Fast and Sample-Efficient Federated Low Rank Matrix Recovery from column-wise Linear and Quadratic Projections

Seyedehsara (Sara) Nayer and Namrata Vaswani
Dept. of Electrical and Computer Engineering, Iowa State University, USA.
Email: namrata@iastate.edu

***Abstract*—We study the following lesser-known low rank (LR) recovery problem: recover an $n \times q$ rank-r matrix, $\boldsymbol{X}^* = [\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, ..., \boldsymbol{x}_q^*]$, with $r \ll \min(n, q)$, from m independent linear projections of each of its q columns, i.e., from $\boldsymbol{y}_k := \boldsymbol{A}_k \boldsymbol{x}_k^*, k \in [q]$, when $\boldsymbol{y}_k$ is an m-length vector with $m < n$. The matrices $\boldsymbol{A}_k$ are known and mutually independent for different k. We introduce a novel gradient descent (GD) based solution called AltGD-Min. We show that, if the $\boldsymbol{A}_k$s are i.i.d. with i.i.d. Gaussian entries, and if the right singular vectors of $\boldsymbol{X}^*$ satisfy the incoherence assumption, then ϵ-accurate recovery of $\boldsymbol{X}^*$ is possible with order $(n+q)r^2 \log(1/\epsilon)$ total samples and order $mqnr \log(1/\epsilon)$ time. Compared with existing work, this is the fastest solution. For $\epsilon < r^{1/4}$, it also has the best sample complexity. A simple extension of AltGD-Min also provably solves LR Phase Retrieval, which is a magnitude-only generalization of the above problem.**

AltGD-Min factorizes the unknown $\boldsymbol{X}$ as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{B}$ where $\boldsymbol{U}$ and $\boldsymbol{B}$ are matrices with r columns and rows respectively. It alternates between a (projected) GD step for updating $\boldsymbol{U}$, and a minimization step for updating $\boldsymbol{B}$. Its each iteration is as fast as that of regular projected GD because the minimization over $\boldsymbol{B}$ decouples column-wise. At the same time, we can prove exponential error decay for it, which we are unable to for projected GD. Finally, it can also be efficiently federated with a communication cost of only nr per node, instead of nq for projected GD.

I. Introduction

This work develops a sample-efficient, fast, and communication-efficient gradient descent (GD) solution, called AltGD-Min, for provably recovering a low-rank (LR) matrix from a set of mutually independent linear projections of each of its columns. The communication-efficiency considers a federated setting. This problem, which we henceforth refer to as "Low Rank column-wise Compressive Sensing (LRcCS)", is precisely defined below. Unlike the other well-studied LR problems – multivariate regression (MVR) [1], LR matrix sensing [2] and LR matrix completion (LRMC) [3], [2] – LRcCS has received little attention so far in terms of approaches with provable guarantees. There are only two existing provably correct solutions. (1) Its generalization *LR phase retrieval (LRPR)*, was studied in our recent work [4], [5], [6] where we developed a provably correct alternating minimization (AltMin) solution. Since LRPR is a generalization, the algorithm also solves LRcCS. (2) In parallel work, [7] developed and analyzed a convex relaxation (mixed-norm minimization) for LRcCS. Both solutions are much slower than GD-based methods, and, in most practical settings, also have worse sample complexity.

LRcCS occurs in accelerated LR dynamic MRI [8], [9], [10], and in distributed/federated sketching [11], [12], [7]. We explain these in Sec. I-D. We show the speed and performance advantage of AltGD-Min for dynamic MRI in [13].

A. Problem Setting, Notation, and Assumption

Problem definition. The goal is to recover an $n \times q$ rank-r matrix $\boldsymbol{X}^* = [\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \dots, \boldsymbol{x}_q^*]$ from m linear projections (sketches) of each of its q columns, i.e. from

$$\boldsymbol{y}_k := \boldsymbol{A}_k \boldsymbol{x}_k^*, \ k \in [q] \tag{1}$$

where each $\boldsymbol{y}_k$ is an m-length vector, $[q] := \{1, 2, \dots, q\}$, and the measurement/sketching matrices $\boldsymbol{A}_k$ are mutually independent and known. The setting of interest is low-rank (LR), $r \ll \min(n, q)$, and undersampled measurements, $m < n$. Our guarantees assume that each $\boldsymbol{A}_k$ is random-Gaussian: each entry of it is independent and identically distributed (i.i.d.) standard Gaussian.

We also study the magnitude-only measurements' setting, LRPR [4], [5], [6]. This involves recovering $\boldsymbol{X}^*$ from

$$\boldsymbol{y}_{(mag)_k} := |\boldsymbol{A}_k \boldsymbol{x}_k^*|, \ k \in [q].$$

Here $|\boldsymbol{z}|$ takes the entry-wise absolute value of entries of the vector $\boldsymbol{z}$.

Notation. Everywhere, $\|.\|_F$ denotes the Frobenius norm, $\|.\|$ without a subscript denotes the (induced) l_2 norm (often called the operator norm or spectral norm), $\|\boldsymbol{M}\|_{\max}$ is the maximum magnitude entry of the matrix $\boldsymbol{M}$, $^\top$ denotes matrix or vector transpose, and $|\boldsymbol{z}|$ for a vector $\boldsymbol{z}$ denotes element-wise absolute values. $\boldsymbol{I}_n$ (or sometimes just $\boldsymbol{I}$) denotes the $n \times n$ identity matrix. We use $\boldsymbol{e}_k$ to denote the k-th canonical basis vector, i.e., the k-th column of $\boldsymbol{I}$. For any matrix $\boldsymbol{Z}$, $\boldsymbol{z}_k$ denotes its k-th column.

We say $\boldsymbol{U}$ is a *basis matrix* if it contains orthonormal columns. For basis matrices $\boldsymbol{U}_1, \boldsymbol{U}_2$, we use

$$\mathrm{SD}(\boldsymbol{U}_1, \boldsymbol{U}_2) := \|(\boldsymbol{I} - \boldsymbol{U}_1 \boldsymbol{U}_1{}^\top)\boldsymbol{U}_2\|_F$$

as the Subspace Distance (SD) measure. For two r-dimensional subspaces, this is the l_2 norm of the sines of the r principal angles between $\mathrm{span}(\boldsymbol{U}_1)$ and $\mathrm{span}(\boldsymbol{U}_2)$. $\mathrm{SD}(\boldsymbol{U}_1, \boldsymbol{U}_2)$ is symmetric when $\boldsymbol{U}_1, \boldsymbol{U}_2$ are both $n \times r$ basis matrices. Notice here we are using the Frobenius SD, unlike many recent works including our older work [5] that use the induced 2-norm

one. This is done because it enables us to prove the desired guarantees easily. We reuse the letters c, C to denote different numerical constants in each use with the convention that $c < 1$ and $C \geq 1$. The notation $a \in \Omega(b)$ means $a \geq Cb$ while $a \in O(b)$ means $a \leq Cb$. We use $\mathbb{1}_{\text{statement}}$ to denote an indicator function that takes the value 1 if statement is true and zero otherwise.

For a vector $\boldsymbol{w}$, we sometimes use $\boldsymbol{w}(k)$ to denote the k-th entry of $\boldsymbol{w}$. For a vector $\boldsymbol{w}$ and a scalar α, $\mathbb{1}(\boldsymbol{w} \leq \alpha)$ returns a vector of 1s and 0s of the same length as $\boldsymbol{w}$, with 1s where $(\boldsymbol{w}(k) \leq \alpha)$ and zero everywhere else. We use $\circ$ to denote the Hadamard product. Thus $\boldsymbol{z} := \boldsymbol{w} \circ \mathbb{1}(\boldsymbol{w} \leq \alpha)$ zeroes out entries of $\boldsymbol{w}$ larger than α, while keeping the smaller ones as is.

For $\boldsymbol{X}^*$ which is a rank-r matrix, we let

$$\boldsymbol{X}^* \overset{\text{SVD}}{=} \boldsymbol{U}^* \underbrace{\boldsymbol{\Sigma}^* \boldsymbol{V}^*}_{\boldsymbol{B}^*} := \boldsymbol{U}^* \boldsymbol{B}^*$$

denote its reduced (rank r) SVD, i.e., $\boldsymbol{U}^*$ and ${\boldsymbol{V}^*}^\top$ are matrices with orthonormal columns *(basis matrices)*, $\boldsymbol{U}^*$ is $n \times r$ and $\boldsymbol{V}^*$ is $r \times q$, and $\boldsymbol{\Sigma}^*$ is an $r \times r$ diagonal matrix with non-negative entries. We use $\kappa := \sigma^*_{\max}/\sigma^*_{\min}$ to denote the condition number of $\boldsymbol{\Sigma}^*$. This is not the condition number of $\boldsymbol{X}^*$ (whose minimum singular value is zero). We let $\boldsymbol{B}^* := \boldsymbol{\Sigma}^* \boldsymbol{V}^*$ and we use $\boldsymbol{b}^*_k$ to denote its k-th column.

We use the phrase *ϵ-accurate recovery* to refer to $\text{SD}(\boldsymbol{U}, \boldsymbol{U}^*) \leq \epsilon$ or $\|\boldsymbol{X} - \boldsymbol{X}^*\|_F \leq \epsilon \|\boldsymbol{X}^*\|_F$ or both.

Assumption. Another way to understand (1) is as follows: each scalar measurement $\boldsymbol{y}_{ki}$ (i-th entry of $\boldsymbol{y}_k$) satisfies

$$\boldsymbol{y}_{ki} := \langle \boldsymbol{a}_{ki}, \boldsymbol{x}^*_k \rangle, \ i \in [m], \ k \in [q]$$

with ${\boldsymbol{a}_{ki}}^\top$ being the i-th row of $\boldsymbol{A}_k$. Observe that the measurements are not global, i.e., no $\boldsymbol{y}_{ki}$ is a function of the entire matrix $\boldsymbol{X}^*$. They are global for each column ($\boldsymbol{y}_{ki}$ is a function of column $\boldsymbol{x}^*_k$) but not across the different columns. We thus need an assumption that enables correct interpolation across the different columns. The following assumption, which is a slightly weaker version of incoherence (w.r.t. the canonical basis) of right singular vectors suffices for this purpose.

Assumption 1.1 ((Weakened) Right Singular Vectors' Incoherence)**.** *Assume that*

$$\max_k \|\boldsymbol{b}^*_k\| \leq \sigma^*_{\max} \mu \sqrt{r/q}.$$

for a constant $\mu \geq 1$ (μ does not grow with n, q, r). Since $\|\boldsymbol{x}^*_k\| = \|\boldsymbol{b}^*_k\|$, this implies that $\max_k \|\boldsymbol{x}^*_k\| \leq \sigma^*_{\max} \mu \sqrt{r/q}$. Also, since $\sigma^*_{\min} \sqrt{r} \leq \|\boldsymbol{X}^*\|_F$, this also implies that $\max_k \|\boldsymbol{x}^*_k\| \leq \kappa \mu \|\boldsymbol{X}^*\|_F / \sqrt{q}$.

Right singular vectors incoherence is the assumption $\max_k \|\boldsymbol{v}^*_k\| \leq \mu \sqrt{r/q}$. Since $\boldsymbol{b}^*_k = \boldsymbol{\Sigma}^* \boldsymbol{v}^*_k$, this implies that the above holds. Incoherence of both left and right singular vectors was introduced for guaranteeing correct "interpolation" for the LRMC problem [3], [2].

B. Existing Work

Existing solutions for LRcCS and LRPR. Since it is always possible to obtain magnitude-only measurements $\boldsymbol{y}_{(mag)_k}$ from linear ones $\boldsymbol{y}_k$ as $\boldsymbol{y}_{(mag)_k} = |\boldsymbol{y}_k|$, a solution to LRPR also automatically solves LRcCS under the same assumptions. Hence the AltMin algorithm for LRPR from [4], [5] is the first provably correct solution for LRcCS. Of course, since LRcCS is an easier problem than LRPR, we expect a direct solution to LRcCS to need weaker assumptions. As we show in this paper, this is indeed true. A more recent work [7] studied the noisy version of LRcCS and developed a convex relaxation (mixed norm minimization) to provably solve it. Its time complexity is not discussed in the paper, however, it is well known that solvers for convex programs are much slower when compared to direct iterative algorithms: they either require number of iterations proportional to $1/\sqrt{\epsilon}$ or the per-iteration cost has cubic dependence on the problem size (here $((n+q)r)^3$) [2]. Thus, if $q \leq n$, its time complexity $O(mqnr \cdot \min(1/\sqrt{\epsilon}, n^3 r^3))$. In [6], we provided the best possible guarantee for the AltMin algorithm for solving LRPR, and hence LRcCS. We discuss these results in detail in Sec. II-D and summarize them in Table I.

Other well-studied LR recovery problems. The multivariate regression (MVR) problem, studied in [1], is our problem with $\boldsymbol{A}_k = \boldsymbol{A}$. However this is a very different setting than ours because, with $\boldsymbol{A}_k = \boldsymbol{A}$, the different $\boldsymbol{y}_k$'s are no longer mutually independent. As a result, one cannot exploit law of large numbers' arguments over all mq scalar measurements $\boldsymbol{y}_{ki}$. Consequently, the required value of m can never be less than n. The result of [1] shows that m of order $(n+q)r$ is both necessary and sufficient. LRMS involves recovering $\boldsymbol{X}^*$ from $\boldsymbol{y}_i = \langle \boldsymbol{A}_i, \boldsymbol{X}^* \rangle, \ i = 1, 2, \ldots, mq$ with $\boldsymbol{A}_i$ being dense matrices, typically i.i.d. Gaussian [2]. Thus all measurements are i.i.d. and *global*: each contains information about the entire quantity-of-interest, here $\boldsymbol{X}^*$. Because of this, for LRMS, one can prove a LR Restricted Isometry Property (RIP) that simplifies the rest of the analysis. This is what makes it very different from, and easier than, our problem.

LRMC, which involves recovering $\boldsymbol{X}^*$ from a subset of its observed entries, *is the most closely related problem to ours* since it also involves recovery from non-global measurements. The typical model assumed is that each matrix entry is observed with probability p independent of others [3], [2]. Setting unobserved entries to zero, this can be written as $\boldsymbol{y}_{jk} = \delta_{jk} \boldsymbol{X}^*_{jk}$ with $\delta_{jk} \overset{\text{iid}}{\sim} Bernoulli(p)$. LRMC measurements are both row-wise and column-wise local. To allow correct interpolation across both rows and columns, it needs the incoherence assumption on both its left and right singular vectors. For our problem, the measurements are global for each column, but not across the different columns. For this reason, only right singular vectors' incoherence is needed. In fact, because of the nature of our measurements, even if left incoherence were assumed, it would not help. This *asymmetry in our measurement model and the fact that our measurements are unbounded (each $\boldsymbol{y}_{ki}$ is a Gaussian r.v) are two key differences* between LRMC and LRcCS *that prevent us from borrowing LRMC proof techniques for our work*. Here *symmetric* means: if we replace $\boldsymbol{X}^*$ by its transpose, the probability distribution of the set of measurements does not change. *Bounded* means that the measurements' magnitude has a *uniform* bound. This bound is

$\|\boldsymbol{X}^*\|_{\max}$ for LRMC measurements.

Non-convex (iterative, not convex relaxation based) LRMC algorithms with the best sample complexity are GD-based. There are two common approaches for designing GD algorithms in the LR recovery literature, and in particular for LRMC. The first is to use standard projected GD on $\boldsymbol{X}$ (*projGD-X*), also referred to as Iterative Hard Thresholding: at each iteration, perform one step of GD for minimizing the squared loss cost function, $\tilde{f}(\boldsymbol{X})$, w.r.t. $\boldsymbol{X}$, followed by projecting the resulting matrix onto the space of rank r matrices (by SVD). This was studied in [14], [15] for solving LRMC. This is shown to converge geometrically with a constant GD step size, while needing only $\Omega((n+q)r^2\log^2 n\log^2(1/\epsilon))$ samples on average.

The second is to let $\boldsymbol{X}=\boldsymbol{U}\boldsymbol{B}$ where $\boldsymbol{U}$ is $n\times r$ and $\boldsymbol{B}$ is $r\times q$ and perform alternating GD for the cost function $f(\boldsymbol{U},\boldsymbol{B}):=\tilde{f}(\boldsymbol{U}\boldsymbol{B})$, i.e., update $\boldsymbol{B}$ with one step of GD for minimizing $f(\boldsymbol{U},\boldsymbol{B})$ while keeping $\boldsymbol{U}$ fixed at its previous value, and then do the same for $\boldsymbol{U}$ with $\boldsymbol{B}$ fixed, and repeat. Since the $\boldsymbol{X}=\boldsymbol{U}\boldsymbol{B}$ factorization is not unique, i.e., $\boldsymbol{X}=\boldsymbol{U}\boldsymbol{R}^{-1}\boldsymbol{R}\boldsymbol{B}$ for any invertible $r\times r$ matrix $\boldsymbol{R}$, this approach can result in the norm of one of $\boldsymbol{U}$ or $\boldsymbol{B}$ growing in an unbounded fashion, while that of the other decreases at the same rate, causing numerical problems. A typical approach to resolve this issue, and one that was used for LRMC [16], [17], is to change the cost function to minimize to $f(\boldsymbol{U},\boldsymbol{B})+\lambda f_2(\boldsymbol{U},\boldsymbol{B})$ where $f_2(\boldsymbol{U},\boldsymbol{B}):=\|\boldsymbol{U}^\top\boldsymbol{U}-\boldsymbol{B}\boldsymbol{B}^\top\|_F$ is the "norm-balancing term" (helps ensure that norms of $\boldsymbol{U}$ and $\boldsymbol{B}$ remain similar). We henceforth refer to this approach as *altGDnormbal*. The sample complexity bound for this approach is similar to that for projGD-X. But, it needs a GD step size of order $1/r$ or smaller [16], [17]; making it r-times slower than projGD-X.

C. Contributions and Novelty

Contribution to solving LRcCS and LRPR. (1) This work develops a novel GD-based solution to LRcCS, called AltGD-Min, that is fast and communication-efficient. We show that, with high probability (w.h.p.), AltGD-Min obtains an ϵ-accurate estimate in order $\kappa^2\log(1/\epsilon)$ iterations, as long as Assumption 1.1 holds, the matrices $\boldsymbol{A}_k$ are i.i.d., with each containing i.i.d. standard Gaussian entries, $mq\in\Omega(\kappa^6\mu^2(n+q)r^2\log(1/\epsilon))$, and $m\in\Omega(\max(\log q,\log n)\log(1/\epsilon))$. Its time complexity is $O(mqnr\cdot\kappa^2\log(1/\epsilon))$ and its communication complexity per node is $O(nr\cdot\kappa^2\log(1/\epsilon))$. We provide a comparison of our guarantee with those of other works in Table I. This table also summarizes the guarantees for the two most sample-efficient LRMC solutions: projGD-X and altGDnormbal. The former is also the fastest LRMC solution, while the latter is the most communication-efficient. As mentioned earlier, LRMC is the most similar problem to ours that has been extensively studied. Notice that, our sample complexity matches that of the best results for LRMC algorithms that do solve a convex relaxation. (2) We show that a simple extension of AltGD-Min also provides the fastest provable solution to LRPR, as long as the above assumptions hold and $mq\in\Omega(\kappa^6\mu^2nr^2(r+\log(1/\epsilon))$. Its time complexity is the same too.

Contributions / Novelty of algorithm design and proof techniques. As explained earlier, there are three commonly used provably correct iterative algorithms for LR recovery problems – altMin, projGD-X, and altGD (altGDnormbal to be precise). AltMin is slower than GD-based methods because, for updating both $\boldsymbol{U}$ and $\boldsymbol{B}$, it requires solving a minimization problem keeping the other variable fixed. For our specific asymmetric problem, the min step for $\boldsymbol{U}$ is the slow one. ProjGD-X and altGDnormbal are faster, but it is not clear how to analyze them for LRcCS under the desired sample complexity[1]. Our novel altGD-min approach however resolves both issues: it is fast as projGD-X and it can be analyzed. Moreover, its communication complexity for a federated implementation (and its memory complexity) is only nr per node per iteration, instead of nq for projGD-X. As can be seen from Table I, treating κ,μ as numerical constants, it has the best sample-, time-, and communication/memory-complexity among all approaches for LRcCS and all fast (iterative) approaches for LRMC as well. Because of this, an AltGD-Min type algorithm may also be of interest for solving LRMC in a fast, sample-efficient and communication-efficient fashion. In fact, it can be also be useful for other bilinear inverse problems such as blind deconvolution.

AltGDmin algorithm. The main idea is as follows. Express $\boldsymbol{X}$ as $\boldsymbol{X}=\boldsymbol{U}\boldsymbol{B}$ and alternatively update $\boldsymbol{U}$ and $\boldsymbol{B}$ as follows: (a) keeping $\boldsymbol{B}$ fixed at its previous value, update $\boldsymbol{U}$ by a GD step for it for the cost function $f(\boldsymbol{U},\boldsymbol{B})$ followed by projecting the output onto the space of matrices with orthonormal columns; and (b) keeping $\boldsymbol{U}$ fixed at its previous value, update $\boldsymbol{B}$ by minimizing $f(\boldsymbol{U},\boldsymbol{B})$ over it. Because of the column-wise decoupled form of our measurement model, step (b) is as fast as the GD step and thus the per-iteration time complexity of AltGD-Min is equal to that of any other GD method such as projGD-X or altGDnormbal. This decoupling (which means that, given $\boldsymbol{U}$, $\boldsymbol{b}_k$ only depends on $\boldsymbol{x}_k^*$, and not on the other columns of $\boldsymbol{X}^*$) also allows us to get the desired tight-enough bound on $\max_k\|\boldsymbol{b}_k-\boldsymbol{U}^\top\boldsymbol{x}_k^*\|$ and hence on $\max_k\|\boldsymbol{x}_k-\boldsymbol{x}_k^*\|$. This, and the fact that we use the gradient w.r.t. $\boldsymbol{U}$ in our algorithm, means that the summands in the gradient, and in other error bound terms, are *nice-enough sub-exponential random variables (r.v.s)*: sub-exponential r.v.s whose maximum sub-exponential norm is small enough (is proportional to (r/q)), so that the summation can be bounded w.h.p. under the desired sample complexity.

AltGDmin analysis. When we analyzed the AltMin approach for LRPR [5], [6], we could directly modify proof techniques from AltMin for LRMC [2] for getting a bound on $\mathrm{SD}(\boldsymbol{U},\boldsymbol{U}^*)$ in terms of the bound on this distance from the previous iteration. We cannot do this for AltGD-Min because the algorithm itself is different from the two GD approaches studied for solving LRMC. We instead analyze AltGD-Min by

[1]In order to show that a GD-based algorithm converges, one needs to be able to bound the norm of the gradient and show that it goes to zero with iterations. When studying both projGD-X and altGDnormbal, for different reasons, the estimates of the different columns are coupled. Consequently, it is not possible to get a tight enough bound on $\max_k\|\boldsymbol{x}_k^*-\boldsymbol{x}_k\|$. But, due to the form of the LRcCS measurement model, such a bound is needed to get a tight enough bound on the 2-norm of the gradient of the cost function, and show that it decreases sufficiently at each iteration, under the desired sample complexity. Moreover, in case of projGD-X, even if one could somehow get the desired bound, it would not suffice because the summands will still be too heavy tailed. This point is explained in detail in Appendix A.

	Sample Comp. $mq \gtrsim$	Time Comp.	Communic. Comp. per node (predicted)	Holds for all $\boldsymbol{X}^*$?	Column-wise error bound?
Convex [7]	$nr\frac{1}{\epsilon^4}$	linear-time $\cdot \min\left(\frac{1}{\sqrt{\epsilon}}, n^3r^3\right)$	not clear	yes	no
AltMin [4], [5]	$nr^4\log(\frac{1}{\epsilon})$	linear-time $\cdot\, r\log^2(\frac{1}{\epsilon})$	$nr\log(\frac{1}{\epsilon})\cdot r\log^2(\frac{1}{\epsilon})$	no	
AltMin [6]	$nr^2(r+\log(\frac{1}{\epsilon}))$	linear-time $\cdot\, r\log^2(\frac{1}{\epsilon})$	$nr\log(\frac{1}{\epsilon})\cdot r\log^2(\frac{1}{\epsilon})$	no	yes
altGD-Min (proposed)	$\mathbf{nr^2}\log(\frac{1}{\epsilon})$	linear-time $\cdot\, \mathbf{r}\log(\frac{1}{\epsilon})$	$\mathbf{nr}\cdot\mathbf{r}\log(\frac{1}{\epsilon})$	no	yes
Best sample LRMC algorithms among those that do not solve a convex relaxation					
ProjGD-X [15]	$\max(n,q)r^2\log^2 n\log^2(\frac{1}{\epsilon})$	linear-time $\cdot\, \mathbf{r}\log(\frac{1}{\epsilon})$	nq **		
AltGDnormbal [16]	$\max(\mathbf{n},\mathbf{q})\mathbf{r^2}\log\mathbf{n}$	linear-time $\cdot\, r^2\log(\frac{1}{\epsilon})$	$\max(\mathbf{n},\mathbf{q})\mathbf{r}$		

**The communication complexity of ProjGD-X would be nq because the gradient w.r.t. $\boldsymbol{X}$ computed at each node will need to be transmitted by the nodes to the center. The gradient w.r.t. $\boldsymbol{X}$ is not low rank (LR), and hence one cannot transmit just its rank r SVD.

TABLE I: Existing work versus our work. For brevity, this table assumes $q \le n$ and treats κ, μ as numerical constants. All approaches also need $m \ge \max(r, \log q, \log n)$. Column-wise error bound exists means $\max_k \|\boldsymbol{x}_k^* - \boldsymbol{x}_k\|/\|\boldsymbol{x}_k^*\| \le \epsilon$ holds in addition to a similar bound on matrix Frobenius norm error. Linear-time is the time needed to read all algorithm inputs. For LRcCS, this is $\boldsymbol{y}_k, \boldsymbol{A}_k$ for all $k \in [q]$ and thus linear-time is order mnq. For LRMC, this is the set of observed entries and their locations and thus linear-time is order mq. None of the other algorithms have been studied in the federated context and hence the communication complexity (Comm. Comp.) listed in the fourth column is based on our understanding of how one would federate the algorithm. Notice that AltGD-min has the best time and communication complexities; and for $\epsilon^4 < r$, it also has the best sample complexity.

a novel use of the fundamental theorem of calculus [18] that, along with other linear algebra tricks, helps us get a bound on $\mathrm{SD}(\boldsymbol{U}, \boldsymbol{U}^*)$ which has the desired property: the terms in it are sums of *nice-enough sub-exponentials*. See Lemma 3.4 and its proof. The use of this result is motivated by its use in [19], and many earlier works, where it is used in a standard way: to bound the Euclidean distance, $\|\boldsymbol{x} - \boldsymbol{x}^*\|$, for standard GD to solve the PR problem for recovering a single vector $\boldsymbol{x}^*$. Thus, at the true solution $\boldsymbol{x} = \boldsymbol{x}^*$, the gradient of the cost function was zero. In our case, there are two differences: (i) we need to bound the subspace distance error, and (ii) our algorithm is not standard GD, and this means that $\nabla_U f(\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U}, \boldsymbol{B}) \neq 0$. We explain our approach in Sec. III-B.

AltGDmin initialization. The standard LR spectral initialization approach cannot be used because its summands are sub-exponential r.v.s that are not *nice-enough*. We give a detailed explanation in Appendix A. We address this issue by borrowing the truncation idea from the PR literature [20], [21], [5]. But, in our case, truncation is applied to a non-symmetric matrix. Thus the sandwiching arguments developed for symmetric matrices in [20], and modified in [21], [5], cannot be borrowed. We need a different argument which is used for proving Lemma B.2 and is briefly explained in Sec. III-D.

D. Applications

The LRcCS and LRPR problems occur in projection imaging applications involving sets of images, e.g., dynamic MRI [8], [9], [10], federated LR sketching [11], [7], and dynamic Fourier ptychography (LRPR) [22]. In MRI, Fourier projections of the region of interest, e.g., a cross-section of the brain or the heart, are acquired one coefficient at a time, making the scanning (data acquisition) quite slow. Hence, reduced sample complexity enables accelerated scanning. Since medical image sequences are usually slow changing, the LR model is a valid assumption for a time sequence [8], [9], [10]. In our notation, $\boldsymbol{x}_k^*$ is the vectorized version of the k-th image of the sequence and there are a total of q images. The matrices $\boldsymbol{A}_k$ are random Fourier, i.e., $\boldsymbol{A}_k = \boldsymbol{H}_k\boldsymbol{F}$ where $\boldsymbol{F}$ is the $n \times n$ matrix that models computation of the 2D discrete Fourier transform as a matrix-vector operation, and $\boldsymbol{H}_k$ is an $m \times n$ random sampling "mask" matrix that models the frequency selection. In [13], we have shown the power of AltGD-Min for fast undersampled dynamic MRI of medical image sequences. It is both much faster, and in most cases, also provides better reconstructions, than many existing solutions from the MRI literature.

Large scale usage of smartphones results in large amounts of geographically distributed data, e.g., images. There is a need to compress/sketch this data before storing it. Sketch refers to a compression approach where the compression end is low complexity, usually simple linear projections [11], [7]. Consider the setting where different subsets of columns of $\boldsymbol{X}^*$ (each column corresponds to one vectorized image) are available at each of the $\rho \le q$ nodes. The goal is to sketch them so that they can be correctly recovered using a federated algorithm. We can store the sketches $\boldsymbol{y}_k := \boldsymbol{A}_k\boldsymbol{x}_k^*$ with $\boldsymbol{A}_k$'s being i.i.d. Gaussian. This way we store a total of only mq scalars, with mq of order roughly just $(n+q)r^2$. Traditional LR sketching approaches, e.g., [23], are designed for centralized settings and will not be efficient in a distributed setting.

E. Organization

In Sec. II, we develop AltGD-Min, give its guarantee for solving LRcCS, and compare it with existing results. We state

and prove the two theorems that help prove our main result in Sec. III. This section also contains brief proof outlines before the actual proofs. The lemmas used in these proofs are proved in Sec. IV. The extension for solving LRPR is developed, and its guarantee is stated and proved, in Sec. V. We discuss the limitations of our results in Sec. VI. Simulation experiments are provided in Sec. VII. We conclude in Sec. VIII.

II. The Proposed AltGD-Min Algorithm and Guarantee

A. The AltGD-Min algorithm

We would like to design a fast GD algorithm to find the matrix $\boldsymbol{X}$ that minimizes the squared-loss cost function $\tilde{f}(\boldsymbol{X}) := \sum_{k=1}^{q} \|\boldsymbol{y}_k - \boldsymbol{A}_k \boldsymbol{x}_k\|^2$. For reasons described earlier, we decompose $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{B}$ and develop an alternating GD-min (AltGD-Min) approach for the squared loss function,

$$f(\boldsymbol{U}, \boldsymbol{B}) := \tilde{f}(\boldsymbol{U}\boldsymbol{B}) = \sum_k \|\boldsymbol{y}_k - \boldsymbol{A}_k \boldsymbol{U} \boldsymbol{b}_k\|^2.$$

Starting with a careful initialization for $\boldsymbol{U}$ explained below, AltGD-Min proceeds as follows. At each new iteration,

- *Min-B:* update $\boldsymbol{B}$ by solving $\boldsymbol{B} \leftarrow \arg\min_{\tilde{\boldsymbol{B}}} f(\boldsymbol{U}, \tilde{\boldsymbol{B}})$. Since $\boldsymbol{b}_k$ only occurs in the k-th summand of $f(\boldsymbol{U}, \boldsymbol{B})$, this decouples to a much simpler column-wise least squares (LS) problem: $\boldsymbol{b}_k \leftarrow \arg\min_{\tilde{\boldsymbol{b}}_k} \|\boldsymbol{y}_k - \boldsymbol{A}_k \boldsymbol{U} \tilde{\boldsymbol{b}}_k\|^2$. This is solved in closed form as $\boldsymbol{b}_k = (\boldsymbol{A}_k \boldsymbol{U})^{\dagger} \boldsymbol{y}_k$ for each k; here $\boldsymbol{M}^{\dagger} := (\boldsymbol{M}^\top \boldsymbol{M})^{-1} \boldsymbol{M}^\top$.
- *ProjGD-U:* update $\boldsymbol{U}$ by one GD step for it, $\hat{\boldsymbol{U}}^+ \leftarrow \boldsymbol{U} - \eta \nabla_U f(\boldsymbol{U}, \boldsymbol{B})$, followed by projecting $\hat{\boldsymbol{U}}^+$ onto the space of matrices with orthonormal columns to get the updated $\boldsymbol{U}^+$. We get $\boldsymbol{U}^+$ by QR decomposition: $\hat{\boldsymbol{U}}^+ \stackrel{\text{QR}}{=} \boldsymbol{U}^+ \boldsymbol{R}^+$.

