Learning Mixtures of Low-Rank Models

Yanxi Chen* Cong Ma† H. Vincent Poor* Yuxin Chen*

September 23, 2020

Abstract

We study the problem of learning mixtures of low-rank models, i.e. reconstructing multiple low-rank matrices from unlabelled linear measurements of each. This problem enriches two widely studied settings — low-rank matrix sensing and mixed linear regression — by bringing latent variables (i.e. unknown labels) and structural priors (i.e. low-rank structures) into consideration. To cope with the non-convexity issues arising from unlabelled heterogeneous data and low-complexity structure, we develop a three-stage meta-algorithm that is guaranteed to recover the unknown matrices with near-optimal sample and computational complexities under Gaussian designs. In addition, the proposed algorithm is provably stable against random noise. We complement the theoretical studies with empirical evidence that confirms the efficacy of our algorithm.

Keywords: matrix sensing, latent variable models, heterogeneous data, mixed linear regression, non-convex optimization, meta-learning

Contents

1 Introduction **2**
- 1.1 Main contributions . . . 3
- 1.2 Notation . . . 3

2 Algorithm **4**
- 2.1 Stage 1: subspace estimation via a spectral method . . . 4
- 2.2 Stage 2: initialization via low-dimensional mixed linear regression . . . 4
- 2.3 Stage 3: local refinement via scaled truncated gradient descent (ScaledTGD) . . . 6
- 2.4 The full algorithm . . . 6

3 Main results **7**
- 3.1 Models and assumptions . . . 7
- 3.2 Theoretical guarantees . . . 8
- 3.3 Numerical experiments . . . 9

4 Prior work **10**

5 Analysis **11**

6 Discussion **12**

A The tensor method for mixed linear regression **13**

*Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA; email: {yanxic,poor,yuxin.chen}@princeton.edu.

†Department of Electrical Engineering and Computer Science, University of California Berkeley, Berkeley, CA 94720, USA; email: congm@berkeley.edu

B Proofs for Section 5 **14**
B.1 Proof of Theorem 3 . . . 14
B.2 Proof of Theorem 4 . . . 16
B.3 Proof of Proposition 1 . . . 19
B.4 Proof of Theorem 5 . . . 19

C Technical lemmas **24**
C.1 Proof of Lemma 1 . . . 26
C.2 Proof of Lemma 2 . . . 30
C.3 Proof of Lemma 3 . . . 32

D Estimating unknown parameters in Algorithm 4 **33**

1 Introduction

This paper explores a mixture of low-rank models with latent variables, which seeks to reconstruct a couple of low-rank matrices $\boldsymbol{M}_k^\star \in \mathbb{R}^{n_1 \times n_2}$ $(1 \le k \le K)$ from *unlabeled* linear measurements of each. More specifically, what we have available is a collection of N linear measurements $\{y_i\}_{1\le i\le N}$ taking the following form:

$$y_i = \begin{cases} \langle \boldsymbol{A}_i, \boldsymbol{M}_1^\star \rangle, & \text{if } i \in \Omega_1^\star, \\ \dots & \dots \\ \langle \boldsymbol{A}_i, \boldsymbol{M}_K^\star \rangle, & \text{if } i \in \Omega_K^\star, \end{cases} \tag{1}$$

where $\{\boldsymbol{A}_i\}_{1\le i\le N}$ are the sampling/design matrices, $\langle \cdot, \cdot \rangle$ denotes the matrix inner product, and $\{\Omega_k^\star\}_{1\le k\le K}$ represents an unknown partition of the index set $\{1, \dots, N\}$. The aim is to design an algorithm that is guaranteed to recover $\{\boldsymbol{M}_k^\star\}$ efficiently and faithfully, despite the absence of knowledge of $\{\Omega_k^\star\}_{1\le k\le K}$.

This problem of learning mixtures of low-rank models enriches two widely studied settings: (1) it generalizes classical low-rank matrix recovery [RFP10, CLC19] by incorporating heterogeneous data and latent variables (i.e. the labels indicating which low-rank matrices are being measured), and (2) it expands the studies of mixed linear regression [QR78, YCS14] by integrating low-complexity structural priors (i.e. low-rank structures). In addition to the prior work [YC15] that has studied this setting, we single out two broader scenarios that bear relevance to and motivate the investigation of mixtures of low-rank models.

- *Mixed matrix completion.* If each measurement y_i only reveals a single entry of one of the unknown matrices $\{\boldsymbol{M}_k^\star\}$, then the problem is commonly referred to as mixed matrix completion (namely, completing several low-rank matrices from a mixture of unlabeled observations of their entries) [PA18]. One motivating application arises from computer vision, where several problems like joint shape matching can be posed as structured matrix completion [CGH14, CC18a]. When the objects to be matched exhibit certain geometric symmetry, there might exist multiple plausible maps (and hence multiple ground-truth matrices), and the provided observations might become intrinsically unlabeled due to symmetric ambiguities [SLHH18]. Other applications include network topology inference and metagenomics given mixed DNA samples; see [PA18] for details.

- *Multi-task learning and meta-learning.* The model (1) can be viewed as an instance of multi-task learning or meta-learning [Bax00, MPRP16, KSS+20], where the tasks follow a discrete prior distribution supported on a set of K meta parameters, and each training data point $(\boldsymbol{A}_i, y_i)$ is a realization of one task that comes with a single sample. While it is typically assumed in meta-learning that even light tasks have more than one samples, understanding this single-sample model is essential towards tackling more general settings. Additionally, in comparison to meta-learning for mixed linear regression [KSS+20, KSKO20], the model (1) imposes further structural prior on the unknown meta parameters, thereby allowing for potential reduction of sample complexities.

The challenge for learning mixtures of low-rank models primarily stems from the non-convexity issues. While the low-rank structure alone already leads to non-convex optimization landscapes, the presence of heterogeneous data and discrete hidden variables further complicates matters significantly.

1.1 Main contributions

This paper takes a step towards learning mixtures of low-rank models, focusing on the tractable Gaussian design where the $\boldsymbol{A}_i$'s have i.i.d. Gaussian entries; in light of this, we shall also call the problem *mixed matrix sensing*, to be consistent with the terminology used in recent literature [BNS16, CLC19]. In particular, we propose a meta-algorithm comprising the following three stages:

1. Estimate the joint column and row spaces of $\{\boldsymbol{M}_k^\star\}_{1\le k\le K}$;
2. Transform mixed matrix sensing into low-dimensional mixed linear regression using the above subspace estimates, and invoke a mixed linear regression solver to obtain initial estimates of $\{\boldsymbol{M}_k^\star\}_{1\le k\le K}$;
3. Successively refine the estimates via a non-convex low-rank matrix factorization algorithm (more specifically, an algorithm called *scaled truncated gradient descent* to be described in Algorithm 3).

The details of each stage will be spelled out and elucidated in Section 2.

Encouragingly, the proposed algorithm is guaranteed to succeed under mild conditions (to be specified in Section 3.1). Informally, our contributions are three-fold.

- *Exact recovery in the noiseless case.* In the absence of noise, our algorithm enables exact recovery of $\{\boldsymbol{M}_k^\star\}$ modulo global permutation. The sample complexity required to achieve this scales linearly (up to some log factor) in the dimension $\max\{n_1, n_2\}$ and polynomially in other salient parameters.
- *Stability vis-à-vis random noise.* The proposed algorithm is provably stable against Gaussian noise, in the sense that the estimation accuracy degrades gracefully as the signal-to-noise-ratio decreases.
- *Computational efficiency.* When the number K of components and the maximum rank of the unknown matrices are both constants, the computational cost of our algorithm scales nearly linearly in Nn_1n_2 with N the number of samples — this is proportional to the time taken to read all design matrices.

The precise theorem statements are postponed to Section 3. Empirical evidence will also be provided in Section 3 to corroborate the efficacy of our algorithm.

1.2 Notation

Before we proceed, let us collect some notation that will be frequently used. Throughout this paper, we reserve boldfaced symbols for vectors (lower case) and matrices (upper case). For a vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|_2$ denotes its ℓ_2 norm. For a matrix $\boldsymbol{X}$, $\|\boldsymbol{X}\|$ (resp. $\|\boldsymbol{X}\|_{\mathrm{F}}$) denotes its spectral (resp. Frobenius) norm, $\sigma_k(\boldsymbol{X})$ denotes its k-th largest singular value, and $\mathsf{col}\{\boldsymbol{X}\}$ (resp. $\mathsf{row}\{\boldsymbol{X}\}$) denotes its column (resp. row) space. If $\boldsymbol{U}$ is a matrix with orthonormal columns, we also use the same notation $\boldsymbol{U}$ to represent its column space, and vice versa. For any matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n_1\times n_2}$, let $\langle \boldsymbol{A}, \boldsymbol{B}\rangle \coloneqq \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} A_{ij}B_{ij}$ stand for the matrix inner product. $\boldsymbol{I}_n$ represents the $n\times n$ identity matrix. $\mathsf{vec}(\cdot)$ denotes vectorization of a matrix, and $\mathsf{mat}(\cdot)$ denotes the inverse operation (the corresponding matrix dimensions should often be clear from the context).

We use both $a_n \lesssim b_n$ and $a_n = O(b_n)$ to indicate that $a_n \le C_0 b_n$ for some universal constant $C_0 > 0$; in addition, $a_n \gtrsim b_n$ is equivalent to $b_n \lesssim a_n$, and $a_n \asymp b_n$ means both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold true. Finally, $a_n = o(b_n)$ means that $a_n/b_n \to 0$ as $n \to \infty$.

For a finite set Ω, we denote by $|\Omega|$ its cardinality. For a number $\alpha \in [0,1]$ and a random variable X following some distribution on $\mathbb{R}$, we let $Q_\alpha(X)$ denote the α-quantile function, namely

$$Q_\alpha(X) \coloneqq \inf\big\{t \in \mathbb{R} : \mathbb{P}(X \le t) \ge \alpha\big\}. \tag{2}$$

For a finite set $\mathcal{D}$ of real numbers, with slight abuse of notation, we let $Q_\alpha(\mathcal{D})$ be the α-quantile of $\mathcal{D}$; more precisely, we define $Q_\alpha(\mathcal{D}) \coloneqq Q_\alpha(X_{\mathcal{D}})$, where $X_{\mathcal{D}}$ denotes a random variable uniformly drawn from $\mathcal{D}$.

2 Algorithm

This section formalizes our algorithm design by specifying each stage of our meta-algorithm with a concrete procedure (namely, Algorithms 1, 2, 3 for Stages 1, 2, 3, respectively). It is worth noting that these are definitely not the only choices; in fact, an advantage of our meta-algorithm is its flexibility and modularity, in the sense that one can plug in different sub-routines to address various models and assumptions.

Before continuing, we introduce more notation that will be used throughout. For any $1 \le k \le K$, define

$$p_k := \frac{|\Omega_k^\star|}{N} \qquad \text{and} \qquad r_k := \mathsf{rank}(\boldsymbol{M}_k^\star), \tag{3}$$

which represent the fraction of samples associated with the k-th component and the rank of the k-th ground-truth matrix $\boldsymbol{M}_k^\star$, respectively. In addition, let the compact singular value decomposition (SVD) of $\{\boldsymbol{M}_k^\star\}$ be

$$\boldsymbol{M}_k^\star = \boldsymbol{U}_k^\star \boldsymbol{\Sigma}_k^\star \boldsymbol{V}_k^{\star\top}, \qquad 1 \le k \le K, \tag{4}$$

where $\boldsymbol{U}_k^\star \in \mathbb{R}^{n_1 \times r_k}$ and $\boldsymbol{V}_k^\star \in \mathbb{R}^{n_2 \times r_k}$ consist of orthonormal columns, and $\boldsymbol{\Sigma}_k^\star$ is a diagonal matrix.

2.1 Stage 1: subspace estimation via a spectral method

Procedure. We propose to estimate the following joint column and row spaces:

$$\boldsymbol{U}^\star := \mathsf{col}\big\{[\boldsymbol{U}_1^\star, \ldots, \boldsymbol{U}_K^\star]\big\} \quad \text{and} \quad \boldsymbol{V}^\star := \mathsf{col}\big\{[\boldsymbol{V}_1^\star, \ldots, \boldsymbol{V}_K^\star]\big\} \tag{5}$$

by means of a spectral method. More specifically, we start by forming a data matrix

$$\boldsymbol{Y} := \frac{1}{N}\sum_{i=1}^N y_i \boldsymbol{A}_i, \tag{6}$$

and set $\boldsymbol{U} \in \mathbb{R}^{n_1 \times R}$ (resp. $\boldsymbol{V} \in \mathbb{R}^{n_2 \times R}$) to be a matrix whose columns consist of the top-R left (resp. right) singular vectors of $\boldsymbol{Y}$, where

$$R := \mathsf{rank}\big(\mathbb{E}[\boldsymbol{Y}]\big). \tag{7}$$

This method is summarized in Algorithm 1.

Rationale. To see why this might work, note that if $\{\boldsymbol{A}_i\}$ consist of i.i.d. standard Gaussian entries, then

$$\begin{aligned}\mathbb{E}[\boldsymbol{Y}] &= \sum_{k=1}^K p_k \mathbb{E}\big[\langle \boldsymbol{A}_i, \boldsymbol{M}_k^\star\rangle \boldsymbol{A}_i\big] = \sum_{k=1}^K p_k \boldsymbol{M}_k^\star = \sum_{k=1}^K p_k \boldsymbol{U}_k^\star \boldsymbol{\Sigma}_k^\star \boldsymbol{V}_k^{\star\top} \\ &= \big[\boldsymbol{U}_1^\star, \boldsymbol{U}_2^\star, \ldots, \boldsymbol{U}_K^\star\big] \begin{bmatrix} p_1\boldsymbol{\Sigma}_1^\star & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & p_2\boldsymbol{\Sigma}_2^\star & \boldsymbol{0} & \vdots \\ \vdots & \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \ldots & \boldsymbol{0} & p_K\boldsymbol{\Sigma}_K^\star \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_1^{\star\top} \\ \boldsymbol{V}_2^{\star\top} \\ \vdots \\ \boldsymbol{V}_K^{\star\top} \end{bmatrix}.\end{aligned} \tag{8}$$

Recalling the definitions of $\boldsymbol{U}^\star$ and $\boldsymbol{V}^\star$ in (5), we have

$$\mathsf{col}\big\{\mathbb{E}[\boldsymbol{Y}]\big\} = \boldsymbol{U}^\star, \quad \mathsf{row}\big\{\mathbb{E}[\boldsymbol{Y}]\big\} = \boldsymbol{V}^\star, \quad \mathsf{rank}(\boldsymbol{U}^\star) = \mathsf{rank}(\boldsymbol{V}^\star) = R$$

under some mild conditions (detailed in Section 3). This motivates the development of Algorithm 1.

2.2 Stage 2: initialization via low-dimensional mixed linear regression

Key observations. Suppose that there is an oracle informing us of the subspaces $\boldsymbol{U}^\star$ and $\boldsymbol{V}^\star$ defined in (5). Recognizing the basic relation $\boldsymbol{M}_k^\star = \boldsymbol{U}^\star \boldsymbol{U}^{\star\top} \boldsymbol{M}_k^\star \boldsymbol{V}^\star \boldsymbol{V}^{\star\top}$ and defining

$$\boldsymbol{S}_k^\star := \boldsymbol{U}^{\star\top} \boldsymbol{M}_k^\star \boldsymbol{V}^\star \in \mathbb{R}^{R\times R}, \qquad 1 \le k \le K, \tag{9}$$

Algorithm 1: Subspace estimation via a spectral method

Input: samples $\{\boldsymbol{A}_i, y_i\}_{1\le i\le N}$, rank R.
Compute $\boldsymbol{Y} \leftarrow \frac{1}{N}\sum_{i=1}^{N} y_i \boldsymbol{A}_i$.
Let $\boldsymbol{U} \in \mathbb{R}^{n_1\times R}$ (resp. $\boldsymbol{V} \in \mathbb{R}^{n_2\times R}$) be the matrix consisting of the top-R left (resp. right) singular vectors of $\boldsymbol{Y}$.
Output: $\boldsymbol{U}, \boldsymbol{V}$.

Algorithm 2: Initialization via low-dimensional mixed linear regression

Input: samples $\{\boldsymbol{A}_i, y_i\}_{1\le i\le N}$, subspaces $\boldsymbol{U}, \boldsymbol{V}$, ranks $\{r_k\}_{1\le k\le K}$.
Transform $\boldsymbol{a}_i \leftarrow \mathsf{vec}(\boldsymbol{U}^\top \boldsymbol{A}_i \boldsymbol{V}), 1 \le i \le N$.
Obtain $\{\widehat{\boldsymbol{\beta}}_k\}_{1\le k\le K} \leftarrow$ the output of a black-box mixed linear regression solver (i.e. Algorithm 5) on $\{\boldsymbol{a}_i, y_i\}_{1\le i\le N}$.
for $k = 1, \ldots, K$ **do**
  $\boldsymbol{U}_k \boldsymbol{\Sigma}_k \boldsymbol{V}_k^\top \leftarrow$ rank-r_k SVD of $\boldsymbol{U}\widehat{\boldsymbol{S}}_k\boldsymbol{V}^\top$, where $\widehat{\boldsymbol{S}}_k \coloneqq \mathsf{mat}(\widehat{\boldsymbol{\beta}}_k)$.
  $\boldsymbol{L}_k \leftarrow \boldsymbol{U}_k \boldsymbol{\Sigma}_k^{1/2}, \boldsymbol{R}_k \leftarrow \boldsymbol{V}_k \boldsymbol{\Sigma}_k^{1/2}$.
Output: $\{\boldsymbol{L}_k, \boldsymbol{R}_k\}_{1\le k\le K}$.

we can rewrite the measurements in hand as follows:

$$y_i = \begin{cases} \langle \boldsymbol{A}_i, \boldsymbol{M}_1^\star\rangle = \langle \boldsymbol{A}_i, \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\boldsymbol{M}_1^\star\boldsymbol{V}^\star\boldsymbol{V}^{\star\top}\rangle = \langle \boldsymbol{U}^{\star\top}\boldsymbol{A}_i\boldsymbol{V}^\star, \boldsymbol{S}_1^\star\rangle, & \text{if } i\in\Omega_1^\star, \\ \cdots & \cdots \\ \langle \boldsymbol{A}_i, \boldsymbol{M}_K^\star\rangle = \langle \boldsymbol{U}^{\star\top}\boldsymbol{A}_i\boldsymbol{V}^\star, \boldsymbol{S}_K^\star\rangle, & \text{if } i\in\Omega_K^\star. \end{cases} \tag{10}$$

In other words, the presence of the oracle effectively reduces the original problem into a mixed linear regression problem in lower dimensions — that is, the problem of recovering $\{\boldsymbol{S}_k^\star\}$ from mixed linear measurements. If $\{\boldsymbol{S}_k^\star\}$ can be reliably estimated, then one can hope to recover $\{\boldsymbol{M}_k^\star\}$ via the following relation:

$$\boldsymbol{M}_k^\star = \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\boldsymbol{M}_k^\star\boldsymbol{V}^\star\boldsymbol{V}^{\star\top} = \boldsymbol{U}^\star\boldsymbol{S}_k^\star\boldsymbol{V}^{\star\top}. \tag{11}$$

Procedure. While we certainly have no access to the aforementioned oracle in reality, Stage 1 described above provides us with subspace estimates $\boldsymbol{U}$ and $\boldsymbol{V}$ satisfying $\boldsymbol{U}\boldsymbol{U}^\top \approx \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}$ and $\boldsymbol{V}\boldsymbol{V}^\top \approx \boldsymbol{V}^\star\boldsymbol{V}^{\star\top}$. Treating these as surrogates of $(\boldsymbol{U}^\star, \boldsymbol{V}^\star)$ (so that $\boldsymbol{M}_k^\star \approx \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{M}_k^\star\boldsymbol{V}\boldsymbol{V}^\top$), we can view the measurements as

$$y_i = \begin{cases} \langle \boldsymbol{A}_i, \boldsymbol{M}_1^\star\rangle \approx \langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{M}_K^\star\boldsymbol{V}\boldsymbol{V}^\top\rangle = \langle \boldsymbol{U}^\top\boldsymbol{A}_i\boldsymbol{V}, \boldsymbol{S}_1\rangle = \langle \boldsymbol{a}_i, \boldsymbol{\beta}_1\rangle, & \text{if } i\in\Omega_1^\star, \\ \cdots & \cdots \\ \langle \boldsymbol{A}_i, \boldsymbol{M}_K^\star\rangle \approx \langle \boldsymbol{a}_i, \boldsymbol{\beta}_K\rangle, & \text{if } i\in\Omega_K^\star, \end{cases} \tag{12}$$

which are mixed linear measurements about the following vectors/matrices:

$$\boldsymbol{\beta}_k \coloneqq \mathsf{vec}\big(\boldsymbol{S}_k\big) \in \mathbb{R}^{R^2}, \qquad \boldsymbol{S}_k \coloneqq \boldsymbol{U}^\top\boldsymbol{M}_k^\star\boldsymbol{V} \in \mathbb{R}^{R\times R}, \qquad 1\le k\le K. \tag{13}$$

Here, the equivalent sensing vectors are defined to be $\boldsymbol{a}_i \coloneqq \mathsf{vec}\big(\boldsymbol{U}^\top\boldsymbol{A}_i\boldsymbol{V}\big) \in \mathbb{R}^{R^2}$ for any $1\le i\le N$. All this motivates us to resort to mixed linear regression algorithms for recovering $\{\boldsymbol{\beta}_k\}$. The proposed algorithm thus entails the following steps, with the precise procedure summarized in Algorithm 2.

- Invoke any mixed linear regression algorithm to obtain estimates $\{\widehat{\boldsymbol{\beta}}_k\}_{1\le k\le K}$ for $\{\boldsymbol{\beta}_k\}_{1\le k\le K}$ (up to global permutation). For concreteness, the current paper applies the tensor method (Algorithm 5) originally proposed in [YCS16]; this is a polynomial-time algorithm, with details deferred to Appendix A. To simplify presentation, let us assume here that the global permutation happens to be an identity map, so that $\widehat{\boldsymbol{\beta}}_k$ is indeed a faithful estimate of $\boldsymbol{\beta}_k$ $(1\le k\le K)$. By simple matricization, $\widehat{\boldsymbol{\beta}}_k$ leads to a reliable estimate $\widehat{\boldsymbol{S}}_k$ of $\boldsymbol{S}_k$.

Algorithm 3: Scaled Truncated Gradient Descent (ScaledTGD) for recovering $\boldsymbol{M}_k^\star$

Input: samples $\{\boldsymbol{A}_i, y_i\}_{1\le i\le N}$, initialization $\boldsymbol{L}^0 \in \mathbb{R}^{n_1\times r_k}, \boldsymbol{R}^0 \in \mathbb{R}^{n_2\times r_k}$, step size η, truncating fraction α.

for $t = 0, 1, 2, \ldots, T_0 - 1$ **do**

$$\boldsymbol{L}^{t+1} \leftarrow \boldsymbol{L}^t - \frac{\eta}{N}\sum_{i\in\Omega^t}\big(\langle \boldsymbol{A}_i, \boldsymbol{L}^t(\boldsymbol{R}^t)^\top\rangle - y_i\big)\boldsymbol{A}_i\boldsymbol{R}^t\big((\boldsymbol{R}^t)^\top\boldsymbol{R}^t\big)^{-1},$$

$$\boldsymbol{R}^{t+1} \leftarrow \boldsymbol{R}^t - \frac{\eta}{N}\sum_{i\in\Omega^t}\big(\langle \boldsymbol{A}_i, \boldsymbol{L}^t(\boldsymbol{R}^t)^\top\rangle - y_i\big)\boldsymbol{A}_i^\top\boldsymbol{L}^t\big((\boldsymbol{L}^t)^\top\boldsymbol{L}^t\big)^{-1},$$

where $\Omega^t \coloneqq \{1\le i\le N : |\langle \boldsymbol{A}_i, \boldsymbol{L}^t(\boldsymbol{R}^t)^\top\rangle - y_i| \le \tau_t\}$, $\tau_t \coloneqq Q_\alpha(\{|\langle \boldsymbol{A}_i, \boldsymbol{L}^t(\boldsymbol{R}^t)^\top\rangle - y_i|\}_{1\le i\le N})$.

Output: $\boldsymbol{L}^{T_0}, \boldsymbol{R}^{T_0}$.

