AN IMPROVED SAMPLE COMPLEXITY FOR RANK-1 MATRIX SENSING

Zhihang Li * **Zhizhou Sha**† **Zhao Song**‡ **Mingda Wan**§

ABSTRACT

Matrix sensing is a problem in signal processing and machine learning that involves recovering a low-rank matrix from a set of linear measurements. The goal is to reconstruct the original matrix as accurately as possible, given only a set of linear measurements obtained by sensing the matrix [Jain, Netrapalli, Sanghavi, STOC 2013]. In this work, we focus on a particular direction of matrix sensing, which is called rank-1 matrix sensing [Zhong, Jian, Dhillon, ALT 2015]. We present an improvement over the original algorithm in [Zhong, Jian, Dhillon, ALT 2015]. It is based on a novel analysis and sketching technique that enables faster convergence rates and better accuracy in recovering low-rank matrices. The algorithm focuses on developing a theoretical understanding of the matrix sensing problem and establishing its advantages over previous methods. The proposed sketching technique allows for efficiently extracting relevant information from the linear measurements, making the algorithm computationally efficient and scalable.

Our novel matrix sensing algorithm improves former result [Zhong, Jian, Dhillon, ALT 2015] on in two senses,

- We improve the sample complexity from $\widetilde{O}(\epsilon^{-2}dk^2)$ to $\widetilde{O}(\epsilon^{-2}(d+k^2))$.
- We improve the running time from $\widetilde{O}(md^2k^2)$ to $\widetilde{O}(md^2k)$.

The proposed algorithm has theoretical guarantees and is analyzed to provide insights into the underlying structure of low-rank matrices and the nature of the linear measurements used in the recovery process. It advances the theoretical understanding of matrix sensing and provides a new approach for solving this important problem.

1 INTRODUCTION

The matrix sensing problem is a fundamental problem in signal processing and machine learning that involves recovering a low-rank matrix from a set of linear measurement. This problem arises in various applications such as image and video processing Fowler et al. (2012); Bouwmans et al. (2018) and sensor networks Middya et al. (2017); Wimalajeewa & Varshney (2017). Mathematically, matrix sensing can be formulated as a matrix view of compressive sensing problem Jain et al. (2013). The rank-1 matrix sensing problem was formally raised in Zhong et al. (2015).

The matrix sensing problem has attracted significant attention in recent years, and several algorithms have been proposed to solve it efficiently. In this paper, we provide a novel improvement over the original algorithm in Zhong et al. (2015), with improvement both on running time and sample complexity.

Matrix sensing is a fundamental problem in signal processing and machine learning that involves recovering a low-rank matrix from a set of linear measurements. Specifically, given a matrix $W_* \in \mathbb{R}^{d\times d}$ of rank k that is not directly accessible, we aim to recover W_* from a set of linear

* lizhihangdll@gmail.com. Huazhong Agricultural University.
† shazz20@mails.tsinghua.edu.cn. Tsinghua University.
‡ magic.linuxkde@gmail.com. The Simons Institute for the Theory of Computing at UC Berkeley.
§ dylan.r.mathison@gmail.com. Anhui University.

measurements $b \in \mathbb{R}^m$ applied to the ground truth matrix W^* where

$$b_i = \mathrm{tr}[A_i^\top W_*], \quad \forall i = 1, \dots, m,$$

where A_i are known linear operators. The measurements b_i are obtained by sensing the matrix W_* using a set of linear measurements, and the goal is to reconstruct the original matrix W_* as accurately as possible. This problem arises in various applications such as image and video processing, sensor networks, and recommendation systems.

The matrix sensing problem is ill-posed since there may exist multiple low-rank matrices that satisfy the given linear measurements. However, the problem becomes well-posed under some assumptions on the underlying matrix, such as incoherence and restricted isometry property (RIP) Candes & Tao (2005); Candes et al. (2006); Gurevich & Hadani (2008) , which ensure unique and stable recovery of the matrix. A well-used method to solve this problem is to use convex optimization techniques that minimize a certain loss function subject to the linear constraints. Specifically, one can solve the following convex optimization problem:

$$\begin{aligned} & \min_{W_*} \ \mathrm{rank}(W_*) \\ & \text{s.t. } \mathrm{tr}[A_i^\top W_*] = b_i, \forall i = 1, \dots, m. \end{aligned}$$

However, this problem is NP-hard Tillmann & Pfetsch (2013) and intractable in general, and hence, various relaxation methods have been proposed, such as nuclear norm minimization and its variants, which provide computationally efficient solutions with theoretical guarantees. In this work, we focus on the *rank-one independent* measurements. Under this setting, the linear operators A_i can be decomposed into the form of $A_i = x_i y_i^\top$, where $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}^d$ are all sampled from zero-mean multivariate Gaussian distribution $\mathcal{N}(0, I_d)$.

Our work on improving the matrix sensing algorithm is based on a novel analysis and sketching technique that enables faster convergence rates and better accuracy in recovering low-rank matrices. We focus on developing a theoretical understanding of the proposed algorithm and establishing its advantages over previous methods. Our analysis provides insights into the underlying structure of the low-rank matrices and the nature of the linear measurements used in the recovery process. To summarize, we improve both the running time of the original algorithm Zhong et al. (2015) from $O(md^2k^2)$ to $O(md^2k)$, and the sample complexity from $\widetilde{O}(\epsilon^{-2}dk^2)$ to $\widetilde{O}(\epsilon^{-2}(d+k^2))$. Formally, we get the following result,

Theorem 1.1 (Informal, combination of Theorem 3.7, Theorem D.7 and Theorem F.8)**.** *Let $\epsilon \in (0,1)$ be some specific parameter. For ground truth matrix $W_* \in \mathbb{R}^{d \times d}$, there is a matrix sensing algorithm (Algorithm 1) that with $\widetilde{O}(\epsilon^{-2}(d+k^2))$ sample complexity, and takes $\widetilde{O}(md^2k)$ time for each iteration, finally outputs a matrix $W \in \mathbb{R}^{d \times d}$ such that*

$$(1-\epsilon)W_* \preceq W \preceq (1+\epsilon)W_*$$

with high probability.

Roadmap. We organize the following paper as follows. In Section 2 we provide some tools and existing results for our work. In Section 3 we state the main result of this paper. In Section 4 we provide the technique overview for our paper. We provide discussion in Section G. In Section 5, we state the conclusion of our results.

2 PRELIMINARY

In this section, we provide preliminaries to be used in our paper. In Section 2.1 we introduce notations we use. In Section 2.2 we introcude the randomness facts.

We state some matrix concentration in Section 2.3. In Section 2.4 we introduce the important definition of restricted isometry property. In Section 2.5 we provide results for rank-one estimation. In Section 2.6 we introduce the rank-one independent Gaussian operator.

2.1 NOTATIONS

Let $x \in \mathbb{R}^n$ and $w \in \mathbb{R}^n_{\geq 0}$, we define the norm $\|x\|_w := (\sum_{i=1}^n w_i x_i^2)^{1/2}$. For $n > k$, for any matrix $A \in \mathbb{R}^{n \times k}$, we denote the spectral norm of A by $\|A\|$. Let $A \in \mathbb{R}^{n \times k}$, we denote the Frobenius

norm of A by $\|A\|_F$. For any square matrix $A \in \mathbb{R}^{n \times n}$, we denote its trace by $\mathrm{tr}[A]$. For any $A \in \mathbb{R}^{n \times d}$ and $B \in \mathbb{R}^{n \times d}$, we denote $\langle A, B \rangle = \mathrm{tr}[A^\top B]$. Let $A \in \mathbb{R}^{n \times d}$ and $x \in \mathbb{R}^d$ be any matrix and vector, we have that $\|Ax\|_2^2 = \langle Ax, Ax \rangle = \langle x, A^\top Ax \rangle = x^\top A^\top Ax$. We denote the Moore-Penrose pseudoinverse matrix of A as $A^\dagger \in \mathbb{R}^{k \times n}$, i.e., $A^\dagger := V \Sigma^{-1} U^\top$. For $A \in \mathbb{R}^{n_1 \times d_1}$, $B \in \mathbb{R}^{n_2 \times d_2}$, we define kronecker product $\otimes$ as $(A \otimes B)_{i_1 + (i_2 - 1) n_1, j_1 + (j_2 - 1) n_2} = A_{i_1, j_1} B_{i_2, j_2}$ for all $i_1 \in [n_1]$, $j_1 \in [d_1]$, $i_2 \in [n_2]$ and $j_2 \in [d_2]$. For any full-rank matrix $A \in \mathbb{R}^{n \times m}$, we define $A = QR$ its QR-decomposition, where $Q \in \mathbb{R}^{m \times n}$ is an orthogonal matrix and $R \in \mathbb{R}^{n \times n}$ is an non-singular lower triangular matrix. We use $R = \mathrm{QR}(A) \in \mathbb{R}^{n \times n}$ to denote the lower triangular matrix obtained by the QR-decomposition of $A \in \mathbb{R}^{m \times n}$. Let $A \in \mathbb{R}^{k \times k}$ be a symmetric matrix. The eigenvalue decomposition of A is $A = U \Lambda U^\top$, where Λ is a diagonal matrix. If a matrix A is positive semidefinite (PSD), we denote it as $A \succeq 0$ Similarly, we say $A \succeq B$ if $x^\top Ax \geq x^\top Bx$ for all vectors x. For any matrix $U \in \mathbb{R}^{n \times k}$, we say U is an orthonormal basis if $\|U_i\| = 1$ for all $i \in [k]$ and for any $i \neq j$, we have $\langle U_i, U_j \rangle = 0$. Here for each $i \in [k]$, we use U_i to denote the i-th column of matrix U. For any $U \in \mathbb{R}^{n \times k}$ (suppose $n > k$) which is an orthonormal basis, we define $U_\perp \in \mathbb{R}^{n \times (n-k)}$ to be another orthonormial basis that, $UU^\top + U_\perp U_\perp^\top = I_n$ and $U^\top U_\perp = \mathbf{0}^{k \times (n-k)}$, where we use $\mathbf{0}^{k \times (n-k)}$ to denote a $k \times (n-k)$ all-zero matrix. We say a vector x lies in the span of U, if there exists a vector y such that $x = Uy$. We say a vector z lies in the complement of span of U, if there exists a vector w such that $z = U_\perp w$. Then it is obvious that $\langle x, z \rangle = x^\top z = z^\top x = 0$. For a matrix A, we define $\sigma_{\min}(A) := \min_x \|Ax\|_2 / \|x\|_2$. Equivalently, $\sigma_{\min}(A) := \min_{x : \|x\|_2 = 1} \|Ax\|_2$. Similarly, we define $\sigma_{\max}(A) := \max_x \|Ax\|_2 / \|x\|_2$. Equivalently, $\sigma_{\max}(A) := \max_{x : \|x\|_2 = 1} \|Ax\|_2$ Let $A_1, \cdots, A_n$ denote a list of square matrices. Let S denote a block diagonal matrix $S = \mathrm{diag}(A_1, A_2, \cdots, A_n)$. Then $\|S\| = \max_{i \in [n]} \|A_i\|$. We use $\Pr[]$ to denote probability. We use $\mathbb{E}[]$ to denote expectation. Let a and b denote two random variables. Let $f(a)$ denote some event that depends on a (for example $f(a)$ can be $a = 0$ or $a \geq 10$.). Let $g(b)$ denote some event that depends on b. We say a and b are independent if $\Pr[f(a) \text{ and } g(b)] = \Pr[f(a)] \cdot \Pr[g(b)]$. We say a and b are not independent if $\Pr[f(a) \text{ and } g(b)] \neq \Pr[f(a)] \cdot \Pr[g(b)]$. Usually if a and b are independent, then we also have $\mathbb{E}[ab] = \mathbb{E}[a] \cdot \mathbb{E}[b]$. We say a random variable x is symmetric if $\Pr[x = u] = \Pr[x = -u]$. For any random variable $x \sim \mathcal{N}(\mu, \sigma^2)$. This means $\mathbb{E}[x] = \mu$ and $\mathbb{E}[x^2] = \sigma^2$. We use $\widetilde{O}(f)$ to denote $f \cdot \mathrm{poly}(\log f)$. We use $\mathcal{T}_{\mathrm{mat}}(a, b, c)$ to denote the time of multiplying an $a \times b$ matrix with another $b \times c$ matrix. We use ω to denote the exponent of matrix multiplication, i.e., $n^\omega = \mathcal{T}_{\mathrm{mat}}(n, n, n)$.

2.2 Randomness Facts

Here we introduce some facts about randomness.

Fact 2.1. *We have*

- *Part 1. Expectation has linearity, i.e.,* $\mathbb{E}[\sum_{i=1}^n x_i] = \sum_{i=1}^n \mathbb{E}[x_i]$.
- *Part 2. For any random vectors x and y, if x and y are independent, then for any fixed function f, we have* $\mathbb{E}_{x,y}[f(x) f(y)] = \mathbb{E}_x[f(x)] \cdot \mathbb{E}_y[f(y)]$.
- *Part 3. Let $A \in \mathbb{R}^{d \times d}$ denote a fixed matrix. For any fixed function $f : \mathbb{R}^d \to \mathbb{R}^{d \times d}$, we have* $\mathbb{E}_x[f(x) \cdot A] = \mathbb{E}_x[f(x)] \cdot A$.
- *Part 4. Given n events $A_1, A_2, \cdots A_n$. For each $i \in [n]$, if* $\Pr[A_i] \geq 1 - \delta_i$. *Then taking a union bound over all the n events, we have*

$$\Pr[A_1 \textit{ and } A_2 \cdots A_n] \geq 1 - \sum_{i=1}^n \delta_i.$$

2.3 Matrix Concentration

We now discuss matrix concentration inequalities, which form the basis for analyzing the behavior of sums of random matrices, particularly through tools such as the Matrix Bernstein Inequality.

Theorem 2.2 (Matrix Bernstein Inequality, Theorem 1.6 of Tropp (2012))**.** *Given a finite sequence $X_1, \cdots, X_m \subset \mathbb{R}^{n_1 \times n_2}$ of independent random matrices, all with dimension $n_1 \times n_2$, let $Z =$*

$\sum_{i=1}^m X_i$. *Assume that*

$$\mathbb{E}[X_i] = 0, \forall i \in [m], \|X_i\| \leq M, \forall i \in [m].$$

Let $\mathrm{Var}[Z]$ *be the matrix variance statistic of the sum:*

$$\mathrm{Var}[Z] = \max\{\|\sum_{i=1}^m \mathbb{E}[X_i X_i^\top]\|, \|\sum_{i=1}^m \mathbb{E}[X_i^\top X_i]\|\}.$$

Then the following results hold:

$$\mathbb{E}[\|Z\|] \leq (2\,\mathrm{Var}[Z] \cdot \log(n_1+n_2))^{1/2} + M\log(n_1+n_2)/3.$$

Further, for all t ¿ 0 ,

$$\Pr[\|Z\| \geq t] \leq (n_1+n_2) \cdot \exp\left(-\frac{t^2/2}{\mathrm{Var}[Z] + Mt/3}\right).$$

2.4 Restricted Isometry Property

This subsection introduces the Restricted Isometry Property (RIP), which characterizes the near-isometric behavior of linear operators on low-rank matrices and serves as a foundational tool in compressive sensing and matrix recovery.

