Efficient Federated Low Rank Matrix Recovery via Alternating GD and Minimization: A Simple Proof

Namrata Vaswani, *Fellow, IEEE*

***Abstract*— This note provides a significantly simpler and shorter proof of our sample complexity guarantee for solving the low rank column-wise sensing problem using the Alternating Gradient Descent (GD) and Minimization (AltGDmin) algorithm. AltGDmin was developed and analyzed for solving this problem in our recent work. We also provide an improved guarantee.**

***Index Terms*— Low rank column-wise sensing (LRCS), federated learning, multi-task representation learning.**

I. Introduction

WE STUDY the low rank column-wise sensing (LRCS) problem which involves recovering a low rank matrix from independent compressive measurements of each of its columns. This problem occurs in dynamic MRI [1] and in multi-task linear representation learning for few-shot learning [2]. The alternating gradient descent (GD) and minimization (AltGDmin) algorithm for solving it in a fast, communication-efficient and private fashion was developed and analyzed in our recent work [3]. This short paper provides a significantly simpler and shorter proof of our sample complexity guarantee for AltGDmin. In fact, it also improves the sample complexity needed by the AltGDmin iterations by a factor of r.

II. Problem Statement, Notation, and Algorithm

A. Problem Statement and Assumption and Notation

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

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

where each $\boldsymbol{y}_k$ is an m-length vector, $[q] := \{1, 2, \ldots, q\}$, 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, $m < n$. Each $\boldsymbol{A}_k$ is assumed to be random-Gaussian: each entry of it is independent and identically distributed (i.i.d.) standard Gaussian. Let $\boldsymbol{X}^\star \overset{\text{SVD}}{=} \boldsymbol{U}^\star \boldsymbol{\Sigma}^* \boldsymbol{V}^\star := \boldsymbol{U}^\star \boldsymbol{B}^\star$ denote its reduced (rank r) SVD, and $\kappa := \sigma_{\max}^\star / \sigma_{\min}^\star$ the condition number of $\boldsymbol{\Sigma}^*$. We let $\boldsymbol{B}^\star := \boldsymbol{\Sigma}^* \boldsymbol{V}^\star$.

Since no measurement $\boldsymbol{y}_{ki}$ is a global function of the entire matrix, $\boldsymbol{X}^\star$, we need the following assumption, borrowed from LR matrix completion literature, to make our problem well-posed (allow for correct interpolation across columns).

Assumption 1 (Incoherence of Right Singular Vectors): Assume that $\|\boldsymbol{b}_k^\star\|^2 \le \mu^2 r {\sigma_{\max}^\star}^2 / q$ for a numerical constant μ.

In our discussion of communication complexity and privacy, we assume a vertically federated setting: different subsets of $\boldsymbol{y}_k, \boldsymbol{A}_k$ are available at different nodes.

1) Notation: We use $\|.\|_F$ to denote the Frobenius norm, $\|.\|$ without a subscript for the (induced) l_2 norm, $^\top$ to denote matrix or vector transpose, $\boldsymbol{e}_k$ to denote the k-th canonical basis vector (k-th column of $\boldsymbol{I}$), and $\boldsymbol{M}^\dagger := (\boldsymbol{M}^\top \boldsymbol{M})^{-1} \boldsymbol{M}^\top$. For two $n \times r$ matrices $\boldsymbol{U}_1, \boldsymbol{U}_2$ that have orthonormal columns, we use

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

as the Subspace Distance (SD) measure. In our previous work [3], we used the Frobenius norm SD,

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

Clearly, $\text{SD}_F(\boldsymbol{U}_1, \boldsymbol{U}_2) \le \sqrt{r} \text{SD}_2(\boldsymbol{U}_1, \boldsymbol{U}_2)$. We reuse the letters c, C to denote different numerical constants in each use with the convention that $c < 1$ and $C \ge 1$. We use $\sum_k$ as a shortcut for the summation over $k = 1$ to q and $\sum_{ki}$ for the summation over $i = 1$ to m and $k = 1$ to q. *We use whp to refer to "with high probability" and this means that the claim holds with probability (w.p.) at least* $1 - n^{-10}$.

B. Review of AltGDmin Algorithm [3]

AltGDmin, summarized in Algorithm 1, imposes the LR constraint by factorizing the unknown matrix $\boldsymbol{X}$ as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{B}$ with $\boldsymbol{U}$ being an $n \times r$ matrix and $\boldsymbol{B}$ an $r \times q$ matrix. It minimizes $f(\boldsymbol{U}, \boldsymbol{B}) := \sum_{k=1}^q \|\boldsymbol{y}_k - \boldsymbol{U}\boldsymbol{b}_k\|^2$ as follows:

1) *Truncated spectral initialization:* Initialize $\boldsymbol{U}$ (see below).
2) At each iteration, update $\boldsymbol{B}$ and $\boldsymbol{U}$ as follows:
   a) *Minimization for $\boldsymbol{B}$:* keeping $\boldsymbol{U}$ fixed, update $\boldsymbol{B}$ by solving $\min_{\boldsymbol{B}} f(\boldsymbol{U}, \boldsymbol{B})$. Due to the form of the LRCS model, this minimization decouples across columns, making it a cheap least squares problem of recovering q different r length vectors. It is solved as $\boldsymbol{b}_k = (\boldsymbol{A}_k \boldsymbol{U})^\dagger \boldsymbol{y}_k$ for each $k \in [q]$.

Manuscript received 3 July 2023; revised 26 October 2023; accepted 10 February 2024. Date of publication 1 March 2024; date of current version 18 June 2024. This work was supported in part by NSF under Grant CCF-2115200.

The author is with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA (e-mail: namrata@iastate.edu).

Communicated by Y. Chi, Associate Editor for Machine Learning and Statistics.

Digital Object Identifier 10.1109/TIT.2024.3365795

b) *Projected-GD for $\boldsymbol{U}$:* keeping $\boldsymbol{B}$ fixed, update $\boldsymbol{U}$ by a GD step, followed by orthonormalizing its columns: $\boldsymbol{U}^+ = QR(\boldsymbol{U} - \eta \nabla_{\boldsymbol{U}} f(\boldsymbol{U}, \boldsymbol{B}))$. Here $QR(.)$ orthonormalizes the columns of its input.

We initialize $\boldsymbol{U}$ by computing the top r singular vectors of

$$\boldsymbol{X}_0 := \sum_k \boldsymbol{A}_k^\top \boldsymbol{y}_{k,trnc} \boldsymbol{e}_k^\top, \ \boldsymbol{y}_{k,trnc} := \text{trunc}(\boldsymbol{y}_k, \alpha)$$

Here $\alpha := \tilde{C} \sum_k \|\boldsymbol{y}_k\|^2/mq$ with $\tilde{C} := 9\kappa^2\mu^2$. The function trunc truncates (zeroes out) all entries of $\boldsymbol{y}_k$ with magnitude greater than $\sqrt{\alpha}$, i.e., for all $j \in [n]$, $\text{trunc}(\boldsymbol{y}, \alpha)_j = (\boldsymbol{y})_j \mathbb{1}_{|\boldsymbol{y}_j| \le \sqrt{\alpha}})$, with $\mathbb{1}$ being the indicator function.

Sample-splitting is assumed, i.e., each new update of $\boldsymbol{U}$ and $\boldsymbol{B}$ uses a new independent set of measurements and measurement matrices, $\boldsymbol{y}_k, \boldsymbol{A}_k$.