- Given the observation that

$$\boldsymbol{U}\widehat{\boldsymbol{S}}_k\boldsymbol{V}^\top \approx \boldsymbol{U}\boldsymbol{S}_k\boldsymbol{V}^\top = \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{M}_k^\star\boldsymbol{V}\boldsymbol{V}^\top \approx \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\boldsymbol{M}_k^\star\boldsymbol{V}^\star\boldsymbol{V}^{\star\top} = \boldsymbol{M}_k^\star, \tag{14}$$

we propose to compute the rank-r_k SVD — denoted by $\boldsymbol{U}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k^\top$ — of the matrix $\boldsymbol{U}\widehat{\boldsymbol{S}}_k\boldsymbol{V}^\top$ for each $1\le k\le K$. This in turn leads to our initial estimate for the low-rank factors

$$\boldsymbol{L}_k \coloneqq \boldsymbol{U}_k\boldsymbol{\Sigma}_k^{1/2} \in \mathbb{R}^{n_1\times r_k}, \quad \text{and} \quad \boldsymbol{R}_k \coloneqq \boldsymbol{V}_k\boldsymbol{\Sigma}_k^{1/2} \in \mathbb{R}^{n_2\times r_k}. \tag{15}$$

2.3 Stage 3: local refinement via scaled truncated gradient descent (ScaledTGD)

Suppose that an initial point $\boldsymbol{L}^0(\boldsymbol{R}^0)^\top$ lies within a reasonably small neighborhood of $\boldsymbol{M}_k^\star$ for some $1\le k\le K$. Stage 3 serves to locally refine this initial estimate, moving it closer to our target $\boldsymbol{M}_k^\star$. Towards this end, we propose to deploy the following update rule termed *scaled truncated gradient descent* (ScaledTGD):

$$\boldsymbol{L}^{t+1} = \boldsymbol{L}^t - \frac{\eta}{N}\sum_{i\in\Omega^t}\big(\langle \boldsymbol{A}_i, \boldsymbol{L}^t(\boldsymbol{R}^t)^\top\rangle - y_i\big)\boldsymbol{A}_i\boldsymbol{R}^t\big((\boldsymbol{R}^t)^\top\boldsymbol{R}^t\big)^{-1}, \tag{16a}$$

$$\boldsymbol{R}^{t+1} = \boldsymbol{R}^t - \frac{\eta}{N}\sum_{i\in\Omega^t}\big(\langle \boldsymbol{A}_i, \boldsymbol{L}^t(\boldsymbol{R}^t)^\top\rangle - y_i\big)\boldsymbol{A}_i^\top\boldsymbol{L}^t\big((\boldsymbol{L}^t)^\top\boldsymbol{L}^t\big)^{-1}, \tag{16b}$$

where $\eta > 0$ denotes the step size. Here, $\Omega^t \subseteq \{1, 2, \cdots, N\}$ is an adaptive and iteration-varying index set designed to mimic the index set $\Omega_k^\star$. Indeed, if $\Omega^t = \Omega_k^\star$, the aforementioned update rule reduces to the ScaledGD method developed for vanilla low-rank matrix sensing (see [TMC20]), which is guaranteed to converge to $\boldsymbol{M}_k^\star$ in the presence of a suitable initialization. Here, the rescaling matrix $\big((\boldsymbol{R}^t)^\top\boldsymbol{R}^t\big)^{-1}$ (resp. $\big((\boldsymbol{L}^t)^\top\boldsymbol{L}^t\big)^{-1}$) acts as a pre-conditioner of the conventional gradient $\sum_{i\in\Omega^t}\big(\langle \boldsymbol{A}_i, \boldsymbol{L}^t(\boldsymbol{R}^t)^\top\rangle - y_i\big)\boldsymbol{A}_i\boldsymbol{R}^t$ (resp. $\sum_{i\in\Omega^t}\big(\langle \boldsymbol{A}_i, \boldsymbol{L}^t(\boldsymbol{R}^t)^\top\rangle - y_i\big)\boldsymbol{A}_i^\top\boldsymbol{L}^t$), which effectively accelerates convergence when $\boldsymbol{M}_k^\star$ is ill-conditioned. See [TMC20] for more intuitions and justifications of this rescaling strategy.

Viewed in this light, the key to ensuring effectiveness of ScaledTGD lies in the design of the index set Ω^t. If we know *a priori* that $\boldsymbol{L}^t(\boldsymbol{R}^t)^\top \approx \boldsymbol{M}_k^\star$, then it is intuitively clear that $|\langle \boldsymbol{A}_i, \boldsymbol{L}^t(\boldsymbol{R}^t)^\top\rangle - y_i|$ typically has a smaller scale for a sample $i\in\Omega_k^\star$ when compared with those $i\notin\Omega_k^\star$. This motivates us to include in Ω^t a certain fraction (denoted by $0<\alpha<1$) of samples enjoying the smallest empirical loss $|\langle \boldsymbol{A}_i, \boldsymbol{L}^t(\boldsymbol{R}^t)^\top\rangle - y_i|$. Intuitively, the fraction α should not be too large in which case Ω^t is likely to contain samples outside $\Omega_k^\star$; on the other hand, α should not be chosen too small in order not to waste information. As it turns out, choosing $0.6p_k \le \alpha \le 0.8p_k$ strikes a suitable balance and works well for our purpose. See Algorithm 3 for a precise description.

2.4 The full algorithm

With the three stages fully described, we can specify the whole algorithm in Algorithm 4, with the choices of algorithmic parameters listed in Table 1. Note that the discussion in Section 2.3 focuses on estimating a

Algorithm 4: A fully specified three-stage algorithm for mixed matrix sensing

1 **Input:** independent samples $\{\boldsymbol{A}_i, y_i\}_{1\le i\le N}, \{\boldsymbol{A}'_i, y'_i\}_{1\le i\le N_{\mathsf{MLR}}}$, parameters $R, \{r_k, \eta_k, \alpha_k\}_{1\le k\le K}$ (see Table 1).
2 Run Algorithm 1 with $\{\boldsymbol{A}_i, y_i\}_{1\le i\le N}$ and R to obtain $\boldsymbol{U}, \boldsymbol{V}$.
3 Run Algorithm 2 with $\{\boldsymbol{A}'_i, y'_i\}_{1\le i\le N_{\mathsf{MLR}}}, \boldsymbol{U}, \boldsymbol{V}$ and $\{r_k\}_{1\le k\le K}$ to obtain $\{\boldsymbol{L}_k, \boldsymbol{R}_k\}_{1\le k\le K}$.
4 **for** $k = 1, 2, \ldots, K$ **do**
5 Run Algorithm 3 on $\{\boldsymbol{A}_i, y_i\}_{1\le i\le N}$ with $(\boldsymbol{L}^0, \boldsymbol{R}^0) \leftarrow (\boldsymbol{L}_k, \boldsymbol{R}_k), \eta_k, \alpha_k$ to obtain $\boldsymbol{L}^{T_0}, \boldsymbol{R}^{T_0}$.
6 Set $\boldsymbol{M}_k \leftarrow \boldsymbol{L}^{T_0}(\boldsymbol{R}^{T_0})^\top$.
7 **Output:** $\{\boldsymbol{M}_k\}_{1\le k\le K}$.

Table 1: Our choices of the algorithmic parameters in Algorithm 4.

Algorithm 1	Rank $R = \mathsf{rank}(\sum_k p_k \boldsymbol{M}_k^\star)$.
Algorithm 2	Ranks $r_k = \mathsf{rank}(\boldsymbol{M}_k^\star), 1 \le k \le K$.
Algorithm 3 (for $\boldsymbol{M}_k^\star$)	Step size $0 < \eta_k \le 1.3/p_k$, truncating fraction $0.6p_k \le \alpha_k \le 0.8p_k$.

single component; in order to recover all K components $\{\boldsymbol{M}_k^\star\}_{1\le k\le K}$, we simply need to run Algorithm 3 for K times (which can be executed in parallel). In addition, Algorithm 4 is built upon sample splitting: while Stages 1 and 3 employ the same set of samples $\{\boldsymbol{A}_i, y_i\}_{1\le i\le N}$, Stage 2 (i.e. Line 3 of Algorithm 4) operates upon an *independent* set of samples $\{\boldsymbol{A}'_i, y'_i\}_{1\le i\le N_{\mathsf{MLR}}}$ (where "MLR" stands for "mixed linear regression"), thus resulting in a total sample complexity of $N + N_{\mathsf{MLR}}$. The main purpose of sample splitting is to decouple statistical dependency across stages and facilitate analysis. Finally, the interested reader is referred to Appendix D for a discussion regarding how to estimate certain parameters in Algorithm 4 if they are not known *a priori*.

3 Main results

3.1 Models and assumptions

For notational convenience, let us define the following parameters:

$$n \coloneqq \max\{n_1, n_2\}, \qquad r \coloneqq \max_{1\le k\le K} r_k, \qquad \kappa \coloneqq \max_{1\le k\le K} \kappa(\boldsymbol{M}_k^\star), \qquad \text{and} \quad \Gamma \coloneqq \frac{\max_{1\le k\le K} \|\boldsymbol{M}_k^\star\|_{\mathrm{F}}}{\min_{1\le k\le K} \|\boldsymbol{M}_k^\star\|_{\mathrm{F}}}, \tag{17}$$

where $\kappa(\boldsymbol{M}_k^\star) \coloneqq \sigma_1(\boldsymbol{M}_k^\star)/\sigma_{r_k}(\boldsymbol{M}_k^\star)$ stands for the condition number of $\boldsymbol{M}_k^\star$. This paper focuses on the *Gaussian design*, where the entries of each design matrix $\boldsymbol{A}_i$ are independently drawn from the standard Gaussian distribution. In addition, we assume that the samples drawn from the K components are reasonably *well-balanced* in the sense that for all $1 \le k \le K$,

$$p_k = \frac{|\Omega_k^\star|}{N} \asymp \frac{1}{K}, \tag{18}$$

where $\Omega_k^\star$ is the index set for the k-th component (see (1)). We assume that this well-balancedness assumption holds for both sets of samples $\{\boldsymbol{A}_i, y_i\}_{1\le i\le N}$ and $\{\boldsymbol{A}'_i, y'_i\}_{1\le i\le N_{\mathsf{MLR}}}$.

Next, we introduce an incoherence parameter that plays a crucial role in our theoretical development.

Definition 1. *The incoherence parameter* $\mu \ge 0$ is the smallest quantity that satisfies

$$\left\|\boldsymbol{U}_i^{\star\top}\boldsymbol{U}_j^\star\right\|_{\mathrm{F}} \le \frac{\mu r}{\sqrt{n_1}}, \qquad \text{and} \qquad \left\|\boldsymbol{V}_i^{\star\top}\boldsymbol{V}_j^\star\right\|_{\mathrm{F}} \le \frac{\mu r}{\sqrt{n_2}} \qquad \text{for all } 1 \le i < j \le K. \tag{19}$$

The incoherence parameter μ takes value on $[0, \sqrt{n/r}]$. As an example, if $\{\boldsymbol{U}_k^\star\}_{1\le k\le K}$ (resp. $\{\boldsymbol{V}_k^\star\}_{1\le k\le K}$) are random low-dimensional subspaces in $\mathbb{R}^{n_1}$ (resp. $\mathbb{R}^{n_2}$), then for any $i \ne j$, $\|\boldsymbol{U}_i^{\star\top}\boldsymbol{U}_j^\star\|_{\mathrm{F}}$ (resp. $\|\boldsymbol{V}_i^{\star\top}\boldsymbol{V}_j^\star\|_{\mathrm{F}}$)

is on the order of $\sqrt{r_i r_j/n_1}$ (resp. $\sqrt{r_i r_j/n_2}$), which is further upper bounded by $r/\sqrt{n_1}$ (resp. $r/\sqrt{n_2}$). This observation motivates our definition of the incoherence parameter. One of our main technical assumptions is that the column (resp. row) spaces of the ground-truth matrices are mutually weakly correlated — defined through the parameter μ — which covers a broad range of settings.

Assumption 1. *The incoherence parameter μ is upper bounded by*

$$\mu \le \frac{\sqrt{\min\{n_1, n_2\}}}{2r \max\{K, \sqrt{K\Gamma}\}}. \tag{20}$$

3.2 Theoretical guarantees

Exact recovery in the absence of noise. Our first main result uncovers that, in the noiseless case, Algorithm 4 achieves exact recovery efficiently, in terms of both sample and computational complexities.

Theorem 1 (Exact recovery)**.** *Consider the noiseless case (1) under the assumptions in Section 3.1. Suppose*

$$N \ge C_1 K^3 r^2 \kappa^2 \Gamma^2 \max\{K^2\Gamma^4, r\kappa^2\} \cdot n \log N \quad \text{and} \quad N_{\mathsf{MLR}} \ge C_2 K^8 r^2 \Gamma^{12} \max\{K^2, r\kappa^2\} \cdot \log n \cdot \log^3 N_{\mathsf{MLR}} \tag{21}$$

for some sufficiently large constants $C_1, C_2 > 0$. Then with probability at least $1 - o(1)$, there exists some permutation $\pi : \{1, \ldots, K\} \mapsto \{1, \ldots, K\}$ such that the outputs of Algorithm 4 obey for all $1 \le k \le K$

$$\left\|\boldsymbol{M}_{\pi(k)} - \boldsymbol{M}_k^\star\right\|_{\mathrm{F}} \le \left(1 - c_0 \eta_k p_k\right)^{T_0} \left\|\boldsymbol{M}_k^\star\right\|_{\mathrm{F}} \tag{22}$$

for some universal constant $0 < c_0 < 1/4$, where T_0 is the number of iterations used in Algorithm 3.

The proof can be found in Section 5. Two implications are in order.

- Suppose that the parameters $K, r, \kappa, \Gamma = O(1)$. In order to achieve exact recovery, the sample size N in (21) only needs to scale as $O(n \log n)$, while N_{MLR} only needs to exceed the order of $\log^4 n$.
- By setting the step size $\eta_k = c_1/p_k$ for some constant $0 < c_1 \le 1.3$, we see that the third stage (i.e. ScaledTGD) achieves linear convergence with a *constant* contraction rate, which is independent of the condition number $\kappa(\boldsymbol{M}_k^\star)$ of the matrix $\boldsymbol{M}_k^\star$.

Stability vis-à-vis noise. Moving on to the more realistic case with noise, we consider the following set of samples $\{\boldsymbol{A}_i, y_i\}_{1 \le i \le N}$:

$$\zeta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right), \qquad y_i = \begin{cases} \langle \boldsymbol{A}_i, \boldsymbol{M}_1^\star \rangle + \zeta_i, & \text{if } i \in \Omega_1^\star, \\ \ldots \\ \langle \boldsymbol{A}_i, \boldsymbol{M}_K^\star \rangle + \zeta_i, & \text{if } i \in \Omega_K^\star. \end{cases} \tag{23}$$

The set $\{\boldsymbol{A}_i', y_i'\}_{1 \le i \le N_{\mathsf{MLR}}}$ is independently generated in a similar manner. Our next result reveals that the proposed algorithm is stable against Gaussian noise. The proof is postponed to Section 5.

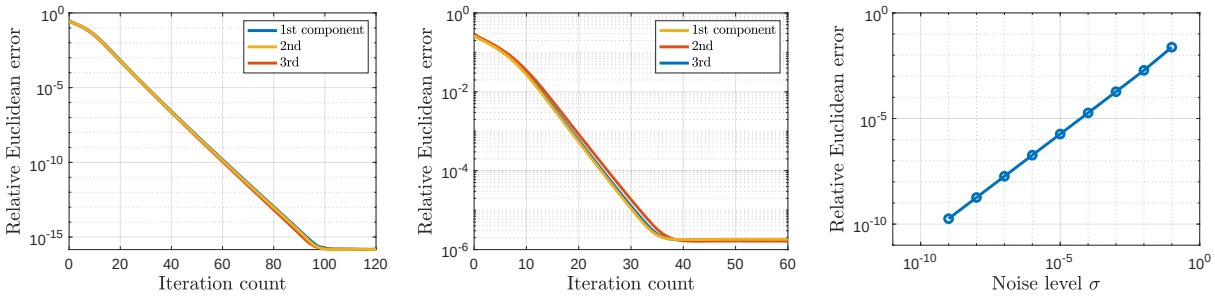
Theorem 2 (Stable recovery)**.** *Consider the noisy model (23) under the assumptions of Section 3.1. Suppose that the sample sizes satisfy (21), and that the noise level satisfies*

$$\sigma \le c \min_{1 \le k \le K} \|\boldsymbol{M}_k^\star\|_{\mathrm{F}} \cdot \min\left\{\frac{1}{K}, \frac{1}{\sqrt{r}\kappa}\right\} \tag{24}$$

for some sufficiently small constant $c > 0$. Then with probability at least $1 - o(1)$, there exists some permutation $\pi : \{1, \ldots, K\} \mapsto \{1, \ldots, K\}$ such that the outputs of Algorithm 4 obey for all $1 \le k \le K$

$$\left\|\boldsymbol{M}_{\pi(k)} - \boldsymbol{M}_k^\star\right\|_{\mathrm{F}} \le \left(1 - c_0 \eta_k p_k\right)^{T_0} \left\|\boldsymbol{M}_k^\star\right\|_{\mathrm{F}} + C_0 \max\left\{\sigma\sqrt{\frac{nrK^3 \log N}{N}}, \frac{K\sigma^2}{\min_{j:j\neq k} \|\boldsymbol{M}_j^\star - \boldsymbol{M}_k^\star\|_{\mathrm{F}}}\right\}, \tag{25}$$

where $0 < c_0 < 1/4$ and $C_0 > 0$ are some universal constants, and T_0 is the number of iterations used in Algorithm 3.



(a) (b) (c)

Figure 1: (a) The relative Euclidean error vs. the iteration count of ScaledTGD in Stage 3 of Algorithm 4 for each of the three components, in the noiseless case. (b) Convergence of ScaledTGD in the noisy case $\sigma = 10^{-5}$. (c) The largest relative Euclidean error (at convergence) of ScaledTGD in Algorithm 4, vs. the noise level σ. Each data point is an average over 10 independent trials.

Theorem 2 asserts that, when initialized using the proposed schemes, the ScaledTGD algorithm converges linearly until an error floor is hit. To interpret the statistical guarantees (25), we find it helpful to define the signal-to-noise-ratio (SNR) w.r.t. $\boldsymbol{M}_k^\star$ as follows:

$$\mathsf{SNR}_k := \frac{\mathbb{E}\left[\left|\langle \boldsymbol{A}_i, \boldsymbol{M}_k^\star\rangle\right|^2\right]}{\mathbb{E}[\zeta_i^2]} = \frac{\|\boldsymbol{M}_k^\star\|_{\mathrm{F}}^2}{\sigma^2}. \tag{26}$$

This together with the simple consequence $\min_{j:j\neq k} \|\boldsymbol{M}_j^\star - \boldsymbol{M}_k^\star\|_{\mathrm{F}} \gtrsim \|\boldsymbol{M}_k^\star\|_{\mathrm{F}}$ of Assumption 1 implies that

$$\frac{\|\boldsymbol{M}_{\pi(k)} - \boldsymbol{M}_k^\star\|_{\mathrm{F}}}{\|\boldsymbol{M}_k^\star\|_{\mathrm{F}}} \lesssim \max\left\{\frac{1}{\sqrt{\mathsf{SNR}_k}}\sqrt{\frac{nrK^3\log N}{N}}, \frac{K}{\mathsf{SNR}_k}\right\} \tag{27}$$

as long as the iteration number T_0 is sufficiently large. Here, the first term on the right-hand side of (27) matches the *minimax lower bound* for low-rank matrix sensing [CP11, Theorem 2.5] (the case with $K = 1$) up to a factor of $K\sqrt{\log N}$. In contrast, the second term on the right-hand side of (27) — which becomes very small as SNR_k grows — is not a function of the sample size N and does not vanish as $N \to \infty$. This term arises since, even at the population level, the point $(\boldsymbol{L}, \boldsymbol{R})$ satisfying $\boldsymbol{L}\boldsymbol{R}^\top = \boldsymbol{M}_k^\star$ is not a fixed point of the ScaledTGD update rule, due to the presence of mislabeled samples.

3.3 Numerical experiments

To validate our theoretical findings, we conduct a series of numerical experiments. To match practice, we do not deploy sample splitting (given that it is merely introduced to simplify analysis), and reuse the same dataset of size N for all three stages. Throughout the experiments, we set $n_1 = n_2 = n = 120$, $r = 2$, and $K = 3$. For each k, we let $p_k = 1/K$ and $\boldsymbol{\Sigma}_k^\star = \boldsymbol{I}_r$, and generate $\boldsymbol{U}_k^\star$ and $\boldsymbol{V}_k^\star$ as random r-dimensional subspaces in $\mathbb{R}^n$. We fix the sample size to be $N = 90\, nrK$. The algorithmic parameters are chosen according to our recommendations in Table 1. For instance, for each run of ScaledTGD, we set the step size as $\eta = 1.3K$ and the truncation fraction as $\alpha = 0.8/K$.

Linear convergence of ScaledTGD. Our first series of experiments aims at verifying the linear convergence of ScaledTGD towards the ground-truth matrices $\{\boldsymbol{M}_k^\star\}$ when initialized using the outputs of Stage 2. We consider both the noiseless case (i.e. $\sigma = 0$) and the noisy case $\sigma = 10^{-5}$. Figures 1(a) and 1(b) plot the relative Euclidean error $\|\boldsymbol{L}^t(\boldsymbol{R}^t)^\top - \boldsymbol{M}_k^\star\|_{\mathrm{F}}/\|\boldsymbol{M}_k^\star\|_{\mathrm{F}}$ versus the iteration count t for each component $1 \le k \le 3$. It is easily seen from Figures 1(a) and 1(b) that ScaledTGD, when seeded with the outputs from Stage 2, converges linearly to the ground-truth matrices $\{\boldsymbol{M}_k^\star\}$ in the absence of noise, and to within a small neighborhood of $\{\boldsymbol{M}_k^\star\}$ in the noisy setting.

Estimation error in the presence of random noise. The second series of experiments investigates the stability of the three-stage algorithm in the presence of random noise. We vary the noise level within $[10^{-9}, 10^{-1}]$. Figure 1(c) plots the largest relative Euclidean error $\max_{1\le k\le K} \|\boldsymbol{M}_k - \boldsymbol{M}_k^\star\|_{\mathrm{F}}/\|\boldsymbol{M}_k^\star\|_{\mathrm{F}}$ (where $\{\boldsymbol{M}_k\}$ are the outputs of Algorithm 4) versus the noise level σ, showing that the recovering error is indeed linear in σ, as predicted by our theory.

4 Prior work

Low-rank matrix recovery. There exists a vast literature on low-rank matrix recovery (e.g. [CR09, KMO10, BNS16, CC17, MWCC20, CCFM19, SL16, CLS15, JNS13, CCF+ar, CFMY19, SQW18, CLL20, DC20, NNS+14, CCG15, CCD+19, ACHL19, ZQW20, ZWYG18, LMZ18, PKCS17]); we refer the readers to [CC18b, CLC19] for an overview of this extensively studied topic. Most related to our work is the problem of matrix sensing (or low-rank matrix recovery from linear measurements). While convex relaxation [CR09, RFP10, CP11] enjoys optimal statistical performance, two-stage non-convex approaches [ZL15, TBS+16, TMC20] have received growing attention in recent years, due to their ability to achieve statistical and computational efficiency at once. Our three-stage algorithm is partially inspired by the two-stage approach along this line. It is worth mentioning that the non-convex loss function associated with low-rank matrix sensing enjoys benign landscape, which in turn enables tractable global convergence of simple first-order methods [BNS16, GJZ17, ZLTW18, LMZ18, LZT19].

Mixed linear regression. Being a classical problem in statistics [QR78], mixed linear regression has attracted much attention due to its broad applications in music perception [DV89, VT02], health care [DH00], trajectory clustering [GS99], plant science [Tur00], neuroscience [YPCR18], to name a few. While computationally intractable in the worst case [YCS14], mixed linear regression can be solved efficiently under certain statistical models on the design matrix. Take the two-component case for instance: efficient methods include alternating minimization with initialization via grid search [YCS14], EM with random initialization [KYB19, KQC+19], and convex reformulations [CYC14, HJ18], where EM further achieves minimax estimation guarantees [CYC14] in the presence of Gaussian noise [KHC20]. Mixed linear regression becomes substantially more challenging when the number K of components is allowed to grow with n. The state-of-the-art method — namely, the Fourier moment method [CLS20] — achieves sub-exponential sample and computational complexities w.r.t. K, whereas other methods (e.g. the method of moments [LL18] and grid search over K-dimensional subspaces [SS19a]) all have exponential dependence on K. It turns out that by restricting the ground-truth vectors to be in "general position" (e.g. linearly independent), tensor methods [YCS16, CL13, SJA16, ZJD16] solve mixed linear regression with polynomial sample and computational complexities in K. It is worth noting that most of the prior work focused on the Gaussian design for theoretical analysis, with a few exceptions [CYC14, HJ18, SS19a]. Another line of work [KC07, SBVDG10, YPCR18, KMMP19, MP20] considered mixed linear regression with sparsity, which is beyond the scope of the current paper.

Mixed low-rank matrix estimation. Moving beyond mixed linear regressions, there are a few papers that tackle mixtures of low-rank models. For example, [YC15] proposed a regularized EM algorithm and applied it to mixed matrix sensing with two symmetric components; however, only local convergence was investigated therein. Additionally, [PA18] was the first to systematically study mixed matrix completion, investigating the identifiability conditions and sample complexities of this problem; however, the heuristic algorithm proposed therein comes without provable guarantees.

Iterative truncated loss minimization. Least trimmed square [Rou84] is a classical method for robust linear regression. Combining the idea of trimming (i.e. selecting a subset of "good" samples) with iterative optimization algorithms (e.g. gradient descent and its variants) leads to a general paradigm of iterative truncated loss minimization — a principled method for improving robustness w.r.t. heavy-tailed data, adversarial outliers, etc. [SS19b, SWS20]. Successful applications of this kind include linear regression [BJK15], mixed linear regression [SS19a], phase retrieval [CC17, ZCL18], matrix sensing [LCZL20], and learning entangled single-sample distributions [YL20], among others.

Multi-task learning and meta-learning. The aim of multi-task learning [Car97, Bax00, BDS03, AZ05, EMP05, AEP07, JSRR10, PLW+20, PL14, MPRP16] is to simultaneously learn a model that connect multiple *related* tasks. Exploiting the similarity across tasks enables improved performance for learning each individual task, and leads to enhance generalization capabilities for unseen but related tasks with limited samples. This paradigm (or its variants) is also referred to in the literature as meta-learning [FAL17, TJJ20] (i.e. learning-to-learn), transfer learning [PY09], and few-shot learning [SSZ17, DHK+20], depending on the specific scenarios of interest. Our study on learning mixture of models is related to the probabilistic approach taken in multi-task learning and meta-learning, in which all the tasks (both the training and the testing ones) are independently sampled from a common environment, i.e. a prior distribution of tasks [Bax00]. See [KSS+20, KSKO20] for recent efforts that make explicit the connection between mixed linear regression and meta-learning.

5 Analysis

In this section, we present the proofs of Theorems 1 and 2. Our analysis is modular in the sense that we deliver the performance guarantees for the three stages separately that are independent of each other. For instance, one can replace the tensor method in Stage 2 by any other mixed linear regression solver with provable guarantees, without affecting Stages 1 and 3.