Notice that, because of the decoupling for $\boldsymbol{B}$, the min step only involves solving q r-dimensional Least Squares (LS) problems, in addition to also first computing the matrices, $\boldsymbol{A}_k \boldsymbol{U}$. Computing the matrices needs time of order mnr, and solving one LS problem needs time of order mr^2. Thus, the LS step needs time $O(q \max(mnr, mr^2)) = O(mqnr)$ since $r \le n$. This is equal to the time needed to compute the gradient w.r.t. $\boldsymbol{U}$; and thus, the per-iteration cost of AltGD-Min is only $O(mqnr)$. The QR decomposition of an $n \times r$ matrix takes time only nr^2.

Since $f(\boldsymbol{U}, \boldsymbol{B})$ is not a convex function of the unknowns $\{\boldsymbol{U}, \boldsymbol{B}\}$, a careful initialization is needed. Borrowing the spectral initialization idea from LRMC and LRMS solutions, we should initialize $\boldsymbol{U}_0$ by computing the top r singular vectors of

$$\boldsymbol{X}_{0,full} = \frac{1}{m}[(\boldsymbol{A}_1^\top \boldsymbol{y}_1), (\boldsymbol{A}_2^\top \boldsymbol{y}_2), \ldots, (\boldsymbol{A}_k^\top \boldsymbol{y}_k), \ldots (\boldsymbol{A}_q^\top \boldsymbol{y}_q)]$$

Clearly the expected value of the k-th column of this matrix equals $\boldsymbol{x}_k^*$ and thus $\mathbb{E}[\boldsymbol{X}_{0,full}] = \boldsymbol{X}^*$. But, as we explain next, it is not clear how to prove that this matrix concentrates around $\boldsymbol{X}^*$. Observe that it can also be written as

$$\boldsymbol{X}_{0,full} := \frac{1}{m} \sum_{k=1}^{q} \sum_{i=1}^{m} \boldsymbol{a}_{ki} \boldsymbol{y}_{ki} \boldsymbol{e}_k^\top$$

Its summands are independent sub-exponential r.v.s with maximum sub-exponential norm $\max_k \|\boldsymbol{x}_k^*\| \le \mu \sqrt{r/q} \sigma^*_{\max}$. This is too large and does not allow us to bound $\|\boldsymbol{X}_{0,full} - \boldsymbol{X}^*\|$ under the desired sample complexity; see Appendix A. To resolve this issue, we borrow the truncation idea from earlier work on PR [20], [5] and initialize $\boldsymbol{U}_0$ as the top r left singular vectors of

$$\begin{aligned} \boldsymbol{X}_0 &:= \frac{1}{m} \sum_{k=1}^{q} \sum_{i=1}^{m} \boldsymbol{a}_{ki} \boldsymbol{y}_{ki} \boldsymbol{e}_k^\top \mathbb{1}_{\{\boldsymbol{y}_{ki}^2 \le \alpha\}} \\ &= \frac{1}{m} \sum_{k=1}^{q} \boldsymbol{A}_k^\top \boldsymbol{y}_{k,trunc}(\alpha) \boldsymbol{e}_k^\top \end{aligned} \tag{2}$$

where $\alpha := \tilde{C} \frac{\sum_{ki} (\boldsymbol{y}_{ki})^2}{mq}$ and $\boldsymbol{y}_{k,trunc}(\alpha) := \boldsymbol{y}_k \circ \mathbb{1}(|\boldsymbol{y}_k| \le \sqrt{\alpha})$. We set $\tilde{C}$ in our main result. Observe that we are summing over only those i, k for which $\boldsymbol{y}_{ki}^2$ is not too large (is not much larger than its empirically computed average value). This truncation filters out the too large (outlier-like) measurements and sums over the rest. Theoretically, this converts the summands into sub-Gaussian r.v.s which have lighter tails than the un-truncated ones. This allows us to prove the desired concentration bound. Different from the above setting, in [20], [5], truncation was applied to symmetric positive definite matrices and was used to convert summands that were heavier-tailed than sub-exponential to sub-exponential.

We summarize the complete algorithm in Algorithm 1. This uses sample-splitting which is a commonly used approach in the LR recovery literature [2], [14], [15] as well as in other compressive sensing settings. It helps ensure that the measurement matrices in each iteration for updating $\boldsymbol{U}$ and $\boldsymbol{B}$ are independent of all previous iterates. This allows one to use concentration bounds for sums of independent r.v.s. We provide a detailed discussion in Sec. VI-A.

1) Practical algorithm and setting algorithm parameters: First, when we implement the algorithm, we use Algorithm 1 with using the full set of measurements for all the steps (no sample-splitting). The algorithm has 4 parameters: η, T, $\tilde{C}$ and the rank r. According to the theorem below, we should set $\eta = c/{\sigma^*_{\max}}^2$ with $c < 0.5$. But $\sigma^*_{\max}$ is not known. The initialization matrix $\boldsymbol{X}_0$ provides an approximation to $\boldsymbol{X}^*$ and hence we can set $\eta = c/\|\boldsymbol{X}_0\|^2$. Consider $\tilde{C}$. The theorem requires setting $\tilde{C} = 9\kappa^2\mu^2$, however κ, μ are functions of $\boldsymbol{X}^*$ which is unknown. Using the definition of μ from Assumption 1.1, we can replace $\kappa^2\mu^2$ by an estimate of its lower bound: $q \cdot \max_k \widehat{\|\boldsymbol{x}_k^*\|^2} / \widehat{\|\boldsymbol{X}^*\|_F^2}$ with $\widehat{\|\boldsymbol{x}_k^*\|^2} = (1/m) \sum_i \boldsymbol{y}_{ki}^2$ and $\widehat{\|\boldsymbol{X}^*\|_F^2} = (1/m) \sum_k \sum_i \boldsymbol{y}_{ki}^2$. To set the total number of algorithm iterations T, we can use a large maximum value along with breaking the loop if a stopping criterion is satisfied. A common stopping criterion for GD is to stop when the iterates do not change much. One way to do this is to stop when $\text{SD}(\boldsymbol{U}_t, \boldsymbol{U}_{t-1}) \le 0.01\sqrt{r}$ for last few iterations.

As explained in [13], we can use the following constraints to set the rank. We need our choice of rank, $\hat{r}$, to be sufficiently small compared to $\min(n, q)$ for the algorithm to take advantage of the LR assumption. Moreover, for the LS step for updating $\boldsymbol{b}_k$'s (which are r-length vectors) to work well (for its error to be small), we also need it to also be small compared with m. One approach that is used often is to use the "b% energy threshold" on singular values. Thus, one good

heuristic that respects the above constraints is to compute the "$b\%$ energy threshold" of the first $\min(n,q,m)/10$ singular values, i.e. compute $\hat{r}$ as the smallest value of r for which

$$\sum_{j=1}^{r} \sigma_j(\boldsymbol{X}_0)^2 \geq (b/100) \cdot \sum_{j=1}^{\min(n,q,m)/10} \sigma_j(\boldsymbol{X}_0)^2$$

for a $b \leq 100$. In our MRI experiments in [13], we used $b = 85$. We also realized from the experiments that the algorithm is not very sensitive to this value as long as $\hat{r} \ll \min(n,q,m)$.

2) Federating the algorithm: Suppose that our sketches $\boldsymbol{y}_k$ are geographically distributed across a set of L nodes. Each node ℓ stores a subset, denoted $\mathcal{S}_\ell$, of the $\boldsymbol{y}_k$s with $|\mathcal{S}_\ell| = q_\ell$. These subsets are mutually disjoint so that $\sum_\ell q_\ell = q$. Typically $L \ll q$. Privacy constraints dictate that we cannot share the $\boldsymbol{y}_k$s with the central server; although summaries computed using the $\boldsymbol{y}_k$s can be shared at each algorithm iteration. This will be done as follows. Consider the GDmin steps of Algorithm 1 first. Line 13 (Update $\boldsymbol{b}_k$s, $\boldsymbol{x}_k$s) is done locally at the node that stores the corresponding $\boldsymbol{y}_k$. For line 14 (Gradient w.r.t $\boldsymbol{U}$), the partial sums over $k \in \mathcal{S}_\ell$ are computed at node ℓ and transmitted to the center which adds all the partial sums to obtain $\nabla_{\boldsymbol{U}} f(\boldsymbol{U},\boldsymbol{B})$. Line 15 (GD step) and line 16 (projection via QR) are done at the center. The updated $\boldsymbol{U}$ is then broadcast to all the nodes for use in the next iteration. The per node time complexity of this algorithm is thus $mnrq_\ell$ at each iteration. The center only performs additions and a QR decomposition (an order nr^2 operation) in each iteration. Thus, the time complexity of the federated solution is only $mnr(\max_\ell q_\ell)T$ per node.

The initialization step can be federated by using the Power Method (PM) [24], [25] to compute the top r eigenvectors of $\boldsymbol{X}_0\boldsymbol{X}_0{}^\top$. Any PM guarantee helps ensure that its output is close in subspace distance to the span of the top r eigenvectors of $\boldsymbol{X}_0\boldsymbol{X}_0{}^\top$ after a sufficient number of iterations. The communication complexity of the federated implementation is thus just nr per node per iteration (need to share the partial gradient sums). Observe also that the information shared with the center is not sufficient to recover $\boldsymbol{X}^*$ centrally. It is only sufficient to recover $\mathrm{span}(\boldsymbol{U}^*)$. The recovery of the columns of $\boldsymbol{B}$, $\boldsymbol{b}_k^*$, is entirely done locally at the node where the corresponding $\boldsymbol{y}_k$ is stored, thus ensuring privacy.

B. Main Result

We can prove the following result.

Theorem 2.1. *Consider Algorithm 1. Let m_t denote the number of samples used in iteration t. Set $\tilde{C} = 9\kappa^2\mu^2$, $\eta = c/{\sigma^*_{\max}}^2$ with a $c \leq 0.5$, and $T = C\kappa^2\log(1/\epsilon)$. Assume that Assumption 1.1 holds and that the $\boldsymbol{A}_k$s are i.i.d. and each contains i.i.d. standard Gaussian entries. If*

$$m_0 q \geq C\kappa^6\mu^2(n+q)r^2,$$

and m_t for $t \geq 1$ satisfies

$$m_t q \geq C\kappa^4\mu^2(n+q)r^2\log\kappa \text{ and } m_t \geq C\max(r,\log q,\log n)$$

then, with probability (w.p.) at least $1 - tn^{-10}$, for all $t \geq 0$,

$$\mathrm{SD}(\boldsymbol{U}_t,\boldsymbol{U}^*) \leq \left(1 - \frac{(\eta{\sigma^*_{\max}}^2)0.4}{\kappa^2}\right)^t \delta_0$$

Algorithm 1 The AltGD-Min algorithm. Let $\boldsymbol{M}^\dagger := (\boldsymbol{M}^\top\boldsymbol{M})^{-1}\boldsymbol{M}^\top$.

1: **Input:** $\boldsymbol{y}_k, \boldsymbol{A}_k, k \in [q]$
2: **Parameters:** Multiplier in specifying α for init step, $\tilde{C}$; GD step size, η; Number of iterations, T
3: **Sample-split:** Partition the measurements and measurement matrices into $2T+1$ equal-sized disjoint sets: one set for initialization and $2T$ sets for the iterations. Denote these by $\boldsymbol{y}_k^{(\tau)}, \boldsymbol{A}_k^{(\tau)}, \tau = 0, 1, \ldots 2T$.
4: **Initialization:**
5: Using $\boldsymbol{y}_k \equiv \boldsymbol{y}_k^{(0)}, \boldsymbol{A}_k \equiv \boldsymbol{A}_k^{(0)}$, set
6: $\alpha = \tilde{C}\frac{1}{mq}\sum_{ki}\left|\boldsymbol{y}_{ki}\right|^2$,
7: $\boldsymbol{y}_{k,trunc}(\alpha) := \boldsymbol{y}_k \circ \mathbb{1}\{|\boldsymbol{y}_k| \leq \sqrt{\alpha}\}$
8: $\boldsymbol{X}_0 := (1/m)\sum_{k\in[q]} \boldsymbol{A}_k^\top \boldsymbol{y}_{k,trunc}(\alpha)\boldsymbol{e}_k^\top$
9: Set $\boldsymbol{U}_0 \leftarrow$ top-r-singular-vectors of $\boldsymbol{X}_0$
10: **GDmin iterations:**
11: **for** $t = 1$ **to** T **do**
12: Let $\boldsymbol{U} \leftarrow \boldsymbol{U}_{t-1}$.
13: **Update $\boldsymbol{b}_k, \boldsymbol{x}_k$:** For each $k \in [q]$, set $(\boldsymbol{b}_k)_t \leftarrow (\boldsymbol{A}_k^{(t)}\boldsymbol{U})^\dagger \boldsymbol{y}_k^{(t)}$ and set $(\boldsymbol{x}_k)_t \leftarrow \boldsymbol{U}(\boldsymbol{b}_k)_t$
14: **Gradient w.r.t. $\boldsymbol{U}$:** With $\boldsymbol{y}_k \equiv \boldsymbol{y}_k^{(T+t)}, \boldsymbol{A}_k \equiv \boldsymbol{A}_k^{(T+t)}$, compute $\nabla_{\boldsymbol{U}} f(\boldsymbol{U},\boldsymbol{B}_t) = \sum_k \boldsymbol{A}_k^\top(\boldsymbol{A}_k\boldsymbol{U}(\boldsymbol{b}_k)_t - \boldsymbol{y}_k)(\boldsymbol{b}_k)_t^\top$
15: **GD step:** Set $\hat{\boldsymbol{U}}^+ \leftarrow \boldsymbol{U} - (\eta/m)\nabla_{\boldsymbol{U}} f(\boldsymbol{U},\boldsymbol{B}_t)$.
16: **Projection step:** Compute $\hat{\boldsymbol{U}}^+ \overset{\mathrm{QR}}{=} \boldsymbol{U}^+\boldsymbol{R}^+$.
17: Set $\boldsymbol{U}_t \leftarrow \boldsymbol{U}^+$.
18: **end for**

with $\delta_0 = 0.09/\kappa^2$. Thus, with $T = C\kappa^2\log(1/\epsilon)$ and $\eta = 0.5/{\sigma^*_{\max}}^2$, w.p. at least $1-(T+1)n^{-10}$,

$$\mathrm{SD}(\boldsymbol{U}_T,\boldsymbol{U}^*) \leq \epsilon,\ \|(\boldsymbol{x}_k)_T - \boldsymbol{x}_k^*\| \leq \epsilon\|\boldsymbol{x}_k^*\|, \text{ for all } k \in [q],$$
$$\|\boldsymbol{X}_T - \boldsymbol{X}^*\|_F \leq 1.4\epsilon\|\boldsymbol{X}^*\|$$

Sample complexity The sample complexity (total number of samples needed to achieve ϵ-accurate recovery) is $m_{\mathrm{tot}} = \sum_{\tau=0}^{T} m_\tau \geq m_0 + T\min_{t\geq 1} m_t$. From the above result, this needs to satisfy $m_{tot}q \geq C\kappa^6\mu^2(n+q)r^2\log(1/\epsilon)\log(\kappa)$ and $m_{tot} > C\kappa^2\max(r,\log q,\log n)\log(1/\epsilon)$.

Time complexity Let $m \equiv m_t$. The initialization step needs time mqn for computing $\boldsymbol{X}_0$; and time of order nqr times the number of iterations used in the r-SVD step. Since we only need a δ_0-accurate initial estimate of $\mathrm{span}(\boldsymbol{U}^*)$, with $\delta_0 = c/\kappa^2$, order $\log(\kappa)$ number of iterations suffice for this SVD step. Thus the complexity is $O(nq(m+r)\cdot\log\kappa) = O(mqn\cdot\log\kappa)$ since $m \geq r$. One gradient computation needs time $O(mqnr)$. The QR decomposition needs time of order nr^2. The update of columns of $\boldsymbol{B}$ by LS also needs time $O(mqnr)$ (explained earlier). As we prove above, we need to repeat these steps $T = O(\kappa^2\log(1/\epsilon))$ times. Thus the total time complexity is $O(mqn\log\kappa + \max(mqnr, nr^2, mqnr)\cdot T) = O(\kappa^2 mqnr\log(1/\epsilon)\log\kappa)$.

Communication complexity The communication complexity per node per iteration for a federated implementation is just order nr. Thus, the total is $O(nr\cdot\kappa^2\log(1/\epsilon))$.

Thus, we have the following corollary.

Corollary 2.2 (AltGD-Min)**.** *In the setting of Theorem 2.1, if Assumption 1.1 holds, and if*

$$m_{tot}q \geq C\kappa^6\mu^2(n+q)r^2\log(1/\epsilon)\log(\kappa)$$

and $m_{tot} > C\kappa^2\max(r,\log q,\log n)\log(1/\epsilon)$*, then, w.p. at least* $1-(C\kappa^2\log(1/\epsilon))n^{-10}$*,* $\|\boldsymbol{X}-\boldsymbol{X}^*\|_F \leq 1.4\epsilon\|\boldsymbol{X}^*\|$ *and* $\|\boldsymbol{x}_k-\boldsymbol{x}_k^*\| \leq \epsilon\|\boldsymbol{x}_k^*\|$ *for all* $k\in[q]$*. The time complexity is* $Cmqnr\log(1/\epsilon)\log\kappa$ *and the communication complexity is* $O(nr\cdot\kappa^2\log(1/\epsilon))$*.*

Observe that the above results show that after $T = C\kappa^2\log(1/\epsilon)$ iterations, $\mathrm{SD}(\boldsymbol{U}_T,\boldsymbol{U}^*) \leq \epsilon$, $\|\boldsymbol{x}_k-\boldsymbol{x}_k^*\| \leq \epsilon\|\boldsymbol{x}_k^*\|$, and $\|\boldsymbol{X}_T-\boldsymbol{X}^*\|_F \leq 1.4\epsilon\|\boldsymbol{X}^*\|$. The RHS in the third bound does indeed contain $\|\boldsymbol{X}^*\|$ (the induced 2-norm). This is correct because, $\mathrm{SD}(.,.)$ is a Frobenius norm subspace distance. We explain this in Sec. III-B.

C. Discussion and comparison with the best LRMC results

An algorithm is called linear time if its time complexity is the same order as the time needed to load all input data. In our case, this is $O(mqn)$. Treating κ as a constant, the AltGD-Min complexity is worse than linear-time by a factor of only $r\log(1/\epsilon)$. As can be seen from Table I, the same is also true for the fastest LRMC solution, projGD-X [15]. For LRMC, linear time is $O(mq)$. To our best knowledge, this is the case for the fastest algorithms for all LR problems.

Consider the sample complexity. The degrees of freedom (number of unknowns) of a rank-r $n\times q$ matrix are $(n+q)r$. A sample complexity of $\Omega((n+q)r)$ samples (or, sometimes this times log factors) is called "optimal". Thus, ignoring the log factors, our sample complexity of $m_{tot}q \gtrsim (n+q)r^2$ is sub-optimal only by a factor of r. As can also be seen from Table I, this suboptimality matches that of the best results for LRMC solutions that are not convex relaxation based [15], [16], [17]. The need for exploiting incoherence while obtaining the high probability bounds on the recovery error terms is what introduces the extra factor of r for both LRMC and LRcCS. LRMC has been extensively studied for over a decade and there does not seem to be a way to obtain an (order-) optimal sample complexity guarantee for it except when studying convex relaxation solutions (which are much slower).

In addition, we also need $m \gtrsim \max(r,\log q,\log n)$. This is redundant except for very large q,n. This is needed because, the recovery of each column of $\boldsymbol{B}^*$ is a decoupled r-dimensional LS problem. We analyze this step in Lemma 3.3; notice that the bound on the recovery error of column k holds w.p. at least $1-\exp(r-cm)$. By union bound, it holds for all q columns w.p. at least at least $1-q\exp(r-cm) = 1-\exp(\log q+r-cm)$. This probability is at least $1-n^{-10} = 1-\exp(-10\log n)$ if $m \gtrsim \max(r,\log q,\log n)$.

D. Detailed comparison with existing LRcCS results

There are two existing solutions for LRcCS – AltMin [4], [5], [6] and the convex relaxation (mixed norm minimization) [7]. Mixed norm is defined as $\|\boldsymbol{X}\|_{mixed} := \inf_{\{\boldsymbol{U},\boldsymbol{V}:\boldsymbol{U}\boldsymbol{V}=\boldsymbol{X}\}}\|\boldsymbol{U}\|_F\max_{k\in[q]}\|\boldsymbol{v}_k\|$, where $\boldsymbol{U}$ is $n\times r$ and $\boldsymbol{V} := [\boldsymbol{v}_1,\boldsymbol{v}_2,\dots\boldsymbol{v}_q]$ is an $r\times q$ matrix. In our notation, for the noise-free case ($\sigma=0$), their main result states the following.

Proposition 2.3 (Convex relaxation (mixed norm min) in the $\sigma=0$ (noise-free) setting [7])**.** *Consider a matrix* $\boldsymbol{X}^* \in \{\boldsymbol{X}^*: \max_k\|\boldsymbol{x}_k^*\|^2 \leq \alpha^2, \|\boldsymbol{X}^*\|_{mixed} \leq R \leq \alpha\sqrt{r}\}$*. Then, w.p.* $1-\exp(-c_2nR^2/\alpha^2)$*,* $\frac{\|\boldsymbol{X}-\boldsymbol{X}^*\|_F^2}{\|\boldsymbol{X}^*\|_F^2} \leq c_1\frac{\alpha^2}{\|\boldsymbol{X}^*\|_F^2/q}\sqrt{\frac{(n+q)r\log^6 n}{m_{tot}q}}$ *Under our Assumption 1.1,* $\max_k\|\boldsymbol{x}_k^*\|^2 \leq \mu^2(r/q)\sigma_{\max}^{*\,2} = (\mu^2\kappa^2)(r/q)\sigma_{\min}^{*\,2} \leq (\kappa^2\mu^2)\|\boldsymbol{X}^*\|_F^2/q$*, i.e.* $\frac{\alpha^2}{\|\boldsymbol{X}^*\|_F^2/q} = (\kappa^2\mu^2)$*. Thus, the above result can also be stated as:*

For all matrices $\boldsymbol{X}^*$ *that satisfy Assumption 1.1 and for which* $\|\boldsymbol{X}^*\|_{mixed} \leq \sqrt{r}\cdot\kappa\mu\|\boldsymbol{X}^*\|_F/\sqrt{q}$*, if*

$$m_{tot}q \geq C_1\kappa^4\mu^4(n+q)r\log^6 n\cdot\frac{1}{\epsilon^4},$$

then, w.p. at least $1-\exp(-c_2n)$*,* $\|\boldsymbol{X}-\boldsymbol{X}^*\|_F \leq \epsilon\|\boldsymbol{X}^*\|_F$*. The time complexity is* $Cmqnr\min(\frac{1}{\sqrt{\epsilon}}, n^3r^3)$ *(explained earlier in Sec. I-B).*

Notice that both the sample and the time complexity of the convex solution depend on powers of $1/\sqrt{\epsilon}$: the sample complexity grows as $1/\epsilon^4$ while the time complexity grows as $1/\sqrt{\epsilon}$. However, its sample complexity has an order-optimal dependence on r. For AltGD-Min, both sample and time complexities depend only logarithmically on ϵ only as $\log(1/\epsilon)$. But its sample complexity depends sub-optimally on r, it grows as r^2. In summary, the time complexity of the convex solution is always much worse, its sample complexity is worse when a solution with accuracy level $\epsilon < 1/r^{1/4}$ is needed. A second point to mention is that our result for AltGD-Min provides a column-wise error bound (bounds $\|\boldsymbol{x}_k^*-\boldsymbol{x}_k\|/\boldsymbol{x}_k^*\|$). The convex result only provides a bound on the Frobenius norm of the entire matrix. Thus it is possible that some columns have much larger recovery error than others. This can be problematic in applications such as dynamic MRI where each column corresponds to one signal/image of a time sequence and where the goal is to ensure accurate-enough recovery of all columns. On the other hand, the advantage of the convex guarantee is that it holds w.h.p. for all matrices $\boldsymbol{X}^*$ in the specified set, where as our result only holds w.h.p. for *a* matrix $\boldsymbol{X}^*$ satisfying Assumption 1.1. The reason for these last two points and the reason that we cannot avoid using sample-splitting is the same: the update of $\boldsymbol{B}$ is a column-wise LS problem. We explain the reasoning carefully in Sec. VI-A where we discuss the limitations of our approach. A second advantage of the convex result is that it directly studies the noisy version of the LRcCS problem. This should be possible for AltGD-Min too, we postpone it to future work.

The best result for AltMin is from [6], it states the following.

Proposition 2.4 (AltMin [6])**.** *Under Assumption 1.1, if*

$$m_{tot}q \geq C\kappa^8\mu^2nr^2(r+\log(1/\epsilon)) \text{ and } m_{tot} > \max(r,\log q,\log n),$$

then, w.p. at least $1-(\log(1/\epsilon))n^{-10}$*,* $\|\boldsymbol{X}-\boldsymbol{X}^*\| \leq \epsilon\|\boldsymbol{X}^*\|$ *and* $\|\boldsymbol{x}_k-\boldsymbol{x}_k^*\| \leq \epsilon\|\boldsymbol{x}_k^*\|$ *for all* $k\in[q]$*. The time complexity is* $Cmqnr\log^2(1/\epsilon)$*.*

Treating κ as a numerical constant, compared with the above result for AltMin, the sample complexity of AltGD-Min is

either better by a factor of r or is as good. It is better when $r > \log(1/\epsilon)$. Also, the time complexity is always better by a factor $\log(1/\epsilon)$. As a function of κ, the AltGD-Min sample complexity is better by a factor of κ^2, but its time is worse by a factor of κ^2 compared to that of AltMin. The reason is that its error decays as $(1-c/\kappa^2)^t$. For AltMin the error decays as c^t. Experimentally, GD is usually much faster than AltMin because the constants in its time complexity are also lower.

III. Proving Theorem 2.1

A. Two key results for proving Theorem 2.1 and its proof

Theorem 2.1 is an almost immediate consequence of the following two results.

Theorem 3.1 (Initialization). *Pick a $\delta_0 < 0.1$. If $mq \geq C\kappa^4\mu^2(n+q)r^2/\delta_0^2$, then w.p. at least $1-\exp(-c(n+q))$,*

$$\mathrm{SD}(\boldsymbol{U}^*, \boldsymbol{U}_0) \leq \delta_0.$$

Proof: See Sec. III-E (simpler proof with sample-splitting for α) or Appendix B (proof without sample-splitting). Proof outline is given in Sec. III-D. ∎

Theorem 3.2 (GD Descent). *If, at each iteration t, $mq \geq C\kappa^4\mu^2(n+q)r^2\log\kappa$ and $m > C\max(\log q, \log n)$; if $\mathrm{SD}(\boldsymbol{U}^*, \boldsymbol{U}_0) \leq \delta_0 = c/\kappa^2$ for a $c \leq 0.1/1.1$; and if $\eta \leq 0.5/{\sigma_{\max}^*}^2$, then w.p. at least $1-(t+1)n^{-10}$,*

$$\mathrm{SD}(\boldsymbol{U}^*, \boldsymbol{U}_{t+1}) \leq \delta_{t+1} := \left(1-(\eta{\sigma_{\max}^*}^2)\tfrac{0.4}{\kappa^2}\right)^{t+1}\delta_0.$$

If $\eta = 0.5{\sigma_{\max}^*}^2$, this simplifies to $\mathrm{SD}(\boldsymbol{U}^*, \boldsymbol{U}_{t+1}) \leq (1-0.2/\kappa^2)^{t+1}\delta_0$.

Also, with the above probability,

$$\|(1/m)\nabla_U f(\boldsymbol{U}_t, \boldsymbol{B}_{t+1})\| \leq 1.6\delta_t{\sigma_{\max}^*}^2.$$

with δ_t defined in the $\mathrm{SD}(\boldsymbol{U}^*, \boldsymbol{U}_{t+1})$ bound above.

Since δ_t decays exponentially with t, the same is also true for the gradient norm at iteration t, $\|(1/m)\nabla_U f(\boldsymbol{U}_t, \boldsymbol{B}_{t+1})\|$.

Proof: See Sec. III-C. Proof outline is given in Sec. III-B. ∎

Proof of Theorem 2.1: The SD(.) bound is an immediate consequence of Theorems 3.1 and 3.2. To apply Theorem 3.2, we need $\delta_0 = c/\kappa^2$. By Theorem 3.1, if $mq \geq C\kappa^6\mu^2(n+q)r^2$, then, w.p. at least $1-n^{-10}$, $\mathrm{SD}(\boldsymbol{U}^*, \boldsymbol{U}_0) \leq \delta_0 = c/\kappa^2$. With this, if, at each iteration, $mq \geq C\kappa^4\mu^2(n+q)r^2\log\kappa$ and $m \geq C\max(\log q, \log n)$, then by Theorem 3.2, w.p. at least $1-(t+1)n^{-10}$, the stated bound on $\mathrm{SD}(\boldsymbol{U}^*, \boldsymbol{U}_{t+1})$ holds. By setting $T = C\kappa^2\log(1/\epsilon)$ in this, we can guarantee $\left(1-\frac{c_1}{\kappa^2}\right)^T \leq \epsilon$. This proves the $\mathrm{SD}(\boldsymbol{U}_T, \boldsymbol{U}^*)$ bound. The bounds on $\|\boldsymbol{x}_k - \boldsymbol{x}_k^*\|$ and $\|\boldsymbol{X}-\boldsymbol{X}^*\|_F$ follow by Lemma 3.3 given in Sec. III-C. ∎

B. Proof outline (and novelty) for Theorem 3.2

For proving exponential error decay, we need to show this: at iteration t, if $\mathrm{SD}(\boldsymbol{U}, \boldsymbol{U}^*) \leq \delta_t$ with $\delta_t < \delta_0 = c/\kappa^2$. Then, $\mathrm{SD}(\boldsymbol{U}^+, \boldsymbol{U}^*) \leq c\delta_t$ for a $c < 1$. We explain how to do this next. Suppose that, at iteration t, $\mathrm{SD}(\boldsymbol{U}, \boldsymbol{U}^*) \leq \delta_t < \delta_0 = 0.1/\kappa^2$.

Analyzing the minimization step for updating B (Lemma 3.3). Recall from Algorithm 1 that $\boldsymbol{b}_k = (\boldsymbol{A}_k\boldsymbol{U})^\dagger\boldsymbol{y}_k$, $\boldsymbol{x}_k = \boldsymbol{U}\boldsymbol{b}_k$, and $\boldsymbol{x}_k^* = \boldsymbol{U}^*\boldsymbol{b}_k^*$. Using standard results from [26], we can show that the estimates $\boldsymbol{b}_k$ satisfy $\|\boldsymbol{b}_k - \boldsymbol{U}^\top\boldsymbol{x}_k^*\| \leq 0.4\|(\boldsymbol{I}-\boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{U}^*\boldsymbol{b}_k^*\|$. This then implies that (i) $\boldsymbol{b}_k$'s are incoherent, i.e., $\|\boldsymbol{b}_k\| \leq 1.1\mu\sigma_{\max}^*\sqrt{r/q}$; and (ii) $\|\boldsymbol{x}_k - \boldsymbol{x}_k^*\| \leq 1.4\|(\boldsymbol{I}-\boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{U}^*\boldsymbol{b}_k^*\| \leq 1.4\delta_t\max_k\|\boldsymbol{x}_k^*\|$, i.e., we can get the desired column-wise error bound. Also (iii) $\|\boldsymbol{X}-\boldsymbol{X}^*\|_F \leq 1.4\delta_t\sigma_{\max}^*$ (notice this bound does not contain r). We get this as follows:

$$\begin{aligned}
\|\boldsymbol{X}-\boldsymbol{X}^*\|_F &= \sqrt{\sum_k \|\boldsymbol{x}_k - \boldsymbol{x}_k^*\|^2} \\
&\leq \sqrt{1.4^2\sum_k \|(\boldsymbol{I}-\boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{U}^*\boldsymbol{b}_k^*\|^2} \\
&= 1.4\|(\boldsymbol{I}-\boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{U}^*\boldsymbol{B}^*\|_F \\
&\leq 1.4\|(\boldsymbol{I}-\boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{U}^*\|_F\sigma_{\max}^*
\end{aligned}$$

Similarly, $\|\boldsymbol{B}-\boldsymbol{U}^\top\boldsymbol{X}^*\|_F \leq 0.4\delta_t\sigma_{\max}^*$. (iv) Using Weyl's inequality and $\delta_t < 0.1/\kappa^2$, this then implies that $\sigma_{\max}(\boldsymbol{B}) \leq 1.1\sigma_{\max}^*$ and $\sigma_{\min}(\boldsymbol{B}) \geq 0.9\sigma_{\min}^*$.

Bounding $\mathrm{SD}(\boldsymbol{U}^+, \boldsymbol{U}^*)$ by a novel use of fundamental theorem of calculus (Lemma 3.4). Recall from Algorithm 1 that $\hat{\boldsymbol{U}}^+ = \hat{\boldsymbol{U}} - (\eta/m)\nabla_U f(\boldsymbol{U}, \boldsymbol{B})$ and $\hat{\boldsymbol{U}}^+ \overset{\mathrm{QR}}{=} \boldsymbol{U}^+\boldsymbol{R}^+$. We bound $\mathrm{SD}(\boldsymbol{U}^+, \boldsymbol{U}^*)$ using the fundamental theorem of calculus [18, Chapter XIII, Theorem 4.2],[19], summarized in Theorem 4.2. The use of this result is motivated by its use in [19], and many earlier works, where it is used in a standard way: to bound the Euclidean norm error $\|\boldsymbol{x}-\boldsymbol{x}^*\|$ for standard GD to solve the PR problem for recovering a single vector $\boldsymbol{x}^*$. Thus, at the true solution $\boldsymbol{x} = \boldsymbol{x}^*$, the gradient of the cost function was zero. In our case, there are two differences: (i) we need to bound the subspace distance error, and (ii) our algorithm is not standard GD; in particular, this means that $\nabla_U f(\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U}, \boldsymbol{B}) \neq 0$.