Definition 2.3 (Restricted isometry property (RIP), see Definition 1 in Zhong et al. (2015))**.** *A linear operator* $\mathcal{A} : \mathbb{R}^{d\times d} \to \mathbb{R}^m$ *satisfies the* RIP *if, for all* $W \in \mathbb{R}^{d\times d}$ *such that* $\mathrm{rank}(W) \leq k$*, the following holds:*

$$(1-\epsilon_k) \cdot \|W\|_F^2 \leq \|\mathcal{A}(W)\|_F^2 \leq (1+\epsilon_k) \cdot \|W\|_F^2$$

where $\epsilon_k > 0$ *is a constant that depends only on* k*.*

2.5 Rank-one Estimation

The goal of matrix sensing is to design a linear operator $\mathcal{A} : \mathbb{R}^{d\times d} \to \mathbb{R}^m$ and a recovery algorithm so that a low-rank matrix $W_* \in \mathbb{R}^{d\times d}$ can be recovered exactly using $\mathcal{A}(W_*)$. Then we define:

Definition 2.4 (Low-rank matrix estimation using rank one measurements)**.** *Given a ground-truth matrix* $W_* \in \mathbb{R}^{d\times d}$*. Let* $(x_1, y_1), \cdots, (x_m, y_m) \in \mathbb{R}^d \times \mathbb{R}^d$ *denote* m *pair of feature vectors. Let* $b \in \mathbb{R}^m$ *be defined*

$$b_i = x_i^\top W_* y_i, \quad \forall i \in [m].$$

The goal is to use $b \in \mathbb{R}^m$ *and* $\{(x_i, y_i)\}_{i\in[m]} \subset \mathbb{R}^d \times \mathbb{R}^d$ *to recover* $W_* \in \mathbb{R}^{d\times d}$*.*

2.6 Rank-one Independent Gaussian Operator

We formally define Gaussian independent operator, here.

Definition 2.5 (Gaussian Independent (GI) Operator)**.** *Let* $(x_1, y_1), \cdots, (x_m, y_m) \subset \mathbb{R}^d \times \mathbb{R}^d$ *denote i.i.d. samples from Gaussian distribution. For each* $i \in [m]$*,* $A_i := x_i y_i^\top$*. We define* $\mathcal{A}_{\mathrm{GI}} \in \mathbb{R}^{d\times md}$ *as (*GI *denotes Gaussian Independent):*

$$\mathcal{A}_{\mathrm{GI}} := [A_1 \quad A_2 \quad \cdots \quad A_m].$$

2.7 Matrix Angle and Distance

We list several basic definitions and tools in literature.

Definition 2.6 (Definition 4.1 in Gu et al. (2023))**.** *Let* $X, Y \in \mathbb{R}^{n\times k}$ *denote two matrices.*

For any matrix X*, and for orthonormal matrix* Y *(*$Y^\top Y = I_k$*) we define*

- $\tan\theta(Y, X) := \|Y_\perp^\top X (Y^\top X)^{-1}\|$

For orthonormal matrices Y *and* X *(*$Y^\top Y = I_k$ *and* $X^\top X = I_k$*), we define*

- $\cos\theta(Y,X) := \sigma_{\min}(Y^\top X)$.
  - *It is obvious that* $\cos(Y,X) = 1/\|(Y^\top X)^{-1}\|$ *and* $\cos(Y,X) \leq 1$.
- $\sin\theta(Y,X) := \|(I - YY^\top)X\|$.
  - *It is obvious that* $\sin\theta(Y,X) = \|Y_\perp Y_\perp^\top X\| = \|Y_\perp^\top X\|$ *and* $\sin\theta(Y,X) \leq 1$.
  - *From Lemma 2.8, we know that* $\sin^2\theta(Y,X) + \cos^2\theta(Y,X) = 1$.
- $\mathrm{dist}(Y,X) := \sin\theta(Y,X)$

Lemma 2.7 (Lemma A.7 in Gu et al. (2023))**.** *Let $X, Y \in \mathbb{R}^{n\times k}$ be orthogonal matrices, then*

$$\tan\theta(Y,X) = \frac{\sin\theta(Y,X)}{\cos\theta(Y,X)}.$$

Lemma 2.8 (Lemma A.8 in Gu et al. (2023))**.** *Let $X, Y \in \mathbb{R}^{n\times k}$ be orthogonal matrices, then*

$$\sin^2\theta(Y,X) + \cos^2\theta(Y,X) = 1.$$

3 Main Result

This section presents the main theoretical results of this work, including the foundational definitions, operator properties, and the convergence guarantees. Section 3.1 introduces critical definitions such as the target matrix, its condition number, and the measurement framework. Section 3.2 formalizes the concentration properties of operators and the initialization criteria essential for our analysis. Finally, Section 3.3 states and proves the main convergence theorem, which guarantees the success of the alternating minimization method under specific conditions on the initialization and operator properties.

3.1 Key concepts

This subsection establishes the key concepts that will be used in subsequent theoretical developments, including the characterization of matrix components, condition numbers, and measurement definitions.

Definition 3.1. *We define $W_* \in \mathbb{R}^{d\times d}$ as $W_* = U_*\Sigma_*V_*^\top$, where $U_* \in \mathbb{R}^{n\times k}$ are orthonormal columns, and $V_* \in \mathbb{R}^{n\times k}$ are orthonormal columns. Let $\sigma_1^*, \sigma_2^*, \cdots \sigma_k^*$ denote the diagonal entries of diagonal matrix $\Sigma_* \in \mathbb{R}^{d\times d}$.*

Definition 3.2 (Condition number)**.** *Let W_* be defined as Definition 3.1. We define κ to the condition number of W_*, i.e., $\kappa := \sigma_1/\sigma_k$. It is obvious that $\kappa \geq 1$.*

Definition 3.3 (Measurements)**.** *For each $i \in [m]$, let x_i, y_i denote samples from $\mathcal{N}(0, I_d)$.*

For each $i \in [m]$, we define $A_i = x_i y_i^\top$ and $b_i = x_i^\top W_* y_i$.

3.2 Properties of Operators

This subsection defines the initialization and concentration properties of operators, providing tools for analyzing their behavior in high-dimensional settings.

Definition 3.4 (Initialization)**.** *For each $i \in [m]$, let A_i and b_i be defined as Definition 3.3.*

We define

$$W_0 := \frac{1}{m}\sum_{i=1}^m b_i A_i.$$

We say initialization matrix $W_0 \in \mathbb{R}^{d\times d}$ is an ϵ-good operator if

$$\|W_0 - W_*\| \leq \|W_*\| \cdot \epsilon.$$

Definition 3.5 (Concentration of operators B_x, B_y)**.** *For any vectors u, v, we define*

- $B_x := \frac{1}{m} \sum_{l=1}^m (y_l^\top v)^2 x_l x_l^\top$
- $B_y := \frac{1}{m} \sum_{l=1}^m (x_l^\top u)^2 y_l y_l^\top$

We say $B = (B_x, B_y)$ is ϵ-operator if $\|B_x - I\| \leq \epsilon$ and $\|B_y - I\| \leq \epsilon$.

Definition 3.6 (Concentration of operators G_x, G_y)**.** *For any vectors $u, v \in \mathbb{R}^d$. We define*

- $G_x := \frac{1}{m} \sum_{l=1}^m (y_l^\top v)(y_l^\top v_\perp) x_l x_l^\top$
- $G_y := \frac{1}{m} \sum_{l=1}^m (x_l^\top u)(x_l^\top u_\perp) y_l y_l^\top$

$u, u_\perp \in \mathbb{R}^d, v, v_\perp \in \mathbb{R}^d$ are unit vectors, s.t., $u^\top u_\perp = 0$ and $v^\top v_\perp = 0$. We say $G = (G_x, G_y)$ is ϵ-operator if $\|G_x\| \leq \epsilon$ and $\|G_y\| \leq \epsilon$.

3.3 Main Result

Now, we prove our main convergence result as follows:

Theorem 3.7 (Formal version of Theorem 1.1)**.** *Let $W_* \in \mathbb{R}^{d \times d}$ be defined as Definition 3.1, and W_* has rank-k and condition number κ. Also, let $\mathcal{A} : \mathbb{R}^{d \times d} \to \mathbb{R}^m$ be a linear measurement operator parameterized by m matrices, i.e., $\mathcal{A} = \{A_1, A_2, \cdots, A_m\}$ where $A_l = x_l y_l^\top$. Let $\mathcal{A}(W)$ be as given by*

$$b = \mathcal{A}(W) = \begin{bmatrix} \mathrm{tr}[A_1^\top W] & \mathrm{tr}[A_2^\top W] & \cdots & \mathrm{tr}[A_m^\top W] \end{bmatrix}^\top.$$

Let $\epsilon = 0.001/(k^{1.5}\kappa)$ Let $T = 100 \log(\kappa k/\epsilon_0)$ Let $\{(b_i, A_i)\}_{i \in [m]}$ be an ϵ-good operator (Definition 3.4). Let B be an ϵ-operator (Definition 3.5). Let G be an ϵ-operator (Definition 3.6). Then, after T-iterations of the alternating minimization method (Algorithm 1), we obtain $W_T = U_T V_T^\top$ s.t.,

$$\|W_T - W_*\| \leq \epsilon_0.$$

Proof. Note that using the initialization property (first property mentioned in Theorem 3.7), we get, $\|W_0 - W_*\| \leq \epsilon \sigma_1^* \leq \frac{\sigma_k^*}{100}$. Now, using the standard sin theta theorem for singular vector perturbation Li (1994), we get $\mathrm{dist}(U_0, U_*) \leq \frac{1}{100}$ and $\mathrm{dist}(V_0, V_*) \leq \frac{1}{100}$.

After T iteration (via Lemma C.1), we obtain

$$\mathrm{dist}(U_T, U_*) \leq (1/4)^T \text{ and } \mathrm{dist}(V_T, V_*) \leq (1/4)^T$$

which implies that $\|W_T - W_*\| \leq \epsilon_0$.

□

4 Technique Overview

This section introduces the techniques and methodologies underlying the key contributions of this work. Section 4.1 outlines the tighter analysis and its reduction to sample complexity, leveraging random matrix theory and concentration inequalities. Section 4.2 demonstrates the correctness of the iterative algorithm through induction, detailing the convergence properties of the alternating minimization method. Section 4.3 presents the sketching-based acceleration strategy, which significantly reduces the computational cost of each iteration by reformulating the optimization problem and applying a fast low-rank matrix completion approach. These components collectively establish the foundation for our theoretical and practical advancements.

4.1 Tighter Analysis Implies Reduction to Sample Complexity

Our approach achieves this improvement by using a new sketching technique that compresses the original matrix into a smaller one while preserving its low-rank structure. This compressed version

can then be used to efficiently extract relevant information from linear measurements of the original matrix.

To analyze the performance of our approach, we use tools from random matrix theory and concentration inequalities. Specifically, we use the Bernstein's inequality for matrices to establish bounds on the error of our recovery algorithm. We first define our measurements and operators, for each $i \in [m]$, let x_i, y_i denotes samples from $\mathcal{N}(0, I_d)$. We define

- $A_i := x_i y_i^\top$;
- $b_i := x_i^\top W_* y_i$;
- $W_0 := \frac{1}{m} \sum_{i=1}^m b_i A_i$;
- $B_x := \frac{1}{m} \sum_{i=1}^m (y_i^\top v)^2 x_i x_i^\top$;
- $B_y := \frac{1}{m} \sum_{i=1}^m (x_i^\top v)^2 y_i y_i^\top$;
- $G_x := \frac{1}{m} \sum_{i=1}^m (y_i^\top v)(y_i^\top v_\perp) x_i x_i^\top$;
- $G_y := \frac{1}{m} \sum_{i=1}^m (x_i^\top v)(x_i^\top v_\perp) y_i y_i^\top$.

We need to argue that our measurements are *good* under our choices of m, here the word "good" means that

- $\|W_0 - W_*\| \leq \epsilon \cdot \|W_*\|$;
- $\|B_x - I\| \leq \epsilon$ and $\|B_y - I\| \leq \epsilon$;
- $\|G_x\| \leq \epsilon$ and $\|G_y\| \leq \epsilon$.

In our analysis we need to first bound $\|Z_i\|$ and $\|\mathbb{E}[Z_i Z_i^\top]\|$, where $Z_i := x_i x_i^\top U_* \Sigma_* V_*^\top y_i y_i^\top$. With an analysis, we are able to show that (Lemma D.5 and Lemma D.6)

$$\begin{aligned} \Pr[\|Z_i\| \leq C^2 k^2 \log^2(d/\delta) \sigma^4 \cdot \sigma_1^*] &\geq 1 - \delta / \operatorname{poly}(d) \\ \|\mathbb{E}[Z_i Z_i^\top]\| &\leq C^2 k^2 \sigma^4 (\sigma_1^*)^2. \end{aligned}$$

Now, applying these two results and by Bernstein's inequality, we are able to show that our operators are all "good" (Theorem D.7).

4.2 Induction Implies Correctness

To get the final error bounded, we show that the iterates are getting closer and closer to the ground truth. Here we let U_* and V_* be the decomposition of ground truth W_*, i.e., $W_* = U_* \Sigma V_*^\top$. We show that, when iteratively applying our alternating minimization method, if U_t and V_t are closed to U_* and V_* respectively, then the output of next iteration $t+1$ is close to U_* and V_*. Specifically, we show that, if $\operatorname{dist}(U_t, U_*) \leq \frac{1}{4} \cdot \operatorname{dist}(V_t, V_*)$, then it yields

$$\operatorname{dist}(V_{t+1}, V_*) \leq \frac{1}{4} \cdot \operatorname{dist}(U_t, U_*). \tag{1}$$

Similarly, from the other side, if $\operatorname{dist}(V_{t+1}, V_*) \leq \frac{1}{4} \cdot \operatorname{dist}(U_t, U_*)$, we have

$$\operatorname{dist}(U_{t+1}, U_*) \leq \frac{1}{4} \cdot \operatorname{dist}(V_{t+1}, V_*). \tag{2}$$

This two recurrence relations together give the guarantee that, if the starting error $U_0 - U_*$ and $V_0 - V_*$ is bounded, the distance from V_t and U_t to V_* and U_* will be bounded to, respectively.

To prove the result, we first define the value of ϵ_d as $1/10$. Then, by the algorithm, we have the following relationship between V_{t+1} and $\widehat{V}_{t+1} R^{-1}$,

$$V_{t+1} = \widehat{V}_{t+1} R^{-1} = (W_*^\top U_t - F) R^{-1},$$

where the second step follows from the definition of $\widehat{V}$ and defining F as Definition E.1. Now we show that, $\|F\|$ and $\|R^{-1}\|$ can be bounded respectively,

$$\|F\| \leq 2\epsilon k^{1.5} \cdot \sigma_1^* \cdot \operatorname{dist}(U_t, U_*) \qquad \text{Lemma E.4}$$

$$\|R^{-1}\| \leq 10/\sigma_{k^*} \qquad \text{Lemma E.5}$$

Note that the bound of R^{-1} need $\mathrm{dist}(U_t, U_*) \leq \frac{1}{4} \cdot \mathrm{dist}(V_t, V_*)$.

With these bounds, we are able to show the bound for $\mathrm{dist}(V_{t+1}, V_*)$. We first notice that, $\mathrm{dist}(V_{t+1}, V_*)$ can be represented as $(V_{*,\perp})^\top V_{t+1}$, where $V_{*,\perp} \in \mathbb{R}^{d\times(d-k)}$ is a fixed orthonormal basis of the subspace orthogonal to $\mathrm{span}(V_*)$. Then we show that (Claim E.3)

$$(V_{*,\perp})^\top V_{t+1} = -(V_{*,\perp})^\top F R^{-1}.$$

Now, by turning $\mathrm{dist}(V_{t+1}, V_*)$ to the term of F and R, and using the bound for $\|F\|$ and $\|R^{-1}\|$, we are finally able to reach the bound

$$\begin{aligned}
\mathrm{dist}(V_{t+1}, V_*) &= \|FR^{-1}\| \\
&\leq \|F\| \cdot \|R^{-1}\| \\
&\leq 2\epsilon k^{1.5} \cdot \sigma_1^* \cdot \mathrm{dist}(U_t, U_*) \cdot \|R^{-1}\| \\
&\leq 2\epsilon k^{1.5} \cdot \sigma_1^* \cdot \mathrm{dist}(U_t, U_*) \cdot 10/\sigma_k^* \\
&\leq 0.01 \cdot \mathrm{dist}(U_t, U_*).
\end{aligned}$$

By a similar analysis, we can show Eq. (2).

