Learning Tensor Representations for Meta-Learning

Samuel Deng, Yilin Guo, Daniel Hsu, and Debmalya Mandal

Columbia University

September 9, 2021

Abstract

We introduce a tensor-based model of shared representation for meta-learning from a diverse set of tasks. Prior works on learning linear representations for meta-learning assume that there is a common shared representation across different tasks, and do not consider the additional task-specific observable side information. In this work, we model the meta-parameter through an order-3 tensor, which can adapt to the observed task features of the task. We propose two methods to estimate the underlying tensor. The first method solves a tensor regression problem and works under natural assumptions on the data generating process. The second method uses the method of moments under additional distributional assumptions and has an improved sample complexity in terms of the number of tasks. We also focus on the meta-test phase, and consider estimating task-specific parameters on a new task. Substituting the estimated tensor from the first step allows us estimating the task-specific parameters with very few samples of the new task, thereby showing the benefits of learning tensor representations for meta-learning. Finally, through simulation and several real-world datasets, we evaluate our methods and show that it improves over previous linear models of shared representations for meta-learning.

1 Introduction

One of the major challenges in modern machine learning is training a model with limited amounts of data. This is particularly important in settings where data is scarce and new data is costly to acquire. In recent years, several techniques like data augmentation, transfer learning have been proposed to address problems with limited data. The focus of this paper is meta-learning, which has turned out to be an important framework to address such problems. The main idea behind meta-learning is to design learning algorithms that can leverage prior learning experience to adapt to a new problem quickly, and learn a useful algorithm with few samples. Such approaches have been quite successful in diverse applications like natural language processing [19], robotics [22], and healthcare [36].

Meta-learning algorithms are often given a family of related tasks and attempt to use few samples on a new related task by utilizing the overlap between the new test task and already seen training tasks. In that sense, a meta-learning algorithm is learning to learn on new tasks, and performance improves with experience and number of tasks [26]. Despite immense success, we are yet to fully

E-mail: , sd3013@columbia.edu, yg2553@columbia.edu, djhsu@cs.columbia.edu, dm3557@columbia.edu

understand the theoretical foundations of meta-learning algorithms. The most promising theoretical direction stems from representation learning. The main idea is that the tasks share a common shared representation and a task-specific representation [29, 28]: if the shared representation is learned from training tasks, then the task-specific representation for the new task can be learned with few samples.

Current models of shared representations for meta-learning do not take into account two observations – (1) the training tasks are often heterogeneous, and the shared representation cannot be captured by a single parameter, (2) tasks often come with additional task-specific observable side information, and they should be part of any representation-based model of meta-learning. The first situation often arises in robotics, and various reinforcement learning environments [33], while the latter is prevalent in recommender system [31], where items (tasks) that users rate often come with observable features.

We aim to understand meta-learning of features for settings where task-specific observable features affect the outcome. In particular, we are interested in the following questions. (1) What is the appropriate generalization of meta-learning of linear representation with task-specific observable features? (2) Moreover, given samples from T tasks, how can we efficiently learn such a representation and does it improve sample efficiency on a new task?

1.1 Contributions

Tensor Based Model. We propose a tensor based model of representations for meta-learning representations for a diverse set of tasks. In particular, we model the meta-parameter through a tensor of order-3, which can be thought of as a multi-linear function mapping a tuple of (input feature, observed task feature, unobserved task feature) to a real-valued output. As our model considers task-specific observed features, the meta-parameter can adapt to particular task and generalizes the matrix-based linear representations proposed by [28].

Estimation. We first determine the identifiable component of the shared representation based model, and estimate the first two factors of the underlying tensor in the meta-training phase. We propose two methods – (1) tensor regression based method works with natural assumptions on the data generating process, and (2) method of moments based estimation works under additional distributional assumptions, but has improved sample complexity in terms of the number of tasks.

Meta-Test Phase. After estimating the shared parameters, we focus on the meta-test phase, where a new task is given. We show that substituting the estimated factors from the first step provably improves error in estimating the task-specific parameters on a new task. In particular, the excess test error on the new task is bounded by $O\left(\frac{r^2}{N_2}\right)$ where r is the rank of the underlying tensor and N_2 is the number of samples from the new task. As tensor rank r can be quite small compared to the dimensions, this highlights the benefits of learning task-adaptive representations in meta-learning. Finally, through a simulated dataset and several real-world datasets, we evaluate our methods and show that it improves over previous models of learning shared representations for meta-learning.

1.2 Related Work

Baxter [6] was the first to prove generalization bound for multitask learning problem. However, they considered a model of multitask learning where tasks with shared representation are sampled from a generative model. Pontil et al., Maurer et al. [24, 20] developed general uniform-convergence based

framework to analyze multitask representation learning. However, they assume oracle access to a global empirical risk minimizer. On the other hand, we provide specific algorithms and also consider task-specific side information.

The work closest to ours is [28], who proposed a linear model for learning representation in meta-learning. Our model can be thought of as a general model of theirs as we do not assume a fixed low-dimensional representation across tasks, and can adapt to observable side-information of the tasks. We also note that Tripuraneni et al. [29] generalized the linear model [28] to consider transfer learning with general class of functions, however, they assume oracle access to a global empirical risk minimizer, and the common representation (a shared function) does not adapt to observable features of the tasks. Finally, Du et al. [10] also considered the problem of learning shared representations and obtained similar results. Compared to [29], they consider general non-linear representations, but the representation again does not depend on the observable features of the task.

Our work is also related to the conditional meta-learning framework introduced by [35, 9]. Conditional meta-learning aims to learn a conditioning function that maps task-specific side information to a meta-parameter suitable for the task. Denevi et al. [9] studies a biased regularization formulation where the goal is to find task-specific parameter close to a bias vector, possibly dependent on side-information. On the other hand, [35] takes a structured prediction framework, and only proves generalization bounds. Although our framework falls within the conditional meta-learning framework, we want to understand the benefits of representation learning on a new task.

In this work, we aim to understand meta-learning through a representation learning viewpoint. However, in recent years, several works have attempted to improve our understanding of meta-learning from other viewpoints. These include optimization [7, 13], train-validation split [4], and convexity [25]. Additionally, there are several recent works on understanding gradient based meta-learning [11, 12, 8, 5, 14], but their setting is very different from ours.

Finally, we use tensors to model the meta-parameter in the presence of task-specific side information. Our estimation method uses tensor regression [37, 27] and tensor decomposition [1]. For tensor regression, we build upon the algorithm proposed by [27], and for tensor decomposition we use a robust version introduced by [2].

2 Preliminaries

We will consider standard two-stage model of meta-learning, consisting of a meta-training phase and a meta-test phase. In the meta-training stage, we see N samples from T training tasks and learn a meta parameter. In the meta-test stage, we see N_2 samples from a fixed target task (say task 0) and learn a target-specific parameter conditioned on the meta-parameter and features of the new task 0. We first define our response model which specifies the particular model of shared linear representation.

Response model. There are T training tasks and each task is associated with a pair of observed and unobserved task feature vector. Task t is characterized by (Y_t, Z_t) where $Y_t \in \mathbb{R}^{d_2}$ is the observed task feature vector and $Z_t \in \mathbb{R}^{d_3}$ is the unobserved task feature vector for the t-th task. A sample from task t is specified by a tuple (X, Y_t, Z_t), where X is some user feature vector. Given such a

data tuple (X, Y_t, Z_t) the response is given as

$$R = A(X, Y_t, Z_t) + \varepsilon = \sum_{i,j,k} A_{i,j,k} X_i Y_{tj} Z_{tk} + \varepsilon \tag{1}$$

Here the noise variable $\varepsilon \sim \mathrm{N}(0, \sigma^2)$, and $A \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is the system tensor which we treat as a multi-linear real-valued function on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \mathbb{R}^{d_3}$. Note that, our model generalizes the linear model proposed by [28], and the meta-parameter (tensor A) is not a fixed parameter, and adapts to the observed feature / side information for task t. We will assume that the tensor A has CP-rank r i.e. there exist matrices $A^1 \in \mathbb{R}^{d_1 \times r}$, $A^2 \in \mathbb{R}^{d_2 \times r}$, and $A^3 \in \mathbb{R}^{d_3 \times r}$ such that

$$A_{i,j,k} = \sum_{s=1}^{r} A^1_{si} A^2_{sj} A^3_{sk}$$

Following [15], we will write $A = [\![\mathrm{I}_r; A^1, A^2, A^3]\!]$ to denote the rank-r decomposition of the tensor A. Notice that here we assume that all the singular values of the tensor A is one. Without making strong assumptions on the unobserved task features Z_t, general singular values cannot be identified.[1]

Training data. Let P be a distribution over the feature vectors X_i's which we will often refer to as user feature vectors. Let Q be a joint distribution over observed and unobserved task feature vectors. Let $\{X_i : i \in [N]\}$ and $\{(Y_t, Z_t) : t \in [T]\}$ be independent random variables, where $X_1, \ldots, X_N \sim_{\text{iid}} P$ and $(Y_1, Z_1), \ldots, (Y_T, Z_T) \sim_{\text{iid}} Q$. Conditional on these (random) feature vectors, let $R_1, \ldots, R_N$ be independent realizations of R from the response model in Equation (1), where

$$R_i = A(X_i, Y_{t(i)}, Z_{t(i)}) + \varepsilon_i. \tag{2}$$

Here $t : [N] \to [T]$ is a mapping that specifies, for each training instance i, corresponding task $t(i)$. Therefore, the training data is given as $\{X_i, Y_{t(i)}, R_i\}_{i \in [N]}$.

Meta-Test data. At test time we are given a fixed task (say 0) with observed feature Y_0 and unobserved feature Z_0. We are given N_2 instances from this new task, $\{X_i, Y_0, R_i\}_{i \in [N_2]}$ where $X_1, \ldots, X_{N_2} \sim_{\text{iid}} P$. Our goal is to design a predictor $f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ that maps an input feature and an observed task feature to a predicted response. We will evaluate our predictor by its mean squared error on the new task.

$$\mathrm{MSE}(f) = \mathbb{E}\left[(f(X, Y_0) - A(X, Y_0, Z_0) - \varepsilon)^2\right] = \sigma^2 + \mathbb{E}\left[(f(X, Y_0) - A(X, Y_0, Z_0))^2\right].$$

In order to design the predictor on the new task, we need estimates of tensor A, and unobserved task feature on the new task Z_0.

Notations. For a matrix $B \in \mathbb{R}^{d_1 \times d_2}$ we will write $\|B\|_{\text{op}}$ to denote its operator norm, which is defined as $\|B\|_{\text{op}} = \max_{x \in \mathbb{R}^{d_2}} \frac{\|Bx\|_2}{\|x\|_2}$. For matrix B, we will write $\|B\|_F = \sqrt{\sum_{i,j} B_{ij}^2}$ to denote its Frobenius norm. For a tensor $A \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ we write its spectral norm $\|A\|_{\text{op}} =$

[1] We provide a counter-example in the appendix.

$\max_{\|x\|_2=\|y\|_2=\|z\|_2=1} \big|A(x,y,z)\big|$. Like a matrix, we will also write $\|A\|_F = \sqrt{\sum_{i,j,k} A_{i,j,k}^2}$ to denote the Frobenius norm of the tensor A. We sometimes use the tensor by slices, for the 3-order tensor A, we denote its horizontal slices as $A_{j::}$, for $j \in [d_1]$.

We will use two types of special matrix products in our paper. Given matrices $A \in \mathbb{R}^{d_1 \times d_2}$ and $B \in \mathbb{R}^{d_3 \times d_4}$, the Kronecker product $A \otimes B \in \mathbb{R}^{d_1 d_3 \times d_2 d_4}$ is

$$A \otimes B = [a_1 \otimes b_1 \; a_1 \otimes b_2 \ldots a_{d_2} \otimes b_{d_4 - 1} \; a_{d_2} \otimes b_{d_4}]$$

For matrices $A \in \mathbb{R}^{d_1 \times d_2}$ and $B \in \mathbb{R}^{d_3 \times d_2}$, their Khatri-Rao product $A \odot B \in \mathbb{R}^{d_1 d_3 \times d_2}$ is

$$A \odot B = [a_1 \otimes b_1 \; a_2 \otimes b_2 \; \ldots a_{d_2} \otimes b_{d_2}]$$

In addition, we denote standard basis vector as $\mathbf{e}_i$ whose coordinates are all zero, except i-th equals 1.

3 Estimation

We estimate the parameters of our model in two steps – (1) estimate the shared tensor using the meta-training data, and (2) estimate the parameters of the test task using the meta-test data and the estimate of the shared tensor. However, it turns out that even when there is a single task in the meta-training phase, the third factor A^3 cannot be identified for general tensor with orthogonal factors. In the appendix, we construct an example which shows that two different tensors with identical A^1, A^2 but different A^3 leads to the same observed outcomes. Therefore, we estimate the first two factors of the tensor A in the meta-training phase. In the meta-test phase, we substitute estimates of A^1 and A^2 and recover the parameters for the new task. We provide two ways to estimate the factors. The first method uses tensor regression; the second method uses the method-of-moments.

3.1 Tensor-Regression Based Estimation

In order to see why tensor regression can help us recover the shared tensor, we first show an alternate way to write the response R_i, as defined in Equation (2). Define

$$\mathcal{Z} = \begin{bmatrix} Z_1 & \cdots & Z_T \end{bmatrix}^\mathsf{T} \in \mathbb{R}^{T \times d_3} \tag{3}$$

to be the matrix corresponding to the unobserved features of the T training tasks. Then $A \times_3 \mathcal{Z} = A(I_{d_1}, I_{d_2}, \mathcal{Z}) \in \mathbb{R}^{d_1 \times d_2 \times T}$ is the tensor corresponding to unobserved parameters, defined as

$$(A \times_3 \mathcal{Z})_{i,j,t} = \sum_{k=1}^{d} A_{i,j,k} \mathcal{Z}_{t,k}.$$

Additionally, we define a covariate tensor $\mathcal{X}_i \in \mathbb{R}^{d_1 \times d_2 \times T}$ corresponding to the observed features as:

$$\mathcal{X}_i(\cdot, \cdot, t) = \begin{cases} X_i Y_{t(i)}^\mathsf{T} & \text{if } t = t(i) \\ 0_{d_1 \times d_2} & \text{o.w.} \end{cases} \tag{4}$$

Then, according to Equation (2), we have the following linear regression model for the i-th response.

$$R_i = \langle \mathcal{X}_i, A \times_3 \mathcal{Z} \rangle + \varepsilon_i. \tag{5}$$

Therefore, we can use tensor regression to get an estimate of $A \times_3 \mathcal{Z}$. Since the three factors of A are A^1, A^2, and A^3, it can be easily seen that the CP-decomposition of $A \times_3 \mathcal{Z}$ is $[\![\mathrm{I}_r; A^1, A^2, \mathcal{Z}A^3]\!] = [\![G^{-1}; A^1, A^2, \mathcal{Z}A^3G]\!]$. Here G is a diagonal matrix with i-th entry $1/\left\|\mathcal{Z}A_i^3\right\|_2$ and normalizes the columns of $\mathcal{Z}A^3$. Because of this particular form of the tensor $A \times_3 \mathcal{Z}$, we can run a tensor decomposition of the estimate of $A \times_3 \mathcal{Z}$ to recover A^1, A^2, and $\mathcal{Z}A^3G$. However, there is a catch as we have an estimate of $A \times_3 \mathcal{Z}$, instead of the exact tensor. So we need the tensor decomposition method to be *robust* to the estimation error. Algorithm 1 describes the full algorithm for recovering A from the training samples.

ALGORITHM 1: Tensor-Regression Based Estimation

Input: $(X_i, Y_{t(i)}, R_i)$ for $i = 1, \ldots, N$

1. Solve the following tensor regression problem:

$$\widehat{B} = \underset{B \in \mathbb{R}^{d_1 \times d_2 \times T}}{\arg\min} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(R_i - \langle \mathcal{X}_i, B \rangle \right)^2 + \lambda \|B\|_S \right\} \tag{6}$$

1

2. Run a robust tensor decomposition of $\widehat{B}$ of CP-rank r:

$$[\![\widehat{W}; \widehat{B}^1, \widehat{B}^2, \widehat{B}^3]\!] \leftarrow \text{Robust-Tensor-Decomposition}(\widehat{B}, r)$$

Output: Return $\widehat{A \times_3 \mathcal{Z}} = [\![\mathrm{I}_r; \widehat{B}^1, \widehat{B}^2, \widehat{B}^3\widehat{W}]\!]$.

Tensor Regression Details and Guarantees. Throughout this section, we will make the following assumptions about the data generating distribution.

(A1) $X_1, \ldots, X_N \sim_{\text{iid}} \mathrm{N}(0, \Sigma)$.

(A2) $Y_1, \ldots, Y_T \sim_{\text{iid}} \mathrm{N}(0, \Sigma_y)$.

(A3) For each i, $t(i) \sim \text{Unif}\{1, \ldots, T\}$.

Equation (6) is the tensor regression step to obtain an estimate of $B = A \times_3 \mathcal{Z}$. We use a regularized least squared regression, introduced by Tomioka et al. [27]. Here $\|B\|_S$ is the overlapped Schatten-1 norm of the tensor B, which is defined as the average of mode-wise nuclear norms i.e. $\|B\|_S = 1/3 \sum_{k=1}^{3} \|B_{(k)}\|_\star$. Since matrix nuclear norm is a convex function, the tensor regression problem stated in Equation (6) is also a convex problem, and can be solved efficiently.

For a tensor of dimension $d_1 \times d_2 \times T$, we introduce the following notation, which will appear frequently in our bounds.

$$D_1 = \sqrt{d_1} + \sqrt{d_2} + \sqrt{T} + \sqrt{d_1 T} + \sqrt{d_2 T} + \sqrt{d_1 d_2} \tag{7}$$

The next theorem states the guarantees of the tensor regression step.