To deal with (i) and (ii), we proceed as follows. We first bound $\|(\boldsymbol{I}-\boldsymbol{U}^*\boldsymbol{U}^{*\top})\hat{\boldsymbol{U}}^+\|_F$. To do this, we apply Theorem 4.2 on vectorized $\nabla_U f(\boldsymbol{U}, \boldsymbol{B})$ with the pivot being vectorized $\nabla_U f(\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U}, \boldsymbol{B})$, and use this in the equation for $\hat{\boldsymbol{U}}^+$. Next, we project both sides of this expression orthogonal to $\boldsymbol{U}^*$ followed by some careful linear algebra. Notice here that $\nabla_U f(\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U}, \boldsymbol{B}) \neq 0$, because $\boldsymbol{B} \neq \boldsymbol{B}^*$. Because of this, we get an extra term, $\mathrm{Term2} := (\boldsymbol{I}-\boldsymbol{U}^*\boldsymbol{U}^{*\top})\nabla_U f(\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U}, \boldsymbol{B})$, in our bound other than the usual term containing the Hessian. We are able to bound it by $\epsilon\delta_t{\sigma_{\max}^*}^2$ for any constant small enough ϵ, by realizing that $\mathbb{E}[\mathrm{Term2}] = 0$ (conditioned on past measurements), and that its summands are *nice-enough* subexponentials. Next, we bound $\mathrm{SD}(\boldsymbol{U}^*, \boldsymbol{U}^+)$ by using

$$\begin{aligned}
\mathrm{SD}(\boldsymbol{U}^*, \boldsymbol{U}^+) &\leq \|(\boldsymbol{I}-\boldsymbol{U}^*\boldsymbol{U}^{*\top})\hat{\boldsymbol{U}}^+\|_F\|(\boldsymbol{R}^+)^{-1}\| \\
&= \frac{\|(\boldsymbol{I}-\boldsymbol{U}^*\boldsymbol{U}^{*\top})\hat{\boldsymbol{U}}^+\|_F}{\sigma_{\min}(\hat{\boldsymbol{U}}^+)}
\end{aligned}$$

and $\sigma_{\min}(\hat{\boldsymbol{U}}^+) = \sigma_{\min}(\boldsymbol{U} - (\eta/m)\nabla_{\boldsymbol{U}} f(\boldsymbol{U}, \boldsymbol{B})) \geq 1 - (\eta/m)\|\nabla_{\boldsymbol{U}} f(\boldsymbol{U}, \boldsymbol{B})\|$.

Bounding the terms in the $\mathrm{SD}(\boldsymbol{U}^*, \boldsymbol{U}^+)$ bound (Lemma 3.5). Consider $\|\nabla_{\boldsymbol{U}} f(\boldsymbol{U}, \boldsymbol{B})\|$. Using Lemma 3.3, it can be shown that, for unit vectors $\boldsymbol{w}, \boldsymbol{z}$, the maximum sub-exponential norm of any summand of $\boldsymbol{w}^\top \nabla_{\boldsymbol{U}} f(\boldsymbol{U}, \boldsymbol{B}) \boldsymbol{z}$ is bounded by $\|\boldsymbol{x}_k - \boldsymbol{x}_k^*\| \cdot \|\boldsymbol{b}_k\| \le 1.1\mu^2 {\sigma^*_{\max}}^2 \delta_t (r/q)$. Observe that we get this (sufficiently small) bound because of the extra $\boldsymbol{b}_k^\top$ term in the summands of $\nabla_{\boldsymbol{U}} f(\boldsymbol{U}, \boldsymbol{B})$ compared to those in $\nabla_{\boldsymbol{X}} \tilde{f}(\boldsymbol{X})$. This, along with using the sub-exponential Bernstein inequality [26] followed by a standard epsilon-net argument, and bounding $\|\mathbb{E}[\nabla_U f]\|$ using $\|\mathbb{E}[\nabla_U f]\| = \|m(\boldsymbol{X} - \boldsymbol{X}^*)\boldsymbol{B}^\top\| \le m\delta_t {\sigma^*_{\max}}^2$ (by Lemma 3.3), helps guarantee that $\|\nabla_U f\| \lesssim 2m\delta_t {\sigma^*_{\max}}^2$ w.h.p. as long as $mq \gtrsim (n+q)r^2$. We bound $\|\mathrm{Term2}\|_F$ using similar ideas and the key fact that $\mathbb{E}[\mathrm{Term2}] = 0$. This is true because of sample-splitting. We upper and lower bound the eigenvalues of the Hessian, Hess, using similar ideas and the following: for a unit vector $\boldsymbol{w}$ of length nr and its rearranged unit Frobenius norm matrix $\boldsymbol{W}$ of size $n \times r$, $\mathbb{E}[\boldsymbol{w}^\top \mathrm{Hess}\ \boldsymbol{w}] = \mathbb{E}[\sum_{ki} ({\boldsymbol{a}_{ki}}^\top \boldsymbol{W} \boldsymbol{b}_k)^2] = m\|\boldsymbol{W}\boldsymbol{B}\|_F^2$. Using the bounds on $\sigma_i(\boldsymbol{B})$ from Lemma 3.3, this can be upper and lower bounded.

C. Lemmas for proving GD descent Theorem 3.2 and its proof

Let $\boldsymbol{U} \equiv \boldsymbol{U}_t$, $\boldsymbol{B} \equiv \boldsymbol{B}_{t+1}$. The proof follows using the following 3 lemmas.

Lemma 3.3 (Error bound on $\boldsymbol{B}$ and its implications)**.** *Let $U \equiv \boldsymbol{U}_t$, $\boldsymbol{B} \equiv \boldsymbol{B}_{t+1}$, and*

$$\boldsymbol{g}_k := \boldsymbol{U}^\top \boldsymbol{x}_k^*.$$

Assume that $\mathrm{SD}(\boldsymbol{U}^*, \boldsymbol{U}_t) \le \delta_t$ with $\delta_t < \delta_0 = c/\kappa^2$ (this bound on δ_t is needed for the second part of this lemma). Then, w.p. $\ge 1 - q\exp(r - cm)$,

1)
$$\|\boldsymbol{g}_k - \boldsymbol{b}_k\| \le 0.4 \|\left(\boldsymbol{I}_n - \boldsymbol{U}\boldsymbol{U}^\top\right)\boldsymbol{U}^* \boldsymbol{b}_k^*\| \tag{3}$$

2) *This in turn implies all of the following.*
   a) $\|\boldsymbol{x}_k - \boldsymbol{x}_k^*\| \le 1.4 \|\left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top\right)\boldsymbol{U}^* \boldsymbol{b}_k^*\|$
   b) $\|\boldsymbol{G} - \boldsymbol{B}\|_F \le 0.4\delta_t \sigma^*_{\max}$ *and* $\|\boldsymbol{X}^* - \boldsymbol{X}\|_F \le \sqrt{1.16}\delta_t \sigma^*_{\max}$,
   c) $\|\boldsymbol{g}_k - \boldsymbol{b}_k\| \le 0.4\delta_t \|\boldsymbol{b}_k^*\|$ *and* $\|\boldsymbol{x}_k - \boldsymbol{x}_k^*\| \le 1.4\delta_t \|\boldsymbol{x}_k^*\|$,
   d) $\|{\boldsymbol{U}^*}^\top \boldsymbol{U}\boldsymbol{b}_k - \boldsymbol{b}_k^*\| \le 2.4\delta_t \|\boldsymbol{b}_k^*\|$,
   e) $\|\boldsymbol{b}_k\| \le 1.1\mu\sigma^*_{\max}\sqrt{r/q}$.
   f) $\sigma_{\min}(\boldsymbol{B}) \ge 0.9\sigma^*_{\min}$ *and* $\sigma_{\max}(\boldsymbol{B}) \le 1.1\sigma^*_{\max}$,

Proof: See Sec. IV-D. ∎

Lemma 3.4. *Let $\boldsymbol{U} \equiv \boldsymbol{U}_t$, $\boldsymbol{B} \equiv \boldsymbol{B}_{t+1}$. Let $\otimes$ denote the Kronecker product. We have*

$$\mathrm{SD}(\boldsymbol{U}_{t+1}, \boldsymbol{U}^*) \le \frac{\|\boldsymbol{I}_{nr} - (\eta/m)\mathrm{Hess}\| \cdot \mathrm{SD}(\boldsymbol{U}^*, \boldsymbol{U}) + (\eta/m)\|\mathrm{Term2}\|_F}{1 - (\eta/m)\|\mathrm{GradU}\|},$$

where,

$$\mathrm{GradU} := \nabla_{\boldsymbol{U}} f(\boldsymbol{U}, \boldsymbol{B}) = \sum_{ki} (\boldsymbol{y}_{ki} - {\boldsymbol{a}_{ki}}^\top \boldsymbol{U}\boldsymbol{b}_k)\boldsymbol{a}_{ki}{\boldsymbol{b}_k}^\top$$

$$\mathrm{Term2} := (\boldsymbol{I} - \boldsymbol{U}^*{\boldsymbol{U}^*}^\top)\nabla_{\boldsymbol{U}} f((\boldsymbol{U}^*{\boldsymbol{U}^*}^\top \boldsymbol{U}), \boldsymbol{B})$$
$$= (\boldsymbol{I} - \boldsymbol{U}^*{\boldsymbol{U}^*}^\top) \sum_{ki} (\boldsymbol{y}_{ki} - {\boldsymbol{a}_{ki}}^\top \boldsymbol{U}^*{\boldsymbol{U}^*}^\top \boldsymbol{U}\boldsymbol{b}_k)\boldsymbol{a}_{ki}{\boldsymbol{b}_k}^\top$$

$$\mathrm{Hess} := \sum_{ki} (\boldsymbol{a}_{ki} \otimes \boldsymbol{b}_k)(\boldsymbol{a}_{ki} \otimes \boldsymbol{b}_k)^\top$$

Proof: See Sec. IV-B ∎

Lemma 3.5. *Assume $\mathrm{SD}(\boldsymbol{U}^*, \boldsymbol{U}) \le \delta_t < \delta_0 = c/\kappa^2$. Then,*

1) *w.p. at least $1 - \exp((n+r) - cmq\epsilon_1^2/r\mu^2) - \exp(\log q + r - cm)$,*
$$\|\mathrm{GradU}\| \le 1.5(1.1 + \epsilon_1)m\delta_t {\sigma^*_{\max}}^2;$$

2) *w.p. at least $1 - \exp(nr - cmq\epsilon_2^2/r\mu^2) - \exp(\log q + r - cm)$,*
$$\|\mathrm{Term2}\|_F \le 1.1m\epsilon_2\delta_t {\sigma^*_{\max}}^2;$$

3) *w.p. at least $1 - \exp(nr\log\kappa - cmq\epsilon_3^2/r\kappa^4\mu^2) - \exp(\log q + r - cm)$,*
$$m(0.65 - 1.2\epsilon_3){\sigma^*_{\min}}^2 \le \lambda_{\min}(\mathrm{Hess}) \le \lambda_{\max}(\mathrm{Hess}) \le m(1.1 + \epsilon_3){\sigma^*_{\max}}^2.$$

Proof: See Sec. IV-C. ∎

Proof of Theorem 3.2: The proof follows by induction. Base case for $t = 0$ is true by assumption. Induction assumption: Assume that, w.p. at least $1 - tn^{-10}$, $\mathrm{SD}(\boldsymbol{U}^*, \boldsymbol{U}_t) \le \delta_t$ with $\delta_t \le \delta_0 = c_0/\kappa^2$.

Set $\epsilon_1 = 0.1$, $\epsilon_3 = 0.01$, $\epsilon_2 = 0.01/1.1\kappa^2$ and, $c_0 = 0.1/1.5(1.1 + 0.1)$.

The upper bound on $\lambda_{\max}(\mathrm{Hess})$ and using $\eta \le 0.5/{\sigma^*_{\max}}^2$ implies that $\lambda_{\min}(\boldsymbol{I}_{nr} - (\eta/m)\mathrm{Hess}) = 1 - (\eta/m)\lambda_{\max}(\mathrm{Hess}) \ge 1 - \frac{0.5(1.1+0.01)m{\sigma^*_{\max}}^2}{m{\sigma^*_{\max}}^2} > 1 - 0.555 > 0$ i.e. $\boldsymbol{I}_{nr} - (\eta/m)\mathrm{Hess}$ is positive definite. Thus, $\|\boldsymbol{I}_{nr} - (\eta/m)\mathrm{Hess}\| = \lambda_{\max}(\boldsymbol{I}_{nr} - (\eta/m)\mathrm{Hess}) = 1 - (\eta/m)\lambda_{\min}(\mathrm{Hess}) \le 1 - (\eta/m)m(0.65 - 1.2\epsilon_3){\sigma^*_{\min}}^2 \le 1 - (\eta{\sigma^*_{\max}}^2)0.63/\kappa^2$.

By Lemma 3.4, Lemma 3.5, and the above, w.p. at least $1 - tn^{-10} - \exp((n+q) - cmq/r\mu^2) - \exp(nr - cmq/r\kappa^4\mu^2) - \exp(nr\log\kappa - cmq/r\kappa^4\mu^2) - \exp(\log q + r - cm)$,

$$\mathrm{SD}(\boldsymbol{U}^*, \boldsymbol{U}_{t+1}) \le \frac{(1 - (\eta{\sigma^*_{\max}}^2)0.63/\kappa^2) \cdot \delta_t + (\eta/m)1.1m\epsilon_2{\sigma^*_{\max}}^2\delta_t}{1 - (\eta/m)1.5(1.1 + \epsilon_1)m\delta_t{\sigma^*_{\max}}^2}$$
$$\le \left(\frac{1 - (\eta{\sigma^*_{\max}}^2)0.63/\kappa^2 + (\eta{\sigma^*_{\max}}^2)0.01/\kappa^2}{1 - (\eta{\sigma^*_{\max}}^2)0.1/\kappa^2}\right)\delta_t$$
$$\le \left(1 - (\eta{\sigma^*_{\max}}^2)\frac{0.42}{\kappa^2}\right)\delta_t$$

The second inequality substituted the values of ϵ_j's and used $\delta_t < \delta_0 = 0.1/(1.5(1.1 + 0.1)\kappa^2)$ for its denominator term. The third inequality used $(1 - (\eta{\sigma^*_{\max}}^2)0.1/\kappa^2)^{-1} \le (1 + (\eta{\sigma^*_{\max}}^2)0.2/\kappa^2)$ (for $0 < x < 1$, $1/(1-x) \le 1 + 2x$).

By plugging in the epsilon values in the probability, the above holds w.p. $\ge 1 - tn^{-10} - 0.2\exp((n+q) - cmq/r\mu^2) -$

$0.2\exp(nr-cmq/r\mu^2\kappa^4)-0.2\exp(nr\log\kappa-cmq/r\mu^2\kappa^4)-\exp(\log q+r-cm)$. If $mq \geq C\kappa^4(n+q)r^2\log\kappa$ and $m \geq C\max(r,\log q,\log n)$ for a C large enough, then, this probability is $\geq 1-tn^{-10}-0.2\exp(-c(n+q))-0.4\exp(-cnr)-n^{-10} > 1-(t+1)n^{-10}$. ∎

D. Proof outline (and novelty) for Initialization Theorem 3.1

Recall that we compute $\boldsymbol{U}_0$ as the top r left singular vectors of $\boldsymbol{X}_0$ defined in (2) and that this is a truncated version of $\boldsymbol{X}_{0,full}$. As noted there, we cannot use $\boldsymbol{X}_{0,full}$ because its summands are not *nice-enough sub-exponentials*. Truncation converts the summands into sub-Gaussian r.v.s. For these, we can use the sub-Gaussian Hoeffding inequality [26, Chap 2] which needs a small enough bound on only the squared sum of the sub-Gaussian norms of the mq summands, and not on their maximum value (as needed by the sub-exponential Bernstein inequality). This is an easier requirement that gets satisfied for our problem. Of course, truncation also means that the summands of $\boldsymbol{X}_0$ are not mutually independent (each summand depends on the truncation threshold α which is computed using all measurements $\boldsymbol{y}_{ki}$) and that $\mathbb{E}[\boldsymbol{X}_0] \neq \boldsymbol{X}^*$. There are two ways to resolve this issue. The first and simpler approach, but one that assumes more sample-splitting is given below in Sec III-E. This assumes that α is a computed using a different independent set of measurements than those used to define the rest of $\boldsymbol{X}_0$. With this, $\mathbb{E}[\boldsymbol{X}_0|\alpha] = \boldsymbol{X}^*\boldsymbol{D}(\alpha)$, where $\boldsymbol{D}$ is a diagonal matrix defined below in Lemma 3.6 and the summands are independent conditioned on α. Thus, we can apply Wedin's $\sin\Theta$ theorem [27], [28] (given in Proposition 4.1) on $\boldsymbol{X}_0$ and $\mathbb{E}[\boldsymbol{X}_0|\alpha]$ to bound $\mathrm{SD}(\boldsymbol{U}_0,\boldsymbol{U}^*)$, followed by subGaussian Hoeffding and a standard epsilon-net argument, to bound the terms in this bound.

To avoid sample-splitting for α, we need to significantly modify the sandwiching arguments from [20], [5] for our setting. This is done in Appendix B. In the previous works, sandwiching was used for a symmetric positive definite (p.d.) matrix. Here we need such an argument for a non-symmetric matrix. Briefly, we do this as follows. We define a matrix $\boldsymbol{X}_+$ that is such that the span of top r left singular vectors of its expected value equals that of $\boldsymbol{U}^*$ and that can be shown to be close to $\boldsymbol{X}_0$. $\boldsymbol{X}_+$ is $\boldsymbol{X}_0$ with α replaced by $\tilde{C}(1+\epsilon)\|\boldsymbol{X}^*\|_F^2/q$. We bound $\|\boldsymbol{X}_0-\mathbb{E}[\boldsymbol{X}_+]\|$ by bounding $\|\boldsymbol{X}_+-\boldsymbol{X}_0\|$ and $\|\boldsymbol{X}_+-\mathbb{E}[\boldsymbol{X}_+]\|$. Bounding the latter is simple. Bounding $\|\boldsymbol{X}_+-\boldsymbol{X}_0\|$ requires bounding $\boldsymbol{w}^\top(\boldsymbol{X}_+-\boldsymbol{X}_0)\boldsymbol{z}$ for unit vectors $\boldsymbol{w},\boldsymbol{z}$ and this is not straightforward because its summands are not mutually independent. To deal with this, we first bound each summand by its absolute value, and then bound the indicator function term to get a new one that is non-random so that the summands of this new term are mutually independent. But, its summands are no longer zero mean (because of taking the absolute values), and hence more work is needed to get the desired small enough bound on the expected value of this term.

E. Simpler proof of Theorem 3.1 that assumes independent measurements used for computing α

For the simpler proof given here, assume that we use a different independent set of measurements for computing α than those used for the rest of $\boldsymbol{X}_0$, i.e., let

$$\alpha = \tilde{C}\frac{\sum_{ki}(\boldsymbol{y}_{ki}^{nrmX})^2}{mq}$$

with $\boldsymbol{y}_{ki}^{nrmX}$ independent of $\{\boldsymbol{A}_k^{(0)},\boldsymbol{y}_k^{(0)}\}$. With this change, it is possible to compute $\mathbb{E}[\boldsymbol{X}_0|\alpha]$ easily. But, it does not affect the sample complexity order and so it does not change our theorem statement. The proof follows by combining the two lemmas and facts given next.

Lemma 3.6. *Conditioned on α, we have the following conclusions.*

1) *Let ζ be a scalar standard Gaussian r.v.. Define*

$$\beta_k(\alpha) := \mathbb{E}[\zeta^2\mathbb{1}_{\{\|\boldsymbol{x}_k^*\|^2\zeta^2\leq\alpha\}}].$$

Then,

$$\mathbb{E}[\boldsymbol{X}_0|\alpha] = \boldsymbol{X}^*\boldsymbol{D}(\alpha),$$
$$\text{where } \boldsymbol{D}(\alpha) := diagonal(\beta_k(\alpha), k\in[q]) \quad (4)$$

i.e. $\boldsymbol{D}(\alpha)$ is a diagonal matrix of size $q\times q$ with diagonal entries β_k defined above.

2) *Let $\mathbb{E}[\boldsymbol{X}_0|\alpha] = \boldsymbol{X}^*\boldsymbol{D}(\alpha) \overset{\text{SVD}}{=} \boldsymbol{U}^*\check{\boldsymbol{\Sigma}}^*\check{\boldsymbol{V}}$ be its r-SVD. Then,*

$$\mathrm{SD}(\boldsymbol{U}_0,\boldsymbol{U}^*) \leq \frac{\sqrt{2}\max\left(\|(\boldsymbol{X}_0-\mathbb{E}[\boldsymbol{X}_0|\alpha])^\top\boldsymbol{U}^*\|_F, \|(\boldsymbol{X}_0-\mathbb{E}[\boldsymbol{X}_0|\alpha])\check{\boldsymbol{V}}^\top\|_F\right)}{\sigma^*_{\min}\min_k\beta_k(\alpha)-\|\boldsymbol{X}_0-\mathbb{E}[\boldsymbol{X}_0|\alpha]\|} \quad (5)$$

as long as the denominator is non-negative.

Proof: See Sec. IV-F ∎

Define the set $\mathcal{E}$ as follows

$$\mathcal{E} := \left\{\tilde{C}(1-\epsilon_1)\frac{\|\boldsymbol{X}^*\|_F^2}{q} \leq \alpha \leq \tilde{C}(1+\epsilon_1)\frac{\|\boldsymbol{X}^*\|_F^2}{q}\right\}. \quad (6)$$

The following fact is an immediate consequence of sub-exponential Bernstein inequality for bounding $|\alpha-\|\boldsymbol{X}^*\|_F^2/q|$.

Fact 3.7. $\Pr(\alpha\in\mathcal{E}) \geq 1-\exp(-\tilde{c}mq\epsilon_1^2) := 1-p_\alpha$. *Here $\tilde{c}=c/\tilde{C}=c/\kappa^2\mu^2$.*

The next lemma bounds the terms of Lemma 3.6.

Lemma 3.8. *Fix $0<\epsilon_1<1$. Then,*

1) *w.p. at least $1-\exp\left[(n+q)-c\epsilon_1^2mq/\mu^2\kappa^2\right]$, conditioned on α, for an $\alpha\in\mathcal{E}$,*

$$\|\boldsymbol{X}_0-\mathbb{E}[\boldsymbol{X}_0|\alpha]\| \leq 1.1\epsilon_1\|\boldsymbol{X}^*\|_F$$

2) *w.p. at least $1-\exp\left[qr-c\epsilon_1^2mq/\mu^2\kappa^2\right]$, conditioned on α, for an $\alpha\in\mathcal{E}$,*

$$\|\left(\boldsymbol{X}_0-\mathbb{E}[\boldsymbol{X}_0|\alpha]\right)^\top\boldsymbol{U}^*\|_F \leq 1.1\epsilon_1\|\boldsymbol{X}^*\|_F$$

3) *w.p. at least $1-\exp\left[nr-c\epsilon_1^2mq/\mu^2\kappa^2\right]$, conditioned on α, for an $\alpha\in\mathcal{E}$,*

$$\|\left(\boldsymbol{X}_0-\mathbb{E}[\boldsymbol{X}_0|\alpha]\right)\check{\boldsymbol{V}}^\top\|_F \leq 1.1\epsilon_1\|\boldsymbol{X}^*\|_F.$$

Proof: See Sec. IV-G ∎

We also need to the following fact.

Fact 3.9. *For any* $\epsilon_1 \leq 0.1$, $\min_k \mathbb{E}\left[\zeta^2 \mathbb{1}_{\left\{|\zeta| \leq \tilde{C}\frac{\sqrt{1-\epsilon_1}\|\boldsymbol{X}^*\|_F}{\sqrt{q}\|\boldsymbol{w}_k^*\|}\right\}}\right] \geq 0.92$.

Proof of Theorem 3.1: Set $\epsilon_1 = 0.4\delta_0/\sqrt{r}\kappa$. Define $p_0 = 2\exp((n+q) - cmq\delta_0^2/r\kappa^2) + 2\exp(nr - cmq\delta_0^2/r\kappa^2) + 2\exp(qr - cmq\delta_0^2/r\kappa^2)$. Recall that $\Pr(\alpha \in \mathcal{E}) \geq 1 - p_\alpha$ with $p_\alpha = \exp(-\tilde{c}mq\epsilon_1^2) = \exp(-cmq\delta_0^2/r\mu^2\kappa^2)$.

Using Lemma 3.8, conditioned on α, for an $\alpha \in \mathcal{E}$,

- w.p. at least $1 - p_0$, $\|\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha]\| \leq 1.1\epsilon_1\|\boldsymbol{X}^*\|_F = 0.44\delta_0\sigma^*_{\min}$, and $\max\left(\|(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha])^\top \boldsymbol{U}^*\|_F, \|(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha])\boldsymbol{V}^\top\|_F\right) \leq 0.44\delta_0\sigma^*_{\min}$
- $\min_k \beta_k(\alpha) \geq \min_k \mathbb{E}\left[\zeta^2 \mathbb{1}_{\{|\zeta| \leq \tilde{C}\frac{\sqrt{1-\epsilon_1}\|\boldsymbol{X}^*\|_F}{\sqrt{q}\|\boldsymbol{w}_k^*\|}\}}\right] \geq 0.9$
  The first inequality is an immediate consequence of $\alpha \in \mathcal{E}$ and the second follows by Fact 3.9.

Plugging the above bounds into (5) of Lemma 3.6, conditioned on α, for any $\alpha \in \mathcal{E}$, w.p. at least $1-p_0$, $\mathrm{SD}(\boldsymbol{U}_0, \boldsymbol{U}^*) \leq \frac{0.44\delta_0}{0.9 - 0.44\delta_0} < \delta_0$ since $\delta_0 < 0.1$. In other words,

$$\Pr\left(\mathrm{SD}(\boldsymbol{U}_0, \boldsymbol{U}^*) \geq \delta_0 | \alpha\right) \leq p_0 \text{ for any } \alpha \in \mathcal{E}. \tag{7}$$

Since (i) $\Pr(\mathrm{SD}(\boldsymbol{U}_0, \boldsymbol{U}^*) \geq \delta_0) \leq \Pr(\mathrm{SD}(\boldsymbol{U}_0, \boldsymbol{U}^*) \geq \delta_0 \text{ and } \alpha \in \mathcal{E}) + \Pr(\alpha \notin \mathcal{E})$, and (ii) $\Pr(\mathrm{SD}(\boldsymbol{U}_0, \boldsymbol{U}^*) \geq \delta_0 \text{ and } \alpha \in \mathcal{E}) \leq \Pr(\alpha \in \mathcal{E}) \max_{\alpha \in \mathcal{E}} \Pr(\mathrm{SD}(\boldsymbol{U}_0, \boldsymbol{U}^*) \leq \delta_0 | \alpha)$, thus, using Fact 3.7 and (7), we can conclude that

$$\Pr\left(\mathrm{SD}(\boldsymbol{U}_0, \boldsymbol{U}^*) \geq \delta_0\right) \leq p_0(1 - p_\alpha) + p_\alpha \leq p_0 + p_\alpha$$

Thus, for a $\delta_0 < 0.1$, $\mathrm{SD}(\boldsymbol{U}_0, \boldsymbol{U}^*) < \delta_0$ w.p. at least $1 - p_0 - p_\alpha = 1 - 2\exp((n+q) - cmq\delta_0^2/r\kappa^2) - 2\exp(nr - cmq\delta_0^2/r\kappa^2) - 2\exp(qr - cmq\delta_0^2/r\kappa^2) - \exp(-cmq\delta_0^2/r\mu^2\kappa^4)$. This is $\geq 1 - 5\exp(-c(n+q))$ if $mq > C\kappa^2\mu^2(n+q)r^2/\delta_0^2$. This finishes our proof. ∎

IV. Proofs of all the lemmas

A. Basic tools used

Our proofs use the following results and definitions:

Theorem 4.1 (Wedin $\sin\Theta$ theorem for Frobenius norm subspace distance [27], [28][Theorem 2.3.1]**.** *] For two $n_1 \times n_2$ matrices $\boldsymbol{M}^*$, $\boldsymbol{M}$, let $\boldsymbol{U}^*, \boldsymbol{U}$ denote the matrices containing their top r singular vectors and let $\boldsymbol{V}^{*\top}, \boldsymbol{V}^\top$ be the matrices of their right singular vectors (recall from problem definition that we defined SVD with the right matrix transposed). Let $\sigma_r^*, \sigma_{r+1}^*$ denote the r-th and $(r+1)$-th singular values of $\boldsymbol{M}^*$. If $\|\boldsymbol{M} - \boldsymbol{M}^*\| \leq \sigma_r^* - \sigma_{r+1}^*$, then*

$$\mathrm{SD}(\boldsymbol{U}, \boldsymbol{U}^*) \leq \frac{\sqrt{2}\max(\|(\boldsymbol{M} - \boldsymbol{M}^*)^\top \boldsymbol{U}^*\|_F, \|(\boldsymbol{M} - \boldsymbol{M}^*)^\top \boldsymbol{V}^{*\top}\|_F)}{\sigma_r^* - \sigma_{r+1}^* - \|\boldsymbol{M} - \boldsymbol{M}^*\|}$$

Theorem 4.2 (Fundamental theorem of calculus [18][Chapter XIII, Theorem 4.2])**.** *, [19]] For two vectors $\boldsymbol{z}_0, \boldsymbol{z}^* \in \Re^d$, and a differentiable vector function $g(\boldsymbol{z}) \in \Re^{d_2}$,*

$$g(\boldsymbol{z}_0) - g(\boldsymbol{z}^*) = \left(\int_{\tau=0}^1 \nabla g(\boldsymbol{z}(\tau)) d\tau\right)(\boldsymbol{z}_0 - \boldsymbol{z}^*),$$

where

$$\boldsymbol{z}(\tau) = \boldsymbol{z}^* + \tau(\boldsymbol{z}_0 - \boldsymbol{z}^*).$$

Observe that $\nabla_{\boldsymbol{z}} g(\boldsymbol{z})$ is a $d_2 \times d$ matrix.