The use of minimization to update $\boldsymbol{B}$ at each iteration is what helps ensure that we can show exponential error decay with a constant step size. At the same time, due to the column-wise decoupled nature of LRCS, the time complexity for this step is only as much as that of computing one gradient w.r.t. $\boldsymbol{U}$. Both steps need time[1] of order $mqnr$. This is only r times more than "linear time" (time needed to read the algorithm inputs, here $\boldsymbol{y}_k, \boldsymbol{A}_k$'s). To our knowledge, r-times linear-time is the best known time complexity for any algorithm for any LR matrix recovery problem. Moreover, due to the use of the $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{B}$ factorization, AltGDmin is also communication-efficient. Each node needs to only send nr scalars (gradients w.r.t $\boldsymbol{U}$) at each iteration.

III. New Guarantee

Let m_0 denote the total number of samples per column needed for initialization and let m_1 denote this number for each GDmin iteration. Then, the total sample complexity per column is $m = m_0 + m_1 T$. Our guarantee given next provides the required minimum value of m.

Theorem 2: Assume that Assumption 1 holds. Set $\eta = 0.4/m{\sigma^\star_{\max}}^2$ and $T = C\kappa^2 \log(1/\epsilon)$. If

$$mq \ge C\kappa^4\mu^2(n+q)r(\kappa^4 r + \log(1/\epsilon))$$

and $m \ge C \max(\log n, \log q, r) \log(1/\epsilon)$, then, with probability (w.p.) at least $1 - n^{-10}$,

$$\text{SD}_2(\boldsymbol{U}, \boldsymbol{U}^\star) \le \epsilon \text{ and } \|\boldsymbol{x}_k - \boldsymbol{x}_k^\star\| \le \epsilon \|\boldsymbol{x}_k^\star\| \text{ for all } k \in [q].$$

The time complexity is $mqnr \cdot T = mqnr \cdot \kappa^2 \log(1/\epsilon)$. The communication complexity is $nr \cdot T = nr \cdot \kappa^2 \log(1/\epsilon)$ per node.

Proof: We prove the three results needed for proving this in Sections IV-B, IV-D, and V below. We use these to prove the above result in Appendix A. □

[1]The LS step time is $\max(q \cdot mnr, q \cdot mr^2) = mqnr$ (maximum of the time needed for computing $\boldsymbol{A}_k\boldsymbol{U}$ for all k, and that for obtaining $\boldsymbol{b}_k$ for all k) while the GD step time is $\max(q \cdot mnr, nr^2) = mqnr$ (maximum of the time needed for computing the gradient w.r.t. $\boldsymbol{U}$, and time for the QR step).

Algorithm 1 The AltGD-Min Algorithm

Input: $\boldsymbol{y}_k, \boldsymbol{A}_k, k \in [q]$
Sample-split: Partition the measurements and measurement matrices into $2T + 2$ equal-sized disjoint sets: two sets for initialization and $2T$ sets for the iterations. Denote these by $\boldsymbol{y}_k^{(\tau)}, \boldsymbol{A}_k^{(\tau)}, \tau = 00, 0, 1, \dots 2T$.
Initialization:
Using $\boldsymbol{y}_k \equiv \boldsymbol{y}_k^{(00)}, \boldsymbol{A}_k \equiv \boldsymbol{A}_k^{(00)}$,
set $\alpha \leftarrow \tilde{C} \frac{1}{mq} \sum_{ki} \left|\boldsymbol{y}_{ki}\right|^2$,
Using $\boldsymbol{y}_k \equiv \boldsymbol{y}_k^{(0)}, \boldsymbol{A}_k \equiv \boldsymbol{A}_k^{(0)}$,
set $\boldsymbol{y}_{k,trnc}(\alpha) \leftarrow \boldsymbol{y}_{k,trnc} := \text{trunc}(\boldsymbol{y}_k, \alpha)$,
set $\boldsymbol{X}_0 \leftarrow (1/m) \sum_{k \in [q]} \boldsymbol{A}_k^\top \boldsymbol{y}_{k,trunc}(\alpha) \boldsymbol{e}_k^\top$
set $\boldsymbol{U}_0 \leftarrow$ top-r-singular-vectors of $\boldsymbol{X}_0$
GDmin iterations:
for $t = 1$ **to** T **do**
  Let $\boldsymbol{U} \leftarrow \boldsymbol{U}_{t-1}$.
  Using $\boldsymbol{y}_k \equiv \boldsymbol{y}_k^{(t)}, \boldsymbol{A}_k \equiv \boldsymbol{A}_k^{(t)}$,
  set $\boldsymbol{b}_k \leftarrow (\boldsymbol{A}_k\boldsymbol{U})^\dagger \boldsymbol{y}_k$, $\boldsymbol{x}_k \leftarrow \boldsymbol{U}\boldsymbol{b}_k$ for all $k \in [q]$
  Using $\boldsymbol{y}_k \equiv \boldsymbol{y}_k^{(T+t)}, \boldsymbol{A}_k \equiv \boldsymbol{A}_k^{(T+t)}$, compute
  set $\nabla_{\boldsymbol{U}} f(\boldsymbol{U}, \boldsymbol{B}) = \sum_k \boldsymbol{A}_k^\top (\boldsymbol{A}_k \boldsymbol{U} \boldsymbol{b}_k - \boldsymbol{y}_k) \boldsymbol{b}_k^\top$
  set $\hat{\boldsymbol{U}}^+ \leftarrow \boldsymbol{U} - (\eta/m) \nabla_{\boldsymbol{U}} f(\boldsymbol{U}, \boldsymbol{B}_t)$.
  compute $\hat{\boldsymbol{U}}^+ \overset{\text{QR}}{=} \boldsymbol{U}^+ \boldsymbol{R}^+$.
  Set $\boldsymbol{U}_t \leftarrow \boldsymbol{U}^+$.
end for

A. Discussion

We use $a \gtrsim b$ to mean that $a \ge C_{\kappa,\mu} b$ where $C_{\kappa,\mu}$ includes terms dependent on κ, μ. Most of our discussion treats κ, μ as numerical constants that do not grow with n, q, r. This assumption is borrowed from the rich past literature on various other LR matrix recovery problems, e.g., see [4] and [5]. Also, below, whp means w.p. at least $1 - n^{-10}$.

In this work (as well as in other works that use matrix factorization to solve LR recovery problems), the goal is to obtain a bound of the form $\text{SD}_2(\boldsymbol{U}, \boldsymbol{U}^\star) \le \epsilon$. If a bound on SD_F is needed, one can use $\text{SD}_F(\boldsymbol{U}, \boldsymbol{U}^\star) \le \sqrt{r}\text{SD}_2(\boldsymbol{U}, \boldsymbol{U}^\star) \le \sqrt{r}\epsilon$. Using Lemma 3 given below and triangle inequality, the bound on SD_2 helps get the bound $\|\boldsymbol{x}_k - \boldsymbol{x}_k^\star\| \lesssim \epsilon \|\boldsymbol{x}_k^\star\|$.

Our older result from [3] used $\text{SD}_F(\boldsymbol{U}, \boldsymbol{U}^\star)$ in its analysis and showed that, to guarantee $\text{SD}_F(\boldsymbol{U}, \boldsymbol{U}^\star) \le \epsilon$ whp, we need

$$mq \gtrsim C\kappa^4\mu^2(n+q)r^2(\kappa^4 + \log(1/\epsilon))$$

Since $\text{SD}_2(\boldsymbol{U}, \boldsymbol{U}^\star) \le \text{SD}_F(\boldsymbol{U}, \boldsymbol{U}^\star)$, this implies that we need this same complexity also to guarantee $\text{SD}_2(\boldsymbol{U}, \boldsymbol{U}^\star) \le \epsilon$. Our new result needs $mq \ge C\kappa^4\mu^2(n+q)r(\kappa^4 r + \log(1/\epsilon))$ to guarantee $\text{SD}_2(\boldsymbol{U}, \boldsymbol{U}^\star) \le \epsilon$. Thus, this new result improves the dependence on r, ϵ from order $r^2 \log(1/\epsilon)$ to order $r \max(r, \log(1/\epsilon))$. This is an improvement over the old result by a factor of $\min(r, \log(1/\epsilon))$. This improvement is obtained because our new guarantee for the GD step (Theorem 6) only needs $mq \gtrsim nr$ at each iteration. On the other hand, the older result needed $mq \gtrsim nr^2$. Both guarantees need $mq \gtrsim nr^2$ for initialization, see Theorem 13.