Now applying the above results and with a detailed analysis, we have the claim proved. Finally, when we prove that the initialization of the parameters are good, we can show that, the final output W_T satisfies

$$\|W_T - W_*\| \leq \epsilon_0.$$

4.3 Speeding up with Sketching Technique

Now we consider the running time at each iteration. At each iteration of our algorithm, we need to solve the following optimization problem:

$$\arg\min_{V\in\mathbb{R}^{d\times k}} \sum_{i=1}^{m} (\mathrm{tr}[A_i^\top U V^\top] - b)^2. \tag{3}$$

When this problem is straightforwardly solved, it costs $O(md^2k^2)$ time, which is very expensive. So from another new direction, we give an analysis such that, this problem can be converted to a minimization problem where the target variable is a vector. To be specific, we show that, above optimization question (3) is equivalent to the following (Lemma F.3),

$$\arg\min_{v\in\mathbb{R}^{dk}} \|Mv - b\|_2^2,$$

where the matrix $M \in \mathbb{R}^{m\times dk}$ is defined to be the reformed matrix of $U^\top A_i$'s, i.e.,

$$M_{i,*} := \mathrm{vec}(U^\top A_i), \ \ \forall i \in [m].$$

When working on this form of optimization problem, inspired by a recent work Gu et al. (2023), we apply the fast sketch-to-solve low-rank matrix completion method. With this technique, we are able to reduce the running time to $\widetilde{O}(md^2k)$ (Theorem F.8), which is much more acceptable.

5 Conclusion

In conclusion, matrix sensing is a fundamental problem in signal processing and machine learning that aims to recover a low-rank matrix from a set of linear measurements. It has various applications in fields such as image and video processing, sensor networks, and recommendation systems. While the matrix sensing problem is ill-posed, under certain assumptions on the underlying matrix, such as incoherence and restricted isometry property, it can be solved using convex optimization techniques that minimize a certain loss function subject to the linear constraints. In this paper, we have proposed a novel improvement over the original algorithm for the rank-1 matrix sensing problem, with improvements in both running time and sample complexity to the original method in Zhong et al. (2015). Our work is based on a novel analysis and sketching technique that enables faster convergence rates and better accuracy in recovering low-rank matrices. Our proposed algorithm is computationally efficient and scalable, and our analysis provides insights into the underlying structure of the low-rank matrices and the nature of the linear measurements used in the recovery process.

REFERENCES

Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. A universal sampling method for reconstructing signals with simple fourier transforms. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1051–1063, 2019.

Mayank Bakshi, Sidharth Jaggi, Sheng Cai, and Minghua Chen. Sho-fa: Robust compressive sensing with order-optimal complexity, measurements, and bits. *IEEE Transactions on Information Theory*, 62(12):7419–7444, 2015.

Thomas Blumensath and Mike E Davies. Iterative thresholding for sparse approximations. *Journal of Fourier analysis and Applications*, 14:629–654, 2008.

Christos Boutsidis and David P Woodruff. Optimal cur matrix decompositions. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 353–362, 2014.

Thierry Bouwmans, Sajid Javed, Hongyang Zhang, Zhouchen Lin, and Ricardo Otazo. On the applications of robust pca in image and video processing. *Proceedings of the IEEE*, 106(8):1427–1457, 2018.

Jan van den Brand, Binghui Peng, Zhao Song, and Omri Weinstein. Training (overparametrized) neural networks in near-linear time. In *ITCS*, 2021.

Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

Xue Chen, Daniel M Kane, Eric Price, and Zhao Song. Fourier-sparse interpolation without a frequency gap. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 741–750. IEEE, 2016.

Kenneth L Clarkson and David P Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, 63(6):1–45, 2017.

Michael B Cohen, Yin Tat Lee, and Zhao Song. Solving linear programs in the current matrix multiplication time. In *STOC*, 2019.

Yichuan Deng, Zhao Song, and Omri Weinstein. Discrepancy minimization in input-sparsity time. *arXiv preprint arXiv:2210.12468*, 2022.

Huaian Diao, Zhao Song, Wen Sun, and David Woodruff. Sketching for kronecker product regression and p-splines. In *International Conference on Artificial Intelligence and Statistics*, pp. 1299–1308. PMLR, 2018.

Huaian Diao, Rajesh Jayaram, Zhao Song, Wen Sun, and David Woodruff. Optimal sketching for kronecker product regression and low rank approximation. *Advances in neural information processing systems*, 32, 2019.

James E Fowler, Sungkwang Mun, Eric W Tramel, et al. Block-based compressed sensing of images and video. *Foundations and Trends® in Signal Processing*, 4(4):297–416, 2012.

Yuzhou Gu and Zhao Song. A faster small treewidth sdp solver. *arXiv preprint arXiv:2211.06033*, 2022.

Yuzhou Gu, Zhao Song, Junze Yin, and Lichen Zhang. Low rank matrix completion via robust alternating minimization in nearly linear time. In *arXiv preprint*. https://arxiv.org/abs/2302.11068, 2023.

Shamgar Gurevich and Ronny Hadani. Incoherent dictionaries and the statistical restricted isometry property. *arXiv preprint arXiv:0809.1687*, 2008.

Haitham Hassanieh, Piotr Indyk, Dina Katabi, and Eric Price. Nearly optimal sparse fourier transform. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pp. 563–578, 2012a.

Haitham Hassanieh, Piotr Indyk, Dina Katabi, and Eric Price. Simple and practical algorithm for sparse fourier transform. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pp. 1183–1194. SIAM, 2012b.

Piotr Indyk and Michael Kapralov. Sample-optimal fourier sampling in any constant dimension. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 514–523. IEEE, 2014.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674, 2013.

Shunhua Jiang, Zhao Song, Omri Weinstein, and Hengjie Zhang. Faster dynamic matrix inverse for faster lps. In *STOC*. arXiv preprint arXiv:2004.07470, 2021.

Shunhua Jiang, Yunze Man, Zhao Song, Zheng Yu, and Danyang Zhuo. Fast graph neural tangent kernel via kronecker sketching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7033–7041, 2022.

Yaonan Jin, Daogao Liu, and Zhao Song. Super-resolution and robust sparse continuous fourier transform in any constant dimension: Nearly linear time and sample complexity. In *SODA*, pp. 4667–4767. SIAM, 2023.

Michael Kapralov. Sparse fourier transform in any constant dimension with nearly-optimal sample complexity in sublinear time. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 264–277, 2016.

Michael Kapralov. Sample efficient estimation and recovery in sparse fft via isolation on average. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 651–662. Ieee, 2017.

Yehuda Koren. The bellkor solution to the netflix grand prize. *Netflix prize documentation*, 81 (2009):1–10, 2009.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pp. 1302–1338, 2000.

Yin Tat Lee, Zhao Song, and Qiuyi Zhang. Solving empirical risk minimization in the current matrix multiplication time. In *Conference on Learning Theory*, pp. 2140–2157. PMLR, 2019.

Ren Cang Li. On perturbations of matrix pencils with real spectra. *Mathematics of Computation*, 62(205):231–265, 1994.

Xiao Li, Dong Yin, Sameer Pawar, Ramtin Pedarsani, and Kannan Ramchandran. Sub-linear time support recovery for compressed sensing using sparse-graph codes. *IEEE Transactions on Information Theory*, 65(10):6580–6619, 2019.

Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):171–184, 2012.

Xiaoqi Liu and Ramji Venkataramanan. Sketching sparse low-rank matrices with near-optimal sample-and time-complexity using message passing. *IEEE Transactions on Information Theory*, 69(9):6071–6097, 2023.

Rajarshi Middya, Nabajit Chakravarty, and Mrinal Kanti Naskar. Compressive sensing in wireless sensor networks–a survey. *IETE technical review*, 34(6):642–654, 2017.

Vasileios Nakos and Zhao Song. Stronger l2/l2 compressed sensing; without iterating. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 289–297, 2019.

Vasileios Nakos, Zhao Song, and Zhengyu Wang. (nearly) sample-optimal sparse fourier transform in any dimension; ripless and filterless. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 1568–1577. IEEE, 2019.

Jelani Nelson and Huy L Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 ieee 54th annual symposium on foundations of computer science*, pp. 117–126. IEEE, 2013.

Luong Trung Nguyen, Junhan Kim, and Byonghyo Shim. Low-rank matrix completion: A contemporary survey. *IEEE Access*, 7:94215–94237, 2019.

Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. In *Artificial Intelligence and Statistics*, pp. 65–74. PMLR, 2017.

Sameer Pawar and Kannan Ramchandran. Computing a k-sparse n-length discrete fourier transform using at most 4k samples and o (k log k) complexity. In *2013 IEEE International Symposium on Information Theory*, pp. 464–468. IEEE, 2013.

Eric Price and Zhao Song. A robust sparse fourier transform in the continuous setting. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pp. 583–600. IEEE, 2015.

Lianke Qin, Zhao Song, Lichen Zhang, and Danyang Zhuo. An online and unified algorithm for projection matrix vector multiplication with application to empirical risk minimization. In *AISTATS*, 2023.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Aravind Reddy, Zhao Song, and Lichen Zhang. Dynamic tensor product regression. *arXiv preprint arXiv:2210.03961*, 2022.

Anshumali Shrivastava, Zhao Song, and Zhaozhuo Xu. Sublinear least-squares value iteration via locality sensitive hashing. In *AISTATS*, 2023.

Zhao Song and Zheng Yu. Oblivious sketching-based central path method for linear programming. In *International Conference on Machine Learning*, pp. 9835–9847. PMLR, 2021.

Zhao Song, David P Woodruff, and Peilin Zhong. Low rank approximation with entrywise l1-norm error. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 688–701, 2017.

Zhao Song, David P Woodruff, and Peilin Zhong. Relative error tensor low rank approximation. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2772–2789. SIAM, 2019.

Zhao Song, David Woodruff, Zheng Yu, and Lichen Zhang. Fast sketching of polynomial kernels of polynomial degree. In *International Conference on Machine Learning*, pp. 9812–9823. PMLR, 2021a.

Zhao Song, Lichen Zhang, and Ruizhe Zhang. Training multi-layer over-parametrized neural network in subquadratic time. *arXiv preprint arXiv:2112.07628*, 2021b.

Zhao Song, Baocheng Sun, Omri Weinstein, and Ruizhe Zhang. Quartic samples suffice for fourier interpolation. *arXiv preprint arXiv:2210.12495*, 2022a.

Zhao Song, Baocheng Sun, Omri Weinstein, and Ruizhe Zhang. Sparse fourier transform over lattices: A unified approach to signal reconstruction. *CoRR*, abs/2205.00658, 2022b.

Zhao Song, Zhaozhuo Xu, Yuanyuan Yang, and Lichen Zhang. Accelerating frank-wolfe algorithm using low-dimensional and adaptive data structures. *arXiv preprint arXiv:2207.09002*, 2022c.

Zhao Song, Zhaozhuo Xu, and Lichen Zhang. Speeding up sparsification using inner product search data structures. *arXiv preprint arXiv:2204.03209*, 2022d.

Andreas M Tillmann and Marc E Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 60(2):1248–1259, 2013.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

Thakshila Wimalajeewa and Pramod K Varshney. Application of compressive sensing techniques in distributed sensor networks: A survey. *arXiv preprint arXiv:1709.10401*, 2017.

Fan Wu and Patrick Rebeschini. Implicit regularization in matrix sensing via mirror descent. *Advances in Neural Information Processing Systems*, 34:20558–20570, 2021.

Zhaozhuo Xu, Zhao Song, and Anshumali Shrivastava. Breaking the linear iteration cost barrier for some well-known conditional gradient methods using maxip data-structures. *Advances in Neural Information Processing Systems*, 34:5576–5589, 2021.

Lichen Zhang. Speeding up optimizations via data structures: Faster search, sample and maintenance. Master's thesis, Carnegie Mellon University, 2022.

Lichen Zhang. Personal communication. In *MIT*, 2023.

Kai Zhong, Prateek Jain, and Inderjit S Dhillon. Efficient matrix sensing using rank-1 gaussian measurements. In *Algorithmic Learning Theory: 26th International Conference, ALT 2015, Banff, AB, Canada, October 4-6, 2015, Proceedings 26*, pp. 3–18. Springer, 2015.

Appendix

CONTENTS

1 Introduction **1**

2 Preliminary **2**

2.1 Notations 2

2.2 Randomness Facts 3

2.3 Matrix Concentration 3

2.4 Restricted Isometry Property 4

2.5 Rank-one Estimation 4

2.6 Rank-one Independent Gaussian Operator 4

2.7 Matrix Angle and Distance 4

3 Main Result **5**

3.1 Key concepts 5

3.2 Properties of Operators 5

3.3 Main Result 6

4 Technique Overview **6**

4.1 Tighter Analysis Implies Reduction to Sample Complexity 6

4.2 Induction Implies Correctness 7

4.3 Speeding up with Sketching Technique 8

5 Conclusion **8**

A Related Work **14**

B Preliminary **15**

B.1 Notations 15

B.2 Algebra Facts 16

C ANALYSIS **16**

C.1 Main Induction Hypothesis 17

D MEASUREMENTS ARE GOOD OPERATOR **18**

D.1 Tools for Gaussian 19

D.2 Bounding $\|Z_i\|$ 20

D.3 Bounding $\|\mathbb{E}[Z_i Z_i^\top]\|$ 21

D.4 Main Results 22

D.5 Initialization Is a Good Operator 22

D.6 Operator B and G is good 25

E ONE SHRINKING STEP **26**

E.1 Definitions of B, C, D, S 26

E.2 Upper Bound on $\|BD - C\|$ 27

E.3 Rewrite V_{t+1} 29

E.4 Upper bound on $\|F\|$ 30

E.5 Upper bound on $\|R^{-1}\|$ 32

F MATRIX SENSING REGRESSION **33**

F.1 Definition and Equivalence 33

F.2 From Sensing Matrix to Regression Matrix 34

F.3 Our Fast Regression Solver 34

F.4 Straightforward Solver 35

F.5 Condition Number 35

G Limitations **36**

Roadmap. We organize the appendix as follows. In Section C we provide the detailed analysis for our algorithm. In Section D we argue that our measurements are good. In Section E we provide analysis for a shrinking step. In Section F we provide the analysis for our techniques used to solve the optimization problem at each iteration.