Theorem 1. *Suppose Assumptions (A1)-(A3) hold, and* $N \geq O\left(\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma_y)} r D_1^2\right)$. *Then, with probability at least* $1 - e^{-\Omega(D_1^2)}$, *we have*

$$\left\|\widehat{B} - B\right\|_F \leq O\left(\frac{\sigma T D_1 \sqrt{r}}{\lambda_{\min}(\Sigma_y)\lambda_{\min}(\Sigma)\sqrt{N}}\right).$$

The full proof is provided in the supplementary material. Here we provide an overview of the main steps of the proof. Our analysis builds upon the work by Tomioka et al. [27], who analyzed the performance of tensor regression with overlapped Schatten-1 norm. The main ingredient of the proof is to show that under certain assumptions *restricted strong convexity* (RSC) holds. This property was introduced by [23] in the context of several matrix estimation problems, and ensures that the loss function has sufficient curvature to ensure consistent recovery of the unknown parameter. Tomioka et al. [27] proves that when the covariate tensors $\mathcal{X}_i$ are normally distributed, RSC holds with a fixed constant. For our setting, the covariate tensors are defined in Equation (4) and are not necessarily distributed from a multivariate Gaussian distribution. However, we can generalize the original proof of [23] to show that under Assumptions (A1) (X_i-s are normally distributed) and (A2) (tasks are sampled uniformly at random), RSC still holds for our setting, but with constant $O(1/T)$. Then we show that the parameter λ can be chosen to be a suitably large constant to get the error bounds of Theorem 1.

Tensor Decomposition Details and Guarantees: Having recovered the tensor $A \times_3 \mathcal{Z}$, we now aim to recover the factors A^1, A^2, and $\mathcal{Z}A^3$. Since we do not have the exact tensor $A \times_3 \mathcal{Z}$, but rather an estimate of the tensor, we apply robust tensor decomposition method (step 2 of Algorithm 1) to recover the factors of $A \times_3 \mathcal{Z}$. For robust tensor decomposition method, we will apply the algorithm of [2]. It is in general impossible to recover the factors of a noisy tensor without making any assumptions. So we will make the following assumptions about the underlying tensor $A = [\mathrm{I}_r; A^1, A^2, A^3]$. We will write $d = \max\{d_1, d_2, T\}$.

(B1) The columns of the factors of A are orthogonal i.e., $\langle A_i^1, A_j^1\rangle = \langle A_i^2, A_j^2\rangle = \langle A_i^3, A_j^3\rangle = 0$ for all $i \neq j$.

(B2) The components have bounded $2 \to p$ for some p i.e. $\exists p < 3$,

$$\max\left\{\left\|A^{1\intercal}\right\|_{2\to p}, \left\|\left\|\,1\right]A^{2\intercal}{}_{2\to p}, \left\|A^{3\intercal}\right\|_{2\to p}\right\} \leq 1 + o(1).$$

(B3) Rank is bounded i.e. $r = o(d)$.

Additionally, recall the definition of $\mathcal{Z}$, the matrix of unobserved features.

$$\mathcal{Z} = \begin{bmatrix} Z_1 & \cdots & Z_T \end{bmatrix}^\intercal \in \mathbb{R}^{T \times d_3}$$

(Z1) $\frac{1}{d_3^{0.5+\gamma}} \mathrm{I}_{d_3} \preccurlyeq \mathcal{Z}^\intercal \mathcal{Z} \preccurlyeq \frac{1}{\sqrt{d_3}} \mathrm{I}_{d_3}$ for some $\gamma > 0$.

(Z2) $\kappa(\mathcal{Z}^\intercal \mathcal{Z}) = \frac{\lambda_{\max}(\mathcal{Z}^\intercal \mathcal{Z})}{\lambda_{\min}(\mathcal{Z}^\intercal \mathcal{Z})} \leq 1 + O(\sqrt{r/d})$.

Although assumptions (Z1) and (Z2) might seem strong requirements on the matrix of unobserved features, they are usually satisfied when the unobserved task feature matrix is drawn from gaussian distribution. For example, if $Z_t \sim_{\text{iid}} \mathrm{N}(0, \nu \mathrm{I}_{d_3})$ then the assumptions hold for small enough ν.

Lemma 1 (Informal Statement). *Suppose tensor A satisfies the assumptions (B1)-(B3), the matrix of unobserved features $\mathcal{Z}$ satisfies assumptions (Z1)-(Z2), and $N \geq \tilde{O}\left(\frac{\sigma^2 T^2 D_1^2 r}{\lambda_{\min}^2(\Sigma_y)\lambda_{\min}^2(\Sigma)}\right)$. Then the tensor $\hat{A} = [\mathrm{I}_r; \widehat{A^1}, \widehat{A^2}, \widehat{\mathcal{Z}A^3}]$ output by Algorithm 1 satisfies*

$$\max\left\{\left\|\widehat{A^1} - A^1\right\|_F, \left\|\widehat{A^2} - A^2\right\|_F\right\} \leq \tilde{O}\left(\frac{\sigma T D_1 r}{\rho\sqrt{N}}\right), \quad \left\|\widehat{\mathcal{Z}A^3} - \mathcal{Z}A^3\right\|_F \leq \tilde{O}\left(\frac{\sigma T D_1 r^{1.5}}{\rho\sqrt{N}}\right)$$

where $\rho = \sqrt{\lambda_{\min}(\mathcal{Z}^\top\mathcal{Z})}\lambda_{\min}(\Sigma_y)\lambda_{\min}(\Sigma)$.

The proof shows that when the assumptions (B1)-(B3) and (Z1)-(Z3) are satisfied, we can apply robust tensor decomposition method to the tensor $A \times_3 \mathcal{Z}$. Note that the bound for the third factor $\mathcal{Z}A^3$ is worse by a factor of $\sqrt{r}$. This is because we recover an estimate of $\mathcal{Z}A^3G$ for a diagonal matrix G from tensor decomposition and then post-multiply this estimate by another diagonal matrix to obtain an estimate of $\mathcal{Z}A^3$.

3.1.1 Meta-Test

During the meta-test phase, we are given a new task (i.e. task 0 with observed feature $Y_0 \in \mathbb{R}^{d_2}$, and hidden feature $Z_0 \in \mathbb{R}^{d_3}$), and our goal is to learn the unobserved parameter of this task with as few samples as possible. As is standard in the meta-learning literature, we get a new training sample from the new task, and our goal is to perform well on the test sample drawn from the new task. There are N_2 training samples from the new task, where the features $X_1, \ldots, X_{N_2}$ are drawn iid from a distribution P. We will assume each feature X_i is mean-zero, has covariance matrix Σ ($\mathbb{E}[X_i X_i^\top] = \Sigma$), and Σ-subgaussian i.e. $\mathbb{E}[\exp(v^\top X_i)] \leq \exp\left(1/2 \left\|\Sigma^{1/2} v\right\|_2^2\right)$. The observed responses on these N_2 points are given as

$$R_i = A(X_i, Y_0, Z_0) + \varepsilon_i$$

where $\varepsilon_i \sim_{\text{iid}} \mathrm{N}(0, 1)$. Define $\mathcal{X}_0$ to be the matrix corresponding to the features on the new task i.e.

$$\mathcal{X}_0 = \begin{bmatrix} X_1 & \cdots & X_{N_2} \end{bmatrix}^T \in \mathbb{R}^{N_2 \times d_1}$$

We aim to estimate $A^{3\top} Z_0$ by substituting the estimates of A^1 and A^2. Notice that the response R_i can also be expressed as

$$R_i = (Y_0^\top A^2 \odot X_i^\top A^1) A^{3\top} Z_0 + \varepsilon_i.$$

Therefore, we can solve the following least square regression problem.

$$\widehat{A^{3\top} Z_0} = \arg\min_{\alpha_0 \in \mathbb{R}^r} \left\| R - (Y_0^\top \hat{A}^2 \odot \mathcal{X}_0 \hat{A}^1)\alpha_0 \right\|_2^2. \tag{8}$$

If we write $\widehat{V} = (Y_0^\top \hat{A}^2 \odot \mathcal{X}_0 \hat{A}^1)$, then the solution of Problem 8 is given as $\widehat{A^{3\top} Z_0} = (\widehat{V}^\top \widehat{V})^{-1} \widehat{V}^\top R$. Now our prediction on a new test instance X_0 from the new task is given as $(Y_0^\top \hat{A}^2 \odot X_0^\top \hat{A}^1)\widehat{A^{3\top} Z_0}$. With slight abuse of notation we will write this prediction as $\hat{A}(X_0, Y_0, \widehat{Z}_0)$. The next theorem bounds the mean squared error in the meta-test phase.

Theorem 2 (Informal Statement). *Suppose* $\max\{\|\hat{A}^1 - A^1\|_F, \|\hat{A}^2 - A^2\|_F\} \le \delta$. *Additionally,* $N_2 \ge \tilde{O}(r)$ *and* $|Y_0^\intercal \hat{A}_i^2| \ge \eta \|Y_0\|_2$ *for all* $i \in [r]$. *Then for dimension-independent constants* B_1, *and* B_2 *we have*

$$\mathbb{E}_{X_0}\left[\left(R_0 - \hat{A}(X_0, Y_0, \widehat{Z}_0)\right)^2\right] = O\left(\sigma^2 + \frac{B_1}{\eta^2} r^2 \delta^2 + \frac{B_2}{\eta^2}\frac{r^2}{N_2}\right)$$

with high probability.

The proof of the theorem shows that the mean squared error can be bounded as $O\left(r \left\|\widehat{A^{3\intercal} Z_0} - A^{3\intercal} Z_0\right\|_2^2\right) + O\left(\delta^2 \left\|\widehat{A^{3\intercal} Z_0}\right\|_2^2\right)$. Then we write down the first term as a sum of bias and variance term and establish respective bounds of $O(r^2/N_2)$ and $O(r^2\delta^2)$. Finally, we show that the L_2-norm of $\widehat{A^{3\intercal} Z_0}$ cannot be too large and is bounded by $O(r)$. Substituting these three bounds on the upper bound on the mean squared error gives us the desired result. Note that the theorem requires a lower bound on the inner product between the new task feature Y_0 and the columns of $\hat{A}^2$. This can be avoided with a slightly worse dependence on r. First, we can eliminate all columns i such that $|Y_0^\intercal \hat{A}_i^2| \ge \eta \|Y_0\|_2$. If there are r' such columns, we work with a tensor of rank $r - r'$ in the meta-test phase. The reduction in rank increases mean squared error by at most $O(r^2\eta^2)$. Now if we choose $\eta = \sigma/r$ we get a bound of $O(B_1 r^4 \delta^2 + B_2 r^4/N_2)$ on the excess error.

This theorem implies that for a new task, the number of samples needed is $N_2 = O(r^2/\epsilon)$ if we want to achieve a test error of ϵ on the new task. If we were to run a least squares regression on the new task from scratch, the required number of samples would have been $O((d_1 + d_2)/\epsilon)$. As the CP-rank of the tensor A can be smaller (often a constant) than the dimension of the unobserved features, transfer of the knowledge of the tensor A provides a significant reduction in the number of samples on the new task.

4 Method-of-Moments Based Estimation

In this section, we provide a new algorithm that estimates the underlying tensor A and also has optimal dependence on the number of tasks (T) under some additional distributional assumptions. In particular, we will assume $X_i \sim_{\text{iid}} \mathrm{N}(0, \mathrm{I}_{d_1})$, $Y_t \sim_{\text{iid}} \mathrm{N}(0, \mathrm{I}_{d_2})$, and $Z_t \sim_{\text{iid}} \mathrm{N}(0, \mathrm{I}_{d_3})$. Our algorithm is based on repeated applications a method-of-moments based estimator proposed in [28], and we briefly summarize that estimator. Suppose the i-th response is given as $R_i = X_i^\intercal B \alpha_{t(i)} + \varepsilon_i$ and each $X_i \sim_{\text{iid}} \mathrm{N}(0, I_{d_1})$, and $B \in \mathbb{R}^{d_1 \times r}$ has orthonormal columns. Then it is possible to recover B from the top r singular values of the statistic $\frac{1}{N}\sum_{i=1}^N R_i^2 X_i X_i^\intercal$.

Recovering A^1. For our setting, the i-th response is given as $R_i = A(X_i, Y_{t(i)}, Z_{t(i)}) + \varepsilon_i$. If we want to recover the first factor A^1 then we can rewrite the i-th response as

$$R_i = X_i^\intercal A_{(1)}(Z_{t(i)} \otimes Y_{t(i)}) + \varepsilon_i = X_i^\intercal A^1 \underbrace{(A^3 \odot A^2)^\intercal (Z_{t(i)} \otimes Y_{t(i)})}_{:=\alpha_{t(i)}} + \varepsilon_i.$$

Since each X_i is drawn from a standard normal distribution, we can recover A^1 from the top-r singular values of the statistic $\frac{1}{N}\sum_{i=1}^N R_i^2 X_i X_i^\intercal$.

Recovering A^2. We can recover A^2 through a similar method. We can rewrite the i-th response as

$$R_i = Y_{t(i)}^\top A_{(2)}(Z_{t(i)} \otimes X_i) + \varepsilon_i = Y_{t(i)}^\top A^2 \underbrace{W(A^3 \odot A^1)^\top (Z_{t(i)} \otimes X_i)}_{:=\alpha_{t(i)}} + \varepsilon_i$$

Since each Y_t is drawn from a standard normal distribution, we can recover A^2 from the top-r singular values of the statistic $\frac{1}{N}\sum_{i=1}^N R_i^2 Y_{t(i)} Y_{t(i)}^\top$.

ALGORITHM 2: Method-of-Moments Based Estimation

Input: $(X_i, Y_{t(i)}, R_i)$ for $i = 1, \ldots, N$.

1. $UDU^\top \leftarrow \text{top} - r$ SVD of $\frac{1}{N}\sum_{i=1}^N R_i^2 X_i X_i^\top$. Set $\hat{A}^1 = U$.
2. $UDU^\top \leftarrow \text{top} - r$ SVD of $\frac{1}{N}\sum_{i=1}^N R_i^2 Y_{t(i)} Y_{t(i)}^\top$. Set $\hat{A}^2 = U$.

Output: Return $\hat{A}^1$ and $\hat{A}^2$.

Theorem 3. *Suppose $X_i \sim_{\text{iid}} \mathrm{N}(0, \mathrm{I}_{d_1})$, $Y_t \sim_{\text{iid}} \mathrm{N}(0, I_{d_2})$, and $Z_t \sim_{\text{iid}} \mathrm{N}(0, I_{d_3})$. Then the factors $\hat{A}^1$ and $\hat{A}^2$ returned by Algorithm 2 satisfies the following guarantees*

$$\sin\theta(\hat{A}^1, A^1) \le O\left(\sqrt{\frac{d_1 r}{TN}}\right), \text{ and } \sin\theta(\hat{A}^2, A^2) \le O\left(\sqrt{\frac{d_2 r}{N}}\right)$$

with probability at least $1 - T\exp(-\Omega(\min\{d_1, d_2\}))$.

Once Algorithm 2 estimates $\hat{A}^1$ and $\hat{A}^2$, we again estimate $A^{3\top} Z_0$ in the meta-test phase. We can show a meta-test theorem similar to Theorem 2, and the details are provided in the appendix.

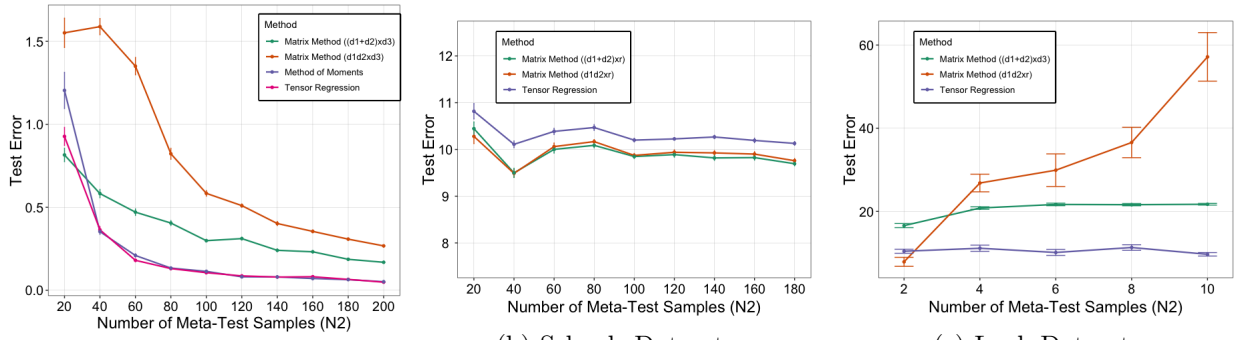
5 Experiments

We first evaluate our tensor-based representation learning through a simulation setup. For this experiment, we generated data from a low-rank tensor of order-3. We chose a tensor of dimension $100 \times 50 \times 50$ and of CP-rank 10. We generated a training dataset of $N = 1000$ points and estimated the factors A^1 and A^2 using both the tensor regression (Algorithm 1) and the method of moments (Algorithm 2). For the meta-test phase, we selected a new test task with observed feature Y_0 of dimension 50 and unobserved feature Z_0 of dimension 50. As described in Section 3.1.1, we estimate $\widehat{A^{3\top} Z_0}$ by substituting the estimated factors from the meta-training step.

We plot the meta-test error for various values of N_2, the number of samples available from the new task. As we increase N_2, test error for predicting outcome on a new test instance X_0 decreases significantly, as shown in Figure 1a. We compare our method with the matrix-based representations for meta learning developed by [28]. They assume that the response from a task t with unobserved feature $Z_t \in \mathbb{R}^r$ and i-th feature X_i is given as

$$R_i = X_i B Z_t + \varepsilon_i$$

where matrix $B \in \mathbb{R}^{d \times r}$. Recall that, for our setting, each training instance is given as $(X_i, Y_{t(i)}, R_i)$. Since [28] assume that there is no available side-information for the tasks, the most natural comparison



(a) Synthetic Dataset (b) Schools Dataset (c) Lenk Dataset

Figure 1: Test-error vs the number of samples from a new task (N_2)

would be to ignore the observable task features Y_t and consider each input as (X_i, R_i). So we consider two natural dimensions of the matrix B. First, we estimate a matrix of dimension $d_1 d_2 \times d_3$ where $X_i \otimes Y_{t(i)}$ is the i-th feature. Second, we estimate a matrix of dimension $(d_1 + d_2) \times d_3$ where $[X_i Y_{t(i)}]$ is the i-th input feature. We compare these two different types of matrix based methods with both tensor regression and method-of-moments based method. As Figure 1a shows both tensor methods perform equally well, but they are significantly better than the matrix methods.

We now consider two real-world datasets. Both the datasets were used in the context of *conditional meta-learning* to show the benefits of task-specific side-information [9].