Definition 4.3. *For any $n \times r$ matrix $\boldsymbol{Z}$, let $\boldsymbol{Z}_{vec}$ denote the nr length vector formed by arranging all r columns of $\boldsymbol{Z}$ one below the other. Thus, for n-length and r-length vectors $\boldsymbol{a}$ and $\boldsymbol{b}$,*

- $(\boldsymbol{a}\boldsymbol{b}^\top)_{vec} = \boldsymbol{a} \otimes \boldsymbol{b}$ *with $\otimes$ being the Kronecker product;*
- $\boldsymbol{a}^\top \boldsymbol{U} \boldsymbol{b} = \mathrm{trace}(\boldsymbol{a}^\top \boldsymbol{U} \boldsymbol{b}) = \mathrm{trace}(\boldsymbol{b}\boldsymbol{a}^\top \boldsymbol{U}) = \langle (\boldsymbol{a}\boldsymbol{b}^\top), \boldsymbol{U} \rangle = \langle \boldsymbol{a} \otimes \boldsymbol{b}, \boldsymbol{U}_{vec} \rangle$;

$f(\boldsymbol{U}_{vec}, \boldsymbol{B}) = \sum_{ki}((\boldsymbol{a}_{ki} \otimes \boldsymbol{b}_k)^\top \boldsymbol{U}_{vec} - \boldsymbol{y}_{ki})^2$ and

$$(\nabla_{\boldsymbol{U}} f(\boldsymbol{U}, \boldsymbol{B}))_{vec} = \nabla_{\boldsymbol{U}_{vec}} f(\boldsymbol{U}_{vec}, \boldsymbol{B}) \tag{8}$$

Definition 4.4. *At various places, $\nabla f(\boldsymbol{U}, \boldsymbol{B})$ is short for $\nabla_{\boldsymbol{U}} f(\boldsymbol{U}, \boldsymbol{B}) = \sum_{ki} \boldsymbol{a}_{ki}\boldsymbol{b}_k{}^\top(\boldsymbol{a}_{ki}{}^\top \boldsymbol{U}\boldsymbol{b}_k - \boldsymbol{y}_{ki})$ and similarly $\nabla f(\boldsymbol{U}_{vec}, \boldsymbol{B})$ is short for $\nabla_{\boldsymbol{U}_{vec}} f(\boldsymbol{U}_{vec}, \boldsymbol{B}) = \sum_{ki}(\boldsymbol{a}_{ki} \otimes \boldsymbol{b}_k)((\boldsymbol{a}_{ki} \otimes \boldsymbol{b}_k)^\top \boldsymbol{U}_{vec} - \boldsymbol{y}_{ki})$.*

Definition 4.5. *For any vector $\boldsymbol{w}$, we use $\boldsymbol{w}(k)$ to denote its k-th entry.*

Definition 4.6. *Everywhere we use $\mathcal{S}_{nr}$ to denote both the set of matrices $\{\boldsymbol{W} \in \Re^{n \times r} : \|\boldsymbol{W}\|_F = 1\}$ and the set of these matrices vectorized $\{\boldsymbol{w} \in \Re^{nr} : \|\boldsymbol{w}\| = 1\}$. We also switch between the two sometimes. In the entire writing below, $\boldsymbol{w} = \boldsymbol{W}_{vec}$.*

All the high probability bounds for initialization use sub-Gaussian Hoeffding inequality, while those for GD lemmas use the sub-exponential Bernstein inequality, both are from [26]. In addition, these lemmas also use the following results to "epsilon-net" extend a bound holding for a fixed unit norm $\boldsymbol{W}$ (or $\boldsymbol{w}$) to all unit norm $\boldsymbol{W}$s (or $\boldsymbol{w}$s)

Proposition 4.7 (Epsilon-netting for bounding $\max_{\boldsymbol{w} \in \mathcal{S}_n, \boldsymbol{z} \in \mathcal{S}_r} |\boldsymbol{w}^\top \boldsymbol{M} \boldsymbol{z}|$)**.** *For an $n \times r$ matrix $\boldsymbol{M}$ and fixed vectors $\boldsymbol{w}, \boldsymbol{z}$ with, $\boldsymbol{w} \in \mathcal{S}_n$ and $\boldsymbol{z} \in \mathcal{S}_r$, suppose that $|\boldsymbol{w}^\top \boldsymbol{M} \boldsymbol{z}| \leq b_0$ w.p. at least $1 - p_0$. Consider an ϵ_{net} net covering $\mathcal{S}_n$ and $\mathcal{S}_r$, $\bar{\mathcal{S}}_n$, $\bar{\mathcal{S}}_r$ Then w.p. at least $1 - (1 + 2/\epsilon_{net})^{n+r} p_0$,*

- $\max_{\boldsymbol{w} \in \bar{\mathcal{S}}_n, \boldsymbol{z} \in \bar{\mathcal{S}}_r} |\boldsymbol{w}^\top \boldsymbol{M} \boldsymbol{z}| \leq b_0$ *and*
- $\max_{\boldsymbol{w} \in \mathcal{S}_n, \boldsymbol{z} \in \mathcal{S}_r} |\boldsymbol{w}^\top \boldsymbol{M} \boldsymbol{z}| \leq \frac{1}{1 - 2\epsilon_{net} - \epsilon_{net}^2} b_0$.

Using $\epsilon_{net} = 1/8$, this implies the following simpler conclusion:
W.p. at least $1 - 17^{n+r} p_0 = 1 - \exp((\log 17)(n+r)) \cdot p_0$, $\max_{\boldsymbol{w} \in \mathcal{S}_n, \boldsymbol{z} \in \mathcal{S}_r} |\boldsymbol{w}^\top \boldsymbol{M} \boldsymbol{z}| \leq 1.4 b_0$.

Proof: The proof follows that of Lemma 4.4.1 of [26] ∎

Proposition 4.8 (Epsilon-netting for bounding $\max_{\boldsymbol{W} \in \mathcal{S}_{nr}} \langle \boldsymbol{M}, \boldsymbol{W} \rangle$)**.** *For an $n \times r$ matrix $\boldsymbol{M}$ and a fixed $n \times r$ matrix $\boldsymbol{W} \in \mathcal{S}_{nr}$ (unit Frobenius norm matrix), suppose that $\langle \boldsymbol{M}, \boldsymbol{W} \rangle \leq b_0$ w.p. at least $1 - p_0$. Consider an ϵ_{net} net covering $\mathcal{S}_{nr}$, $\bar{\mathcal{S}}_{nr}$. Then w.p. at least $1 - (1 + 2/\epsilon_{net})^{nr} p_0$,*

- $\max_{\boldsymbol{W} \in \bar{\mathcal{S}}_{nr}} \langle \boldsymbol{M}, \boldsymbol{W} \rangle \leq b_0$ *and*
- $\max_{\boldsymbol{W} \in \mathcal{S}_{nr}} \langle \boldsymbol{M}, \boldsymbol{W} \rangle \leq \frac{1}{1 - \epsilon_{net}} b_0$.

Using $\epsilon_{net} = 1/8$, this implies the following simpler conclusion:

w.p. at least $1-17^{nr}p_0 = 1-\exp((\log 17)(nr))\cdot p_0$, $\max_{\boldsymbol{W}\in\mathcal{S}_{nr}}\langle \boldsymbol{M},\boldsymbol{W}\rangle \le 1.2b_0$.

Proof: The proof follows exactly as that of Exercise 4.4.3 of [26] ∎

Proposition 4.9 (Epsilon-netting for upper and lower bounding $\sum_{ki}\langle \boldsymbol{M}_{ki},\boldsymbol{W}\rangle^2$ over all $\boldsymbol{W}\in\mathcal{S}_{nr}$)**.** *For an $n\times r$ matrices $\boldsymbol{M}_{ki}$ and a fixed $\boldsymbol{W}\in\mathcal{S}_{nr}$, suppose that, w.p. at least $1-p_0$,*

$$b_1 \le \sum_{ki}\langle \boldsymbol{M}_{ki},\boldsymbol{W}\rangle^2 \le b_2$$

Consider an ϵ_{net} net covering $\mathcal{S}_{nr}$, $\bar{\mathcal{S}}_{nr}$. Then, w.p. at least $1-(1+2/\epsilon_{net})^{nr}p_0$,

$$\max_{\boldsymbol{W}\in\mathcal{S}_{nr}}\sum_{ki}\langle \boldsymbol{M}_{ki},\boldsymbol{W}\rangle^2 \le \frac{1}{1-\epsilon_{net}^2-2\epsilon_{net}}b_2$$

and

$$\min_{\boldsymbol{W}\in\mathcal{S}_{nr}}\sum_{ki}\langle \boldsymbol{M}_{ki},\boldsymbol{W}\rangle^2 \ge b_1 - 2\epsilon_{net}\cdot\frac{1}{1-\epsilon_{net}^2-2\epsilon_{net}}b_2$$

Picking $\epsilon_{net} = b_1/(8b_2)$ guarantees that the above lower bound is non-negative. In particular, it implies the following: w.p. at least $1-(24b_2/b_1)^{nr}p_0 = 1-\exp(Cnr\log(b_2/b_1))\cdot p_0$, $0.8b_1 \le \min_{\boldsymbol{W}\in\mathcal{S}_{nr}}\sum_{ki}\langle \boldsymbol{M}_{ki},\boldsymbol{W}\rangle^2 \le \max_{\boldsymbol{W}\in\mathcal{S}_{nr}}\sum_{ki}\langle \boldsymbol{M}_{ki},\boldsymbol{W}\rangle^2 \le 1.4b_2$

Proof: By union bound, for all $\bar{\boldsymbol{W}}\in\bar{\mathcal{S}}_{nr}$, $b_1 \le \sum_{ki}\langle \boldsymbol{M}_{ki},\bar{\boldsymbol{W}}\rangle^2 \le b_2$ holds w.p. at least $1-(1+2/\epsilon_{net})^{nr}p_0$.

Proof for the upper bound: Let $\gamma^* = \max_{\boldsymbol{W}\in\mathcal{S}_{nr}}\sum_{ki}\langle \boldsymbol{M}_{ki},\boldsymbol{W}\rangle^2$. Writing $\boldsymbol{W} = \bar{\boldsymbol{W}} + (\boldsymbol{W}-\bar{\boldsymbol{W}})$ where $\bar{\boldsymbol{W}}$ is the closest point to $\boldsymbol{W}$ on $\bar{\mathcal{S}}_{nr}$, we have $\sum_{ki}\langle \boldsymbol{M}_{ki},\boldsymbol{W}\rangle^2 = \sum_{ki}\langle \boldsymbol{M}_{ki},\bar{\boldsymbol{W}}\rangle^2 + \sum_{ki}\langle \boldsymbol{M}_{ki},(\boldsymbol{W}-\bar{\boldsymbol{W}})\rangle^2 + 2\sum_{ki}\langle \boldsymbol{M}_{ki},\bar{\boldsymbol{W}}\rangle\cdot\sum_{ki}\langle \boldsymbol{M}_{ki},(\boldsymbol{W}-\bar{\boldsymbol{W}})\rangle$ and $\|(\boldsymbol{W}-\bar{\boldsymbol{W}})\|_F \le \epsilon_{net}$.

Rewriting $(\boldsymbol{W}-\bar{\boldsymbol{W}}) = (\boldsymbol{W}-\bar{\boldsymbol{W}})\cdot(\boldsymbol{W}-\bar{\boldsymbol{W}})/\|(\boldsymbol{W}-\bar{\boldsymbol{W}})\|_F$ and using the fact that $(\boldsymbol{W}-\bar{\boldsymbol{W}})/\|(\boldsymbol{W}-\bar{\boldsymbol{W}})\|_F \in \mathcal{S}_{nr}$ and $\|(\boldsymbol{W}-\bar{\boldsymbol{W}})\|_F \le \epsilon_{net}$ and using Cauchy-Schwarz for the third term in the above expression, we have

$$\gamma^* \le b_2 + \epsilon_{net}^2\gamma^* + 2\sqrt{\gamma^*\cdot\epsilon_{net}^2\gamma^*} = b_2 + \epsilon_{net}^2\gamma^* + 2\epsilon_{net}\gamma^*$$

Thus, $\gamma^* \le 1/(1-\epsilon_{net}^2-2\epsilon_{net})\cdot b_2$.

Proof for the lower bound: Let $\beta^* = \min_{\boldsymbol{W}\in\mathcal{S}_{nr}}\sum_{ki}\langle \boldsymbol{M}_{ki},\boldsymbol{W}\rangle^2$. Proceeding as above, we have

$$\beta^* \ge b_1 - 2\sqrt{\gamma^*\cdot\epsilon_{net}^2\gamma^*} = b_1 - 2\epsilon_{net}\gamma^*$$

∎

B. Proving GD iterations' lemmas: Proof of Lemma 3.4 (algebra lemma)

Recall that $\boldsymbol{U}_{vec}$ denotes the vectorized $\boldsymbol{U}$. We use this so that we can apply the simple vector version of the fundamental theorem of calculus [18, Chapter XIII, Theorem 4.2],[19, Lemma 2 proof] (given in Theorem 4.2) on the nr length vector $\nabla f(\boldsymbol{U}_{vec},\boldsymbol{B})$, and so that the Hessian can be expressed as an $nr\times nr$ matrix.

We apply Theorem 4.2 with $\boldsymbol{z}_0 \equiv \boldsymbol{U}_{vec}$, $\boldsymbol{z}^* \equiv (\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U})_{vec}$, and $g(\boldsymbol{z}) = \nabla f(\boldsymbol{z},\boldsymbol{B})$. Thus $d = d_2 = nr$ and $\nabla g(\boldsymbol{z})$ is the Hessian of $f(\boldsymbol{z},\boldsymbol{B})$ computed at $\boldsymbol{z}$. Let $\boldsymbol{U}(\tau) := \boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U} + \tau(\boldsymbol{U}-\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U})$. Applying the theorem,

$$\begin{aligned}&\nabla f(\boldsymbol{U}_{vec},\boldsymbol{B}) - \nabla f((\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U})_{vec},\boldsymbol{B})\\ &= \Big(\int_{\tau=0}^{1}\nabla^2_{\boldsymbol{U}_{vec}} f(\boldsymbol{U}(\tau)_{vec},\boldsymbol{B})d\tau\Big)(\boldsymbol{U}_{vec}-(\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U})_{vec})\end{aligned} \tag{9}$$

where

$$\nabla^2_{\boldsymbol{U}_{vec}} f(\boldsymbol{U}(\tau)_{vec},\boldsymbol{B}) = \sum_{ki}(\boldsymbol{a}_{ki}\otimes\boldsymbol{b}_k)(\boldsymbol{a}_{ki}\otimes\boldsymbol{b}_k)^\top := \text{ Hess} \tag{10}$$

This is an $nr\times nr$ matrix. Because the cost function is quadratic, the Hessian is constant w.r.t. τ. Henceforth, we refer to it as Hess . With this, the above simplifies to

$$\begin{aligned}&\nabla f(\boldsymbol{U}_{vec},\boldsymbol{B}) - \nabla f((\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U})_{vec},\boldsymbol{B})\\ &= \text{ Hess }(\boldsymbol{U}_{vec}-(\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U})_{vec}) = \text{Hess }(\boldsymbol{P}\boldsymbol{U})_{vec}\end{aligned} \tag{11}$$

with

$$\boldsymbol{P} := \boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top}$$

denoting the $n\times n$ projection matrix to project orthogonal to $\boldsymbol{U}^*$. This proof is motivated by a similar approach used in [19, Lemma 2 proof] to analyze GD for standard PR. However, there the application was much simpler because $f(.)$ was a function of one variable and at the true solution the gradient was zero, i.e., $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$. In our case $\nabla f(\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U},\boldsymbol{B}) \neq \boldsymbol{0}$ because $\boldsymbol{B}\neq\boldsymbol{B}^*$. But we can show that $\mathbb{E}[(\boldsymbol{I}-\boldsymbol{U}^*\boldsymbol{U}^{*\top})\nabla f(\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U},\boldsymbol{B})] = \boldsymbol{0}$ and this helps us get the final desired result.

From Algorithm 1, recall that $\hat{\boldsymbol{U}}^+ = \boldsymbol{U} - (\eta/m)\nabla f(\boldsymbol{U},\boldsymbol{B})$. Vectorizing this equation, and using (11), we get

$$\begin{aligned}(\hat{\boldsymbol{U}}^+)_{vec} &= \boldsymbol{U}_{vec} - (\eta/m)\nabla f(\boldsymbol{U}_{vec},\boldsymbol{B})\\ &= \boldsymbol{U}_{vec} - (\eta/m)\text{ Hess }(\boldsymbol{P}\boldsymbol{U})_{vec}\\ &\quad - (\eta/m)\nabla f((\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U})_{vec},\boldsymbol{B}))\end{aligned} \tag{12}$$

We can prove our final result by using (8) and the following simple facts:

1) For an $n\times n$ matrix $\boldsymbol{M}$, let $\text{big}(\boldsymbol{M}) := \boldsymbol{I}_r\otimes\boldsymbol{M}$. be an $nr\times nr$ block diagonal matrix with $\boldsymbol{M}$ in the diagonal blocks. For any $n\times r$ matrix $\boldsymbol{Z}$,
$$\text{big}(\boldsymbol{M})\boldsymbol{Z}_{vec} = (\boldsymbol{M}\boldsymbol{Z})_{vec} \tag{13}$$
2) Since $\boldsymbol{P}$ is idempotent, $\boldsymbol{P} = \boldsymbol{P}^2$. Also, because of its block diagonal structure, $\text{big}(\boldsymbol{M}^2) = (\text{big}(\boldsymbol{M}))^2$. Thus,
$$\text{big}(\boldsymbol{P}) = \text{big}(\boldsymbol{P}^2) = (\text{big}(\boldsymbol{P}))^2 = \text{big}(\boldsymbol{P}))\boldsymbol{I}_{nr}(\text{big}(\boldsymbol{P}) \tag{14}$$

Left multiplying both sides of (12) by $\text{big}(\boldsymbol{P})$, and using (13), (14), and (8),

$$\begin{aligned}&\text{big}(\boldsymbol{P})(\hat{\boldsymbol{U}}^+)_{vec} = \text{big}(\boldsymbol{P})\boldsymbol{U}_{vec} - (\eta/m)\text{big}(\boldsymbol{P})\text{ Hess }(\boldsymbol{P}\boldsymbol{U})_{vec}\\ &\quad - (\eta/m)\text{big}(\boldsymbol{P})\nabla f((\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U})_{vec},\boldsymbol{B})\\ &= \text{big}(\boldsymbol{P})\boldsymbol{I}_{nr}\text{big}(\boldsymbol{P})\boldsymbol{U}_{vec} - (\eta/m)\text{big}(\boldsymbol{P})\text{ Hess big}(\boldsymbol{P})\boldsymbol{U}_{vec}\\ &\quad - (\eta/m)\text{big}(\boldsymbol{P})\nabla f((\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U})_{vec},\boldsymbol{B})\\ &= \text{big}(\boldsymbol{P})(\boldsymbol{I}_{nr} - (\eta/m)\text{ Hess})\text{big}(\boldsymbol{P})\boldsymbol{U}_{vec}\\ &\quad - (\eta/m)\text{big}(\boldsymbol{P})\nabla f((\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U})_{vec},\boldsymbol{B}).\end{aligned}$$

Thus, using $\|\text{big}(\boldsymbol{P})\| = \|\boldsymbol{P}\| = 1$, (13), and (8),

$$\|(\boldsymbol{P}\hat{\boldsymbol{U}}^+)_{vec}\| \le \|\boldsymbol{I}_{nr} - (\eta/m)\ \text{Hess}\ \|\ \|(\boldsymbol{P}\boldsymbol{U})_{vec}\| + (\eta/m)\|(\nabla f((\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U}), \boldsymbol{B}))_{vec}\| \quad (15)$$

Converting the vectors to matrices, using $||\boldsymbol{M}_{vec}|| = ||\boldsymbol{M}||_F$, and substituting for $\boldsymbol{P}$,

$$\|(\boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\hat{\boldsymbol{U}}^+\|_F \le \|\boldsymbol{I}_{nr} - (\eta/m)\ \text{Hess}\ \|\ \|(\boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\boldsymbol{U}\|_F + (\eta/m)\|(\boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\nabla f((\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U}), \boldsymbol{B})\|_F$$

Since $\hat{\boldsymbol{U}}^+ \overset{\text{QR}}{=} \boldsymbol{U}^+\boldsymbol{R}^+$ and since $\|\boldsymbol{M}_1\boldsymbol{M}_2\|_F \le \|\boldsymbol{M}_1\|_F\|\boldsymbol{M}_2\|$, this means that

$$\text{SD}(\boldsymbol{U}^*, \boldsymbol{U}^+) \le \|(\boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\hat{\boldsymbol{U}}^+\|_F\|(\boldsymbol{R}^+)^{-1}\|.$$

Since $\|(\boldsymbol{R}^+)^{-1}\| = 1/\sigma_{\min}(\boldsymbol{R}^+) = 1/\sigma_{\min}(\hat{\boldsymbol{U}}^+)$, using $\hat{\boldsymbol{U}}^+ = \boldsymbol{U} - (\eta/m)\nabla f(\boldsymbol{U}, \boldsymbol{B})$,

$$\|(\boldsymbol{R}^+)^{-1}\| = \frac{1}{\sigma_{\min}(\boldsymbol{U} - (\eta/m)\nabla f(\boldsymbol{U}, \boldsymbol{B}))} \le \frac{1}{1 - (\eta/m)\|\nabla f(\boldsymbol{U}, \boldsymbol{B})\|}$$

where we used $\sigma_{\min}(\boldsymbol{U} - (\eta/m)\nabla f(\boldsymbol{U}, \boldsymbol{B})) \ge \sigma_{\min}(\boldsymbol{U}) - (\eta/m)\|\nabla f(\boldsymbol{U}, \boldsymbol{B})\| = 1 - (\eta/m)\|\nabla f(\boldsymbol{U}, \boldsymbol{B})\|$ for the last inequality. Combining the last three equations above proves our lemma.

C. Proof of GD iterations' lemmas: Proof of Lemma 3.5

1) Upper and Lower bounding the Hessian eigenvalues and hence HessTerm: First assume the event that implies that the conclusions of Lemma 3.3 hold.

Recall from (10) that $\text{Hess} := \nabla^2_{\tilde{\boldsymbol{U}}_{vec}} f(\tilde{\boldsymbol{U}}_{vec}; \boldsymbol{B}) = \sum_{ki}(\boldsymbol{a}_{ki} \otimes \boldsymbol{b}_k)(\boldsymbol{a}_{ki} \otimes \boldsymbol{b}_k)^\top$. Since Hess is a positive semi-definite matrix, $\lambda_{\min}(\text{Hess}) = \min_{\boldsymbol{w}\in\mathcal{S}_{nr}} \boldsymbol{w}^\top \text{Hess}\ \boldsymbol{w}$ and $\lambda_{\max}(\text{Hess}) = \max_{\boldsymbol{w}\in\mathcal{S}_{nr}} \boldsymbol{w}^\top \text{Hess}\ \boldsymbol{w}$. For a fixed $\boldsymbol{w} \in \mathcal{S}_{nr}$,

$$\boldsymbol{w}^\top \text{Hess}\ \boldsymbol{w} = \sum_{ki}(\boldsymbol{a}_{ki}{}^\top \boldsymbol{W}\boldsymbol{b}_k)^2$$

where $\boldsymbol{W}$ is an $n \times r$ matrix with $\|\boldsymbol{W}\|_F = 1$. Clearly $(\boldsymbol{a}_{ki}{}^\top \boldsymbol{W}\boldsymbol{b}_k)^2$ are mutually independent sub-exponential random variables (r.v.) with sub-exponential norm $K_{ki} \le \|\boldsymbol{W}\boldsymbol{b}_k\|^2$. Also, $\mathbb{E}[(\boldsymbol{a}_{ki}{}^\top \boldsymbol{W}\boldsymbol{b}_k)^2] = \|\boldsymbol{W}\boldsymbol{b}_k\|^2$ and thus $\mathbb{E}[\sum_{ki}(\boldsymbol{a}_{ki}{}^\top \boldsymbol{W}\boldsymbol{b}_k)^2] = m\|\boldsymbol{W}\boldsymbol{B}\|_F^2$. Applying the sub-exponential Bernstein inequality, Theorem 2.8.1 of [26], for a fixed $\boldsymbol{W} \in \mathcal{S}_{nr}$ yields

$$\Pr\left\{\left|\sum_{ki}\left|\boldsymbol{a}_{ki}{}^\top \boldsymbol{W}\boldsymbol{b}_k\right|^2 - m\|\boldsymbol{W}\boldsymbol{B}\|_F^2\right| \ge t\right\} \le \exp\left[-c\min\left(\frac{t^2}{\sum_{ki} K_{ki}^2}, \frac{t}{\max_{ki} K_{ki}}\right)\right].$$

We set $t = \epsilon_3 m{\sigma^*_{\min}}^2$. By Lemma 3.3, $\|\boldsymbol{b}_k\|^2 \le 1.1\mu^2{\sigma^*_{\max}}^2(r/q) = 1.1\kappa^2\mu^2{\sigma^*_{\min}}^2(r/q)$. Thus,

$$\frac{t^2}{\sum_{ki} K_{ki}^2} \ge \frac{\epsilon_3^2 m^2{\sigma^*_{\min}}^4}{\sum_{ki}\|\boldsymbol{W}\boldsymbol{b}_k\|^4} \ge \frac{\epsilon_3^2 m{\sigma^*_{\min}}^4}{\max_k\|\boldsymbol{b}_k\|^2\sum_k\|\boldsymbol{W}\boldsymbol{b}_k\|^2} \ge \frac{\epsilon_3^2 m{\sigma^*_{\min}}^4}{\mu^2{\sigma^*_{\max}}^2(r/q)1.1.{\sigma^*_{\max}}^2} = c\epsilon_3^2 mq/r\mu^2\kappa^4$$

Here we used $\sum_k\|\boldsymbol{W}\boldsymbol{b}_k\|^2 = \|\boldsymbol{W}\boldsymbol{B}\|_F^2 \le \|\boldsymbol{W}\|_F\|\boldsymbol{B}\|_2 \le 1.1.\sigma^*_{\max}$ using the bound on $\|\boldsymbol{B}\|_2$ from Lemma 3.3. Also,

$$\frac{t}{\max_{ki} K_{ki}} \ge \frac{\epsilon_3 m{\sigma^*_{\min}}^2}{\max_{ki}\|\boldsymbol{W}\boldsymbol{b}_k\|^2} \ge \frac{\epsilon_3 m{\sigma^*_{\min}}^2}{1.1\mu^2{\sigma^*_{\max}}^2(r/q)} = c\epsilon_3 mq/r\mu^2\kappa^2.$$

Therefore, for a fixed $\boldsymbol{W} \in \mathcal{S}_{nr}$, w.p. $1 - \exp\left[-c\epsilon_3^2 mq/r\mu^2\kappa^4\right]$ we have

$$\left|\sum_{ki}\left|\boldsymbol{a}_{ki}{}^\top \boldsymbol{W}\boldsymbol{b}_k\right|^2 - m\|\boldsymbol{W}\boldsymbol{B}\|_F^2\right| \le \epsilon_3 m{\sigma^*_{\min}}^2. \quad (16)$$

and hence, by Lemma 3.3, w.p. $1 - \exp\left[-c\epsilon_3^2 mq/r\mu^2\kappa^4\right]$,

$$\sum_{ki}\left|\boldsymbol{a}_{ki}{}^\top \boldsymbol{W}\boldsymbol{b}_k\right|^2 \le m\|\boldsymbol{W}\boldsymbol{B}\|_F^2 + \epsilon_3 m{\sigma^*_{\min}}^2 \le m\|\boldsymbol{B}\|^2 + \epsilon_3 m{\sigma^*_{\min}}^2 \le m(1.1 + \epsilon_3/\kappa^2){\sigma^*_{\max}}^2. \quad (17)$$

and

$$\sum_{ki}\left|\boldsymbol{a}_{ki}{}^\top \boldsymbol{W}\boldsymbol{b}_k\right|^2 \ge m\|\boldsymbol{W}\boldsymbol{B}\|_F^2 - \epsilon_3 m{\sigma^*_{\min}}^2 \ge 0.9m{\sigma^*_{\min}}^2 + \epsilon_3 m{\sigma^*_{\min}}^2 \ge m(0.9 - \epsilon_3){\sigma^*_{\min}}^2. \quad (18)$$

To extend these bounds to all $\boldsymbol{W} \in \mathcal{S}_{nr}$ we apply Proposition 4.9 with $b_1 \equiv m(0.9 - \epsilon_3){\sigma^*_{\min}}^2$ and $b_2 \equiv m(1.1 + \epsilon_3/\kappa^2){\sigma^*_{\max}}^2$. Applying it we can conclude that, given the event that the claims of Lemma 3.3 holds, w.p. at least $1 - \exp(nr\log\kappa - cmq\epsilon_3^2/r\mu^2\kappa^4)$,

$$m(0.7 - 1.2\epsilon_3){\sigma^*_{\min}}^2 \le \lambda_{\min}(\text{Hess}) \le \lambda_{\max}(\text{Hess}) \le m(1.1 + \epsilon_3){\sigma^*_{\max}}^2$$

Using the probability from Lemma 3.3, the above bound holds w.p. at least $1 - \exp(nr\log\kappa - cmq\epsilon_3^2/r\mu^2\kappa^4) - \exp(\log q + r - cm)$.

2) Bounding the GradU Term: We have $\|\nabla f(\boldsymbol{U}, \boldsymbol{B})\| = \max_{\boldsymbol{z}\in\mathcal{S}_n, \boldsymbol{w}\in\mathcal{S}_r} \boldsymbol{z}^\top\nabla f(\boldsymbol{U}, \boldsymbol{B})\boldsymbol{w}$. For a fixed $\boldsymbol{z} \in \mathcal{S}_n, \boldsymbol{w} \in \mathcal{S}_r$ we have

$$\boldsymbol{z}^\top\left(\nabla f(\boldsymbol{U}, \boldsymbol{B}) - \mathbb{E}[\nabla f(\boldsymbol{U}, \boldsymbol{B})]\right)\boldsymbol{w} = \sum_{ki}\left[\left(\boldsymbol{a}_{ki}{}^\top \boldsymbol{U}\boldsymbol{b}_k - \boldsymbol{y}_{ki}\right)\left(\boldsymbol{a}_{ki}{}^\top \boldsymbol{z}\right)\left(\boldsymbol{w}^\top \boldsymbol{b}_k\right) - \mathbb{E}[.]\right]$$

where $\mathbb{E}[.]$ is the expected value of the first term. Clearly, the summands are independent sub-exponential r.v.s with norm $K_{ki} \le C\|\boldsymbol{x}_k - \boldsymbol{x}_k^*\|\|\boldsymbol{b}_k\|$. We apply the sub-exponential Bernstein inequality, Theorem 2.8.1 of [26], with $t = \epsilon_1\delta_t m{\sigma^*_{\max}}^2$. To apply this, we use bounds on $\|\boldsymbol{b}_k\|$, $\|\boldsymbol{X}^* - \boldsymbol{X}\|_F$ and $\|\boldsymbol{x}_k - \boldsymbol{x}_k^*\|$ from Lemma 3.3 to show that

$$\frac{t^2}{\sum_{ki} K_{ki}^2} \ge c\frac{\epsilon_1^2\delta_t^2 m^2{\sigma^*_{\max}}^4}{m\max_k\|\boldsymbol{b}_k\|^2\sum_k\|\boldsymbol{x}_k - \boldsymbol{x}_k^*\|^2} \ge c\frac{\epsilon_1^2\delta_t^2 m{\sigma^*_{\max}}^4}{C\mu^2{\sigma^*_{\max}}^2(r/q)\|\boldsymbol{X} - \boldsymbol{X}^*\|_F^2} \ge c\frac{\epsilon_1^2\delta_t^2 mq{\sigma^*_{\max}}^4}{C\mu^2{\sigma^*_{\max}}^2 r\delta_t^2{\sigma^*_{\max}}^2} = c\epsilon_1^2\frac{mq}{r\mu^2}.$$

and

$$\frac{t}{\max_{ki} K_{ki}} \ge c\frac{\epsilon_1\delta_t m{\sigma^*_{\max}}^2}{C\delta_t{\sigma^*_{\max}}^2\mu^2(r/q)} \ge c\epsilon_1\frac{mq}{r\mu^2}.$$

Therefore, for a fixed $\boldsymbol{z} \in \mathcal{S}_n, \boldsymbol{w} \in \mathcal{S}_r$ w.p. $1 - \exp(-c\epsilon_1^2 mq/r\mu^2)$,

$$\boldsymbol{z}^\top \left(\nabla f(\boldsymbol{U},\boldsymbol{B}) - \mathbb{E}[\nabla f(\boldsymbol{U},\boldsymbol{B})]\right) \boldsymbol{w} \leq \epsilon_1 \delta_t m {\sigma^*_{\max}}^2$$

Since $\nabla f(\boldsymbol{U},\boldsymbol{B}) = \sum_{ki} \boldsymbol{a}_{ki}\boldsymbol{a}_{ki}^\top(\boldsymbol{x}_k - \boldsymbol{x}_k^*)\boldsymbol{b}_k^\top$,

$$\mathbb{E}[\nabla f(\boldsymbol{U},\boldsymbol{B})] = m\sum_k (\boldsymbol{x}_k - \boldsymbol{x}_k^*)\boldsymbol{b}_k^\top = m\left(\boldsymbol{X} - \boldsymbol{X}^*\right)\boldsymbol{B}^\top.$$

Using the bounds on $\|\boldsymbol{X}^* - \boldsymbol{X}\|_F$ and $\|\boldsymbol{B}\|$ from Lemma 3.3,

$$\begin{aligned}\|\mathbb{E}[\nabla f(\boldsymbol{U},\boldsymbol{B})]\| &= m\|(\boldsymbol{X} - \boldsymbol{X}^*)\boldsymbol{B}^\top\| \\ &\leq m\|\boldsymbol{X} - \boldsymbol{X}^*\|\ \|\boldsymbol{B}\| \\ &\leq m\|\boldsymbol{X} - \boldsymbol{X}^*\|_F\ \|\boldsymbol{B}\| \\ &\leq 1.1 m\delta_t {\sigma^*_{\max}}^2\end{aligned}$$

Hence, for a fixed $\boldsymbol{z} \in \mathcal{S}_n, \boldsymbol{w} \in \mathcal{S}_r$ w.p. $1 - \exp\left[-c\epsilon_1^2 mq/r\mu^2\right]$ we have

$$|\boldsymbol{z}^\top \nabla f(\boldsymbol{U},\boldsymbol{B})\boldsymbol{w}| \leq (1.1 + \epsilon_1) m\delta_t {\sigma^*_{\max}}^2.$$

Applying Proposition 4.7, this implies that, w.p. $1 - \exp((n+r)(\log 17) - c\epsilon_1^2 mq/r\mu^2)$, $\max_{\boldsymbol{z}\in\mathcal{S}_n, \boldsymbol{w}\in\mathcal{S}_r} \boldsymbol{z}^\top \nabla f(\boldsymbol{U},\boldsymbol{B})\boldsymbol{w} \leq 1.4(1.1+\epsilon_1) m\delta_t {\sigma^*_{\max}}^2$.