A Related Work

Matrix Sensing The matrix sensing problem has attracted significant attention in recent years, and several algorithms have been proposed to solve it efficiently. One of the earliest approaches is the convex optimization-based algorithm proposed by Candès and Recht in 2009 Candes & Recht (2012), which minimizes the nuclear norm of the matrix subject to the linear constraints. This approach has been shown to achieve optimal recovery guarantees under certain conditions on the linear operators, such as incoherence and RIP. Since then, various algorithms have been proposed that improve upon the original approach in terms of computational efficiency and theoretical guarantees. For instance, the iterative hard thresholding algorithm (IHT) proposed by Blumensath and Davies in 2009 Blumensath & Davies (2008), and its variants, such as the iterative soft thresholding algorithm (IST), provide computationally efficient solutions with improved recovery guarantees. In the work by Recht, Fazel, and Parrilo Recht et al. (2010), they gave some measurement operators satisfying the RIP and proved that, with $O(kd \log d)$ measurements, a rank-k matrix $W_* \in \mathbb{R}^{d \times d}$ can be recovered. Moreover, later works have proposed new approaches that exploit additional structure in the low-rank matrix, such as sparsity or group sparsity, to further improve recovery guarantees and efficiency. For instance, the sparse plus low-rank ($S + L$) approach proposed by Liu et al. (2012), and its variants, such as the robust principal component analysis (RPCA) and the sparse subspace clustering (SSC), provide efficient solutions with improved robustness to outliers and noise. More recently, Park et al. (2017) considers the non-square matrix sensing under RIP assumptions, and show that matrix factorization does not introduce any spurious local minima under RIP. Wu & Rebeschini (2021) studies the technique of discrete-time mirror descent utilized to address the unregularized empirical risk in matrix sensing.

Compressive Sensing Compressive sensing has been a widely studied topic in signal processing and theoretical computer science field Hassanieh et al. (2012a;b); Pawar & Ramchandran (2013); Indyk & Kapralov (2014); Price & Song (2015); Bakshi et al. (2015); Kapralov (2016); Chen et al. (2016); Kapralov (2017); Nakos & Song (2019); Nakos et al. (2019); Avron et al. (2019); Li et al. (2019); Liu & Venkataramanan (2023). Hassanieh et al. (2012a) gave a fast algorithm (runs in

time $O(k \log n \log(n/k))$ for general inputs and $O(k \log n \log(n/k))$ for at most k non-zero Fourier coefficients input) for k-sparse approximation to the discrete Fourier transform of an n-dimensional signal. Kapralov (2016) provided an algorithm that uses $O_d(k \log N \log \log N)$ samples of signal and runs in time $O_d(k \log^{d+3} N)$ for k-sparse approximation to the Fourier transform of a length of N signal. Later work Kapralov (2017) proposed a new technique for analysing noisy hashing schemes that arise in Sparse FFT, which is called isolation on average, and applying it, it achieves sample-optimal results in $k \log^{O(1)} n$ time for estimating the values of a list of frequencies using few samples and computing Sparse FFT itself. Nakos & Song (2019) gave the first sublinear-time ℓ_2/ℓ_2 compressed sensing which achieves the optimal number of measurements without iterating. After that, Nakos et al. (2019) provided an algorithm which uses $O(k \log k \log n)$ samples to compute a k-sparse approximation to the d-dimensional Fourier transform of a length n signal. Later by Song et al. (2022a) provided an efficient Fourier Interpolation algorithm that improves the previous best algorithm Chen et al. (2016) on sample complexity, time complexity and output sparsity. And in Song et al. (2022b) they presented a unified framework for the problem of band-limited signal reconstruction and achieves high-dimensional Fourier sparse recovery and high-accuracy Fourier interpolation. Recent work Jin et al. (2023) designed robust algorithms for super-resolution imaging that are efficient in terms of both running time and sample complexity for any constant dimension under the same noise model as Price & Song (2015), based on new techniques in Sparse Fourier transform.

Faster Iterative Algorithm via Sketching Low rank matrix completion is a well-known problem in machine learning with various applications in practical fields such as recommender systems, computer vision, and signal processing. Some notable surveys of this problem are provided in Koren (2009); Nguyen et al. (2019). While Candes and Recht Candes & Recht (2012) first proved the sample complexity for low rank matrix completion, other works such as Candès & Tao (2010) and Jain et al. (2013) have provided improvements and guarantees on convergence for heuristics. In recent years, sketching has been applied to various machine learning problems such as linear regression Clarkson & Woodruff (2017); Nelson & Nguyên (2013), low-rank approximation Clarkson & Woodruff (2017); Nelson & Nguyên (2013), weighted low rank approximation, matrix CUR decomposition Boutsidis & Woodruff (2014); Song et al. (2017; 2019), and tensor regression Diao et al. (2018; 2019); Song et al. (2021a); Reddy et al. (2022), leading to improved efficiency of optimization algorithms in many problems. For example, linear programming Cohen et al. (2019); Song & Yu (2021); Jiang et al. (2021); Gu & Song (2022), matrix completion Gu et al. (2023), empirical risk minimization Lee et al. (2019); Qin et al. (2023), training over-parameterized neural network Brand et al. (2021); Song et al. (2021b); Jiang et al. (2022); Zhang (2022), discrepancy algorithm Zhang (2022); Song et al. (2022d); Deng et al. (2022), frank-wolfe method Xu et al. (2021); Song et al. (2022c), and reinforcement learning Shrivastava et al. (2023).

B PRELIMINARY

In Section B.1 we state our notations. In Section B.2 we provide some algebra facts.

B.1 NOTATIONS

Let $x \in \mathbb{R}^n$ and $w \in \mathbb{R}^n_{\geq 0}$, we define the norm $\|x\|_w := (\sum_{i=1}^n w_i x_i^2)^{1/2}$. For $n > k$, for any matrix $A \in \mathbb{R}^{n \times k}$, we denote the spectral norm of A by $\|A\|$. Let $A \in \mathbb{R}^{n \times k}$, we denote the Frobenius norm of A by $\|A\|_F$. For any square matrix $A \in \mathbb{R}^{n \times n}$, we denote its trace by $\mathrm{tr}[A]$. For any $A \in \mathbb{R}^{n \times d}$ and $B \in \mathbb{R}^{n \times d}$, we denote $\langle A, B \rangle = \mathrm{tr}[A^\top B]$. Let $A \in \mathbb{R}^{n \times d}$ and $x \in \mathbb{R}^d$ be any matrix and vector, we have that $\|Ax\|_2^2 = \langle Ax, Ax \rangle = \langle x, A^\top A x \rangle = x^\top A^\top A x$. Let the SVD decomposition of $A \in \mathbb{R}^{n \times k}$ to be $A = U \Sigma V^\top$, where $U \in \mathbb{R}^{n \times k}$ and $V \in \mathbb{R}^{k \times k}$ have orthonormal columns and $\Sigma \in \mathbb{R}^{k \times k}$ be diagonal matrix. We say the columns of U are the singular vectors of A. We denote the Moore-Penrose pseudoinverse matrix of A as $A^\dagger \in \mathbb{R}^{k \times n}$, i.e., $A^\dagger := V \Sigma^{-1} U^\top$. We call the diagonal entries $\sigma_1, \sigma_2, \ldots, \sigma_k$ of Σ to be the eigenvalues of A. We assume they are sorted from largest to lowest, so σ_i denotes its i-th largest eigenvalue, and we can write it as $\sigma_i(A)$. For $A \in \mathbb{R}^{n_1 \times d_1}$, $B \in \mathbb{R}^{n_2 \times d_2}$. We define kronecker product $\otimes$ as $(A \otimes B)_{i_1 + (i_2 - 1) n_1, j_1 + (j_2 - 1) n_2} = A_{i_1, j_1} B_{i_2, j_2}$ for all $i_1 \in [n_1]$, $j_1 \in [d_1]$, $i_2 \in [n_2]$ and $j_2 \in [d_2]$. For any non-singular matrix $A \in \mathbb{R}^{n \times n}$, we define $A = QR$ its QR-decomposition, where $Q \in \mathbb{R}^{n \times n}$ is an orthogonal matrix and $R \in \mathbb{R}^{n \times n}$ is

an non-singular lower triangular matrix. For any full-rank matrix $A \in \mathbb{R}^{n \times m}$, we define $A = QR$ its QR-decomposition, where $Q \in \mathbb{R}^{m \times n}$ is an orthogonal matrix and $R \in \mathbb{R}^{n \times n}$ is an non-singular lower triangular matrix. We use $R = \mathrm{QR}(A) \in \mathbb{R}^{n \times n}$ to denote the lower triangular matrix obtained by the QR-decomposition of $A \in \mathbb{R}^{m \times n}$. Let $A \in \mathbb{R}^{k \times k}$ be a symmetric matrix. The eigenvalue decomposition of A is $A = U \Lambda U^\top$, where Λ is a diagonal matrix. If a matrix A is positive semidefinite (PSD), we denote it as $A \succeq 0$, which means $x^\top A x \geq 0$ for all x. Similarly, we say $A \succeq B$ if $x^\top A x \geq x^\top B x$ for all vectors x. For any matrix $U \in \mathbb{R}^{n \times k}$, we say U is an orthonormal basis if $\|U_i\| = 1$ for all $i \in [k]$ and for any $i \neq j$, we have $\langle U_i, U_j \rangle = 0$. Here for each $i \in [k]$, we use U_i to denote the i-th column of matrix U. For any $U \in \mathbb{R}^{n \times k}$ (suppose $n > k$) which is an orthonormal basis, we define $U_\perp \in \mathbb{R}^{n \times (n-k)}$ to be another orthonormial basis that, $UU^\top + U_\perp U_\perp^\top = I_n$ and $U^\top U_\perp = \mathbf{0}^{k \times (n-k)}$, where we use $\mathbf{0}^{k \times (n-k)}$ to denote a $k \times (n-k)$ all-zero matrix. We say a vector x lies in the span of U, if there exists a vector y such that $x = Uy$. We say a vector z lies in the complement of span of U, if there exists a vector w such that $z = U_\perp w$. Then it is obvious that $\langle x, z \rangle = x^\top z = z^\top x = 0$. For a matrix A, we define $\sigma_{\min}(A) := \min_x \|Ax\|_2 / \|x\|_2$. Equivalently, $\sigma_{\min}(A) := \min_{x : \|x\|_2 = 1} \|Ax\|_2$. Similarly, we define $\sigma_{\max}(A) := \max_x \|Ax\|_2 / \|x\|_2$. Equivalently, $\sigma_{\max}(A) := \max_{x : \|x\|_2 = 1} \|Ax\|_2$ Let $A_1, \cdots, A_n$ denote a list of square matrices. Let S denote a block diagonal matrix $S = \mathrm{diag}(A_1, A_2, \cdots, A_n)$. Then $\|S\| = \max_{i \in [n]} \|A_i\|$. We use $\Pr[]$ to denote probability. We use $\mathbb{E}[]$ to denote expectation. Let a and b denote two random variables. Let $f(a)$ denote some event that depends on a (for example $f(a)$ can be $a = 0$ or $a \geq 10$.). Let $g(b)$ denote some event that depends on b. We say a and b are independent if $\Pr[f(a) \text{ and } g(b)] = \Pr[f(a)] \cdot \Pr[g(b)]$. We say a and b are not independent if $\Pr[f(a) \text{ and } g(b)] \neq \Pr[f(a)] \cdot \Pr[g(b)]$. Usually if a and b are independent, then we also have $\mathbb{E}[ab] = \mathbb{E}[a] \cdot \mathbb{E}[b]$. We say a random variable x is symmetric if $\Pr[x = u] = \Pr[x = -u]$. For any random variable $x \sim \mathcal{N}(\mu, \sigma^2)$. This means $\mathbb{E}[x] = \mu$ and $\mathbb{E}[x^2] = \sigma^2$. We use $\widetilde{O}(f)$ to denote $f \cdot \mathrm{poly}(\log f)$. We use $\mathcal{T}_{\mathrm{mat}}(a, b, c)$ to denote the time of multiplying an $a \times b$ matrix with another $b \times c$ matrix. We use ω to denote the exponent of matrix multiplication, i.e., $n^\omega = \mathcal{T}_{\mathrm{mat}}(n, n, n)$.

B.2 Algebra Facts

We state some standard facts and omit their proofs, since they're very standard.

Fact B.1. *We have*

- *For any orthonormal basis $U \in \mathbb{R}^{n \times k}$ and a vector $x \in \mathbb{R}^k$, we have $\|Ux\|_2 = \|x\|_2$.*
- *For any orthonornal basis $U \in \mathbb{R}^{n \times k}$, we have $\|U\|_F \leq \sqrt{k}$.*
- *For any diagonal matrix $\Sigma \in \mathbb{R}^{k \times k}$ and any vector $x \in \mathbb{R}^k$, we have $\|\Sigma x\|_2 \geq \sigma_{\min}(\Sigma) \|x\|_2$.*
- *For symmetric matrix A, we have $\sigma_{\min}(A) = \min_{z : \|z\|_2 = 1} z^\top A z$ and $\|A\| \geq z^\top A z$ for all $\|z\|_2 = 1$.*
- *For symmetric matrix A, we have $\sigma_{\min}(A) \|z\|_2^2 \leq z^\top A z$ for all vectors z.*
- *For symmetric matrix A, we have $\sigma_{\max}(A) \|z\|_2^2 \geq z^\top A z$ for all vectors z.*
- *For any matrix A, we have $\|A\| \leq \|A\|_F$.*
- *For any square matrix $A \in \mathbb{R}^{k \times k}$ and vector $x \in \mathbb{R}^k$, we have $x^\top A x = \sum_{i=1}^k \sum_{j=1}^k x_i A_{i,j} x_j = \sum_{i=1}^k x_i A_{i,i} x_i + \sum_{i \neq j} x_i A_{i,j} x_j$.*
- *For any square and invertible matrix $R \in \mathbb{R}^{n \times n}$, we have $\|R^{-1}\| = \sigma_{\min}(R)^{-1}$*
- *For any matrix $A \in \mathbb{R}^{n \times k}$ and for any unit vector $x \in \mathbb{R}^k$, we have $\|A\| \geq \|Ax\|_2$.*
- *For any matrix $A \in \mathbb{R}^{n \times k}$, $\|AA^\top\| = \|A^\top A\|$.*

C ANALYSIS

Here in this section, we provide analysis for our proposed algorithm.