We are able to improve our result because we now use a simpler proof technique that works for the LRCS problem, but

not for its LR phase retrieval (LRPR) generalization. LRPR involves recovering $\boldsymbol{X}^\star$ from $\boldsymbol{z}_k := |\boldsymbol{y}_k|, k \in [q]$. In [3], we were attempting to solve both problems. There are two differences in our new proof compared with that of [3]: (i) we use the 2-norm subspace distance $\mathrm{SD}_2(\boldsymbol{U}, \boldsymbol{U}^\star)$ instead of $\mathrm{SD}_F(\boldsymbol{U}, \boldsymbol{U}^\star)$, and (ii) we do not use the fundamental theorem of calculus [6], [7] for analyzing the GD step, but instead use a much simpler direct approach. If we only use (ii) but not (i), we will still get a simpler proof, but we will not get the sample complexity gain. In [3], we used the Frobenius norm SD because it helped obtain the desired nr^3 guarantee for LRPR.[2] Also, in hindsight, the use of the fundamental theorem of calculus was unnecessary. It has been used in earlier work [7] for analyzing a GD based algorithm for standard PR and for LR matrix completion and that originally motivated us to adapt the same approach for LRCS and LRPR.

Both the current result and our old one have the same dependence on κ. After initialization, the sample complexity grows roughly as κ^4. A similar dependence on κ also exists in all past sample complexity guarantees for iterative – AltMin or GD – algorithms for LR matrix sensing, LR matrix completion or robust PCA, e.g., see [4] and [5]; and also for our older work on AltMin for a generalization of LRCS, LR phase retrieval [8]. One way to improve this dependence is to develop a stage-wise algorithm similar to that introduced in [4] and [9] for LR matrix sensing and completion respectively. Developing this for the current LRCS problem is an open future work question. A second option is to use the convex relaxation approaches which usually depend on lower powers of κ. However, these need a much higher iteration complexity, making them much slower.

IV. New Proof: GDMin Iterations

A. Definitions and Preliminaries

Let $\boldsymbol{U}$ be the estimate at the t-th iteration. Define

$$\boldsymbol{g}_k := \boldsymbol{U}^\top \boldsymbol{x}_k^\star, k \in [q], \text{ and } \boldsymbol{G} := \boldsymbol{U}^\top \boldsymbol{X}^\star,$$
$$\boldsymbol{P} := \boldsymbol{I} - \boldsymbol{U}^\star \boldsymbol{U}^{\star\top},$$
$$\mathrm{GradU} := \nabla_{\boldsymbol{U}} f(\boldsymbol{U}, \boldsymbol{B}) = \sum_k \boldsymbol{A}_k^\top (\boldsymbol{A}_k \boldsymbol{U} \boldsymbol{b}_k - \boldsymbol{y}_k) \boldsymbol{b}_k^\top$$
$$= \sum_{ki} (\boldsymbol{y}_{ki} - \boldsymbol{a}_{ki}{}^\top \boldsymbol{U} \boldsymbol{b}_k) \boldsymbol{a}_{ki} \boldsymbol{b}_k{}^\top$$

For an $n_1 \times n_2$ matrix, $\boldsymbol{Z}$, $\sigma_{\min}(\boldsymbol{Z}) = \sigma_{\min}(\boldsymbol{Z}^\top) = \sigma_{\min(n_1,n_2)}(\boldsymbol{Z})$. Thus, if $\boldsymbol{A}$ is tall, then $\sigma_{\min}(\boldsymbol{A}) = \sqrt{\lambda_{\min}(\boldsymbol{A}^\top \boldsymbol{A})}$. Using this, it follows that, if $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C}$ and $\boldsymbol{A}$ and $\boldsymbol{B}$ are tall (or square), then $\sigma_{\min}(\boldsymbol{A}) \geq \sigma_{\min}(\boldsymbol{B}) \sigma_{\min}(\boldsymbol{C})$.

B. Minimization Step

Assume $\mathrm{SD}_2(\boldsymbol{U}, \boldsymbol{U}^\star) \leq \delta_t$ with $\delta_t < 0.02$.

We use the following lemma from [3].

Lemma 3 [3]: Let $\boldsymbol{g}_k := \boldsymbol{U}^\top \boldsymbol{x}_k^\star$. Then, w.p. at least $1 - \exp(\log q + r - c\epsilon m)$, for all $k \in [q]$,

$$\|\boldsymbol{g}_k - \boldsymbol{b}_k\| \leq 1.2\epsilon \| \left(\boldsymbol{I}_n - \boldsymbol{U}\boldsymbol{U}^\top \right) \boldsymbol{U}^\star \boldsymbol{b}_k^\star \|$$

[2]Our older guarantee for LRPR [3] needed $mq \gtrsim nr^2(r + \log(1/\epsilon))$. If we use the new approach developed here, it will need $mq \gtrsim nr(r^3 + \log(1/\epsilon))$.

Proof: We provide a proof of this lemma in Appendix B to emphasise a slightly more general version of this lemma. In particular, our proof shows that we can replace 0.4 in the previous version of this lemma by any $\epsilon > 0$, and the bound holds w.p. at least $1 - \exp(\log q + r - c\epsilon m)$. □

By Lemma 3 with $\epsilon = 0.3$, if $m \gtrsim \max(\log q, \log n, r)$, then, whp, for all $k \in [q]$,

$$\|\boldsymbol{b}_k - \boldsymbol{g}_k\| \leq 0.4 \|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{U}^\star \boldsymbol{b}_k^\star\|$$

Using $\mathrm{SD}_2(\boldsymbol{U}, \boldsymbol{U}^\star) \leq \delta_t$, this directly implies that

1) $\|\boldsymbol{b}_k - \boldsymbol{g}_k\| \leq 0.4\delta_t \|\boldsymbol{b}_k^\star\|$
2) $\|\boldsymbol{b}_k\| \leq \|\boldsymbol{g}_k\| + 0.4 \cdot 0.02 \|\boldsymbol{b}_k^\star\| \leq 1.1 \|\boldsymbol{b}_k^\star\|$
3) $\|\boldsymbol{x}_k - \boldsymbol{x}_k^\star\| \leq 1.4\delta_t \|\boldsymbol{b}_k^\star\|$

(to obtain the last bound, we need to add and subtract $\boldsymbol{U}\boldsymbol{g}_k$ and then use triangle inequality). Using these bounds,

$$\|\boldsymbol{B} - \boldsymbol{G}\|_F \leq 0.4\delta_t \sqrt{\sum_k \|\boldsymbol{b}_k^\star\|^2} = 0.4\delta_t \sqrt{r}\sigma_{\max}^\star$$

and we can get a similar bound on $\|\boldsymbol{X} - \boldsymbol{X}^\star\|_F$. Thus,

1) $\|\boldsymbol{B} - \boldsymbol{G}\|_F \leq 0.4\delta_t \|\boldsymbol{B}^\star\|_F \leq 0.4\sqrt{r}\delta_t \sigma_{\max}^\star$
2) $\|\boldsymbol{X} - \boldsymbol{X}^\star\|_F \leq 1.4\sqrt{r}\delta_t \sigma_{\max}^\star$