Stage 1. The first result confirms that given enough samples, Algorithm 1 outputs reasonable estimates of the subspaces $(\boldsymbol{U}^\star, \boldsymbol{V}^\star)$ (cf. (5)). The proof is deferred to Appendix B.1.

Theorem 3. *Consider the model (23) under the assumptions in Section 3.1. Recall the definitions of κ and Γ in (17). For any $0 < \delta < 1$, the estimates $\boldsymbol{U}$ and $\boldsymbol{V}$ returned by Algorithm 1 satisfy*

$$\max\left\{\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\right\|, \left\|\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{V}^\star\boldsymbol{V}^{\star\top}\right\|\right\} \lesssim \delta K\sqrt{r}\kappa\left(\Gamma + \frac{1}{\sqrt{Kr}}\frac{\sigma}{\min_k \|\boldsymbol{M}_k^\star\|_{\mathrm{F}}}\right) \tag{28}$$

with probability at least $1 - Ce^{-cn}$ for some universal constants $C, c > 0$, provided that the sample size obeys

$$N \geq C_0 \frac{nrK}{\delta^2}\log\frac{1}{\delta} \tag{29}$$

for some sufficiently large constant $C_0 > 0$.

Stage 2. Next, we demonstrate that the tensor method employed in Algorithm 2 reliably solves the intermediate mixed linear regression problem defined in (12). The proof is postponed to Appendix B.2.

Theorem 4. *Consider the model (23) under the assumptions in Section 3.1. Suppose that the subspace estimates $\boldsymbol{U}$ and $\boldsymbol{V}$ are independent of $\{\boldsymbol{A}_i, y_i\}_{1\leq i\leq N}$ and obey $\max\{\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\|, \|\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{V}^\star\boldsymbol{V}^{\star\top}\|\} \leq c_1/(K\Gamma^2)$ for some sufficiently small constant $c_1 > 0$. Let $\{\widehat{\boldsymbol{\beta}}_k\}_{1\leq k\leq K}$ be the estimates returned by Line 3 of Algorithm 2. Given any $0 < \epsilon \leq c_2/K$, there exists a permutation $\pi(\cdot) : \{1, \ldots, K\} \mapsto \{1, \ldots, K\}$ such that*

$$\left\|\widehat{\boldsymbol{\beta}}_{\pi(k)} - \boldsymbol{\beta}_k\right\|_2 \leq \epsilon \cdot \max_{1\leq j\leq K}\left\|\boldsymbol{M}_j^\star\right\|_{\mathrm{F}} \qquad \text{for all } 1 \leq k \leq K \tag{30}$$

with probability at least $1 - O(1/\log n)$, provided that the sample size obeys

$$N \geq C\frac{K^8 r^2}{\epsilon^2}\left(\Gamma^{10} + \frac{\sigma^{10}}{\min_k \|\boldsymbol{M}_k^\star\|_{\mathrm{F}}^{10}}\right)\log n \cdot \log^3 N. \tag{31}$$

Here, $c_2 > 0$ (resp. $C > 0$) is some sufficiently small (resp. large) constant.

From now on, we shall assume without loss of generality that $\pi(\cdot)$ is an identity map (i.e. $\pi(k) = k$) to simplify the presentation. Our next result transfers the estimation error bounds for $\boldsymbol{U}, \boldsymbol{V}$ and $\{\widehat{\boldsymbol{\beta}}_k\}$ to that for $\{\boldsymbol{L}_k\boldsymbol{R}_k^\top\}$, thus concluding the analysis of Stage 2; see Appendix B.3 for a proof.

Proposition 1. *The estimates $\{\boldsymbol{L}_k, \boldsymbol{R}_k\}_{k=1}^K$ computed in Lines 4-6 of Algorithm 2 obey*

$$\left\|\boldsymbol{L}_k\boldsymbol{R}_k^\top - \boldsymbol{M}_k^\star\right\|_{\mathrm{F}} \leq 2\max\left\{\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\right\|, \left\|\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{V}^\star\boldsymbol{V}^{\star\top}\right\|\right\}\left\|\boldsymbol{M}_k^\star\right\|_{\mathrm{F}} + 2\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_2 \tag{32}$$

for all $1 \leq k \leq K$.

Stage 3. The last result guarantees that Algorithm 3 — when suitably initialized — converges linearly towards $\boldsymbol{M}_k^\star$ up to a certain error floor. Here $\boldsymbol{M}_k^\star$ is the closest among $\{\boldsymbol{M}_j^\star\}_{1\le j\le K}$ to the point $\boldsymbol{L}^0(\boldsymbol{R}^0)^\top$. The proof can be found in Appendix B.4.

Theorem 5. *Consider the model (23) under the assumptions in Section 3.1. Suppose that the noise level obeys (24). Choose the step size η and truncating fraction α such that $0<\eta\le 1.3/p_k$ and $0.6p_k\le\alpha\le 0.8p_k$. Given any $0<\delta<c_0/K$, if $\boldsymbol{L}^0\in\mathbb{R}^{n_1\times r_k}$ and $\boldsymbol{R}^0\in\mathbb{R}^{n_2\times r_k}$ obey*

$$\left\|\boldsymbol{L}^0(\boldsymbol{R}^0)^\top-\boldsymbol{M}_k^\star\right\|_{\mathrm{F}}\le c_1\left\|\boldsymbol{M}_k^\star\right\|_{\mathrm{F}}\cdot\min\left\{\frac{1}{\sqrt{r}\kappa},\frac{1}{K}\right\}, \tag{33}$$

then with probability at least $1-Ce^{-cn}$ the iterates of Algorithm 3 satisfy

$$\left\|\boldsymbol{L}^t(\boldsymbol{R}^t)^\top-\boldsymbol{M}_k^\star\right\|_{\mathrm{F}}\le(1-c_2\eta p_k)^t\left\|\boldsymbol{L}^0(\boldsymbol{R}^0)^\top-\boldsymbol{M}_k^\star\right\|_{\mathrm{F}}+C_2\max\left\{K\sigma\delta,\frac{K\sigma^2}{\min_{j:j\neq k}\|\boldsymbol{M}_j^\star-\boldsymbol{M}_k^\star\|_{\mathrm{F}}}\right\} \tag{34}$$

for all $t\ge 0$, provided that the sample size exceeds $N\ge C_0\frac{nrK}{\delta^2}\log N$. Here, $0<c_2<1/4$ and $C,c,C_2>0$ are some universal constants, and $c_0,c_1>0$ (resp. $C_0>0$) are some sufficiently small (resp. large) constants.

Putting pieces together: proof of Theorems 1 and 2. With the above performance guarantees in place, we are ready to establish the main theorems. Note that due to sample splitting in Algorithm 4, we shall apply Theorems 3 and 5 to the dataset $\{\boldsymbol{A}_i,y_i\}_{1\le i\le N}$, and Theorem 4 to the dataset $\{\boldsymbol{A}_i',y_i'\}_{1\le i\le N_{\mathsf{MLR}}}$. Set

$$\delta\le c_3\frac{1}{K\sqrt{r}\kappa\Gamma}\min\left\{\frac{1}{\sqrt{r}\kappa},\frac{1}{K\Gamma^2}\right\},\qquad\text{and}\qquad\epsilon\le c_4\frac{1}{\Gamma}\min\left\{\frac{1}{\sqrt{r}\kappa},\frac{1}{K}\right\},$$

for some sufficiently small constants $c_3,c_4>0$ in Theorems 3 and 4. These choices — in conjunction with our assumption on σ in Theorem 2, as well as Proposition 1 — guarantee that the initialization $\boldsymbol{L}^0(\boldsymbol{R}^0)^\top$ lies in the neighborhood of $\boldsymbol{M}_k^\star$ as required by (33). This allows us to invoke Theorem 5 to conclude the proof of Theorem 2. Finally, Theorem 1 follows by simply setting the noise level $\sigma=0$ in Theorem 2.

6 Discussion

This paper develops a three-stage algorithm for the mixed low-rank matrix sensing problem, which is provably efficient in terms of both sample and computational complexities. Having said this, there are numerous directions that are worthy of further investigations; we single out a few in the following.

To begin with, while our required sample complexity scales linearly (and optimally) w.r.t. the matrix dimension $\max\{n_1,n_2\}$, its dependency on other salient parameters — e.g. the number K of components, the ranks $\{r_k\}$ of the ground-truth matrices $\{\boldsymbol{M}_k^\star\}$ — is likely sub-optimal. Improving the sample efficiency in these aspects is certainly an interesting direction to explore. In addition, in the presence of *random* noise, the performance of ScaledTGD saturates after the number of samples exceeds a certain threshold. It would be helpful to investigate other algorithms like expectation-maximization to see whether there is any performance gain one can harvest. Furthermore, our current theory builds upon the Gaussian designs $\{\boldsymbol{A}_i\}$, which often does not capture the practical scenarios. It is of great practical importance to develop efficient algorithms that can accommodate a wider range of design matrices $\{\boldsymbol{A}_i\}$ — for instance, the case of mixed low-rank matrix completion. Last but not least, it would be of interest to study more general meta-learning settings in the presence of both light and heavy tasks (beyond the current single-sample setting) [KSS+20], and see how sample complexities can be reduced (compared to meta-learning for mixed regression) by exploiting such low-complexity structural priors .

Acknowledgements

Y. Chen is supported in part by the grants AFOSR YIP award FA9550-19-1-0030, ONR N00014-19-1-2120, ARO YIP award W911NF-20-1-0097, ARO W911NF-18-1-0303, NSF CCF-1907661, IIS-1900140 and DMS-2014279, and the Princeton SEAS Innovation Award. H. V. Poor is supported in part by NSF CCF-1908308,

Algorithm 5: The tensor method for mixed linear regression [YCS16, Algorithm 1]

Input: $\{\boldsymbol{a}_i, y_i\}_{1\le i\le N}$.
Randomly split the samples into two disjoint sets $\{\boldsymbol{a}_i, y_i\}_{1\le i\le N_1}, \{\boldsymbol{a}'_i, y'_i\}_{1\le i\le N_2}$ such that $N = N_1 + N_2$, by assigning each sample to either dataset with probability 0.5.
Compute $m_0 \leftarrow \frac{1}{N_1}\sum_{i=1}^{N_1} y_i^2, \boldsymbol{m}_1 \leftarrow \frac{1}{6N_2}\sum_{i=1}^{N_2} {y'_i}^3 \boldsymbol{a}'_i$.
Compute $\boldsymbol{M}_2 \leftarrow \frac{1}{2N_1}\sum_{i=1}^{N_1} y_i^2 \boldsymbol{a}_i\boldsymbol{a}_i^\top - \frac{1}{2}m_0\boldsymbol{I}_d, \boldsymbol{M}_3 \leftarrow \frac{1}{6N_2}\sum_{i=1}^{N_2} {y'_i}^3 \boldsymbol{a}_i'^{\otimes 3} - \mathcal{T}(\boldsymbol{m}_1)$, where $\mathcal{T}$ is defined in (36).
Denote the rank-K SVD of $\boldsymbol{M}_2$ as $\boldsymbol{U}_2\boldsymbol{\Sigma}_2\boldsymbol{V}_2^\top$, and compute the whitening matrix $\boldsymbol{W} \leftarrow \boldsymbol{U}_2\boldsymbol{\Sigma}_2^{-1/2}$.
Compute $\widetilde{\boldsymbol{M}}_3 \leftarrow \boldsymbol{M}_3(\boldsymbol{W}, \boldsymbol{W}, \boldsymbol{W})$.
Run the robust tensor power method [YCS16, Algorithm 2] on $\widetilde{\boldsymbol{M}}_3$ to obtain K eigenvalue/eigenvector pairs $\{\widetilde{\omega}_k, \widetilde{\boldsymbol{\beta}}_k\}_{1\le k\le K}$.
Compute $\omega_k \leftarrow 1/\widetilde{\omega}_k^2, \boldsymbol{\beta}_k \leftarrow \widetilde{\omega}_k \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\widetilde{\boldsymbol{\beta}}_k, 1 \le k \le K$.
Output: $\{\omega_k, \boldsymbol{\beta}_k\}_{1\le k\le K}$.

and in part by a Princeton Schmidt Data-X Research Award. We would like to thank Qixing Huang who taught us the symmetry synchronoziation problem in computer vision that largely inspired this research.

A The tensor method for mixed linear regression

This section reviews the tensor method proposed in [YCS16] for solving mixed linear regression. For simplicity of exposition, we consider the noiseless case where we have access to the samples $\{\boldsymbol{a}_i, y_i\}_{1\le i\le N}$ obeying

$$y_i = \begin{cases} \langle \boldsymbol{a}_i, \boldsymbol{\beta}_1^\star\rangle, & \text{if } i \in \Omega_1^\star, \\ \dots & \dots \\ \langle \boldsymbol{a}_i, \boldsymbol{\beta}_K^\star\rangle, & \text{if } i \in \Omega_K^\star. \end{cases} \tag{35}$$

Our goal is to recover the ground truths $\boldsymbol{\beta}_k^\star \in \mathbb{R}^d, 1 \le k \le K$, without knowing the index sets $\{\Omega_k^\star\}$.

Notation for tensors. For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, denote by $\boldsymbol{A}\otimes\boldsymbol{B}$ their Kronecker product, and let $\boldsymbol{A}^{\otimes 3}$ represent $\boldsymbol{A}\otimes\boldsymbol{A}\otimes\boldsymbol{A}$. For a symmetric tensor $\boldsymbol{T} \in \mathbb{R}^{d\times d\times d}$ and matrices $\boldsymbol{A} \in \mathbb{R}^{d\times d_1}, \boldsymbol{B} \in \mathbb{R}^{d\times d_2}, \boldsymbol{C} \in \mathbb{R}^{d\times d_3}$, let $\boldsymbol{T}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \in \mathbb{R}^{d_1\times d_2\times d_3}$ denote the multi-linear matrix multiplication such that

$$\big[\boldsymbol{T}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\big]_{m,n,p} = \sum_{1\le i,j,k\le d} T_{i,j,k}A_{i,m}B_{j,n}C_{k,p}, \quad 1 \le m \le d_1,\ 1 \le n \le d_2,\ 1 \le p \le d_3.$$

In addition, let $\|\boldsymbol{T}\|$ stand for the operator norm of $\boldsymbol{T}$, namely, $\|\boldsymbol{T}\| := \sup_{\boldsymbol{x}:\|\boldsymbol{x}\|_2=1} \big|\boldsymbol{T}(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x})\big|$.

The tensor method: algorithm and rationale. We summarize the tensor method in Algorithm 5, which is mostly the same as [YCS16, Algorithm 1] and included here for completeness.

In the following, we explain the intuitions behind its algorithmic design. Given data $\{\boldsymbol{a}_i, y_i\}_{1\le i\le N}$ generated according to (35), we compute the following empirical moments:

$$m_0 := \frac{1}{N}\sum_{i=1}^N y_i^2 \in \mathbb{R}, \quad \boldsymbol{m}_1 := \frac{1}{6N}\sum_{i=1}^N y_i^3\boldsymbol{a}_i \in \mathbb{R}^d,$$

$$\boldsymbol{M}_2 := \frac{1}{2N}\sum_{i=1}^N y_i^2\boldsymbol{a}_i\boldsymbol{a}_i^\top - \frac{1}{2}m_0\boldsymbol{I}_d \in \mathbb{R}^{d\times d}, \quad \boldsymbol{M}_3 := \frac{1}{6N}\sum_{i=1}^N y_i^3\boldsymbol{a}_i^{\otimes 3} - \mathcal{T}(\boldsymbol{m}_1) \in \mathbb{R}^{d\times d\times d};$$

here, letting $\{\boldsymbol{e}_i\}_{1\le i\le d}$ be the canonical basis of $\mathbb{R}^d$, we define the operator $\mathcal{T}(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^{d\times d\times d}$ as

$$\mathcal{T}(\boldsymbol{m}) := \sum_{i=1}^d \left(\boldsymbol{m}\otimes\boldsymbol{e}_i\otimes\boldsymbol{e}_i + \boldsymbol{e}_i\otimes\boldsymbol{m}\otimes\boldsymbol{e}_i + \boldsymbol{e}_i\otimes\boldsymbol{e}_i\otimes\boldsymbol{m}\right), \quad \text{where } \boldsymbol{m} \in \mathbb{R}^d. \tag{36}$$

The key observation is that: under the Gaussian design (i.e. $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$), $\boldsymbol{M}_2$ and $\boldsymbol{M}_3$ reveal crucial second-order and third-order moments of $\{\boldsymbol{\beta}_k^\star\}$ since (cf. [YCS16, Lemma 1])

$$\mathbb{E}[\boldsymbol{M}_2] = \sum_{k=1}^{K} p_k \boldsymbol{\beta}_k^\star (\boldsymbol{\beta}_k^\star)^\top \quad \text{and} \quad \mathbb{E}[\boldsymbol{M}_3] = \sum_{k=1}^{K} p_k (\boldsymbol{\beta}_k^\star)^{\otimes 3},$$

where we recall $p_k = |\Omega_k^\star|/N$. This motivates one to apply tensor decomposition [AGH+14] on $\boldsymbol{M}_2$ and $\boldsymbol{M}_3$ in order to estimate $\{\boldsymbol{\beta}_k^\star\}$ and $\{p_k\}$. Indeed, the estimates $\{\boldsymbol{\beta}_k\}$ and $\{\omega_k\}$ returned by Algorithm 5 serve as our estimates of $\{\boldsymbol{\beta}_k^\star\}$ and $\{p_k\}$, respectively.

Remark 1 (Sample splitting). Similar to [YCS16], we assume that m_0 and $\boldsymbol{M}_2$ are computed using one set of data, while $\boldsymbol{M}_1$ and $\boldsymbol{M}_3$ are obtained based on another *independent* set of samples. This sample splitting strategy ensures that the whitening matrix $\boldsymbol{W}$ is independent of $\boldsymbol{M}_3$, thus simplifying theoretical analysis.

B Proofs for Section 5

For notational simplicity, we use $\mathsf{dist}_{\boldsymbol{U},\boldsymbol{V}}$ throughout to denote the following subspace estimation error:

$$\mathsf{dist}_{\boldsymbol{U},\boldsymbol{V}} \coloneqq \max\Big\{\big\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\big\|, \big\|\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{V}^\star\boldsymbol{V}^{\star\top}\big\|\Big\}. \tag{37}$$

B.1 Proof of Theorem 3

The proof is decomposed into two steps: we first develop an upper bound $\|\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]\|$ (where $\boldsymbol{Y}$ is as defined in Algorithm 1), and then combine this with Wedin's Theorem to control the subspace distance $\mathsf{dist}_{\boldsymbol{U},\boldsymbol{V}}$.

Step 1: controlling $\|\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]\|$. We start by decomposing $\boldsymbol{Y}$ into $\boldsymbol{Y} = \boldsymbol{Y}_A + \boldsymbol{Y}_\zeta$, where we define

$$\boldsymbol{Y}_A \coloneqq \sum_{k=1}^{K} \frac{p_k}{|\Omega_k^\star|} \sum_{i\in\Omega_k^\star} \langle \boldsymbol{A}_i, \boldsymbol{M}_k^\star\rangle \boldsymbol{A}_i \qquad \text{and} \qquad \boldsymbol{Y}_\zeta \coloneqq \frac{1}{N}\sum_{i=1}^{N} \zeta_i \boldsymbol{A}_i.$$

Lemma 1 asserts that: with probability at least $1 - Ce^{-cn}$ for some universal constants $C, c > 0$, we have

$$\left\| \frac{1}{|\Omega_k^\star|} \sum_{i\in\Omega_k^\star} \langle \boldsymbol{A}_i, \boldsymbol{M}_k^\star\rangle \boldsymbol{A}_i - \boldsymbol{M}_k^\star \right\| \le \delta \|\boldsymbol{M}_k^\star\|_{\mathrm{F}}, \quad 1 \le k \le K,$$

as long as the sample size N satisfies (29), which together with the triangle inequality further implies

$$\|\boldsymbol{Y}_A - \mathbb{E}[\boldsymbol{Y}_A]\| \le \sum_{k=1}^{K} p_k \left\| \frac{1}{|\Omega_k^\star|} \sum_{i\in\Omega_k^\star} \langle \boldsymbol{A}_i, \boldsymbol{M}_k^\star\rangle \boldsymbol{A}_i - \boldsymbol{M}_k^\star \right\| \le \delta \sum_{k=1}^{K} p_k \|\boldsymbol{M}_k^\star\|_{\mathrm{F}} \le \delta \max_{1\le k\le K} \|\boldsymbol{M}_k^\star\|_{\mathrm{F}}.$$

In addition, [CP11, Lemma 1.1] reveals that with probability at least $1 - Ce^{-cn}$ for some constants $C, c > 0$,

$$\|\boldsymbol{Y}_\zeta\| \lesssim \sigma\sqrt{\frac{n}{N}} \lesssim \sigma\frac{\delta}{\sqrt{Kr}}$$

holds under the sample size condition (29). Given that $\mathbb{E}[\boldsymbol{Y}_A] = \mathbb{E}[\boldsymbol{Y}] = \sum_k p_k \boldsymbol{M}_k^\star$, we have established the existence of some universal constant $C_1 > 0$ such that

$$\|\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]\| \le \|\boldsymbol{Y}_A - \mathbb{E}[\boldsymbol{Y}_A]\| + \|\boldsymbol{Y}_\zeta\| \le C_1\delta\left(\max_{1\le k\le K}\|\boldsymbol{M}_k^\star\|_{\mathrm{F}} + \frac{\sigma}{\sqrt{Kr}}\right) =: \Delta. \tag{38}$$

Step 2: controlling $\mathsf{dist}_{\boldsymbol{U},\boldsymbol{V}}$. Before embarking on controlling $\mathsf{dist}_{\boldsymbol{U},\boldsymbol{V}}$, we make the following claim.

Claim 1. *Under the assumptions of Theorem 3, we have*

$$\mathsf{col}\left\{\sum_{k=1}^{K} p_k \boldsymbol{M}_k^\star\right\} = \mathsf{col}\Big\{[\boldsymbol{U}_1^\star, \ldots, \boldsymbol{U}_K^\star]\Big\}, \quad \mathsf{row}\left\{\sum_{k=1}^{K} p_k \boldsymbol{M}_k^\star\right\} = \mathsf{col}\Big\{[\boldsymbol{V}_1^\star, \ldots, \boldsymbol{V}_K^\star]\Big\}, \quad R = \sum_{k=1}^{K} r_k, \tag{39a}$$

$$\text{and} \qquad \sigma_R\left(\sum_{k=1}^{K} p_k \boldsymbol{M}_k^\star\right) \gtrsim \frac{1}{K} \min_k \sigma_{r_k}(\boldsymbol{M}_k^\star). \tag{39b}$$

With this claim in place, we are ready to apply Wedin's Theorem [Wed72] to obtain

$$\mathsf{dist}_{\boldsymbol{U},\boldsymbol{V}} \le \frac{\|\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]\|}{\sigma_R(\mathbb{E}[\boldsymbol{Y}]) - \|\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]\|} \le \frac{\Delta}{\sigma_R(\mathbb{E}[\boldsymbol{Y}]) - \Delta} \le \frac{2\Delta}{\sigma_R(\mathbb{E}[\boldsymbol{Y}])} = 2C_1 \frac{\delta\left(\max_k \|\boldsymbol{M}_k^\star\|_{\mathrm{F}} + \frac{\sigma}{\sqrt{Kr}}\right)}{\sigma_R\left(\sum_k p_k \boldsymbol{M}_k^\star\right)}, \tag{40}$$

with the proviso that Δ defined in (38) obeys $\Delta \le \frac{1}{2}\sigma_R(\mathbb{E}[\boldsymbol{Y}])$. On the other hand, if instead one has $\Delta > \frac{1}{2}\sigma_R(\mathbb{E}[\boldsymbol{Y}])$, then we claim that (40) trivially holds; this can be seen by observing that $\mathsf{dist}_{\boldsymbol{U},\boldsymbol{V}} \le 1$, while the right-hand side of (40) is greater than 1 if $\Delta > \frac{1}{2}\sigma_R(\mathbb{E}[\boldsymbol{Y}])$. Finally, Claim 1 tells us that

$$\sigma_R\left(\sum_k p_k \boldsymbol{M}_k^\star\right) \gtrsim \frac{1}{K} \min_k \sigma_{r_k}(\boldsymbol{M}_k^\star) \gtrsim \frac{1}{K\sqrt{r}\kappa} \min_k \|\boldsymbol{M}_k^\star\|_{\mathrm{F}}.$$

Substituting this relation into (40) immediately leads to the advertised bound (28) in Theorem 3.