3) Bounding Term2: First, since $\text{Term2} = (\boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\sum_{ki} \boldsymbol{a}_{ki}(\boldsymbol{a}_{ki}^\top \boldsymbol{U}^*(\boldsymbol{U}^{*\top}\boldsymbol{U}\boldsymbol{b}_k - \boldsymbol{b}_k^*))\boldsymbol{b}_k^\top$, and $\mathbb{E}[\boldsymbol{a}_{ki}\boldsymbol{a}_{ki}^\top] = \boldsymbol{I}$,

$$\mathbb{E}[\text{Term2}] = 0$$

We have

$$\begin{aligned}&\|(\boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\nabla f((\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U}),\boldsymbol{B})\|_F \\ &\quad = \max_{\boldsymbol{W}\in\mathcal{S}_{nr}} \langle (\boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\nabla f((\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U}),\boldsymbol{B}),\ \boldsymbol{W}\rangle\end{aligned}$$

For a fixed $n \times r$ matrix $\boldsymbol{W}$ with unit Frobenius norm,

$$\begin{aligned}&\langle (\boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\nabla f((\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U}),\boldsymbol{B}),\ \boldsymbol{W}\rangle \\ &\quad = \sum_{ki} \left(\boldsymbol{a}_{ki}^\top \boldsymbol{U}^*(\boldsymbol{U}^{*\top}\boldsymbol{U}\boldsymbol{b}_k - \boldsymbol{b}_k^*)\right)\left(\boldsymbol{a}_{ki}^\top(\boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\boldsymbol{W}\boldsymbol{b}_k\right)\end{aligned}$$

Observe that the summands are independent, zero mean, sub-exponential r.v.s with sub-exponential norm $K_{ki} \leq C\|\boldsymbol{U}^{*\top}\boldsymbol{U}\boldsymbol{b}_k - \boldsymbol{b}_k^*\|\|(\boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\boldsymbol{W}\boldsymbol{b}_k\| \leq \|\boldsymbol{U}^{*\top}\boldsymbol{U}\boldsymbol{b}_k - \boldsymbol{b}_k^*\|\|\boldsymbol{W}\boldsymbol{b}_k\|$. We can now apply the sub-exponential Bernstein inequality Theorem 2.8.1 of [26]. Let $t = \epsilon_2\delta_t m{\sigma^*_{\max}}^2$. Using the bound on $\|\boldsymbol{U}^{*\top}\boldsymbol{U}\boldsymbol{b}_k - \boldsymbol{b}_k^*\|$ from Lemma 3.3 followed by Assumption 1.1 (right incoherence), and also the bound on $\|\boldsymbol{B}\|$ from Lemma 3.3,

$$\begin{aligned}\frac{t^2}{\sum_{ki} K_{ki}^2} &\geq \frac{\epsilon_2^2\delta_t^2 m^2 {\sigma^*_{\max}}^4}{\delta_t^2{\sigma^*_{\max}}^2\mu^2(r/q)\sum_{ki}\|\boldsymbol{W}\boldsymbol{b}_k\|^2} \\ &\geq \frac{\epsilon_2^2 m^2{\sigma^*_{\max}}^2}{C\mu^2(r/q)m\|\boldsymbol{W}\boldsymbol{B}\|_F^2} \geq \frac{\epsilon_2^2 m^2{\sigma^*_{\max}}^2}{\mu^2(r/q)m{\sigma^*_{\max}}^2} \\ &\geq c\epsilon_2^2 mq/r\mu^2,\end{aligned}$$

and

$$\frac{t}{\max_{ki} K_{ki}} \geq \frac{\epsilon_2\delta_t m{\sigma^*_{\max}}^2}{C\delta_t\kappa^2\mu^2{\sigma^*_{\max}}^2(r/q)} \geq c\epsilon_2 mq/(r\kappa^2\mu^2).$$

Thus, by the sub-exponential Bernstein inequality, for a fixed $\boldsymbol{W} \in \mathcal{S}_{nr}$, w.p. $1 - \exp(-c\epsilon_2^2 mq/r\kappa^2\mu^2)$,

$$\langle (\boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\nabla f((\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U}),\boldsymbol{B}),\ \boldsymbol{W}\rangle \leq \epsilon_2\delta_t m{\sigma^*_{\max}}^2.$$

Applying Proposition 4.8, w.p. at least $1 - \exp(nr - c\epsilon_2^2 mq/r\kappa^2\mu^2)$, $\max_{\boldsymbol{W}\in\mathcal{S}_{nr}} \langle (\boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\nabla f((\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U}),\boldsymbol{B}),\boldsymbol{W}\rangle \leq 1.2\epsilon_2\delta_t m{\sigma^*_{\max}}^2$.

D. Proof of GD iterations' lemmas: Proof of Lemma 3.3, all parts other than the first part

Recall that $\boldsymbol{g}_k = \boldsymbol{U}^\top\boldsymbol{x}_k^* = \boldsymbol{U}^\top\boldsymbol{U}^*\boldsymbol{b}_k^*$, and $\boldsymbol{G} = \boldsymbol{U}^\top\boldsymbol{U}^*\boldsymbol{B}^*$.

Using the SD bound and the first part, $\|\boldsymbol{g}_k - \boldsymbol{b}_k\| \leq 0.4\delta_t\|\boldsymbol{b}_k^*\|$.

Since $\boldsymbol{x}_k^* - \boldsymbol{x}_k = \boldsymbol{U}\boldsymbol{g}_k + (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^* - \boldsymbol{U}\boldsymbol{b}_k = \boldsymbol{U}(\boldsymbol{g}_k - \boldsymbol{b}_k) + (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^*$, using (3),

$$\|\boldsymbol{x}_k^* - \boldsymbol{x}_k\| \leq \|\boldsymbol{g}_k - \boldsymbol{b}_k\| + \|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{U}^*\boldsymbol{b}_k^*\| \leq 1.4\delta_t\|\boldsymbol{b}_k^*\|.$$

$\|\boldsymbol{U}^{*\top}\boldsymbol{U}\boldsymbol{b}_k - \boldsymbol{b}_k^*\| = \|\boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{U}\boldsymbol{b}_k - \boldsymbol{U}^*\boldsymbol{b}_k^*\| = \|\boldsymbol{U}\boldsymbol{b}_k - (\boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\boldsymbol{U}\boldsymbol{b}_k - \boldsymbol{U}^*\boldsymbol{b}_k^*\| = \|\boldsymbol{x}_k - (\boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\boldsymbol{U}\boldsymbol{b}_k - \boldsymbol{x}_k^*\| \leq \|\boldsymbol{x}_k - \boldsymbol{x}_k^*\| + \|(\boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\boldsymbol{U}\boldsymbol{b}_k\| \leq 2.4\delta_t\|\boldsymbol{b}_k^*\|$

Bounding $\|\boldsymbol{G} - \boldsymbol{B}\|_F$ and $\|\boldsymbol{X}^* - \boldsymbol{X}\|_F$: Since $\sum_k \|\boldsymbol{M}\boldsymbol{b}_k^*\|^2 = \|\boldsymbol{M}\boldsymbol{B}^*\|_F^2 \leq \|\boldsymbol{M}\|_F^2\|\boldsymbol{B}^*\|^2 = \|\boldsymbol{M}\|_F^2{\sigma^*_{\max}}^2$, we can use the first bound from (3) to conclude that

$$\begin{aligned}\|\boldsymbol{G} - \boldsymbol{B}\|_F^2 &= \sum_k \|\boldsymbol{g}_k - \boldsymbol{b}_k\|^2 \\ &\leq 0.4^2\sum_k \|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{U}^*\boldsymbol{b}_k^*\|^2 \\ &= 0.4^2\|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{U}^*\boldsymbol{B}^*\|_F^2 \leq 0.4^2\delta_t^2{\sigma^*_{\max}}^2\end{aligned}$$

and, similarly,

$$\begin{aligned}\|\boldsymbol{X}^* - \boldsymbol{X}\|_F^2 &\leq \sum_k \|\boldsymbol{g}_k - \boldsymbol{b}_k\|^2 + \sum_k \|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{U}^*\boldsymbol{b}_k^*\|^2 \\ &\leq (0.4^2 + 1^2)\delta_t^2{\sigma^*_{\max}}^2\end{aligned}$$

Incoherence of $\boldsymbol{b}_k$'s: Using the bound on $\|\boldsymbol{b}_k - \boldsymbol{g}_k\|$, and using $\|\boldsymbol{g}_k\| \leq \|\boldsymbol{b}_k^*\|$ and the right incoherence assumption,

$$\|\boldsymbol{b}_k\| = \|(\boldsymbol{b}_k - \boldsymbol{g}_k + \boldsymbol{g}_k)\| \leq (1 + 0.4\delta_t)\|\boldsymbol{b}_k^*\| \leq 1.04\sigma^*_{\max}\sqrt{r/q}.$$

Lower and Upper Bounds on $\sigma_i(\boldsymbol{B})$): Using the bound on $\|\boldsymbol{G} - \boldsymbol{B}\|_F$ and using $\text{SD}(\boldsymbol{U},\boldsymbol{U}^*) \leq \delta_t < c/\kappa$,

$$\begin{aligned}\sigma_{\min}(\boldsymbol{B}) &\geq \sigma_{\min}(\boldsymbol{G}) - \|\boldsymbol{G} - \boldsymbol{B}\| \\ &\geq \sigma_{\min}(\boldsymbol{U}^\top\boldsymbol{U}^*)\sigma_{\min}(\boldsymbol{B}^*) - \|\boldsymbol{G} - \boldsymbol{B}\|_F \\ &\geq \sqrt{1 - \|\boldsymbol{U}^{*\perp\top}\boldsymbol{U}\|^2}\sigma^*_{\min} - 0.4\delta_t\sigma^*_{\max} \\ &\geq \sqrt{1 - \delta_t^2}\sigma^*_{\min} - 0.4\delta_t\sigma^*_{\max} \geq 0.9\sigma^*_{\min}\end{aligned}$$

since we assumed $\delta_t \leq \delta_0 < 0.1/\kappa$. Similarly,

$$\begin{aligned}\|\boldsymbol{B}\| = \sigma_{\max}(\boldsymbol{B}) &\leq \sigma_{\max}(\boldsymbol{U}^\top\boldsymbol{U}^*)\sigma_{\max}(\boldsymbol{B}^*) + \|\boldsymbol{G} - \boldsymbol{B}\|_F \\ &\leq \sigma^*_{\max} + 0.4\delta_t\sigma^*_{\max} \leq 1.1\sigma^*_{\max}\end{aligned}$$

E. Proof of GD iterations' lemmas: Proof of Lemma 3.3, first part

We bound $\|\boldsymbol{g}_k - \boldsymbol{b}_k\|$ here. Recall that $\boldsymbol{g}_k = \boldsymbol{U}^\top \boldsymbol{x}_k^*$. Since $\boldsymbol{y}_k = \boldsymbol{A}_k \boldsymbol{x}_k^* = \boldsymbol{A}_k \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}_k^* + \boldsymbol{A}_k(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^*$, therefore

$$\begin{aligned}
\boldsymbol{b}_k &= \left(\boldsymbol{U}^\top \boldsymbol{A}_k{}^\top \boldsymbol{A}_k \boldsymbol{U}\right)^{-1} (\boldsymbol{U}^\top \boldsymbol{A}_k{}^\top) \boldsymbol{A}_k \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}_k^* \\
&\quad + \left(\boldsymbol{U}^\top \boldsymbol{A}_k{}^\top \boldsymbol{A}_k \boldsymbol{U}\right)^{-1} (\boldsymbol{U}^\top \boldsymbol{A}_k{}^\top) \boldsymbol{A}_k (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^*, \\
&= \left(\boldsymbol{U}^\top \boldsymbol{A}_k{}^\top \boldsymbol{A}_k \boldsymbol{U}\right)^{-1} \left(\boldsymbol{U}^\top \boldsymbol{A}_k{}^\top \boldsymbol{A}_k \boldsymbol{U}\right) \boldsymbol{U}^\top \boldsymbol{x}_k^* \\
&\quad + \left(\boldsymbol{U}^\top \boldsymbol{A}_k{}^\top \boldsymbol{A}_k \boldsymbol{U}\right)^{-1} (\boldsymbol{U}^\top \boldsymbol{A}_k{}^\top) \boldsymbol{A}_k (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^*, \\
&= \boldsymbol{g}_k + \left(\boldsymbol{U}^\top \boldsymbol{A}_k{}^\top \boldsymbol{A}_k \boldsymbol{U}\right)^{-1} (\boldsymbol{U}^\top \boldsymbol{A}_k{}^\top) \boldsymbol{A}_k (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^*.
\end{aligned}$$

Thus,

$$\begin{aligned}
\|\boldsymbol{b}_k - \boldsymbol{g}_k\| &\le \| \left(\boldsymbol{U}^\top \boldsymbol{A}_k{}^\top \boldsymbol{A}_k \boldsymbol{U}\right)^{-1} \| \\
&\quad \times \|\boldsymbol{U}^\top \boldsymbol{A}_k{}^\top \boldsymbol{A}_k (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^*\|. \qquad (19)
\end{aligned}$$

Using standard results from [26], one can show the following:

1) W.p. $\ge 1 - q\exp(r - cm)$, for all $k \in [q]$, $\min_{\boldsymbol{w}\in\mathcal{S}_r} \sum_i \left|\boldsymbol{a}_{ki}{}^\top \boldsymbol{U}\boldsymbol{w}\right|^2 \ge 0.7m$ and so
$$\begin{aligned}
\| \left(\boldsymbol{U}^\top \boldsymbol{A}_k{}^\top \boldsymbol{A}_k \boldsymbol{U}\right)^{-1} \| &= \frac{1}{\sigma_{\min}\left(\boldsymbol{U}^\top \boldsymbol{A}_k{}^\top \boldsymbol{A}_k \boldsymbol{U}\right)} \\
&= \frac{1}{\min_{\boldsymbol{w}\in\mathcal{S}_r} \sum_i \langle \boldsymbol{U}^\top \boldsymbol{a}_{ki}, \boldsymbol{w}\rangle^2} \\
&\le \frac{1}{0.7m}
\end{aligned}$$
2) W.p. at least $1 - q\exp(r - cm)$, for all $k \in [q]$,
$$\|\boldsymbol{U}^\top \boldsymbol{A}_k{}^\top \boldsymbol{A}_k (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^*\| \le 0.15m \|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^*\|$$

Combining the above two bounds and (19), w.p. at least $1 - 2\exp(\log q + r - cm)$, for all $k \in [q]$,

$$\|\boldsymbol{g}_k - \boldsymbol{b}_k\| \le 0.4 \| \left(\boldsymbol{I}_n - \boldsymbol{U}\boldsymbol{U}^\top\right) \boldsymbol{U}^* \boldsymbol{b}_k^* \|.$$

This completes the proof. We explain next how to get the above two bounds.

The first bound above follows by a restatement of Theorem 4.6.1 of [26]. Or, it follows more directly by using $\mathbb{E}[\sum_i \left|\boldsymbol{a}_{ki}{}^\top \boldsymbol{U}\boldsymbol{w}\right|^2] = m$, applying the sub-exponential Bernstein inequality [29, Theorem 2.8.1] to bound the deviation from this mean, and then applying Proposition 4.9 with $n \equiv 1, r \equiv r$ (epsilon net argument).

The second bound is obtained as follows. Notice that

$$\begin{aligned}
&\|\boldsymbol{U}^\top \boldsymbol{A}_k{}^\top \boldsymbol{A}_k (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^*\| \\
&= \max_{\boldsymbol{w}\in\mathcal{S}_r} \boldsymbol{w}^\top \boldsymbol{U}^\top \boldsymbol{A}_k{}^\top \boldsymbol{A}_k (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^* \\
&= \max_{\boldsymbol{w}\in\mathcal{S}_r} \sum_i (\boldsymbol{a}_{ki}{}^\top \boldsymbol{U}\boldsymbol{w})(\boldsymbol{a}_{ki}{}^\top (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^*)
\end{aligned}$$

Clearly $\mathbb{E}\left[\boldsymbol{U}^\top \boldsymbol{A}_k{}^\top \boldsymbol{A}_k (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^*\right] = \boldsymbol{U}^\top(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^* = 0$. Moreover, the summands are products of sub-Gaussian r.v.s and are thus sub-exponential. Also, the different summands are mutually independent and zero mean. Applying sub-exponential Bernstein with $t = \epsilon_0 m \|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^*\|$ for a fixed $\boldsymbol{w} \in \mathcal{S}_r$,

$$|\sum_i (\boldsymbol{a}_{ki}{}^\top \boldsymbol{U}\boldsymbol{w})(\boldsymbol{a}_{ki}{}^\top (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^*)| \le \epsilon_0 m \|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^*\|$$

w.p. at least $1 - \exp(-c\epsilon_0^2 m)$. Setting $\epsilon_0 = 0.1$, this implies that the above is bounded by $0.1m\|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^*\|$ w.p. at least $1 - \exp(-cm)$. By Proposition 4.8 with $n \equiv 1, r \equiv r$, the above is bounded by $0.12m\|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_k^*\|$ for all $\boldsymbol{w} \in \mathcal{S}_r$ w.p. at least $1 - \exp(r - cm)$. Using a union bound over all q columns, the bound holds for all q columns w.p. at least $1 - q\exp(r - cm)$.

F. Proof of Initialization lemmas/facts: Proof of Lemma 3.6

To see why (4) holds, it suffices to show that $\mathbb{E}[(\boldsymbol{X}_0)_k|\alpha] = \boldsymbol{x}_k^* \beta_k(\alpha)$ for each k. The easiest way to see this is to express $\boldsymbol{x}_k^* = \|\boldsymbol{x}_k^*\| \boldsymbol{Q}_k \boldsymbol{e}_1$ where $\boldsymbol{Q}_k$ is an $n \times n$ unitary matrix with first column $\boldsymbol{x}_k^*/\|\boldsymbol{x}_k^*\|$; and to use the fact that $\tilde{\boldsymbol{a}}_{ki} := \boldsymbol{Q}_k^\top \boldsymbol{a}_{ki}$ has the same distribution as $\boldsymbol{a}_{ki}$, both are $\mathcal{N}(0, \boldsymbol{I}_n)$. Using $\boldsymbol{Q}_k \boldsymbol{Q}_k^\top = \boldsymbol{I}$, $(\boldsymbol{X}_0)_k = (1/m)\sum_i \boldsymbol{Q}_k \boldsymbol{Q}_k^\top \boldsymbol{a}_{ki} \boldsymbol{a}_{ki}^\top \|\boldsymbol{x}_k^*\| \boldsymbol{Q}_k \boldsymbol{e}_1 \mathbb{1}_{\|\boldsymbol{x}_k^*\| |\boldsymbol{a}_{ki}^\top \boldsymbol{Q}_k \boldsymbol{e}_1| \le \sqrt{\alpha}} = (1/m)\sum_i \boldsymbol{Q}_k \|\boldsymbol{x}_k^*\| \tilde{\boldsymbol{a}}_{ki} \tilde{\boldsymbol{a}}_{ki}(1) \mathbb{1}_{|\tilde{\boldsymbol{a}}_{ki}(1)| \le \sqrt{\alpha}/\|\boldsymbol{x}_k^*\|}$. Thus $\mathbb{E}[((\boldsymbol{X}_0)_k] = (1/m) m \boldsymbol{Q}_k \|\boldsymbol{x}_k^*\| \boldsymbol{e}_1 \mathbb{E}[\zeta^2 \mathbb{1}_{|\zeta| < \sqrt{\alpha}/\|\boldsymbol{x}_k^*\|}]$. This follows because $\mathbb{E}[\boldsymbol{a}\boldsymbol{a}(1) \mathbb{1}_{|\boldsymbol{a}(1)| < \beta} = \boldsymbol{e}_1 \mathbb{E}[\boldsymbol{a}(1)^2 \mathbb{1}_{|\boldsymbol{a}(1)| < \beta}]$.

Recall that $\tilde{C} = 9\kappa^2\mu^2$ and $\tilde{c} = c/\tilde{C}$ for a $c < 1$. Recall also that $\boldsymbol{X}^* \overset{\text{SVD}}{=} \boldsymbol{U}^* \boldsymbol{\Sigma}^* \boldsymbol{V}^*$ and $\mathbb{E}[\boldsymbol{X}_0|\alpha] \overset{\text{SVD}}{=} \boldsymbol{U}^* \check{\boldsymbol{\Sigma}}^* \check{\boldsymbol{V}}$. Thus, using (4), $\check{\boldsymbol{\Sigma}}^* = \boldsymbol{\Sigma}^* \boldsymbol{V}^* \boldsymbol{D} \check{\boldsymbol{V}}^\top$. Hence,

$$\begin{aligned}
\sigma_r(\mathbb{E}[\boldsymbol{X}_0|\alpha]) &= \sigma_{\min}(\check{\boldsymbol{\Sigma}}^*) \\
&= \sigma_{\min}(\boldsymbol{\Sigma}^* \boldsymbol{V}^* \boldsymbol{D} \check{\boldsymbol{V}}^\top) \\
&\ge \sigma_{\min}(\boldsymbol{\Sigma}^*) \sigma_{\min}(\boldsymbol{V}^*) \sigma_{\min}(\boldsymbol{D}) \sigma_{\min}(\check{\boldsymbol{V}}^\top) \\
&= \sigma_{\min}^* \cdot 1 \cdot (\min_k \beta_k(\alpha)) \cdot 1
\end{aligned}$$

Also, $\sigma_{r+1}(\mathbb{E}[\boldsymbol{X}_0]) = 0$ since it is a rank r matrix. Thus, using Wedin's $\sin\Theta$ theorem for the Frobenius norm subspace distance SD [27], [28][Theorem 2.3.1, second row] (specified in Theorem 4.1 above) applied with $\boldsymbol{M} \equiv \boldsymbol{X}_0$, $\boldsymbol{M}^* \equiv \mathbb{E}[\boldsymbol{X}_0]$ we get (5).

G. Proof of Initialization lemmas and facts: Proof of Lemma 3.8

Proof of first part of Lemma 3.8: The proof involves an application of the sub-Gaussian Hoeffding inequality, Theorem 2.6.2 of [26], followed by an epsilon-net argument. The application of sub-Gaussian Hoeffding uses conditioning on α for $\alpha \in \mathcal{E}$. For $\alpha \in \mathcal{E}$, $\alpha \le \sqrt{\tilde{C}(1+\epsilon_1)}\|\boldsymbol{X}^*\|_F/\sqrt{q}$ and this helps get a simple probability bound. Since α is independent of all $\boldsymbol{a}_{ki}, \boldsymbol{y}_{ki}$'s used in defining $\boldsymbol{X}_0$, the conditioning does not change anything else in our proof. For example, the different summands are mutually independent even conditioned on it.

We have,

$$\|\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha]\| = \max_{\boldsymbol{z}\in\mathcal{S}_n, \boldsymbol{w}\in\mathcal{S}_q} \langle \boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha],\ \boldsymbol{z}\boldsymbol{w}^\top \rangle.$$

For a fixed $\boldsymbol{z} \in \mathcal{S}_n, \boldsymbol{w} \in \mathcal{S}_q$, we have

$$\begin{aligned}
&\langle \boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha],\ \boldsymbol{z}\boldsymbol{w}^\top \rangle \\
&= \frac{1}{m} \sum_{ki} \boldsymbol{w}(k) \boldsymbol{y}_{ki} (\boldsymbol{a}_{ki}{}^\top \boldsymbol{z}) \mathbb{1}_{\{|\boldsymbol{y}_{ki}|^2 \le \alpha\}} \\
&\quad - \mathbb{E}\left[\boldsymbol{w}(k) \boldsymbol{y}_{ki} (\boldsymbol{a}_{ki}{}^\top \boldsymbol{z}) \mathbb{1}_{\{|\boldsymbol{y}_{ki}|^2 \le \alpha\}}\right].
\end{aligned}$$

The summands are mutually independent, zero mean sub-Gaussian r.v.s with sub-Gaussian norm $K_{ki} \le C|\boldsymbol{w}(k)|\sqrt{\alpha}/m$. For $\alpha \in \mathcal{E}$, $\alpha \le \sqrt{\tilde{C}(1+\epsilon_1)}\|\boldsymbol{X}^*\|_F/m\sqrt{q}$. Let $t = \epsilon_1\|\boldsymbol{X}^*\|_F$. Then, for any $\alpha \in \mathcal{E}$,

$$\frac{t^2}{\sum_{ki} K_{ki}^2} \ge \frac{\epsilon_1^2\|\boldsymbol{X}^*\|_F^2}{\sum_{ki} \tilde{C}(1+\epsilon_1)\boldsymbol{w}(k)^2\|\boldsymbol{X}^*\|_F^2/m^2q} \ge \frac{\epsilon_1^2 mq}{C\mu^2\kappa^2}$$

since $\sum_k \boldsymbol{w}(k)^2 = \|\boldsymbol{w}\|^2 = 1$. Thus, for a fixed $\boldsymbol{z} \in \mathcal{S}_n, \boldsymbol{w} \in \mathcal{S}_q$, by sub-Gaussian Hoeffding, we conclude that, conditioned on α, for any $\alpha \in \mathcal{E}$, w.p. at least $1 - \exp\left[-c\epsilon_1^2 mq/\mu^2\kappa^2\right]$,

$$\langle \boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha],\ \boldsymbol{z}\boldsymbol{w}^\top\rangle \le C\epsilon_1\|\boldsymbol{X}^*\|_F.$$

The rest of the proof follows by a standard epsilon net argument summarized in Proposition 4.7. Applying it, conditioned on α, for any $\alpha \in \mathcal{E}$, w.p. at least $1 - \exp\left[(n+q) - c\epsilon_1^2 mq/\mu^2\kappa^2\right]$, $\max_{\boldsymbol{z}\in\mathcal{S}_n,\boldsymbol{w}\in\mathcal{S}_q}\langle \boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha],\ \boldsymbol{z}\boldsymbol{w}^\top\rangle \le 1.4C\epsilon_1\|\boldsymbol{X}^*\|_F$. ■

Proof of second part of Lemma 3.8: We have

$$\|\left(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha]\right)^\top \boldsymbol{U}^*\|_F = \max_{\boldsymbol{W}\in\mathcal{S}_{qr}} \langle \boldsymbol{W}, \left(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha]\right)^\top \boldsymbol{U}^*\rangle$$

For a fixed $\boldsymbol{W} \in \mathcal{S}_{qr}$,

$$\begin{aligned}&\langle \boldsymbol{W}, \left(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha]\right)^\top \boldsymbol{U}^*\rangle \\ &\quad= \text{trace}\left(\boldsymbol{W}^\top\left(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha]\right)^\top \boldsymbol{U}^*\right) \\ &\quad= \frac{1}{m}\sum_{ki}\left(\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top \boldsymbol{U}^*\boldsymbol{w}_k)\mathbb{1}_{\{|\boldsymbol{y}_{ki}|^2\le\alpha\}} - \mathbb{E}[.]\right)\end{aligned}$$

Conditioned on α, for an $\alpha \in \mathcal{E}$, the summands are independent zero mean sub-Gaussian r.v.s with subGaussian norm $K_{ki} \le \sqrt{\alpha}\|\boldsymbol{w}_k\|/m \le \sqrt{\tilde{C}(1+\epsilon_1)}\|\boldsymbol{X}^*\|_F\|\boldsymbol{w}_k\|/m\sqrt{q}$. Thus,

$$\sum_{ki} K_{ki}^2 \le m\tilde{C}(1+\epsilon_1)\|\boldsymbol{W}\|_F^2\|\boldsymbol{X}^*\|_F^2/m^2q = \tilde{C}\|\boldsymbol{X}^*\|_F^2/mq$$

Applying the sub-Gaussian Hoeffding inequality Theorem 2.6.2 of [26], for a fixed $\boldsymbol{W} \in \mathcal{S}_{qr}$, conditioned on α, for an $\alpha \in \mathcal{E}$, w.p. $1 - \exp\left[-\epsilon_1^2 mq/C\mu^2\kappa^2\right]$,

$$\text{trace}\left(\boldsymbol{W}^\top\left(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha]\right)^\top \boldsymbol{U}^*\right) \le \epsilon_1\|\boldsymbol{X}^*\|_F.$$

The rest of the proof follows by a standard epsilon net argument summarized in Proposition 4.8. Applying Proposition 4.8, conditioned on α, for an $\alpha \in \mathcal{E}$, w.p. at least $1 - \exp\left[qr - c\epsilon_1^2 mq/\mu^2\kappa^2\right]$, $\max_{\boldsymbol{W}\in\mathcal{S}_{qr}} \text{trace}\left(\boldsymbol{W}^\top\left(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha]\right)^\top \boldsymbol{U}^*\right) < 1.2\epsilon_1\|\boldsymbol{X}^*\|_F$. ■

Proof of third part of Lemma 3.8: We have

$$\|\left(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha]\right)\check{\boldsymbol{V}}^\top\|_F = \max_{\boldsymbol{W}\in\mathcal{S}_{nr}} \langle \left(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha]\right)\check{\boldsymbol{V}}^\top,\ \boldsymbol{W}\rangle$$

For a fixed $\boldsymbol{W} \in \mathcal{S}_{nr}$ we have,

$$\begin{aligned}&\langle \left(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha]\right)\check{\boldsymbol{V}}^\top,\ \boldsymbol{W}\rangle \\ &\quad= \frac{1}{m}\sum_{ki}\left(\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top \boldsymbol{W}\check{\boldsymbol{v}}_k)\mathbb{1}_{\{|\boldsymbol{y}_{ki}|^2\le\alpha\}} - \mathbb{E}[.]\right)\end{aligned}$$

where $\mathbb{E}[.]$ is the expected value of the first term. Conditioned on α, for an $\alpha \in \mathcal{E}$, the summands are independent, zero mean, sub-Gaussian r.v.s with subGaussian norm $K_{ki} \le C\sqrt{\alpha}\|\boldsymbol{W}\check{\boldsymbol{v}}_k\| \le C\sqrt{\tilde{C}(1+\epsilon_1)}\|\boldsymbol{X}^*\|_F\|\boldsymbol{W}\check{\boldsymbol{v}}_k\|/m\sqrt{q}$. Thus, by applying the sub-Gaussian Hoeffding inequality Theorem 2.6.2 of [26], with $t = \epsilon_1\|\boldsymbol{X}^*\|_F$, and using $\|\boldsymbol{W}\check{\boldsymbol{V}}\|_F = 1$ (holds since $\check{\boldsymbol{V}}$ contains orthormal rows which are right singular vectors of $\mathbb{E}[\boldsymbol{X}_0|\alpha]$), conditioned on α, for an $\alpha \in \mathcal{E}$, we will get that,

$$\frac{t^2}{\sum_{ki} K_{ki}^2} \ge \frac{m^2\epsilon_1^2\|\boldsymbol{X}^*\|_F^2}{\sum_{ki} \tilde{C}(1+\epsilon_1)\|\boldsymbol{X}^*\|_F^2\|\boldsymbol{W}\check{\boldsymbol{v}}_k\|^2/q} = \frac{mq\epsilon_1^2}{C\mu^2\kappa^2},$$

w.p. $1 - \exp\left[-c\epsilon_1^2 mq/(\mu^2\kappa^2)\right]$. Here we used the fact that $\check{\boldsymbol{V}}\check{\boldsymbol{V}}^\top = \boldsymbol{I}$ and thus $\|\boldsymbol{W}\check{\boldsymbol{V}}\|_F^2 = 1$.

$$\langle \left(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha]\right)\check{\boldsymbol{V}}^\top,\ \boldsymbol{W}\rangle \le C\epsilon_1\|\boldsymbol{X}^*\|_F.$$

Applying Proposition 4.8, conditioned on α, for an $\alpha \in \mathcal{E}$, w.p. at least $1 - \exp\left[nr - c\epsilon_1^2 mq/(\mu^2\kappa^2)\right]$, $\max_{\boldsymbol{W}\in\mathcal{S}_{nr}}\langle \left(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha]\right)\check{\boldsymbol{V}}^\top,\ \boldsymbol{W}\rangle \le 1.2C\epsilon_1\|\boldsymbol{X}^*\|_F$. ■

H. Proof of Initialization lemmas and facts: Proof of Facts

Proof of Fact 3.7: Apply sub-exponential Bernstein. ■

Proof of Fact 3.9: Let $\gamma_k = \frac{\sqrt{\tilde{C}(1-\epsilon_1)}\|\boldsymbol{X}^*\|_F}{\sqrt{q}\|\boldsymbol{x}_k^*\|}$. Since $\tilde{C} = 9\mu^2\kappa^2$ and $\|\boldsymbol{x}_k^*\|^2 \le \mu^2\kappa^2\|\boldsymbol{X}^*\|_F^2/q$ (Assumption 1.1) thus

$$\gamma_k \ge 3.$$

Now,

$$\begin{aligned}\mathbb{E}\left[\zeta^2\mathbb{1}_{\{|\zeta|\le\gamma_k\}}\right] &= 1 - \mathbb{E}\left[\zeta^2\mathbb{1}_{\{|\zeta|\ge\gamma_k\}}\right] \\ &\ge 1 - \frac{2}{\sqrt{2\pi}}\int_3^\infty z^2\exp(-z^2/2)dz \\ &\ge 1 - \frac{2e^{-1/2}}{\sqrt{\pi}}\int_3^\infty z\exp(-z^2/4)dz \\ &= 1 - \frac{2e^{-11/4}}{\sqrt{\pi}} \ge 0.92.\end{aligned}$$

The first inequality used $\gamma_k \ge 3$. The second used the fact that $z\exp(-z^2/4) \le \sqrt{2e}$ for all $z \in \Re$. ■

In all the proofs above, notice that the only thing we used about $\check{\boldsymbol{V}}$ is the fact that its rows contain singular vectors and thus $\check{\boldsymbol{V}}\check{\boldsymbol{V}}^\top = \boldsymbol{I}$ and so $\sigma_r(\check{\boldsymbol{V}}) = \sigma_1(\check{\boldsymbol{V}}) = 1$. We never required incoherence for it

V. Extension to Low Rank Phase Retrieval (LRPR)

In LRPR, recall that, we measure $\boldsymbol{y}_{(mag)_k} = |\boldsymbol{A}_k\boldsymbol{x}_k^*|$. This problem commonly occurs in dynamic phaseless imaging applications such as Fourier ptychography. Because of the magnitude-only measurements, we can recover each column only up to a global phase uncertainty. We use $\text{dist}(\boldsymbol{x}^*, \boldsymbol{x}) := \min_{\theta\in[-\pi,\pi]}\|\boldsymbol{x}^* - e^{-j\theta}\boldsymbol{x}\|$ to quantify this phase invariant distance [30], [21]. Also, for a complex number, z, we use $\bar{z}$ to denote its conjugate and we use $phase(z) := z/|z|$.