Furthermore,

$$\sigma_{\min}(\boldsymbol{B}) \geq \sigma_{\min}(\boldsymbol{G}) - \|\boldsymbol{B} - \boldsymbol{G}\| \geq \sigma_{\min}(\boldsymbol{G}) - \|\boldsymbol{B} - \boldsymbol{G}\|_F$$

To lower bound $\sigma_{\min}(\boldsymbol{G})$, observe that

$$\sigma_{\min}(\boldsymbol{G}) = \sigma_{\min}(\boldsymbol{G}^\top) \geq \sigma_{\min}^\star \sigma_{\min}(\boldsymbol{U}^{\star\top}\boldsymbol{U})$$

and

$$\sigma_{\min}(\boldsymbol{U}^{\star\top}\boldsymbol{U}) = \sqrt{1 - \|\boldsymbol{P}\boldsymbol{U}\|^2} \geq \sqrt{1 - \delta_t^2}.$$

This follows using $\sigma_{\min}^2(\boldsymbol{U}^{\star\top}\boldsymbol{U}) = \lambda_{\min}(\boldsymbol{U}^\top \boldsymbol{U}^\star \boldsymbol{U}^{\star\top}\boldsymbol{U}) = \lambda_{\min}(\boldsymbol{U}^\top(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{U}) = \lambda_{\min}(\boldsymbol{I} - \boldsymbol{U}^\top \boldsymbol{P}\boldsymbol{U}) = \lambda_{\min}(\boldsymbol{I} - \boldsymbol{U}^\top \boldsymbol{P}^2 \boldsymbol{U}) = 1 - \lambda_{\max}(\boldsymbol{U}^\top \boldsymbol{P}^2 \boldsymbol{U}) = 1 - \|\boldsymbol{P}\boldsymbol{U}\|^2$.

Combining the above three bounds and using the bound on $\|\boldsymbol{B} - \boldsymbol{G}\|_F$, if $\delta_t < 0.02/\sqrt{r}\kappa$, then

$$\sigma_{\min}(\boldsymbol{B}) \geq \sqrt{1 - \delta_t^2}\sigma_{\min}^\star - 0.4\sqrt{r}\delta_t \sigma_{\max}^\star \geq 0.9\sigma_{\min}^\star$$

and

$$\sigma_{\max}(\boldsymbol{B}) \leq \|\boldsymbol{G}\| + 0.4\sqrt{r}\delta_t \sigma_{\max}^\star \leq 1.1\sigma_{\max}^\star$$

since $\|\boldsymbol{G}\| \leq \|\boldsymbol{B}^\star\| = \sigma_{\max}^\star$. Thus, we have proved the following claim:

Theorem 4: Assume that $\mathrm{SD}_2(\boldsymbol{U}, \boldsymbol{U}^\star) \leq \delta_t$. If $\delta_t \leq 0.02/\sqrt{r}\kappa$, and if $m \gtrsim \max(\log q, \log n, r)$, then whp,

1) $\|\boldsymbol{b}_k - \boldsymbol{g}_k\| \leq 0.4\delta_t \|\boldsymbol{b}_k^\star\|$
2) $\|\boldsymbol{b}_k\| \leq \|\boldsymbol{g}_k\| + 0.4 \cdot 0.02 \|\boldsymbol{b}_k^\star\| \leq 1.1 \|\boldsymbol{b}_k^\star\|$
3) $\|\boldsymbol{B} - \boldsymbol{G}\|_F \leq 0.4\delta_t \|\boldsymbol{B}^\star\|_F \leq 0.4\sqrt{r}\delta_t \sigma_{\max}^\star$
4) $\|\boldsymbol{x}_k - \boldsymbol{x}_k^\star\| \leq 1.4\delta_t \|\boldsymbol{b}_k^\star\|$
5) $\|\boldsymbol{X} - \boldsymbol{X}^\star\|_F \leq 1.4\sqrt{r}\delta_t \sigma_{\max}^\star$
6) $\sigma_{\min}(\boldsymbol{B}) \geq 0.9\sigma_{\min}^\star$
7) $\sigma_{\max}(\boldsymbol{B}) \leq 1.1\sigma_{\max}^\star$

(only the last two bounds require the upper bound on δ_t).

C. New Bounds on the Expected Gradient and Deviation From it

Using independence of $\boldsymbol{A}_k$ and $\{\boldsymbol{U}, \boldsymbol{b}_k\}$ (due to sample splitting),

$$\mathbb{E}[\text{GradU}] = \sum_k m(\boldsymbol{x}_k - \boldsymbol{x}_k^\star)\boldsymbol{b}_k{}^\top$$

Using bounds on $\|\boldsymbol{B}\|$ and $\|\boldsymbol{X}^\star - \boldsymbol{X}\|_F$ from Theorem 4, if $\delta_t < \frac{c}{\sqrt{r}\kappa}$,

$$\begin{aligned}\|\mathbb{E}[\text{GradU}]\| &= \|\sum_k m(\boldsymbol{x}_k - \boldsymbol{x}_k^\star)\boldsymbol{b}_k{}^\top\| = m\|(\boldsymbol{X}-\boldsymbol{X}^\star)\boldsymbol{B}^\top\| \\ &\le m\|\boldsymbol{X}-\boldsymbol{X}^\star\|\cdot\|\boldsymbol{B}\| \\ &\le m\|\boldsymbol{X}-\boldsymbol{X}^\star\|_F\cdot\|\boldsymbol{B}\| \le 1.1m\delta_t\sqrt{r}{\sigma_{\max}^\star}^2.\end{aligned}$$

w.p. $1-\exp(\log q + r - cm)$.

Next, we bound $\|\text{GradU} - \mathbb{E}[\text{GradU}]\| = \max_{\|\boldsymbol{w}\|=1,\|\boldsymbol{z}\|=1} \boldsymbol{w}^\top(\sum_k\sum_i \boldsymbol{a}_{ki}\boldsymbol{a}_{ki}^\top(\boldsymbol{x}_k - \boldsymbol{x}_k^\star)\boldsymbol{b}_k^\top - \mathbb{E}[\cdot])\boldsymbol{z}$. This also uses independence of $\boldsymbol{A}_k$ and $\{\boldsymbol{U}, \boldsymbol{b}_k\}$.

We bound the above for fixed unit norm $\boldsymbol{w}, \boldsymbol{z}$ using sub-exponential Bernstein inequality (Theorem 2.8.1 of [10]). We extend the bound to all unit norm $\boldsymbol{w}, \boldsymbol{z}$ by using a standard epsilon-net argument. For fixed unit norm $\boldsymbol{w}, \boldsymbol{z}$, consider

$$\sum_k\sum_i \Big((\boldsymbol{w}^\top\boldsymbol{a}_{ki})(\boldsymbol{b}_k^\top\boldsymbol{z})\boldsymbol{a}_{ki}^\top(\boldsymbol{x}_k - \boldsymbol{x}_k^\star) - \mathbb{E}[\cdot]\Big)$$