Schools Dataset [3]. This dataset consists of examination records from $T = 139$ schools (task). The number of samples per task (n_t) varied from 24 to 251. Each instance represents an individual student, and is represented by a feature of dimension $d_1 = 26$. The outcomes are their exam scores. As task specific feature of task t we use $Y_t = \frac{1}{n_t} \sum_{i=1}^{n_t} \phi(x_i)$ where $\phi(x_i)$ is a vector of dimension $d_2 = 50$ constructed from a random Fourier feature map. This is built as follows. First sample v from $\mathrm{Unif}[0, 2\pi]^{d_2}$. Then a matrix $U \in \mathbb{R}^{d_2 \times d_1}$ is sampled from $\mathrm{N}(0, \sigma^2 \mathrm{I})$. Finally, we set

$$\phi(x_i) = \sqrt{\frac{2}{d_2}} \cos\left(U x_i + v\right) \in \mathbb{R}^{d_2}.$$

Lenk Dataset [21, 18]. This is a computer survey data where $T = 180$ people (tasks) rated the likelihood of purchasing one of 20 different personal computers. So there are 20 different samples from each task. The input has dimension $d_1 = 13$ and represents different computers' characteristics, while the output is an integer rating from 0 to 10. As task specific feature of a task t we use $Y_t = \frac{1}{20} \sum_{i=1}^{20} \phi(z_i)$ where $\phi(z_i) = \mathrm{vec}(x_i (R_i, 1)^\intercal)$; the sum is over all z_i-s belonging to the task t.

To construct the meta-training set, we sampled 50 tasks uniformly at random and then sampled n_t ($n_t = 20$ for Schools and $n_t = 10$ for Lenk) responses from each task. Since we do not know the value of r, we also constructed a meta-evaluation set by selecting another set of n_t samples from the selected tasks. The meta-evaluation set was used to select the best value of r during meta-training phase. The meta-test set was constructed by selecting a fixed task and then gradually increasing the number of samples from that task. Figures 1b and 1c respectively compare our method with two different types of matrix based representation learning for different values of N_2. We found that the tensor regression method performs better than the method-of-moments based estimator and only results for Algorithm 1 are shown. Our method performs significantly better than the matrix based methods for Lenk. Although our method performs slightly worse on Schools, the test error

increases by at most 5%. Overall, the performance on the synthetic dataset and two real-world datasets demonstrate the benefits of using tensor based representations for meta-learning.

6 Conclusion and Open Questions

In this work, we develop a tensor-based model of shared representation for learning from a diverse set of tasks. The main difference with previous models on shared representations for meta-learning is that our model incorporates the observable side information of the tasks. We designed two methods to estimate the underlying tensor and compared them in terms of recovery guarantees, required assumptions on the tensor, and mean squared error on a new task.

There are many interesting directions for future work. An interesting direction is to generalize our model and consider non-linear models of shared representations that incorporates the observable side-information of the tasks. Finally, we just leveraged the framework of order-3 tensor in this work, and it would be interesting to see if we can leverage higher order tensors for learning shared representations for meta-learning.

References

[1] A Anandkumar, R Ge, D Hsu, SM Kakade, and M Telgarsky. "Tensor decompositions for learning latent variable models". In: *Journal of Machine Learning Research* 15 (2014), pp. 2773–2832.

[2] Animashree Anandkumar, Rong Ge, and Majid Janzamin. "Guaranteed Non-Orthogonal Tensor Decomposition via Alternating Rank-1 Updates". In: *arXiv preprint arXiv:1402.5180* (2014).

[3] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. "Convex multi-task feature learning". In: *Machine learning* 73.3 (2008), pp. 243–272.

[4] Yu Bai, Minshuo Chen, Pan Zhou, Tuo Zhao, Jason D Lee, Sham Kakade, Huan Wang, and Caiming Xiong. "How Important is the Train-Validation Split in Meta-Learning?" In: *arXiv preprint arXiv:2010.05843* (2020).

[5] Maria-Florina Balcan, Mikhail Khodak, and Ameet Talwalkar. "Provable guarantees for gradient-based meta-learning". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 424–433.

[6] Jonathan Baxter. "A model of inductive bias learning". In: *Journal of artificial intelligence research* 12 (2000), pp. 149–198.

[7] Alberto Bernacchia. "Meta-learning with negative learning rates". In: *arXiv preprint arXiv:2102.00940* (2021).

[8] Giulia Denevi, Carlo Ciliberto, Riccardo Grazzi, and Massimiliano Pontil. "Learning-to-learn stochastic gradient descent with biased regularization". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 1566–1575.

[9] Giulia Denevi, Massimiliano Pontil, and Carlo Ciliberto. "The Advantage of Conditional Meta-Learning for Biased Regularization and Fine Tuning". In: *Advances in Neural Information Processing Systems* 33 (2020).

[10] Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. "Few-shot learning via learning the representation, provably". In: *arXiv preprint arXiv:2002.09434* (2020).

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1126–1135.

[12] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. "Online meta-learning". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 1920–1930.

[13] Katelyn Gao and Ozan Sener. "Modeling and Optimization Trade-off in Meta-learning". In: *Advances in Neural Information Processing Systems* 33 (2020).

[14] M Khodak, M Balcan, and A Talwalkar. "Adaptive Gradient-Based Meta-Learning Methods". In: *Neural Information Processing Systems*. 2019.

[15] Tamara G Kolda and Brett W Bader. "Tensor decompositions and applications". In: *SIAM review* 51.3 (2009), pp. 455–500.

[16] Aditya Krishnan, Sidhanth Mohanty, and David P Woodruff. "On Sketching the q to p Norms". In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques* (2018).

[17] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[18] Peter J Lenk, Wayne S DeSarbo, Paul E Green, and Martin R Young. "Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs". In: *Marketing Science* 15.2 (1996), pp. 173–191.

[19] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. "Multi-Task Deep Neural Networks for Natural Language Understanding". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 4487–4496.

[20] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. "The benefit of multitask representation learning". In: *Journal of Machine Learning Research* 17.81 (2016), pp. 1–32.

[21] Andrew M McDonald, Massimiliano Pontil, and Dimitris Stamos. "New perspectives on k-support and cluster norms". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 5376–5413.

[22] Anusha Nagabandi, Kurt Konolige, Sergey Levine, and Vikash Kumar. "Deep dynamics models for learning dexterous manipulation". In: *Conference on Robot Learning*. PMLR. 2020, pp. 1101–1112.

[23] Sahand Negahban and Martin J Wainwright. "Estimation of (near) low-rank matrices with noise and high-dimensional scaling". In: *The Annals of Statistics* (2011), pp. 1069–1097.

[24] Massimiliano Pontil and Andreas Maurer. "Excess risk bounds for multitask learning with trace norm regularization". In: *Conference on Learning Theory*. PMLR. 2013, pp. 55–76.

[25] Nikunj Saunshi, Yi Zhang, Mikhail Khodak, and Sanjeev Arora. "A sample complexity separation between non-convex and convex meta-learning". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8512–8521.

[26] Sebastian Thrun and Lorien Pratt. "Learning to learn: Introduction and overview". In: *Learning to learn*. Springer, 1998, pp. 3–17.

[27] Ryota Tomioka, Taiji Suzuki, Kohei Hayashi, and Hisashi Kashima. "Statistical performance of convex tensor decomposition". In: *Advances in neural information processing systems*. 2011, pp. 972–980.

[28] Nilesh Tripuraneni, Chi Jin, and Michael I Jordan. "Provable meta-learning of linear representations". In: *International Conference on Machine Learning* (2021).

[29] Nilesh Tripuraneni, Michael Jordan, and Chi Jin. "On the Theory of Transfer Learning: The Importance of Task Diversity". In: *Advances in Neural Information Processing Systems* 33 (2020).

[30] Joel A Tropp. "An Introduction to Matrix Concentration Inequalities". In: *Foundations and Trends in Machine Learning* 8.1-2 (2015), pp. 1–230.

[31] Manasi Vartak, Arvind Thiagarajan, Conrado Miranda, Jeshua Bratman, and Hugo Larochelle. "A meta-learning perspective on cold-start recommendations for items". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 6907–6917.

[32] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.

[33] Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J Lim. "Multimodal Model-Agnostic Meta-Learning via Task-Aware Modulation". In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[34] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019.

[35] Ruohan Wang, Yiannis Demiris, and Carlo Ciliberto. "Structured Prediction for Conditional Meta-Learning". In: *Advances in Neural Information Processing Systems* 33 (2020).

[36] Xi Sheryl Zhang, Fengyi Tang, Hiroko H Dodge, Jiayu Zhou, and Fei Wang. "Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 2487–2495.

[37] Hua Zhou, Lexin Li, and Hongtu Zhu. "Tensor regression with applications in neuroimaging data analysis". In: *Journal of the American Statistical Association* 108.502 (2013), pp. 540–552.

A Non-identifiability of General Model

We first show that the general response model is not identifiable unless all the singular values are one.

Lemma 2. *Consider the response model $R_i = A(X_i, Y_{t(i)}, Z_{t(i)}) + \varepsilon_i$ specified in (1). Then the underlying tensor A is not identifiable if the singular values are not all ones.*

Proof. We show that the statement is also true for simpler matrix based linear representations for multitask learning. In that case, the responses are generated as $R = x^\top Bz$ for an orthonormal matrix B. Now consider the model $R = x^\top BWz$ for a diagonal matrix W. Even if we assume that $\|z\|_2 = 1$, given a choice of W and z, one can choose $W' \neq W$ and $z' \neq z$ s.t. $x^\top BWz = x^\top BW'z'$. A possible choice is $W'(1,1) = \lambda_1 W(1,1)$, $W'(2,2) = W(2,2)/\lambda_2$, $z'_1 = z_1/\lambda_1$, $z'_2 = \lambda_2 z_2$ and $z_1^2/z_2^2 = (\lambda_2^2 - 1)/(1 - 1/\lambda_1^2)$. Note that this choice guarantees that $\left\|z'\right\|_2 = 1$. □

Lemma 3. *Consider the response model $R_i = A(X_i, Y_{t(i)}, Z_{t(i)}) + \varepsilon_i$ specified in (1). Then it is impossible to approximate A^3 either in terms of Frobenius norm or in terms of $\sin\theta$ distance.*

Proof. We construct an example where $d_1 = d_2 = d_3 = d$ and rank $r = d/2$. First consider the tensor $A = [\![\mathrm{I}_r; A^1, A^2, A^3]\!]$ where $A^1 = A^2 = A^3 = \begin{bmatrix} \mathrm{I}_{r\times r} \\ 0_{(d-r)\times r} \end{bmatrix}$. Suppose the observed feature vector $Y = \left(\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}, \ldots, \frac{1}{\sqrt{d}}\right)$ and the unobserved feature vector $Z_1 = \left(\frac{1}{\sqrt{r}}, \frac{1}{\sqrt{r}}, \ldots, \frac{1}{\sqrt{r}}, 0, \ldots, 0\right)$. Then for any feature vector X the expected response on this task is given as

$$R = A(X, Y, Z_1) = \sum_{i=1}^{r} X(i)Y(i)Z_1(i) = \frac{1}{\sqrt{dr}} \sum_{i=1}^{r} X(i) = \frac{\sqrt{2}}{d} \sum_{i=1}^{r} X(i)$$

where the last equality uses $r = d/2$. We now consider a new tensor $B = [\![\mathrm{I}_r; A^1, A^2, B^3]\!]$. The first two factors of B are the same as the first two factors of A, but the third factor is different. Let C^3 be a bidiagonal matrix of dimension $(r+1) \times r$ with the leading diagonal and the diagonal entries just below the leading diagonal entries consisting of all ones.

$$C^3 = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 & \ldots & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & \ldots & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \frac{1}{\sqrt{2}} \end{bmatrix}$$

Then $B^3 = \begin{bmatrix} C^3 \\ 0_{(r-1)\times r} \end{bmatrix}$. The observed task feature Y remains as it was but the new unobserved task feature is given as $Z_2 = \left(\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}, \ldots, \frac{1}{\sqrt{d}}\right)$. Then it can be checked that the new responses are given as

$$R = B(X, Y, Z_2) = \sum_{i=1}^{r} X(i) \frac{1}{\sqrt{d}} \left(\frac{1}{\sqrt{2d}} + \frac{1}{\sqrt{2d}}\right) = \frac{\sqrt{2}}{d} \sum_{i=1}^{r} X(i)$$

Therefore, we have two instances where the first two factors of the underlying tensors (A^1 and A^2) and the observable task feature (Y) are the same, but different choices of the third factor and hidden feature vector give the same response. Moreover, for the given choices of A^3 and B^3 it can be easily verified that $\left\|A^3 - B^3\right\|_F = O(r)$ and $\sin\theta(A^3, B^3) = \left\|A^{3\top}_{\perp} B^3\right\|_{\text{op}} = \frac{1}{\sqrt{2}} = \sin(\pi/4)$. Therefore, even if we exactly know the factors A^1 and A^2, it is impossible to approximate A^3 either in terms of Frobenius norm or in terms of $\sin\theta$ distance. □

B Proof of Theorem 1

Our analysis builds upon the work by Tomioka et al. [27], who analyzed the performance of tensor regression with overlapped Schatten-1 norm. Recall the definition of the term $D_1 = \sqrt{d_1} + \sqrt{d_2} + \sqrt{T} + \sqrt{d_1 d_2} + \sqrt{d_1 T} + \sqrt{d_2 T}$. Tomioka et al. [27] showed that when (1) the true tensor B has multi-way rank bounded by r, i.e. $\max\left\{\text{rank}(B_{(1)}), \text{rank}(B_{(2)}), \text{rank}(B_{(3)})\right\} \leq r$, (2) the number of samples $N \geq c_1 r D_1^2$,and (3) the covariate tensors X_i are drawn iid from standard Gaussian distribution, then choosing $\lambda \geq c_2 \frac{\sigma D_1}{\sqrt{N}}$ guarantees the following:

$$\left\|B - \hat{B}\right\|_F \leq O\left(\frac{\sigma\sqrt{r} D_1}{\sqrt{N}}\right) \tag{9}$$

with high probability. In order to state the main ideas behind the proof and how they can be adapted for our setting, we introduce the following notations.

- $\mathfrak{X} : \mathbb{R}^{d_1 \times d_2 \times T} \to \mathbb{R}^N$ defined as $\mathfrak{X}(W)_i = \langle \mathcal{X}_i, W \rangle$.
- Adjoint operator $\mathfrak{X}^* : \mathbb{R}^N \to \mathbb{R}^{d_1 \times d_2 \times T}$ defined as $\mathfrak{X}(\overrightarrow{\varepsilon}) = \sum_{i=1}^N \varepsilon_i \mathcal{X}_i$.
- Given a tensor $\Delta \in \mathbb{R}^{d_1 \times d_2 \times T}$ write its k-th mode as $\Delta_{(k)}$ as $\Delta_{(k)} = \Delta'_{(k)} + \Delta''_{(k)}$ where the row and column space of $\Delta'_{(k)}$ are orthogonal to the row and column spaces of $B_{(k)}$ respectively.
- A constraint set $\mathcal{C} = \left\{\Delta \in \mathbb{R}^{d_1 \times d_2 \times T} : (1)\text{rank}(\Delta'_{(k)}) \leq 2r \ \forall k \text{ and } (2) \sum_k \left\|\Delta''_{(k)}\right\|_\star \leq 3 \sum_k \left\|\Delta'_{(k)}\right\|_\star\right\}$.

Definition 1 (Restricted Strong Convexity)**.** *There exists a constant $\kappa(\mathfrak{X})$ such that for all tensors in $\Delta \in \mathcal{C}$, we have*

$$\frac{\left\|\mathfrak{X}(\Delta)\right\|_2^2}{N} \geq \kappa(\mathfrak{X}) \|\Delta\|_F^2 .$$

With this definition, Tomioka et al. [27] proves the guarantee in eq. 9 in three steps.

1. If the restricted strong convexity is satisfied with a constant $\kappa(\mathfrak{X})$ and λ is chosen to be at least $\frac{2}{N}\left\|\mathfrak{X}^*(\overrightarrow{\varepsilon})\right\|_{\text{mean}}$[2], then we have the following guarantee:

$$\left\|B - \hat{B}\right\|_F \leq O\left(\frac{\lambda\sqrt{r}}{\kappa(\mathfrak{X})}\right). \tag{10}$$

[2] $\|\cdot\|_{\text{mean}}$ is the dual norm of $\|\cdot\|_S$ and is defined as $\|A\|_{\text{mean}} = 1/3 \sum_{k=1}^3 \left\|W_{(k)}\right\|_{\text{op}}$

2. Gaussian design (i.e. $\mathcal{X}_i \sim \mathrm{N}(0, I_{d_1 \times d_2 \times T})$ satisfies restricted strong convexity with constant $\kappa(\mathfrak{X}) = O(1)$.

3. Additionally, Gaussian design satisfies $\left\|\mathfrak{X}^*(\overrightarrow{\varepsilon})\right\|_{\text{mean}} = O(\sigma D_1 \sqrt{N})$ with high probability.

We now carry out these steps for our setting. First, lemma 4 proves that our setting satisfies *restricted strong convexity* with high probability. As a result of this lemma, we see that our setting satisfies restricted strong convexity with constant $\kappa(\mathfrak{X}) = \frac{\lambda_{\min}(\Sigma_y)\lambda_{\max}(\Sigma)}{36T}$. Compared to [27], we don't get a constant independent of the number of tasks T and it gets worse with increasing T. The constant is $O(1/T)$ because of uniform sampling, where each individual samples one task uniformly at random out of T tasks. For other assignment scheme, the constant could be adjusted appropriately.

Recall, that we need to choose $\lambda > \frac{2}{N}\left\|\mathfrak{X}^*(\overrightarrow{\varepsilon})\right\|_{\text{mean}}$. Lemma 5 lemma provides a lower bound of $O(\sigma D_1/\sqrt{N})$ on $\left\|\mathfrak{X}^*(\overrightarrow{\varepsilon})\right\|_{\text{mean}}$. Now we substitute, $\lambda = O\left(\frac{\sigma D_1}{\sqrt{N}}\right)$ and $\kappa(\mathfrak{X}) = \left(\frac{\lambda_{\min}(\Sigma_y)\lambda_{\max}(\Sigma)}{T}\right)$ in equation 10 to get the main result for our setting. If we fix d_1 and d_2, then the bound scales as $\frac{T^{3/2}}{\sqrt{N}}$. This is worse by a factor of $\sqrt{T}$ compared to the result of [27]. Because of uniform sampling the number of effective samples is $\sqrt{N/T}$, and one should expect a bound of $\frac{\sqrt{T}}{\sqrt{N/T}} = \frac{T}{\sqrt{N}}$.