A. AltGD-Min-LRPR algorithm

With three simple changes that we explain next, the AltGD-Min approach also solves LRPR and provides the fastest existing solution for it. First, observe that because of the magnitude-only measurements, we cannot use $\boldsymbol{X}_0$ with $\boldsymbol{y}_{ki}$ replaced by $\boldsymbol{y}_{(mag)_{ki}}$ for initialization. The reason is $\mathbb{E}[\boldsymbol{a}_{ki}\boldsymbol{y}_{(mag)_{ki}}] = 0$

Algorithm 2 The AltGD-Min-LRPR algorithm.

1: **Input:** $\boldsymbol{y}_{(mag)_k}, \boldsymbol{A}_k, k \in [q]$
2: **Parameters:** GD step size, η; Number of iterations, T
3: **Sample-split:** Partition the measurements and measurement matrices into $2T+1$ equal-sized disjoint sets: one set for initialization and $2T$ sets for the iterations. Denote these by ${\boldsymbol{y}_{(mag)_k}}^{(\tau)}, \boldsymbol{A}_k^{(\tau)}, \tau = 0, 1, \dots 2T$.
4: **Initialization:**
5: Compute $\boldsymbol{U}_0$ as the top r singular vectors of $\boldsymbol{Y}_U := \frac{1}{mq}\sum_{ki}(\boldsymbol{y}_{(mag)_{ki}})^2 \boldsymbol{a}_{ki}\boldsymbol{a}_{ki}^\top \mathbb{1}_{\{(\boldsymbol{y}_{(mag)_{ki}})^2 \le \tilde{C}\frac{1}{mq}\sum_{ki}(\boldsymbol{y}_{(mag)_{ki}})^2\}}$ with $\boldsymbol{y}_{(mag)_{ki}} \equiv {\boldsymbol{y}_{(mag)_k}}^{(0)}, \boldsymbol{a}_{ki} \equiv \boldsymbol{a}_{ki}^{(0)}$.
6: **GDmin Iterations:**
7: **for** $t = 1$ **to** T **do**
8: Let $\boldsymbol{U} \leftarrow \boldsymbol{U}_{t-1}$.
9: **Update $\boldsymbol{b}_k, \boldsymbol{x}_k$:** For each $k \in [q]$, set $(\boldsymbol{b}_k)_t \leftarrow \mathrm{RWF}({\boldsymbol{y}_{(mag)_k}}^{(t)}, (\boldsymbol{U}^\top \boldsymbol{A}_k^{(t)}), T_{RWF,t})$. Set $(\boldsymbol{x}_k)_t \leftarrow \boldsymbol{U}(\boldsymbol{b}_k)_t$
10: **Estimate gradient w.r.t. $\boldsymbol{U}$:** With $\boldsymbol{y}_{(mag)_{ki}} \equiv {\boldsymbol{y}_{(mag)_{ki}}}^{(T+t)}, \boldsymbol{a}_{ki} \equiv \boldsymbol{a}_{ki}^{(T+t)}$,
- compute $\hat{\boldsymbol{y}}_{ki} := \boldsymbol{y}_{(mag)_{ki}}\hat{\boldsymbol{c}}_{ki}$ with $\hat{\boldsymbol{c}}_{ki} = phase({\boldsymbol{a}_{ki}}^\top \boldsymbol{x}_k)$ and
- compute $\widehat{\mathrm{GradU}} = \sum_{ki}(\hat{\boldsymbol{y}}_{ki} - {\boldsymbol{a}_{ki}}^\top \boldsymbol{x}_k)_t)\boldsymbol{a}_{ki}{(\boldsymbol{b}_k)_t}^\top$

11: Set $\hat{\boldsymbol{U}}^+ \leftarrow \boldsymbol{U} - (\eta/m)\widehat{\mathrm{GradU}}$
12: **Orthornormalize to get new $\boldsymbol{U}$:** Compute $\hat{\boldsymbol{U}}^+ \overset{\mathrm{QR}}{=} \boldsymbol{U}^+\boldsymbol{R}^+$. Set $\boldsymbol{U}_t \leftarrow \boldsymbol{U}^+$.
13: **end for**

and so $\mathbb{E}[\boldsymbol{a}_{ki}\boldsymbol{y}_{(mag)_{ki}}\mathbb{1}_{\boldsymbol{y}_{(mag)_{ki}} \le \sqrt{\alpha}}] = 0$ too. In fact, because of this, it is not even possible to define a different matrix $\boldsymbol{X}$ whose expected value can be shown to be close to $\boldsymbol{X}^*$. Instead, we have to use the initialization approach of [5]. This is given in line 5 of Algorithm 2. The matrix $\boldsymbol{Y}_U$ is such that its expected value is close to $\boldsymbol{X}^*\boldsymbol{X}^{*\top} + c\boldsymbol{I}$. This fact is used to argue that its top r singular vectors span a subspace that is close to that spanned by columns of $\boldsymbol{U}^*$.

Next, consider the GDmin iterations. We use the following idea to deal with the magnitude-only measurements: $\boldsymbol{y}_{(mag)_{ki}} := |\boldsymbol{y}_{ki}|$. Let $\boldsymbol{c}_{ki} := \mathrm{phase}({\boldsymbol{a}_{ki}}^\top \boldsymbol{x}_k^*)$. Then, clearly,

$$\boldsymbol{y}_{ki} = \boldsymbol{c}_{ki}\boldsymbol{y}_{(mag)_{ki}}$$

and $\boldsymbol{y}_{(mag)_{ki}} = \bar{\boldsymbol{c}}_{ki}\boldsymbol{y}_{ki}$. We do not observe $\boldsymbol{c}_{ki}$, but we can estimate it using $\boldsymbol{x}_k$ which is an estimate of $\boldsymbol{x}_k^*$. Using the estimated phase, we can get an estimate $\hat{\boldsymbol{y}}_{ki}$ of $\boldsymbol{y}_{ki}$. We replace $\nabla_U f(\boldsymbol{U}, \boldsymbol{B})$ by its estimate which uses $\hat{\boldsymbol{y}}_{ki} = \boldsymbol{y}_{(mag)_{ki}}\hat{\boldsymbol{c}}_{ki}$, with $\hat{\boldsymbol{c}}_{ki} = phase({\boldsymbol{a}_{ki}}^\top \boldsymbol{x}_k)$, to replace $\boldsymbol{y}_{ki}$. See line 10 of Algorithm 2.

Lastly, because of the magnitude-only measurements, the update step for updating $\boldsymbol{b}_k$s is no longer an LS problem. We now need to solve an r-dimensional standard PR problem: $\min_{\boldsymbol{b}} \|\boldsymbol{y}_{(mag)_k} - |\boldsymbol{A}_k\boldsymbol{U}\boldsymbol{b}|\|^2$. This can be solved using any of the order-optimal algorithms for standard PR, e.g., Truncated Wirtinger Flow (TWF) [20] or Reshaped WF (RWF) [21]. For concreteness, we assume that RWF is used. We should point out here that we only need to run $T_{RWF,t}$ iterations of RWF at outer loop iteration t, with $T_{RWF,t}$ set below in our theorem (we set this to ensure that the error level of this step is of order δ_t). The entire algorithm, AltGD-Min-LRPR, is summarized in Algorithm 2.

B. Main Result

We can prove the following result with simple changes to the proof of Theorem 2.1.

Theorem 5.1. *Consider Algorithm 2. Set $\eta = c/{\sigma_{\max}^*}^2$, $\tilde{C} = 9\kappa^2\mu^2$, $T = C\kappa^2\log(1/\epsilon)$, and $T_{RWF,t} = C(t + c\log r)$. Assume that Assumption 1.1 holds. If*

$$mq \ge C\kappa^6\mu^2(n+q)r^2(r + \log(1/\epsilon)\log\kappa)$$

and $m \ge C\max(\log q, \log n)\log(1/\epsilon)$, then, w.p. $1 - n^{-10}$, $\mathrm{SD}(\boldsymbol{U}^*, \boldsymbol{U}_T) \le \epsilon$, $\mathrm{dist}((\boldsymbol{x}_k)_T, \boldsymbol{x}_k^*) \le \epsilon\|\boldsymbol{x}_k^*\|$ for all $k \in [q]$, and $\sum_k \mathrm{dist}^2((\boldsymbol{x}_k)_T, \boldsymbol{x}_k^*) \le \epsilon^2{\sigma_{\max}^*}^2$.

We prove this result in Sec. V-C. Notice the $\log(1/\epsilon)$ in the sample complexity of Theorem 2.1 is now replaced by $(r + \log(1/\epsilon))$. The reason is because of the different initialization approach which needs nr^3 samples instead of nr^2. This is needed because PR is a more difficult problem: we cannot define a matrix $\boldsymbol{X}_0$ for it for which $\mathbb{E}[\boldsymbol{X}_0]$ is close to $\boldsymbol{X}^*$.

Observe that AltGD-Min-LRPR has the same sample complexity as that for the AltMin solution from [6]. But its time complexity is better by a factor of $\log(1/\epsilon)$ making it the fastest solution for LRPR. Also, we should mention here that, for solutions to the two related problems – sparse PR (phaseless but global measurements) and LRMC (linear but non-global measurements) – that have been extensively studied for nearly a decade, the best sample complexity guarantees for iterative (and hence fast) algorithms are sub-optimal. The best sparse PR guarantee [31] requires m to be of order s^2 for the initialization step. Here s is the sparsity level. LRPR has both phaseless and non-global measurements. This is why its initialization step needs two extra factors of r compared to the optimal. Once initialized close enough to the true solution, it is well known that a PR problem behaves like a linear one. This is true for AltGD-Min-LRPR too.

Consider a comparison with use of a standard PR approach to recover each column of $\boldsymbol{X}^*$ individually. If TWF [20] or RWF [21] were used for this, this would require $m \gtrsim n$. In comparison, ignoring log factors, our solution for LRPR needs $m \gtrsim (n/q)r^3$. Thus, the use of altGD-min is a better idea when the rank, r, of the matrix $\boldsymbol{X}^*$ is small enough so that $q \gtrsim r^3$.

C. Proof of Theorem 5.1

For the initialization, we use the bound from [5].

Lemma 5.2 ([5]). *Let $\mathrm{SD}_2(\boldsymbol{U}_0, \boldsymbol{U}^*) = \|(\boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\boldsymbol{U}_0\|$. Pick a $\delta_{\text{init}} < 0.1$. Then, w.p. at least $1 - 2\exp\left(n(\log 17) - c\frac{\delta_{\text{init}}^2 mq}{\kappa^4 r^2}\right) - 2\exp\left(-c\frac{\delta_{\text{init}}^2 mq}{\kappa^4\mu^2 r^2}\right)$,*

$$\mathrm{SD}_2(\boldsymbol{U}_0, \boldsymbol{U}^*) \le \delta_{\text{init}} \text{ and so } \mathrm{SD}(\boldsymbol{U}_0, \boldsymbol{U}^*) \le \sqrt{r}\delta_{\text{init}}.$$

For the iterations, without loss of generality, as also done in past works on PR, e.g., [30], [20], [21], [6], to make things simpler, we assume that, for each k, $\boldsymbol{x}_k^*$ is replaced by $\bar{z}\boldsymbol{x}_k^*$

where $z = \text{phase}(\langle \boldsymbol{x}_k^*, \boldsymbol{x}_k \rangle)$. With this, $\text{dist}(\boldsymbol{x}_k^*, \boldsymbol{x}_k) = \|\boldsymbol{x}_k^* - \boldsymbol{x}_k\|$.

We modify Lemma 3.4 using the following idea. Let $\boldsymbol{U} = \boldsymbol{U}_t$ and $\boldsymbol{B} = \boldsymbol{B}_t$. For LRPR, the GD step uses an approximate gradient w.r.t. the old cost function $f(\boldsymbol{U}, \boldsymbol{B})$. Let

$$\text{Err} := \widehat{\text{GradU}} - \text{GradU}.$$

Here $\widehat{\text{GradU}} = \sum_{ki}(\hat{\boldsymbol{y}}_{ki} - \boldsymbol{a}_{ki}{}^\top \boldsymbol{x}_k))\boldsymbol{a}_{ki}\boldsymbol{b}_k{}^\top$ and $\text{GradU} = \nabla_{\boldsymbol{U}} f(\boldsymbol{U}, \boldsymbol{B}) = \sum_{ki}(\boldsymbol{y}_{ki} - \boldsymbol{a}_{ki}{}^\top \boldsymbol{x}_k))\boldsymbol{a}_{ki}\boldsymbol{b}_k{}^\top$ is the same as earlier. Thus,

$$\begin{aligned}
\text{Err} &= \sum_{ki}(\hat{\boldsymbol{y}}_{ki} - \boldsymbol{y}_{ki})\boldsymbol{a}_{ki}\boldsymbol{b}_k{}^\top \\
&= \sum_{ki}(\hat{\boldsymbol{c}}_{ki} - \boldsymbol{c}_{ki})|\boldsymbol{a}_{ki}^\top \boldsymbol{x}_k^*|\boldsymbol{a}_{ki}\boldsymbol{b}_k{}^\top \\
&= \sum_{ki}(\hat{\boldsymbol{c}}_{ki}\bar{\boldsymbol{c}}_{ki} - 1)(\boldsymbol{a}_{ki}^\top \boldsymbol{x}_k^*)\boldsymbol{a}_{ki}\boldsymbol{b}_k{}^\top
\end{aligned}$$

Proceeding as in the proof of Lemma 3.4, and using $\|(\boldsymbol{I} - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\text{Err}\|_F \le \|\text{Err}\|_F$ and $\|\text{Err}\| \le \|\text{Err}\|_F$, we can conclude the following

$$\text{SD}(\boldsymbol{U}^*, \boldsymbol{U}^+) \le \frac{\|(I - (\eta/m)\text{Hess}\| \cdot \text{SD}(\boldsymbol{U}^*, \boldsymbol{U}) + (\eta/m)\|\text{Term2}\|_F + (\eta/m)\|\text{Err}\|_F}{1 - (\eta/m)\|\text{GradU}\| - (\eta/m)\|\text{Err}\|_F}$$

where the expressions for $\text{GradU}, \text{Term2}, \text{Hess}$ are the same as before with one change: $\boldsymbol{b}_k$ is now obtained by solving a noisy r-dimensional PR problem (instead of a LS problem) using RWF [21]. Thus, to complete the proof, (i) we need to bound

$$\|\text{Err}\|_F = \max_{\boldsymbol{W} \in \mathcal{S}_{nr}} \sum_{ki}(\hat{\boldsymbol{c}}_{ki}\bar{\boldsymbol{c}}_{ki} - 1)(\boldsymbol{a}_{ki}^\top \boldsymbol{x}_k^*)(\boldsymbol{a}_{ki}^\top \boldsymbol{W}\boldsymbol{b}_k)$$

and (ii) we need bounds on the three other terms that were also bounded earlier for the linear case.

The term $\|\text{Err}\|_F$, is bounded in Lemma 4 of [6] . We repeat the lemma below.

Lemma 5.3. *Assume that* $\text{SD}(\boldsymbol{U}_t, \boldsymbol{U}^*) \le \delta_t$ *with* $\delta_t < c/\kappa^2$. *Then, w.p. at least* $1 - 2\exp\left(nr\log(17) - c\frac{mq\epsilon_2^2}{\mu^2\kappa r}\right) - \exp(\log q + r - cm)$,

$$\|\text{Err}\|_F \le Cm(\epsilon_2 + \sqrt{\delta_t})\delta_t {\sigma_{\max}^*}^2$$

Consider the other three terms: $\text{GradU}, \text{Term2}, \text{Hess}$. These were bounded in Lemma 3.5 for the linear case. The statement and proof of this lemma remain the same as earlier because of the following reason. Its proof uses the bounds on $\boldsymbol{b}_k$, $\boldsymbol{x}_k$ from Lemma 3.3. The statement of this lemma also remains the same with one change: we replace $\|\boldsymbol{x}^* - \boldsymbol{x}\|$ by $\text{dist}(\boldsymbol{x}^*, \boldsymbol{x})$ and $\|\boldsymbol{X}^* - \boldsymbol{X}\|_F^2$ by $\sum_{k=1}^q \text{dist}^2(\boldsymbol{x}_k^*, \boldsymbol{x}_k)$, and the same for $\boldsymbol{b}_k^*, \boldsymbol{g}_k$. The first part of Lemma 3.3 now follows by the first part of [6, Lemma 3.3]. All the subparts of the second part of Lemma 3.3 follow exactly as given in its proof in Sec. IV-D.

VI. Limitations of our results

Our results have three limitations: (i) the algorithm that is analyzed needs sample-splitting, even though, in numerical experiments this is not needed; (ii) our bound holds w.h.p. for a single matrix $\boldsymbol{X}^*$ satisfying Assumption 1.1 (and not for all such matrices); and (iii) for obtaining exactly zero error, we need an infinite number of samples. We explain here the reasons why we are unable to address these issues. We should mention here that, since all computers are finite precision, (iii) is entirely a theoretical curiosity. Also, many other results in the LR recovery literature, e.g., [2], [14], [15], also have all these limitations.

A. Need for sample-splitting

In Algorithm 1, sample-splitting (line 3) helps ensure that the measurement matrices in each iteration for updating each of $\boldsymbol{U}$ and $\boldsymbol{B}$ are independent of all previous iterates: we split our sample set into $2T + 1$ subsets, we use one subset for initialization of $\boldsymbol{U}$ and one subset each for T iterations of updating $\boldsymbol{B}$ and updating $\boldsymbol{U}$. This helps prove the desired error decay bound by applying the sub-exponential Bernstein inequality [26] which requires the summands to be mutually independent. This becomes true in our case because, conditioned on past measurement matrices, the current set of $\boldsymbol{a}_{ki}$'s are independent of the last updated values of $\boldsymbol{U}, \boldsymbol{B}$; and the $\boldsymbol{a}_{ki}$s for different (i, k) are mutually independent by definition. Thus, under the conditioning, the summands are mutually independent. Since we prove convergence in order $\log(1/\epsilon)$ iterations, this only adds a multiplicative factor of $\log(1/\epsilon)$ in the sample complexity. Sample-splitting and the above overall idea is a standard approach used in many older works; in fact it is assumed for most of the LRMC guarantees for solutions that do not solve a convex relaxation (are iterative algorithms) [2], [14], [15]. An exception is [16].

There are a few commonly used approaches to avoid sample splitting. (1) One is using the leave-one-out strategy as done in [19]. But this means that the sample complexity dependence on r worsens: the LRMC sample complexity with this approach is $(n+q)r^3$ times log factors. Also, it is not clear how to develop this approach for alternating $\boldsymbol{U}, \boldsymbol{B}$ updates. (2) The second is to try to prove error decay for all matrices that are close enough to the true $\boldsymbol{X}^*$ and that satisfy the other assumptions of the guarantee. There are at least two different approaches to doing this. (2a) The first, which was used in [16], works for LRMC since its measurements are bounded and symmetric: the authors are able to utilize i.i.d. Bernoulli sampling and left and right singular vectors' incoherence to prove key probabilistic bounds for all matrices of the form $\boldsymbol{UV}$ with $\boldsymbol{U}, \boldsymbol{V}$ both being incoherent. This does not work in our case because our measurements are asymmetric and unbounded (which means for example that $\boldsymbol{y}_{ki}$ times its estimate is heavier-tailed than $\boldsymbol{y}_{ki}$).

(2b) An alternative approach is the following overall idea, which has been successfully used for analyzing standard PR algorithms, e.g., see [20], [21], but does not always work for other problems. In our setting, this means the following: At iteration $t + 1$, suppose that the previous estimate $\boldsymbol{U}_t$ satisfies

$\mathrm{SD}(\boldsymbol{U}_t, \boldsymbol{U}^*) \le \delta_t$. We need to try to show that, for all $\boldsymbol{U}$ that are a subspace distance δ_t away from the true subspace, the next iterate (which is a function of $\boldsymbol{U}$ and of the current $\boldsymbol{A}_k, \boldsymbol{y}_k$ for all k) is a distance $c\delta_t$ away with a $c < 1$. To be precise, for all $\boldsymbol{U} \in \mathcal{T} := \{\boldsymbol{U} : \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$ and $\mathrm{SD}(\boldsymbol{U}, \boldsymbol{U}^*) \le \delta_t\}$, we need $\boldsymbol{U}^+(\boldsymbol{U}) = orth(\boldsymbol{U} - \eta \nabla_U f(\boldsymbol{U}, \boldsymbol{B}))$ to satisfy $\mathrm{SD}(\boldsymbol{U}^+, \boldsymbol{U}^*) \le c\delta_t$ for a $c < 1$. Here $orth(\boldsymbol{M})$ is a matrix with orthonormal columns spanning the same subspace as those of $\boldsymbol{M}$. Also recall that the columns of $\boldsymbol{B}$ are $\boldsymbol{b}_k := (\boldsymbol{A}_k \boldsymbol{U})^\dagger \boldsymbol{y}_k$ for all $k \in [q]$. One can show this for all $\boldsymbol{U} \in \mathcal{T}$ by covering $\mathcal{T}$ by a net containing a finite number of points that are such that any point in $\mathcal{T}$ is with a subspace distance $0.25\delta_t$ of some point in the net, and first proving that this bound holds for all $\boldsymbol{U}$ in the net. The first step for proving such a bound is to bound the error in the estimates $\boldsymbol{b}_k$ for all $\boldsymbol{U}$ in this net. Because of the decoupled column-wise recovery of the $\boldsymbol{b}_k$'s, for *one* $\boldsymbol{U}$ in this net, the bound on $\|\boldsymbol{b}_k(\boldsymbol{U}) - \boldsymbol{U}^\top \boldsymbol{x}_k^*\|$ holds w.p. $\ge 1 - q\exp(r - cm)$. This is proved in Lemma 3.3. If we want this bound to hold for all $\boldsymbol{U}$'s in the net covering $\mathcal{T}$, we will need a union bound over all points in the net. The smallest sized net to cover $\mathcal{T}$ with accuracy $\epsilon_{net} = 0.25\delta_t$ has size upper bounded by C^{nr} [26]. With using this, the probability lower bound becomes $1 - \exp(nr + \log q + r - cm)$. For this to even just be non-negative, we need $m > Cnr$ which is too large and makes our guarantee useless.

B. Why we cannot prove our result for all X^*

The inability to obtain a useful union bound over a net of size C^{nr} explained above is also why we cannot do this.

C. Why sample complexity depends on the desired final accuracy ϵ

Observe from our result that the number of samples required to achieve a certain accuracy ϵ grows as $\log(1/\epsilon)$. This means that, for the algorithm to achieve zero error, we need an infinite number of samples. We should mention that this problem is not unique to our result. It is often seen for results that use sample-splitting, e.g., [2], [15]. An exception is [14] for LRMC, where the following basic idea is used: one tries to show that after enough iterations, e.g., when the recovery error is $\epsilon_0 = 1/n$ or smaller, one can start reusing the same samples and still prove error decay. This is also the idea used in [19]. Briefly, the reason we are unable to circumvent this problem using a similar idea to that of [14] is that our algorithm is not a regular GD or projected GD method.

To use a similar idea in our setting, we would need to proceed as follows. We use independent samples until the error is below an ϵ_0 that is small enough. Pick $\epsilon_0 = 1/(\kappa^2 n^2)$. This happens after $T(\epsilon_0) = C\kappa^2 \log(n)\log(\kappa)$ iterations. Consider $t = T+1$. At this time, $\delta_t = \epsilon_0 = 1/(\kappa^2 n^2)$. Thus, by Lemma 3.3, $\|\boldsymbol{b}_k - \boldsymbol{U}^\top \boldsymbol{x}_k^*\| \lesssim (1/(\kappa^2 n^2))\|\boldsymbol{x}_k^*\|$ and all the other bounds also hold with δ_t replaced by ϵ_0. We try to show error decay by applying Lemma 3.4. For this to work, we need to be able to show all of the following without using independence between $\boldsymbol{U}, \boldsymbol{B}$ and the $\boldsymbol{A}_k$s: (i) upper and lower bound the eigenvalues of $\mathrm{Hess} = \sum_{ki} (\boldsymbol{a}_{ki} \otimes \boldsymbol{b}_k)(.)^\top$ as those proved earlier, (ii) bound $\|\nabla_{\boldsymbol{U}} f(\boldsymbol{U}, \boldsymbol{B})\|/m$ by $c_0 {\sigma_{\min}^*}^2$ for a small constant $c_0 < 1$ (in fact even in our main proof, such a bound is sufficient since this term only appears in the denominator), and (iii) bound $\|\mathrm{Term2}\|_F/m$ by $(c_2/\kappa^2)\delta_t {\sigma_{\max}^*}^2$ with a c_2 sufficiently less than one.

As we explain next, (i) and (ii) can be obtained easily, but (iii) cannot. We can obtain (i) by showing that Hess is close to $\mathrm{Hess}^* = \sum_{ki} (\boldsymbol{a}_{ki} \otimes (\boldsymbol{U}^\top \boldsymbol{U}^* \boldsymbol{b}_k^*))(.)^\top$; and Hess^* can be bounded almost exactly as done in our proof earlier since $\boldsymbol{A}_k$s are independent of $\boldsymbol{x}_k^*$s. The $\boldsymbol{U}$ in the expression for Hess^* does not matter because $\boldsymbol{U}^\top \boldsymbol{U}^*$ is an $r \times r$ rotation matrix and one can take a maximum over all rotation matrices. Using the loose bounds $\|\boldsymbol{a}_{ki}\| \le 5\sqrt{n}$ w.h.p., one can show that $\|\mathrm{Hess}^* - \mathrm{Hess}\| \le mq \max_{ki}[\max_{\boldsymbol{W} \in \mathcal{S}_{nr}} |\boldsymbol{a}_{ki}^\top \boldsymbol{W} \boldsymbol{g}_k| \cdot \max_{\boldsymbol{W} \in \mathcal{S}_{nr}} |\boldsymbol{a}_{ki}^\top \boldsymbol{W}(\boldsymbol{g}_k - \boldsymbol{b}_k)|] \lesssim mq\sqrt{n}\mu\sqrt{r/q}\sigma_{\max}^* \cdot \sqrt{n}\epsilon_0 \mu\sqrt{r/q}\sigma_{\max}^* \le m\mu^2(r/n){\sigma_{\min}^*}^2$. Similarly, for (ii), $\sum_{ki} \|\boldsymbol{a}_{ki}\boldsymbol{a}_{ki}^\top(\boldsymbol{x}_k^* - \boldsymbol{x}_k)\boldsymbol{b}_k^\top\| \lesssim mq \cdot \sqrt{n} \cdot \sqrt{n} \cdot \epsilon_0 \cdot (\mu^2 r/q){\sigma_{\max}^*}^2 = m(\mu^2 r/n){\sigma_{\min}^*}^2$. Using $(\mu^2 r/n) \ll 1$, claims (i) and (ii) follow. However, proving (iii) seems to be impossible without using the fact that $\mathbb{E}[\mathrm{Term2}] = 0$. But this expected value is zero only when $\boldsymbol{A}_k$s are independent of $\boldsymbol{U}, \boldsymbol{B}$.

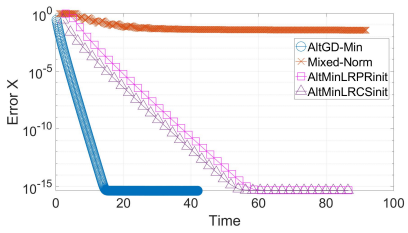
Possible ways to prove (iii). For bounding Term2 for times $t > T(\epsilon_0)$, we can try one of the following ideas. (1) Try to use Cauchy-Schwarz in a way that the projection orthogonal to $\boldsymbol{U}^*$ is used. There does not seem to be a way to make this work. (2) Try to use the leave-one-out strategy of [19] only for $t > T(\epsilon_0)$.

VII. Numerical Experiments

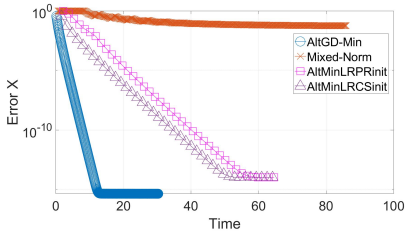
Our first experiment compares AltGD-Min with the mixed norm minimization solution from [7] (mixed-norm-min) and with the AltMin algorithm [4], [5], [6] modified for the linear LRcCS problem (replace the PR step for updating $\boldsymbol{b}_k$'s by a simple LS step). We implement this with using two possible initializations: the initialization developed in [4], [5], [6] for LRPR (AltMinLin-LRPRinit), and with the initialization approach developed in this work (AltMinLin-LRCSinit). For mixed norm min, we used the code downloaded from https://www.dropbox.com/sh/lywtzc0y9awpvgz/AABbjuiuLWPy_8y7C3GQKo8pa?dl=0, which is provided by the authors. For AltMin, we used the code from https://github.com/praneethmurthy/. We implemented AltGD-Min with $\eta = 0.4/\|\boldsymbol{X}_0\|^2$ and $\tilde{C} = 9$. Also, we used *one* set of measurements for all its iterations.

For chosen values of n, q, r and m, we simulated the data as follows. We simulated $\boldsymbol{U}^*$ by orthogonalizing an $n \times r$ standard Gaussian matrix; and $\boldsymbol{b}_k^*$s were generated i.i.d. from $\mathcal{N}(0, \boldsymbol{I}_r)$. These were generated once. For each of 100 Monte Carlo runs, the measurement matrices $\boldsymbol{A}_k$ contained i.i.d. standard Gaussian entries. We obtained $\boldsymbol{y}_k = \boldsymbol{A}_k \boldsymbol{U}^* \boldsymbol{b}_k^*$, $k \in [q]$. For the LRPR experiment, we used $\boldsymbol{y}_{(mag)_k} = |\boldsymbol{y}_k|$ as the measurements. We plot the empirical average of $\|\boldsymbol{X} - \boldsymbol{X}^*\|_F / \|\boldsymbol{X}^*\|_F$ at each iteration t on the y-axis (labeled "Error-X" in the plots) and the time taken by the algorithm until iteration t on the x-axis.

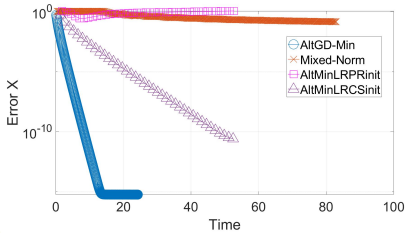
For our first experiment, shown in Fig. 1a, we used $n = 600, q = 600, r = 4$ and $m = 80$. In this case, mixed-norm-min error decays to about 2-5% but does not reduce any further.



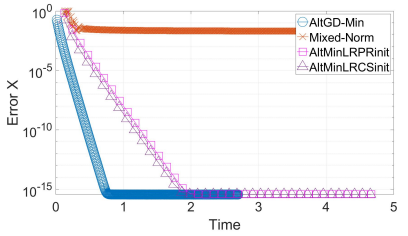
(a) $m = 80,\ n = q = 600, r = 4$



(b) $m = 50,\ n = q = 600, r = 4$



(c) $m = 30,\ n = q = 600, r = 4$



(d) $m = 90,\ n = 100, q = 120, r = 2$

Fig. 1: *Comparing the proposed algorithm with existing approaches for solving LRcCS.*

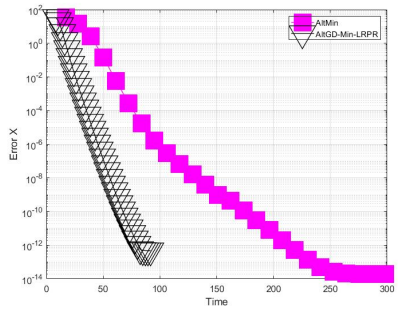


Fig. 2: *Comparing the proposed algorithm with existing approach for solving LRPR. We used* $n = 600, q = 1000, r = 4$ *and* $m = 250$.