Observe that the summands are independent, zero mean, sub-exponential r.v.s with sub-exponential norm $K_{ki} \le C\|\boldsymbol{w}\|\cdot|\boldsymbol{z}^\top\boldsymbol{b}_k|\cdot\|\boldsymbol{x}_k - \boldsymbol{x}_k^\star\| = C|\boldsymbol{z}^\top\boldsymbol{b}_k|\cdot\|\boldsymbol{x}_k - \boldsymbol{x}_k^\star\|$. We apply the sub-exponential Bernstein inequality, Theorem 2.8.1 of [10], with $t = \epsilon_1\delta_t m{\sigma_{\min}^\star}^2$. We have

$$\begin{aligned}\frac{t^2}{\sum_{ki}K_{ki}^2} &\ge \frac{\epsilon_1^2\delta_t^2 m^2{\sigma_{\min}^\star}^4}{m\sum_k\|\boldsymbol{x}_k-\boldsymbol{x}_k^\star\|^2(\boldsymbol{z}^\top\boldsymbol{b}_k)^2} \\ &\ge \frac{\epsilon_1^2\delta_t^2 m^2{\sigma_{\min}^\star}^4}{m\max_k\|\boldsymbol{x}_k-\boldsymbol{x}_k^\star\|^2\sum_k(\boldsymbol{z}^\top\boldsymbol{b}_k)^2} \\ &= \frac{\epsilon_1^2\delta_t^2 m^2{\sigma_{\min}^\star}^4}{m\max_k\|\boldsymbol{x}_k-\boldsymbol{x}_k^\star\|^2\|\boldsymbol{z}^\top\boldsymbol{B}\|^2} \\ &\ge \frac{\epsilon_1^2\delta_t^2 m^2{\sigma_{\min}^\star}^4 q}{1.4^2 m\mu^2\delta_t^2 r{\sigma_{\max}^\star}^2\|\boldsymbol{B}\|^2} \\ &\ge \frac{\epsilon_1^2\delta_t^2 m^2{\sigma_{\min}^\star}^4 q}{1.4^2 m\mu^2\delta_t^2 r{\sigma_{\max}^\star}^2 1.1{\sigma_{\max}^\star}^2} = c\frac{\epsilon_1^2 mq}{\kappa^4\mu^2 r}\end{aligned}$$

$$\frac{t}{\max_{ki}K_{ki}} \ge \frac{\epsilon_1\delta_t m{\sigma_{\min}^\star}^2}{\max_k\|\boldsymbol{b}_k\|\max_k\|\boldsymbol{x}_k-\boldsymbol{x}_k^\star\|} \ge c\frac{\epsilon_1 mq}{\kappa^2\mu^2 r}$$

In the above, we used (i) $\sum_k(\boldsymbol{z}^\top\boldsymbol{b}_k)^2 = \|\boldsymbol{z}^\top\boldsymbol{B}\|^2 \le \|\boldsymbol{B}\|^2$ since $\boldsymbol{z}$ is unit norm, (ii) Theorem 4 to bound $\|\boldsymbol{B}\| \le 1.1\sigma_{\max}^\star$, and (iii) Theorem 4 followed by Assumption 1 (right incoherence) to bound $\|\boldsymbol{x}_k - \boldsymbol{x}_k^\star\| \le \delta_t\cdot\mu\sigma_{\max}^\star\sqrt{r/q}$ and $|\boldsymbol{z}^\top\boldsymbol{b}_k| \le \|\boldsymbol{b}_k\| \le 1.1\|\boldsymbol{b}_k^\star\| \le 1.1\mu\sigma_{\max}^\star\sqrt{r/q}$.

For $\epsilon_1 < 1$, the first term above is smaller (since $1/\kappa^4 \le 1/\kappa^2$), i.e., $\min(\frac{t^2}{\sum_{ki}K_{ki}^2}, \frac{t}{\max_{ki}K_{ki}}) = c\frac{\epsilon_1^2 mq}{\kappa^4\mu^2 r}$. Thus, by sub-exponential Bernstein, w.p. at least $1-\exp(-c\frac{\epsilon_1^2 mq}{\kappa^4\mu^2 r}) - \exp(\log q + r - cm)$, for a given $\boldsymbol{w}, \boldsymbol{z}$,

$$\boldsymbol{w}^\top(\text{GradU} - \mathbb{E}[\text{GradU}])\boldsymbol{z} \le \epsilon_1\delta_t m{\sigma_{\min}^\star}^2$$

Using a standard epsilon-net argument to bound the maximum of the above over all unit norm $\boldsymbol{w}, \boldsymbol{z}$, e.g., using [3, Proposition 4.7], we can conclude that

$$\|\text{GradU} - \mathbb{E}[\text{GradU}]\| \le 1.1\epsilon_1\delta_t m{\sigma_{\min}^\star}^2$$

w.p. at least $1-\exp(C(n+r)-c\frac{\epsilon_1^2 mq}{\kappa^4\mu^2 r})-\exp(\log q+r-cm)$. The factor of $\exp(C(n+r))$ is due to the epsilon-net over $\boldsymbol{w}$ and that over $\boldsymbol{z}$: $\boldsymbol{w}$ is an n-length unit norm vector while $\boldsymbol{z}$ is an r-length unit norm vector. The smallest epsilon net covering the hyper-sphere of all $\boldsymbol{w}$s is of size $(1+2/\epsilon_{net})^n = C^n$ with $\epsilon_{net} = c$ while that for $\boldsymbol{z}$ is of size C^r. Union bounding over both thus gives a factor of C^{n+r}. By replacing ϵ_1 by $\epsilon_1/1.1$, our bound becomes simpler (and $1/1.1^2$ gets incorporated into the factor c). We have thus proved the following.

Lemma 5: Assume that $\text{SD}_2(\boldsymbol{U}, \boldsymbol{U}^\star) \le \delta_t$. The following hold:

1) $\mathbb{E}[\text{GradU}] = m(\boldsymbol{X}-\boldsymbol{X}^\star)\boldsymbol{B}^\top = m(\boldsymbol{U}\boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{X}^\star\boldsymbol{B}^\top)$
2) $\|\mathbb{E}[\text{GradU}]\| \le 1.1m\delta_t\sqrt{r}{\sigma_{\max}^\star}^2$
3) If $\delta_t < \frac{c}{\sqrt{r}\kappa}$, then, w.p. at least
$1-\exp(C(n+r)-c\frac{\epsilon_1^2 mq}{\kappa^4\mu^2 r})-\exp(\log q+r-cm)$,
$$\|\text{GradU} - \mathbb{E}[\text{GradU}]\| \le \epsilon_1\delta_t m{\sigma_{\min}^\star}^2$$

The above lemma is an improvement over the bounds given in [3] because δ_t is now the bound on the 2-norm SD, and still it only needs $mq \gtrsim nr/\epsilon_1^2$.

D. GD Step

Assume that $\text{SD}_2(\boldsymbol{U}, \boldsymbol{U}^\star) \le \delta_t$ with $\delta_t < 0.02$.

Recall the Projected GD step for $\boldsymbol{U}$:

$$\widetilde{\boldsymbol{U}}^+ = \boldsymbol{U} - \eta\text{GradU} \text{ and } \widetilde{\boldsymbol{U}}^+ \overset{\text{QR}}{=} \boldsymbol{U}^+\boldsymbol{R}^+$$

Since $\boldsymbol{U}^+ = \widetilde{\boldsymbol{U}}^+(\boldsymbol{R}^+)^{-1}$ and since $\|(\boldsymbol{R}^+)^{-1}\| = 1/\sigma_{\min}(\boldsymbol{R}^+) = 1/\sigma_{\min}(\widetilde{\boldsymbol{U}}^+)$, thus, $\text{SD}_2(\boldsymbol{U}^+, \boldsymbol{U}^\star) = \|\boldsymbol{P}\boldsymbol{U}^+\|$ can be bounded as

$$\text{SD}_2(\boldsymbol{U}^+, \boldsymbol{U}^\star) \le \frac{\|\boldsymbol{P}\widetilde{\boldsymbol{U}}^+\|}{\sigma_{\min}(\widetilde{\boldsymbol{U}}^+)} \le \frac{\|\boldsymbol{P}\widetilde{\boldsymbol{U}}^+\|}{\sigma_{\min}(\boldsymbol{U}) - \eta\|\text{GradU}\|} \quad (2)$$