Lemma 4. *Suppose $X_1, \ldots, X_N \sim_{iid} \mathrm{N}(0, \Sigma)$, $Y_1, \ldots, Y_T \sim_{\text{iid}} \mathrm{N}(0, \Sigma_y)$, and $t(i) \sim Unif\{1, \ldots, T\}$ for each i. If $N \geq O(rD_1^2 \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma_y))$, then for any $\Delta \in \mathcal{C}$, the following holds*

$$\frac{\left\|\mathfrak{X}(\Delta)\right\|_2}{\sqrt{N}} \geq \frac{\sqrt{\lambda_{\min}(\Sigma_y)\lambda_{\min}(\Sigma)}}{6\sqrt{T}}\|\Delta\|_F$$

with probability at least $1 - e^{-\Omega(N/T)}$.

Proof. We first assume $X_1, \ldots, X_N \sim_{\text{iid}} \mathrm{N}(0, \mathrm{I})$ and derive our result. We will then see how a standard trick handles the case of general covariance matrix.

Since Σ_y is a positive-definite matrix, we can right its eigen-decomposition as $\Sigma_y = U^\top D U$ where $U \in\in \mathbb{R}^{d_2 \times d_2}$ is an orthonormal matrix. This implies that there exists a matrix $M = UD^{1/2}$ such that $\Sigma_y = M^\top M$. Moreover the columns of M form an orthogonal basis of $\mathbb{R}^{d_2}$ and L_2 norm of any column of M is at least $\lambda_{\min}^{1/2}(\Sigma_y)$. Given a tensor $\Delta \in \mathcal{C}$ let us define a new tensor $\Delta_M \in \mathbb{R}^{d_1 \times d_2 \times T}$ defined as $\Delta_M(a, b, t) = \Delta_{a:t}^\top M_b$. We first prove the following result.

$$\frac{\left\|\mathfrak{X}(\Delta)\right\|_2}{\sqrt{N}} \geq \frac{\|\Delta_M\|_F}{4\sqrt{T}} - \frac{D_1}{3\sqrt{TN}}\|\Delta_M\|_S \tag{11}$$

We can assume that $\|\Delta_M\|_F = 1$. Otherwise, we construct a new tensor $\tilde{\Delta} = \Delta/\|\Delta_M\|_F$, and the new tensor has $\left\|\tilde{\Delta}_M\right\|_S = 1/3\sum_k \left\|\tilde{\Delta}_{M(k)}\right\|_\star = 1/(3\|\Delta_M\|_F)\sum_k \left\|\Delta_{(k)}\right\|_\star = \|\Delta\|_S / \|\Delta_M\|_F$, and the claim is valid upto rescaling by $\|\Delta\|_F$. We now proceed similar to the proof of proposition 1 in [23]. First, by a peeling argument very similar to the proof of proposition 1 in [23], it is enough to consider the case $\|\Delta_M\|_S \leq t$ and show the following:

$$\frac{\left\|\mathfrak{X}(\Delta)\right\|_2}{\sqrt{N}} \geq \frac{1}{4\sqrt{T}} - \frac{tD_1}{\sqrt{TN}}$$

for all tensors Δ in the set $\mathcal{R}(t) = \left\{\Gamma \in \mathbb{R}^{d_1 \times d_2 \times T} : \|\Gamma_M\|_F = 1 \text{ and } \|\Gamma_M\|_S \leq t\right\}$. Let $S^{N-1} = \left\{u \in \mathbb{R}^N : \|u\|_2 = 1\right\}$ and for all $u \in S^{N-1}$ we define $Z_{u,\Delta} = \left\langle u, \mathfrak{X}(\Delta)\right\rangle$ for any $\Delta \in \mathbb{R}^{d_1 \times d_2 \times T}$. Note that,

$$Z_{u,\Delta} = \sum_{i=1}^{N} u_i \left\langle \mathcal{X}_i, \Delta \right\rangle = \sum_{i=1}^{N} u_i \left\langle X_i Y_{t(i)}^\top, \Delta_{::t(i)} \right\rangle.$$

Moreover,

$$\begin{aligned}
\mathbb{E}\left[(Z_{u,\Delta} - Z_{u',\Delta'})^2\right] &= \frac{1}{T}\sum_{i,a,t} \mathbb{E}\left[\left\{\sum_b Y_t(b)(u_i\Delta(a,b,t) - u_i'\Delta'(a,b,t))\right\}^2 |Y_t\right] \\
&= \frac{1}{T}\sum_{i,a,t}\sum_b \Sigma_y(b,b)(u_i\Delta(a,b,t) - u_i'\Delta'(a,b,t))^2 \\
&+ \frac{1}{T}\sum_{i,a,t}\sum_{b\neq b'} \Sigma_y(b,b')(u_i\Delta(a,b,t) - u_i'\Delta'(a,b,t))^2(u_i\Delta(a,b',t) - u_i'\Delta'(a,b',t))^2
\end{aligned}$$

We now use the eigen-decomposition of $\Sigma_y = M^\top M$ to get the following result.

$$\begin{aligned}
\mathbb{E}\left[(Z_{u,\Delta} - Z_{u',\Delta'})^2\right] &= \frac{1}{T}\sum_{i,a,t} \left\|u_i\Delta_{a:t}M - u_i'\Delta'_{a:t}M\right\|_2^2 \\
&= \frac{1}{T}\sum_{i,a,t,b}\left(u_i\Delta_{a:t}^\top M_b - u_i'\Delta'_{a:t}{}^\top M_b\right)^2 \\
&= \frac{1}{T}\left\|u\otimes\Delta_M - u'\otimes\Delta'_M\right\|_F^2
\end{aligned}$$

where in the last line we write Δ_M for the tensor $\Delta_M(a,b,t) = \Delta_{a:t}^\top M_b$. We now consider a second mean-zero gaussian process $W_{u,\Delta} = \frac{1}{\sqrt{T}}\left(\langle g, u\rangle + \langle G, \Delta_M\rangle\right)$, where $g \in \mathbb{R}^N$ and $G \in \mathbb{R}^{d_1\times d_2\times T}$ are iid with $N(0,1)$ entries. We have

$$\mathbb{E}\left[(W_{u,\Delta} - W_{u',\Delta'})^2\right] = \frac{1}{T}\left\|u - u'\right\|_2^2 + \frac{1}{T}\left\|\Delta_M - \Delta'_M\right\|_F^2.$$

We now verify that the two gaussian processes $(Z_{u,\Delta})$ and $(W_{u,\Delta})$ satisfy the requried conditions of Gordon-Slepian's inquaility (lemma 6). We always have the following inequality $\left\|u\otimes\Delta_M - u'\otimes\Delta'_M\right\|_F^2 \leq \left\|u-u'\right\|_2^2 + \left\|\Delta_M - \Delta'_M\right\|_F^2$ for all pairs (u,Δ) and (u',Δ'). Moreover, if $\Delta = \Delta'$, then $\Delta_M = \Delta'_M$ and equality holds.

Therefore, the two required conditions of Gordon-Slepian inequality(lemma 6) are satisfied for the gaussian process $(W_{\Delta,u})_{\Delta\in\mathcal{R}(t),u\in S^{N-1}}$ and $(Z_{\Delta,u})_{\Delta\in\mathcal{R}(t),u\in S^{N-1}}$ we get the following inequality:

$$\mathbb{E}\inf_{\Delta\in\mathcal{R}(t)}\sup_{u\in S^{N-1}} W_{\Delta,u} \leq \mathbb{E}\inf_{\Delta\in\mathcal{R}(t)}\sup_{u\in S^{N-1}} Z_{\Delta,u}$$

which helps us bound $\inf_{\Delta\in\mathcal{R}(t)}\left\|\mathfrak{X}(\Delta)\right\|_2$.

$$\begin{aligned}
\mathbb{E}\left[\inf_{\Delta\in\mathcal{R}(t)}\left\|\mathfrak{X}(\Delta)\right\|_2\right] &= \mathbb{E}\left[\inf_{\Delta\in\mathcal{R}(t)}\sup_{u\in S^{N-1}} Z_{u,\Delta}\right] \geq \mathbb{E}\left[\inf_{\Delta\in\mathcal{R}(t)}\sup_{u\in S^{N-1}} W_{u,\Delta}\right] \\
&= \mathbb{E}\left[\sup_{u\in S^{N-1}}\frac{1}{\sqrt{T}}\langle g,u\rangle\right] + \mathbb{E}\left[\inf_{\Delta\in\mathcal{R}(t)}\frac{1}{\sqrt{T}}\langle G,\Delta_M\rangle\right] \\
&= \frac{1}{\sqrt{T}}\mathbb{E}\left[\|g\|_2\right] - \frac{1}{\sqrt{T}}\mathbb{E}\left[\sup_{\Delta\in\mathcal{R}(t)}\langle G,\Delta_M\rangle\right] \\
&\geq \frac{\sqrt{N}}{2\sqrt{T}} - \frac{t}{\sqrt{T}}\mathbb{E}\left[\|G\|_{\text{mean}}\right]
\end{aligned}$$

Here the last inequality uses $\langle G,\Delta_M\rangle \leq \|G\|_{\text{mean}}\|\Delta_M\|_S \leq t\|G\|_{\text{mean}}$. Moreover, for a random gaussian matrix of dimension $m_1\times m_2$ the expected value of its operator norm is bounded by $\sqrt{m_1}+\sqrt{m_2}$. This gives us $\mathbb{E}\left[\|G\|_{\text{mean}}\right] = \frac{1}{3}\sum_k \mathbb{E}\left[\left\|G_(k)\right\|_{\text{op}}\right] = D_1/3$.

$$\frac{\mathbb{E}\left[\inf_{\Delta\in\mathcal{R}(t)}\left\|\mathfrak{X}(\Delta)\right\|_2\right]}{\sqrt{N}} \geq \frac{1}{2\sqrt{T}} - \frac{tD_1}{3\sqrt{TN}}$$

Now the function $f(\{X_i\}_{i\in[N]}) = \inf_{\Delta\in\mathcal{R}(t)}\frac{\left\|\mathfrak{X}(\Delta)\right\|_2}{\sqrt{N}}$ is $1/\sqrt{N}$-Lipschitz. Therefore for all $\delta>0$, we have

$$P\left(\inf_{\Delta\in\mathcal{R}(t)}\frac{\left\|\mathfrak{X}(\Delta)\right\|_2}{\sqrt{N}} \leq \frac{1}{2\sqrt{T}} - \frac{tD_1}{3\sqrt{TN}} - \delta\right) \leq 2\exp\left(-\frac{\delta^2 N}{2}\right)$$

Now substituting $\delta = 1/(4\sqrt{T})$ we get that the identity defined in eq. 11 holds. We now relate the norms of Δ_M and Δ. Let $k=1$ and $\Delta_{(1)} = U_1 D_1 V_1^\top$ be the corresponding singular value decomposition. Then $\left\|\Delta_{(1)}\right\|_\star = \text{Tr}(D_1)$. If we define $\tilde{V}_1$ a new matrix with s-th column $\tilde{v}_{1,s}(b,t) = \sum_{b'} v_{1,s}(b',t)M(b',b)$, then we have $\Delta_{M,(1)} = U_1 D_1 \tilde{V}_1^\top$. This implies that $\left\|\Delta_{(1)}\right\|_\star = \left\|\Delta_{M,(1)}\right\|_\star$. Similarly, it can be shown that $\left\|\Delta_{(2)}\right\|_\star = \left\|\Delta_{M,(2)}\right\|_\star$ and $\left\|\Delta_{(3)}\right\|_\star = \left\|\Delta_{M,(3)}\right\|_\star$. This implies that $\|\Delta\|_S = \|\Delta_M\|_S$. For the Frobenius norm we use the fact that the columns of M_b form an orthogonal basis of $\mathbb{R}^{d_2}$ and get $\|\Delta_M\|_F^2 = \sum_{a,b,t}(\Delta_{a:t}^\top M_b)^2 \geq \lambda_{\min}(\Sigma_y)\sum_{a,t}\|\Delta_{a:t}\|_2^2 = \lambda_{\min}(\Sigma_y)\|\Delta\|_F^2$. The previous two relations give us the following bound.

$$\frac{\left\|\mathfrak{X}(\Delta)\right\|_2}{\sqrt{N}} \geq \frac{\|\Delta_M\|_F}{4\sqrt{T}} - \frac{D_1}{3\sqrt{TN}}\|\Delta_M\|_S \geq \frac{\lambda_{\min}^{1/2}(\Sigma_y)\|\Delta\|_F}{4\sqrt{T}} - \frac{D_1}{3\sqrt{TN}}\|\Delta\|_S$$

On the other hand, from the definition of the constraint set $\mathcal{C}$ we get $\|\Delta\|_S = \frac{1}{3}\sum_k\left\|\Delta_{(k)}\right\|_\star \leq \frac{2}{3}\sum_k\left\|\Delta'_{(k)}\right\|_\star \leq \frac{2}{3}\sqrt{2r}\sum_k\left\|\Delta'_{(k)}\right\|_F \leq \frac{2}{3}\sqrt{2r}\sum_k\left\|\Delta_{(k)}\right\|_F = \sqrt{2r}\|\Delta\|_F$. Therefore we have,

$$\frac{\left\|\mathfrak{X}(\Delta)\right\|_2}{\sqrt{N}} \geq \frac{\lambda_{\min}^{1/2}(\Sigma_y)\|\Delta\|_F}{4\sqrt{T}} - \frac{D_1\sqrt{2r}}{3\sqrt{TN}}\|\Delta\|_F \geq \frac{\lambda_{\min}^{1/2}(\Sigma_y)}{6\sqrt{T}}\|\Delta\|_F$$

as long as $N \geq O(rD_1^2/\lambda_{\min}(\Sigma_y)$.

Finally, we consider the case when $X_1, \ldots, X_N \sim_{\text{iid}} \mathrm{N}(0, \Sigma)$ for a general covariance matrix Σ. We define the following operator $T_\Sigma : \mathbb{R}^{d_1 \times d_2 \times T} \to \mathbb{R}^{d_1 \times d_2 \times T}$ defined as $T_\Sigma(\Delta)_{(1)} = \sqrt{\Sigma}\Delta_{(1)}$. We also define a gaussian random operator $\mathfrak{X}' : \mathbb{R}^{d_1 \times d_2 \times T} \to \mathbb{R}^N$ defined as $\mathfrak{X}'_i = \langle \mathcal{X}'_i, T_\Sigma(\Delta) \rangle$. Here for each i, we define $\mathcal{X}'_i$ as:

$$\mathcal{X}'_i(\cdot,\cdot,t) = \begin{cases} \Sigma^{-1/2} X_i & \text{if } t(i) = t \\ 0 & \text{o.w.} \end{cases}$$

Since each $\Sigma^{-1/2} X_i$ is drawn from standard gaussian distribution, we have

$$\frac{\|\mathfrak{X}'(\Delta)\|_2}{\sqrt{N}} \geq \frac{\lambda_{\min}^{1/2}(\Sigma_y)}{6\sqrt{T}} \|T_\Sigma(\Delta)\|_F$$

as long as $N \geq O(rD_1^2 \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma_y)$. In deriving the above result, we use the inequality $\|T_\Sigma(\Delta)\|_S \leq \lambda_{\max}^{1/2}(\Sigma)\|\Delta\|_S$. Now, from the definition $\mathfrak{X}'(\Delta)_i = \langle \mathcal{X}'_i, T_\Sigma(\Delta) \rangle = \langle \mathcal{X}_i, \Delta \rangle = \mathfrak{X}(\Delta)_i$. Moreover, $\|T_\Sigma(\Delta)\|_F = \left\|\sqrt{\Sigma}\Delta_{(1)}\right\|_F \geq \lambda_{\min}^{1/2}(\Sigma)\left\|\Delta_{(1)}\right\|_F = \lambda_{\min}^{1/2}(\Sigma)\|\Delta\|_F$. Substituting this bound on the Frobenius norm gives us the desired result. □

Lemma 5.

$$P\left(\|\mathfrak{X}^*(\overrightarrow{\varepsilon})\|_{mean} \leq 20\sigma\sqrt{N}D_1\right) \geq 1 - 2e^{-\Omega(D_1^2)}.$$

Proof. As $\varepsilon_1, \ldots, \varepsilon_N$ are iid drawn from $N(0, \sigma^2)$ and the Euclidean norm $\|\overrightarrow{\varepsilon}\|_2$ is 1-Lipschitz we get,

$$P\left(\left|\|\overrightarrow{\varepsilon}\|_2 - \mathbb{E}\|\overrightarrow{\varepsilon}\|_2\right| > \sigma\delta\right) \leq 2\exp\left(-\delta^2/2\right)$$

Substituting $\delta = \sqrt{N}$ and observing that $\mathbb{E}\|\overrightarrow{\varepsilon}\|_2 \leq 4\sigma\sqrt{N}$, we get that with probability at least $1 - \exp(-\Omega(N))$, $\|\overrightarrow{\varepsilon}\|_2$ is bounded by $5\sigma\sqrt{N}$. We will write $\mathcal{E}$ to denote this event.

$$\|\mathfrak{X}^*(\overrightarrow{\varepsilon})\|_{\text{mean}} = \frac{1}{3}\sum_{k=1}^{3}\left\|\mathfrak{X}^*(\overrightarrow{\varepsilon})_{(k)}\right\|_{\text{op}}$$

We now bound the operator norm of each of the three modes of $\mathfrak{X}^*(\overrightarrow{\varepsilon})$ separately. Our proof follows the main ideas of the proof of Corollary 10.10 of [34]. Since $\mathfrak{X}^*(\overrightarrow{\varepsilon})_{(1)} \in \mathbb{R}^{d_1 \times d_2 T}$, we choose 1/4-cover $\left\{u^1, \ldots, u^{M_1}\right\}$ of the set $S^{d_1-1} = \left\{u \in \mathbb{R}^{d_1} : \|u\|_2 = 1\right\}$, and 1/4-cover $\left\{v^1, \ldots, v^{M_2}\right\}$ of the set $S^{d_2T-1} = \left\{v \in \mathbb{R}^{d_2T} : \|v\|_2 = 1\right\}$. Note that, we can always choose the covers so that $M_1 \leq 9^{d_1}$ and $M_2 \leq 9^{d_2T}$.

$$\left\|\mathfrak{X}^*(\overrightarrow{\varepsilon})_{(1)}\right\|_{\text{op}} = \sup_{v \in S^{d_2T-1}}\left\|\mathfrak{X}^*(\overrightarrow{\varepsilon})_{(1)}v\right\|_2 \leq \frac{1}{4}\left\|\mathfrak{X}^*(\overrightarrow{\varepsilon})_{(1)}\right\|_{\text{op}} + \max_{l \in [M_2]}\left\|\mathfrak{X}^*(\overrightarrow{\varepsilon})_{(1)}v^l\right\|_2$$

