee b)} \cdot \sqrt{p} \cdot \sqrt{n} \cdot \sqrt{\mu(S) \vee \mu(Q)}\] (Bernstein’s ineq.)
\[
\leq \alpha \cdot \kappa(Q) \left(1 + \sqrt{\frac{n}{b}}\right) \cdot \sqrt{\log(n \vee b)} \cdot \sqrt{p} \cdot \sqrt{n} \cdot \sqrt{\mu(S) \vee \mu(Q)}\] (Bernstein’s ineq.)
\[
\leq \alpha \cdot \kappa(Q) \left(1 + \sqrt{\frac{n}{b}}\right) \cdot \sqrt{\log(n \vee b)} \cdot \sqrt{p} \cdot \sqrt{n} \cdot \sqrt{\mu(S) \vee \mu(Q)}\]

provided that \(p\) satisfies (31). This completes the proof of Lemma 1.

### References