Consider the numerator. Adding/subtracting $\mathbb{E}[\text{GradU}]$, left multiplying both sides by $\boldsymbol{P}$, and using Lemma 5 (first part),

$$\begin{aligned}\widetilde{\boldsymbol{U}}^+ &= \boldsymbol{U} - \eta\mathbb{E}[\text{GradU}] + \eta(\mathbb{E}[\text{GradU}] - \text{GradU}), \text{ thus,} \\ \boldsymbol{P}\widetilde{\boldsymbol{U}}^+ &= \boldsymbol{P}\boldsymbol{U} - \eta m\boldsymbol{P}\boldsymbol{U}\boldsymbol{B}\boldsymbol{B}^\top + \eta\boldsymbol{P}(\mathbb{E}[\text{GradU}] - \text{GradU})\end{aligned}$$

The last row used $\boldsymbol{P}\boldsymbol{X}^\star = 0$. Thus,

$$\|\boldsymbol{P}\widetilde{\boldsymbol{U}}^+\| \le \|\boldsymbol{P}\boldsymbol{U}\|\|\boldsymbol{I}-\eta m\boldsymbol{B}\boldsymbol{B}^\top\|+\eta\|\mathbb{E}[\text{GradU}] - \text{GradU}\| \quad (3)$$

Using Theorem 4, we get

$$\lambda_{\min}(\boldsymbol{I} - \eta m\boldsymbol{B}\boldsymbol{B}^\top) = 1 - \eta m\|\boldsymbol{B}\|^2 \ge 1 - 1.2\eta m{\sigma_{\max}^\star}^2$$

Thus, if $\eta < 0.5/m{\sigma_{\max}^\star}^2$, then the above matrix is p.s.d. This along with Theorem 4 then implies that

$$\|\boldsymbol{I} - \eta m\boldsymbol{B}\boldsymbol{B}^\top\| = \lambda_{\max}(\boldsymbol{I} - \eta m\boldsymbol{B}\boldsymbol{B}^\top) \le 1 - 0.9\eta m{\sigma_{\min}^\star}^2$$

Using the above, (3), and the bound on $\|\mathbb{E}[\text{GradU}] - \text{GradU}\|$ from Lemma 5, we conclude the following: If $\eta \leq 0.5/m{\sigma^\star_{\max}}^2$, and $\delta_t \leq c/\sqrt{r}\kappa$, then

$$\begin{aligned}\|\boldsymbol{P}\tilde{\boldsymbol{U}}^+\| &\leq \|\boldsymbol{P}\boldsymbol{U}\|\|\boldsymbol{I}-\eta m\boldsymbol{B}\boldsymbol{B}^\top\|+ !\eta\|\mathbb{E}[\text{GradU}] - \text{GradU}\| \\ &\leq \delta_t(1-0.9\ \eta m{\sigma^\star_{\min}}^2) + \eta m\epsilon_1\delta_t{\sigma^\star_{\min}}^2 \end{aligned} \tag{4}$$

w.p. at least $1-\exp(C(n+r)-c\frac{\epsilon_1^2 mq}{\kappa^4\mu^2 r})-\exp(\log q+r-cm)$. This probability is at least $1-n^{-10}$ if $mq \gtrsim \kappa^4\mu^2 nr/\epsilon_1^2$ and $m \gtrsim \max(\log n, \log q, r)$.

Next we use (4) with $\epsilon_1 = 0.1$ and Lemma 5 to bound the right hand side of (2). Set $\eta = c_\eta/m{\sigma^\star_{\max}}^2$. If $c_\eta \leq 0.5$, if $\delta_t \leq c/\sqrt{r}\kappa^2$, and lower bounds on m from above hold, (2) implies that, whp,

$$\begin{aligned}&\text{SD}_2(\boldsymbol{U}^+, \boldsymbol{U}^\star) \\ &\leq \frac{\|\boldsymbol{P}\tilde{\boldsymbol{U}}^+\|}{\sigma_{\min}(\tilde{\boldsymbol{U}}^+)} \\ &\leq \frac{\|\boldsymbol{P}\tilde{\boldsymbol{U}}^+\|}{\sigma_{\min}(\boldsymbol{U}) - \eta\|\text{GradU}\|} \\ &\leq \frac{\|\boldsymbol{P}\boldsymbol{U}\|\|\boldsymbol{I}-\eta m\boldsymbol{B}\boldsymbol{B}^\top\| + \eta\|\mathbb{E}[\text{GradU}] - \text{GradU}\|}{1-\eta\|\mathbb{E}[\text{GradU}]\| - \eta\|\text{GradU} - \mathbb{E}[\text{GradU}]} \\ &\leq \frac{\delta_t(1-\eta m{\sigma^\star_{\min}}^2(0.9-0.1))}{1-\eta\|\mathbb{E}[\text{GradU}]\| - \eta\|\text{GradU} - \mathbb{E}[\text{GradU}]\|} \\ &\leq \frac{\delta_t(1-0.8\eta m{\sigma^\star_{\min}}^2)}{1-\eta m\delta_t\sqrt{r}{\sigma^\star_{\max}}^2(1.4+\frac{0.1}{\kappa^2\sqrt{r}})} \\ &\leq \frac{\delta_t(1-0.8\eta m{\sigma^\star_{\min}}^2)}{1-1.5\eta m\delta_t\sqrt{r}{\sigma^\star_{\max}}^2} \\ &\leq \delta_t(1-0.8\eta m{\sigma^\star_{\min}}^2)(1+1.5\eta m\delta_t\sqrt{r}{\sigma^\star_{\max}}^2) \\ &\leq \delta_t(1-0.8\eta m{\sigma^\star_{\min}}^2+1.5\eta m\delta_t\sqrt{r}{\sigma^\star_{\max}}^2) \\ &= \delta_t(1-\eta m{\sigma^\star_{\min}}^2(0.8-1.5\delta_t\sqrt{r}\kappa^2)) \\ &\leq \delta_t(1-\eta m{\sigma^\star_{\min}}^2(0.8-0.15)) \\ &\leq \delta_t(1-0.6\eta m{\sigma^\star_{\max}}^2/\kappa^2) = \delta_t(1-0.6c_\eta/\kappa^2)\end{aligned}$$

In the above we used $\kappa^2\sqrt{r} > 1$, $(1-x)^{-1} < (1+x)$ if $|x| < 1$, and $\delta_t < 0.1/\sqrt{r}\kappa^2$ (used in second-last inequality).

Thus, we have proved the following result.

Theorem 6: Assume that Assumption 1 holds and $\text{SD}_2(\boldsymbol{U}, \boldsymbol{U}^\star) \leq \delta_t$. If $\delta_t \leq 0.02/\sqrt{r}\kappa^2$, if $\eta = c_\eta/m{\sigma^\star_{\max}}^2$ with $c_\eta \leq 0.5$, and if $mq \gtrsim \kappa^4\mu^2 nr$ and $m \gtrsim \max(\log n, \log q, r)$, then, whp

$$\text{SD}_2(\boldsymbol{U}^+, \boldsymbol{U}^\star) \leq \delta_{t+1} := \delta_t(1-0.6c_\eta/\kappa^2)$$

V. New Proof: Initialization

We need a new result for the initialization step because we need a tight bound on $\text{SD}_2(\boldsymbol{U}_0, \boldsymbol{U}^\star)$.

A. Results Taken From [3]

Recall from Algorithm 1 that α uses a different set of measurements that is independent of those used for $\boldsymbol{X}_0$. We use the following four results from [3].