Similarly one can show that

$$\left\|\mathfrak{X}^*(\overrightarrow{\varepsilon})_{(1)}v^l\right\|_2 \leq \frac{1}{4}\left\|\mathfrak{X}^*(\overrightarrow{\varepsilon})_{(1)}\right\|_{\text{op}} + \max_{j \in [M_1]}\left\langle u^j, \mathfrak{X}^*(\overrightarrow{\varepsilon})_{(1)}v^l\right\rangle$$

This establishes the following bound on the operator norm in terms of the covers.

$$\left\|\mathfrak{X}^*(\overrightarrow{\varepsilon})_{(1)}\right\|_{\text{op}} \leq 2 \max_{j\in[M_1],l\in[M_2]} \left|Z^{jl}\right| \quad \text{where } Z^{jl} = \left\langle u^j, \mathfrak{X}^*(\overrightarrow{\varepsilon})_{(1)} v^l \right\rangle$$

Using the definition of $\mathfrak{X}^*(\overrightarrow{\varepsilon})$, we get

$$Z^{jl} = \sum_{i=1}^{N} \varepsilon_i \left\langle u^j, \mathcal{X}_{i,(1)} v^l \right\rangle = \sum_{i=1}^{N} \varepsilon_i \sum_{a,b} X_i(a) Y_{t(i)}(b) v^l(b,t(i)) u^j(a) \tag{12}$$

Since each entry of X_i is drawn iid from $N(0,1)$, Z^{jl} is a zero mean gaussian random variable with variance

$$\sum_{i=1}^{N} \varepsilon_i^2 \sum_a \{u^j(a)\}^2 \left(\sum_b v^l(b,t(i)) Y_{t(i)}(b)\right)^2 \leq \sum_{i=1}^{N} \varepsilon_i^2 \sum_a \{u^j(a)\}^2 \sum_{b_1} Y_{t(i)}^2(b_1) \sum_{b_2} \{v^l(b_2,t(i))\}^2 \leq \sum_{i=1}^{N} \varepsilon_i^2$$

The last inequality uses – the observed task features are normalized, $u \in S^{d_1-1}$ and $v \in S^{d_2 T-1}$. Conditioned on the event the variance of each Z^{jl} is bounded by $5\sigma\sqrt{N}$. Now we can provide a high probability bound on the operator norm.

$$\begin{aligned} P\left(\left\|\mathfrak{X}^*(\overrightarrow{\varepsilon})_{(1)}\right\|_{\text{op}} \geq T_N\right) &\leq P\left(2 \max_{j\in[M_1],l\in[M_2]} \left|Z^{jl}\right| \geq T_N\right) \\ &\leq \sum_{j\in[M_1]} \sum_{j\in[M_2]} P\left(\left|Z^{jl}\right| \geq T_N/2\right) \\ &\leq 2M_1 M_2 \exp\left\{-\frac{T_N^2}{50\sigma^2 N}\right\} \leq 2\exp\left\{-\frac{T_N^2}{50\sigma^2 N} + (d_1+d_2T)\log 9\right\} \end{aligned}$$

If we choose $T_N \geq 20\sigma\sqrt{N} D_1$, we get

$$P\left(\left\|\mathfrak{X}^*(\overrightarrow{\varepsilon})_{(1)}\right\|_{\text{op}} \geq 20\sigma\sqrt{N} D_1\right) \leq 2\exp\left\{-2D_1^2\right\}$$

By a similar argument, we can bound the operator norm of the other two modes of $\mathfrak{X}^*(\overrightarrow{\varepsilon})$.

□

Lemma 6 (Gordon's Inequality)**.** *Let $(X_{ut})_{u\in U, t\in T}$ and $(Y_{ut})_{u\in U, t\in T}$ be two mean zero Gaussian processes indexed by pairs of points (u,t) in a product space $U \times T$. Assume that we have*

1. $\mathbb{E}(X_{ut} - X_{us})^2 \leq \mathbb{E}(Y_{ut} - Y_{us})^2$ *for all* u, t, s.
2. $\mathbb{E}(X_{ut} - X_{vs})^2 \geq \mathbb{E}(Y_{ut} - Y_{vs})^2$ *for all* $u \neq v$ *and* t, s.

Then we have

$$\mathbb{E} \inf_{u\in U} \sup_{t\in T} X_{ut} \leq \mathbb{E} \inf_{u\in U} \sup_{t\in T} Y_{ut}$$

Proof. See [17], chapter 3. □

C Formal Statement and Proof of Lemma 1

First, we state weaker set of assumptions under which the bounds of lemma 1 holds. We will make the following assumptions about the underlying tensor $A = [\mathrm{I}_r; A^1, A^2, A^3]$.

(A1) The columns of the factors of A are orthogonal i.e. $\left\langle A^1_i, A^1_j \right\rangle = \left\langle A^2_i, A^2_j \right\rangle = \left\langle A^3_i, A^3_j \right\rangle = 0$ for all $i \neq j$.

(A2) The components have bounded norm i.e. $\exists p < 3$, $\max\left\{\left\|A^{1^\top}\right\|_{2\to p}, \left\|A^{2^\top}\right\|_{2\to p}, \left\|A^{3^\top}\right\|_{2\to p}\right\} \leq 1 + o(1)$.

(A3) Rank is bounded i.e. $r = o(d)$.

Recall the definition of $\mathcal{Z}$, the matrix of unobserved features.

$$\mathcal{Z} = \begin{bmatrix} Z_1 & \cdots & Z_T \end{bmatrix}^T \in \mathbb{R}^{T \times d_3} \tag{13}$$

Let $\mathcal{Z}(s)$ denote the s-th column of the matrix $\mathcal{Z}$. We will make the following assumptions about $\mathcal{Z}$.

(Z1) $\frac{1}{d_3^{0.5+\gamma}} \mathrm{I}_{d_3} \preccurlyeq \mathcal{Z}^\top \mathcal{Z} \preccurlyeq \frac{1}{\sqrt{d_3}} \mathrm{I}_{d_3}$ for some $\gamma > 0$.

(Z2) $\kappa(\mathcal{Z}^\top \mathcal{Z}) = \frac{\lambda_{\max}(\mathcal{Z}^\top \mathcal{Z})}{\lambda_{\min}(\mathcal{Z}^\top \mathcal{Z})} \leq 1 + O(\sqrt{r/d})$.

Lemma 7. *Suppose tensor A has rank r CP-decomposition $A = [\mathrm{I}_r; A^1, A^2, A^3]$ and satisfies the assumptions (A1)-(A3), the matrix of unobserved features $\mathcal{Z}$ satisfies assumptions (Z1)-(Z2), and $N = \Omega\left(\frac{\sigma^2 T^2 D_1^2 r}{\lambda^2_{\min}(\Sigma_y)\lambda^2_{\min}(\Sigma)} \min\left\{\frac{1}{36}, \frac{\log r}{d}\right\}\right)$. Then we have the following guarantees:*

$$\max\left\{\left\|\widehat{A^1} - A^1\right\|_F, \left\|\widehat{A^2} - A^2\right\|_F\right\} \leq \tilde{O}\left(\frac{\sigma T D_1 r}{\sqrt{\lambda_{\min}(\mathcal{Z}^\top \mathcal{Z})}\lambda_{\min}(\Sigma_y)\lambda_{\min}(\Sigma)\sqrt{N}}\right),$$
$$\left\|\widehat{\mathcal{Z}A^3} - \mathcal{Z}A^3\right\|_F \leq \tilde{O}\left(\frac{\sigma\sqrt{\lambda_{\max}(\mathcal{Z}^\top \mathcal{Z})} T D_1 r^{1.5}}{\sqrt{\lambda_{\min}(\mathcal{Z}^\top \mathcal{Z})}\lambda_{\min}(\Sigma_y)\lambda_{\min}(\Sigma)\sqrt{N}}\right)$$

Proof. We will be using the robust tensor decomposition algorithm proposed by [2]. We first review the necessary conditions and the guarantees of their main algorithm. We are given a tensor $\hat{S} = S + \Psi$ where $S \in \mathbb{R}^{d_1 \times d_2 \times T}$ has rank-r decomposition $S = [W; S^1, S^2, S^3]$ and Ψ is a noise tensor with spectral norm $\psi = \|\Psi\|$. We will write the singular values as $w_1 \geq w_2 \geq \ldots \geq w_r > 0$ with $\gamma = w_1/w_r$. Let $d = \max\{d_1, d_2, T\}$. Moreover, suppose the tensor S satisfies the following conditions.

(S1) The components are incoherent i.e. $\max_{i \neq j}\left\{\left|\left\langle s^1_i, s^1_j \right\rangle\right|, \left|\left\langle s^2_i, s^2_j \right\rangle\right|, \left|\left\langle s^3_i, s^3_j \right\rangle\right|\right\} \leq \frac{\mathrm{polylog}(d)}{\sqrt{d}}$.

(S2) The components have bounded norm i.e. $\max\left\{\left\|S^1\right\|_{\mathrm{op}}, \left\|S^2\right\|_{\mathrm{op}}, \left\|S^3\right\|_{\mathrm{op}}\right\} \leq 1 + O\left(\sqrt{r/d}\right)$ and for some $p < 3$, $\max\left\{\left\|S^{1^\top}\right\|_{2\to p}, \left\|S^{1^\top}\right\|_{2\to p}, \left\|S^{1^\top}\right\|_{2\to p}\right\} \leq 1 + o(1)$. [3]

[3] For a matrix $M \in \mathbb{R}^{m \times n}$, define $\|M\|_{q\to p} = \sup_{\|u\|_q = 1} \|Mu\|_p$.

(S3) Rank is bounded i.e. $r = o(d^{1.5}/\text{polylog}(d))$.

(S4) $\psi \leq \min\left\{\frac{1}{6}, O\left(\sqrt{\frac{\log r}{d}}\right)\right\}$.

(S5) Tensor norm of S is bounded i.e. $\|S\| \leq O(w_1)$ and $\left\|\sum_{i\neq j} w_i \langle s_i^1, s_j^1\rangle \langle s_i^2, s_j^2\rangle s_j^3\right\| \leq \frac{w_1 \text{polylog}(d)\sqrt{r}}{d}$.

(S6) The maximum ratio of the weights satisfy $\gamma = O\left(\min\left\{\sqrt{d}, d^{1.5}/r\right\}\right)$.

When the underlying tensor S satisfies the conditioned above, [2] proposed an algorithm that returns an estimate $[\widehat{W}; \widehat{S^1}, \widehat{S^2}, \widehat{S^3}]$ with the following guarantees:

$$\max\left\{\left\|\widehat{S^1} - S^1\right\|_F, \left\|\widehat{S^2} - S^2\right\|_F, \left\|\widehat{S^3} - S^3\right\|_F\right\} \leq \tilde{O}\left(\frac{\sqrt{r}\psi}{w_r}\right) \text{ and } \left\|\widehat{W} - W\right\|_2 \leq \tilde{O}(\sqrt{r}\psi)$$

Consider the tensor $B = A \times_3 \mathcal{Z}$. We now check that the conditions (S1)-(S6) are also satisfied when we consider the tensor B. B has the following rank r CP-decomposition $B = [G^{-1}; A^1, A^2, \mathcal{Z}A^3G]$ where the i-th entry of the diagonal matrix G is $G_i = 1/\left\|\mathcal{Z}A_i^3\right\|_2$. This means that the rank of B is also r and (S3) is satisfied. The singular values of B are given by $\left\|\mathcal{Z}A_i^3\right\|_2$ for $i \in [r]$. As each column of A^3 is normalized, the following result holds for any i.

$$\lambda_{\min}(\mathcal{Z}^\top\mathcal{Z}) \leq \left\|\mathcal{Z}A_i^3\right\|_2^2 \leq \lambda_{\max}(\mathcal{Z}^\top\mathcal{Z})$$

Therefore, the maximum ratio of singular values of the tensor B is bounded by $\sqrt{\lambda_{\max}(\mathcal{Z}^\top\mathcal{Z})/\lambda_{\min}(\mathcal{Z}^\top\mathcal{Z})}$ which is bounded by $\sqrt{d}$ and assumption (S6) is satisfied.

We will write C to denote the matrix $\mathcal{Z}A^3G$. Note that the i-th column of C is given as $\mathcal{Z}A_i^3/\left\|\mathcal{Z}A_i^3\right\|_2$. In order to check condition (S1), we need to verify $\left|\langle C_i, C_j\rangle\right| \leq \frac{\text{polylog}(d)}{\sqrt{d}}$. Note that $\left|\langle C_i, C_j\rangle\right| = \frac{\left|\langle \mathcal{Z}A_i^3, \mathcal{Z}A_j^3\rangle\right|}{\left\|\mathcal{Z}A_i^3\right\|_2\left\|\mathcal{Z}A_j^3\right\|_2} \leq \frac{\left|\langle \mathcal{Z}A_i^3, \mathcal{Z}A_j^3\rangle\right|}{\lambda_{\min}(\mathcal{Z}^\top\mathcal{Z})}$.

$$\frac{1}{2}(A_i^3 + A_j^3)^\top\mathcal{Z}^\top\mathcal{Z}\frac{1}{2}(A_i^3 + A_j^3) = A_i^{3\top}\mathcal{Z}^\top\mathcal{Z}A_i^3 + A_j^{3\top}\mathcal{Z}^\top\mathcal{Z}A_j^3 + 2A_i^{3\top}\mathcal{Z}^\top\mathcal{Z}A_j^3$$

Using assumption (Z2) we get,

$$2A_i^{3\top}\mathcal{Z}^\top\mathcal{Z}A_j^3 \leq \frac{1}{\sqrt{d_3}} - A_i^{3\top}\mathcal{Z}^\top\mathcal{Z}A_i^3 - A_j^{3\top}\mathcal{Z}^\top\mathcal{Z}A_j^3 \leq \frac{1}{\sqrt{d_3}} - \frac{2}{d_3^{0.5+\gamma}} = O\left(\frac{1}{\sqrt{d_3}}\right)$$

In order to check (S2), notice that $\left\|B^1\right\|_{\text{op}} = \left\|A^1\right\|_{\text{op}} \leq 1 + O\left(\sqrt{r/d}\right)$. Same result holds for B^2. For the third factor we have, $\left\|B^3\right\|_{\text{op}} = \left\|\mathcal{Z}A^3G\right\|_{\text{op}} \leq \|\mathcal{Z}\|_{\text{op}}\left\|A^3\right\|_{\text{op}} \max_i \frac{1}{\left\|\mathcal{Z}A_i^3\right\|_2} \leq \sqrt{\frac{\lambda_{\max}(\mathcal{Z}^\top\mathcal{Z})}{\lambda_{\min}(\mathcal{Z}^\top\mathcal{Z})}}\left\|A^3\right\|_{\text{op}} \leq \left(1 + O(\sqrt{r/d})\right)$. For the second part of (S2), we just need to bound

$\left\|(\mathcal{Z}A^3G)^\top\right\|_{2\to p}$.

$$\begin{aligned}
\left\|(\mathcal{Z}A^3G)^\top\right\|_{2\to p} &= \left\|\mathcal{Z}A^3G\right\|_{\frac{p}{p-1}\to 2} \quad \text{[By lemma 8 of [16]} \\
&= \max_{x:\|x\|_{p/(p-1)}=1}\left\|\mathcal{Z}A^3Gx\right\|_2 \le \|\mathcal{Z}\|_{\text{op}} \max_{x:\|x\|_{p/(p-1)}=1}\left\|A^3Gx\right\|_2 \\
&= \|\mathcal{Z}\|_{\text{op}}\left\|A^3G\right\|_{\frac{p}{p-1}\to 2} = \|\mathcal{Z}\|_{\text{op}}\left\|GA^{3^\top}\right\|_{2\to p} \\
&\le \|\mathcal{Z}\|_{\text{op}}\|G\|_p\left\|A^{3^\top}\right\|_{2\to p} \le \sqrt{\frac{\lambda_{\max}(\mathcal{Z}^\top\mathcal{Z})}{\lambda_{\min}(\mathcal{Z}^\top\mathcal{Z})}}\left\|A^{3^\top}\right\|_{2\to p} \le 1+o(1)
\end{aligned}$$

The last line uses (A2), (A3), and (Z2).

If we write $\hat{B} = B + \Psi$, from the guarantees of tensor regression (theorem 1) we have $\psi = \|\Psi\| \le \|\Psi\|_F \le O\left(\frac{\sigma T D_1\sqrt{r}}{\lambda_{\min}(\Sigma_y)\lambda_{\min}(\Sigma)\sqrt{N}}\right)$. So as long as, $N \ge O\left(\frac{\sigma^2T^2D_1^2r}{\lambda_{\min}(\Sigma_y)^2\lambda_{\min}(\Sigma)^2}\min\left\{\frac{1}{36}, \frac{\log r}{d}\right\}\right)$, condition (S4) is satisfied.

We now verify condition (S5). Fix three vectors $a \in \mathbb{R}^{d_1}, b \in \mathbb{R}^{d_2}$, and $c \in \mathbb{R}^T$ with $\|a\|_2 = \|b\|_2 = \|c\|_2 = 1$.

$$\begin{aligned}
B(x,y,z) &= \sum_{i=1}^r G_i^{-1}(A^{1^\top}a)_i(A^{2^\top}b)_i((\mathcal{Z}AG)^\top c)_i \\
&\le \max_i G_i^{-1}\left\|A^{1^\top}a\right\|_3\left\|A^{2^\top}b\right\|_3\left\|(\mathcal{Z}AG)^\top c\right\|_3 \\
&\le \max_i G_i^{-1}\left\|A^{1^\top}\right\|_{2\to 3}\|a\|_2\left\|A^{2^\top}\right\|_{2\to 3}\|b\|_2\left\|(\mathcal{Z}AG)^\top\right\|_{2\to 3}\|c\|_2 \\
&\le \max_i G_i^{-1}\left\|A^{1^\top}\right\|_{2\to p}\left\|A^{2^\top}\right\|_{2\to p}\left\|(\mathcal{Z}AG)^\top\right\|_{2\to p} = O(\max_i G_i^{-1})
\end{aligned}$$

The first inequality uses Corollary 3 from [2], which applies Hölder's inequality three times. The inequality on the following fact. For any matrix M, $\|M\|_{2\to 3} \le \|M\|_{2\to p}$ which follows from the definition of $\|\cdot\|_{2\to p}$ and $p < 3$. Finally, the second part of condition (S5) follows immediately as the columns of A^1 and A^2 are orthonormal.