Proof of Claim 1. Recall that we can write $\sum_k p_k \boldsymbol{M}_k^\star$ in terms of $\{\boldsymbol{U}_k^\star, \boldsymbol{\Sigma}_k^\star, \boldsymbol{V}_k^\star\}$, in the form of (8). Therefore, to prove (39a), it suffices to show that $\min\{\sigma_{R'}([\boldsymbol{U}_1^\star, \ldots, \boldsymbol{U}_K^\star]), \sigma_{R'}([\boldsymbol{V}_1^\star, \ldots, \boldsymbol{V}_K^\star])\} \ge 1/\sqrt{2}$, where $R' \coloneqq \sum_k r_k$. We only prove this for $\sigma_{R'}([\boldsymbol{U}_1^\star, \ldots, \boldsymbol{U}_K^\star])$, since the proof for $\sigma_{R'}([\boldsymbol{V}_1^\star, \ldots, \boldsymbol{V}_K^\star])$ is identical. Denoting $\boldsymbol{W} \coloneqq [\boldsymbol{U}_1^\star, \ldots, \boldsymbol{U}_K^\star]$ for notational convenience, we have

$$\boldsymbol{W}^\top \boldsymbol{W} = \begin{bmatrix} \boldsymbol{U}_1^{\star\top} \\ \vdots \\ \boldsymbol{U}_K^{\star\top} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_1^\star & \ldots & \boldsymbol{U}_K^\star \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_{r_1} & \boldsymbol{U}_1^{\star\top}\boldsymbol{U}_2^\star & \ldots & \boldsymbol{U}_1^{\star\top}\boldsymbol{U}_K^\star \\ \boldsymbol{U}_2^{\star\top}\boldsymbol{U}_1^\star & \boldsymbol{I}_{r_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \boldsymbol{U}_K^{\star\top}\boldsymbol{U}_1^\star & \ldots & \ldots & \boldsymbol{I}_{r_K} \end{bmatrix}.$$

This together with Assumption 1 gives

$$\|\boldsymbol{W}^\top \boldsymbol{W} - \boldsymbol{I}_{R'}\|_{\mathrm{F}}^2 = \sum_{i \ne j} \|\boldsymbol{U}_i^{\star\top} \boldsymbol{U}_j^\star\|_{\mathrm{F}}^2 \le K^2 \left(\frac{1}{2K}\right)^2 \le \frac{1}{4}.$$

Apply Weyl's inequality to obtain

$$\sigma_{R'}(\boldsymbol{W}^\top \boldsymbol{W}) \ge 1 - \|\boldsymbol{W}^\top \boldsymbol{W} - \boldsymbol{I}_{R'}\| \ge 1 - \|\boldsymbol{W}^\top \boldsymbol{W} - \boldsymbol{I}_{R'}\|_{\mathrm{F}} \ge \frac{1}{2},$$

thus indicating that $\sigma_{R'}(\boldsymbol{W}) = \sqrt{\sigma_{R'}(\boldsymbol{W}^\top \boldsymbol{W})} \ge 1/\sqrt{2}$. This completes the proof of (39a).

Next, we turn attention to (39b). Denote the SVD of $[\boldsymbol{U}_1^\star, \ldots, \boldsymbol{U}_K^\star]$ (resp. $[\boldsymbol{V}_1^\star, \ldots, \boldsymbol{V}_K^\star]$) as $\boldsymbol{U}_{\mathsf{left}} \boldsymbol{\Sigma}_{\mathsf{left}} \boldsymbol{V}_{\mathsf{left}}^\top$ (resp. $\boldsymbol{U}_{\mathsf{right}} \boldsymbol{\Sigma}_{\mathsf{right}} \boldsymbol{V}_{\mathsf{right}}^\top$), where $\boldsymbol{V}_{\mathsf{left}}$ (resp. $\boldsymbol{V}_{\mathsf{right}}$) is a $R \times R$ orthonormal matrix. Substitution into (8) yields

$$\sum_{k=1}^{K} p_k \boldsymbol{M}_k^\star = \boldsymbol{U}_{\mathsf{left}} \boldsymbol{\Sigma}_{\mathsf{left}} \boldsymbol{V}_{\mathsf{left}}^\top \mathsf{diag}\big(\{p_k \boldsymbol{\Sigma}_k^\star\}_{1 \le k \le K}\big) \boldsymbol{V}_{\mathsf{right}} \boldsymbol{\Sigma}_{\mathsf{right}} \boldsymbol{U}_{\mathsf{right}}^\top,$$

where $\mathsf{diag}\big(\{p_k \boldsymbol{\Sigma}_k^\star\}_{1 \le k \le K}\big)$ is a $R \times R$ full-rank diagonal matrix, with blocks $p_1 \boldsymbol{\Sigma}_1^\star, \ldots, p_K \boldsymbol{\Sigma}_K^\star$ on the diagonal. This implies that

$$\sigma_R\left(\sum_{k=1}^{K} p_k \boldsymbol{M}_k^\star\right) = \sigma_R\Big(\boldsymbol{\Sigma}_{\mathsf{left}} \boldsymbol{V}_{\mathsf{left}}^\top \mathsf{diag}\big(\{p_k \boldsymbol{\Sigma}_k^\star\}_{1 \le k \le K}\big) \boldsymbol{V}_{\mathsf{right}} \boldsymbol{\Sigma}_{\mathsf{right}}\Big) \ge \sigma_R(\boldsymbol{\Sigma}_{\mathsf{left}}) \sigma_R(\boldsymbol{\Sigma}_{\mathsf{right}}) \cdot \min_k \big\{p_k \sigma_{r_k}(\boldsymbol{M}_k^\star)\big\}$$

$$\geq \left(\frac{1}{\sqrt{2}}\right)^2 \min_k \left\{p_k \sigma_{r_k}(\boldsymbol{M}_k^\star)\right\} \gtrsim \frac{1}{K} \min_k \sigma_{r_k}(\boldsymbol{M}_k^\star),$$

where the last inequality uses the assumption that $p_k \gtrsim 1/K$. This establishes (39b). □

B.2 Proof of Theorem 4

Step 1: basic properties of the auxiliary mixed linear regression problem. We begin by formally characterizing the intermediate mixed linear regression problem in Stage 2. It is easily seen from Section 2.2 that for $i \in \Omega_k^\star$, one has

$$y_i = \langle \boldsymbol{A}_i, \boldsymbol{M}_k^\star \rangle + \zeta_i = \langle \boldsymbol{a}_i, \boldsymbol{\beta}_k \rangle + \underbrace{z_i + \zeta_i}_{=:\xi_i}, \tag{41}$$

where the additional term

$$z_i \coloneqq \langle \boldsymbol{A}_i, \boldsymbol{M}_k^\star \rangle - \langle \boldsymbol{a}_i, \boldsymbol{\beta}_k \rangle = \langle \boldsymbol{A}_i, \boldsymbol{M}_k^\star - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{M}_k^\star \boldsymbol{V}\boldsymbol{V}^\top \rangle \tag{42}$$

accounts for the subspace estimation error. In words, the observations $\{y_i\}$ can be equivalently written in the mixed linear regression form, where $\{\boldsymbol{\beta}_k\}$ constitutes the underlying parameters, $\{\boldsymbol{a}_i\}$ the measurement vectors and $\{\xi_i\}$ the measurement noise. We then focus on characterizing the properties of $\boldsymbol{a}_i$ and ξ_i.

Recall from Algorithm 2 that $\boldsymbol{a}_i = \mathsf{vec}(\boldsymbol{U}^\top \boldsymbol{A}_i \boldsymbol{V})$. In view of the independence between $\{\boldsymbol{A}_i\}$ and $\boldsymbol{U}, \boldsymbol{V}$, one can deduce that

$$\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d), \quad 1 \leq i \leq N,$$

where $d \coloneqq R^2$. Again, leveraging the independence between $\{\boldsymbol{A}_i, \zeta_i\}$ and $\boldsymbol{U}, \boldsymbol{V}$, we have

$$\xi_i = \langle \boldsymbol{A}_i, \boldsymbol{M}_k^\star - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{M}_k^\star \boldsymbol{V}\boldsymbol{V}^\top \rangle + \zeta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0, \|\boldsymbol{M}_k^\star - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{M}_k^\star \boldsymbol{V}\boldsymbol{V}^\top\|_{\mathrm{F}}^2 + \sigma^2\big).$$

For notational convenience, we shall denote the variance to be

$$\sigma_k^2 \coloneqq \|\boldsymbol{M}_k^\star - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{M}_k^\star \boldsymbol{V}\boldsymbol{V}^\top\|_{\mathrm{F}}^2 + \sigma^2, \quad 1 \leq k \leq K. \tag{43}$$

More importantly, the measurement vectors $\{\boldsymbol{a}_i\}$ are independent of the measurement noise $\{\xi_i\}$. To see this, one has

$$\begin{aligned}
\mathbb{E}[\xi_i \boldsymbol{a}_i] &= \mathbb{E}[\zeta_i \boldsymbol{a}_i] + \mathbb{E}[z_i \boldsymbol{a}_i] = \boldsymbol{0} + \mathsf{vec}\left(\mathbb{E}[\langle \boldsymbol{A}_i, \boldsymbol{M}_k^\star - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{M}_k^\star \boldsymbol{V}\boldsymbol{V}^\top \rangle \boldsymbol{U}^\top \boldsymbol{A}_i \boldsymbol{V}]\right) \\
&= \mathsf{vec}\left(\boldsymbol{U}^\top \left(\boldsymbol{M}_k^\star - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{M}_k^\star \boldsymbol{V}\boldsymbol{V}^\top\right) \boldsymbol{V}\right) = \boldsymbol{0}.
\end{aligned}$$

Here the second equality follows from the independence between ζ_i and $\boldsymbol{A}_i, \boldsymbol{U}, \boldsymbol{V}$, whereas the last line utilizes the independence between $\boldsymbol{A}_i$ and $\boldsymbol{U}, \boldsymbol{V}$ and the isotropic property of $\boldsymbol{A}_i$.

In conclusion, in Line 3 of Algorithm 2, we are equivalently faced with a d-dimensional mixed linear regression problem with data $\{\boldsymbol{a}_i, y_i\}_{1 \leq i \leq N}$, which satisfies that for $i \in \Omega_k^\star$,

$$y_i = \left\langle \boldsymbol{a}_i, \boldsymbol{\beta}_k \right\rangle + \xi_i, \qquad \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_k^2\right), \ \boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d) \tag{44}$$

with ξ_i being independent from $\boldsymbol{a}_i$.

Step 2: performance of the tensor method. Next, we characterize the performance of the tensor method for solving the above mixed linear regression problem. Our proof follows closely that of [YCS16, Theorem 1], with minor modifications to accommodate the noise $\{\xi_i\}$. Therefore we only provide a sketch here.

Recall that in Algorithm 5, we randomly split the input data $\{\boldsymbol{a}_i, y_i\}_{1 \leq i \leq N}$ into two sets $\{\boldsymbol{a}_i, y_i\}_{1 \leq i \leq N_1}$ and $\{\boldsymbol{a}_i', y_i'\}_{1 \leq i \leq N_2}$ (with slight abuse of notation). This sample splitting strategy is adopted merely to decouple statistical dependence and facilitate analysis. The high-level idea of the proof of [YCS16, Theorem 1] is simple to state: if the quantities

$$\left\|\boldsymbol{M}_2 - \sum_{k=1}^K p_k \boldsymbol{\beta}_k \boldsymbol{\beta}_k^\top\right\| \quad \text{and} \quad \left\|\left(\boldsymbol{M}_3 - \sum_{k=1}^K p_k \boldsymbol{\beta}_k{}^{\otimes 3}\right)(\boldsymbol{W}, \boldsymbol{W}, \boldsymbol{W})\right\| \tag{45}$$

are sufficiently small, then the tensor method returns reliable estimates of $\{\boldsymbol{\beta}_k\}$; see [YCS16, Eq. (24) in Section 5.4.1]. Here, the empirical moments $\boldsymbol{M}_2, \boldsymbol{M}_3$ and the whitening matrix $\boldsymbol{W}$ are defined in Algorithm 5.

With this connection in place, it suffices to control the quantities in (45). While the analysis in [YCS16, Section 5.4.2] only applies to the noiseless mixed linear regression problem, we can easily modify it to accommodate our noisy case (44). The trick is to *augment* $\{\boldsymbol{\beta}_k\}$ and $\{\boldsymbol{a}_i\}$ as follows:

$$\boldsymbol{\beta}_k^{\mathsf{aug}} \coloneqq \begin{bmatrix}\boldsymbol{\beta}_k \\ \sigma_k\end{bmatrix} \in \mathbb{R}^{d+1}, \quad 1 \le k \le K; \quad \boldsymbol{a}_i^{\mathsf{aug}} \coloneqq \begin{bmatrix}\boldsymbol{a}_i \\ \xi_i/\sigma_k\end{bmatrix} \in \mathbb{R}^{d+1}, \quad i \in \Omega_k^\star. \tag{46}$$

The advantage is clear: the noisy mixed linear regression problem (44) can be equivalently phrased as a noiseless one, that is for all $i \in \Omega_k^\star$,

$$\boldsymbol{a}_i^{\mathsf{aug}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d+1}) \qquad \text{and} \qquad y_i = \big\langle \boldsymbol{a}_i^{\mathsf{aug}}, \boldsymbol{\beta}_k^{\mathsf{aug}} \big\rangle. \tag{47}$$

Similarly, we can define $\boldsymbol{a}_i^{\mathsf{aug}'}$ analogously, and introduce the augmented versions of the empirical moments as follows:

$$m_0^{\mathsf{aug}} \coloneqq \frac{1}{N_1}\sum_{i=1}^{N_1} y_i^2 \in \mathbb{R}, \quad \boldsymbol{m}_1^{\mathsf{aug}} \coloneqq \frac{1}{6N_2}\sum_{i=1}^{N_2} y_i'^3 \boldsymbol{a}_i^{\mathsf{aug}'} \in \mathbb{R}^{d+1}, \tag{48a}$$

$$\boldsymbol{M}_2^{\mathsf{aug}} \coloneqq \frac{1}{2N_1}\sum_{i=1}^{N_1} y_i^2 \boldsymbol{a}_i^{\mathsf{aug}} (\boldsymbol{a}_i^{\mathsf{aug}})^\top - \frac{1}{2} m_0^{\mathsf{aug}} \boldsymbol{I}_{d+1} \in \mathbb{R}^{(d+1)\times(d+1)}, \tag{48b}$$

$$\boldsymbol{M}_3^{\mathsf{aug}} \coloneqq \frac{1}{6N_2}\sum_{i=1}^{N_2} y_i'^3 (\boldsymbol{a}_i^{\mathsf{aug}'})^{\otimes 3} - \mathcal{T}^{\mathsf{aug}}(\boldsymbol{m}_1^{\mathsf{aug}}) \in \mathbb{R}^{(d+1)\times(d+1)\times(d+1)}, \tag{48c}$$

where $\mathcal{T}^{\mathsf{aug}}(\cdot)$ is defined analogously as in (36). By virtue of the augmentation procedure, $\boldsymbol{M}_2$ (resp. $\boldsymbol{M}_3$) is a a sub-matrix (resp. sub-tensor) of $\boldsymbol{M}_2^{\mathsf{aug}}$ (resp. $\boldsymbol{M}_3^{\mathsf{aug}}$). Consequently, we have

$$\begin{aligned}
\Big\| \boldsymbol{M}_2 - \sum_{k=1}^K p_k \boldsymbol{\beta}_k \boldsymbol{\beta}_k^\top \Big\| &\le \Big\| \boldsymbol{M}_2^{\mathsf{aug}} - \sum_{k=1}^K p_k \boldsymbol{\beta}_k^{\mathsf{aug}} (\boldsymbol{\beta}_k^{\mathsf{aug}})^\top \Big\|; \\
\Big\| \Big(\boldsymbol{M}_3 - \sum_{k=1}^K p_k \boldsymbol{\beta}_k{}^{\otimes 3} \Big)(\boldsymbol{W}, \boldsymbol{W}, \boldsymbol{W}) \Big\| &= \Big\| \Big(\boldsymbol{M}_3^{\mathsf{aug}} - \sum_{k=1}^K p_k (\boldsymbol{\beta}_k^{\mathsf{aug}})^{\otimes 3} \Big)(\boldsymbol{W}^{\mathsf{aug}}, \boldsymbol{W}^{\mathsf{aug}}, \boldsymbol{W}^{\mathsf{aug}}) \Big\|,
\end{aligned}$$

where $\boldsymbol{W}^{\mathsf{aug}} \coloneqq [\boldsymbol{W}^\top, \boldsymbol{0}]^\top$.

With the above augmented vectors/matrices/tensors in place, one can follow the analysis in [YCS16, Section 5.4.2] to upper bound the quantities above. One subtle issue is that our sampling scheme is slightly different from the one in [YCS16], where each sample has i.i.d. labeling; nevertheless, it is easy to check that this difference is minor, and does not affect the result of the analysis. Indeed, repeating the analysis in [YCS16, Section 5.4] yields the conclusion that: in order to achieve ϵ errors (30) with probability at least $1-\gamma$, it suffices to require the sample complexities to exceed (analogous to [YCS16, Eq. (13)])

$$N_1 \ge C_1 \left(\frac{d}{(\min_k p_k)\,\epsilon^2} \frac{\max_k \big\|\boldsymbol{\beta}_k^{\mathsf{aug}}\big\|_2^{10}}{\sigma_K(\sum_k p_k \boldsymbol{\beta}_k \boldsymbol{\beta}_k^\top)^5} \log\frac{12K}{\gamma} \log^2 N_1 + \frac{K}{(\min_k p_k)\,\gamma} \right) \tag{49a}$$

$$\overset{\text{(i)}}{\asymp} \frac{dK^6}{\epsilon^2} \left(\Gamma^{10} + \frac{\sigma^{10}}{\min_k \|\boldsymbol{M}_k^\star\|_{\mathrm{F}}^{10}} \right) \log(K\log n) \log^2 N_1 + K^2 \log n, \tag{49b}$$

$$N_2 \ge C_2 \left(\frac{(K^2+d)}{(\min_k p_k)\,\epsilon^2} \frac{\max_k \big\|\boldsymbol{\beta}_k^{\mathsf{aug}}\big\|_2^{6}}{\sigma_K(\sum_k p_k \boldsymbol{\beta}_k \boldsymbol{\beta}_k^\top)^3} \log\frac{12K}{\gamma} \log^3 N_2 + \frac{K}{(\min_k p_k)\,\gamma} \right) \tag{49c}$$

$$\overset{\text{(ii)}}{\asymp} \frac{dK^4}{\epsilon^2} \left(\Gamma^{6} + \frac{\sigma^{6}}{\min_k \|\boldsymbol{M}_k^\star\|_{\mathrm{F}}^{6}} \right) \log(K\log n) \log^3 N_2 + K^2 \log n. \tag{49d}$$

Here $C_1, C_2 > 0$ are some sufficiently large constants, and the simplifications (i) (ii) hold due to the following facts: (i) $d = R^2 \ge K^2$, (ii) $\min_k p_k \asymp 1/K$, (iii) we choose $\gamma = O(1/\log n)$, (iv) $\|\boldsymbol{\beta}_k^{\mathsf{aug}}\|_2^2 = \|\boldsymbol{\beta}_k\|_2^2 + \sigma_k^2 = \|\boldsymbol{\beta}_k\|_2^2 + \|\boldsymbol{M}_k^\star - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{M}_k^\star \boldsymbol{V}\boldsymbol{V}^\top\|_{\mathrm{F}}^2 + \sigma^2$, and (v) the following claim (in particular, (51) and (52) therein).

Claim 2. *Instate the assumptions of Theorem 4.*

1. *The ground-truth matrices* $\{\boldsymbol{M}_k^\star\}_{1\le k\le K}$ *satisfy that for all* $1\le i,j\le K, i\neq j$,

$$\big|\langle \boldsymbol{M}_i^\star, \boldsymbol{M}_j^\star\rangle\big| \le \frac{1}{4K\Gamma^2}\|\boldsymbol{M}_i^\star\|_{\mathrm{F}}\|\boldsymbol{M}_j^\star\|_{\mathrm{F}}, \quad \text{and} \quad \big\|\boldsymbol{M}_i^\star - \boldsymbol{M}_j^\star\big\|_{\mathrm{F}} \gtrsim \|\boldsymbol{M}_i^\star\|_{\mathrm{F}} + \|\boldsymbol{M}_j^\star\|_{\mathrm{F}}. \tag{50}$$

2. *In addition, the parameters* $\{\boldsymbol{\beta}_k\}_{1\le k\le K}$ *obey that for all* $1\le k,i,j\le K, i\neq j$,

$$0.9\|\boldsymbol{M}_k^\star\|_{\mathrm{F}} \le \|\boldsymbol{\beta}_k\|_2 \le \|\boldsymbol{M}_k^\star\|_{\mathrm{F}}, \quad \text{and} \quad \big|\langle \boldsymbol{\beta}_i, \boldsymbol{\beta}_j\rangle\big| \le \frac{1}{2K\Gamma^2}\big\|\boldsymbol{\beta}_i\big\|_2\big\|\boldsymbol{\beta}_j\big\|_2. \tag{51}$$

3. *In the end, we have*

$$\sigma_K\left(\sum_{k=1}^K p_k\boldsymbol{\beta}_k\boldsymbol{\beta}_k^\top\right) \asymp \frac{1}{K}\min_{1\le k\le K}\|\boldsymbol{M}_k^\star\|_{\mathrm{F}}^2. \tag{52}$$

Armed with (49b) and (49d), we can plug in the bounds $d = R^2 \le K^2r^2$ and $\log(K\log n)\lesssim \log n$ to complete the proof of Theorem 4.

Proof of Claim 2. With regards to the first part of (50), it is seen that

$$\begin{aligned}\big|\langle \boldsymbol{M}_i^\star, \boldsymbol{M}_j^\star\rangle\big| &= \big|\langle \boldsymbol{U}_i^\star\boldsymbol{\Sigma}_i^\star\boldsymbol{V}_i^{\star\top}, \boldsymbol{U}_j^\star\boldsymbol{\Sigma}_j^\star\boldsymbol{V}_j^{\star\top}\rangle\big| = \big|\langle \boldsymbol{\Sigma}_i^\star, \boldsymbol{U}_i^{\star\top}\boldsymbol{U}_j^\star\boldsymbol{\Sigma}_j^\star\boldsymbol{V}_j^{\star\top}\boldsymbol{V}_i^\star\rangle\big| \le \|\boldsymbol{U}_i^{\star\top}\boldsymbol{U}_j^\star\|_{\mathrm{F}}\|\boldsymbol{V}_i^{\star\top}\boldsymbol{V}_j^\star\|_{\mathrm{F}}\|\boldsymbol{\Sigma}_i^\star\|_{\mathrm{F}}\|\boldsymbol{\Sigma}_j^\star\|_{\mathrm{F}}\\ &\le \left(\frac{1}{2\sqrt{K}\Gamma}\right)^2\|\boldsymbol{\Sigma}_i^\star\|_{\mathrm{F}}\|\boldsymbol{\Sigma}_j^\star\|_{\mathrm{F}} = \frac{1}{4K\Gamma^2}\|\boldsymbol{M}_i^\star\|_{\mathrm{F}}\|\boldsymbol{M}_j^\star\|_{\mathrm{F}},\end{aligned}$$

where the second line utilizes Assumption 1. The second part of (50) follows immediately from the first part and some elementary calculations.

Next, we turn to proving (51). Recall the definitions $\boldsymbol{\beta}_k = \mathsf{vec}(\boldsymbol{S}_k) = \mathsf{vec}(\boldsymbol{U}^\top\boldsymbol{M}_k^\star\boldsymbol{V})$ and $\mathsf{dist}_{\boldsymbol{U},\boldsymbol{V}} = \max\{\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\|, \|\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{V}^\star\boldsymbol{V}^{\star\top}\|\}$. We have the upper bound $\|\boldsymbol{\beta}_k\|_2 = \|\boldsymbol{U}^\top\boldsymbol{M}_k^\star\boldsymbol{V}\|_{\mathrm{F}} \le \|\boldsymbol{M}_k^\star\|_{\mathrm{F}}$ as well as the lower bound

$$\begin{aligned}\big\|\boldsymbol{\beta}_k\big\|_2 = \big\|\boldsymbol{U}^\top\boldsymbol{M}_k^\star\boldsymbol{V}\big\|_{\mathrm{F}} &= \|\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{M}_k^\star\boldsymbol{V}\boldsymbol{V}^\top\|_{\mathrm{F}} \ge \|\boldsymbol{M}_k^\star\|_{\mathrm{F}} - \|\boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\boldsymbol{M}_k^\star\boldsymbol{V}^\star\boldsymbol{V}^{\star\top} - \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{M}_k^\star\boldsymbol{V}\boldsymbol{V}^\top\|_{\mathrm{F}}\\ &\ge \|\boldsymbol{M}_k^\star\|_{\mathrm{F}} - \|(\boldsymbol{U}^\star\boldsymbol{U}^{\star\top} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{M}_k^\star\boldsymbol{V}^\star\boldsymbol{V}^{\star\top}\|_{\mathrm{F}} - \|\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{M}_k^\star(\boldsymbol{V}^\star\boldsymbol{V}^{\star\top} - \boldsymbol{V}\boldsymbol{V}^\top)\|_{\mathrm{F}}\\ &\ge (1 - 2\,\mathsf{dist}_{\boldsymbol{U},\boldsymbol{V}})\|\boldsymbol{M}_k^\star\|_{\mathrm{F}} \ge 0.9\|\boldsymbol{M}_k^\star\|_{\mathrm{F}},\end{aligned}$$

where the last inequality uses the assumption that $\mathsf{dist}_{\boldsymbol{U},\boldsymbol{V}} \le c_1/(K\Gamma^2) \le 0.05$; this justifies the first part of (51). To prove the second part of (51), we start with the decomposition

$$\begin{aligned}\big\langle \boldsymbol{\beta}_i, \boldsymbol{\beta}_j\big\rangle &= \big\langle \boldsymbol{U}^\top\boldsymbol{M}_i^\star\boldsymbol{V}, \boldsymbol{U}^\top\boldsymbol{M}_j^\star\boldsymbol{V}\big\rangle = \big\langle \boldsymbol{M}_i^\star, \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{M}_j^\star\boldsymbol{V}\boldsymbol{V}^\top\big\rangle\\ &= \big\langle \boldsymbol{M}_i^\star, \boldsymbol{M}_j^\star\big\rangle + \big\langle \boldsymbol{M}_i^\star, \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{M}_j^\star\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\boldsymbol{M}_j^\star\boldsymbol{V}\boldsymbol{V}^\top\big\rangle\\ &\quad + \big\langle \boldsymbol{M}_i^\star, \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\boldsymbol{M}_j^\star\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\boldsymbol{M}_j^\star\boldsymbol{V}^\star\boldsymbol{V}^{\star\top}\big\rangle,\end{aligned}$$

which together with the triangle inequality yields

$$\big|\big\langle \boldsymbol{\beta}_i, \boldsymbol{\beta}_j\big\rangle\big| \le |\langle \boldsymbol{M}_i^\star, \boldsymbol{M}_j^\star\rangle| + 2\,\mathsf{dist}_{\boldsymbol{U},\boldsymbol{V}}\|\boldsymbol{M}_i^\star\|_{\mathrm{F}}\|\boldsymbol{M}_j^\star\|_{\mathrm{F}}.$$

In light of the first part of (50), the first part of (51), and our assumption on $\mathsf{dist}_{\boldsymbol{U},\boldsymbol{V}}$, this establishes the second part of (51).