Algorithm 1 Our Faster Matrix Sensing Algorithm

procedure FASTMATRIXSENSING($\mathcal{A}_{all} \subset \mathbb{R}^{d\times d}, b_{all} \subset \mathbb{R}, \epsilon_0 \in (0, 0.1), \epsilon \in (0, 0.1), \delta \in (0, 0.1)$) ▷ Theorem 1.1
  ▷ Let b_{all} scalar measurements
  ▷ Let $\mathcal{A}_{all}$ sensing matrices
  ▷ Let $W_* \in \mathbb{R}^{d\times d}$ denote a rank-k matrix
  ▷ Let σ_1^* denote the largest singular value of W_*
  ▷ Let κ denote the condition number of W_*
  $T \leftarrow \Theta(\log(k\kappa\sigma_1^*/\epsilon_0))$
  $m \leftarrow \Theta(\epsilon^{-2}(d + k^2)\log(d/\delta))$
  Split $(\mathcal{A}_{all}, b_{all})$ into $2T+1$ sets (each of size m) with t-th set being $\mathcal{A}^t \subset \mathbb{R}^{d\times d}$ and $b^t \in \mathbb{R}$
  $U_0 \leftarrow$ top-k left singular vectors of $\frac{1}{m}\sum_{l=1}^m b_l^0 A_l^0$
  for $t \leftarrow 0$ **to** $T-1$ **do**
    $b \leftarrow b^{2t+1}, \mathcal{A} \leftarrow \mathcal{A}^{2t+1}$
    $\widehat{V}_{t+1} \leftarrow \arg\min_{V\in\mathbb{R}^{d\times k}} \sum_{l=1}^m (b_l - x_l^\top U_t V^\top y_l)^2$ ▷ Using Lemma F.7
    $V_{t+1} \leftarrow \mathrm{QR}(\widehat{V}_{t+1})$ ▷ orthonormalization of $\widehat{V}_{t+1}$
    $b \leftarrow b^{2t+2}, \mathcal{A} \leftarrow \mathcal{A}^{2t+2}$
    $\widehat{U}_{t+1} \leftarrow \arg\min_{U\in\mathbb{R}^{d\times k}} \sum_{l=1}^m (b_l - x_l^\top U V_{t+1}^\top y_l)^2$ ▷ Using Lemma F.7
    $U_{t+1} \leftarrow \mathrm{QR}(\widehat{U}_{t+1})$ ▷ orthonormalization of $\widehat{U}_{t+1}$
  end for
  $W_T \leftarrow U_T(\widehat{V}_T)^\top$
  return W_T
end procedure

C.1 MAIN INDUCTION HYPOTHESIS

Lemma C.1 (Induction hypothesis)**.** *We define $\epsilon_d := 1/10$. We assume that $\epsilon = 0.001/(k^{1.5}\kappa)$. For all $t \in [T]$, we have the following results.*

- *Part 1. If $\mathrm{dist}(U_t, U_*) \leq \frac{1}{4}\mathrm{dist}(V_t, V_*) \leq \epsilon_d$, then we have*
  - $\mathrm{dist}(V_{t+1}, V_*) \leq \frac{1}{4}\mathrm{dist}(U_t, U_*) \leq \epsilon_d$
- *Part 2. If $\mathrm{dist}(V_{t+1}, V_*) \leq \frac{1}{4}\mathrm{dist}(U_t, U_*) \leq \epsilon_d$, then we have*
  - $\mathrm{dist}(U_{t+1}, U_*) \leq \frac{1}{4}\mathrm{dist}(V_{t+1}, V_*) \leq \epsilon_d$

Proof. **Proof of Part 1.**

Recall that for each $i \in [n]$, we have

$$b_i = x_i^\top W_* y_i = \langle x_i y_i^\top, W_* \rangle = \langle A_i, W_* \rangle = \mathrm{tr}[A_i^\top W_*].$$

Recall that

$$\begin{aligned}\widehat{V}_{t+1} = & \arg\min_{V\in\mathbb{R}^{d\times k}} \sum_{i=1}^m (b_i - x_i^\top U_t V^\top y_i)^2 \\ = & \arg\min_{V\in\mathbb{R}^{d\times k}} \sum_{i=1}^m (x_i^\top W_* y_i - x_i^\top U_t V^\top y_i)^2\end{aligned}$$

Hence, by setting gradient of this objective function to zero and let $F \in \mathbb{R}^{d\times k}$ be defined as Definition E.1. We have $\widehat{V}_{t+1} \in \mathbb{R}^{d\times k}$ can be written as follows:

$$\widehat{V}_{t+1} = W_*^\top U_t - F \tag{4}$$

where $F \in \mathbb{R}^{d\times k}$ is the error matrix

$$F = [F_1 \quad F_2 \quad \cdots \quad F_k]$$

where $F_i \in \mathbb{R}^d$ for each $i \in [k]$.

Then, using the definitions of $F \in \mathbb{R}^{d \times k}$ and Definition E.1, we get:

$$\begin{bmatrix} F_1 \\ \vdots \\ F_k \end{bmatrix} = B^{-1}(BD - C)S \cdot \mathrm{vec}(V_*) \tag{5}$$

where $\mathrm{vec}(V_*) \in \mathbb{R}^{dk}$ is the vectorization of matrix $V_* \in \mathbb{R}^{d \times k}$.

Now, recall that in the $t+1$-th iteration of Algorithm 1, $V_{t+1} \in \mathbb{R}^{d \times k}$ is obtained by QR decomposition of $\widehat{V}_{t+1} \in \mathbb{R}^{d \times k}$. Using notation mentioned above,

$$\widehat{V}_{t+1} = V_{t+1} R \tag{6}$$

where $R \in \mathbb{R}^{k \times k}$ denotes the lower triangular matrix $R_{t+1} \in \mathbb{R}^{k \times k}$ obtained by the QR decomposition of $V_{t+1} \in \mathbb{R}^{d \times k}$.

We can rewrite $V_{t+1} \in \mathbb{R}^{d \times k}$ as follows

$$\begin{aligned} V_{t+1} &= \widehat{V}_{t+1} R^{-1} \\ &= (W_*^\top U_t - F) R^{-1} \end{aligned} \tag{7}$$

where the first step follows from Eq. (6) , and the last step follows from Eq. (4).

Multiplying both the sides by $V_{*,\perp} \in \mathbb{R}^{d \times (d-k)}$, where $V_{*,\perp} \in \mathbb{R}^{d \times (d-k)}$ is a fixed orthonormal basis of the subspace orthogonal to $\mathrm{span}(V_*)$, using Claim E.3

$$(V_{*,\perp})^\top V_{t+1} = -(V_{*,\perp})^\top F R^{-1} \tag{8}$$

Thus, we get:

$$\begin{aligned} \mathrm{dist}(V_{t+1}, V_*) &= \|(V_{*,\perp})^\top V_{t+1}\| \\ &= \|(V_{*,\perp})^\top F R^{-1}\| \\ &= \|F R^{-1}\| \\ &\leq \|F\| \cdot \|R^{-1}\| \\ &\leq 0.001 \sigma_k^* \, \mathrm{dist}(U_t, U_*) \cdot \|R^{-1}\| \\ &\leq 0.001 \sigma_k^* \, \mathrm{dist}(U_t, U_*) \cdot 2(\sigma_k^*)^{-1} \\ &\leq 0.01 \cdot \mathrm{dist}(U_t, U_*) \end{aligned}$$

where the first step follows from definition of dist (see Definition 2.6), the second step follows from Eq. (8), the third step follows from $V_{*,\perp}$ is an orthonormal basis, and the forth step follows from Fact B.1, the fifth step follows from Lemma. E.4, the sixth step follows from Lemma E.5 (In order to run this lemma, we need to the condition of Part 1 statement to be holding), the last step follows from simple algebra.

Proof of Part 2.

Similarly, we can prove this as Part 1.

□

D MEASUREMENTS ARE GOOD OPERATOR

In this section, we provide detailed analysis for our operators. First Section D.1 we introduce some standard results for truncated Gaussian. In Section D.2 and Section D.3 we bound the term $\|Z_i\|$ and $\|\mathbb{E}[Z_i Z_i^\top]\|$ respectively. In Section D.4 we state our main lemma. In Section D.5 we show that out initialization is good. In Section D.6 we show our two operators are good.

D.1 TOOLS FOR GAUSSIAN

We state a standard tool from literature,

Lemma D.1 (Lemma 1 in Laurent & Massart (2000)). *Let $X \sim \mathcal{X}_k^2$ be a chi-squared distributed random variable with k degrees of freedom. Each one has zero means and σ^2 variance.*

Then it holds that

$$\Pr[X - k\sigma^2 \geq (2\sqrt{kt} + 2t)\sigma^2] \leq \exp(-t)$$
$$\Pr[k\sigma^2 - X \geq 2\sqrt{kt}\sigma^2] \leq \exp(-t)$$

Further if $k \geq \Omega(\epsilon^{-2}t)$ and $t \geq \Omega(\log(1/\delta))$, then we have

$$\Pr[|X - k\sigma^2| \leq \epsilon k\sigma^2] \leq \delta.$$

We state a standard fact for the 4-th moment of Gaussian distribution.

Fact D.2. *Let $x \sim \mathcal{N}(0, \sigma^2)$, then it holds that $\mathbb{E}_{x\sim\mathcal{N}(0,\sigma^2)}[x^4] = 3\sigma^4$.*

Lemma D.3. *Let $x \sim \mathcal{N}(0, \sigma^2 I_d)$ denote a random Gaussian vector. Then we have*

- *Part 1*

$$\mathbb{E}[xx^\top xx^\top] = (d+2)\sigma^4$$

- *Part 2*

$$\|\mathbb{E}[xx^\top xx^\top]\| = (d+2)\sigma^4$$

Proof. We define $A := xx^\top xx^\top$. Then we have

$$A_{i,j} = x_i \sum_{l=1}^{d} x_l x_l x_j$$

For $i = j$, we have

$$\begin{aligned}
\mathbb{E}[A_{i,i}] = & \mathbb{E}[x_i \sum_{l=1}^{d} x_l x_l x_i] \\
= & \mathbb{E}[x_i (\sum_{l=1}^{i-1} x_l x_l + x_i x_i + \sum_{l=i+1}^{d} x_l x_l) x_i] \\
= & \mathbb{E}[x_i^4] + \sum_{l\in[d]\setminus i} \mathbb{E}[x_l^2 x_i^2] \\
= & \mathbb{E}[x_i^4] + \sum_{l\in[d]\setminus i} \mathbb{E}[x_l^2]\,\mathbb{E}[x_i^2] \\
= & \mathbb{E}[x_i^4] + (d-1)\sigma^4 \\
= & 3\sigma^4 + (d-1)\sigma^4 \\
= & (d+2)\sigma^4
\end{aligned}$$

where the third step follows from linearity of expectation (Fact 2.1), the forth step follows from x_l and x_i are independent, the fifth step follows from $\mathbb{E}[x_l^2] = \sigma^2$, the sixth step follows $\mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2)}[z^4] = 3\sigma^4$.

For $i \neq j$, we have

$$\mathbb{E}[A_{i,j}] = \mathbb{E}[x_i \sum_{l=1}^{d} x_l x_l x_j]$$

$$= \mathbb{E}[x_i x_j^3] + \mathbb{E}[x_i^3 x_j] + \sum_{l \in [d] \setminus i,j} \mathbb{E}[x_i x_l^2 x_j]$$
$$= 0$$

where the second step follows from linearity of expectation (Fact 2.1), and the last step follows from $\mathbb{E}[x_i] = 0$.

□

Fact D.4 (Rotation invariance property of Gaussian)**.** *Let $A^\top \in \mathbb{R}^{d \times k}$ with $k < d$ denote an orthonormal basis (i.e., $AA^\top = I_k$). Then for a Gaussian $x \sim \mathcal{N}(0, \sigma^2 I_d)$, we have*

$$Ax \sim \mathcal{N}(0, \sigma^2 I_k).$$

Proof. Let $y := Ax \in \mathbb{R}^k$, then

$$y_i = \sum_{j=1}^d A_{ij} x_j, \ \ \forall i \in [k].$$

By definition of Gaussian distribution

$$y_i \sim \mathcal{N}(0, \sigma^2 \sum_{j=1}^d A_{ij}^2).$$

Recall that $A^\top$ is an orthonormal basis.

We have

$$\sum_{j=1}^d A_{ij}^2 = 1.$$

Thus we have

$$y \sim \mathcal{N}(0, \sigma^2 I_k),$$

□

D.2 Bounding $\|Z_i\|$

Lemma D.5. *Let x_i denote a random Gaussian vector samples from $\mathcal{N}(0, \sigma^2 I_d)$. Let y_i denote a random Gaussian vector samples from $\mathcal{N}(0, \sigma^2 I_d)$.*

Let $U_*, V_* \in \mathbb{R}^{d \times k}$.

We define

$$Z_i := x_i x_i^\top U_* \Sigma_* V_*^\top y_i y_i^\top, \quad \forall i \in [m]$$

- *Part 1. We have*

$$\Pr[\|Z_i\| \leq C^2 k^2 \log^2(d/\delta) \sigma^4 \cdot \sigma_1^*] \geq 1 - \delta / \operatorname{poly}(d).$$

- *Part 2. If $k \geq \Omega(\log(d/\delta))$ We have*

$$\Pr[\|Z_i\| \leq C^2 k^2 \sigma^4 \cdot \sigma_1^*] \geq 1 - \delta / \operatorname{poly}(d).$$

Proof. **Proof of Part 1.**

We define

$$a_i := U_*^\top x_i \in \mathbb{R}^k$$

$$b_i := V_*^\top y_i \in \mathbb{R}^k$$

Since U_* and V_* are orthornormal basis, due to rotation invariance property of Gaussian (Fact D.4) , we know that $a_i \sim \mathcal{N}(0, \sigma^2 I_k)$ and $b_i \sim \mathcal{N}(0, \sigma^2 I_k)$.

We also know that

$$\begin{aligned} x_i &= (U_*^\top)^\dagger a_i = U_* a_i \\ y_i &= (V_*^\top)^\dagger b_i = V_* b_i \end{aligned}$$

Thus, by replacing x_i, y_i with a_i, b_i, we have

$$\begin{aligned} \|Z_i\| &= \|x_i x_i^\top U_* \Sigma_* V_*^\top y_i y_i^\top\| \\ &= \|U_* a_i a_i^\top U_*^\top U_* \Sigma_* V_*^\top V_* b_i b_i^\top V_*^\top\| \\ &= \|U_* a_i a_i^\top \Sigma_* b_i b_i^\top V_*^\top\| \\ &\leq \|U_*\| \cdot \|a_i a_i^\top\| \cdot \|\Sigma_*\| \cdot \|b_i b_i^\top\| \cdot \|V_*^\top\| \\ &\leq \sigma_1^* \cdot \|a_i\|_2^2 \cdot \|b_i\|_2^2 \end{aligned}$$

where the second step follows from replacing x, y by a, b, the third step follows from $U_*^\top U_* = I$ and $V_*^\top V_* = I$, the forth step follows from Fact B.1, the last step follows from $\|\Sigma_*\| = \sigma_1^*$ and $\|U_*\|, \|V^*\| \leq 1$ (since they are orthonormal basis).

Due to property of Gaussian, we know that

$$\Pr[|a_{i,j}| > \sqrt{C \log(d/\delta)}\sigma] \leq \delta / \operatorname{poly}(d)$$

Taking a union bound over k coordinates, we know that

$$\Pr[\|a_i\|_2^2 \leq Ck \log(d/\delta)\sigma^2] \geq 1 - \delta / \operatorname{poly}(d)$$

Similarly, we can prove it for $\|b_i\|_2^2$.

Proof of Part 2. Since $k \geq \Omega(\log(d/\delta))$, then we can use Lemma D.1 to obtain a better bound.

□

D.3 Bounding $\|\mathbb{E}[Z_i Z_i^\top]\|$

Lemma D.6. *We can show that*

$$\|\mathbb{E}[Z_i Z_i^\top]\| \leq C^2 k^2 \sigma^4 (\sigma_1^*)^2.$$

Proof. Using Lemma D.3

$$\|\mathop{\mathbb{E}}_{a \sim \mathcal{N}(0, \sigma^2 I_k)}[a_i a_i^\top a_i a_i^\top]\| \leq Ck\sigma^2.$$

Thus, we have

$$\mathbb{E}[a_i a_i^\top a_i a_i^\top] \preceq Ck\sigma^2 \cdot I_k$$

Then, we have

$$\begin{aligned} \|\mathbb{E}[Z_i Z_i^\top]\| &= \|\mathop{\mathbb{E}}_{x,y}[x_i x_i^\top U_* \Sigma_* V_*^\top y_i y_i^\top y_i y_i^\top V_* \Sigma_* U_*^\top x_i x_i^\top]\| \\ &= \|\mathop{\mathbb{E}}_{a,b}[U_* a_i a_i^\top U_*^\top U_* \Sigma_* V_*^\top V_* b_i b_i^\top V_*^\top V_* b_i b_i^\top V_*^\top V_* \Sigma_* U_*^\top U_* a_i a_i^\top U_*^\top]\| \\ &= \|\mathop{\mathbb{E}}_{a,b}[U_* a_i a_i^\top \Sigma_* b_i b_i^\top V_*^\top V_* b_i b_i^\top \Sigma_* a_i a_i^\top U_*^\top]\| \end{aligned}$$

$$
\begin{aligned}
&= \| \mathop{\mathbb{E}}_{a,b}[U_* a_i a_i^\top \Sigma_* b_i b_i^\top b_i b_i^\top \Sigma_* a_i a_i^\top U_*^\top] \| \\
&\leq \| \mathop{\mathbb{E}}_{a,b}[a_i a_i^\top \Sigma_* b_i b_i^\top b_i b_i^\top \Sigma_* a_i a_i^\top] \| \\
&\leq \| \mathop{\mathbb{E}}_{a}[a_i a_i^\top \Sigma_* \mathop{\mathbb{E}}_{b}[b_i b_i^\top b_i b_i^\top] \Sigma_* a_i a_i^\top] \| \\
&\leq C^2 k^2 \sigma^4 (\sigma_1^*)^2 \qquad (9)
\end{aligned}
$$

where the first step follows from the definition of Z_i, the second step follows from replacing x_i, y_i with a_i, b_i, the third step follows from U_*, V_* are orthonormal columns, the fourth step follows from V_* are orthonormal columns, the fifth step follows from $\|U_*\| \leq 1$, the sixth step follows from simple algebra, the seventh step follows from using Lemma D.3 twice. □

D.4 Main Results

We prove our main result for measurements.