Therefore, we conclude that the tensor $B = A \times_3 \mathcal{Z}$ satisfies assumptions (S1)-(S6) and we can apply robust tensor decomposition algorithm from [2]. As we can write B as $B = [\![G^{-1}; A^1, A^2, \mathcal{Z}A^3G]\!]$, we get the following guarantees.

$$\max\left\{\left\|\widehat{A^1} - A^1\right\|_F, \left\|\widehat{A^2} - A^2\right\|_F, \left\|\widehat{\mathcal{Z}A^3G} - \mathcal{Z}A^3G\right\|_F\right\} \le \tilde{O}\left(\frac{\sigma T D_1 r}{\sqrt{\lambda_{\min}(\mathcal{Z}^\top\mathcal{Z})}\lambda_{\min}(\Sigma_y)\lambda_{\min}(\Sigma)\sqrt{N}}\right) \quad \text{and}$$

$$\left\|\widehat{G^{-1}} - G^{-1}\right\|_2 \le \tilde{O}\left(\frac{\sigma T D_1 r}{\lambda_{\min}(\Sigma_y)\lambda_{\min}(\Sigma)\sqrt{N}}\right)$$

Since we also have an estimate of G^{-1} we can estimate $\mathcal{Z}A^3$ by $\widehat{\mathcal{Z}A^3G}\widehat{G^{-1}}$. Then we have the

following guarantee.

$$\begin{aligned}
\left\|\widehat{\mathcal{Z}A^3} - \mathcal{Z}A^3\right\|_F &= \left\|\widehat{\mathcal{Z}A^3G}\widehat{G^{-1}} - \mathcal{Z}A^3GG^{-1}\right\|_F \\
&= \left\|\widehat{\mathcal{Z}A^3G}\widehat{G^{-1}} - \mathcal{Z}A^3G\widehat{G^{-1}} + \mathcal{Z}A^3G\widehat{G^{-1}} - \mathcal{Z}A^3GG^{-1}\right\|_F \\
&\le \left\|\widehat{\mathcal{Z}A^3G} - \mathcal{Z}A^3G\right\|_F \left\|\widehat{G^{-1}}\right\|_F + \left\|\mathcal{Z}A^3G\right\|_F \left\|\widehat{G^{-1}} - G^{-1}\right\|_2 \\
&\le \left\|\widehat{\mathcal{Z}A^3G} - \mathcal{Z}A^3G\right\|_F \left(\left\|\widehat{G^{-1}} - G^{-1}\right\|_F + \left\|G^{-1}\right\|_F\right) + \sqrt{r}\left\|\mathcal{Z}A^3G\right\|_{\text{op}} \left\|\widehat{G^{-1}} - G^{-1}\right\|_2 \\
&= \tilde{O}\left(\frac{\sqrt{\kappa(\mathcal{Z}^\top\mathcal{Z})}\sigma T D_1 r^{1.5}}{\lambda_{\min}(\Sigma_y)\lambda_{\min}(\Sigma)\sqrt{N}}\right)
\end{aligned}$$

□

D Formal Statement and Proof of Theorem 2

Theorem 4. *Each covariate vector X_i is mean-zero, satisfies $\mathbb{E}[X_iX_i^\top] = \Sigma$ and Σ-sub-gaussian, and $\max\left\{\left\|\hat{A}^1 - A^1\right\|_F, \left\|\hat{A}^2 - A^2\right\|_F\right\} \le \delta$. Additionally, suppose that $N_2 \ge O\left(r\left(\|Y_0\|_2^2 \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\right)^2 \log(2/\delta_1)\right)$, and $\left|Y_0^\top \hat{A}^2{}_i\right| \ge \eta\|Y_0\|_2$ for all $i \in [r]$. Then with probability at least $1-\delta_1$ we have*

$$\mathbb{E}_{X_0}\left[\left(A(X_0,Y_0,Z_0) - \hat{A}(X_0,Y_0,\widehat{Z}_0\right)^2\right] = O\left(\frac{B_1}{\eta^2}r^2\delta^2 + \frac{B_2}{\eta^2}\frac{r^2}{N_2}\right),$$

for $B_1 = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\mathbb{E}[\|X_0\|_2^2]\|Y_0\|_2^2\|Z_0\|_2^2$ and $B_2 = \frac{\mathbb{E}[\|X_0\|_2^2]}{\lambda_{\min}(\Sigma)}$.

Proof. Mean squared error is given as

$$\begin{aligned}
&\mathbb{E}_{X_0}\left[\left(\hat{A}(X_0,Y_0,\hat{Z}_0) - A(X_0,Y_0,Z_0)\right)^2\right] \\
&= \mathbb{E}_{X_0}\left[\left((Y_0^\top\hat{A}^2 \odot X_0^\top\hat{A}^1)\widehat{A^{3\top}Z_0} - (Y_0^\top A^2 \odot X_0^\top A^1)A^{3\top}Z_0\right)^2\right] \qquad (14)
\end{aligned}$$

We will write $u \in \mathbb{R}^r$ to denote the vector $(Y_0^\top A^2 \odot X_0A^1)$ and $\hat{u}$ to denote its estimate $(Y_0^\top\hat{A}^2 \odot X_0\hat{A}^1)$.

$$\begin{aligned}
&\mathbb{E}_{X_0}\left[\left(\hat{u}^\top\widehat{A^{3\top}Z_0} - u^\top A^{3\top}Z_0\right)^2\right] = E_{X_0}\left[\left((\hat{u}-u)^\top\widehat{A^{3\top}Z_0} + u^\top(\widehat{A^{3\top}Z_0} - A^{3\top}Z_0)\right)^2\right] \\
&\le 2\mathbb{E}_{X_0}[\|\hat{u}-u\|_2^2]\left\|\widehat{A^{3\top}Z_0}\right\|_2^2 + 2\mathbb{E}_{X_0}[\|u\|_2^2]\left\|\widehat{A^{3\top}Z_0} - A^{3\top}Z_0\right\|_2^2
\end{aligned}$$

Now, $\|u\|_2^2 = \sum_{i=1}^r (Y_0^\top A_i^2)^2(X_0^\top A_i^1)^2 \le \sum_{i=1}^r \|Y_0\|_2^2\|A_i^2\|_2^2\|X_0\|_2^2\|A_i^1\|_2^2 = r\|Y_0\|_2^2\|X_0\|_2^2$. As X_0 is drawn from a zero-mean, Σ-subgaussian distribution, we have $\mathbb{E}[\|u\|_2^2] = O(\|Y_0\|_2^2 r\mathbb{E}[\|X_0\|_2^2])$.

Moreover,

$$\begin{aligned}
\|\hat{u}-u\|_2^2 &= \sum_{i=1}^{r}\left[(Y_0^\top A_i^2)(X_0^\top A_i^1)-(Y_0^\top \hat{A}_i^2)(X_0^\top \hat{A}_i^1)\right]^2 \\
&= \sum_{i=1}^{r}\left[Y_0^\top A_i^2(X_0^\top A_i^1 - X_0^\top \hat{A}_i^1)+X_0^\top \hat{A}_i^1(Y_0^\top A_i^2 - Y_0^\top \hat{A}_i^2)\right]^2 \\
&\le 2\sum_{i=1}^{r}(Y_0^\top A_i^2)^2(X_0^\top A_i^1 - X_0^\top \hat{A}_i^1)^2 + 2\sum_{i=1}^{r}(X_0^\top \hat{A}_i^1)^2(Y_0^\top A_i^2 - Y_0^\top \hat{A}_i^2)^2 \\
&\le 2\sum_{i=1}^{r}\|Y_0\|_2^2\left\|A_i^2\right\|_2^2\|X_0\|_2^2\left\|A_i^1\hat{A}_i^1\right\|_2^2 + 2\sum_{i=1}^{r}\|X_0\|_2^2\left\|\hat{A}_i^1\right\|_2^2\|Y_0\|_2^2\left\|\hat{A}_i^2 - A_i^2\right\|_2^2 \\
&= 2\|X_0\|_2^2\|Y_0\|_2^2\left(\left\|A^1-\hat{A}^1\right\|_F^2+\left\|A^2-\hat{A}^2\right\|_F^2\right) \\
&\le 4\|X_0\|_2^2\|Y_0\|_2^2\,\delta^2
\end{aligned}$$

Therefore, $\mathbb{E}[\|\hat{u}-u\|_2^2] = O(\|Y_0\|_2^2\,\mathbb{E}[\|X_0\|_2^2]\delta^2)$.

This gives us a bound of

$$O\left(\|Y_0\|_2^2\,\mathbb{E}[\|X_0\|_2^2]\left(\delta^2\left\|\widehat{A^{3^\top}Z_0}\right\|_2^2 + r\left\|\widehat{A^{3^\top}Z_0} - A^{3^\top}Z_0\right\|_2^2\right)\right) \tag{15}$$

on the mean-squared error. We first bound $\left\|\widehat{A^{3^\top}Z_0} - A^{3^\top}Z_0\right\|_2^2$. Recall that if we write $\widehat{V} = (Y_0^\top\hat{A}^2 \odot \mathcal{X}_0\hat{A}^1)$, then we can write $\widehat{A^{3^\top}Z_0}$ as $\left(\widehat{V}^\top\widehat{V}\right)^{-1}\widehat{V}^\top R$.

$$\begin{aligned}
\widehat{A^{3^\top}Z_0} - A^{3^\top}Z_0 &= \left(\widehat{V}^\top\widehat{V}\right)^{-1}\widehat{V}^\top R - A^{3^\top}Z_0 \\
&= \left(\widehat{V}^\top\widehat{V}\right)^{-1}\widehat{V}^\top(VA^{3^\top}Z_0+\boldsymbol{\varepsilon}) - A^{3^\top}Z_0 \\
&= \left(\widehat{V}^\top\widehat{V}\right)^{-1}\widehat{V}^\top\boldsymbol{\varepsilon} \\
&\quad + \left(\widehat{V}^\top\widehat{V}\right)^{-1}\widehat{V}^\top VA^{3^\top}Z_0 - A^{3^\top}Z_0
\end{aligned}$$

Lemmas 8 and 9 respectively bound the bias and the variance term. Substituting these bounds we get $\left\|\widehat{A^{3^\top}Z_0} - A^{3^\top}Z_0\right\|_2^2 = O\left(\frac{C_1}{\eta^2}\frac{r}{N_2} + \frac{C_2}{\eta^2}r\delta^2\right)$ for $C_1 = \frac{1}{\|Y_0\|_2^2\lambda_{\min}(\Sigma)}$ and $C_2 = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\|Z_0\|_2^2$. We now consider the remaining term $\left\|\hat{Z}_0\right\|_2^2$ in the upper bound on MSE (eq. (15)).

$$\begin{aligned}
\left\|\widehat{A^{3^\top}Z_0}\right\|_2^2 &= \left\|\left(\widehat{V}^\top\widehat{V}\right)^{-1}\widehat{V}^\top R\right\|_2^2 = \left\|\left(\widehat{V}^\top\widehat{V}\right)^{-1}\widehat{V}^\top(VA^{3^\top}Z_0+\boldsymbol{\varepsilon})\right\|_2^2 \\
&\le 2\left\|\left(\widehat{V}^\top\widehat{V}\right)^{-1}\widehat{V}^\top\boldsymbol{\varepsilon}\right\|_2^2 + 2\left\|\left(\widehat{V}^\top\widehat{V}\right)^{-1}\widehat{V}^\top VA^{3^\top}Z_0\right\|_2^2
\end{aligned}$$

The first term can be bounded by $\frac{C_1}{\eta^2}\frac{r}{N_2}$ by lemma 8. The second term can be bounded as follows.

$$
\begin{aligned}
&\left\|\left(\widehat{V}^\top\widehat{V}\right)^{-1}\widehat{V}^\top V A^{3^\top} Z_0\right\|_2^2 \\
&\leq \left\|(\widehat{V}^\top\widehat{V})^{-1}\widehat{V}^\top\right\|_{\text{op}}^2 \|V\|_{\text{op}}^2 \left\|A^{3^\top} Z_0\right\|_2^2 \\
&\leq \left\|(\widehat{V}^\top\widehat{V})^{-1}\right\|_{\text{op}} O(\|Y_0\|_2^2 N_2 \lambda_{\max}(\Sigma)) r \|Z_0\|_2^2 \\
&\left[\because \left\|(\widehat{V}^\top\widehat{V})^{-1}\widehat{V}^\top\right\|_{\text{op}}^2 = \left\|\widehat{V}(\widehat{V}^\top\widehat{V})^{-2}\widehat{V}^\top\right\|_{\text{op}} = \left\|(\widehat{V}^\top\widehat{V})^{-1}\right\|_{\text{op}} \text{ and lemma 10}\right] \\
&= O\left(\frac{1}{N_2\eta^2\|Y_0\|_2^2\,\lambda_{\min}(\Sigma)}\right) O(\|Y_0\|_2^2 N_2 \lambda_{\max}(\Sigma)) r \|Z_0\|_2^2 \\
&= O\left(C_2 \frac{r}{\eta^2}\right) \quad \text{for } C_2 = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\|Z_0\|_2^2
\end{aligned}
$$

Therefore, we have bound $\left\|\widehat{A^{3^\top} Z_0}\right\|_2^2$ by $\frac{C_1}{\eta^2}\frac{r}{N_2} + C_2\frac{r}{\eta^2}$. Substituting the upper bounds on $\left\|\widehat{A^{3^\top} Z_0}\right\|_2^2$ and $\left\|\widehat{A^{3^\top}\hat{Z}_0} - A^{3^\top} Z_0\right\|_2^2$ in equation 15 establishes the desired bound. □

Lemma 8. *Each covariate vector X_i is mean-zero, satisfies $\mathbb{E}[X_i X_i^\top] = \Sigma$ and Σ-sub-gaussian. Additionally, suppose that$N_2 \geq O\left(r\left(\frac{\|Y_0\|_2^2}{\eta^2}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\right)^2 \log(2/\delta_1)\right)$, and $\left|Y_0^\top \hat{A^2}_i\right| \geq \eta\|Y_0\|_2$ for all $i \in [r]$. Then with probability at least $1-\delta_1$ we have*

$$
\left\|\left(\widehat{V}^\top\widehat{V}\right)^{-1}\widehat{V}^\top\boldsymbol{\varepsilon}\right\|_2^2 \leq \tilde{O}\left(\frac{r}{N_2\eta^2\|Y_0\|_2^2\,\lambda_{\min}(\Sigma)}\right)
$$

Proof. The bias term is given as $\left\|(\widehat{V}^\top\widehat{V})^{-1}\widehat{V}^\top\boldsymbol{\varepsilon}\right\|_2^2 = \boldsymbol{\varepsilon}^\top \underbrace{\widehat{V}(\widehat{V}^\top\widehat{V})^{-2}\widehat{V}^\top}_{:=M}\boldsymbol{\varepsilon}$. By the Hanson-Wright inequality ([32], lemma 6.2.1) we have

$$
P\left(\left|\boldsymbol{\varepsilon}^\top M\boldsymbol{\varepsilon} - E[\boldsymbol{\varepsilon}^\top M\boldsymbol{\varepsilon}]\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{\|M\|_F^2}, \frac{t}{\|M\|_{\text{op}}}\right)\right).
$$

Therefore, we have $\boldsymbol{\varepsilon}^\top M\boldsymbol{\varepsilon} \leq E[\boldsymbol{\varepsilon}^\top M\boldsymbol{\varepsilon}] + O\left(\|M\|_F\sqrt{\log(2/\delta_1)}\right) + O\left(\|M\|_{\text{op}}\log(2/\delta_1)\right)$ with probability at least $1-\delta_1/2$. From the singular value decomposition of $\widehat{V}$, it is easy to see that $\|M\|_{\text{op}} = \left\|\widehat{V}(\widehat{V}^\top\widehat{V})^{-2}\widehat{V}^\top\right\|_{\text{op}} = \left\|(\widehat{V}^\top\widehat{V})^{-1}\right\|_{\text{op}}$. Moreover, lemma 13 proves that with probability at least $1-\delta_1/2$, the matrix M is invertible and $\|M\|_{\text{op}} \leq O\left(\frac{1}{N_2\eta^2\|Y_0\|_2^2\lambda_{\min}(\Sigma)}\right)$ as long as $N_2 \geq O\left(r\left(\frac{\|Y_0\|_2^2}{\eta^2}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\right)^2\log(2/\delta_1)\right)$.