Lemma 7 [3]: Conditioned on α, we have the following conclusions. Let ζ be a scalar standard Gaussian r.v.. Define

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

Then,

$$\begin{aligned}&\mathbb{E}[\boldsymbol{X}_0|\alpha] = \boldsymbol{X}^\star\boldsymbol{D}(\alpha), \\ &\text{where } \boldsymbol{D}(\alpha) := diagonal(\beta_k(\alpha), k \in [q])\end{aligned}$$

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

Fact 8 [3]: Let

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

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

Fact 9 [3]: For any $\epsilon_1 \leq 0.1$, $\min_k \mathbb{E}\left[\zeta^2 \mathbb{1}_{\left\{|\zeta| \leq \tilde{C}\frac{\sqrt{1-\epsilon_1}\|\boldsymbol{X}^\star\|_F}{\sqrt{q}\|\boldsymbol{x}_k^\star\|}\right\}}\right] \geq 0.92.$

Lemma 10 [3]: Fix $0 < \epsilon_1 < 1$. Then, w.p. at least $1-\exp\left[(n+q)-c\epsilon_1^2 mq/\mu^2\kappa^2\right]$, conditioned on α, for an $\alpha \in \mathcal{E}$,

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

Facts 8 and 9 together imply that, w.p. at least $1-\exp(-\tilde{c}mq\epsilon_1^2)$,

$$\min_k \beta_k(\alpha) \geq \min_k \mathbb{E}\left[\zeta^2 \mathbb{1}_{\{|\zeta| \leq \tilde{C}\frac{\sqrt{1-\epsilon_1}\|\boldsymbol{X}^\star\|_F}{\sqrt{q}\|\boldsymbol{x}_k^\star\|}\}}\right] \geq 0.92. \tag{5}$$

The first inequality is an immediate consequence of Fact 8 ($\alpha \in \mathcal{E}$) and the second follows by Fact 9.

By setting $\epsilon_1 = \epsilon_0/1.1\sqrt{r}\kappa$ in Lemma 10, and using Fact 8, we get the following corollary.

Corollary 11: Fix $0 < \epsilon_1 < 1$. Then, w.p. at least $1-\exp\left[(n+q)-c\frac{\epsilon_0^2 mq}{\mu^2\kappa^4 r}\right]-\exp(-cmq\epsilon_1^2/\kappa^2\mu^2)$,

$$\|\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha]\| \leq \epsilon_0\sigma^\star_{\min}$$

with $\mathbb{E}[\boldsymbol{X}_0|\alpha]$ being as given on Lemma 7.

B. Obtaining the SD Error Bound for Initialization

By Lemma 7, $\mathbb{E}[\boldsymbol{X}_0|\alpha] = \boldsymbol{X}^\star\boldsymbol{D}(\alpha)$ with $\boldsymbol{D}(\alpha)$ as defined there. Clearly, its rank is r or less. To obtain the bound, we apply Wedin's $\sin\theta$ theorem [11] for SD_2 with $\boldsymbol{M} = \boldsymbol{X}_0$, $\boldsymbol{M}^* = \mathbb{E}[\boldsymbol{X}_0|\alpha] = \boldsymbol{X}^\star\boldsymbol{D}$. Recall that $\boldsymbol{U}_0$ is the matrix of top r singular vectors of $\boldsymbol{X}_0$. Also, the span of top r singular vectors of $\mathbb{E}[\boldsymbol{X}_0|\alpha] = \boldsymbol{X}^\star\boldsymbol{D}$ equals the column span of $\boldsymbol{U}^\star$. Thus applying Wedin will help us bound $\text{SD}_2(\boldsymbol{U}_0, \boldsymbol{U}^\star)$. To do this, we need to define the SVD of $\mathbb{E}[\boldsymbol{X}_0|\alpha]$. Let

$$\mathbb{E}[\boldsymbol{X}_0|\alpha] = \boldsymbol{X}^\star\boldsymbol{D}(\alpha) \overset{\text{SVD}}{=} (\boldsymbol{U}^\star\boldsymbol{Q})\check{\boldsymbol{\Sigma}}^*\check{\boldsymbol{V}}$$

where $\boldsymbol{Q}$ is a $r \times r$ unitary matrix, $\check{\boldsymbol{\Sigma}}^*$ is an $r \times r$ diagonal matrix with non-negative entries (singular values) and $\check{\boldsymbol{V}}$ is an $r \times q$ matrix with orthonormal rows, i.e. $\check{\boldsymbol{V}}\check{\boldsymbol{V}}^\top = \boldsymbol{I}$. Observe also that $\sigma_{r+1}(\mathbb{E}[\boldsymbol{X}_0|\alpha]) = 0$ since it is a rank r matrix. Also, from above,

$$\sigma_r(\mathbb{E}[\boldsymbol{X}_0|\alpha]) = \sigma_{\min}(\check{\boldsymbol{\Sigma}}^*) \geq \sigma^\star_{\min}\sigma_{\min}(\boldsymbol{D}) = \sigma^\star_{\min}\min_k \beta_k(\alpha)$$

This follows since $\check{\boldsymbol{\Sigma}}^* = \boldsymbol{Q}^\top \boldsymbol{B}^\star \boldsymbol{D} \check{\boldsymbol{V}}^\top$ and so $\sigma_{\min}(\check{\boldsymbol{\Sigma}}^*) \geq \sigma_{\min}(\boldsymbol{B}^\star \boldsymbol{D} \check{\boldsymbol{V}}^\top) \geq \sigma_{\min}(\boldsymbol{B}^\star \boldsymbol{D}) \cdot 1 \geq \sigma_{\min}(\boldsymbol{D} \boldsymbol{B}^{\star\top}) \geq \sigma_{\min}(\boldsymbol{D}) \cdot \sigma^\star_{\min}$ and $\sigma_{\min}(\boldsymbol{D}) = \min_k \beta_k$. The last equality follows since $\boldsymbol{D}$ is diagonal with entries β_k. Thus,

Fact 12: $\sigma_r(\mathbb{E}[\boldsymbol{X}_0|\alpha]) \geq \sigma^\star_{\min} \min_k \beta_k(\alpha)$, $\sigma_{r+1}(\mathbb{E}[\boldsymbol{X}_0|\alpha]) = 0$.

Applying Wedin's $\sin\theta$ theorem, and then using Corollary 11, Fact 12 and (5), we get

$$\begin{aligned}
&\mathrm{SD}_2(\boldsymbol{U}_0, \boldsymbol{U}^\star) \\
&\leq \sqrt{2} \frac{\max(\|(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha])^\top \boldsymbol{U}^\star\|, \|(\boldsymbol{X}_0 - \mathbb{E}[X_0|\alpha])\check{\boldsymbol{V}}\|)}{\sigma^\star_{\min} \min_k \beta_k(\alpha) - 0 - \|\boldsymbol{X}_0 - \mathbb{E}[X_0|\alpha]\|} \\
&\lesssim \sqrt{2} \frac{\epsilon_0 \sigma^\star_{\min}}{0.92\sigma^\star_{\min} - \epsilon_0 \sigma^\star_{\min}} \lesssim \sqrt{2} \frac{\epsilon_0}{0.9} < 1.6\epsilon_0
\end{aligned}$$

if $\epsilon_0 < 0.02$ and $mq \gtrsim (n+q)r/\epsilon_0^2$. In the above we used $\|(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha])^\top \boldsymbol{U}^\star\| \leq \|\boldsymbol{X}_0 - \mathbb{E}[X_0|\alpha]\| \cdot 1$ and $\|(\boldsymbol{X}_0 - \mathbb{E}[\boldsymbol{X}_0|\alpha])\check{\boldsymbol{V}}\| \leq \|\boldsymbol{X}_0 - \mathbb{E}[X_0|\alpha]\| \cdot 1$. Setting $\epsilon_0 = 0.5\delta_0$ with $\delta_0 < 0.02$, we obtain our desired result.