Finally, it remains to prove (52). In view of the assumption that $p_k \asymp 1/K$ $(1\le k\le K)$, one has

$$\sigma_K\left(\sum_{k=1}^K p_k\boldsymbol{\beta}_k\boldsymbol{\beta}_k^\top\right) \asymp \frac{1}{K}\sigma_K\left(\sum_{k=1}^K \boldsymbol{\beta}_k\boldsymbol{\beta}_k^\top\right). \tag{53}$$

Therefore, it suffices to show that $\sigma_K(\sum_k \boldsymbol{\beta}_k\boldsymbol{\beta}_k^\top) \asymp \min_k\|\boldsymbol{M}_k^\star\|_{\mathrm{F}}^2$. Towards this, we find it helpful to define $\boldsymbol{B} \coloneqq [\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K] \in \mathbb{R}^{d\times K}$, and decompose $\boldsymbol{B}^\top\boldsymbol{B}$ as $\boldsymbol{B}^\top\boldsymbol{B} = \boldsymbol{D} + \boldsymbol{O}$. Here, $\boldsymbol{D}$ stands for the diagonal

part of $\boldsymbol{B}^\top\boldsymbol{B}$ with $D_{kk} = \|\boldsymbol{\beta}_k\|_2^2$, while $\boldsymbol{O}$ is the off-diagonal part of $\boldsymbol{B}^\top\boldsymbol{B}$. Note that for any $i \neq j$, $[\boldsymbol{O}]_{ij} = [\boldsymbol{B}^\top\boldsymbol{B}]_{ij} = \langle \boldsymbol{\beta}_i, \boldsymbol{\beta}_j \rangle$, which combined with (51) gives

$$\|\boldsymbol{O}\|_{\mathrm{F}}^2 = \sum_{i\neq j} \langle \boldsymbol{\beta}_i, \boldsymbol{\beta}_j \rangle^2 \le K^2 \left(\frac{1}{2K\Gamma^2}\right)^2 \max_k \|\boldsymbol{\beta}_k\|_2^4 = \frac{1}{4\Gamma^4} \max_k \|\boldsymbol{\beta}_k\|_2^4 \le \frac{1}{2} \min_k \|\boldsymbol{\beta}_k\|_2^4;$$

the last inequality follows from the definition $\Gamma = \max_k \|\boldsymbol{M}_k^\star\|_{\mathrm{F}} / \min_k \|\boldsymbol{M}_k^\star\|_{\mathrm{F}}$, and the first part of (51). This together with Weyl's inequality implies that

$$\left|\sigma_K\left(\boldsymbol{B}^\top\boldsymbol{B}\right) - \sigma_K(\boldsymbol{D})\right| = \left|\sigma_K\left(\boldsymbol{B}^\top\boldsymbol{B}\right) - \min_k \|\boldsymbol{\beta}_k\|_2^2\right| \le \|\boldsymbol{O}\|_{\mathrm{F}} \le \frac{1}{\sqrt{2}} \min_k \|\boldsymbol{\beta}_k\|_2^2. \tag{54}$$

As a result, we arrive at

$$\sigma_K\left(\sum_k \boldsymbol{\beta}_k\boldsymbol{\beta}_k^\top\right) = \sigma_K\left(\boldsymbol{B}\boldsymbol{B}^\top\right) = \sigma_K\left(\boldsymbol{B}^\top\boldsymbol{B}\right) \asymp \min_k \|\boldsymbol{\beta}_k\|_2^2 \asymp \min_k \|\boldsymbol{M}_k^\star\|_{\mathrm{F}}^2,$$

which in conjunction with (53) completes the proof of (52). □

B.3 Proof of Proposition 1

To begin with, the triangle inequality gives

$$\|\boldsymbol{L}_k\boldsymbol{R}_k^\top - \boldsymbol{M}_k^\star\|_{\mathrm{F}} \le \|\boldsymbol{U}\boldsymbol{S}_k\boldsymbol{V}^\top - \boldsymbol{M}_k^\star\|_{\mathrm{F}} + \|\boldsymbol{L}_k\boldsymbol{R}_k^\top - \boldsymbol{U}\boldsymbol{S}_k\boldsymbol{V}^\top\|_{\mathrm{F}}. \tag{55}$$

Regarding the first term on the right-hand side of (55), we plug in the definition (13) of $\boldsymbol{S}_k$ to obtain

$$\begin{aligned}
\|\boldsymbol{U}\boldsymbol{S}_k\boldsymbol{V}^\top - \boldsymbol{M}_k^\star\|_{\mathrm{F}} &= \|\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{M}_k^\star\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\boldsymbol{M}_k^\star\boldsymbol{V}^\star\boldsymbol{V}^{\star\top}\|_{\mathrm{F}} \\
&\le \|(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top})\boldsymbol{M}_k^\star\boldsymbol{V}\boldsymbol{V}^\top\|_{\mathrm{F}} + \|\boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\boldsymbol{M}_k^\star(\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{V}^\star\boldsymbol{V}^{\star\top})\|_{\mathrm{F}} \\
&\le 2\max\left\{\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\|, \|\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{V}^\star\boldsymbol{V}^{\star\top}\|\right\}\|\boldsymbol{M}_k^\star\|_{\mathrm{F}}.
\end{aligned}$$

With regards to the second term on the right-hand side of (55), we observe that

$$\begin{aligned}
\|\boldsymbol{L}_k\boldsymbol{R}_k^\top - \boldsymbol{U}\boldsymbol{S}_k\boldsymbol{V}^\top\|_{\mathrm{F}} &\le \|\boldsymbol{L}_k\boldsymbol{R}_k^\top - \boldsymbol{U}\widehat{\boldsymbol{S}}_k\boldsymbol{V}^\top\|_{\mathrm{F}} + \|\boldsymbol{U}\widehat{\boldsymbol{S}}_k\boldsymbol{V}^\top - \boldsymbol{U}\boldsymbol{S}_k\boldsymbol{V}^\top\|_{\mathrm{F}} \\
&\overset{\text{(i)}}{\le} 2\|\boldsymbol{U}(\widehat{\boldsymbol{S}}_k - \boldsymbol{S}_k)\boldsymbol{V}^\top\|_{\mathrm{F}} \overset{\text{(ii)}}{\le} 2\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_2.
\end{aligned}$$

Here, (i) follows since $\boldsymbol{L}_k\boldsymbol{R}_k^\top$ is the best rank-r_k approximation of $\boldsymbol{U}\widehat{\boldsymbol{S}}_k\boldsymbol{V}^\top$ and $\boldsymbol{U}\boldsymbol{S}_k\boldsymbol{V}^\top$ is also rank-r_k; (ii) holds since $\widehat{\boldsymbol{\beta}}_k = \mathsf{vec}(\widehat{\boldsymbol{S}}_k)$ and $\boldsymbol{\beta}_k = \mathsf{vec}(\boldsymbol{S}_k)$. Substitution into (55) establishes (32).

B.4 Proof of Theorem 5

We shall only prove the local convergence w.r.t. the matrix $\boldsymbol{M}_1^\star$; the proof for other components is identical and hence is omitted. Our proof is decomposed into three steps.

1. Study the ScaledTGD dynamics (particularly the population-level dynamics), and control the effects of mislabeling and finite-sample errors.
2. Show that if the estimation error is larger than the error floor (namely, the last term in (34)), then one step of the ScaledTGD update contracts the error by a constant factor.
3. Show that, once the estimation error gets smaller than this error floor, then the estimation errors remain small in subsequent iterations.

Before continuing, we note that Condition (33) with $k = 1$ implies the existence of some constant $c_1 > 0$ such that

$$(\boldsymbol{L}^0, \boldsymbol{R}^0) \in \mathcal{B},$$

where

$$\mathcal{B} \coloneqq \left\{ (\boldsymbol{L}, \boldsymbol{R}) \in \mathbb{R}^{n_1 \times r_1} \times \mathbb{R}^{n_2 \times r_1} : \left\| \boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{M}_1^\star \right\|_{\mathrm{F}} \le c_1 \min \left\{ \sigma_{r_1}(\boldsymbol{M}_1^\star),\ \frac{1}{K} \min_{j \neq 1} \left\| \boldsymbol{M}_j^\star - \boldsymbol{M}_1^\star \right\|_{\mathrm{F}} \right\} \right\}. \tag{56}$$

This arises from the inequalities $\sigma_{r_1}(\boldsymbol{M}_1^\star) \ge \|\boldsymbol{M}_1^\star\|_{\mathrm{F}}/(\sqrt{r}\kappa)$ and $\min_{j\neq 1} \|\boldsymbol{M}_j^\star - \boldsymbol{M}_1^\star\|_{\mathrm{F}} \gtrsim \|\boldsymbol{M}_1^\star\|_{\mathrm{F}}$ (due to Assumption 1). We isolate Condition (56) since it is more convenient to work with in the analysis.

Notation. To simplify presentation, we shall often let $(\boldsymbol{L}, \boldsymbol{R})$ denote an iterate lying within $\mathcal{B}$ (cf. (56)), and define the corresponding estimation errors as

$$\boldsymbol{\Delta}_k \coloneqq \boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{M}_k^\star, \quad 1 \le k \le K. \tag{57}$$

The truncating level for a prescribed truncating fraction α is denoted by

$$\tau \coloneqq Q_\alpha \left(\left\{ \left| \langle \boldsymbol{A}_i, \boldsymbol{L}\boldsymbol{R}^\top \rangle - y_i \right| \right\}_{1 \le i \le N} \right), \tag{58}$$

where Q_α is the α-quantile defined in Section 1.2. We also define the following functions and quantities:

$$\mathbb{1}(a; b) \coloneqq \mathbb{1}(|a| \le b), \quad a, b \in \mathbb{R}, \tag{59}$$

$$w(x) \coloneqq \int_{-x}^{x} t^2 \phi(t)\, \mathrm{d}t, \quad x \ge 0, \qquad w_k \coloneqq w\left(\frac{\tau}{\sqrt{\|\boldsymbol{\Delta}_k\|_{\mathrm{F}}^2 + \sigma^2}} \right), \quad 1 \le k \le K, \tag{60}$$

where ϕ stands for the probability density function of a standard Gaussian random variable.

Step 1: characterizing the ScaledTGD dynamic. The above notation allows one to express the ScaledTGD update rule (16) as

$$\boldsymbol{L}^+ = \boldsymbol{L} - \frac{\eta}{N} \sum_{i=1}^{N} \big(\langle \boldsymbol{A}_i, \boldsymbol{L}\boldsymbol{R}^\top \rangle - y_i \big) \mathbb{1}\big(\langle \boldsymbol{A}_i, \boldsymbol{L}\boldsymbol{R}^\top \rangle - y_i; \tau \big) \boldsymbol{A}_i \boldsymbol{R} \big(\boldsymbol{R}^\top \boldsymbol{R} \big)^{-1}, \tag{61a}$$

$$\boldsymbol{R}^+ = \boldsymbol{R} - \frac{\eta}{N} \sum_{i=1}^{N} \big(\langle \boldsymbol{A}_i, \boldsymbol{L}\boldsymbol{R}^\top \rangle - y_i \big) \mathbb{1}\big(\langle \boldsymbol{A}_i, \boldsymbol{L}\boldsymbol{R}^\top \rangle - y_i; \tau \big) \boldsymbol{A}_i^\top \boldsymbol{L} \big(\boldsymbol{L}^\top \boldsymbol{L} \big)^{-1}. \tag{61b}$$

Recall that for any $i \in \Omega_k^\star$, we have $y_i = \langle \boldsymbol{A}_i, \boldsymbol{M}_k^\star \rangle + \zeta_i$, and thus

$$\langle \boldsymbol{A}_i, \boldsymbol{L}\boldsymbol{R}^\top \rangle - y_i = \langle \boldsymbol{A}_i, \boldsymbol{\Delta}_k \rangle - \zeta_i, \qquad \text{for all } i \in \Omega_k^\star. \tag{62}$$

The following result makes apparent a useful decomposition of the ScaledTGD update rule.

Claim 3. *Recall the notation (59) and (60). The ScaledTGD update rule (61) can be written as*

$$\boldsymbol{L}^+ = \boldsymbol{L}_{\mathsf{pop}}^+ - \eta \boldsymbol{E}_{\boldsymbol{L}}, \qquad \boldsymbol{R}^+ = \boldsymbol{R}_{\mathsf{pop}}^+ - \eta \boldsymbol{E}_{\boldsymbol{R}}. \tag{63}$$

Here, $(\boldsymbol{L}_{\mathsf{pop}}^+, \boldsymbol{R}_{\mathsf{pop}}^+)$ represents the population-level update from $\Omega_1^\star$

$$\boldsymbol{L}_{\mathsf{pop}}^+ \coloneqq \boldsymbol{L} - \eta p_1 w_1 \boldsymbol{\Delta}_1 \boldsymbol{R} (\boldsymbol{R}^\top \boldsymbol{R})^{-1}, \qquad \boldsymbol{R}_{\mathsf{pop}}^+ \coloneqq \boldsymbol{R} - \eta p_1 w_1 \boldsymbol{\Delta}_1^\top \boldsymbol{L} (\boldsymbol{L}^\top \boldsymbol{L})^{-1}, \tag{64}$$

and the residual components are given by

$$\boldsymbol{E}_{\boldsymbol{L}} \coloneqq (\boldsymbol{\Delta}_{\mathsf{mis}} + \boldsymbol{\Delta}_{\mathsf{fs}}) \boldsymbol{R} (\boldsymbol{R}^\top \boldsymbol{R})^{-1}, \qquad \boldsymbol{E}_{\boldsymbol{R}} \coloneqq (\boldsymbol{\Delta}_{\mathsf{mis}} + \boldsymbol{\Delta}_{\mathsf{fs}})^\top \boldsymbol{L} (\boldsymbol{L}^\top \boldsymbol{L})^{-1}$$

with

$$\boldsymbol{\Delta}_{\mathsf{mis}} \coloneqq \sum_{k \neq 1} p_k w_k \boldsymbol{\Delta}_k, \quad \boldsymbol{\Delta}_{\mathsf{fs}} \coloneqq \sum_{k=1}^{K} p_k \left(\frac{1}{|\Omega_k^\star|} \sum_{i \in \Omega_k^\star} \big(\langle \boldsymbol{A}_i, \boldsymbol{\Delta}_k \rangle - \zeta_i \big) \mathbb{1}\big(\langle \boldsymbol{A}_i, \boldsymbol{\Delta}_k \rangle - \zeta_i; \tau \big) \boldsymbol{A}_i - w_k \boldsymbol{\Delta}_k \right). \tag{65}$$

Before moving on, we note that it is crucial to control the sizes of $\boldsymbol{\Delta}_{\mathsf{mis}}$ and $\boldsymbol{\Delta}_{\mathsf{fs}}$, where "$\mathsf{mis}$" stands for "mislabeling", and "fs" stands for "finite sample". Regarding $\boldsymbol{\Delta}_{\mathsf{mis}}$, Fact 1 tells us that for all $k \neq 1$,

$$w_k\|\boldsymbol{\Delta}_k\|_{\mathrm{F}} = w_1\frac{w_k}{w_1}\|\boldsymbol{\Delta}_k\|_{\mathrm{F}} \le w_1\frac{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2+\sigma^2}{\|\boldsymbol{\Delta}_k\|_{\mathrm{F}}^2+\sigma^2}\|\boldsymbol{\Delta}_k\|_{\mathrm{F}} \le \frac{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2+\sigma^2}{\|\boldsymbol{\Delta}_k\|_{\mathrm{F}}} \le 2\frac{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2+\sigma^2}{\|\boldsymbol{M}_k^\star-\boldsymbol{M}_1^\star\|_{\mathrm{F}}}.$$

Here, the last inequality holds since

$$\|\boldsymbol{\Delta}_k\|_{\mathrm{F}} = \|\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{M}_k^\star\|_{\mathrm{F}} \ge \|\boldsymbol{M}_1^\star-\boldsymbol{M}_k^\star\|_{\mathrm{F}} - \|\boldsymbol{\Delta}_1\|_{\mathrm{F}} \ge 0.5\|\boldsymbol{M}_1^\star-\boldsymbol{M}_k^\star\|_{\mathrm{F}},$$

where we have used $\|\boldsymbol{\Delta}_1\|_{\mathrm{F}} \le c_2\sigma_{r_1}(\boldsymbol{M}_1^\star) \le 0.5\|\boldsymbol{M}_1^\star-\boldsymbol{M}_k^\star\|_{\mathrm{F}}$ due to the assumption that $(\boldsymbol{L},\boldsymbol{R}) \in \mathcal{B}$ defined in (56). Consequently, we obtain

$$\|\boldsymbol{\Delta}_{\mathsf{mis}}\|_{\mathrm{F}} = \Big\|\sum_{k\neq 1} p_k w_k \boldsymbol{\Delta}_k\Big\|_{\mathrm{F}} \le \sum_{k\neq 1} p_k w_k\|\boldsymbol{\Delta}_k\|_{\mathrm{F}} \le 2\frac{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2+\sigma^2}{\min_{k\neq 1}\|\boldsymbol{M}_k^\star-\boldsymbol{M}_1^\star\|_{\mathrm{F}}}. \tag{66}$$

Next, we turn to the term $\boldsymbol{\Delta}_{\mathsf{fs}}$. Note that $\mathsf{rank}(\boldsymbol{\Delta}_k) \le 2r$. Therefore, Lemmas 1 and 2 (see Remark 2) imply that, with probability at least $1-Ce^{-cn}$ for some constants $c, C > 0$, the following holds simultaneously for all $(\boldsymbol{L},\boldsymbol{R}) \in \mathcal{B}$ (cf. (56)):

1. the truncating level τ obeys
$$0.54 < \frac{\tau}{\sqrt{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2+\sigma^2}} < 1.35; \tag{67}$$
2. for any real matrix $\boldsymbol{W}$ with n_2 rows and of rank at most r, we have
$$\|\boldsymbol{\Delta}_{\mathsf{fs}}\boldsymbol{W}\|_{\mathrm{F}} \le \sum_{k=1}^K p_k\delta\tau\|\boldsymbol{W}\| = \delta\tau\|\boldsymbol{W}\| \le 1.35\delta\sqrt{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2+\sigma^2}\|\boldsymbol{W}\|. \tag{68}$$

The above-mentioned bounds will play a useful role in subsequent steps.

Step 2: per-iteration improvement above the error floor (34). Let us look at the Euclidean error

$$\big\|\boldsymbol{L}^+(\boldsymbol{R}^+)^\top-\boldsymbol{M}_1^\star\big\|_{\mathrm{F}} = \big\|(\boldsymbol{L}_{\mathsf{pop}}^+-\eta\boldsymbol{E}_{\boldsymbol{L}})(\boldsymbol{R}_{\mathsf{pop}}^+-\eta\boldsymbol{E}_{\boldsymbol{R}})^\top-\boldsymbol{M}_1^\star\big\|_{\mathrm{F}} \tag{69a}$$

$$\le \big\|\boldsymbol{L}_{\mathsf{pop}}^+(\boldsymbol{R}_{\mathsf{pop}}^+)^\top-\boldsymbol{M}_1^\star\big\|_{\mathrm{F}} + \eta\Big(\big\|\boldsymbol{E}_{\boldsymbol{L}}(\boldsymbol{R}_{\mathsf{pop}}^+)^\top\big)\big\|_{\mathrm{F}} + \big\|\boldsymbol{L}_{\mathsf{pop}}^+(\boldsymbol{E}_{\boldsymbol{R}})^\top\big\|_{\mathrm{F}} + \eta\big\|\boldsymbol{E}_{\boldsymbol{L}}(\boldsymbol{E}_{\boldsymbol{R}})^\top\big\|_{\mathrm{F}}\Big). \tag{69b}$$

Since $\boldsymbol{L}_{\mathsf{pop}}^+$ and $\boldsymbol{R}_{\mathsf{pop}}^+$ (64) are exactly the same as the update rule of scaled gradient descent for low-rank matrix factorization, [TMC20, Theorem 3] tells us that if $0 < \eta p_1 w_1 \le 2/3$ (which holds true under our choices of $\eta \le 1.3/p_1$ and $\alpha \le 0.8p_1$), then

$$\big\|\boldsymbol{L}_{\mathsf{pop}}^+(\boldsymbol{R}_{\mathsf{pop}}^+)^\top-\boldsymbol{M}_1^\star\big\|_{\mathrm{F}} \le \big(1-0.7\eta p_1 w_1\big)\big\|\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{M}_1^\star\big\|_{\mathrm{F}}. \tag{70}$$

It remains to control the perturbation terms in (69b), accomplished as follows.

Claim 4. *Denoting*

$$B \coloneqq 2\left(\delta\sqrt{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2+\sigma^2} + \frac{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2+\sigma^2}{\min_{k\neq 1}\|\boldsymbol{M}_k^\star-\boldsymbol{M}_1^\star\|_{\mathrm{F}}}\right), \tag{71}$$

one has

$$\max\Big\{\big\|\boldsymbol{E}_{\boldsymbol{L}}(\boldsymbol{R}_{\mathsf{pop}}^+)^\top\big\|_{\mathrm{F}}, \big\|\boldsymbol{L}_{\mathsf{pop}}^+(\boldsymbol{E}_{\boldsymbol{R}})^\top\big\|_{\mathrm{F}}\Big\} \le 2B, \quad \big\|\boldsymbol{E}_{\boldsymbol{L}}(\boldsymbol{E}_{\boldsymbol{R}})^\top\big\|_{\mathrm{F}} \le \frac{2}{\sigma_{r_1}(\boldsymbol{M}_1^\star)}B^2. \tag{72}$$

Putting (70) and (72) back to (69) and denoting $\boldsymbol{\Delta}_1^+ \coloneqq \boldsymbol{L}^+(\boldsymbol{R}^+)^\top-\boldsymbol{M}_1^\star$, we have

$$\|\boldsymbol{\Delta}_1^+\|_{\mathrm{F}} \le \big(1-0.7\eta p_1 w_1\big)\|\boldsymbol{\Delta}_1\|_{\mathrm{F}} + \eta\left(4B + \frac{2\eta}{\sigma_{r_1}(\boldsymbol{M}_1^\star)}B^2\right). \tag{73}$$

It remains to control B. First, the relations $\delta \le c_0/K$ and $\|\boldsymbol{\Delta}_1\|_{\mathrm{F}} \ge C_2 K\delta\sigma$ (for some sufficiently large constant $C_2 > 0$) imply that

$$\delta\sqrt{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}} + \sigma^2} \le \delta\|\boldsymbol{\Delta}_1\|_{\mathrm{F}} + \delta\sigma \le \frac{c_3}{K}\|\boldsymbol{\Delta}_1\|_{\mathrm{F}} \tag{74}$$

for some sufficiently small constant $c_3 > 0$. Moreover, observing that

$$\frac{C_2 K\sigma^2}{\min_{k\neq 1}\|\boldsymbol{M}_k^\star - \boldsymbol{M}_1^\star\|_{\mathrm{F}}} \le \|\boldsymbol{\Delta}_1\|_{\mathrm{F}} \le \frac{c_1}{K}\min_{k\neq 1}\|\boldsymbol{M}_k^\star - \boldsymbol{M}_1^\star\|_{\mathrm{F}},$$

we have

$$\frac{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2 + \sigma^2}{\min_{k\neq 1}\|\boldsymbol{M}_k^\star - \boldsymbol{M}_1^\star\|_{\mathrm{F}}} = \frac{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2}{\min_{k\neq 1}\|\boldsymbol{M}_k^\star - \boldsymbol{M}_1^\star\|_{\mathrm{F}}} + \frac{\sigma^2}{\min_{k\neq 1}\|\boldsymbol{M}_k^\star - \boldsymbol{M}_1^\star\|_{\mathrm{F}}} \le \frac{c_4}{K}\|\boldsymbol{\Delta}_1\|_{\mathrm{F}} \tag{75}$$

for some sufficiently small constant $c_4 > 0$. Putting (74) and (75) back into (71), we have

$$B \le \frac{2(c_3 + c_4)}{K}\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}, \tag{76}$$

which together with $\|\boldsymbol{\Delta}_1\|_{\mathrm{F}} \le c_1\sigma_{r_1}(\boldsymbol{M}_1^\star)$ implies the existence of some small constant $c_5 > 0$ such that

$$4B + \frac{2\eta}{\sigma_{r_1}(\boldsymbol{M}_1^\star)}B^2 = 4B\left(1 + \frac{\eta}{2\sigma_{r_1}(\boldsymbol{M}_1^\star)}B\right) \le 8B \le \frac{c_5}{K}\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}.$$

Substituting this into (73), we arrive at the desired bound

$$\|\boldsymbol{\Delta}_1^+\|_{\mathrm{F}} \le (1 - c_2\eta p_1)\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}$$

for some constant $c_2 > 0$; this is because in (73), we have $p_1 \asymp 1/K$ by assumption, and $w_1 \gtrsim 1$ according to (67).

Step 3: no blowing up below the error floor (34). Suppose that the estimation error satisfies

$$\|\boldsymbol{\Delta}_1\|_{\mathrm{F}} \lesssim \max\left\{K\sigma\delta, \frac{K\sigma^2}{\min_{k\neq 1}\|\boldsymbol{M}_k^\star - \boldsymbol{M}_1^\star\|_{\mathrm{F}}}\right\}. \tag{77}$$

We intend to show that, in this case, the estimation error of the next iterate $\|\boldsymbol{\Delta}_1^+\|_{\mathrm{F}}$ satisfies the same upper bound (77); if this claim were true, then combining this with our results in Step 2 would complete the convergence analysis of ScaledTGD.