Since $\operatorname{rank}(M) \leq \operatorname{rank}(\widehat{\Gamma}) \leq r$, we have $\|M\|_F \leq \sqrt{r}\|M\|_{\text{op}} \leq O\left(\frac{\sqrt{r}}{N_2\eta^2\|Y_0\|_2^2\lambda_{\min}(\Sigma)}\right)$. By a similar argument we get $\mathbb{E}[\boldsymbol{\varepsilon}^\top M\boldsymbol{\varepsilon}] = \operatorname{Tr}(M) \leq r\|M\|_{\text{op}} \leq O\left(\frac{r}{N_2\eta^2\|Y_0\|_2^2\lambda_{\min}(\Sigma)}\right)$. This gives us $\boldsymbol{\varepsilon}^\top M\boldsymbol{\varepsilon} \leq \tilde{O}\left(\frac{r}{N_2\eta^2\|Y_0\|_2^2\lambda_{\min}(\Sigma)}\right)$. □

Lemma 9. *Each covariate vector X_i is mean-zero, satisfies $\mathbb{E}[X_iX_i^\top] = \Sigma$ and Σ-sub-gaussian. Additionally, assume that $\max\left\{\left\|\hat{A}^1 - A^1\right\|_F, \left\|\hat{A}^2 - A^2\right\|_F\right\} \leq \delta$, and $\sin\theta(A^3, \hat{A}^3) \leq \delta\sqrt{r}$. If $N_2 \geq O\left(r\left(\frac{\|Y_0\|_2^2}{\eta^2}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\right)^2\log(2/\delta_1)\right)$, and $\left|Y_0^\top \hat{A}^2{}_i\right| \geq \eta\|Y_0\|_2$ for all $i \in [r]$, then with probability at least $1-\delta_1$ we have*

$$\left\|\left(\widehat{V}^\top\widehat{V}\right)^{-1}\widehat{V}^\top V {A^3}^\top Z_0 - {A^3}^\top Z_0\right\|_2^2 = O\left(\frac{\lambda_{\max}(\Sigma)}{\eta^2\lambda_{\min}(\Sigma)}\|Z_0\|_2^2 r\delta^2\right)$$

Proof. Our proof resembles the proof of Lemma 19 of [28], but there are some important differences. First note that, by lemma 11 we can write $V = \widehat{V} + E_V$ for a matrix E_V with $\|E_V\|_{\text{op}} \leq O(\|Y_0\|_2\sqrt{N_2\lambda_{\max}(\Sigma)}\delta)$. This gives us the following bound on the variance.

$$\begin{aligned}
&\left\|\left(\widehat{V}^\top\widehat{V}\right)^{-1}\widehat{V}^\top V {A^3}^\top Z_0 - {A^3}^\top Z_0\right\|_2^2 \\
&= \left\|\left(\widehat{V}^\top\widehat{V}\right)^{-1}\widehat{V}^\top\widehat{V} {A^3}^\top Z_0 - {A^3}^\top Z_0 + \left(\widehat{V}^\top\widehat{V}\right)^{-1}\widehat{V}^\top E_V {A^3}^\top Z_0\right\|_2^2 \\
&= \left\|\left(\widehat{V}^\top\widehat{V}\right)^{-1}\widehat{V}^\top E_V {A^3}^\top Z_0\right\|_2^2 \\
&\leq \left\|\left(\widehat{V}^\top\widehat{V}\right)^{-1}\widehat{V}^\top\right\|_{\text{op}}^2 \|E_V\|_{\text{op}}^2 \left\|{A^3}^\top Z_0\right\|_2^2 \\
&\leq \left\|\widehat{V}\left(\widehat{V}^\top\widehat{V}\right)^{-2}\widehat{V}^\top\right\|_{\text{op}} O(\|Y_0\|_2^2 N_2\lambda_{\max}(\Sigma)\delta^2) r\|Z_0\|_2^2 \qquad (16)
\end{aligned}$$

The last line uses $\left\|{A^3}^\top Z_0\right\|_2^2 = \sum_{i=1}^r ({A^3}_i^\top Z_0)^2 \leq \sum_{i=1}^r \left\|A^3\right\|_2^2\|Z_0\|_2^2 = r\|Z_0\|_2^2$. Now $\left\|\widehat{V}\left(\widehat{V}^\top\widehat{V}\right)^{-2}\widehat{V}^\top\right\|_{\text{op}} = \left\|\left(\widehat{V}^\top\widehat{V}\right)^{-1}\right\|_{\text{op}}$ and lemma 13 proves that with probability at least $1-\delta_1/2$, the matrix $\widehat{V}^\top\widehat{V}$ is invertible and $\left\|\left(\widehat{V}^\top\widehat{V}\right)^{-1}\right\|_{\text{op}} \leq O\left(\frac{1}{N_2\eta^2\|Y_0\|_2^2\lambda_{\min}(\Sigma)}\right)$ as long as $N_2 \geq O\left(r\left(\frac{\|Y_0\|_2^2}{\eta^2}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\right)^2\log(2/\delta_1)\right)$. Substituting the upper bound on the operator norm of $\left(\widehat{V}^\top\widehat{V}\right)^{-1}$ gives the desired bound.

□

Lemma 10. *If $N_2 \geq O(r\log(1/\delta_1))$ then we have*

$$\left\|\widehat{V}\right\|_{op} \leq O\left(\|Y_0\|_2\sqrt{N_2\lambda_{\max}(\Sigma)}\right)$$

with probability at least $1-\delta_1$.

Proof.

$$\begin{aligned}\left\|\widehat{V}\right\|_{\text{op}}^2 &= \lambda_{\max}\left((Y_0^\top \hat{A}^2 \odot \mathcal{X}\hat{A}^1)^\top (Y_0^\top \hat{A}^2 \odot \mathcal{X}\hat{A}^1)\right) \\ &= N_2\lambda_{\max}\left(U^\top(\frac{1}{N_2}\mathcal{X}^\top\mathcal{X})U\right)\end{aligned}$$

where in the last line we write $U \in \mathbb{R}^{d_1\times r}$ to denote the matrix with columns $U_i = (Y_0^\top \hat{A}_i^2)\hat{A}_i^1$. The matrix U has orthogonal columns and $\|U\|_{\text{op}} \le \|Y_0\|_2$. Therefore, we can apply lemma 12 to obtain that as long as $N_2 \ge r\log(1/\delta_1)$, we have $\lambda_{\max}\left(U^\top(\frac{1}{N_2}\mathcal{X}^\top\mathcal{X})U\right)$ is bounded by $O\left(\left\|U^\top\Sigma U\right\|_{\text{op}} + \lambda_{\max}(\Sigma)\|Y_0\|_2^2\right)$ with probability at least $1-\delta_1$. Moreover, $\left\|U^\top\Sigma U\right\|_{\text{op}}$ is bounded by $\lambda_{\max}(\Sigma)\|Y_0\|_2^2$. This establishes a bound of $O(N_2\lambda_{\max}(\Sigma)\|Y_0\|_2^2)$ on $\left\|\widehat{V}\right\|_{\text{op}}^2$. □

Lemma 11. *Suppose,* $\max\left\{\left\|\hat{A}^1 - A^1\right\|_F, \left\|\hat{A}^2 - A^2\right\|_F\right\} \le \delta$. *If* $N_2 \ge O(r\log(1/\delta_1))$ *then we have*

$$\left\|\widehat{V} - V\right\|_{op} \le O\left(\|Y_0\|_2\sqrt{N_2\lambda_{\max}(\Sigma)}\delta\right)$$

with probability at least $1-\delta_1$.

Proof. We will write $\hat{A}^1 = A^1 + E^1$, and $\hat{A}^2 = A^2 + E^2$. Note that we have $\left\|E^1\right\|_F \le \delta, \left\|E^2\right\|_F \le \delta$.

$$\begin{aligned}\left\|\widehat{V} - V\right\|_{\text{op}} &= \left\|(Y_0^\top \hat{A}^2 \odot \mathcal{X}\hat{A}^1) - (Y_0^\top A^2 \odot \mathcal{X}A^1)\right\|_{\text{op}} \\ &= \left\|(Y_0^\top (A^2+E^2) \odot \mathcal{X}(A^1+E^1)) - (Y_0^\top A^2 \odot \mathcal{X}A^1)\right\|_{\text{op}} \\ &\le \left\|(Y_0^\top E^2) \odot (\mathcal{X}\hat{A}^1)\right\|_{\text{op}} + \left\|(Y_0^\top A^2) \odot (\mathcal{X}E^1)\right\|_{\text{op}} + \left\|(Y_0^\top E^2) \odot (\mathcal{X}E^1)\right\|_{\text{op}}\end{aligned}$$

Consider the first term.

$$\begin{aligned}&\left\|(Y_0^\top E^2) \odot (\mathcal{X}\hat{A}^1)\right\|_{\text{op}}^2 \\ &= \lambda_{\max}\left(\left((Y_0^\top E^2) \odot (\mathcal{X}\hat{A}^1)\right)^\top (Y_0^\top E^2) \odot (\mathcal{X}\hat{A}^1)\right) \\ &= N_2\lambda_{\max}\left(U^\top(\frac{1}{N_2}\mathcal{X}^\top\mathcal{X})U\right)\end{aligned}$$

where in the last line we write $U \in \mathbb{R}^{d_1\times r}$ to denote the matrix with columns $U_i = (Y_0^\top E_i^2)\hat{A^1}_i$. Note that U has orthogonal columns as the columns of $\hat{A}^1$ are orthogonal. Moreover, $\|U\|_{\text{op}} \le \|U\|_F \le \|Y_0\|_2\left\|E^2\right\|_F \le \|Y_0\|_2\,\delta$. Therefore, we can apply lemma 12 to get that as long as $N_2 \ge r\log(1/\delta_1)$,

we have $\lambda_{\max}(U^\top(\frac{1}{N_2}\mathcal{X}^\top\mathcal{X})U)$ is bounded by $O\left(\left\|U^\top\Sigma U\right\|_{\text{op}} + \lambda_{\max}(\Sigma)\delta^2\|Y_0\|_2^2\right)$ with probability at least $1-\delta_1$. Moreover, $\left\|U^\top\Sigma U\right\|_{\text{op}} = \lambda_{\max}(\Sigma)\|U\|_F^2 \le \lambda_{\max}(\Sigma)\delta^2\|Y_0\|_2^2$. This establishes a bound of $O(\sqrt{\lambda_{\max}(\Sigma)N_2}\delta\|Y_0\|_2)$ on $\left\|(Y_0^\top E^2)\odot(\mathcal{X}\hat{A}^1)\right\|_{\text{op}}$. By a similar argument, one can establish a bound of $O(\sqrt{\lambda_{\max}(\Sigma)N_2}\delta\|Y_0\|_2)$ on the second term $\left\|(Y_0^\top A^2)\odot(\mathcal{X}E^1)\right\|_{\text{op}}$, and a bound of $O(\sqrt{\lambda_{\max}(\Sigma)N_2}\delta\|Y_0\|_2)$ on the third term $\left\|(Y_0^\top A^2)\odot(\mathcal{X}A^1)E^0\right\|_{\text{op}}$. □

Lemma 12. *Suppose each covariate x_i is mean-zero, satisfies $\mathbb{E}[xx^\top]=\Sigma$ and Σ-subgaussian. Moreover, A and B are rank r matrices with orthogonal columns. Then the following holds*

$$\left\|A^\top\frac{\mathcal{X}^\top\mathcal{X}}{n}B - A^\top\Sigma B\right\|_{op} \le O\left(\lambda_{\max}(\Sigma)\max\{\|A\|_{op}^2,\|B\|_{op}^2\}\left(\sqrt{\frac{r}{n}}+\frac{r}{n}+\sqrt{\frac{\log(1/\delta)}{n}}+\frac{\log(1/\delta)}{n}\right)\right)$$

with probability at least $1-\delta$.

Proof. The proof is very similar to the proof of lemma 20 from [28]. □

Lemma 13. *Suppose, $N_2 \ge O\left(r\left(\frac{\|Y_0\|_2^2}{\eta^2}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\right)^2\log(1/\delta_2)\right)$, and $\left|Y_0^\top\hat{A}^2{}_i\right| \ge \eta\|Y_0\|_2$ for all $i\in[r]$. Then the matrix $(\widehat{V}^\top\widehat{V})$ is invertible and*

$$\left\|(\widehat{V}^\top\widehat{V})^{-1}\right\|_{op} \le O\left(\frac{1}{N_2\eta^2\|Y_0\|_2^2\,\lambda_{\min}(\Sigma)}\right)$$

with probability at least $1-\delta_2$.

Proof. From the definition of the matrix $\widehat{V}$, we have

$$\widehat{V}^\top\widehat{V} = (Y_0^\top\hat{A}^2\odot\mathcal{X}\hat{A}^1)^\top(Y_0^\top\hat{A}^2\odot\mathcal{X}\hat{A}^1).$$

If we define a matrix $U\in\mathbb{R}^{d_1\times r}$ with columns $U_i = (Y_0^\top\hat{A}_i^2)\hat{A}_i^1$, then it can be verified that $\frac{1}{N_2}\widehat{V}^\top\widehat{V} = U^\top\left(\frac{1}{N_2}\mathcal{X}^\top\mathcal{X}\right)U$. This gives us $\mathbb{E}[\frac{1}{N_2}\widehat{V}^\top\widehat{V}] = U^\top\Sigma U$. Therefore, we can write $\frac{1}{N_2}\widehat{V}^\top\widehat{V} = \mathcal{E}+U^\top\Sigma U$, for a matrix $\mathcal{E}$ with $\mathbb{E}[\mathcal{E}]=0$. Since matrix U has orthogonal columns and $\|U\|_{\text{op}} \le \|Y_0\|_2$ we can apply lemma 12 to conclude that as long as $N_2 \ge O(r\log(1/\delta_1))$ we have $\|\mathcal{E}\|_{\text{op}} \le O\left(\lambda_{\max}(\Sigma)\|Y_0\|_2^2\sqrt{r/N_2}\right)$. On the other hand,

$$\lambda_{\min}(U^\top\Sigma U) = \min_{x\in\mathbb{R}^{d_3},x\neq0}\frac{x^\top U^\top\Sigma U x}{x^\top x} \ge \eta^2\|Y_0\|_2^2\min_{w\in\mathbb{R}^{d_1},w\neq0}\frac{w^\top\Sigma w}{w^\top w} = \eta^2\|Y_0\|_2^2\,\lambda_{\min}(\Sigma).$$

The first inequality follows from substituting $w=Ux$ and observing that $w^\top w = x^\top U^\top U x \ge$

$\min_i \left| Y_0^\top \hat{A}^2{}_i \right|^2 x^\top x \geq \|Y_0\|_2^2 \eta^2 x^\top x$. Therefore,

$$\begin{aligned}\lambda_{\min}\left(\frac{1}{N_2}\widehat{V}^\top\widehat{V}\right) &\geq \lambda_{\min}(U^\top\Sigma U) - \lambda_{\max}(\mathcal{E}) \\ &\geq O\left(\|Y_0\|_2^2\,\eta^2\lambda_{\min}(\Sigma)\right) - \|\mathcal{E}\|_{\text{op}} \\ &\geq O\left(\|Y_0\|_2^2\,(\eta^2\lambda_{\min}(\Sigma) - \lambda_{\max}(\Sigma)\sqrt{r/N_2})\right)\end{aligned}$$

Therefore, as long as $N_2 \geq r\left(\frac{1}{\eta^2}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\right)^2$, $\frac{1}{N_2}\widehat{V}^\top\widehat{V}$ is invertible and so is $\widehat{V}^\top\widehat{V}$.

Now, $(\widehat{V}^\top\widehat{V})^{-1} = \frac{1}{N_2}(\frac{1}{N_2}\widehat{V}^\top\widehat{V})^{-1} = \frac{1}{N_2}(\mathcal{E} + U^\top\Sigma U)^{-1}$. Moreover,

$$\left\|(U^\top\Sigma U)^{-1}\mathcal{E}\right\|_{\text{op}} \leq \left\|(U^\top\Sigma U)^{-1}\right\|_{\text{op}}\|\mathcal{E}\|_{\text{op}} = \frac{\|\mathcal{E}\|_{\text{op}}}{\lambda_{\min}(U^\top\Sigma U)} = O\left(\frac{1}{\eta^2}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\sqrt{\frac{r}{N_2}}\right)$$

Therefore, as long as $N_2 \geq O\left(\frac{r}{\eta^2}\frac{\lambda_{\max}^2(\Sigma)}{\lambda_{\min}^2(\Sigma)}\right)$ we have, $\left\|(U^\top\Sigma U)^{-1}\mathcal{E}\right\|_{\text{op}} \leq 1/4$. Therefore, we can apply lemma 14 to conclude that $(\frac{1}{N_2}\widehat{V}^\top\widehat{V})^{-1} = (U^\top\Sigma U)^{-1} + F$ where $\|F\|_{\text{op}} \leq \frac{1}{3}\left\|(U^\top\Sigma U)^{-1}\right\|_{\text{op}}$. Therefore, $\left\|(\widehat{V}^\top\widehat{V})^{-1}\right\|_{\text{op}} \leq \frac{4}{3N_2}\left\|(U^\top\Sigma U)^{-1}\right\|_{\text{op}} = \frac{4}{3N_2}\frac{1}{\lambda_{\min}(U^\top\Sigma U)} \leq \frac{4}{3N_2\eta^2\|Y_0\|_2^2\lambda_{\min}(\Sigma)}$. □

Lemma 14 (Restated lemma 23 from [28]). *Let A be a positive-definite matrix and E is another matrix satisfying $\left\|EA^{-1}\right\| \leq \frac{1}{4}$. Then $(A+E)^{-1} = A^{-1} + F$ where $\|F\|_{op} \leq \frac{4}{3}\left\|A^{-1}\right\|_{op}\left\|EA^{-1}\right\|_{op}$.*

E Proof of Theorem 3

We first recall the method of moments estimator from [28]. If the response $R_i = X_i^\top B\alpha_{t(i)}$ and each $X_i \sim_{\text{iid}} \text{N}(0, I_{d_1})$ then we have $\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N R_i^2 X_i X_i^\top\right] = 2\bar{\Gamma} + (1 + \text{Tr}(\bar{\Gamma}))I_{d_1}$ where $\bar{\Gamma} = \frac{1}{N}\sum_{i=1}^N B\alpha_{t(i)}\alpha_{t(i)}^\top B^\top$. If we write $\bar{\Lambda} = \frac{1}{N}\sum_{i=1}^N \alpha_{t(i)}\alpha_{t(i)}^\top$ to be the empirical task matrix we have $\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N R_i^2 X_i X_i^\top\right] = B(2\bar{\Lambda})B^\top + B_\perp(1 + \text{Tr}(\bar{\Gamma}))\,\text{I}_r\,B_\perp^\top$. So that we can recover B from the top r singular values of the statistic $\frac{1}{N}\sum_{i=1}^N R_i^2 X_i X_i^\top$. Moreover, theorem 3 of [28] proves that such an estimate $\hat{B}$ satisfies $\sin\theta(\hat{B}, B) \leq \sqrt{\frac{\kappa}{\nu}\frac{d_1 r}{N}}$, where $\nu = \sigma_r(\bar{\Lambda})$ and $\kappa = \text{Tr}(\bar{\Lambda})/(r\nu)$.