Theorem D.7 (Formal of Theorem 1.1, Measurements are good operator)**.** *Let $\{A_i, b_i\}_{i \in [m]}$ denote measurements be defined as Definition 3.3.*

Assuming the following conditions are holding

- $k = \Omega(\log(d/\delta))$
- $m = \Omega(\epsilon^{-2}(d + k^2)\log(d/\delta))$

Then,

- *The property in Definition 3.4, initialization is a ϵ-operator*
- *The property in Definition 3.5, B are ϵ-operator.*
- *The property in Definition 3.6, G are ϵ-operator.*

holds with probability at least $1 - \delta/\operatorname{poly}(d)$.

Proof. Using Lemma D.8 and Lemma D.9, we complete the proof. □

D.5 Initialization Is a Good Operator

Lemma D.8. *We define matrix $S \in \mathbb{R}^{d \times d}$ as follows*

$$
S := \frac{1}{m} \sum_{i=1}^{m} b_i A_i.
$$

If the following two condition holds

- *Condition 1.* $k = \Omega(\log(d/\delta))$,
- *Condition 2.* $m = \Omega(\epsilon^{-2} k^2 \log(d/\delta))$.

Then we have

$$
\Pr[\|S - W_*\| \leq \epsilon \cdot \|W_*\|] \geq 1 - \delta.
$$

Proof. (Initialization in Definition 3.4) Now, we have:

$$
S = \frac{1}{m} \sum_{i=1}^{m} b_i A_i
$$

$$
\begin{aligned}
&= \frac{1}{m} \sum_{i=1}^{m} b_i x_i y_i^\top \\
&= \frac{1}{m} \sum_{i=1}^{m} x_i b_i y_i^\top \\
&= \frac{1}{m} \sum_{i=1}^{m} x_i x_i^\top W_* y_i y_i^\top \\
&= \frac{1}{m} \sum_{i=1}^{m} x_i x_i^\top U_* \Sigma_* V_*^\top y_i y_i^\top,
\end{aligned}
$$

where the first step follows from Definition 3.4, the second step follows from $A_i = x_i y_i^\top$, the third step follows from b_i is a scalar, the forth step follows from $b_i = x_i^\top W_* y_i$, the fifth step follows from $W_* = U_* \Sigma_* V_*^\top$.

For each $i \in [m]$, we define matrix $Z_i \in \mathbb{R}^{d \times d}$ as follows:

$$
Z_i := x_i x_i^\top U_* \Sigma_* V_*^\top y_i y_i^\top,
$$

then we can rewrite $S \in \mathbb{R}^{d \times d}$ in the following sense,

$$
S = \frac{1}{m} \sum_{i=1}^{m} Z_i
$$

Note that, we can compute $\mathbb{E}[Z_i] \in \mathbb{R}^{d \times d}$

$$
\begin{aligned}
\mathop{\mathbb{E}}_{x_i, y_i}[Z_i] = & \mathop{\mathbb{E}}_{x_i, y_i}[\underbrace{x_i x_i^\top}_{d \times d} \underbrace{U_* \Sigma_* V_*^\top}_{d \times d} \underbrace{y_i y_i^\top}_{d \times d}] \\
= & \mathop{\mathbb{E}}_{x_i}[\underbrace{x_i x_i^\top}_{d \times d} \underbrace{U_* \Sigma_* V_*^\top}_{d \times d}] \cdot \mathop{\mathbb{E}}_{y_i}[\underbrace{y_i y_i^\top}_{d \times d}] \\
= & \mathop{\mathbb{E}}_{x_i}[x_i x_i^\top] \cdot U_* \Sigma_* V_*^\top \cdot \mathop{\mathbb{E}}_{y_i}[y_i y_i^\top] \\
= & U_* \Sigma_* V_*^\top
\end{aligned}
$$

where the first step follows definition of Z_i, the second step follows from x_i and y_i are independent and Fact 2.1, the third step follows from Fact 2.1 the forth step follows from $\mathbb{E}[x_i x_i^\top] = I_d$ and $\mathbb{E}[y_i y_i^\top] = I_d$.

As $S \in \mathbb{R}^{d \times d}$ is a sum of m random matrices, the goal is to apply Theorem 2.2 to show that S is close to

$$
\begin{aligned}
\mathbb{E}[S] = & \mathbb{E}[Z_i] \\
= & U_* \Sigma_* V_*^\top
\end{aligned}
$$

for large enough m.

Using Lemma D.5 (Part 2) with choosing Gaussian variance $\sigma^2 = 1$, we have

$$
\Pr[\|Z_i\| \leq C^2 k^2 \sigma_1^*, \forall i \in [m]] \geq 1 - \delta / \operatorname{poly}(d) \tag{10}
$$

Using Lemma D.6 with choosing Gaussian variance $\sigma^2 = 1$, we can bound $\| \mathbb{E}[Z_i Z_i^\top] \|$ as follows

$$
\| \mathbb{E}[Z_i Z_i^\top] \| \leq C^2 k^2 (\sigma_1^*)^2 \tag{11}
$$

Similarly, we can prove $\| \mathbb{E}[Z_i^\top Z_i] \|$ with the same bound.

Let $Z = \sum_{i=1}^{m}(Z_i - W_*)$.

Applying Theorem 2.2 we get

$$\Pr[\|Z\| \geq t] \leq 2d \cdot \exp(-\frac{t^2/2}{\mathrm{Var}[Z] + Mt/3}) \tag{12}$$

where

$$\begin{aligned} Z &= mS - mW_* \\ \mathrm{Var}[Z] &= m \cdot C^2 k^2 (\sigma_1^*)^2, && \text{by Eq. (11)} \\ M &= C^2 k^2 \sigma_1^* && \text{by Eq. (10)} \end{aligned}$$

Replacing $t = \epsilon \sigma_1^* m$ and $Z = mS - mW_*$ inside $\Pr[]$ in Eq. (12), we have

$$\Pr[\|S - W^*\| \geq \epsilon \sigma_1^*] \leq 2d \cdot \exp\left(- \frac{t^2/2}{\mathrm{Var}[Z] + Mt/3} \right)$$

Our goal is to choose m sufficiently large such that the above quantity is upper bounded by $2d \cdot \exp(-\Omega(\log(d/\delta)))$.

First, we need

$$\begin{aligned} \frac{t^2}{\mathrm{Var}[Z]} &= \frac{\epsilon^2 m^2 (\sigma_1^*)^2}{m \cdot C^2 k^2 (\sigma_1^*)^2} \\ &= \frac{\epsilon^2 m}{C^2 k^2} \\ &\geq \log(d/\delta) \end{aligned}$$

where the first step follows from choice of t and bound for $\mathrm{Var}[Z]$.

This requires

$$m \geq C^2 \epsilon^{-2} k^2 \log(d/\delta)$$

Second, we need

$$\begin{aligned} \frac{t^2}{Mt} &= \frac{\epsilon m \sigma_1^*}{M} \\ &= \frac{\epsilon m \sigma_1^*}{C^2 k^2 \sigma_1^*} \\ &= \frac{\epsilon m}{C^2 k^2} \\ &\geq \log(d/\delta) \end{aligned}$$

where the first step follows from choice of t and the second step follows from bound on M.

This requires

$$m \geq C^2 \epsilon^{-2} k^2 \log(d/\delta)$$

Finally, we should choose

$$m \geq 10 C^2 \epsilon^{-2} k^2 \log(d/\delta),$$

Which implies that

$$\Pr[\|S - W_*\| \leq \epsilon \cdot \sigma_1^*] \geq 1 - \delta/\operatorname{poly}(d). \tag{13}$$

Taking the union bound with all $\|Z_i\|$ are upper bounded, then we complete the proof.

□

D.6 OPERATOR B AND G IS GOOD

Lemma D.9. *If the following two conditions hold*

- *Condition 1.* $d = \Omega(\log(d/\delta))$
- *Condition 2.* $m = \Omega(\epsilon^{-2} d \log(d/\delta))$

Then operator B (see Definition 3.5) is ϵ good, i.e.,

$$\Pr[\|B_x - I_d\| \leq \epsilon] \geq 1 - \delta/\operatorname{poly}(d)$$
$$\Pr[\|B_y - I_d\| \leq \epsilon] \geq 1 - \delta/\operatorname{poly}(d)$$

Similar results hold for operator G (see Definition 3.6).

Proof. Recall that $B_x := \frac{1}{m}\sum_{l=1}^m (y_l^\top v)^2 x_l x_l^\top$.

Recall that $B_y := \frac{1}{m}\sum_{l=1}^m (x_l^\top u)^2 y_l y_l^\top$.

Now, as x_i, y_i are rotationally invariant random variables , wlog, we can assume $u = e_1$.

We use $x_{i,1} \in \mathbb{R}$ to denote the first entry of $x_i \in \mathbb{R}^d$.

Thus,

$$(x_i^\top u u^\top x_i) = x_{i,1}^2$$

Then

$$\mathbb{E}[(x_i^\top u u^\top x_i)^2] = \mathbb{E}[x_{i,1}^4] = 3$$

We define

$$Z_i = (x_i^\top u)^2 y_i y_i^\top$$

then

$$\mathbb{E}[Z_i] = I_d$$

Using similar idea in Lemma D.5, we have

$$\Pr[\|Z_i\| \leq Cd, \forall i \in [m]] \geq 1 - \delta/\operatorname{poly}(d)$$

We can bound

$$\begin{aligned}
\|\mathbb{E}[Z_i Z_i^\top]\| &= \|\mathop{\mathbb{E}}_{x,y}[(x_i^\top u)^2 y_i y_i^\top y_i y_i^\top (x_i^\top u)^2]\| \\
&= \|\mathop{\mathbb{E}}_{x}[(x_i^\top u)^2 \mathop{\mathbb{E}}_{y}[y_i y_i^\top y_i y_i^\top](x_i^\top u)^2]\| \\
&= (d+2) \cdot |\mathop{\mathbb{E}}_{x}[(x_i^\top u)^2 (x_i^\top u)^2]| \\
&= (d+2) \cdot 3 \\
&\leq Cd
\end{aligned}$$

where the fourth step follows from $C \geq 1$ is a sufficiently large constant.

Let $Z = \sum_{i=1}^m (Z_i - I_d)$.

Applying Theorem 2.2 we get

$$\Pr[\|Z\| \geq t] \leq 2d \cdot \exp(-\frac{t^2/2}{\operatorname{Var}[Z] + Mt/3}),$$

where

$$\begin{aligned}
Z &= m \cdot B - m \cdot I \\
\operatorname{Var}[Z] &= Cmd
\end{aligned}$$

$$M = Cd$$

Using $t = m\epsilon$ and $Z = \sum_{i=1}^m (Z_i - I_d)$, and $B = \frac{1}{m}\sum_{i=1}^m Z_i$, we have

$$\begin{aligned}\Pr[\|Z\| \geq t] = & \Pr[\|\sum_{i=1}^m (Z_i - I_d)\| \geq m\epsilon] \\ = & \Pr[\|\frac{1}{m}\sum_{i=1}^m Z_i - I_d\| \geq \epsilon] \\ = & \Pr[\|B - I_d\| \geq \epsilon]\end{aligned}$$

By choosing $t = m\epsilon$ and $m = \Omega(\epsilon^{-2} d \log(d/\delta))$ we have

$$\Pr[\|B - I_d\| \geq \epsilon] \leq \delta / \operatorname{poly}(d).$$

where B can be either B_x or B_y.

Similarly, we can prove

$$\begin{aligned}&\Pr[\|G_x\| \leq \epsilon] \geq 1 - \delta, \\ &\Pr[\|G_y\| \leq \epsilon] \geq 1 - \delta.\end{aligned}$$

□

E ONE SHRINKING STEP

In this section, we provide a shirking step for our result. In Section E.1 we define the matrices B, C, D, S to be used in analysis. In Section E.2 we upper bound the norm of $BD - C$. In Section E.3 we show the update term V_{t+1} can be written in a different way. In Section E.4 and Section E.5 we upper bounded $\|F\|$ and $\|R^{-1}\|$ respectively.

E.1 DEFINITIONS OF B, C, D, S

Definition E.1. *For each $p \in [k]$, let $u_{*,p} \in \mathbb{R}^n$ denotes the p-th column of matrix $U_* \in \mathbb{R}^{n \times k}$.*

For each $p \in [k]$, let $u_{t,p}$ denote the p-th column of matrix $U_t \in \mathbb{R}^{n \times k}$.

We define block matrices $B, C, D, S \in \mathbb{R}^{kd \times kd}$ as follows: For each $(p, q) \in [k] \times [k]$

- *Let $B_{p,q} \in \mathbb{R}^{d \times d}$ denote the (p, q)-th block of B*

$$B_{p,q} = \sum_{i=1}^m \underbrace{y_i y_i^\top}_{d \times d \text{ matrix}} \cdot \underbrace{(x_i^\top u_{t,p})}_{\text{scalar}} \cdot \underbrace{(x_i^\top u_{t,q})}_{\text{scalar}}$$

- *Let $C_{p,q} \in \mathbb{R}^{d \times d}$ denote the (p, q)-th block of C,*

$$C_{p,q} = \sum_{i=1}^m \underbrace{y_i y_i^\top}_{d \times d \text{ matrix}} \cdot \underbrace{(x_i^\top u_{t,p})}_{\text{scalar}} \cdot \underbrace{(x_i^\top u_{*q})}_{\text{scalar}}$$

- *Let $D_{p,q} \in \mathbb{R}^{d \times d}$ denote the (p, q)-th block of D,*

$$D_{p,q} = u_{t,p}^\top u_{*q} I$$

- *Let $S_{p,q} \in \mathbb{R}^{d\times d}$ denote the (p,q)-th block of S,*

$$S_{p,q} = \begin{cases} \sigma_p^* I, & \text{if } p = q; \\ 0, & \text{if } p \neq q. \end{cases}$$

  Here $\sigma_1^*, \cdots \sigma_k^*$ are singular values of $W_* \in \mathbb{R}^{d\times d}$.

- *We define $F \in \mathbb{R}^{d\times k}$ as follows*

$$\underbrace{\mathrm{vec}(F)}_{d\times 1} := \underbrace{B^{-1}}_{d\times d} \underbrace{(BD - C)}_{d\times d} \underbrace{S}_{d\times d} \cdot \underbrace{\mathrm{vec}(V_*)}_{d\times 1}.$$

E.2 Upper Bound on $\|BD - C\|$

Claim E.2. *Let B, C and D be defined as Definition E.1. Then we have*

$$\|BD - C\| \leq \epsilon \cdot \mathrm{dist}(U, U_*) \cdot k$$

Proof. Let $z_1, \cdots, z_k \in \mathbb{R}^d$ denote k vectors. Let $z = \begin{bmatrix} z_1 \\ \vdots \\ z_k \end{bmatrix}$.