But, for our algorithm, AltGD-Min, and for both versions of AltMin, the error decays to 10^{-15}. Notice also that AltGD-Min is much faster than all the other approaches. Fig. 1b reduced m to $m = 50$. Here a similar trend is observed, except that the error decays to only around 10^{-13} for AltGD-Min and 10^{-11} for the two AltMin approaches. Finally, for Fig. 1c, we reduced m to $m = 30$. In this case, only AltGDmin and AltMin-LRCSinit work, while mixed-norm-min and AltMin-LRPRinit errors do not decrease at all. The reason is both these need a higher sample complexity (see Table I). Finally, we also tried an experiment with very large m: $n = 100, q = 120, r = 2$ and $m = 0.9n = 90$, see Fig. 1d. Even for such a large value of m (compared to n), observe that the mixed-norm-min error saturates at around 1-2%. The likely reason for this that, in the guarantee for mixed-norm-min [7] (summarized for the noiseless case in Proposition 2.3 given earlier), even for $m = n$, the error is bounded by a multiplier (more than 1) times $\sqrt{r/q}$.

For the comparisons for the LRPR problem shown in Fig. 2, we need a much larger q and m since LRPR requires mq to scale as nr^3 both for initialization and for the GDmin iterations and the multiplying constants are also much larger for LRPR. We used $n = 600, q = 1000, r = 4$ and $m = 250$. Notice that altGD-Min-LRPR is faster than AltMin-LRPR. We implemented altGD-Min-LRPR with $\eta = 0.9/\|\boldsymbol{X}_0\|^2$, $\tilde{C} = 9$, and $T_{RWF,t} = \max(5+t, 40)$ in the RWF code (code for [21], downloaded from the specified site). Also, here again, we used *one* set of measurements for all its iterations.

VIII. Conclusions

This work developed a sample-efficient and fast gradient descent (GD) solution, called AltGD-Min, for provably recovering a low-rank (LR) matrix from mutually independent column-wise linear projections. This problem, which we refer to as "Low Rank column-wise Compressive Sensing (LRcCS)", frequently occurs in LR-based accelerated low rank dynamic MRI and in federated sketching. If used in a federated setting, AltGD-Min is also communication-efficient. The LRcCS problem has not received little attention in the theoretical literature unlike the other well-studied LR recovery problems (matrix completion, sensing, or multivariate regression).

References

[1] S. Negahban, M. J. Wainwright, *et al.*, "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *The Annals of Statistics*, vol. 39, no. 2, pp. 1069–1097, 2011.

[2] P. Netrapalli, P. Jain, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Annual ACM Symp. on Th. of Comp. (STOC)*, 2013.

[3] E. J. Candes and B. Recht, "Exact matrix completion via convex optimization," *Found. of Comput. Math*, no. 9, pp. 717–772, 2008.

[4] S. Nayer, P. Narayanamurthy, and N. Vaswani, "Phaseless PCA: Low-rank matrix recovery from column-wise phaseless measurements," in *Intl. Conf. Machine Learning (ICML)*, 2019.

[5] ——, "Provable low rank phase retrieval," *IEEE Trans. Info. Th.*, March 2020.

[6] S. Nayer and N. Vaswani, "Sample-efficient low rank phase retrieval," *IEEE Trans. Info. Th.*, 2021.

[7] R. S. Srinivasa, K. Lee, M. Junge, and J. Romberg, "Decentralized sketching of low rank matrices," in *Neur. Info. Proc. Sys. (NeurIPS)*, 2019, pp. 10 101–10 110.

[8] Z.-P. Liang, "Spatiotemporal imaging with partially separable functions," in *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2007, pp. 988–991.

[9] S. G. Lingala, Y. Hu, E. DiBella, and M. Jacob, "Accelerated dynamic mri exploiting sparsity and low-rank structure: kt slr," *IEEE Transactions on Medical Imaging*, vol. 30, no. 5, pp. 1042–1054, 2011.

[10] J. Yao, Z. Xu, X. Huang, and J. Huang, "An efficient algorithm for dynamic mri using low-rank and total variation regularizations," *Medical Image Analysis*, vol. 44, pp. 14–27, 2018.

[11] F. P. Anaraki and S. Hughes, "Memory and computation efficient pca via very sparse random projections," in *Intl. Conf. Machine Learning (ICML)*, 2014, pp. 1341–1349.

[12] A. Krishnamurthy, M. Azizyan, and A. Singh, "Subspace learning from extremely compressed measurements," *arXiv preprint arXiv:1404.0751*, 2014.

[13] S. Babu, S. Nayer, S. G. Lingala, and N. Vaswani, "Fast low rank compressive sensing for accelerated dynamic mri," in *IEEE Intl. Conf. Acoustics, Speech, Sig. Proc. (ICASSP)*, 2022, to appear.

[14] P. Jain and P. Netrapalli, "Fast exact matrix completion with finite samples," in *Conf. on Learning Theory*, 2015, pp. 1007–1034.

[15] Y. Cherapanamjeri, K. Gupta, and P. Jain, "Nearly-optimal robust matrix completion," *ICML*, 2016.

[16] X. Yi, D. Park, Y. Chen, and C. Caramanis, "Fast algorithms for robust pca via gradient descent," in *Neur. Info. Proc. Sys. (NeurIPS)*, 2016.

[17] Q. Zheng and J. Lafferty, "Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent," *arXiv preprint arXiv:1605.07051*, 2016.

[18] S. Lang, *Real and Functional Analysis*. Springer-Verlag, New York 10:11–13, 1993.

[19] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution," in *Intl. Conf. Machine Learning (ICML)*, 2018.

[20] Y. Chen and E. Candes, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," in *Neur. Info. Proc. Sys. (NeurIPS)*, 2015, pp. 739–747.

[21] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi, "A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 5164–5198, 2017.

[22] G. Jagatap, Z. Chen, S. Nayer, C. Hegde, and N. Vaswani, "Sample efficient fourier ptychography for structured data," *IEEE Trans. Comput. Imaging*, vol. 6, pp. 344–357, 2020.

[23] Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and stable covariance estimation from quadratic sampling via convex programming," *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 4034–4059, 2015.

[24] G. H. Golub and C. F. Van Loan, "Matrix computations," *The Johns Hopkins University Press, Baltimore, USA*, 1989.

[25] M. Hardt and E. Price, "The noisy power method: A meta algorithm with applications," in *Neur. Info. Proc. Sys. (NeurIPS)*, 2014, pp. 2861–2869.

[26] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018, vol. 47.

[27] P.-Å. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT Numerical Mathematics*, vol. 12, no. 1, pp. 99–111, 1972.

[28] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Spectral methods for data science: A statistical perspective," *arXiv preprint arXiv:2012.08496*, 2020.

[29] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*. Cambridge Univ. Press, Cambridge, 2012.

[30] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Neur. Info. Proc. Sys. (NeurIPS)*, 2013, pp. 2796–2804.

[31] T. Cai, X. Li, and Z. Ma, "Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow," *The Annals of Statistics*, vol. 44, no. 5, pp. 2221–2251, 2016.

[32] L. Erdos, A. Knowles, H. Yau, and J. Yin, "Spectral statistics of erdos–rényi graphs i: Local semicircle law," *The Annals of Probability*, vol. 41, no. 3B, pp. 2279–2375, 2013.

[33] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Found. Comput. Math.*, vol. 12, no. 4, 2012.

Author Biographies

Seyedehsara Nayer (Email: sarana@iastate.edu) recently completed her Ph.D. in ECE at Iowa State University. She has an M.S. from Sharif University in Iran. She works as a Senior Engineer at ASML in Santa Clara, CA. Her research interests are around various aspects of information science and focuses on Signal Processing, and Statistical Machine Learning.

Namrata Vaswani (Email: namrata@iastate.edu) received a B.Tech from IIT-Delhi in India in 1999 and a Ph.D. from the University of Maryland, College Park in 2004, both in Electrical Engineering. Since Fall 2005, she has been with the Iowa State University where she is currently the Anderlik Professor of Electrical and Computer Engineering. Her research interests lie in a data science, with a particular focus on Statistical Machine Learning and Signal Processing. She has served two terms as an Associate Editor for the IEEE Transactions on Signal Processing; as a lead guest-editor for a 2018 Proceedings of the IEEE Special Issue (Rethinking PCA for modern datasets); and as an Area Editor for the IEEE Signal Processing Magazine (2018-2020). Vaswani is a recipient of the Iowa State Early Career Engineering Faculty Research Award (2014), the Iowa State University Mid-Career Achievement in Research Award (2019) and University of Maryland's ECE Distinguished Alumni Award (2019). She also received the 2014 IEEE Signal Processing Society Best Paper Award for her 2010 IEEE Transactions on Signal Processing paper co-authored with her student Wei Lu on "Modified-CS: Modifying compressive sensing for problems with partially known support". She is a Fellow of the IEEE Fellow (class of 2019).

Appendix A
Understanding why LRMC-style GD approaches cannot be easily analyzed for LRcCS

A. Gradient Descent

The iterates of a gradient descent (GD) algorithm converge when the gradient approaches zero. Thus, in order to show its convergence, one needs to be able to bound the norm of the gradient and show that it goes to zero with iterations. In order to show fast enough convergence (reach ϵ error in order $\log(1/\epsilon)$ iterations), one further needs to show that this bound on the gradient norm decreases sufficiently with each iteration. Consider projGD-X which was studied in [15] for solving LRMC. ProjGD-X iterations involve computing $\boldsymbol{X}^+ \leftarrow \mathcal{P}_r(\boldsymbol{X} - \nabla_{\boldsymbol{X}} \tilde{f}(\boldsymbol{X}))$, here $\mathcal{P}_r(\boldsymbol{M})$ projects its argument onto the space of rank-r matrices. To bound $\|\nabla_{\boldsymbol{X}} \tilde{f}(\boldsymbol{X})\|$, we need to bound $|\boldsymbol{w}^\top \nabla_{\boldsymbol{X}} \tilde{f}(\boldsymbol{X}) \boldsymbol{z}|$ for any unit norm vectors $\boldsymbol{w}, \boldsymbol{z}$. We show the cost function $\tilde{f}(\boldsymbol{X})$ and its gradient for both LRMC and LRcCS in Table II. Observe that, for LRcCS, $\boldsymbol{w}^\top \nabla_{\boldsymbol{X}} \tilde{f}(\boldsymbol{X}) \boldsymbol{z}$ is a sum of sub-exponential r.v.s with sub-exponential norms bounded by $K_e = \max_k \|\boldsymbol{w}\| \cdot \|\boldsymbol{x}_k^* - \boldsymbol{x}_k\| \cdot |\boldsymbol{z}_k| \le \max_k \|\boldsymbol{x}_k^* - \boldsymbol{x}_k\|$. Thus, in order to get a small enough bound on $|\boldsymbol{w}^\top \nabla_{\boldsymbol{X}} \tilde{f}(\boldsymbol{X}) \boldsymbol{z}|$ by applying the sub-exponential Bernstein inequality [26], we need a small enough bound on $\max_k \|\boldsymbol{x}_k^* - \boldsymbol{x}_k\|$ (column-wise error bound). It is not clear how to get this because the projection step introduces coupling between the different columns of the estimated matrix $\boldsymbol{X}$ [2].

[2]Let $\boldsymbol{H} := \boldsymbol{X} - \boldsymbol{X}^*$, $\tilde{\boldsymbol{H}} := (\boldsymbol{X} - \eta \nabla f(\boldsymbol{X})) - \boldsymbol{X}^* = \boldsymbol{H} - \eta \nabla f(\boldsymbol{X})$, and $\boldsymbol{H}^+ = \boldsymbol{X}^+ - \boldsymbol{X}^* = \mathcal{P}_r(\boldsymbol{X} - \nabla f(\boldsymbol{X})) - \boldsymbol{X}^* = \mathcal{P}_r(\boldsymbol{X}^* + \tilde{\boldsymbol{H}}) - \boldsymbol{X}^*$. To bound the LRMC projGD-X errors, one needs an entry-wise bound of the form $\|\boldsymbol{H}^+\|_{\max} \le \delta_t \|\boldsymbol{X}^*\|_{\max}$ with δ_t decaying exponentially. We show the expressions for $\tilde{\boldsymbol{H}}$ in the table. For LRMC, notice that different summands of $\tilde{\boldsymbol{H}}$ are mutually independent and each depends on only one entry of $\boldsymbol{H}$. This fact is carefully exploited in [15, Lemma 1] and [14, Lemma 1]. By borrowing ideas from the literature on spectral statistics of Erdos-Renyi graphs [32], the authors are able to obtain expressions for higher powers of $(\tilde{\boldsymbol{H}}\tilde{\boldsymbol{H}}^\top)$. These expressions help them get the desired bound under the desired sample complexity. For LRcCS, using the gradient expression, we need a bound on $\max_k \|\boldsymbol{h}_k^+\|$ in terms of $\|\boldsymbol{h}_k\|$ in order to show its exponential decay. Since the different entries of $\tilde{\boldsymbol{H}}$ are not mutually independent and not bounded, the LRMC proof approach cannot be borrowed.

TABLE II: Understanding why LRMC style projected-GD on $\boldsymbol{X}$ does not work in our case.

	LRMC	Our Problem, LRcCS
$\tilde{f}(\boldsymbol{X})$	$\sum_{k=1}^{q}\sum_{j=1}^{n}(\boldsymbol{y}_{jk}-\delta_{jk}\boldsymbol{X}_{jk})^2$	$\sum_{k=1}^{q}\sum_{i=1}^{m}(\boldsymbol{y}_{ki}-\boldsymbol{a}_{ki}^\top\boldsymbol{x}_k)^2$
	$\delta_{jk}\overset{\text{iid}}{\sim} Bernoulli(p)$	$\boldsymbol{a}_{ki}\overset{\text{iid}}{\sim}\mathcal{N}(0,\boldsymbol{I}_n)$
$\nabla_X\tilde{f}(\boldsymbol{X})$	$\sum_{k=1}^{q}\sum_{j=1}^{n}\delta_{jk}(\boldsymbol{y}_{jk}-\delta_{jk}\boldsymbol{X}_{jk})\boldsymbol{e}_j\boldsymbol{e}_k^\top$	$\sum_{k=1}^{q}\sum_{i=1}^{m}(\boldsymbol{y}_{ki}-\boldsymbol{a}_{ki}^\top\boldsymbol{x}_k)\boldsymbol{a}_{ki}\boldsymbol{e}_k^\top$
	$=\sum_{k=1}^{q}\sum_{j=1}^{n}\delta_{jk}(\boldsymbol{X}^*_{jk}-\boldsymbol{X}_{jk})\boldsymbol{e}_j\boldsymbol{e}_k^\top$	$=\sum_{k=1}^{q}\sum_{i=1}^{m}\boldsymbol{a}_{ki}^\top(\boldsymbol{x}_k^*-\boldsymbol{x}_k)\boldsymbol{a}_{ki}\boldsymbol{e}_k^\top$
$\tilde{\boldsymbol{H}}:=\boldsymbol{H}-\eta\nabla f(\boldsymbol{X})$	$\sum_{k=1}^{q}\sum_{j=1}^{n}(1-\frac{\delta_{jk}}{p})\boldsymbol{H}_{jk}\boldsymbol{e}_j{\boldsymbol{e}_k}^\top$	$\frac{1}{m}\sum_{k=1}^{q}\sum_{i=1}^{m}(\boldsymbol{I}-\boldsymbol{a}_{ki}{\boldsymbol{a}_{ki}}^\top){\boldsymbol{h}_k}{\boldsymbol{e}_k}^\top$

Moreover, even if we could somehow get such a bound, in the best case, it would be proportional to $\delta_t\max_k\|\boldsymbol{x}_k^*\|$ with $\delta_t<1$ and decaying exponentially with t. Using Assumption 1.1, this would then imply that $K_e\le\delta_t\max_k\|\boldsymbol{x}_k^*\|\le\delta_t\mu\sqrt{r/q}\sigma^*_{\max}$. But, this is not small enough. We need it to be proportional to $\delta_t(r/q)$ in order to be able to bound the gradient norm under the desired sample complexity.

Consider altGDnormbal studied in [17], [16] for LRMC. In this case again, the desired column-wise error bound cannot be obtained because the update step for $\boldsymbol{B}$ involves GD w.r.t. $f(\boldsymbol{U},\boldsymbol{B})+f_2(\boldsymbol{U},\boldsymbol{B})$. The gradient w.r.t f_2 (norm-balancing term) introduces coupling between the different columns of $\boldsymbol{B}$, and hence, also between columns of $\boldsymbol{X}=\boldsymbol{U}\boldsymbol{B}$. Thus, once again, it is not clear how to get a tight bound on $\max_k\|\boldsymbol{x}_k^*-\boldsymbol{x}_k\|$.

For AltGD-Min, because the min step for updating $\boldsymbol{B}$ is a decoupled LS problem, it is possible to get the desired column-wise error bound. Secondly, because we use GD w.r.t $\boldsymbol{U}$, there is an extra $\boldsymbol{b}_k^\top$ term in the gradient summands. This makes the gradient (and its deviation from its expected value), a sum of *nice-enough* sub-exponential r.v.s as explained in Sec. III-B.

B. Initialization

The standard approach used for initializing iterative algorithms for LRMC (as well as other linear LRR problems) is to compute the top r left singular vectors of the matrix $\boldsymbol{X}_{0,full}$ that satisfies $(\boldsymbol{X}_{0,full})_{vec}=\mathcal{A}^\top(\boldsymbol{y}_{all})$, where $\boldsymbol{y}_{all}$ is the mq-length vector of all measurements and $\mathcal{A}$ denotes the linear mapping from $(\boldsymbol{X}^*)_{vec}$ to $\boldsymbol{y}_{all}$. In case of LRMC and LRcCS, this is computed is as given in Table III. It is not hard to see that, in both cases, $\mathbb{E}[\boldsymbol{X}_{0,full}]=\boldsymbol{X}^*$. To show that this approach works, one typically uses a $\sin\Theta$ theorem, e.g., Davis-Kahan or Wedin, to bound $\mathrm{SD}(\boldsymbol{U}^*,\boldsymbol{U}_0)$ as a function of terms that depend on $\boldsymbol{H}_0:=\boldsymbol{X}_{0,full}-\boldsymbol{X}^*$. Thus a first requirement is to bound $\|\boldsymbol{H}_0\|$. For LRMC, this can be done easily since $\boldsymbol{H}_0$ is a sum of the independent one-sparse random matrices shown in the table with each matrix containing an i.i.d. Bernoulli r.v. times $\boldsymbol{X}^*_{jk}$ (jk-th entry of $\boldsymbol{X}^*$) as its nonzero entry. Using the left and right singular vectors' incoherence (assumed in all LRMC guarantees), and $\boldsymbol{X}^*_{jk}=\boldsymbol{e}_j^\top\boldsymbol{X}^*\boldsymbol{e}_k$, one can argue that, for unit vectors $\boldsymbol{w},\boldsymbol{z}$, each summand of $|\boldsymbol{w}^\top\boldsymbol{H}_0\boldsymbol{z}|$ is of order at most $(1/p)\sigma^*_{\max}r/\sqrt{nq}$. This bound, along with a bound on the "variance parameter" needed for applying matrix Bernstein [33],[26, Chap 5] helps show that $\|\boldsymbol{H}_0\|\le c\sigma^*_{\max}$ w.h.p., under the desired sample complexity bound. For LRcCS, the summands of $\boldsymbol{X}_{0,full}$, and hence of $\boldsymbol{H}_0$, are sub-exponential r.v.s. These can be bounded using the sub-exponential Bernstein inequality [26, Chap 2]. This requires a bound on the maximum sub-exponential norm of any summand. Denote this bound by K_e. In order to show that $\|\boldsymbol{H}_0\|\le c\sigma^*_{\max}$ w.h.p, under the desired sample complexity, we need K_e to be of order (r/q) or smaller. However, for our summands, we can only guarantee $K_e\le(1/m)\max_k\|\boldsymbol{x}_k^*\|\le(1/m)\mu\sqrt{r/q}\sigma^*_{\max}$. This is not small enough, i.e., the summands are not *nice-enough* subexponentials. It will require $mq\gtrsim(n+q)r\cdot\sqrt{q}$ which is too large.

Appendix B
Proof of Initialization Theorem 3.1 without sample-splitting

Consider the initialization using $\boldsymbol{X}_0$ defined in (2). We we want to bound the initialization error without sample-splitting. This means that the threshold α is not independent of the $\boldsymbol{a}_{ki},\boldsymbol{y}_{ki}$ used in the expression for $\boldsymbol{X}_0$ and thus, it is not clear how to compute its expected value even if we condition on α. However, the following slightly more complicated approach can be used. Using Fact 3.7 and Assumption 1.1, it is possible to show that $\boldsymbol{X}_0$ is close to a matrix, $\boldsymbol{X}_+(\epsilon_1)$ given next for which $\mathbb{E}[\boldsymbol{X}_+]$ is easily computed: Let

$$\alpha_+:=\tilde{C}(1+\epsilon_1)\frac{\|\boldsymbol{X}^*\|_F^2}{q}$$

and define

$$\begin{aligned}\boldsymbol{X}_+(\epsilon_1)&:=\frac{1}{m}\sum_{ki}\boldsymbol{a}_{ki}\boldsymbol{y}_{ki}{\boldsymbol{e}_k}^\top\mathbb{1}_{\{\boldsymbol{y}_{ki}^2\le\alpha_+\}}.\ \text{Then,}\\ \mathbb{E}[\boldsymbol{X}_+]&=\boldsymbol{X}^*\boldsymbol{D}(\epsilon_1),\\ \boldsymbol{D}&:=diagonal(\beta_k(\epsilon_1)),\\ \beta_k(\epsilon_1)&:=\mathbb{E}\left[\zeta^2\mathbb{1}_{\left\{\zeta^2\le\frac{\alpha_+}{\|\boldsymbol{x}_k^*\|^2}\right\}}\right]\end{aligned}\tag{20}$$

with ζ being a scalar standard Gaussian. Thus $\boldsymbol{X}_+$ is $\boldsymbol{X}_0$ with the threshold α replaced by α_+ which is deterministic. Consequently $\mathbb{E}[\boldsymbol{X}_+]$ has a similar form too and is obtained as explained in the proof of Lemma 3.6 given in Sec. IV-F.

Next, recall that $\boldsymbol{X}^*\overset{\text{SVD}}{=}\boldsymbol{U}^*\boldsymbol{\Sigma}^*\boldsymbol{V}^*$ and $\tilde{C}=9\kappa^2\mu^2$. Let $\tilde{c}=c/\tilde{C}$ for a $c<1$. Clearly, the span of the top r singular

TABLE III: Why the LRMC initialization approach cannot be directly borrowed?

	LRMC	Our Problem, LRcCS
$\boldsymbol{X}_{0,full} =$	$\sum_k \sum_j \frac{\delta_{jk}}{p} \boldsymbol{y}_{jk} \boldsymbol{e}_j \boldsymbol{e}_k{}^\top$	$\frac{1}{m} \sum_k \sum_i \boldsymbol{a}_{ki} \boldsymbol{y}_{ki} \boldsymbol{e}_k{}^\top$
	$\delta_{jk} \overset{\text{iid}}{\sim} Bernoulli(p)$	$\boldsymbol{a}_{ki} \overset{\text{iid}}{\sim} \mathcal{N}(0, \boldsymbol{I}_n)$
$\boldsymbol{H}_0 = \boldsymbol{X}_{0,full} - \boldsymbol{X}^*$	$\sum_{k=1}^{q} \sum_{j=1}^{n} (1 - \frac{\delta_{jk}}{p}) \boldsymbol{X}^*_{jk} \boldsymbol{e}_j \boldsymbol{e}_k{}^\top$	$\frac{1}{m} \sum_{k=1}^{q} \sum_{i=1}^{m} (\boldsymbol{I} - \boldsymbol{a}_{ki} \boldsymbol{a}_{ki}{}^\top) \boldsymbol{x}^*_k \boldsymbol{e}_k{}^\top$
Each summand is	nicely bounded by $\mu^2 \sigma^*_{\max}(r/\sqrt{nq})$	unbounded & sub-expo. norm** is $\mu \sigma^*_{\max} \sqrt{r/q}$ (too large, need r/q)
Concen. ineq.	Matrix Bernstein [33] gives desired sample comp.	Sub-expo Bernstein [26] does not give desired sample comp.

**: "max sub-expo. norm": max sub-exponential norm of $(\boldsymbol{a}_{ki}{}^\top \boldsymbol{w})(\boldsymbol{a}_{ki}{}^\top \boldsymbol{x}^*_k)(\boldsymbol{e}_k^\top \boldsymbol{z})$ for any unit vectors $\boldsymbol{w}, \boldsymbol{z}$.

vectors of $\mathbb{E}[\boldsymbol{X}_+] = \boldsymbol{X}^* \boldsymbol{D}$ equals $\text{span}(\boldsymbol{U}^*)$ and it is rank r matrix. Let,

$$\mathbb{E}[\boldsymbol{X}_+] = \boldsymbol{X}^* \boldsymbol{D} \overset{\text{SVD}}{=} \boldsymbol{U}^* \check{\boldsymbol{\Sigma}}^* \check{\boldsymbol{V}}$$

be its r-SVD (here $\check{\boldsymbol{V}}$ is an $r \times q$ matrix with its rows containing the r right singular vectors). We thus have

$$\begin{aligned}\sigma_r(\mathbb{E}[\boldsymbol{X}_+]) &= \sigma_{\min}(\check{\boldsymbol{\Sigma}}^*) = \sigma_{\min}(\boldsymbol{\Sigma}^* \boldsymbol{V}^* \boldsymbol{D} \check{\boldsymbol{V}}^\top) \\ &\geq \sigma_{\min}(\boldsymbol{\Sigma}^*) \sigma_{\min}(\boldsymbol{V}^*) \sigma_{\min}(\boldsymbol{D}) \sigma_{\min}(\check{\boldsymbol{V}}^\top) \\ &= \sigma^*_{\min} \cdot 1 \cdot (\min_k \beta_k) \cdot 1\end{aligned}$$

Fact 3.9 given earlier shows that $(\min_k \beta_k) \geq 0.9$ and thus,

$$\sigma_r(\mathbb{E}[\boldsymbol{X}_+]) \geq 0.9 \sigma^*_{\min}$$

Also, $\sigma_{r+1}(\mathbb{E}[\boldsymbol{X}_+]) = 0$ since it is a rank r matrix. Thus, using Wedin's $\sin\Theta$ theorem for SD (summarized in Theorem 4.1) applied with $\boldsymbol{M} \equiv \boldsymbol{X}_0$, $\boldsymbol{M}^* \equiv \mathbb{E}[\boldsymbol{X}_+]$ gives

$$\begin{aligned}&\text{SD}(\boldsymbol{U}_0, \boldsymbol{U}^*) \\ &\leq \frac{\sqrt{2} \max\left(\|(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_+])^\top \boldsymbol{U}^*\|_F, \|(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_+]) \check{\boldsymbol{V}}^\top\|_F \right)}{0.9 \sigma^*_{\min} - \|\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_+]\|}\end{aligned} \tag{21}$$

In the next three subsections, we prove a set of six lemmas that help bound the three terms in the expression above. *The main new ideas over the proof given earlier in Sec III-E, are in the proof of the first lemma, Lemma B.2 given below, and in the proof of Claim B.1 that is used in this proof.*

Claim B.1. *Let $\boldsymbol{x}^* \in \Re^n$, $\boldsymbol{z} \in \Re^n$ be two deterministic vectors and let α be a deterministic scalar. Let $\boldsymbol{a} \sim \mathcal{N}(0, \boldsymbol{I}_n)$ be a standard Gaussian vector and define $\boldsymbol{y} := \boldsymbol{a}^\top \boldsymbol{x}^*$. For an $0 < \epsilon < 1$,*

$$\mathbb{E}\left[|\boldsymbol{y}(\boldsymbol{a}^\top \boldsymbol{z})| \mathbb{1}_{\{\boldsymbol{y}^2 \in [1 \pm \epsilon] \alpha\}} \right] \leq C \epsilon \|\boldsymbol{z}\| \sqrt{\alpha}.$$

Combining Lemmas B.3 and B.2 and using Fact 3.7, and setting $\epsilon_1 = c\delta_0/\sqrt{r}\kappa$, we conclude that, w.p. at least $1 - 2\exp((n+q) - \tilde{c}\epsilon_1^2 mq) - \exp(-\tilde{c}mq\epsilon_1^2) \geq 1 - 2\exp((n+q) - \tilde{c}mq\delta_0^2/r\kappa^2) - \exp(-\tilde{c}mq\delta_0^2/r\kappa^2)$,

$$\|\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_+]\| \lesssim \epsilon_1 \|\boldsymbol{X}^*\|_F \lesssim c\delta_0 \sigma^*_{\min}$$

By combining Lemmas B.4, B.5, B.6, and B.7 and using Fact 3.7, and setting $\epsilon_1 = c\delta_0/\sqrt{r}\kappa$, we conclude that, w.p. at least $1 - 2\exp(nr - \tilde{c}mq\delta_0^2/r\kappa^2) - 2\exp(qr - \tilde{c}mq\delta_0^2/r\kappa^2) - \exp(-\tilde{c}mq\delta_0^2/r\kappa^2)$,

$$\max\left(\|(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_+])^\top \boldsymbol{U}^*\|_F, \|(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_+]) \check{\boldsymbol{V}}^\top\|_F \right) \lesssim c\delta_0 \sigma^*_{\min}$$

Plugging these into (21) proves Theorem 3.1

A. Bounding the denominator term

By triangle inequality, $\|\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_+]\| \leq \|\boldsymbol{X}_+ - \mathbb{E}[\boldsymbol{X}_+]\| + \|\boldsymbol{X}_0 - \boldsymbol{X}_+\|$. The next two lemmas bound these two terms. The lemmas assume the claim of Fact 3.7 holds, i.e., that $\frac{1}{mq} \sum_{ki} \boldsymbol{y}_{ki}^2 \in [1 \pm \epsilon_1] \tilde{C} \|\boldsymbol{X}^*\|_F^2/q$ where $\tilde{C} = 9\mu^2\kappa^2$.