Note that (73) remains valid when $\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}$ is below the error floor, which implies that

$$\|\boldsymbol{\Delta}_1^+\|_{\mathrm{F}} \lesssim \|\boldsymbol{\Delta}_1\|_{\mathrm{F}} + KB\left(1 + \frac{KB}{\sigma_{r_1}(\boldsymbol{M}_1^\star)}\right). \tag{78}$$

Recalling the definition of B in (71), one has

$$KB \lesssim K\big(\|\boldsymbol{\Delta}_1\|_{\mathrm{F}} + \sigma\big)\left(\delta + \frac{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}} + \sigma}{\min_{k\neq 1}\|\boldsymbol{M}_k^\star - \boldsymbol{M}_1^\star\|_{\mathrm{F}}}\right).$$

By the assumption that $\delta \lesssim 1/K$ and $\sigma \lesssim \min_k\|\boldsymbol{M}_k^\star\|_{\mathrm{F}}/K$, we have $\|\boldsymbol{\Delta}_1\|_{\mathrm{F}} \lesssim \sigma$ according to (77), and thus $KB/\sigma_{r_1}(\boldsymbol{M}_1^\star) \lesssim \sigma/\sigma_{r_1}(\boldsymbol{M}_1^\star) \lesssim 1$. Consequently, on the right-hand side of (78) we have

$$KB\left(1 + \frac{KB}{\sigma_{r_1}(\boldsymbol{M}_1^\star)}\right) \lesssim KB \lesssim K\sigma\left(\delta + \frac{\sigma}{\min_{k\neq 1}\|\boldsymbol{M}_k^\star - \boldsymbol{M}_1^\star\|_{\mathrm{F}}}\right),$$

which has exactly the same form as the error floor in (77). This completes our proof for this step.

Proof of Claim 4. We shall only prove the first part of (72) concerning $\|\boldsymbol{E}_{\boldsymbol{L}}(\boldsymbol{R}_{\mathsf{pop}}^{+})^{\top}\|_{\mathrm{F}}$; the analysis for $\|\boldsymbol{L}_{\mathsf{pop}}^{+}(\boldsymbol{E}_{\boldsymbol{R}})^{\top}\|_{\mathrm{F}}$ is essentially the same. By the triangle inequality, we have

$$\big\|\boldsymbol{E}_{\boldsymbol{L}}(\boldsymbol{R}_{\mathsf{pop}}^{+})^{\top}\big\|_{\mathrm{F}} \le \big\|\boldsymbol{E}_{\boldsymbol{L}}\boldsymbol{R}^{\top}\big\|_{\mathrm{F}} + \eta p_1 w_1 \big\|\boldsymbol{E}_{\boldsymbol{L}}(\boldsymbol{L}^{\top}\boldsymbol{L})^{-1}\boldsymbol{L}^{\top}\boldsymbol{\Delta}_1\big\|_{\mathrm{F}}. \tag{79}$$

We utilize (66) and (68) from Step 1 to control the terms above. For the first term of (79), recognizing that $\|\boldsymbol{R}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{R}^{\top}\| \le 1$ we have

$$\begin{aligned}\big\|\boldsymbol{E}_{\boldsymbol{L}}\boldsymbol{R}^{\top}\big\|_{\mathrm{F}} &= \big\|(\boldsymbol{\Delta}_{\mathsf{fs}}+\boldsymbol{\Delta}_{\mathsf{mis}})\boldsymbol{R}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{R}^{\top}\big\|_{\mathrm{F}} \le \big\|\boldsymbol{\Delta}_{\mathsf{fs}}\boldsymbol{R}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{R}^{\top}\big\|_{\mathrm{F}} + \big\|\boldsymbol{\Delta}_{\mathsf{mis}}\big\|_{\mathrm{F}} \\ &\le 2\left(\delta\sqrt{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2+\sigma^2} + \frac{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2+\sigma^2}{\min_{k\neq 1}\|\boldsymbol{M}_k^{\star}-\boldsymbol{M}_1^{\star}\|_{\mathrm{F}}}\right) = B.\end{aligned}$$

Regarding the second term of (79), we observe that

$$\begin{aligned}\big\|\boldsymbol{E}_{\boldsymbol{L}}(\boldsymbol{L}^{\top}\boldsymbol{L})^{-1}\boldsymbol{L}^{\top}\boldsymbol{\Delta}_1\big\|_{\mathrm{F}} &= \big\|(\boldsymbol{\Delta}_{\mathsf{fs}}+\boldsymbol{\Delta}_{\mathsf{mis}})\boldsymbol{R}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}(\boldsymbol{L}^{\top}\boldsymbol{L})^{-1}\boldsymbol{L}^{\top}\boldsymbol{\Delta}_1\big\|_{\mathrm{F}} \\ &\le \Big(\big\|\boldsymbol{\Delta}_{\mathsf{fs}}\boldsymbol{R}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}(\boldsymbol{L}^{\top}\boldsymbol{L})^{-1}\boldsymbol{L}^{\top}\big\|_{\mathrm{F}} + \big\|\boldsymbol{\Delta}_{\mathsf{mis}}\big\|_{\mathrm{F}}\big\|\boldsymbol{R}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}(\boldsymbol{L}^{\top}\boldsymbol{L})^{-1}\boldsymbol{L}^{\top}\big\|\Big)\|\boldsymbol{\Delta}_1\|_{\mathrm{F}} \\ &\overset{\text{(i)}}{\le} 2\left(\delta\sqrt{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2+\sigma^2} + \frac{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2+\sigma^2}{\min_{k\neq 1}\|\boldsymbol{M}_k^{\star}-\boldsymbol{M}_1^{\star}\|_{\mathrm{F}}}\right)\frac{2}{\sigma_{r_1}(\boldsymbol{M}_1^{\star})}\cdot c_1\sigma_{r_1}(\boldsymbol{M}_1^{\star}) \\ &= 4c_1\left(\delta\sqrt{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2+\sigma^2} + \frac{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2+\sigma^2}{\min_{k\neq 1}\|\boldsymbol{M}_k^{\star}-\boldsymbol{M}_1^{\star}\|_{\mathrm{F}}}\right) = 2c_1 B,\end{aligned}$$

where (i) follows from $\|\boldsymbol{\Delta}_1\|_{\mathrm{F}} \le c_1\sigma_{r_1}(\boldsymbol{M}_1^{\star})$ (see (56)) as well as the following fact (which will be proved at the end of this section): for any $\boldsymbol{L}\in\mathbb{R}^{n_1\times r_1}$ and $\boldsymbol{R}\in\mathbb{R}^{n_2\times r_1}$,

$$\text{if}\quad \big\|\boldsymbol{L}\boldsymbol{R}^{\top}-\boldsymbol{M}_1^{\star}\big\|_{\mathrm{F}} \le \frac{\sigma_{r_1}(\boldsymbol{M}_1^{\star})}{2},\quad \text{then}\quad \big\|\boldsymbol{L}(\boldsymbol{L}^{\top}\boldsymbol{L})^{-1}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{R}^{\top}\big\| \le \frac{2}{\sigma_{r_1}(\boldsymbol{M}_1^{\star})}. \tag{80}$$

Combining these with (79) establishes that $\|\boldsymbol{E}_{\boldsymbol{L}}(\boldsymbol{R}_{\mathsf{pop}}^{+})^{\top}\|_{\mathrm{F}} \le 2B$, which is the first part of (72).

Finally, for the second part of (72), we can apply similar techniques to reach

$$\begin{aligned}\big\|\boldsymbol{E}_{\boldsymbol{L}}(\boldsymbol{E}_{\boldsymbol{R}})^{\top}\big\|_{\mathrm{F}} &= \big\|(\boldsymbol{\Delta}_{\mathsf{fs}}+\boldsymbol{\Delta}_{\mathsf{mis}})\boldsymbol{R}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}(\boldsymbol{L}^{\top}\boldsymbol{L})^{-1}\boldsymbol{L}^{\top}(\boldsymbol{\Delta}_{\mathsf{fs}}+\boldsymbol{\Delta}_{\mathsf{mis}})\big\|_{\mathrm{F}} \\ &\le \frac{2}{\sigma_{r_1}(\boldsymbol{M}_1^{\star})}\cdot 4\left(\delta\sqrt{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2+\sigma^2} + \frac{\|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2+\sigma^2}{\min_{k\neq 1}\|\boldsymbol{M}_k^{\star}-\boldsymbol{M}_1^{\star}\|_{\mathrm{F}}}\right)^2 = \frac{2}{\sigma_{r_1}(\boldsymbol{M}_1^{\star})}B^2.\end{aligned}$$

□

Proof of (80). Weyl's inequality tells us that

$$\sigma_{r_1}\big(\boldsymbol{L}\boldsymbol{R}^{\top}\big) \ge \sigma_{r_1}\big(\boldsymbol{M}_1^{\star}\big) - \big\|\boldsymbol{L}\boldsymbol{R}^{\top}-\boldsymbol{M}_1^{\star}\big\|_{\mathrm{F}} \ge \frac{\sigma_{r_1}\big(\boldsymbol{M}_1^{\star}\big)}{2}, \tag{81}$$

which further implies that both $\boldsymbol{L}$ and $\boldsymbol{R}$ have full column rank r_1. Consequently, we denote the SVD of $\boldsymbol{L}$ and $\boldsymbol{R}$ as $\boldsymbol{L}=\boldsymbol{U}_L\boldsymbol{\Sigma}_L\boldsymbol{V}_L^{\top}$ and $\boldsymbol{R}=\boldsymbol{U}_R\boldsymbol{\Sigma}_R\boldsymbol{V}_R^{\top}$, where $\boldsymbol{V}_L,\boldsymbol{V}_R$ are $r_1\times r_1$ orthonormal matrices. With the SVD representations in place, it is easy to check that

$$\boldsymbol{L}\boldsymbol{R}^{\top} = \boldsymbol{U}_L\boldsymbol{\Sigma}_L\boldsymbol{V}_L^{\top}\boldsymbol{V}_R\boldsymbol{\Sigma}_R\boldsymbol{U}_R^{\top},\quad \text{and}\quad \boldsymbol{L}(\boldsymbol{L}^{\top}\boldsymbol{L})^{-1}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{R}^{\top} = \boldsymbol{U}_L\boldsymbol{\Sigma}_L^{-1}\boldsymbol{V}_L^{\top}\boldsymbol{V}_R\boldsymbol{\Sigma}_R^{-1}\boldsymbol{U}_R^{\top}.$$

In addition, the orthonormality of $\boldsymbol{V}_L$ and $\boldsymbol{V}_R$ implies

$$(\boldsymbol{\Sigma}_L\boldsymbol{V}_L^{\top}\boldsymbol{V}_R\boldsymbol{\Sigma}_R)^{-1} = \boldsymbol{\Sigma}_R^{-1}(\boldsymbol{V}_L^{\top}\boldsymbol{V}_R)^{-1}\boldsymbol{\Sigma}_L^{-1} = \boldsymbol{\Sigma}_R^{-1}\boldsymbol{V}_R^{\top}\boldsymbol{V}_L\boldsymbol{\Sigma}_L^{-1} = (\boldsymbol{\Sigma}_L^{-1}\boldsymbol{V}_L^{\top}\boldsymbol{V}_R\boldsymbol{\Sigma}_R^{-1})^{\top},$$

thus indicating that

$$\begin{aligned}\big\|\boldsymbol{L}(\boldsymbol{L}^{\top}\boldsymbol{L})^{-1}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{R}^{\top}\big\| &= \big\|\boldsymbol{U}_L\boldsymbol{\Sigma}_L^{-1}\boldsymbol{V}_L^{\top}\boldsymbol{V}_R\boldsymbol{\Sigma}_R^{-1}\boldsymbol{U}_R^{\top}\big\| = \big\|\boldsymbol{\Sigma}_L^{-1}\boldsymbol{V}_L^{\top}\boldsymbol{V}_R\boldsymbol{\Sigma}_R^{-1}\big\| \\ &= \big\|(\boldsymbol{\Sigma}_L\boldsymbol{V}_L^{\top}\boldsymbol{V}_R\boldsymbol{\Sigma}_R)^{-1}\big\| = \frac{1}{\sigma_{r_1}\big(\boldsymbol{\Sigma}_L\boldsymbol{V}_L^{\top}\boldsymbol{V}_R\boldsymbol{\Sigma}_R\big)} = \frac{1}{\sigma_{r_1}\big(\boldsymbol{L}\boldsymbol{R}^{\top}\big)}.\end{aligned}$$

Combining this with (81) completes the proof. □

C Technical lemmas

This section collects several technical lemmas that are helpful for our analysis (particularly for the analysis of Stage 3). For notational convenience, we define the set of low-rank matrices as

$$\mathcal{R}_r := \big\{ \boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2} : \mathsf{rank}(\boldsymbol{X}) \le r \big\}. \tag{82}$$

We remind the reader of the definitions $\mathbb{1}(a;b) = \mathbb{1}(|a| \le b)$ for $a, b \in \mathbb{R}$ and $w(x) = \int_{-x}^{x} t^2 \phi(t)\,\mathrm{d}t$ for $x \ge 0$.

Variants of matrix-RIP. We recall the standard notion of restricted isometry property (RIP) from the literature of matrix sensing, and introduce a variant called *truncated* RIP (TRIP).

Definition 2. Let $\{\boldsymbol{A}_i\}_{i=1}^m$ be a set of matrices in $\mathbb{R}^{n_1 \times n_2}$. Consider $1 \le r \le \min\{n_1, n_2\}$ and $0 < \delta < 1$.

1. We say that $\{\boldsymbol{A}_i\}_{1 \le i \le m}$ satisfy (r,δ)-RIP if

$$\left| \frac{1}{m} \sum_{i=1}^{m} \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle \langle \boldsymbol{A}_i, \boldsymbol{Z} \rangle - \langle \boldsymbol{X}, \boldsymbol{Z} \rangle \right| \le \delta \|\boldsymbol{X}\|_{\mathrm{F}} \|\boldsymbol{Z}\|_{\mathrm{F}} \tag{83}$$

   holds simultaneously for all $\boldsymbol{X}, \boldsymbol{Z} \in \mathcal{R}_r$.

2. We say that $\{\boldsymbol{A}_i\}_{1 \le i \le m}$ satisfy (r,δ)-TRIP if

$$\left| \frac{1}{m} \sum_{i=1}^{m} \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle \mathbb{1}\big(\langle \boldsymbol{A}_i, \boldsymbol{X} \rangle; \tau \|\boldsymbol{X}\|_{\mathrm{F}}\big) \langle \boldsymbol{A}_i, \boldsymbol{Z} \rangle - w(\tau) \langle \boldsymbol{X}, \boldsymbol{Z} \rangle \right| \le \delta \tau \|\boldsymbol{X}\|_{\mathrm{F}} \|\boldsymbol{Z}\|_{\mathrm{F}} \tag{84}$$

   holds simultaneously for all $\boldsymbol{X}, \boldsymbol{Z} \in \mathcal{R}_r$ and for all $0 \le \tau \le 1.35$.

As it turns out, the Gaussian design satisfies the above notion of RIP and TRIP, as formalized below.

Lemma 1. *Let $\{\boldsymbol{A}_i\}_{1 \le i \le m}$ be random matrices in $\mathbb{R}^{n_1 \times n_2}$ with i.i.d. $\mathcal{N}(0,1)$ entries, and denote $n := \max\{n_1, n_2\}$. There exist some sufficiently large constants $C_1, C_3 > 0$ and some other constants $C_2, c_2, C_4, c_4 > 0$ such that*

1. *If $m \ge C_1 nr\delta^{-2} \log(1/\delta)$, then with probability at least $1 - C_2 e^{-c_2 n}$, $\{\boldsymbol{A}_i\}_{1 \le i \le m}$ satisfy (r,δ)-RIP.*
2. *If $m \ge C_3 nr\delta^{-2} \log m$, then with probability at least $1 - C_4 e^{-c_4 n}$, $\{\boldsymbol{A}_i\}_{1 \le i \le m}$ satisfy (r,δ)-TRIP.*

Empirical quantiles. Our next technical lemma is a uniform concentration result for empirical quantiles. Given the design matrices $\{\boldsymbol{A}_i\}_{1 \le i \le N}$, the index sets $\{\Omega_k^\star\}_{1 \le k \le K}$ and the low-rank matrices $\{\boldsymbol{X}_k\}_{1 \le k \le K}$, we define several sets as follows:

$$\mathcal{D}_k := \Big\{ \big|\langle \boldsymbol{A}_i, \boldsymbol{X}_k \rangle\big| \Big\}_{i \in \Omega_k^\star}, \quad 1 \le k \le K; \qquad \mathcal{D} := \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K. \tag{85}$$

In addition, let us introduce the following set of low-rank matrices:

$$\mathcal{T}_1 := \Big\{ (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K) : \boldsymbol{X}_k \in \mathcal{R}_r, 1 \le k \le K; 0 < \|\boldsymbol{X}_1\|_{\mathrm{F}} \le \frac{c_0}{K} \min_{k \ne 1} \|\boldsymbol{X}_k\|_{\mathrm{F}} \Big\}, \tag{86}$$

where $c_0 > 0$ is some sufficiently small constant. Recall that $Q_\alpha(\mathcal{D})$ denotes the α-quantile of $\mathcal{D}$, as defined in (2).

Lemma 2. *Let $\{\boldsymbol{A}_i\}_{1 \le i \le N}$ be random matrices in $\mathbb{R}^{n_1 \times n_2}$ with i.i.d. $\mathcal{N}(0,1)$ entries. Set $n = \max\{n_1, n_2\}$, and suppose the index sets $\{\Omega_k^\star\}_{1 \le k \le K}$ are disjoint and satisfy the condition (18). If $0.6p_1 \le \alpha \le 0.8p_1$ and $N \ge C_0 nrK^3 \log N$ for some sufficiently large constant $C_0 > 0$, then there exist some universal constants $C, c > 0$ such that: with probability at least $1 - Ce^{-cn}$,*

$$0.54 < \frac{Q_\alpha(\mathcal{D})}{\|\boldsymbol{X}_1\|_{\mathrm{F}}} < 1.35$$

holds simultaneously for all $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K) \in \mathcal{T}_1$, where $\mathcal{D}$ is defined in (85).

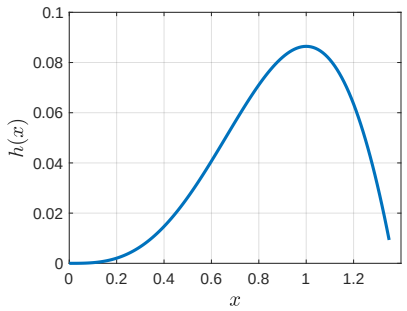


Figure 2: The function $h(\cdot)$ defined in (89) is nonnegative over the interval $(0, 1.35]$.

Remark 2. We can further incorporate additional Gaussian noise $\{\zeta_i\}$ into Lemmas 1 and 2, where $\zeta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. For example, we claim that, with the same sample complexity m as in Lemma 1, we have the following noisy version of (r, δ)-TRIP (84):

$$\left| \frac{1}{m} \sum_{i=1}^{m} \left(\langle \boldsymbol{A}_i, \boldsymbol{X} \rangle - \zeta_i \right) \mathbb{1} \left(\langle \boldsymbol{A}_i, \boldsymbol{X} \rangle - \zeta_i ; \tau \sqrt{\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \sigma^2} \right) \langle \boldsymbol{A}_i, \boldsymbol{Z} \rangle - w(\tau) \langle \boldsymbol{X}, \boldsymbol{Z} \rangle \right| \le \delta \tau \sqrt{\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \sigma^2} \|\boldsymbol{Z}\|_{\mathrm{F}}. \quad (87)$$

To see this, let us define the augmented matrices

$$\boldsymbol{X}^{\mathsf{aug}} := \begin{bmatrix} \boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{0} & -\sigma \end{bmatrix}, \quad \boldsymbol{Z}^{\mathsf{aug}} := \begin{bmatrix} \boldsymbol{Z} & \boldsymbol{0} \\ \boldsymbol{0} & 0 \end{bmatrix}, \quad \boldsymbol{A}_i^{\mathsf{aug}} := \begin{bmatrix} \boldsymbol{A}_i & * \\ * & \zeta_i / \sigma \end{bmatrix}, \quad 1 \le i \le m,$$

where $*$ stands for some auxiliary i.i.d. $\mathcal{N}(0,1)$ entries. Observe that $\{\boldsymbol{A}_i^{\mathsf{aug}}\}_{1 \le i \le N}$ are random matrices with i.i.d. $\mathcal{N}(0,1)$ entries; in addition, $\mathsf{rank}(\boldsymbol{X}^{\mathsf{aug}}) = \mathsf{rank}(\boldsymbol{X}) + 1$, $\mathsf{rank}(\boldsymbol{Z}^{\mathsf{aug}}) = \mathsf{rank}(\boldsymbol{Z})$, and $\|\boldsymbol{X}^{\mathsf{aug}}\|_{\mathrm{F}}^2 = \|\boldsymbol{X}\|_{\mathrm{F}}^2 + \sigma^2$; finally, $\langle \boldsymbol{A}_i, \boldsymbol{X} \rangle - \zeta_i = \langle \boldsymbol{A}_i^{\mathsf{aug}}, \boldsymbol{X}^{\mathsf{aug}} \rangle$, $\langle \boldsymbol{A}_i, \boldsymbol{Z} \rangle = \langle \boldsymbol{A}_i^{\mathsf{aug}}, \boldsymbol{Z}^{\mathsf{aug}} \rangle$, and $\langle \boldsymbol{X}, \boldsymbol{Z} \rangle = \langle \boldsymbol{X}^{\mathsf{aug}}, \boldsymbol{Z}^{\mathsf{aug}} \rangle$. Therefore, the left-hand side of (87) can be equivalently written as in the noiseless form (84), in terms of these augmented matrices, thus allowing us to apply Lemma 1 to prove (87). This trick of augmentation can be applied to Lemma 2 as well, which we omit here for brevity.

One miscellaneous result. Further, we record below a basic property concerning the function $w(\cdot)$.

Fact 1. *The function $w(\cdot)$ defined in (60) satisfies*

$$\frac{w(x)}{w(y)} \le \frac{x^2}{y^2}, \qquad 0 < x \le y \le 1.35. \quad (88)$$

Proof. This result is equivalent to saying $w(x)/x^2 \le w(y)/y^2$ for any $0 < x \le y \le 1.35$. Hence, it suffices to show that the function $g(x) := w(x)/x^2$ is nondecreasing over $(0, 1.35]$, or equivalently,

$$h(x) := \sqrt{\frac{2}{\pi}} x^3 e^{-\frac{x^2}{2}} - 2w(x), \quad g'(x) = \frac{1}{x^3} h(x) \ge 0, \quad 0 < x \le 1.35. \quad (89)$$

This can be verified numerically (see Figure 2), which completes the proof. □

The rest of this section is devoted to proving Lemmas 1 and 2. We use the standard notions (e.g. the subgaussian norm $\|\cdot\|_{\psi_2}$) and properties related to subgaussian random variables (cf. [Ver18, Section 2]). For notational convenience, we define the normalized version of $\mathcal{R}_r$ defined in (82), as follows:

$$\mathcal{R}_r^{\mathsf{norm}} := \left\{ \boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2} : \mathsf{rank}(\boldsymbol{X}) \le r, \|\boldsymbol{X}\|_{\mathrm{F}} = 1 \right\}. \quad (90)$$

Before moving on, we record two results that will be useful throughout the proof.

Lemma 3. *Let $\{\boldsymbol{A}_i\}_{i=1}^m$ be a set of random matrices in $\mathbb{R}^{n_1\times n_2}$ with i.i.d. $\mathcal{N}(0,1)$ entries. Denote $n \coloneqq \max\{n_1,n_2\}$, and let Z be a random variable having the same distribution as $|\mathcal{N}(0,1)|$. For all $t>0$ and $0<\epsilon<1$, with probability at least $1-(9/\epsilon)^{3nr}\exp(-c_1mt^2/(\tau+t))-C_2e^{-c_2n}$, the following*

$$\frac{1}{m}\sum_{i=1}^m \mathbb{1}\Big(|\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle|\le\tau\Big)\le \mathbb{P}\left(Z\le 1.01\tau\right)+t+\frac{200\epsilon}{\tau}$$

holds simultaneously for all $\boldsymbol{X}\in\mathcal{R}_r^{\mathsf{norm}}$, provided that $m\ge Cnr\log m$. Here, $c_1,C_2,c_2>0$ are universal constants, and $C>0$ is some sufficiently large constant.

Proposition 2. *Consider $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim}\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_d), 1\le i\le m$. There exist some universal constants $C,c>0$ such that with probability at least $1-Ce^{-cd}$, we have*

$$\max_{1\le i\le m}\|\boldsymbol{a}_i\|_2\lesssim\sqrt{d}+\sqrt{\log m}.$$

Proof. This result follows from [Ver18, Corollary 7.3.3] and the union bound. □

C.1 Proof of Lemma 1

The first result on RIP has been established in the literature (e.g. [CP11, Theorem 2.3]), and hence we only need to prove the second result on TRIP. We first restrict to the case

$$m^{-100}\le\tau\le 1.35;$$

at the end of this subsection, we will prove TRIP for the case $0\le\tau<m^{-100}$ separately. By homogeneity, it is sufficient to show that

$$\left|\frac{1}{m}\sum_{i=1}^m\langle\boldsymbol{A}_i,\boldsymbol{X}\rangle\mathbb{1}\big(\langle\boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau\big)\langle\boldsymbol{A}_i,\boldsymbol{Z}\rangle-w(\tau)\langle\boldsymbol{X},\boldsymbol{Z}\rangle\right|\le\delta\tau \tag{91}$$

holds simultaneously for all $(\boldsymbol{X},\boldsymbol{Z},\tau)\in\mathcal{T}_{\mathsf{TRIP}}$, where

$$\mathcal{T}_{\mathsf{TRIP}}\coloneqq\big\{(\boldsymbol{X},\boldsymbol{Z},\tau):\boldsymbol{X},\boldsymbol{Z}\in\mathcal{R}_r^{\mathsf{norm}},m^{-100}\le\tau\le 1.35\big\}.$$

The proof consists of two steps: (1) we replace the discontinuous function $\mathbb{1}$ by a Lipschitz continuous surrogate χ and establish a uniform concentration result for χ; (2) we show that the discrepancy incurred by replacing $\mathbb{1}$ with χ is uniformly small. Our proof argument is conditioned on the high-probability event that $\{\boldsymbol{A}_i\}_{i=1}^m$ satisfy $(2r,\delta)$-RIP.