We define $f(z) := z^\top (BD - C) z$

We define $f(z, p, q) = z_p^\top (BD - C)_{p,q} z_q$.

Then we can rewrite

$$\begin{aligned} z^\top (BD - C) z &= \sum_{p=1}^k \sum_{q=1}^k z_p^\top (BD - C)_{p,q} z_q \\ &= \sum_{p=1}^k \sum_{q=1}^k z_p^\top (B_{p,:} D_{:,q} - C_{p,q}) z_q \\ &= \sum_{p=1}^k \sum_{q=1}^k z_p^\top (\sum_{l=1}^k B_{p,l} D_{l,q} - C_{p,q}) z_q \end{aligned}$$

By definition, we know

$$\begin{aligned} B_{p,l} &= \sum_{i=1}^m y_i y_i^\top (x_i^\top u_{t,p}) \cdot (u_{t,l}^\top x_i) \\ D_{l,q} &= (u_{*,q}^\top u_{t,l}) I_d \\ C_{p,q} &= \sum_{i=1}^m y_i y_i^\top (x_i^\top u_{t,p}) \cdot (u_{*,q}^\top x_i) \end{aligned}$$

We can rewrite $C_{p,q}$ as follows

$$C_{p,q} = \sum_{i=1}^m y_i y_i^\top \cdot (x_i^\top u_{t,p}) \cdot (u_{*,q}^\top I_d x_i) \tag{14}$$

Let us compute

$$\begin{aligned} B_{p,l} D_{l,q} &= \sum_{i=1}^m y_i y_i^\top (x_i^\top u_{t,p}) \cdot (u_{t,l}^\top x_i) \cdot (u_{*,q}^\top u_{t,l}) \\ &= \sum_{i=1}^m y_i y_i^\top (x_i^\top u_{t,p}) \cdot (u_{*,q}^\top u_{t,l}) \cdot (u_{t,l}^\top x_i) \end{aligned}$$

where the second step follows from $a \cdot b = b \cdot a$ for any two scalars.

Taking the summation over all $l \in [k]$, we have

$$\begin{aligned}
\sum_{l=1}^{k} B_{p,l} D_{l,q} &= \sum_{l=1}^{k} \sum_{i=1}^{m} y_i y_i^\top (x_i^\top u_{t,p}) \cdot (u_{*,q}^\top u_{t,l}) \cdot (u_{t,l}^\top x_i) \\
&= \sum_{i=1}^{m} y_i y_i^\top (x_i^\top u_{t,p}) \cdot u_{*,q}^\top \sum_{l=1}^{k} (u_{t,l} \cdot u_{t,l}^\top) x_i \\
&= \sum_{i=1}^{m} \underbrace{y_i y_i^\top}_{\text{matrix}} \cdot \underbrace{(x_i^\top u_{t,p})}_{\text{scalar}} \cdot \underbrace{u_{*,q}^\top U_t U_t^\top x_i}_{\text{scalar}}
\end{aligned} \tag{15}$$

where first step follows from definition of B and D.

Then, we have

$$\begin{aligned}
\sum_{l=1}^{k} B_{p,l} D_{l,q} - C_{p,q} &= (\sum_{i=1}^{m} \underbrace{y_i y_i^\top}_{\text{matrix}} \cdot \underbrace{(x_i^\top u_{t,p})}_{\text{scalar}} \cdot \underbrace{u_{*,q}^\top U_t U_t^\top x_i}_{\text{scalar}}) - C_{p,q} \\
&= (\sum_{i=1}^{m} \underbrace{y_i y_i^\top}_{\text{matrix}} \cdot \underbrace{(x_i^\top u_{t,p})}_{\text{scalar}} \cdot \underbrace{u_{*,q}^\top U_t U_t^\top x_i}_{\text{scalar}}) - (\sum_{i=1}^{m} y_i y_i^\top \cdot (x_i^\top u_{t,p}) \cdot (u_{*,q}^\top I_d x_i)) \\
&= \sum_{i=1}^{m} \underbrace{y_i y_i^\top}_{\text{matrix}} \cdot \underbrace{(x_i^\top u_{t,p})}_{\text{scalar}} \cdot \underbrace{u_{*,q}^\top (U_t U_t^\top - I_d) x_i}_{\text{scalar}}
\end{aligned}$$

where the first step follows from Eq. (15), the second step follows from Eq. (14), the last step follows from merging the terms to obtain $(U_t U_t^\top - I_d)$.

Thus,

$$\begin{aligned}
f(z, p, q) &= z_p^\top (\sum_{l=1}^{k} B_{p,l} D_{l,q} - C_{p,q}) z_q \\
&= \sum_{i=1}^{m} \underbrace{(z_p^\top y_i)}_{\text{scalar}} \underbrace{(y_i^\top z_q)}_{\text{scalar}} \cdot \underbrace{(x_i^\top u_{t,p})}_{\text{scalar}} \cdot \underbrace{u_{*,q}^\top (U_t U_t^\top - I_d) x_i}_{\text{scalar}}
\end{aligned}$$

For easy of analysis, we define $v_t := u_{*,q}^\top (U_t U_t^\top - I_d)$. This means v_t lies in the complement of span of U_t.

Then

$$\begin{aligned}
\|v_t\|_2 &= \|u_{*,q}^\top (U_t U_t^\top - I_d)\|_2 \\
&= \|e_q^\top U_*^\top (U_t U_t^\top - I_d)\| \\
&\leq \|U_*^\top (U_t U_t^\top - I_d)\| \\
&= \mathrm{dist}(U_*, U_t).
\end{aligned} \tag{16}$$

where the second step follows from $u_{*,q}^\top = e_q^\top U_*^\top$ ($e_q \in \mathbb{R}^k$ is the vector q-th location is 1 and all other locations are 0s), third step follows from Fact B.1.

We want to apply Definition 3.6, but the issue is z_p, z_q and v_t are not unit vectors. So normalize them. Let $\overline{z}_p = z_p / \|z_p\|_2$, $\overline{z}_q = z_q / \|z_q\|_2$ and $\overline{v}_t = v_t / \|v_t\|_2$.

In order to apply for Definition 3.6, we also need $v_t^\top u_{t,p} = 0$.

Since $u_{t,p}$ is one of the column of U_t, thus $u_{t,p}$ lies in the span of U_t.

This is obvious true, since v_t lies in the complement of span of U_t and $u_{t,p}$ in the span of U_t.

We define

$$G := \sum_{i=1}^{m} \underbrace{(x_i^\top u_{t,p})}_{\text{scalar}} \cdot \underbrace{(x_i^\top \overline{v}_t)}_{\text{scalar}} \cdot \underbrace{y_i y_i^\top}_{\text{matrix}}$$

By Definition 3.6, we know that

$$\|G\| \leq \epsilon.$$

By definition of spectral norm, we have for any unit vector $\overline{z}_p$ and $\overline{z}_q$, we know that

$$|\overline{z}_p^\top G \overline{z}_q| \leq \|G\| \leq \epsilon.$$

where the first step follows from definition of spectral norm (Fact B.1), and the last step follows from Definition 3.6.

Note that

$$\begin{aligned} f(p,q,z) = & \sum_{i=1}^{m} \underbrace{(x_i^\top u_{t,p}) \cdot (x_i^\top \overline{v}_t)}_{\text{scalar}} \cdot \underbrace{(\overline{z}_p^\top y_i) \cdot (y_i^\top \overline{z}_q)}_{\text{scalar}} \cdot \underbrace{\|z_p\|_2 \cdot \|z_q\|_2 \cdot \|v_t\|_2}_{\text{scalar}} \\ = & \underbrace{\overline{z}_p^\top}_{1\times d} \cdot \Big(\sum_{i=1}^{m} \underbrace{(x_i^\top u_{t,p}) \cdot (x_i^\top \overline{v}_t)}_{\text{scalar}} \cdot \underbrace{y_i y_i^\top}_{d\times d} \Big) \cdot \underbrace{\overline{z}_q}_{d\times 1} \cdot \underbrace{\|z_p\|_2 \cdot \|z_q\|_2 \cdot \|v_t\|_2}_{\text{scalar}} \\ = & \underbrace{\overline{z}_p^\top}_{1\times d} \cdot \underbrace{G}_{d\times d} \cdot \underbrace{\overline{z}_q}_{d\times 1} \cdot \underbrace{\|z_p\|_2 \cdot \|z_q\|_2 \cdot \|v_t\|_2}_{\text{scalar}} \end{aligned}$$

where the second step follows from rewrite the second scalar $(\overline{z}_p^\top y_i)(y_i^\top \overline{z}_q) = \overline{z}_p^\top (y_i y_i^\top) \overline{z}_q$, the last step follows from definition of G.

Then,

$$\begin{aligned} |f(z,p,q)| = & |\sum_{i=1}^{m} \overline{z}_p^\top G \overline{z}_q| \cdot \|z_p\|_2 \|z_q\|_2 \|v_t\|_2 \\ \leq & \epsilon \|z_p\|_2 \|z_q\|_2 \cdot \|v_t\|_2 \\ \leq & \epsilon \|z_p\|_2 \|z_q\|_2 \cdot \mathrm{dist}(U_t, U_*) \end{aligned}$$

where the last step follows from Eq. (16).

Finally, we have

$$\begin{aligned} \|BD - C\| = & \max_{z, \|z\|_2 = 1} |z^\top (BD - C) z| \\ = & \max_{z, \|z\|_2 = 1} \Big| \sum_{p \in [k], q \in [k]} f(z,p,q) \Big| \\ \leq & \max_{z, \|z\|_2 = 1} \sum_{p \in [k], q \in [k]} |f(z,p,q)| \\ \leq & \epsilon \cdot \mathrm{dist}(U_t, U_*) \max_{z, \|z\|_2 = 1} \sum_{p \in [k], q \in [k]} \|z_p\|_2 \|z_q\|_2 \\ \leq & \epsilon \cdot \mathrm{dist}(U, U_*) \cdot k \end{aligned} \tag{17}$$

where the first step follows from Fact B.1, the last step step follows from $\sum_{p=1}^{k} \|z_p\|_2 \leq \sqrt{k} (\sum_{p=1}^{k} \|z_p\|_2^2)^{1/2} = \sqrt{k}$.

□

E.3 REWRITE V_{t+1}

Claim E.3. *If*

$$V_{t+1} = (W_*^\top U_t - F) R^{-1}$$

then,

$$(V_{*,\perp})^\top V_{t+1} = -(V_{*,\perp})^\top F R^{-1}$$

Proof. Multiplying both sides by $V_{*,\perp} \in \mathbb{R}^{d\times(d-k)}$:

$$\begin{aligned} V_{t+1} &= (W_*^\top U_t - F)R^{-1} \\ (V_{*,\perp})^\top V_{t+1} &= (V_{*,\perp})^\top (W_*^\top U_t - F)R^{-1} \\ (V_{*,\perp})^\top V_{t+1} &= (V_{*,\perp})^\top W_*^\top R^{-1} - (V_{*,\perp})^\top F R^{-1} \end{aligned}$$

We just need to show $(V_{*,\perp})^\top W_*^\top R^{-1} = 0$.

By definition of $V_{*,\perp}$, we know:

$$V_{*,\perp}^\top V_* = \mathbf{0}_{k\times(n-k)}$$

Thus, we have:

$$\begin{aligned} (V_{*,\perp})^\top W_*^\top &= V_{*,\perp}^\top V_* \Sigma_* U_*^\top \\ &= 0 \end{aligned}$$

□

E.4 Upper bound on $\|F\|$

Lemma E.4 (A variation of Lemma 2 in Zhong et al. (2015))**.** *Let $\mathcal{A}$ be a rank-one measurement operator where $A_i = x_i u_i^\top$. Let κ be defined as Definition 3.2.*

Then, we have

$$\|F\| \leq 2\epsilon k^{1.5} \cdot \sigma_1^* \cdot \mathrm{dist}(U_t, U_*)$$

Further, if $\epsilon \leq 0.001/(k^{1.5}\kappa)$

$$\|F\| \leq 0.01 \cdot \sigma_k^* \cdot \mathrm{dist}(U_t, U_*).$$

Proof. Recall that

$$\mathrm{vec}(F) = B^{-1}(BD - C)S \cdot \mathrm{vec}(V_*).$$

Here, we can upper bound $\|F\|$ as follows

$$\begin{aligned} \|F\| &\leq \|F\|_F \\ &= \|\mathrm{vec}(F)\|_2 \\ &\leq \|B^{-1}\| \cdot \|BD - C\| \cdot \|S\| \cdot \|\mathrm{vec}(V_*)\|_2 \\ &= \|B^{-1}\| \cdot \|(BD - C)\| \cdot \|S\| \cdot \sqrt{k} \\ &\leq \|B^{-1}\| \cdot \|(BD - C)\| \cdot \sigma_1^* \cdot \sqrt{k} \end{aligned} \tag{18}$$

where the first step follows from $\|\cdot\| \leq \|\cdot\|_F$ (Fact B.1), the second step follows vectorization of F is a vector, the third step follows from $\|Ax\|_2 \leq \|A\| \cdot \|x\|_2$, the forth step follows from $\|\mathrm{vec}(V_*)\|_2 = \|V_*\|_F \leq \sqrt{k}$ (Fact B.1) and the last step follows from $\|S\| \leq \sigma_1^*$ (see Definition E.1).

Now, we first bound $\|B^{-1}\| = 1/(\sigma_{\min}(B))$ (Fact B.1). Let $Z = [z_1 \quad z_2 \quad \cdots \quad z_k]$ and let $z = \mathrm{vec}(Z)$. Note that $B_{p,q}$ denotes the (p,q)-th block of B.

Also, we define

$$\mathcal{B} := \{x \in \mathbb{R}^{kd} \mid \|x\|_2 = 1\}.$$

Then

$$\begin{aligned}\sigma_{\min}(B) = & \min_{z \in \mathcal{B}} z^\top B z \\ = & \min_{z \in \mathcal{B}} \sum_{p \in [k], q \in [k]} z_p^\top B_{pq} z_q \\ = & \min_{z \in \mathcal{B}} \sum_{p=1}^{k} z_p^\top B_{p,p} z_p + \sum_{p \neq q} z_p^\top B_{p,q} z_q. \end{aligned} \tag{19}$$

where the first step follows from Fact B.1, the second step follows from simple algebra, the last step follows from (Fact B.1).

We can lower bound $z_p^\top B_{p,p} z_p$ as follows

$$\begin{aligned} z_p^\top B_{p,p} z_p \geq & \sigma_{\min}(B_{p,p}) \cdot \|z_p\|_2^2 \\ \geq & (1-\epsilon) \cdot \|z_p\|_2^2 \end{aligned} \tag{20}$$

where the first step follows from Fact B.1 , the last step follows from Definition 3.5 .

We can upper bound $|z^\top B_{p,q} z_q|$ as follows,

$$\begin{aligned} |z_p^\top B_{p,q} z_q| \leq & \|z_p\|_2 \cdot \|B_{p,q}\| \cdot \|z_q\|_2 \\ \leq & \epsilon \cdot \|z_p\|_2 \cdot \|z_q\|_2 \end{aligned} \tag{21}$$

where the first step follows from Fact B.1, the last step follows from Definition 3.5 .