Lemma B.2. *Assume that $\frac{1}{mq} \sum_{ki} \boldsymbol{y}_{ki}^2 \in [1 \pm \epsilon_1] \tilde{C} \|\boldsymbol{X}^*\|_F^2/q$ (claim of Fact 3.7 holds). Then, w.p. $1 - \exp(C(n+q) - \epsilon_1^2 mq/\mu^2\kappa^2)$,*

$$\|\boldsymbol{X}_0 - \boldsymbol{X}_+\| \leq C\epsilon_1 \mu\kappa \|\boldsymbol{X}^*\|_F.$$

Proof of Lemma B.2: We have

$$\begin{aligned}\|\boldsymbol{X}_+ - \boldsymbol{X}_0\| &= \max_{\boldsymbol{z} \in \mathcal{S}^n, \ \boldsymbol{w} \in \mathcal{S}^q} \boldsymbol{z}^\top (\boldsymbol{X}_+ - \boldsymbol{X}_0) \boldsymbol{w} \\ &= \max_{\boldsymbol{z} \in \mathcal{S}^n, \ \boldsymbol{w} \in \mathcal{S}^q} \frac{1}{m} \sum_{ki} \boldsymbol{w}(k) \boldsymbol{y}_{ki} (\boldsymbol{a}_{ki}{}^\top \boldsymbol{z}) \\ &\qquad \times \mathbb{1}_{\left\{ \frac{\tilde{C}}{mq} \sum_{ki} \boldsymbol{y}_{ki}^2 \leq \boldsymbol{y}_{ki}^2 \leq \frac{\tilde{C}(1+\epsilon_1)}{q} \|\boldsymbol{X}^*\|_F^2 \right\}}.\end{aligned}$$

For the last expression above, we have used the assumption $\sum_{ki} \boldsymbol{y}_{ki}^2/m \leq \tilde{C}(1+\epsilon_1)\|\boldsymbol{X}^*\|_F^2$. Consider the RHS for a fixed unit norm $\boldsymbol{z}$ and $\boldsymbol{w}$. The lower threshold of the indicator function is itself a r.v.. To convert it into a deterministic bound, we need the following sequence of bounding steps: To use our assumption that $\sum_{ki} \boldsymbol{y}_{ki}^2/m \geq (1-\epsilon_1)\tilde{C}\|\boldsymbol{X}^*\|_F^2$, we first need to bound the summands by their absolute values. This is done as follows:

$$\begin{aligned}|\boldsymbol{z}^\top (\boldsymbol{X}_+ - \boldsymbol{X}_0) \boldsymbol{w}| &\leq \frac{1}{m} \sum_{ki} \left| \boldsymbol{w}(k) \boldsymbol{y}_{ki} (\boldsymbol{a}_{ki}{}^\top \boldsymbol{z}) \right| \\ &\qquad \times \mathbb{1}_{\left\{ \frac{\tilde{C}}{mq} \sum_{ki} \boldsymbol{y}_{ki}^2 \leq |\boldsymbol{y}_{ki}|^2 \leq \frac{\tilde{C}(1+\epsilon_1)}{q} \|\boldsymbol{X}^*\|_F^2 \right\}}, \\ &\leq \frac{1}{m} \sum_{ki} \left| \boldsymbol{w}(k) \boldsymbol{y}_{ki} (\boldsymbol{a}_{ki}{}^\top \boldsymbol{z}) \right| \\ &\qquad \times \mathbb{1}_{\left\{ |\boldsymbol{y}_{ki}|^2 \in [1 \pm \epsilon_1] \frac{\tilde{C}}{q} \|\boldsymbol{X}^*\|_F^2 \right\}},\end{aligned}$$

where in the last line we used our assumption that $\sum_{ki} \boldsymbol{y}_{ki}^2/m \geq (1-\epsilon_1)\tilde{C}\|\boldsymbol{X}^*\|_F^2$. This final expression is

a sum of mutually independent sub-Gaussian r.v.s with sub-Gaussian norm $K_{ki} \leq C|\boldsymbol{w}(k)|\sqrt{\tilde{C}(1+\epsilon_1)}\|\boldsymbol{X}^*\|_F/\sqrt{q} \leq \sqrt{\tilde{C}}|\boldsymbol{w}(k)|\|\boldsymbol{X}^*\|_F/\sqrt{q}$. Thus, by applying the sub-Gaussian Hoeffding inequality, Theorem 2.6.2 of [26],

$$\Pr\Bigg\{\Big|\sum_{ki}\big|\boldsymbol{w}(k)\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{z})\big|\mathbb{1}_{\left\{|\boldsymbol{y}_{ki}|^2\in[1\pm\epsilon_1]\frac{\tilde{C}}{q}\|\boldsymbol{X}^*\|_F^2\right\}} - \mathbb{E}\left[\sum_{ki}\big|\boldsymbol{w}(k)\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{z})\big|\mathbb{1}_{\left\{|\boldsymbol{y}_{ki}|^2\in[1\pm\epsilon_1]\frac{\tilde{C}}{q}\|\boldsymbol{X}^*\|_F^2\right\}}\right]\Big| \geq t\Bigg\} \leq 2\exp\left[-c\frac{t^2}{\sum_{ki}K_{ki}^2}\right].$$

By setting $t=\epsilon_1 m\|\boldsymbol{X}^*\|_F$,

$$\frac{t^2}{\sum_{ki}K_{ki}^2} \geq \frac{m^2q\epsilon_1^2\|\boldsymbol{X}^*\|_F^2}{\sum_{ki}\tilde{C}\|\boldsymbol{X}^*\|_F^2|\boldsymbol{w}(k)|^2} = \frac{\epsilon_1^2 mq}{\tilde{C}}.$$

Since $\tilde{C}=9\mu^2\kappa^2$, thus, w.p. $1-\exp(-c\epsilon_1^2 mq/\mu^2\kappa^2)$, for a fixed $\boldsymbol{z}$ and $\boldsymbol{w}$,

$$\boldsymbol{z}^\top(\boldsymbol{X}_0-\boldsymbol{X}_+)\boldsymbol{w} \leq \epsilon_1\|\boldsymbol{X}^*\|_F + \mathbb{E}\left[\frac{1}{m}\sum_{ki}\big|\boldsymbol{w}(k)\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{z})\big|\mathbb{1}_{\left\{|\boldsymbol{y}_{ki}|^2\in[1\pm\epsilon_1]\frac{\tilde{C}}{q}\|\boldsymbol{X}^*\|_F^2\right\}}\right]$$

By using Claim B.1 and $|\boldsymbol{w}(k)|\|\boldsymbol{z}\| = |\boldsymbol{w}(k)|$ we have

$$\mathbb{E}\left[\frac{1}{m}\sum_{ki}\big|\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{z})\boldsymbol{w}(k)\big|\mathbb{1}_{\left\{|\boldsymbol{y}_{ki}|^2\in[1\pm\epsilon_1]\frac{\tilde{C}}{q}\|\boldsymbol{X}^*\|_F^2\right\}}\right] \leq \sqrt{\tilde{C}(1+\epsilon_1)}\epsilon_1\|\boldsymbol{X}^*\|_F\sum_k\big|\boldsymbol{w}(k)\big|/\sqrt{q} \leq C\epsilon_1\mu\kappa\|\boldsymbol{X}^*\|_F,$$

where in the last inequality we used Cauchy-Schwarz to show that $\sum_k\big|\boldsymbol{w}(k)\big|/\sqrt{q} \leq \sqrt{\sum_k\big|\boldsymbol{w}(k)\big|^2\sum_k(1/q)} = 1$. Or this also follows by $\|\boldsymbol{w}\|_1/\sqrt{q} \leq \|\boldsymbol{w}\| = 1$. Also, we used $\sqrt{\tilde{C}} = C\kappa\mu$.

Thus, w.p. $1-\exp(-c\epsilon_1^2 mq/\mu^2\kappa^2)$, for a fixed $\boldsymbol{z}$ and $\boldsymbol{w}$, $\boldsymbol{z}^\top(\boldsymbol{X}_0-\boldsymbol{X}_+)\boldsymbol{w} \leq C\epsilon_1\mu\kappa\|\boldsymbol{X}^*\|_F$.

By Proposition 4.8, $\max_{\boldsymbol{z}\in\mathcal{S}^n,\ \boldsymbol{w}\in\mathcal{S}^q}\boldsymbol{z}^\top(\boldsymbol{X}_0-\boldsymbol{X}_+)\boldsymbol{w} \leq 1.4C\epsilon_1\mu\kappa\|\boldsymbol{X}^*\|_F$ w.p. at least $1-\exp((n+q)\log(17) - c\epsilon_1^2 mq/\mu^2\kappa^2)$. ∎

Lemma B.3. *Consider $\boldsymbol{X}_+$. Fix $1<\epsilon_1<1$. Then, w.p. $1-\exp\left[C(n+q)-c\epsilon_1^2 mq/\mu^2\kappa^2\right]$*

$$\|\boldsymbol{X}_+ - \mathbb{E}[\boldsymbol{X}_+]\| \leq C\epsilon_1\|\boldsymbol{X}^*\|_F.$$

Proof of Lemma B.3: The proof involves an application of the sub-Gaussian Hoeffding inequality followed by an epsilon-net argument, both almost the same as those used in the proof of Lemma B.2 given above. We have,

$$\|\boldsymbol{X}_+ - \mathbb{E}[\boldsymbol{X}_+]\| = \max_{\boldsymbol{z}\in\mathcal{S}_n,\boldsymbol{w}\in\mathcal{S}_q}\langle\boldsymbol{X}_+ - \mathbb{E}[\boldsymbol{X}_+],\ \boldsymbol{z}\boldsymbol{w}^\top\rangle.$$

For a fixed $\boldsymbol{z}\in\mathcal{S}_n, \boldsymbol{w}\in\mathcal{S}_q$, we have

$$\langle\boldsymbol{X}_+ - \mathbb{E}[\boldsymbol{X}_+],\ \boldsymbol{z}\boldsymbol{w}^\top\rangle = \frac{1}{m}\sum_{ki}\left(\boldsymbol{w}(k)\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{z})\mathbb{1}_{\left\{|\boldsymbol{y}_{ki}|^2\leq\frac{\tilde{C}(1+\epsilon_1)}{q}\|\boldsymbol{X}^*\|_F^2\right\}} - \mathbb{E}\left[\boldsymbol{w}(k)\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{z})\mathbb{1}_{\left\{|\boldsymbol{y}_{ki}|^2\leq\frac{\tilde{C}(1+\epsilon_1)}{q}\|\boldsymbol{X}^*\|_F^2\right\}}\right]\right).$$

The summands are mutually independent, zero mean sub-Gaussian r.v.s with norm $K_{ki} \leq C|\boldsymbol{w}(k)|\sqrt{\tilde{C}(1+\epsilon_1)}\|\boldsymbol{X}^*\|_F/\sqrt{q}$. We will again apply the sub-Gaussian Hoeffding inequality Theorem 2.6.2 of [26]. Let $t=\epsilon_1 m\|\boldsymbol{X}^*\|_F$. Then

$$\frac{t^2}{\sum_{ki}K_{ki}^2} \geq \frac{\epsilon_1^2m^2\|\boldsymbol{X}^*\|_F^2}{\sum_{ki}\tilde{C}(1+\epsilon_1)\|\boldsymbol{X}^*\|_F^2/q} \geq \frac{\epsilon_1^2 mq}{C\mu^2\kappa^2}$$

Thus, for a fixed $\boldsymbol{z}\in\mathcal{S}_n, \boldsymbol{w}\in\mathcal{S}_q$, by sub-Gaussian Hoeffding, we conclude that, w.p. at least $1-\exp\left[-c\epsilon_1^2 mq/\mu^2\kappa^2\right]$,

$$\langle\boldsymbol{X}_+ - \mathbb{E}[\boldsymbol{X}_+],\ \boldsymbol{z}\boldsymbol{w}^\top\rangle \leq C\epsilon_1\|\boldsymbol{X}^*\|_F.$$

By Proposition 4.7, the above bound holds w.p. at least $1-\exp\left[(n+q)-c\epsilon_1^2 mq/\mu^2\kappa^2\right]$. ∎

B. Bounding the $\check{\boldsymbol{V}}$ numerator term

We bound $\|(\boldsymbol{X}_0-\mathbb{E}[\boldsymbol{X}_+])\check{\boldsymbol{V}}^\top\|_F$ in this section. By triangle inequality. it is bounded by $\|(\boldsymbol{X}_0-\boldsymbol{X}_+)\check{\boldsymbol{V}}^\top\|_F + \|(\boldsymbol{X}_+-\mathbb{E}[\boldsymbol{X}_+])\check{\boldsymbol{V}}^\top\|_F$.

Lemma B.4. *Assume that $\frac{1}{m}\sum_{ki}\boldsymbol{y}_{ki}^2 \in [1\pm\epsilon_1]\|\boldsymbol{X}^*\|_F^2$. Then, w.p. $1-\exp\left[nr-c\epsilon_1^2 mq/\mu^2\kappa^2\right]$,*

$$\|(\boldsymbol{X}_0-\boldsymbol{X}_+)\check{\boldsymbol{V}}^\top\|_F \leq C\epsilon_1\mu\kappa\|\boldsymbol{X}^*\|_F.$$

Proof of Lemma B.4: The initial part of the proof is very similar to the that of the proof of Lemma B.2. We have, $\|(\boldsymbol{X}_0-\boldsymbol{X}_+)\check{\boldsymbol{V}}^\top\|_F = \max_{\boldsymbol{W}\in\mathcal{S}_{nr}}\langle\boldsymbol{W},\ (\boldsymbol{X}-\boldsymbol{X}_+)\check{\boldsymbol{V}}^\top\rangle$. For a fixed $\boldsymbol{W}\in\mathcal{S}_{nr}$,

$$\langle\boldsymbol{W},\ (\boldsymbol{X}_0-\boldsymbol{X}_+)\check{\boldsymbol{V}}^\top\rangle = \frac{1}{m}\sum_{ki}\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{W}\check{\boldsymbol{v}}_k)\mathbb{1}_{\left\{\frac{\tilde{C}}{mq}\sum_{ki}\boldsymbol{y}_{ki}^2\leq|\boldsymbol{y}_{ki}|^2\leq\frac{\tilde{C}(1+\epsilon_1)}{q}\|\boldsymbol{X}^*\|_F^2\right\}}$$

Proceeding as in the proof of Lemma B.2,

$$\frac{1}{m}\sum_{ki}\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{W}\check{\boldsymbol{v}}_k)\mathbb{1}_{\left\{\frac{\tilde{C}}{mq}\sum_{ki}\boldsymbol{y}_{ki}^2\leq|\boldsymbol{y}_{ki}|^2\leq\frac{\tilde{C}(1+\epsilon_1)}{q}\|\boldsymbol{X}^*\|_F^2\right\}} \leq \frac{1}{m}\sum_{ki}\big|\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{W}\check{\boldsymbol{v}}_k)\big|\mathbb{1}_{\left\{\frac{\tilde{C}}{mq}\sum_{ki}\boldsymbol{y}_{ki}^2\leq|\boldsymbol{y}_{ki}|^2\leq\frac{\tilde{C}(1+\epsilon_1)}{q}\|\boldsymbol{X}^*\|_F^2\right\}},$$

$$\leq \frac{1}{m}\sum_{ki}|\boldsymbol{y}_{ki}||(\boldsymbol{a}_{ki}{}^\top\boldsymbol{W}\check{\boldsymbol{v}}_k)|\mathbb{1}_{\left\{|\boldsymbol{y}_{ki}|^2\in[1\pm\epsilon_1]\frac{\tilde{C}}{q}\|\boldsymbol{X}^*\|_F^2\right\}}.$$

The summands are mutually independent sub-Gaussian r.v.s with norm $K_{ki} \leq C\sqrt{\tilde{C}(1+\epsilon_1)}\|\boldsymbol{W}\check{\boldsymbol{v}}_k\|\|\boldsymbol{X}^*\|_F/\sqrt{q}$. Thus, we can apply the sub-Gaussian Hoeffding inequality Theorem 2.6.2 of [26]. Set $t=\epsilon_1 m\|\boldsymbol{X}^*\|_F$. Then we have

$$\frac{t^2}{\sum_{ki}K_{ki}^2} \geq \frac{\epsilon_1^2m^2\|\boldsymbol{X}^*\|_F^2}{(\sum_{ki}\|\boldsymbol{W}\check{\boldsymbol{v}}_k\|^2)\tilde{C}(1+\epsilon_1)\|\boldsymbol{X}^*\|_F^2/q} \geq \frac{\epsilon_1^2 mq}{C\mu^2\kappa^2},$$

where we used the fact that $\check{\boldsymbol{V}}\check{\boldsymbol{V}}^\top = \boldsymbol{I}$ ($\check{\boldsymbol{V}}^\top$ contains right singular vectors of a matrix) and thus $\|\boldsymbol{W}\check{\boldsymbol{V}}\|_F = 1$. Applying sub-Gaussian Hoeffding, we can conclude that, w.p., $1-\exp\left[-c\epsilon_1^2 mq/\mu^2\kappa^2\right]$

$$\frac{1}{m}\sum_{ki}\big|\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{W}\check{\boldsymbol{v}}_k)\big|\mathbb{1}_{\left\{|\boldsymbol{y}_{ki}|^2\in[1\pm\epsilon_1]\frac{\tilde{C}}{q}\|\boldsymbol{X}^*\|_F^2\right\}} \leq \epsilon_1\|\boldsymbol{X}^*\|_F + \frac{1}{m}\sum_{ki}\mathbb{E}\left[\big|\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{W}\check{\boldsymbol{v}}_k)\big|\mathbb{1}_{\left\{|\boldsymbol{y}_{ki}|^2\in[1\pm\epsilon_1]\frac{\tilde{C}}{q}\|\boldsymbol{X}^*\|_F^2\right\}}\right].$$

We use Claim B.1 to bound the expectation term. Using this lemma with $\alpha^2 \equiv \tilde{C}(1+\epsilon_1)\|\boldsymbol{X}^*\|_F^2/q$, $\boldsymbol{z} \equiv \boldsymbol{W}\check{\boldsymbol{v}}_k$

$$\frac{1}{m}\sum_{ki}\mathbb{E}\left[\left|\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top \boldsymbol{W}\check{\boldsymbol{v}}_k)\right|\mathbb{1}_{\left\{|\boldsymbol{y}_{ki}|^2\in[1\pm\epsilon_1]\frac{\tilde{C}}{q}\|\boldsymbol{X}^*\|_F^2\right\}}\right] \le \frac{1}{m}\sum_{ki}\sqrt{\tilde{C}(1+\epsilon_1)}\epsilon_1\|\boldsymbol{X}^*\|_F\|\boldsymbol{W}\check{\boldsymbol{v}}_k\|/\sqrt{q} \le C\epsilon_1\mu\kappa\|\boldsymbol{X}^*\|_F.$$

where the last inequality used Cauchy-Schwarz on $\sum_k\|\boldsymbol{W}\check{\boldsymbol{v}}_k\|/\sqrt{q}$ to conclude that $\sum_k\|\boldsymbol{W}\check{\boldsymbol{v}}_k\|(1/\sqrt{q}) \le \sqrt{\sum_k\|\boldsymbol{W}\check{\boldsymbol{v}}_k\|^2\sum_k(1/q)} = \sqrt{\|\boldsymbol{W}\check{\boldsymbol{V}}\|_F^2\cdot 1} = 1$ since $\|\boldsymbol{W}\check{\boldsymbol{V}}\|_F = 1$.

By Proposition 4.8, the above bound holds for all $\boldsymbol{W}\in\mathcal{S}_{nr}$, w.p. at least $1-\exp\left[nr\log(1+2/\epsilon_{net}) - c\epsilon_1^2 mq/\mu^2\kappa^2\right]$. ■

Lemma B.5. *Consider* $0<\epsilon_1<1$. *Then, w.p.* $1-\exp\left[nr-\epsilon_1^2 mq/\mu^2\kappa^2\right]$

$$\|\left(\boldsymbol{X}_+ - \mathbb{E}[\boldsymbol{X}_+]\right)\check{\boldsymbol{V}}^\top\|_F \le C\epsilon_1\|\boldsymbol{X}^*\|_F.$$

Proof of Lemma B.5: The proof is quite similar to the previous one. For a fixed $\boldsymbol{W}\in\mathcal{S}_{nr}$ we have,

$$\langle\left(\boldsymbol{X}_+ - \mathbb{E}[\boldsymbol{X}_+]\right)\check{\boldsymbol{V}}^\top,\ \boldsymbol{W}\rangle = \frac{1}{m}\sum_{ki}\left(\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{W}\check{\boldsymbol{v}}_k)\mathbb{1}_{\left\{|\boldsymbol{y}_{ki}|^2\le\frac{\tilde{C}(1+\epsilon_1)}{q}\|\boldsymbol{X}^*\|_F^2\right\}} - \mathbb{E}[.]\right)$$

where $\mathbb{E}[.]$ is the expected value of the first term. The summands are independent, zero mean, sub-Gaussian r.v.s with subGaussian norm less than $K_{ki}\le C\sqrt{\tilde{C}(1+\epsilon_1)}\|\boldsymbol{X}^*\|_F\|\boldsymbol{W}\boldsymbol{b}_k\|/\sqrt{q}$. Thus, by applying the sub-Gaussian Hoeffding inequality Theorem 2.6.2 of [26], with $t=\epsilon_1 m\|\boldsymbol{X}^*\|_F$, and using $\|\boldsymbol{W}\check{\boldsymbol{V}}\|_F = 1$, we can conclude that, w.p. $1-\exp\left[-\epsilon_1^2 mq/(C\mu^2\kappa^2)\right]$,

$$\langle\left(\boldsymbol{X}_+ - \mathbb{E}[\boldsymbol{X}_+]\right)\check{\boldsymbol{V}}^\top,\ \boldsymbol{W}\rangle \le C\epsilon_1\|\boldsymbol{X}^*\|_F.$$

By Proposition 4.8, the above bound holds for all $\boldsymbol{W}\in\mathcal{S}_{nr}$ w.p. $1-\exp\left[nr-\epsilon_1^2 mq/(C\mu^2\kappa^2)\right]$. ■

C. Bounding the U* numerator term

We bound $\|(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_+])^\top\boldsymbol{U}^*\|_F$ here. By triangle inequality, it is bounded by $\|\left(\boldsymbol{X}_0-\boldsymbol{X}_+\right)^\top\boldsymbol{U}^*\|_F + \|\left(\boldsymbol{X}_+ - \mathbb{E}[\boldsymbol{X}_+]\right)^\top\boldsymbol{U}^*\|_F$.

Lemma B.6. *Assume that* $\frac{1}{mq}\sum_{ki}\boldsymbol{y}_{ki}^2 \in [1\pm\epsilon_1]\|\boldsymbol{X}^*\|_F^2/q$. *Then, w.p.* $1-\exp\left[qr - c\epsilon_1^2 mq/\mu^2\kappa^2\right]$

$$\|\left(\boldsymbol{X}_0-\boldsymbol{X}_+\right)^\top\boldsymbol{U}^*\|_F \le C\epsilon_1\mu\kappa\|\boldsymbol{X}^*\|_F.$$

Proof of Lemma B.6: The proof is similar to that of Lemmas B.2 and B.4. We have, $\|\left(\boldsymbol{X}_0-\boldsymbol{X}_+\right)^\top\boldsymbol{U}^*\|_F = \max_{\boldsymbol{W}\in\mathcal{S}_{qr}}\langle\boldsymbol{W},\ \left(\boldsymbol{X}-\boldsymbol{X}_+\right)^\top\boldsymbol{U}^*\rangle$. For a fixed $\boldsymbol{W}\in\mathcal{S}_{qr}$, using the same approach as in Lemma B.2, and letting $\boldsymbol{w}_k$ be the k-th column of the $r\times q$ matrix $\boldsymbol{W}$,

$$\begin{aligned}\langle\boldsymbol{W},\ \left(\boldsymbol{X}_0-\boldsymbol{X}_+\right)^\top\boldsymbol{U}^*\rangle &\le \frac{1}{m}\sum_{ki}\left|\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{U}^*\boldsymbol{w}_k)\right|\mathbb{1}_{\left\{\frac{\tilde{C}}{mq}\sum_{ki}|\boldsymbol{y}_{ki}|^2\le|\boldsymbol{y}_{ki}|^2\le\frac{\tilde{C}(1+\epsilon_1)}{q}\|\boldsymbol{X}^*\|_F^2\right\}}\\ &\le \frac{1}{m}\sum_{ki}\left|\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{U}^*\boldsymbol{w}_k)\right|\mathbb{1}_{\left\{|\boldsymbol{y}_{ki}|^2\in[1\pm\epsilon_1]\frac{\tilde{C}}{q}\|\boldsymbol{X}^*\|_F^2\right\}}.\end{aligned}$$

The summands are now mutually independent sub-Gaussian r.v.s with norm $K_{ki}\le\sqrt{\tilde{C}(1+\epsilon_1)}\|\boldsymbol{w}_k\|\|\boldsymbol{X}^*\|_F/\sqrt{q}$. Thus, we can apply the sub-Gaussian Hoeffding inequality Theorem 2.6.2 of [26], to conclude that, for a fixed $\boldsymbol{W}\in\mathcal{S}_{qr}$, w.p. $1-\exp\left[-c\epsilon_1^2 mq/\mu^2\kappa^2\right]$,

$$\begin{aligned}&\frac{1}{m}\sum_{ki}\left|\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{U}^*\boldsymbol{w}_k)\right|\mathbb{1}_{\left\{|\boldsymbol{y}_{ki}|^2\in[1\pm\epsilon_1]\frac{\tilde{C}}{q}\|\boldsymbol{X}^*\|_F^2\right\}}\\ &\le \epsilon_1\|\boldsymbol{X}^*\|_F + \frac{1}{m}\sum_k\mathbb{E}\left[\left|\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{U}^*\boldsymbol{w}_k)\right|\mathbb{1}_{\left\{|\boldsymbol{y}_{ki}|^2\in[1\pm\epsilon_1]\frac{\tilde{C}}{q}\|\boldsymbol{X}^*\|_F^2\right\}}\right]\end{aligned}$$

By Claim B.1, and using $\sum_k\|\boldsymbol{w}_k\|/\sqrt{q} \le \sqrt{\sum_k\|\boldsymbol{w}_k\|^2}\sqrt{\sum_k 1/q} = 1$,

$$\begin{aligned}&\frac{1}{m}\sum_k\mathbb{E}\left[\left|\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{U}^*\boldsymbol{w}_k)\right|\mathbb{1}_{\left\{|\boldsymbol{y}_{ki}|^2\in[1\pm\epsilon_1]\frac{\tilde{C}}{q}\|\boldsymbol{X}^*\|_F^2\right\}}\right]\\ &\le \frac{1}{m}\sum_{ki}\epsilon_1\|\boldsymbol{w}_k\|\sqrt{\tilde{C}(1+\epsilon_1)/q}\|\boldsymbol{X}^*\|_F,\\ &\le C\epsilon_1\mu\kappa\|\boldsymbol{X}^*\|_F,\end{aligned}$$

By Proposition 4.8 (epsilon net argument), the bound holds for all unit norm $\boldsymbol{W}$ w.p. $1-\exp\left[qr - c\epsilon_1^2 mq/\mu^2\kappa^2\right]$. ■

Lemma B.7. *Consider* $0<\epsilon_1<1$. *Then, w.p.* $1-\exp\left[qr-\epsilon_1^2/mq\mu^2\kappa^2\right]$

$$\|\left(\boldsymbol{X}_+ - \mathbb{E}[\boldsymbol{X}_+]\right)^\top\boldsymbol{U}^*\|_F \le C\epsilon_1\|\boldsymbol{X}^*\|_F.$$

Proof of Lemma B.7: For fixed $\boldsymbol{W}\in\mathcal{S}_{qr}$,

$$\begin{aligned}&\mathrm{trace}\left(\boldsymbol{W}^\top\left(\boldsymbol{X}_+ - \mathbb{E}[\boldsymbol{X}_+]\right)^\top\boldsymbol{U}^*\right)\\ &= \frac{1}{m}\sum_{ki}\left(\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{U}^*\boldsymbol{w}_k)\mathbb{1}_{\left\{|\boldsymbol{y}_{ki}|^2\le\frac{\tilde{C}(1+\epsilon_1)}{q}\|\boldsymbol{X}^*\|_F^2\right\}}\right.\\ &\left. - \mathbb{E}\left[\boldsymbol{y}_{ki}(\boldsymbol{a}_{ki}{}^\top\boldsymbol{U}^*\boldsymbol{w}_k)\mathbb{1}_{\left\{|\boldsymbol{y}_{ki}|^2\le\frac{\tilde{C}(1+\epsilon_1)}{q}\|\boldsymbol{X}^*\|_F^2\right\}}\right]\right)\end{aligned}$$

The summands are independent zero mean sub-Gaussian r.v.s with norm less than $K_{ki}\le\sqrt{\tilde{C}(1+\epsilon_1)}\|\boldsymbol{X}^*\|_F\|\boldsymbol{w}_k\|/\sqrt{q}$. Thus, by applying the sub-Gaussian Hoeffding inequality Theorem 2.6.2 of [26], with $t=\epsilon_1 m\|\boldsymbol{X}^*\|_F$, we can conclude that, for a fixed $\boldsymbol{W}\in\mathcal{S}_{qr}$, w.p. $1-\exp\left[-\epsilon_1^2 mq/C\mu^2\kappa^2\right]$,

$$\mathrm{trace}\left(\boldsymbol{W}^\top\left(\boldsymbol{X}_+ - \mathbb{E}[\boldsymbol{X}_+]\right)^\top\boldsymbol{U}^*\right) \le \epsilon_1\|\boldsymbol{X}^*\|_F.$$

By Proposition 4.8 (epsilon net argument), the bound holds for all unit norm $\boldsymbol{W}$ w.p. $1-\exp\left[qr-\epsilon_1^2 mq/C\mu^2\kappa^2\right]$. ■

D. Proof of Claim B.1

Proof: We can write $\boldsymbol{x}^* = \|\boldsymbol{x}^*\|\boldsymbol{Q}\boldsymbol{e}_1$ where $\boldsymbol{Q}$ is a unitary matrix with first column proportional to $\boldsymbol{x}_k^*$. We need to bound

$$\begin{aligned}&\mathbb{E}[\|\boldsymbol{x}^*\|\cdot|(\boldsymbol{a}^\top\boldsymbol{Q}\boldsymbol{e}_1)(\boldsymbol{a}^\top\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{z})|\mathbb{1}_{\{\|\boldsymbol{x}^*\|^2|\boldsymbol{a}^\top\boldsymbol{Q}\boldsymbol{e}_1|^2\in[1\pm\epsilon]\alpha\}}]\\ &= \|\boldsymbol{x}^*\|\cdot\|\boldsymbol{z}\|\cdot\mathbb{E}[|\tilde{\boldsymbol{a}}(1)\tilde{\boldsymbol{a}}^\top\bar{\boldsymbol{z}}_Q|\mathbb{1}_{\{|\tilde{\boldsymbol{a}}(1)|^2\in[1\pm\epsilon]\beta^2\}}],\end{aligned}$$

where $\bar{\boldsymbol{z}}_Q := \boldsymbol{Q}^\top\boldsymbol{z}/\|\boldsymbol{z}\|$, $\tilde{\boldsymbol{a}} := \boldsymbol{Q}^\top\boldsymbol{a}$ and $\beta := \sqrt{\alpha}/\|\boldsymbol{x}^*\|$. Since $\boldsymbol{Q}$ is unitary and $\boldsymbol{a}$ Gaussian, thus $\tilde{\boldsymbol{a}}$ has the same distribution as $\boldsymbol{a}$. Let $\tilde{\boldsymbol{a}}(1)$ be its first entry and $\tilde{\boldsymbol{a}}(\mathrm{rest})$ be the

$(n-1)$-length vector with the rest of the $n-1$ entries and similarly for $\bar{\boldsymbol{z}}_Q$. Then, $\tilde{\boldsymbol{a}}^\top \bar{\boldsymbol{z}}_Q = \tilde{\boldsymbol{a}}(1)\cdot \bar{\boldsymbol{z}}_Q(1) + \tilde{\boldsymbol{a}}(\mathrm{rest})^\top \bar{\boldsymbol{z}}_Q(\mathrm{rest})$. Since $\tilde{\boldsymbol{a}}(1)$ and $\tilde{\boldsymbol{a}}(\mathrm{rest})$ are independent,

$$\begin{aligned}
&\mathbb{E}[|\tilde{\boldsymbol{a}}(1)\tilde{\boldsymbol{a}}^\top \bar{\boldsymbol{z}}_Q| \mathbb{1}_{|\tilde{\boldsymbol{a}}(1)|^2\in[1\pm\epsilon]\beta^2}] \\
&\le |\bar{\boldsymbol{z}}_Q(1)| \mathbb{E}[|\tilde{\boldsymbol{a}}(1)^2| \mathbb{1}_{|\tilde{\boldsymbol{a}}(1)|^2\in[1\pm\epsilon]\beta^2}] \\
&\qquad + \mathbb{E}[|\tilde{\boldsymbol{a}}(\mathrm{rest})^\top \bar{\boldsymbol{z}}_Q(\mathrm{rest})|]\ \mathbb{E}[|\tilde{\boldsymbol{a}}(1)| \mathbb{1}_{|\tilde{\boldsymbol{a}}(1)|^2\in[1\pm\epsilon]\beta^2}] \\
\le& \mathbb{E}[|\tilde{\boldsymbol{a}}(1)^2| \mathbb{1}_{|\tilde{\boldsymbol{a}}(1)|^2\in[1\pm\epsilon]\beta^2}] + 2\mathbb{E}[|\tilde{\boldsymbol{a}}(1)| \mathbb{1}_{|\tilde{\boldsymbol{a}}(1)|^2\in[1\pm\epsilon]\beta^2}] \\
\le& \epsilon\beta + 2\epsilon\beta = 3\epsilon\beta = C\epsilon\frac{\sqrt{\alpha}}{\|\boldsymbol{x}^*\|}.
\end{aligned}$$

The second inequality used the facts that (i) $|\bar{\boldsymbol{z}}_Q(1)| \le \|\bar{\boldsymbol{z}}_Q\| = 1$ by definition and (ii) $\zeta := \tilde{\boldsymbol{a}}(\mathrm{rest})^\top \bar{\boldsymbol{z}}_Q(\mathrm{rest})$ is a scalar standard Gaussian r.v. and so $\mathbb{E}[|\zeta|] \le 2$. The third one relies on the following two bounds:

1)
$$\begin{aligned}
&\mathbb{E}\left[|\boldsymbol{a}(1)|^2 \mathbb{1}_{\{|\boldsymbol{a}(1)|^2\in[1\pm\epsilon]\beta^2\}}\right] \\
&= \frac{2}{\sqrt{2\pi}} \int_{\sqrt{1-\epsilon}\beta}^{\sqrt{1+\epsilon}\beta} z^2 \exp(-z^2/2) dz, \\
&\le \frac{2e^{-1/2}}{\sqrt{2\pi}} \int_{\sqrt{1-\epsilon}\beta}^{\sqrt{1+\epsilon}\beta} dz \le \frac{2e^{-1/2}}{\sqrt{2\pi}} \epsilon\beta \le \epsilon\beta/3
\end{aligned}$$
where we used the facts that $z^2\exp(-z^2/2) \le \exp(-1/2)$ for all $z \in \Re$; $\sqrt{1-\epsilon} \ge 1-\epsilon/2$ and $\sqrt{1+\epsilon} \le 1+\epsilon/2$ for $0<\epsilon<1$.

2) Similarly, we can show that
$$\begin{aligned}
&\mathbb{E}\left[|\boldsymbol{a}(1)| \mathbb{1}_{\{|\boldsymbol{a}(1)|^2\in[1\pm\epsilon]\beta^2\}}\right] \\
&= \frac{2}{\sqrt{2\pi}} \int_{\sqrt{1-\epsilon}\beta}^{\sqrt{1+\epsilon}\beta} z \exp(-z^2/2) dz, \\
&\le \frac{2e^{-1/2}}{\sqrt{2\pi}} \int_{\sqrt{1-\epsilon}\beta}^{\sqrt{1+\epsilon}\beta} dz = \frac{2e^{-1/2}}{\sqrt{2\pi}} \epsilon\beta \le \epsilon\beta/3
\end{aligned}$$

The claim follows by combining the two equations given above. ■