Step 1: replacing $\mathbb{1}$ with χ. Define an auxiliary function χ as follows: for all $a\in\mathbb{R}$ and $\tau>0$,

$$\chi(a;\tau)\coloneqq\begin{cases}1, & |a|\le(1-c_\chi)\tau;\\ 0, & |a|\ge\tau;\\ \frac{\tau-|a|}{c_\chi\tau}, & (1-c_\chi)\tau<|a|<\tau.\end{cases} \tag{92}$$

Here we set the parameter

$$c_\chi\coloneqq c_0\delta^2m^{-100} \tag{93}$$

for some sufficiently small constant $c_0>0$, whose rationale will be made apparent in Step 2. It is easily seen that χ enjoys the following properties:

- (Continuity) For any $\tau>0$, $\chi(\cdot;\tau)$ is piecewise linear and $1/(c_\chi\tau)$-Lipschitz continuous.
- (Closeness to $\mathbb{1}$) For any $\tau>0$ and $a\in\mathbb{R}$, $\chi(a;\tau)\le\mathbb{1}(a;\tau)\le\chi(a;\tau/(1-c_\chi))$.
- (Homogeneity) For any $\tau>0$, $a\in\mathbb{R}$ and $c_0>0$, $\chi(a;\tau)=\chi(a/c_0;\tau/c_0)$.

- If $0 \le \epsilon_\tau \le c_\chi \tau$ and $\tau - \epsilon_\tau \le \tau_0 \le \tau$, then $\|\chi(\cdot;\tau) - \chi(\cdot;\tau_0)\|_\infty = \chi(\tau_0;\tau) = (\tau - \tau_0)/(c_\chi \tau) \le \epsilon_\tau/(c_\chi \tau)$.
- The function $f(a) \coloneqq a \cdot \chi(a;\tau)$ is $1/c_\chi$-Lipschitz continuous.

For notational convenience, define

$$E_m^\chi(\boldsymbol{X}, \boldsymbol{Z}, \tau) \coloneqq \frac{1}{m} \sum_{i=1}^m \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle \chi\big(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle; \tau\big) \langle \boldsymbol{A}_i, \boldsymbol{Z}\rangle, \tag{94a}$$

$$E^\chi(\boldsymbol{X}, \boldsymbol{Z}, \tau) \coloneqq \mathbb{E}\big[\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle \chi(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle; \tau) \langle \boldsymbol{A}_i, \boldsymbol{Z}\rangle\big], \tag{94b}$$

where the expectation is taken w.r.t. $\{\boldsymbol{A}_i\}$ while assuming that $(\boldsymbol{X}, \boldsymbol{Z}, \tau)$ are fixed. With these preparations in place, we set out to prove that: if $m \ge C_0 nr\delta^{-2} \log m$, then with probability at least $1 - Ce^{-cn}$,

$$\big|E_m^\chi(\boldsymbol{X}, \boldsymbol{Z}, \tau) - E^\chi(\boldsymbol{X}, \boldsymbol{Z}, \tau)\big| \le \delta\tau/2 \tag{95}$$

holds simultaneously for all $(\boldsymbol{X}, \boldsymbol{Z}, \tau) \in \mathcal{T}_{\mathsf{TRIP}}$; here $C_0 > 0$ is some sufficiently large constant, and $C, c > 0$ are some universal constants.

First, consider any fixed point $(\boldsymbol{X}, \boldsymbol{Z}, \tau) \in \mathcal{T}_{\mathsf{TRIP}}$. Note that $|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle \chi\big(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle; \tau\big)| \le \tau$ is bounded, and that the subgaussian norm of $\langle \boldsymbol{A}_i, \boldsymbol{Z}\rangle$ obeys $\|\langle \boldsymbol{A}_i, \boldsymbol{Z}\rangle\|_{\psi_2} = \|\mathcal{N}(0,1)\|_{\psi_2} \lesssim 1$. As a result,

$$\big\|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle \chi\big(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle; \tau\big) \langle \boldsymbol{A}_i, \boldsymbol{Z}\rangle - E^\chi(\boldsymbol{X}, \boldsymbol{Z}, \tau)\big\|_{\psi_2} \lesssim \tau.$$

Invoking [Ver18, Theorem 2.6.2] tells us that for all $t \ge 0$,

$$\mathbb{P}\Big(\big|E_m^\chi(\boldsymbol{X}, \boldsymbol{Z}, \tau) - E^\chi(\boldsymbol{X}, \boldsymbol{Z}, \tau)\big| \ge t\tau\Big) \le 2\exp\left(-c_1 m t^2\right)$$

holds for some constant $c_1 > 0$. Next, we construct an ϵ-net to cover $\mathcal{T}_{\mathsf{TRIP}}$. In view of [CP11, Lemma 3.1], the set $\mathcal{R}_r^{\mathsf{norm}}$ defined in (90) has an ϵ-net (in terms of $\|\cdot\|_{\mathrm{F}}$ distance) of cardinality at most $(9/\epsilon)^{3nr}$. In addition, we can cover the interval $[m^{-100}, 1.35]$ with precision ϵ_τ using no more than $2/\epsilon_\tau$ equidistant points. Putting all this together, we can construct a set $\mathcal{M}_{\mathsf{TRIP}} \subseteq \mathcal{R}_r^{\mathsf{norm}} \times \mathcal{R}_r^{\mathsf{norm}} \times [0, 1.35]$ of cardinality at most $(9/\epsilon)^{6nr}(2/\epsilon_\tau)$ such that: for any $(\boldsymbol{X}, \boldsymbol{Z}, \tau) \in \mathcal{T}_{\mathsf{TRIP}}$, there exists some point $(\boldsymbol{X}_0, \boldsymbol{Z}_0, \tau_0) \in \mathcal{M}_{\mathsf{TRIP}}$ obeying

$$\|\boldsymbol{X} - \boldsymbol{X}_0\|_{\mathrm{F}} \le \epsilon, \quad \|\boldsymbol{Z} - \boldsymbol{Z}_0\|_{\mathrm{F}} \le \epsilon, \quad \text{and} \quad \tau - \epsilon_\tau \le \tau_0 \le \tau. \tag{96}$$

The union bound then implies that with probability at least $1 - 2\exp(-c_1 m t^2)(9/\epsilon)^{6nr}(2/\epsilon_\tau)$, one has

$$\big|E_m^\chi - E^\chi(\boldsymbol{X}, \boldsymbol{Z}, \tau)\big| \le t\tau, \qquad \text{for all } (\boldsymbol{X}, \boldsymbol{Z}, \tau) \in \mathcal{M}_{\mathsf{TRIP}}. \tag{97}$$

In what follows, we shall choose

$$t = \frac{1}{4}\delta \quad \text{and} \quad m \ge C_3 \frac{1}{\delta^2}\left(nr \log\frac{9}{\epsilon} + \log\frac{2}{\epsilon_\tau}\right) \tag{98}$$

so as to achieve a uniformly small error $t\tau = \delta\tau/4$ in (97) with probability at least $1 - 2\exp(-c_3 m\delta^2)$ for some universal constant $c_3 > 0$.

Now, for any $(\boldsymbol{X}, \boldsymbol{Z}, \tau) \in \mathcal{T}_{\mathsf{TRIP}}$, let $(\boldsymbol{X}_0, \boldsymbol{Z}_0, \tau_0) \in \mathcal{M}_{\mathsf{TRIP}}$ be the point satisfying (96). Then we have

$$\big|E_m^\chi(\boldsymbol{X}, \boldsymbol{Z}, \tau) - E^\chi(\boldsymbol{X}, \boldsymbol{Z}, \tau)\big| \le \underbrace{\big|E_m^\chi(\boldsymbol{X}_0, \boldsymbol{Z}_0, \tau_0) - E^\chi(\boldsymbol{X}_0, \boldsymbol{Z}_0, \tau_0)\big|}_{\text{(A)}} \tag{99a}$$

$$+ \underbrace{\big|E_m^\chi(\boldsymbol{X}, \boldsymbol{Z}, \tau) - E_m^\chi(\boldsymbol{X}_0, \boldsymbol{Z}_0, \tau_0)\big|}_{\text{(B)}} + \underbrace{\big|E^\chi(\boldsymbol{X}, \boldsymbol{Z}, \tau) - E^\chi(\boldsymbol{X}_0, \boldsymbol{Z}_0, \tau_0)\big|}_{\text{(C)}}. \tag{99b}$$

Here, (A) is already bounded by $\delta\tau/4$ by construction. We can control (B) via the following decomposition:

$$\text{(B)} \le \underbrace{\left|\frac{1}{m}\sum_{i=1}^m \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle \chi(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle; \tau) \langle \boldsymbol{A}_i, \boldsymbol{Z} - \boldsymbol{Z}_0\rangle\right|}_{\text{(B.1)}}$$

$$+\underbrace{\left|\frac{1}{m}\sum_{i=1}^{m}\Big(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)-\langle \boldsymbol{A}_i,\boldsymbol{X}_0\rangle\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}_0\rangle;\tau)\Big)\langle \boldsymbol{A}_i,\boldsymbol{Z}_0\rangle\right|}_{\text{(B.2)}}$$

$$+\underbrace{\left|\frac{1}{m}\sum_{i=1}^{m}\langle \boldsymbol{A}_i,\boldsymbol{X}_0\rangle\Big(\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}_0\rangle;\tau)-\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}_0\rangle;\tau_0)\Big)\langle \boldsymbol{A}_i,\boldsymbol{Z}_0\rangle\right|}_{\text{(B.3)}}.$$

In light of the $(2r,\delta)$-RIP, the aforementioned properties of χ, and the Cauchy-Schwarz inequality, we have

$$\text{(B.1)} \overset{\text{(i)}}{\le} \tau\frac{1}{m}\sum_{i=1}^{m}\big|\langle \boldsymbol{A}_i,\boldsymbol{Z}-\boldsymbol{Z}_0\rangle\big| \le \tau\sqrt{\frac{1}{m}\sum_{i=1}^{m}\langle \boldsymbol{A}_i,\boldsymbol{Z}-\boldsymbol{Z}_0\rangle^2} \lesssim \tau\epsilon,$$

$$\begin{aligned}\text{(B.2)} &\le \frac{1}{m}\sum_{i=1}^{m}\big|\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)-\langle \boldsymbol{A}_i,\boldsymbol{X}_0\rangle\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}_0\rangle;\tau)\big|\cdot\big|\langle \boldsymbol{A}_i,\boldsymbol{Z}_0\rangle\big| \\ &\overset{\text{(ii)}}{\le} \frac{1}{c_\chi}\frac{1}{m}\sum_{i=1}^{m}\big|\langle \boldsymbol{A}_i,\boldsymbol{X}-\boldsymbol{X}_0\rangle\big|\cdot\big|\langle \boldsymbol{A}_i,\boldsymbol{Z}_0\rangle\big| \le \frac{1}{c_\chi}\sqrt{\frac{1}{m}\sum_{i=1}^{m}\langle \boldsymbol{A}_i,\boldsymbol{X}-\boldsymbol{X}_0\rangle^2}\sqrt{\frac{1}{m}\sum_{i=1}^{m}\langle \boldsymbol{A}_i,\boldsymbol{Z}_0\rangle^2} \lesssim \frac{\epsilon}{c_\chi},\end{aligned}$$

$$\begin{aligned}\text{(B.3)} &\le \big\|\chi(\cdot;\tau)-\chi(\cdot;\tau_0)\big\|_\infty\frac{1}{m}\sum_{i=1}^{m}\big|\langle \boldsymbol{A}_i,\boldsymbol{X}_0\rangle\big|\cdot\big|\langle \boldsymbol{A}_i,\boldsymbol{Z}_0\rangle\big| \\ &\overset{\text{(iii)}}{\lesssim} \frac{\epsilon_\tau}{c_\chi\tau}\sqrt{\frac{1}{m}\sum_{i=1}^{m}\langle \boldsymbol{A}_i,\boldsymbol{X}_0\rangle^2}\sqrt{\frac{1}{m}\sum_{i=1}^{m}\langle \boldsymbol{A}_i,\boldsymbol{Z}_0\rangle^2} \lesssim \frac{\epsilon_\tau}{c_\chi\tau}.\end{aligned}$$

Here, (i) uses $|\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)|\le\tau$, (ii) follows from the property that the function $f(a)=a\cdot\chi(a;\tau)$ is $1/c_\chi$-Lipschitz continuous, whereas (iii) is due to the property $\|\chi(\cdot;\tau)-\chi(\cdot;\tau_0)\|_\infty\le\epsilon_\tau/(c_\chi\tau)$. The term (C) can be controlled by the same decomposition and thus enjoys the same upper bound. Putting these back into (99), we have for all $(\boldsymbol{X},\boldsymbol{Z},\tau)\in\mathcal{T}_{\mathsf{TRIP}}$,

$$|E_m^\chi(\boldsymbol{X},\boldsymbol{Z},\tau)-E^\chi(\boldsymbol{X},\boldsymbol{Z},\tau)| \le \frac{1}{4}\delta\tau + C_3\left(\tau\epsilon+\frac{\epsilon}{c_\chi}+\frac{\epsilon_\tau}{c_\chi\tau}\right)$$

for some universal constant $C_3>0$. Recalling that $\tau\ge m^{-100}$, and choosing $\epsilon\le c_4\delta c_\chi m^{-100}$ and $\epsilon_\tau\le c_5\delta c_\chi m^{-200}$ for some sufficiently small constants $c_4,c_5>0$, we have

$$\big|E_m^\chi(\boldsymbol{X},\boldsymbol{Z},\tau)-E^\chi(\boldsymbol{X},\boldsymbol{Z},\tau)\big| \le \delta\tau/2.$$

Plugging our choice of ϵ and ϵ_τ into (98) immediately establishes the claim (95) of this step.

Step 2: controlling the errors incurred by using the surrogate χ. Similar to (94), we define

$$\begin{aligned}E_m(\boldsymbol{X},\boldsymbol{Z},\tau) &\coloneqq \frac{1}{m}\sum_{i=1}^{m}\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\mathbb{1}(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle, \\ E(\boldsymbol{X},\boldsymbol{Z},\tau) &\coloneqq \mathbb{E}\big[\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\mathbb{1}(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle\big] = w(\tau)\langle \boldsymbol{X},\boldsymbol{Z}\rangle,\end{aligned}$$

where the expectation is taken assuming independence between $\boldsymbol{A}_i$ and $(\boldsymbol{X},\boldsymbol{Z},\tau)$. In this step, we aim to show that: if $m\ge C_0nr\delta^{-2}\log m$, then with probability at least $1-Ce^{-cn}$,

$$\big|E_m(\boldsymbol{X},\boldsymbol{Z},\tau)-E(\boldsymbol{X},\boldsymbol{Z},\tau)\big| \le \big|E_m^\chi(\boldsymbol{X},\boldsymbol{Z},\tau)-E^\chi(\boldsymbol{X},\boldsymbol{Z},\tau)\big|+\delta\tau/2 \tag{100}$$

holds simultaneously for all $(\boldsymbol{X},\boldsymbol{Z},\tau)\in\mathcal{T}_{\mathsf{TRIP}}$. If this were true, then combining this with (95) would immediately conclude the proof of Lemma 1.

Towards establishing (100), we start with the following decomposition:

$$|E_m(\boldsymbol{X},\boldsymbol{Z},\tau)-E(\boldsymbol{X},\boldsymbol{Z},\tau)| \le \big|E_m^\chi(\boldsymbol{X},\boldsymbol{Z},\tau)-E^\chi(\boldsymbol{X},\boldsymbol{Z},\tau)\big| + \underbrace{\big|E(\boldsymbol{X},\boldsymbol{Z},\tau)-E^\chi(\boldsymbol{X},\boldsymbol{Z},\tau)\big|}_{\text{(A)}} + \underbrace{\big|E_m(\boldsymbol{X},\boldsymbol{Z},\tau)-E_m^\chi(\boldsymbol{X},\boldsymbol{Z},\tau)\big|}_{\text{(B)}}, \tag{101}$$

where we abuse the notation (A) and (B). In the sequel, we shall control (A) and (B) separately.

- Regarding (A), the Cauchy-Schwarz inequality gives

$$\begin{aligned}\text{(A)} &= \Big|\mathbb{E}\big[\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\big(\mathbb{1}(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)-\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)\big)\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle\big]\Big| \\ &\le \sqrt{\mathbb{E}\Big[\big(\mathbb{1}(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)-\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)\big)^2\Big]}\sqrt{\mathbb{E}\Big[\big(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle\big)^2\Big]} \lesssim \sqrt{c_\chi\tau}.\end{aligned}$$

  The last inequality holds since $|\mathbb{1}(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)-\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)| \in [0,1]$ is non-zero only for $|\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle|$ on an interval of length $c_\chi\tau$, over which the probability density function of $\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle \sim \mathcal{N}(0,1)$ is upper bounded by some constant. By our choice of c_χ in (93), we have (A) $\le \delta\tau/4$.

- We then move on to (B). For notational convenience, given any $\tau>0$, we let

$$\tau' = \tau'(\tau) \coloneqq \frac{\tau}{1-c_\chi}, \tag{102}$$

  which clearly satisfies $\chi(a;\tau) \le \mathbb{1}(a;\tau) \le \chi(a;\tau')$. In addition, defining

$$\mathbb{1}_-(a) \coloneqq \mathbb{1}(a<0), \quad \mathbb{1}_+(a) \coloneqq \mathbb{1}(a\ge 0), \quad a\in\mathbb{R},$$

  we can deduce that

$$\begin{aligned}E_m(\boldsymbol{X},\boldsymbol{Z},\tau) &\le E_m^\chi(\boldsymbol{X},\boldsymbol{Z},\tau) + \frac{1}{m}\sum_{i=1}^m \big(\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau')-\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)\big)\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle\mathbb{1}_+(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle), \\ E_m(\boldsymbol{X},\boldsymbol{Z},\tau) &\ge E_m^\chi(\boldsymbol{X},\boldsymbol{Z},\tau) + \frac{1}{m}\sum_{i=1}^m \big(\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau')-\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)\big)\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle\mathbb{1}_-(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle).\end{aligned}$$

  As a consequence,

$$\text{(B)} \le \max\Big\{\underbrace{\Big|\frac{1}{m}\sum_{i=1}^m \big(\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau')-\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)\big)\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle\mathbb{1}_+(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle)\Big|}_{\text{(C)}}, \\ \Big|\frac{1}{m}\sum_{i=1}^m \big(\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau')-\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)\big)\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle\mathbb{1}_-(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle)\Big|\Big\}.$$

  Next, we demonstrate how to analyze the first term (C) above; the analysis for the other term is essentially the same. For notational simplicity, define

$$\begin{aligned}F_m^+(\boldsymbol{X},\boldsymbol{Z},\tau) &\coloneqq \frac{1}{m}\sum_{i=1}^m \chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle\mathbb{1}_+(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle), \\ E^+(\boldsymbol{X},\boldsymbol{Z},\tau) &\coloneqq \mathbb{E}\Big[\big(\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau')-\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)\big)\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle\mathbb{1}_+(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle)\Big],\end{aligned}$$

  where the expectation is again taken assuming that $\boldsymbol{A}_i$ is independent of $\boldsymbol{X}$, $\boldsymbol{Z}$ and τ. Then we have

$$\text{(C)} = \big|F_m^+(\boldsymbol{X},\boldsymbol{Z},\tau')-F_m^+(\boldsymbol{X},\boldsymbol{Z},\tau)\big| \le \big|E^+(\boldsymbol{X},\boldsymbol{Z},\tau)\big| + \big|F_m^+(\boldsymbol{X},\boldsymbol{Z},\tau')-F_m^+(\boldsymbol{X},\boldsymbol{Z},\tau)-E^+(\boldsymbol{X},\boldsymbol{Z},\tau)\big|.$$

Regarding the first term on the right-hand side, we follow an argument similar to our previous analysis for (A) to obtain

$$\left|E^+(\boldsymbol{X},\boldsymbol{Z},\tau)\right| \le \sqrt{\mathbb{E}\Big[\big(\chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau') - \chi(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau)\big)^2\Big]}\sqrt{\mathbb{E}\Big[\big(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle\big)^2\Big]} \lesssim \sqrt{c_\chi \tau} \le c_2\delta\tau$$

for some sufficiently small constant $0 < c_2 < 1/8$. Thus, it remains to show that

$$\left|F_m^+(\boldsymbol{X},\boldsymbol{Z},\tau') - F_m^+(\boldsymbol{X},\boldsymbol{Z},\tau) - E^+(\boldsymbol{X},\boldsymbol{Z},\tau)\right| \le \frac{1}{8}\delta\tau \tag{103}$$

holds simultaneously for all $(\boldsymbol{X},\boldsymbol{Z},\tau) \in \mathcal{T}_{\mathsf{TRIP}}$. Note that by definition, $F_m^+(\boldsymbol{X},\boldsymbol{Z},\tau)$ is the empirical average of some Lipschitz continuous function (in particular, $\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle \mathbb{1}_+(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle)$ is 1-Lipschitz continuous over $\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle$). Therefore, we can prove (103) by a standard covering argument similar to that in Step 1; we omit the details for brevity. Putting the above bounds together, we establish that (B) $\le \delta\tau/4$.

- Combining the above bounds (A) $\le \delta\tau/4$ and (B) $\le \delta\tau/4$ with (101), we finish the proof of (100).

Proof for the case $0 \le \tau < m^{-100}$. It remains to prove that (91) holds simultaneously for all $\boldsymbol{X},\boldsymbol{Z} \in \mathcal{R}_r^{\mathsf{norm}}$ (cf. (90)) and all $0 \le \tau < m^{-100}$. We start with the following decomposition:

$$\left|\frac{1}{m}\sum_{i=1}^m \langle \boldsymbol{A}_i,\boldsymbol{X}\rangle \mathbb{1}\big(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau\big)\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle - w(\tau)\langle \boldsymbol{X},\boldsymbol{Z}\rangle\right| \le \left|\frac{1}{m}\sum_{i=1}^m \langle \boldsymbol{A}_i,\boldsymbol{X}\rangle \mathbb{1}\big(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau\big)\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle\right| + \left|w(\tau)\langle \boldsymbol{X},\boldsymbol{Z}\rangle\right|. \tag{104}$$

The second term on the right-hand side of (104) can be bounded by

$$\left|w(\tau)\langle \boldsymbol{X},\boldsymbol{Z}\rangle\right| \le w(\tau) \overset{\text{(i)}}{\le} \tau^3 \le m^{-200}\tau \overset{\text{(ii)}}{\le} 0.1\delta\tau,$$

where (i) can be seen from the definition of $w(\cdot)$ in (60), and (ii) relies on the observation that our assumption $m \ge C_0 nr\delta^{-2}\log m$ implies $\delta \gtrsim m^{-1/2}$.

It thus remains to show that the first term on the right-hand side of (104) is bounded by $0.9\delta\tau$. In view of $(2r,\delta)$-RIP, the Cauchy-Schwarz inequality, and the observation that $|\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle \mathbb{1}\big(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau\big)| \le \tau$, we have

$$\begin{aligned}\left|\frac{1}{m}\sum_{i=1}^m \langle \boldsymbol{A}_i,\boldsymbol{X}\rangle \mathbb{1}\big(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau\big)\langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle\right| &\le \sqrt{\frac{1}{m}\sum_{i=1}^m \langle \boldsymbol{A}_i,\boldsymbol{X}\rangle^2 \mathbb{1}\big(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau\big)}\cdot\sqrt{\frac{1}{m}\sum_{i=1}^m \langle \boldsymbol{A}_i,\boldsymbol{Z}\rangle^2} \\ &\le 2\sqrt{\frac{1}{m}\sum_{i=1}^m \tau^2 \mathbb{1}\big(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;\tau\big)} \le 2\tau\sqrt{\frac{1}{m}\sum_{i=1}^m \mathbb{1}\big(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;m^{-100}\big)},\end{aligned} \tag{105}$$

where the last inequality uses the assumption that $\tau < m^{-100}$. We can invoke Lemma 3 with $t = 0.01\delta^2$ and $\epsilon = m^{-200}$ to obtain that with probability at least $1 - Ce^{-cn}$ (for some constants $c, C > 0$),

$$\frac{1}{m}\sum_{i=1}^m \mathbb{1}\big(\langle \boldsymbol{A}_i,\boldsymbol{X}\rangle;m^{-100}\big) \le \mathbb{P}(Z_0 \le 1.01m^{-100}) + t + \frac{200\epsilon}{m^{-100}} \le 2t = 0.02\delta^2$$

holds simultaneously for all $\boldsymbol{X} \in \mathcal{R}_r^{\mathsf{norm}}$, provided that $m \ge C_0 nr\delta^{-2}\log m$; here, Z_0 denotes a random variable having the same distribution as $|\mathcal{N}(0,1)|$. Plugging this into (105) confirms that the first term on the right-hand side of (104) is bounded by $0.9\delta\tau$, thus concluding the proof for the case with $0 \le \tau < m^{-100}$.

C.2 Proof of Lemma 2

It is easy to check that $Q_\alpha(\mathcal{D})/\|\boldsymbol{X}_1\|_{\mathrm{F}}$ is invariant under a global scaling of $\{\boldsymbol{X}_1,\dots,\boldsymbol{X}_K\}$. Therefore, it suffices to consider a normalized version of $\mathcal{T}_1$ (86) defined as follows:

$$\mathcal{T}_1^{\mathsf{norm}} \coloneqq \mathcal{T}_1 \cap \big\{(\boldsymbol{X}_1,\dots,\boldsymbol{X}_k) : \|\boldsymbol{X}_1\|_{\mathrm{F}} = 1\big\}. \tag{106}$$

In what follows, we shall treat the upper bound and the lower bound separately and invoke a standard covering argument to prove Lemma 2 with $\mathcal{T}_1$ replaced by $\mathcal{T}_1^{\mathsf{norm}}$. Throughout this proof, we denote by Z a random variable following the distribution of $|\mathcal{N}(0,1)|$.