We have

$$\begin{aligned} \sigma_{\min}(B) = & \min_{z, \|z\|_2 = 1} \sum_{p=1}^{k} z_p^\top B_{p,p} z_p + \sum_{p \neq q} z_p^\top B_{p,q} z_q \\ \geq & \min_{z, \|z\|_2 = 1} (1-\epsilon) \sum_{p=1}^{k} \|z_p\|_2^2 + \sum_{p \neq q} z_p^\top B_{p,q} z_q \\ \geq & \min_{z, \|z\|_2 = 1} (1-\epsilon) \sum_{p=1}^{k} \|z_p\|_2^2 - \epsilon \sum_{p \neq q} \|z_p\|_2 \|z_q\|_2 \\ = & \min_{z, \|z\|_2 = 1} (1-\epsilon) - \epsilon \sum_{p \neq q} \|z_p\|_2 \|z_q\|_2 \\ = & \min_{z, \|z\|_2 = 1} (1-\epsilon) - k\epsilon \\ \geq & 1 - 2k\epsilon \\ \geq & 1/2 \end{aligned} \tag{22}$$

where the first step follows from Eq. (19), the second step follows from Eq. (20), the third step follows from Eq. (21), the forth step follows from $\sum_{p=1}^{k} \|z_p\|_2^2 = 1$(which derived from the $\|z\|_2 = 1$ constraint and the definition of $\|z\|_2$), the fifth step follows from $\sum_{p \neq q} \|z_p\|_2 \|z_q\|_2 \leq (\sum_p \|z_p\|_2)^2 \leq (\sum_p \|z_p\|_1)^2 = (\|z\|_1)^2 \leq (\sqrt{k}\|z\|_2)^2 \leq k$, and the last step follows from $\epsilon \leq 0.1/k$.

We can show that

$$\|B^{-1}\| = \sigma_{\min}(B) \leq 2. \tag{23}$$

where the first step follows from Fact B.1, the second step follows from Eq. (22).

Now, consider $BD - C$, using Claim E.2, we have

$$\|BD - C\| \leq k \cdot \epsilon \cdot \mathrm{dist}(U_t, U_*)$$

Now, we have

$$\begin{aligned}\|F\| &\leq \|B^{-1}\| \cdot \|(BD - C)\| \cdot \sigma_1^* \cdot \sqrt{k} \\ &\leq 2 \cdot \|(BD - C)\| \cdot \sigma_1^* \cdot \sqrt{k} \\ &\leq 2 \cdot k \cdot \epsilon \cdot \mathrm{dist}(U_t, U_*) \cdot \sigma_1^* \cdot \sqrt{k}\end{aligned}$$

where the first step follows from Eq .(18), the second step follows from Eq. (23), and the third step follows from Eq. (17). □

E.5 Upper bound on $\|R^{-1}\|$

Lemma E.5 (A variation of Lemma 3 in Zhong et al. (2015))**.** *Let $\mathcal{A}$ be a rank-one measurement operator matrix where $A_i = x_i y_i^\top$. Also, let $\mathcal{A}$ satisfy three properties mentioned in Theorem 3.7.*

If the following condition holds

- $\mathrm{dist}(U_t, U_*) \leq \frac{1}{4} \leq \epsilon_d = 1/10$ *(The condition of Part 1 of Lemma C.1)*

Then,

$$\|R^{-1}\| \leq 10/\sigma_{k^*}$$

Proof. For simplicity, in the following proof, we use V to denote V_{t+1}. We use U to denote U_t.

Using Fact B.1

$$\|R^{-1}\| = \sigma_{\min}(R)^{-1}$$

We can lower bound $\sigma_{\min}(R)$ as follows:

$$\begin{aligned}\sigma_{\min}(R) &= \min_{z, \|z\|_2 = 1} \|Rz\|_2 \\ &= \min_{z, \|z\|_2 = 1} \|VRz\|_2 \\ &= \min_{z, \|z\|_2 = 1} \|V_* \Sigma_* U_*^\top U z - F z\|_2 \\ &\geq \min_{z, \|z\|_2 = 1} \|V_* \Sigma_* U_*^\top U z\|_2 - \|F z\|_2 \\ &\geq \min_{z, \|z\|_2 = 1} \|V_* \Sigma_* U_*^\top U z\|_2 - \|F\| \qquad (24)\end{aligned}$$

where the first step follows from definition of $\sigma_{\min}$, the second step follows from Fact B.1, the third step follows from $V = (W_*^\top U - F)R^{-1} = (V_* \Sigma_* U_*^\top U - F)R^{-1}$ (due to Eq. (7) and Definition 3.1) , the forth step follows from triangle inequality, the fifth step follows from $\|Ax\|_2 \leq \|A\|$ for all $\|x\|_2 = 1$.

Next, we can show that

$$\begin{aligned}\min_{z, \|z\|_2 = 1} \|V_* \Sigma_* U_*^\top U z\|_2 &= \min_{z, \|z\|_2 = 1} \|\Sigma_* U_*^\top U z\|_2 \\ &\geq \min_{z, \|z\|_2 = 1} \sigma_k^* \cdot \|U_*^\top U z\|_2 \\ &= \sigma_k^* \cdot \sigma_{\min}(U^\top U_*)\end{aligned}$$

where the first step follows from Fact B.1, the second step follows from Fact B.1, the third step follows from definition of $\sigma_{\min}$,

Next, we have

$$\begin{aligned}\sigma_{\min}(U^\top U_*) &= \cos \theta(U_*, U) \\ &= \sqrt{1 - \sin^2 \theta(U_*, U)} \\ &\geq \sqrt{1 - \mathrm{dist}(U_*, U)^2}\end{aligned}$$

where the first step follows definition of $\cos$, the second step follows from $\sin^2\theta + \cos^2\theta = 1$ (Lemma 2.8), the third step follows from $\sin \leq \mathrm{dist}$ (see Definition 2.6).

Putting it all together, we have

$$\begin{aligned}\sigma_{\min}(R) &\geq \sigma_k^* \sqrt{1 - \mathrm{dist}(U_*, U)^2} - \|F\| \\ &\geq \sigma_k^* \sqrt{1 - \mathrm{dist}(U_*, U)^2} - 0.001\sigma_k^* \,\mathrm{dist}(U_*, U) \\ &= \sigma_k^* (\sqrt{1 - \mathrm{dist}(U_*, U)^2} - 0.001\,\mathrm{dist}(U_*, U)) \\ &\geq 0.2\sigma_k^*\end{aligned}$$

where the second step follows from Lemma E.4, the last step follows from $\mathrm{dist}(U_*, U) < 1/10$.

□

F MATRIX SENSING REGRESSION

Our algorithm has $O(\log(1/\epsilon_0))$ iterations, in previous section we have proved why is that number of iterations sufficient. In order to show the final running time, we still need to provide a bound for the time we spend in each iteration. In this section, we prove a bound for cost per iteration. In Section F.1 we provide a basic claim that, our sensing problem is equivalent to some regression problem. In Section F.2 we show the different running time of the two implementation of each iteration. In Section F.3 we provide the time analysis for each of the iteration of our solver. In Section F.4 shows the complexity for the straightforward solver. Finally in Section F.5 we show the bound for the condition number.

F.1 DEFINITION AND EQUIVALENCE

In matrix sensing, we need to solve the following problem per iteration:

Definition F.1. *Let $A_1, \ldots, A_m \in \mathbb{R}^{d\times d}$, $U \in \mathbb{R}^{d\times k}$ and $b \in \mathbb{R}^m$ be given. The goal is to solve the following minimization problem*

$$\min_{V\in\mathbb{R}^{d\times k}} \sum_{i=1}^m (\mathrm{tr}[A_i^\top U V^\top] - b_i)^2,$$

We define another regression problem

Definition F.2. *Let $A_1, \ldots, A_m \in \mathbb{R}^{d\times d}$, $U \in \mathbb{R}^{d\times k}$ and $b \in \mathbb{R}^m$ be given.*

We define matrix $M \in \mathbb{R}^{m\times dk}$ as follows

$$M_{i,*} := \mathrm{vec}(U^\top A_i), \quad \forall i \in [m].$$

The goal is to solve the following minimization problem.

$$\min_{v\in\mathbb{R}^{dk}} \|Mv - b\|_2^2,$$

We can prove the following equivalence result

Lemma F.3 (Zhang (2023))**.** *Let $A_1, \ldots, A_m \in \mathbb{R}^{d\times d}$, $U \in \mathbb{R}^{d\times k}$ and $b \in \mathbb{R}^m$ be given.*

If the following conditions hold

- $M_{i,*} := \mathrm{vec}(U^\top A_i), \quad \forall i \in [m]$.
- *The solution matrix $V \in \mathbb{R}^{d\times k}$ can be reshaped through vector $v \in \mathbb{R}^{dk}$, i.e., $v = \mathrm{vec}(V^\top)$.*

Then, the problem (defined in Definition F.1) is equivalent to problem (defined in Definition F.2) .

Proof. Let $X, Y \in \mathbb{R}^{d \times d}$, we want to show that

$$\mathrm{tr}[X^\top Y] = \mathrm{vec}(X)^\top \mathrm{vec}(Y). \tag{25}$$

Note that the RHS is essentially $\sum_{i \in [d]} \sum_{j \in [d]} X_{i,j} Y_{i,j}$, for the LHS, note that

$$(X^\top Y)_{j,j} = \sum_{i \in [d]} X_{i,j} Y_{i,j},$$

the trace is then sum over j.

Thus, we have Eq. (25). This means that for each $i \in [d]$,

$$\mathrm{tr}[A_i^\top U V^\top] = \mathrm{vec}(U^\top A_i)^\top \mathrm{vec}(V^\top).$$

Set $M \in \mathbb{R}^{m \times dk}$ be the matrix where each row is $\mathrm{vec}(U^\top A_i)$, we see Definition F.1 is equivalent to solve the regression problem as in the statement. This completes the proof. □

F.2 From Sensing Matrix to Regression Matrix

Definition F.4. *Let $A_1, \ldots, A_m \in \mathbb{R}^{d \times d}$, $U \in \mathbb{R}^{d \times k}$. We define matrix $M \in \mathbb{R}^{m \times dk}$ as follows*

$$M_{i,*} := \mathrm{vec}(U^\top A_i), \quad \forall i \in [m].$$

Claim F.5. *The naive implementation of computing $M \in \mathbb{R}^{m \times dk}$ takes $m \cdot \mathcal{T}_{\mathrm{mat}}(k, d, d)$ time. Without using fast matrix multiplication, it is $O(md^2k)$ time.*

Proof. For each $i \in [m]$, computing matrix $U^\top \in \mathbb{R}^{k \times d}$ times $A_i \in \mathbb{R}^{d \times d}$ takes $\mathcal{T}_{\mathrm{mat}}(k, d, d)$ time. Thus, we complete the proof. □

Claim F.6. *The batch implementation takes $\mathcal{T}_{\mathrm{mat}}(k, dm, d)$ time. Without using fast matrix multiplication, it takes $O(md^2k)$ time.*

Proof. We can stack all the A_i together, then we matrix multiplication. For example, we construct matrix $A \in \mathbb{R}^{d \times dm}$. Then computing $U^\top A$ takes $\mathcal{T}_{\mathrm{mat}}(k, d, dm)$ time. □

The above two approach only has difference when we use fast matrix multiplication.

F.3 Our Fast Regression Solver

In this section, we provide the results of our fast regression solver. Our approach is basically as in Gu et al. (2023). For detailed analysis, we refer the readers to the Section 5 in Gu et al. (2023).

Lemma F.7 (Main Cost Per Iteration)**.** *Assume $m = \Omega(dk)$. There is an algorithm that runs in time*

$$\widetilde{O}(md^2k + d^3k^3)$$

and outputs a v' such that

$$\|Mv' - b\|_2 \leq (1 + \epsilon) \min_{v \in \mathbb{R}^{dk}} \|Mv - b\|_2$$

Proof. From Claim F.6, writing down $M \in \mathbb{R}^{m \times dk}$ takes $O(md^2k)$ time.

Using Fast regression resolver as Gu et al. (2023), the fast regression solver takes

$$O((m \cdot dk + (dk)^3) \cdot \log(\kappa(M)/\epsilon) \cdot \log^2(n/\delta))$$

□

Lemma F.8 (Formal version of Theorem 1.1)**.** *In each iteration, our requires takes $\widetilde{O}(md^2k)$ time.*

Proof. Finally, in order to run Lemma F.7, we need to argue that $\kappa(M) \leq \mathrm{poly}(k, d, \kappa(W_*))$.

This is true because $\kappa(U) \leq O(\kappa(W_*))$ and condition number of random Gaussian matrices is bounded by $\mathrm{poly}(k, d)$.

Then applying Lemma F.10, we can bound $\kappa(M)$ in each iteration.

Eventually, we just run standard error analysis in Gu et al. (2023). Thus, we should get the desired speedup.

The reason we can drop the $(dk)^3$ is $m \geq dk^2$. □

F.4 Straightforward Solver

Note that from sample complexity analysis, we know that $m = \Omega(dk)$.

Lemma F.9. *Assume $m = \Omega(dk)$. The straightforward implementation of the regression problem (Defintion F.2) takes*

$$O(md^2k^2)$$

time.

Proof. The algorithm has two steps. From Claim F.6, writing down $M \in \mathbb{R}^{m \times dk}$ takes $O(md^2k)$ time.

The first step is writing down the matrix $M \in \mathbb{R}^{m \times dk}$.

The second step is solving regression, it needs to compute $M^\dagger b$ (where $M^\dagger \in \mathbb{R}^{dk \times m}$)

$$M^\dagger b = (M^\top M)^{-1} M b$$

this will take time

$$\begin{aligned} \mathcal{T}_{\mathrm{mat}}(dk, m, dk) + \mathcal{T}_{\mathrm{mat}}(dk, dk, dk) &= md^2k^2 + (dk)^3 \\ &= md^2k^2 \end{aligned}$$

the second step follows from $m = \Omega(dk)$.

Thus, the total time is

$$md^2k + md^2k^2 = O(md^2k^2)$$

□

F.5 Condition Number

Lemma F.10. *We define $B \in \mathbb{R}^{m \times k}$ as follows $B := XU$ and $X \in \mathbb{R}^{m \times d}$ and $U \in \mathbb{R}^{d \times k}$.*

Then, we can rewrite $M \in \mathbb{R}^{m \times dk}$

$$\underbrace{M}_{m \times dk} = \underbrace{B}_{m \times k} \otimes \underbrace{Y}_{m \times d}$$

Then, we know that $\kappa(M) = \kappa(B) \cdot \kappa(Y) \leq \kappa(U)\kappa(X)\kappa(Y)$.

Proof. Recall $U \in \mathbb{R}^{d \times k}$. Then we define $b_i = U^\top x_i$ for each $i \in [m]$.

Then we have

$$M_{i,*} = \mathrm{vec}(U^\top x_i y_i^\top) = \mathrm{vec}(b_i y_i^\top).$$

Thus, it implies

$$M = B \otimes Y$$

□

G LIMITATIONS

In the grand scheme, our contribution significantly propels the theoretical understanding of matrix sensing forward and inaugurates a novel approach for addressing this pivotal problem, intertwining theoretical depth with practical applicability. Notwithstanding, we have astutely identified certain limitations within our work. Specifically, our analysis operates under the assumption that the underlying low-rank matrix adheres to certain incoherence and restricted isometry property (RIP) conditions. While these conditions are both reasonable and prevalently utilized within the existing literature, it is conceivable that there exist matrices that do not conform to these conditions and yet remain of tangible practical interest. Consequently, it becomes a compelling avenue of exploration to ascertain whether our proposed algorithm can be adeptly adapted to accommodate matrices that diverge from these conditions, or alternatively, to forge new algorithms specifically tailored for such matrices.