Theorem 13: Assume that Assumption 1 holds. Pick a $\delta_0 \leq 0.02$. If $mq \gtrsim \kappa^4 \mu^2 (n+q) r / \delta_0^2$ then whp

$$\mathrm{SD}_2(\boldsymbol{U}_0, \boldsymbol{U}^\star) \leq \delta_0$$

Appendix A
Proof of Theorem 2

Theorem 13 tells us that $\mathrm{SD}_2(\boldsymbol{U}_0, \boldsymbol{U}^\star) \leq \delta_0$ whp if $m_0 q \gtrsim (n+q)r/\delta_0^2$. Theorem 6 tells us that if $\delta_t \leq 0.02/\sqrt{r}\kappa^2$, and if $m_1 q \gtrsim nr$, then, whp, δ_t reduces by a factor of $(1 - 0.6c_\eta/\kappa^2)$ at each iteration. In particular, this implies that $\delta_t \leq \delta_0$. Thus, in order to apply Theorem 6, it suffices to require $\delta_0 = 0.02/\sqrt{r}\kappa^2$.

Combining both results, we have shown that if $m_0 q \geq C\kappa^8 \mu^2 (n+q) r^2$ and if $m_1 q \geq C\kappa^4 m u^2 nr$ and $m_1 \geq C \max(r, \log n, \log r)$, and if $\eta = c_\eta/(m{\sigma^\star_{\max}}^2)$ with $c_\eta < 0.5$, then, whp, at each $t \geq 0$,

$$\mathrm{SD}_2(\boldsymbol{U}_t, \boldsymbol{U}^\star) \leq \delta_t := (1 - 0.6\frac{c_\eta}{\kappa^2})^t \delta_0 = (1 - 0.6\frac{c_\eta}{\kappa^2})^t \frac{0.02}{\sqrt{r}\kappa^2}$$

and all bounds of Theorem 4 hold with δ_t as above.

Thus, to guarantee $\mathrm{SD}_2(\boldsymbol{U}_T, \boldsymbol{U}^\star) \leq \epsilon$, we need

$$T \geq C \frac{\kappa^2}{c_\eta} \log(1/\epsilon)$$

This follows by using $\log(1-x) < -x$ for $|x| < 1$ and using $\kappa^2\sqrt{r} \geq 1$. Thus, setting $c_\eta = 0.4$, our sample complexity $m = m_0 + m_1 T$ becomes $mq \geq C\kappa^8 \mu^2 (n+q) r (1 + \log(1/\epsilon))$, and $m \geq C \max(r, \log q, \log n) \log(1/\epsilon)$.

Appendix B
Proof of Lemma 3

Using the expression for $\boldsymbol{b}_k$ and simplifying it,

$$\boldsymbol{b}_k - \boldsymbol{g}_k = \boldsymbol{M}^{-1} \boldsymbol{U}^\top \boldsymbol{A}_k^\top \boldsymbol{A}_k (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top) \boldsymbol{U}^\star \boldsymbol{b}_k^\star$$

with $\boldsymbol{M} := \boldsymbol{U}^\top \boldsymbol{A}_k^\top \boldsymbol{A}_k \boldsymbol{U}$.

Clearly,

$$\mathbb{E}[\boldsymbol{M}] = m\boldsymbol{U}^\top \boldsymbol{U} = m\boldsymbol{I}_r, \ \mathbb{E}[\boldsymbol{U}^\top \boldsymbol{A}_k^\top \boldsymbol{A}_k (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top) \boldsymbol{U}^\star \boldsymbol{b}_k^\star] = 0$$

Thus, using the standard technique (sub-expo Bern ineq followed by an epsilon-net argument), one can show that

1) w.p. at least $1 - \exp(r - \epsilon_2 m)$,

$$\|\boldsymbol{M} - \boldsymbol{I}_r\| \leq 1.1\epsilon_2 m$$

This implies of course that $\sigma^\star_{\min}(\boldsymbol{M}) \geq (1 - 1.1\epsilon_2)m$ and thus $\|\boldsymbol{M}^{-1}\| \leq 1/((1 - 1.1\epsilon_2)m)$.

2) w.p. at least $1 - \exp(r - \epsilon_3 m)$,

$$\begin{aligned}
&\|\boldsymbol{U}^\top \boldsymbol{A}_k^\top \boldsymbol{A}_k (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top) \boldsymbol{U}^\star \boldsymbol{b}_k^\star\| \\
&\qquad \leq 1.1\epsilon_3 m \|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top) \boldsymbol{U}^\star \boldsymbol{b}_k^\star\|
\end{aligned}$$

Thus, setting $\epsilon_2 = 0.1$, and taking union bound over all q vectors, w.p. at least $1 - q\exp(r - \epsilon_3 m)$,

$$\|\boldsymbol{b}_k - \boldsymbol{g}_k\| \leq 1.2\epsilon_3 \|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top) \boldsymbol{U}^\star \boldsymbol{b}_k^\star\| \text{ for all } k \in [q]$$

References

[1] S. Babu, S. G. Lingala, and N. Vaswani, "Fast low rank compressive sensing for accelerated dynamic MRI," *IEEE Trans. Comput. Imag.*, vol. 9, pp. 409–424, 2023, doi: 10.1109/TCI.2023.3263810.

[2] S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei, "Few-shot learning via learning the representation, provably," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[3] S. Nayer and N. Vaswani, "Fast and sample-efficient federated low rank matrix recovery from column-wise linear and quadratic projections," *IEEE Trans. Inf. Theory*, vol. 69, no. 2, pp. 1177–1202, Feb. 2023.

[4] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proc. 45th Annu. ACM Symp. Theory Comput.*, Jun. 2013.

[5] X. Yi, D. Park, Y. Chen, and C. Caramanis, "Fast algorithms for robust PCA via gradient descent," in *Proc. NeurIPS*, 2016.

[6] S. Lang, *Real and Functional Analysis*, vol. 10. New York, NY, USA: Springer, 1993, pp. 11–13.

[7] 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 *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018.

[8] S. Nayer and N. Vaswani, "Sample-efficient low rank phase retrieval," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 2244–2249.

[9] Y. Cherapanamjeri, K. Gupta, and P. Jain, "Nearly-optimal robust matrix completion," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016.

[10] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*, vol. 47. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[11] P. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT Numer. Math.*, vol. 12, no. 1, pp. 99–111, 1972.

Namrata Vaswani (Fellow, IEEE) received the B.Tech. degree from IIT Delhi, India, in 1999, and the Ph.D. degree from the University of Maryland, College Park, in 2004.

She is currently the Anderlik Professor in electrical and computer engineering with Iowa State University, where she is also the Director of the CyMath Program, in which graduate students provided school-year-long math tutoring for under-served grade school students. Her research interests are in data science, with a particular focuses on statistical machine learning for signal processing and computational imaging. She was a recipient of the 2014 IEEE Signal Processing Society Best Paper Award, the Iowa State Early Career Engineering Faculty Research Award in 2014, and the Iowa State Mid-Career Achievement in Research Award in 2019. She has served as an Associate Editor or an Area Editor for IEEE Transactions on Information Theory, IEEE Transactions on Signal Processing, and the *Signal Processing Magazine*. She has also guest-edited special issues for Proceedings of the IEEE and the IEEE Journal of Selected Topics in Signal Processing.