Step 1: upper bounding $Q_\alpha(\mathcal{D})$. Since $\alpha \le 0.8p_1$, we have

$$Q_\alpha(\mathcal{D}) \le Q_{\alpha/p_1}(\mathcal{D}_1) \le Q_{0.8}(\mathcal{D}_1).$$

Now it suffices to upper bound $Q_{0.8}(\mathcal{D}_1)$, which is only related to $\boldsymbol{X}_1 \in \mathcal{R}_r^{\mathsf{norm}}$. Consider any fixed point $\boldsymbol{X}_1 \in \mathcal{R}_r^{\mathsf{norm}}$. Note that the set $\mathcal{D}_1$ defined in (85) contains i.i.d. samples having the same distribution as Z. This combined with the concentration of empirical quantiles [Ser09, Section 2.3.2] gives

$$\mathbb{P}\big(Q_{0.8}(\mathcal{D}_1) \ge Q_{0.8}(Z) + 0.01\big) \le \exp\left(-c_2 N_1\right) \tag{107}$$

for some universal constant $c_2 > 0$. Here, $N_1 \coloneqq |\Omega_1^\star| \asymp N/K$ by the assumption of the well-balancedness property (18). Next, we construct an ϵ-net of $\mathcal{R}_r^{\mathsf{norm}}$ — denoted by $\mathcal{M}$ — whose cardinality is at most $(9/\epsilon)^{3nr}$ (according to [CP11, Lemma 3.1]). Taking the union bound over $\mathcal{M}$ and assuming that

$$N_1 \ge C_0 nr \log\frac{9}{\epsilon}$$

for some sufficiently large constant $C_0 > 0$, we have with probability at least $1 - Ce^{-cn}$, for all $\boldsymbol{X}_1 \in \mathcal{M}$, the dataset $\mathcal{D}_1$ defined in (85) satisfies $Q_{0.8}(\mathcal{D}_1) \le Q_{0.8}(Z) + 0.01$. Finally, consider an arbitrary $\boldsymbol{X}_1 \in \mathcal{R}_r^{\mathsf{norm}}$, and let $\boldsymbol{X}_1^0$ be the point in $\mathcal{M}$ such that $\|\boldsymbol{X}_1^0 - \boldsymbol{X}_1\|_{\mathrm{F}} \le \epsilon$. Denote by $\mathcal{D}_1^0$ the dataset generated by $\boldsymbol{X}_1^0$ analogous to (85). Then we have

$$\big|Q_{0.8}(\mathcal{D}_1) - Q_{0.8}(\mathcal{D}_1^0)\big| \le \max_{i\in\Omega_1^\star} \big|\langle \boldsymbol{A}_i, \boldsymbol{X}_1\rangle - \langle \boldsymbol{A}_i, \boldsymbol{X}_1^0\rangle\big| \le \epsilon \max_{i\in\Omega_1^\star} \|\boldsymbol{A}_i\|_{\mathrm{F}} \lesssim \epsilon\left(n + \sqrt{\log N_1}\right),$$

where the last inequality holds with probability at least $1 - Ce^{-cn}$, according to Proposition 2. Setting $\epsilon = N_1^{-10}$, we further have $|Q_{0.8}(\mathcal{D}_1) - Q_{0.8}(\mathcal{D}_1^0)| \lesssim N_1^{-9} \le 0.01$, as long as N_1 is sufficiently large. In addition, it can be verified numerically that $Q_{0.8}(Z) < 1.30$. These together imply that for any $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K) \in \mathcal{T}_1^{\mathsf{norm}}$, we have

$$Q_\alpha(\mathcal{D}) \le Q_{0.8}(\mathcal{D}_1) \le Q_{0.8}(Z) + 0.02 \le 1.35,$$

which gives rise to the upper bound in Lemma 2.

Step 2: lower bounding $Q_\alpha(\mathcal{D})$. For notational convenience, we denote

$$q \coloneqq \frac{0.7\alpha}{p_1} \in [0.42, 0.56], \quad \text{and} \quad B_N \coloneqq \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\Omega_k^\star} \mathbb{1}\Big(\big|\langle \boldsymbol{A}_i, \boldsymbol{X}_k\rangle\big| \le \frac{Q_q(Z)}{1.01}\Big). \tag{108}$$

Clearly, by the definition of B_N, one has

$$\mathbb{P}\left(Q_\alpha(\mathcal{D}) < \frac{Q_q(Z)}{1.01}\right) \le \mathbb{P}\big(B_N > \alpha\big),$$

where it can be verified numerically that $Q_q(Z)/1.01 \ge 0.54$. Therefore, it suffices to upper bound the probability $\mathbb{P}\big(B_N > \alpha\big)$. To accomplish this, we first upper bound B_N as follows:

$$\begin{aligned} B_N &= \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\Omega_k^\star} \mathbb{1}\left(\left|\left\langle \boldsymbol{A}_i, \frac{\boldsymbol{X}_k}{\|\boldsymbol{X}_k\|_{\mathrm{F}}}\right\rangle\right| \le \frac{Q_q(Z)}{1.01\|\boldsymbol{X}_k\|_{\mathrm{F}}}\right) \\ &\le \frac{1}{N}\sum_{i\in\Omega_1^\star} \mathbb{1}\Big(\big|\langle \boldsymbol{A}_i, \boldsymbol{X}_1\rangle\big| \le \frac{Q_q(Z)}{1.01}\Big) + \frac{1}{N}\sum_{k\ne 1}\sum_{i\in\Omega_k^\star} \mathbb{1}\left(\left|\left\langle \boldsymbol{A}_i, \frac{\boldsymbol{X}_k}{\|\boldsymbol{X}_k\|_{\mathrm{F}}}\right\rangle\right| \le \frac{c_0 Q_q(Z)}{1.01K}\right). \end{aligned} \tag{109}$$

Here, the last line follows from the assumption that $1 = \|\boldsymbol{X}_1\|_{\mathrm{F}} \le (c_0/K)\min_{k\neq 1}\|\boldsymbol{X}_k\|_{\mathrm{F}}$; see the definition of $\mathcal{T}_1^{\mathsf{norm}}$ in (106). Note that $\boldsymbol{X}_1 \in \mathcal{R}_r^{\mathsf{norm}}$, and for all $k \neq 1$, we also have $\boldsymbol{X}_k/\|\boldsymbol{X}_k\|_{\mathrm{F}} \in \mathcal{R}_r^{\mathsf{norm}}$. Therefore, we can invoke Lemma 3 with $m = N_1 = |\Omega_1^\star|, \tau = Q_q(Z)/1.01, t = 0.15\alpha$ and $\epsilon = N_1^{-10}$ to obtain that: with probability at least $1 - Ce^{-cn}$ (provided that $m \ge C_0 nrK^2 \log m$), the following holds simultaneously for all $\boldsymbol{X}_1 \in \mathcal{R}_r^{\mathsf{norm}}$:

$$\frac{1}{N_1}\sum_{i\in\Omega_1^\star}\mathbb{1}\Big(\big|\langle \boldsymbol{A}_i, \boldsymbol{X}_1\rangle\big| \le \frac{Q_q(Z)}{1.01}\Big) \le \mathbb{P}\big(Z \le Q_q(Z)\big) + t + \frac{200\epsilon}{\tau} = q + 0.15\alpha + \frac{202N_1^{-10}}{Q_q(Z)}.$$

Similarly, for all $k \neq 1$, one can apply Lemma 3 with $m = N_k \coloneqq |\Omega_k^\star|, \tau = c_0 Q_q(Z)/(1.01K), t = 0.15\alpha$ and $\epsilon = N_k^{-10}$ to show that: with probability at least $1 - Ce^{-cn}$ (provided $m \ge C_0 nrK^2 \log m$), the following holds simultaneously for all $\boldsymbol{X}_k/\|\boldsymbol{X}_k\|_{\mathrm{F}} \in \mathcal{R}_r^{\mathsf{norm}}$:

$$\frac{1}{N_k}\sum_{i\in\Omega_k^\star}\mathbb{1}\left(\left|\left\langle \boldsymbol{A}_i, \frac{\boldsymbol{X}_k}{\|\boldsymbol{X}_k\|_{\mathrm{F}}}\right\rangle\right| \le \frac{c_0 Q_q(Z)}{1.01K}\right) \le \mathbb{P}\left(Z \le \frac{c_0 Q_q(Z)}{K}\right) + t + \frac{200\epsilon}{\tau} \le \frac{c_0 Q_q(Z)}{K} + 0.15\alpha + \frac{202KN_k^{-10}}{c_0 Q_q(Z)},$$

where the last inequality relies on the property of Z. Combine the above two bounds with (109) to reach

$$\begin{aligned} B_N &\le p_1\left(q + 0.15\alpha + \frac{202N_1^{-10}}{Q_q(Z)}\right) + \sum_{k\neq 1} p_k\left(\frac{c_0 Q_q(Z)}{K} + 0.15\alpha + \frac{202KN_k^{-10}}{c_0 Q_q(Z)}\right) \\ &\le p_1 q + \frac{c_0 Q_q(Z)}{K} + 0.15\alpha + p_1\frac{202N_1^{-10}}{Q_q(Z)} + \sum_{k\neq 1} p_k \frac{202KN_k^{-10}}{c_0 Q_q(Z)}. \end{aligned}$$

Recall that $p_1 q = 0.7\alpha$, $\alpha \asymp p_1 \asymp 1/K$, and observe that $p_1\frac{202N_1^{-10}}{Q_q(Z)} + \sum_{k\neq 1} p_k \frac{202KN_k^{-10}}{c_0 Q_q(Z)} \le 0.05\alpha$ as long as $N_k \gtrsim K$ for all k. Putting these together guarantees that $B_N \le \alpha$ as desired, which further implies

$$Q_\alpha(\mathcal{D}) \ge Q_q(Z)/1.01 \ge 0.54.$$

Combining this lower bound with the upper bound in Step 1 completes our proof of Lemma 2.

C.3 Proof of Lemma 3

Throughout the proof, we assume that the ensemble $\{\boldsymbol{A}_i\}$ obeys $(2r, 1/4)$-RIP. In view of Lemma 1, this happens with probability at least $1 - C_2 e^{-c_2 n}$ for some constants $c_2, C_2 > 0$, as long as $m \ge Cnr$. Recall the definition of χ from Appendix C.1, and set the parameter as $c_\chi = 0.01/1.01$. One then has

$$\frac{1}{m}\sum_{i=1}^m \mathbb{1}\left(|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle| \le \tau\right) \le \frac{1}{m}\sum_{i=1}^m \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle; 1.01\tau\right)$$

In the sequel, we invoke the standard covering argument to upper bound $\frac{1}{m}\sum_{i=1}^m \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle; 1.01\tau\right)$.

First, consider a fixed $\boldsymbol{X} \in \mathcal{R}_r^{\mathsf{norm}}$ independent of $\{\boldsymbol{A}_i\}$. In this case we can bound the expectation as

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^m \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle; 1.01\tau\right)\right] \le \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^m \mathbb{1}\left(|\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle| \le 1.01\tau\right)\right] = \mathbb{P}\left(Z \le 1.01\tau\right),$$

where we recall that Z follows the same distribution as $|\mathcal{N}(0,1)|$. In addition, note that $\frac{1}{m}\sum_{i=1}^m \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle; 1.01\tau\right)$ is the empirical average of m independent random variables, each lying within $[0,1]$ and having variance bounded by 2τ. Therefore, for all $t \ge 0$, one sees from Bernstein's inequality [Ver18, Theorem 2.8.4] that

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle; 1.01\tau\right) \ge \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^m \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle; 1.01\tau\right)\right] + t\right) \le \exp\left(-\frac{c_1 m t^2}{\tau + t}\right),$$

where $c_0, c_1 > 0$ are some universal constants. Let $\mathcal{M} \subseteq \mathcal{R}_r^{\mathsf{norm}}$ be an ϵ-net of $\mathcal{R}_r^{\mathsf{norm}}$, whose cardinality is at most $(9/\epsilon)^{3nr}$. The union bound reveals that: with probability at least $1 - (9/\epsilon)^{3nr}\exp(-c_1 m t^2/(\tau + t))$, one has

$$\sup_{\boldsymbol{X}\in\mathcal{M}} \quad \frac{1}{m}\sum_{i=1}^{m} \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle ; 1.01\tau\right) \le \mathbb{P}\left(Z \le 1.01\tau\right) + t.$$

Next, we move on to account for an arbitrary $\boldsymbol{X} \in \mathcal{R}_r^{\mathsf{norm}}$ (which is not necessarily independent of $\{\boldsymbol{A}_i\}$). Let $\boldsymbol{X}_0$ be a point in $\mathcal{M}$ obeying $\|\boldsymbol{X} - \boldsymbol{X}_0\|_{\mathrm{F}} \le \epsilon$. As a result, one has

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m} \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle ; 1.01\tau\right) - \frac{1}{m}\sum_{i=1}^{m} \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{X}_0\rangle ; 1.01\tau\right) &\le \frac{1}{m}\sum_{i=1}^{m} \left|\chi\left(\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle ; 1.01\tau\right) - \chi\left(\langle \boldsymbol{A}_i, \boldsymbol{X}_0\rangle ; 1.01\tau\right)\right| \\
&\overset{\text{(i)}}{\le} \frac{100}{\tau}\cdot\frac{1}{m}\sum_{i=1}^{m} \left|\langle \boldsymbol{A}_i, \boldsymbol{X} - \boldsymbol{X}_0\rangle\right| \\
&\overset{\text{(ii)}}{\le} \frac{100}{\tau}\cdot\sqrt{\frac{1}{m}\sum_{i=1}^{m} \langle \boldsymbol{A}_i, \boldsymbol{X} - \boldsymbol{X}_0\rangle^2} \\
&\overset{\text{(iii)}}{\le} \frac{200}{\tau}\|\boldsymbol{X} - \boldsymbol{X}_0\|_{\mathrm{F}} \le \frac{200}{\tau}\epsilon.
\end{aligned}$$

Here the inequality (i) holds since $\chi(\cdot; 1.01\tau)$ is Lipschitz with the Lipschitz constant $1/(1.01 c_\chi \tau) = 100/\tau$, the relation (ii) results from the Cauchy-Schwarz inequality, and (iii) follows since $\{\boldsymbol{A}_i\}$ obeys $(2r, 1/4)$-RIP.

Combine the above two inequalities to finish the proof.

D Estimating unknown parameters in Algorithm 4

Throughout the paper, we have assumed the knowledge of several problem-specific parameters, e.g. the proportion p_k of the k-th component, the rank r_k of the low-rank matrix $\boldsymbol{M}_k^\star$ and the rank $R = \mathsf{rank}(\mathbb{E}[\boldsymbol{Y}])$. In the sequel, we specify where we need them and discuss how to estimate them in practice.

- In Line 2 of Algorithm 4, when running Algorithm 1, we need to know $R = \mathsf{rank}(\mathbb{E}[\boldsymbol{Y}])$, which can be estimated faithfully by examining the singular values of the data matrix $\boldsymbol{Y}$.
- In Line 3 of Algorithm 4, when running Algorithm 2, we need to know $\{r_k\}_{1\le k\le K}$, where $r_k = \mathsf{rank}(\boldsymbol{M}_k^\star)$. Recall from (14) that $\boldsymbol{U}\widehat{\boldsymbol{S}}_k\boldsymbol{V}^\top \approx \boldsymbol{M}_k^\star$; therefore, r_k can be estimated accurately by examining the singular values of $\widehat{\boldsymbol{S}}_k$.
- In Line 5 of Algorithm 4, when running Algorithm 3, we need to know p_k to set η_k and α_k appropriately. It turns out that the outputs $\{\omega_k\}$ of the tensor method (see Algorithm 5) satisfy $\omega_k \approx p_k, 1 \le k \le K$.

References

[ACHL19] S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 7413–7424, 2019.

[AEP07] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems*, pages 41–48, 2007.

[AGH+14] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

[AZ05] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

[Bax00] J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

[BDS03] S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pages 567–580. Springer, 2003.

[BJK15] K. Bhatia, P. Jain, and P. Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.

[BNS16] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

[Car97] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[CC17] Y. Chen and E. J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on Pure and Applied Mathematics*, 70(5):822–883, 2017.

[CC18a] Y. Chen and E. J. Candès. The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Communications on Pure and Applied Mathematics*, 71(8):1648–1714, 2018.

[CC18b] Y. Chen and Y. Chi. Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Processing Magazine*, 35(4):14–31, 2018.

[CCD+19] V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *arXiv preprint arXiv:1904.10020*, 2019.

[CCFM19] Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1-2):5–37, 2019.

[CCF+ar] Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM Journal on Optimization*, to appear.

[CCG15] Y. Chen, Y. Chi, and A. J. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

[CFMY19] Y. Chen, J. Fan, C. Ma, and Y. Yan. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937, 2019.

[CGH14] Y. Chen, L. Guibas, and Q. Huang. Near-optimal joint object matching via convex relaxation. In *Proceedings of the International Conference on Machine Learning*, pages 1269–1277, 2014.

[CL13] A. T. Chaganty and P. Liang. Spectral experts for estimating mixtures of linear regressions. In *Proceedings of the International Conference on Machine Learning*, pages 1040–1048, 2013.

[CLC19] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[CLL20] J. Chen, D. Liu, and X. Li. Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. *IEEE Transactions on Information Theory*, 2020.

[CLS15] E. J. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[CLS20] S. Chen, J. Li, and Z. Song. Learning mixtures of linear regressions in subexponential time via fourier moments. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 587–600, 2020.

[CP11] E. J. Candes and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[CR09] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, 2009.

[CYC14] Y. Chen, X. Yi, and C. Caramanis. A convex formulation for mixed regression with two components: Minimax optimal rates. In *Proceedings of the Conference on Learning Theory*, pages 560–604, 2014.

[DC20] L. Ding and Y. Chen. Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Transactions on Information Theory*, 2020.

[DH00] P. Deb and A. M. Holmes. Estimates of use and costs of behavioural health care: a comparison of standard and finite mixture models. *Health Economics*, 9(6):475–489, 2000.

[DHK+20] S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

[DV89] R. D. De Veaux. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8(3):227–245, 1989.

[EMP05] T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6(Apr):615–637, 2005.

[FAL17] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pages 1126–1135. JMLR.org, 2017.

[GJZ17] R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1233–1242. PMLR, 2017.

[GS99] S. Gaffney and P. Smyth. Trajectory clustering with mixtures of regression models. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 63–72, 1999.

[HJ18] P. Hand and B. Joshi. A convex program for mixed linear regression with a recovery guarantee for well-separated data. *Information and Inference: A Journal of the IMA*, 7(3):563–579, 2018.

[JNS13] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, pages 665–674, 2013.

[JSRR10] A. Jalali, S. Sanghavi, C. Ruan, and P. K. Ravikumar. A dirty model for multi-task learning. In *Advances in neural information processing systems*, pages 964–972, 2010.

[KC07] A. Khalili and J. Chen. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102(479):1025–1038, 2007.

[KHC20] J. Kwon, N. Ho, and C. Caramanis. On the minimax optimality of the em algorithm for learning two-component mixed linear regression. *arXiv preprint arXiv:2006.02601*, 2020.

[KMMP19] A. Krishnamurthy, A. Mazumdar, A. McGregor, and S. Pal. Sample complexity of learning mixture of sparse linear regressions. In *Advances in Neural Information Processing Systems*, pages 10532–10541, 2019.

[KMO10] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[KQC+19] J. Kwon, W. Qian, C. Caramanis, Y. Chen, and D. Davis. Global convergence of the em algorithm for mixtures of two component linear regression. In *Proceedings of the Conference on Learning Theory*, pages 2055–2110, 2019.

[KSKO20] W. Kong, R. Somani, S. Kakade, and S. Oh. Robust meta-learning for mixed linear regression with small batches. *arXiv preprint arXiv:2006.09702*, 2020.

[KSS+20] W. Kong, R. Somani, Z. Song, S. Kakade, and S. Oh. Meta-learning for mixed linear regression. *arXiv preprint arXiv:2002.08936*, 2020.

[KYB19] J. M. Klusowski, D. Yang, and W. Brinda. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Transactions on Information Theory*, 65(6):3515–3524, 2019.

[LCZL20] Y. Li, Y. Chi, H. Zhang, and Y. Liang. Non-convex low-rank matrix recovery with arbitrary outliers via median-truncated gradient descent. *Information and Inference: A Journal of the IMA*, 9(2):289–325, 2020.

[LL18] Y. Li and Y. Liang. Learning mixtures of linear regressions with nearly optimal complexity. In *Proceedings of the Conference On Learning Theory*, pages 1125–1144, 2018.

[LMZ18] Y. Li, T. Ma, and H. Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Proceedings of the Conference On Learning Theory*, pages 2–47, 2018.

[LZT19] Q. Li, Z. Zhu, and G. Tang. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1):51–96, 2019.

[MP20] A. Mazumdar and S. Pal. Recovery of sparse signals from a mixture of linear samples. *arXiv preprint arXiv:2006.16406*, 2020.

[MPRP16] A. Maurer, M. Pontil, and B. Romera-Paredes. The benefit of multitask representation learning. *The Journal of Machine Learning Research*, 17(1):2853–2884, 2016.

[MWCC20] C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 20(3):451–632, 2020.

[NNS+14] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain. Non-convex robust pca. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

[PA18] D. Pimentel-Alarcón. Mixture matrix completion. In *Advances in Neural Information Processing Systems*, pages 2193–2203, 2018.

[PKCS17] D. Park, A. Kyrillidis, C. Carmanis, and S. Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. In *Artificial Intelligence and Statistics*, pages 65–74, 2017.

[PL14] A. Pentina and C. Lampert. A pac-bayesian bound for lifelong learning. In *Proceedings of the International Conference on Machine Learning*, pages 991–999, 2014.

[PLW+20] M. Peng, Y. Li, B. Wamsley, Y. Wei, and K. Roeder. cFIT: Integration and transfer learning of single cell transcriptomes, illustrated by fetal brain cell development. *bioRxiv*, 2020.

[PY09] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009.

[QR78] R. E. Quandt and J. B. Ramsey. Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, 73(364):730–738, 1978.

[RFP10] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[Rou84] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

[SBVDG10] N. Städler, P. Bühlmann, and S. Van De Geer. L1-penalization for mixture regression models. *Test*, 19(2):209–256, 2010.

[Ser09] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*, volume 162. John Wiley & Sons, 2009.

[SJA16] H. Sedghi, M. Janzamin, and A. Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics*, pages 1223–1231, 2016.

[SL16] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[SLHH18] Y. Sun, Z. Liang, X. Huang, and Q. Huang. Joint map and symmetry synchronization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 251–264, 2018.

[SQW18] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

[SS19a] Y. Shen and S. Sanghavi. Iterative least trimmed squares for mixed linear regression. In *Advances in Neural Information Processing Systems*, pages 6078–6088, 2019.

[SS19b] Y. Shen and S. Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In *Proceedings of the International Conference on Machine Learning*, pages 5739–5748, 2019.

[SSZ17] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[SWS20] V. Shah, X. Wu, and S. Sanghavi. Choosing the sample with lowest loss makes sgd robust. *arXiv preprint arXiv:2001.03316*, 2020.

[TBS+16] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *Proceedings of the International Conference on Machine Learning*, pages 964–973, 2016.

[TJJ20] N. Tripuraneni, C. Jin, and M. I. Jordan. Provable meta-learning of linear representations. *arXiv preprint arXiv:2002.11684*, 2020.

[TMC20] T. Tong, C. Ma, and Y. Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *arXiv preprint arXiv:2005.08898*, 2020.

[Tur00] T. R. Turner. Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(3):371–384, 2000.

[Ver18] R. Vershynin. *High-dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge university press, 2018.

[VT02] K. Viele and B. Tong. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12(4):315–330, 2002.

[Wed72] P.-Å. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[YC15] X. Yi and C. Caramanis. Regularized em algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 1567–1575, 2015.

[YCS14] X. Yi, C. Caramanis, and S. Sanghavi. Alternating minimization for mixed linear regression. In *Proceedings of the International Conference on Machine Learning*, pages 613–621, 2014.

[YCS16] X. Yi, C. Caramanis, and S. Sanghavi. Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. *arXiv preprint arXiv:1608.05749*, 2016.

[YL20] H. Yuan and Y. Liang. Learning entangled single-sample distributions via iterative trimming. volume 108 of *Proceedings of Machine Learning Research*, pages 2666–2676, Online, 26–28 Aug 2020. PMLR.

[YPCR18] D. Yin, R. Pedarsani, Y. Chen, and K. Ramchandran. Learning mixtures of sparse linear regressions using sparse graph codes. *IEEE Transactions on Information Theory*, 65(3):1430–1451, 2018.

[ZCL18] H. Zhang, Y. Chi, and Y. Liang. Median-truncated nonconvex approach for phase retrieval with outliers. *IEEE Transactions on Information Theory*, 64(11):7287–7310, 2018.

[ZJD16] K. Zhong, P. Jain, and I. S. Dhillon. Mixed linear regression with multiple components. In *Advances in Neural Information Processing Systems*, pages 2190–2198, 2016.

[ZL15] Q. Zheng and J. Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pages 109–117, 2015.

[ZLTW18] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.

[ZQW20] Y. Zhang, Q. Qu, and J. Wright. From symmetry to geometry: Tractable nonconvex problems. *arXiv preprint arXiv:2007.06753*, 2020.

[ZWYG18] X. Zhang, L. Wang, Y. Yu, and Q. Gu. A primal-dual analysis of global optimality in nonconvex low-rank matrix recovery. In *International conference on machine learning*, pages 5862–5871, 2018.