Recovering A^1 . Let us consider the estimation of the first factor A^1. The response of the i-th individual is given as

$$R_i = X_i^\top A_{(1)}(Z_{t(i)} \otimes Y_{t(i)}) + \varepsilon_i = X_i^\top A^1 \underbrace{(A^3 \odot A^2)^\top(Z_{t(i)} \otimes Y_{t(i)})}_{:=P^1_{t(i)}} + \varepsilon_i$$

Therefore, we recover A^1 from the top r singular values of $\frac{1}{N}\sum_{i=1}^N R_i^2 X_i X_i^\top$. In order to obtain a bound on $\sin\theta(A^1, \hat{A}^1)$ we need to bound eigenvalue and trace of the empirical task matrix

$\bar{\Lambda} = \frac{1}{N}\sum_{i=1}^{N} P^1_{t(i)} {P^1_{t(i)}}^\top$. Since each $t(i)$ is a uniform random draw from $\{1, \ldots, T\}$, we have $\Lambda = \mathbb{E}[\bar{\Lambda}] = \frac{1}{T}\sum_{t=1}^{T} P^1_t {P^1_t}^\top = (A^3 \odot A^2)^\top \frac{1}{T}\sum_{t=1}^{T}(Z_t \otimes Y_t)(Z_t \otimes Y_t)^\top (A^3 \odot A^2) = \frac{1}{T}(A^3 \odot A^2)^\top (\mathcal{Z}^\top \odot \mathcal{Y}^\top)(\mathcal{Z}^\top \odot \mathcal{Y}^\top)^\top (A^3 \odot A^2)$. We first bound the eigenvalues of Λ and then use matrix concentration inequality to bound the eigenvalues of the empirical task matrix $\bar{\Lambda}$.

$$\lambda_{\min}(\Lambda) = \frac{1}{T}\lambda_{\min}((\mathcal{Z}^\top \odot \mathcal{Y}^\top)(\mathcal{Z}^\top \odot \mathcal{Y}^\top)^\top) = \frac{1}{T}\lambda_{\min}\left(\mathcal{Z}^\top\mathcal{Z} \otimes \mathcal{Y}^\top\mathcal{Y}\right) = \frac{1}{T}\lambda_{\min}(\mathcal{Z}^\top\mathcal{Z})\lambda_{\min}(\mathcal{Y}^\top\mathcal{Y})$$

The first equality follows from the observation that $A^2 \odot A^1$ has orthonormal columns, and the last equality follows because the eigenvalues of Kronecker product of two matrices are given as the Kronecker product of eigenvalues of the two matrices. Since each $Z_t \sim_{\text{iid}} \mathrm{N}(0, \mathrm{I}_{d_3})$, the minimum singular value of $\mathcal{Z}$ is bounded by $\sqrt{T} - \sqrt{d_3}$ with probability at least $1 - 2\exp(-O(d_3))$ (see e.g. theorem 4.6.1 of [32]). This implies that $\lambda_{\min}(\mathcal{Z}^\top\mathcal{Z}) = \sigma_{\min}(\mathcal{Z})^2 \geq T/4$ with probability at least $1 - 2\exp(-O(d_3))$ as long as $T \geq 4d_3$. Similarly, it can be shown that $\lambda_{\min}(\mathcal{Y}^\top\mathcal{Y}) \geq T/4$ with probability at least $1 - 2\exp(-O(d_2))$ as long as $T \geq 4d_2$. This establishes a high probability lower bound of $T/16$ on $\lambda_{\min}(\Lambda)$.

Moreover, for any $i \in [N]$,

$$\begin{aligned}\lambda_{\max}(P^1_{t(i)}{P^1_{t(i)}}^\top) &= \lambda_{\max}((Z_{t(i)} \otimes Y_{t(i)})(Z_{t(i)} \otimes Y_{t(i)})^\top) = \lambda_{\max}(Z_{t(i)}Z_{t(i)}^\top \otimes Y_{t(i)}Y_{t(i)}^\top) \\ &\leq \lambda_{\max}(Z_{t(i)}Z_{t(i)}^\top)\lambda_{\max}(Y_{t(i)}Y_{t(i)}^\top) \leq \left\|Z_{t(i)}\right\|_2^2 \left\|Y_{t(i)}\right\|_2^2\end{aligned}$$

When $Z_{t(i)}$ is drawn from standard Normal distribution $\left\|Z_{t(i)}\right\|_2 \leq 2\sqrt{d_3}$ with probability at least $1 - \exp(-O(d_3))$. By a union bound over all T tasks we have for all $t \in [T]$, $\|Z_t\|_2 \leq 2\sqrt{d_3}$ with probability at least $1 - T\exp(-O(d_3))$. A similar argument shows that for all $t \in [T]$, $\|Y_t\|_2 \leq 2\sqrt{d_2}$ with probability at least $1 - T\exp(-O(d_2))$. Therefore, we are guaranteed that $\lambda_{\max}(P^1_{t(i)}{P^1_{t(i)}}^\top) \leq 16 d_2 d_3$ for all i, with probability at least $1 - T\exp(-O(\min\{d_2, d_3\}))$. Now we can apply matrix concentration inequality (lemma 15) to derive the following result.

$$P\left(\lambda_{\min}(\bar{\Lambda}) \leq \frac{T}{32}\right) \leq d_1 \exp\left\{-O\left(\frac{TN}{d_2 d_3}\right)\right\}.$$

Similarly, we can establish an upper bound on the maximum eigenvalue of $\bar{\Lambda}$.

$$P\left(\lambda_{\max}(\bar{\Lambda}) \geq 32T\right) \leq d_1 \exp\left\{-O\left(\frac{TN}{d_2 d_3}\right)\right\}.$$

Therefore, $\mathrm{Tr}(\bar{\Lambda}) \leq r\lambda_{\max}(\bar{\Lambda}) \leq 32rT$ with probability at least $1 - d_1 \exp\left\{-O\left(TN/d_2 d_3\right)\right\}$. Moreover, $\kappa = \mathrm{Tr}(\bar{\Lambda})/(r\lambda_{\min}(\bar{\Lambda}) = O(1)$. This implies the following bound on the distance between $\hat{A}^1$ and A^1.

$$\sin\theta\left(\hat{A}^1, A^1\right) \leq O\left(\sqrt{\frac{\kappa d_1 r}{\nu N}}\right) = O\left(\sqrt{\frac{d_1 r}{TN}}\right).$$

Recovering A^2 . We can provide a bound on the error in estimating A^2 through a similar approach. The response of individual i can be written as

$$R_i = Y_{t(i)}^\top A_{(2)}(Z_{t(i)} \otimes X_i) + \varepsilon_i = Y_{t(i)}^\top A^2 \underbrace{W(A^3 \odot A^1)^\top (Z_{t(i)} \otimes X_i)}_{:=P^2_{t(i)}} + \varepsilon_i$$

Therefore, we can recover A^2 from the top r singular values of $\frac{1}{N}\sum_{i=1}^N R_i^2 Y_{t(i)} Y_{t(i)}^\top$. Now the empirical task matrix is $\bar{\Lambda} = \frac{1}{N}\sum_{i=1}^N P^2_{t(i)} {P^2_{t(i)}}^\top$. We now bound the eigenvalue and trace of the empirical task matrix. Since each $t(i)$ is a uniform random draw from $\{1, \ldots, T\}$, we have $\Lambda = \mathbb{E}[\bar{\Lambda}] = \frac{1}{T}\sum_{t=1}^T P_t^2 {P_t^2}^\top = (A^3 \odot A^1)^\top \frac{1}{T}\sum_{t=1}^T \mathbb{E}[(Z_t \otimes X)(Z_t \otimes X)^\top](A^3 \odot A^1) = (A^3 \odot A^1)^\top \frac{1}{T}\sum_{t=1}^T (Z_t \otimes \mathrm{I}_{d_1})(Z_t \otimes \mathrm{I}_{d_1})^\top (A^3 \odot A^1) = \frac{1}{T}(A^3 \odot A^1)^\top (\mathcal{Z}^\top \otimes \mathrm{I}_{d_1})(\mathcal{Z}^\top \otimes \mathrm{I}_{d_1})^\top (A^3 \odot A^1)$

$$\lambda_{\min}(\Lambda) = \frac{1}{T}\sigma_{\min}((\mathcal{Z}^\top \otimes \mathrm{I}_{d_1})(\mathcal{Z}^\top \otimes \mathrm{I}_{d_1})^\top) = \frac{1}{T}\sigma_{\min}(\mathcal{Z}^\top \mathcal{Z} \otimes \mathrm{I}_{d_1}) \geq \frac{1}{T}\sigma_{\min}(\mathcal{Z}^\top \mathcal{Z})$$

Since each $Z_t \sim_{\text{iid}} \mathrm{N}(0, \mathrm{I}_{d_3})$, the minimum singular value of $\mathcal{Z}$ is bounded by $\sqrt{T} - \sqrt{d_3}$ with probability at least $1 - 2\exp(-O(d_3))$ (see e.g. theorem 4.6.1 of [32]). This implies that $\lambda_{\min}(\mathcal{Z}^\top \mathcal{Z}) = \sigma_{\min}(\mathcal{Z})^2 \geq 1/4$ with probability at least $1 - 2\exp(-O(d_3))$ as long as $T \geq 4d_3$. Moreover, for any $i \in [N]$,

$$\begin{aligned}\lambda_{\max}(P^2_{t(i)} {P^2_{t(i)}}^\top) &= \lambda_{\max}((Z_{t(i)} \otimes X_i)(Z_{t(i)} \otimes X_i)^\top) = \lambda_{\max}(Z_{t(i)} Z_{t(i)}^\top \otimes X_i X_i^\top) \\ &\leq \lambda_{\max}(Z_{t(i)} Z_{t(i)}^\top)\lambda_{\max}(X_i X_i^\top) \leq \left\| Z_{t(i)} \right\|_2^2 \|X_i\|_2^2\end{aligned}$$

When $Z_{t(i)}$ is drawn from standard Normal distribution $\left\| Z_{t(i)} \right\|_2 \leq 2\sqrt{d_3}$ with probability at least $1 - \exp(-O(d_3))$. By a union bound over all T tasks we have for all $t \in [T]$, $\|Z_t\|_2 \leq 2\sqrt{d_3}$ with probability at least $1 - T\exp(-O(d_3))$. A similar argument shows that for all $i \in [N]$, $\|X_i\|_2 \leq 2\sqrt{d_1}$ with probability at least $1 - T\exp(-O(d_1))$. Therefore, we are guaranteed that $\lambda_{\max}(P^1_{t(i)} {P^1_{t(i)}}^\top) \leq 16 d_1 d_3$ for all i, with probability at least $1 - N\exp(-O(\min\{d_1, d_3\}))$. Now we can apply matrix concentration inequality (lemma 15) to derive the following result.

$$P\left(\lambda_{\min}(\bar{\Lambda}) \leq \frac{1}{8}\right) \leq d_2 \exp\left\{-O\left(\frac{N}{d_1 d_3}\right)\right\}.$$

Similarly, we can establish an upper bound on the maximum eigenvalue of $\bar{\Lambda}$.

$$P\left(\lambda_{\max}(\bar{\Lambda}) \geq 8\right) \leq d_2 \exp\left\{-O\left(\frac{N}{d_1 d_3}\right)\right\}.$$

Therefore, $\mathrm{Tr}(\bar{\Lambda}) \leq 8r$ with probability at least $1 - d_1 \exp\left\{-O\left(N/d_1 d_3\right)\right\}$. Moreover, $\kappa = \mathrm{Tr}(\bar{\Lambda})/(r\lambda_{\min}(\bar{\Lambda})) = O(1)$. This implies the following bound on the distance between $\hat{A}^2$ and A^2.

$$\sin\theta\left(\hat{A}^2, A^2\right) \leq O\left(\sqrt{\frac{\kappa d_2 r}{\nu N}}\right) = O\left(\sqrt{\frac{d_2 r}{N}}\right).$$

Lemma 15 (Restated theorem 5.1.1 from [30]). *Consider a sequence of $\{X_k\}_{k=1}^N$ independent, random, Hermitian matrices of dimension $d \times d$. Assume that the eigenvalues of each X_k is bounded between $[0, L]$. Let $Y = 1/N \sum_k X_k$, $\mu_{\min} = \lambda_{\min}(\mathbb{E}[Y])$, and $\mu_{\max} = \lambda_{\max}(\mathbb{E}[Y])$. Then we have*

$$P\left(\lambda_{\min}(Y) \leq (1-\varepsilon)\mu_{\min}\right) \leq d\left[\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right]^{\mu_{\min}N/L} \quad \forall \varepsilon \in [0,1)$$

$$P\left(\lambda_{\max}(Y) \geq (1+\varepsilon)\mu_{\max}\right) \leq d\left[\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right]^{\mu_{\max}N/L} \quad \forall \varepsilon \in [0,\infty)$$

F Meta-Test for Method-of-Moments Based Estimation

In the meta-test phase, (X_i, R_i) for $i = 1, \ldots, N_2$ are observed for a task with specific feature Y_0. The model can be expressed as

$$R_i = (Y_0 \otimes X_i)^\top (A^2 \odot A^1) A^{3\top} Z_0 + \varepsilon_i, \quad \text{for } i = 1, \ldots, N_2.$$

If we denote the latent task factor $A^{3\top} Z_0$ as a vector $\alpha \in \mathbb{R}^r$, α can be estimated from the least square problem with A^1 and A^2 substituted by their estimators from the meta-training phase

$$\hat{\alpha} = \text{argmin}_\alpha \left\| \mathbf{R} - (Y_0 \otimes \mathcal{X}^\top)^\top (\hat{A}^2 \odot \hat{A}^1)\alpha \right\|_2.$$

For notation simplicity, throughout this section we denote $(Y_0 \otimes \mathcal{X}^\top)^\top (\hat{A}^2 \odot \hat{A}^1)$ as $\widehat{M}$, and $(Y_0 \otimes \mathcal{X}^\top)^\top (A^2 \odot A^1)$ as M. In addition, we let $\widehat{M}_0$ denote $Y_0^\top \hat{A}^2 \odot X_0^\top \hat{A}^1$, M_0 denote $Y_0^\top A^2 \odot X_0^\top A^1$. Then the least square estimation becomes

$$\hat{\alpha} = [\widehat{M}^\top \widehat{M}]^{-1} \widehat{M}^\top \mathbf{R}.$$

After obtaining the estimation of the task with observable and latent features Y_0 and α, a test sample is collected on this task with input X_0. Then the estimation error can be expressed using the notation as

$$\begin{aligned}
&\mathbb{E}_{X_0}\left[\left(A(X_0, Y_0, \alpha) - \hat{A}(X_0, Y_0, \hat{\alpha})\right)^2\right] \\
&= \mathbb{E}_{X_0}\left[\left((Y_0 \otimes X_0)^\top (A^2 \odot A^1)\alpha - (Y_0 \otimes X_0)^\top (\hat{A}^2 \odot \hat{A}^1)\hat{\alpha}\right)^2\right] \\
&= \mathbb{E}_{X_0}\left[\left\| \widehat{M}_0 \hat{\alpha} - M_0 \alpha \right\|_2^2\right].
\end{aligned}$$

Formally, we have

Theorem 5. *Suppose each covariate x_i is mean-zero, satisfies $\mathbb{E}[xx^\top] = \Sigma$ and Σ-subgaussian, and ε_i's are i.i.d. mean -zero, sub-gaussian variables with variance parameter 1, independent of x_i. If $\left|Y_0^\top \hat{A}^2{}_i\right| \geq \eta \|Y_0\|_2$ for all $i \in [r]$, and $N_2 \geq O\left((r + \log 2/\delta_2)\left(\frac{1}{\eta^2}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\right)^2\right)$, then with probability at least $1 - \delta_2$, we have*

$$\mathbb{E}_{X_0}\left[\left(A(X_0, Y_0, \alpha) - \hat{A}(X_0, Y_0, \hat{\alpha})\right)^2\right] = O\left(C_1 r \delta^2 + C_2 \frac{r}{N_2}\right),$$

for $C_1 = \mathbb{E}\left[\|X_0\|_2^2\right]\|Y_0\|_2^2\left(\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\right)^2\frac{1}{\eta^4}\|\alpha\|_2^2$ *and* $C_2 = \mathbb{E}\left[\|X_0\|_2^2\right]\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}^2(\Sigma)}\frac{1}{\eta^4}$.

Proof. The error can be written as

$$\begin{aligned}
&\widehat{M}_0\hat{\alpha} - M_0\alpha\\
=&\widehat{M}_0(\widehat{M}^\top\widehat{M})^{-1}\widehat{M}^\top(M\alpha+\mathcal{E}) - M_0\alpha\\
=&\widehat{M}_0(\widehat{M}^\top\widehat{M})^{-1}\widehat{M}^\top(\widehat{M}+M-\widehat{M})\alpha + \widehat{M}_0(\widehat{M}^\top\widehat{M})^{-1}\widehat{M}^\top\mathcal{E} - M_0\alpha\\
=&(\widehat{M}_0 - M_0)\alpha - \widehat{M}_0(\widehat{M}^\top\widehat{M})^{-1}\widehat{M}^\top(\widehat{M}-M)\alpha + \widehat{M}_0(\widehat{M}^\top\widehat{M})^{-1}\widehat{M}\mathcal{E}.
\end{aligned}$$

Thus, by Lemma 16, 17, 18, 19, 20, 21, we have

$$\begin{aligned}
&\left\|\widehat{M}_0\hat{\alpha} - M_0\alpha\right\|_2\\
&\le \left\|\widehat{M}_0 - M_0\right\|_{\text{op}}\|\alpha\|_2 + \left\|\widehat{M}_0\right\|_{\text{op}}\left\|(\widehat{M}^\top\widehat{M})^{-1}\right\|_{\text{op}}\left\|\widehat{M}\right\|_{\text{op}}\left\|\widehat{M}-M\right\|_{\text{op}}\|\alpha\|_2\\
&+\left\|\widehat{M}_0(\widehat{M}^\top\widehat{M})^{-1}\widehat{M}^\top\mathcal{E}\right\|_2\\
&\le O\left(\|\alpha\|_2\sqrt{r}\delta\|X_0\|_2\|Y_0\|_2\right) + O\left(\|X_0\|_2\|Y_0\|_2\frac{1}{\eta^2\|Y_0\|_2^2\lambda_{\min}(\Sigma)}\frac{1}{N_2}\right.\\
&\left.\cdot\sqrt{N_2\lambda_{\max}(\Sigma)}\|Y_0\|_2\cdot\|Y_0\|_2\,\delta\sqrt{N_2 r\lambda_{max}(\Sigma)}\|\alpha\|_2\right)\\
&+O\left(\frac{\|X_0\|_2}{\sqrt{\lambda_{\max}(\Sigma)}}\frac{1}{\sqrt{N_2}}\sqrt{r+\log 2/\delta_2}\frac{1}{\eta^2}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\right)\\
&= O\left(\|X_0\|_2\|Y_0\|_2\,\delta\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\frac{1}{\eta^2}\sqrt{r}\|\alpha\|_2\right) + O\left(\|X_0\|_2\frac{\sqrt{\lambda_{\max}(\Sigma)}}{\lambda_{\min}(\Sigma)}\frac{1}{\eta^2}\sqrt{r+\log 2/\delta_2}\frac{1}{\sqrt{N_2}}\right).
\end{aligned}$$

□

Lemma 16. *Suppose each covariate* x_i *is mean-zero, satisfies* $\mathbb{E}[xx^\top] = \Sigma$ *and* Σ*-subgaussian, and* ε_i*'s are i.i.d. mean -zero, sub-gaussian variables with variance parameter 1, independent of* x_i*. If* $\left|Y_0^\top\hat{A}^2{}_i\right| \ge \eta\|Y_0\|_2$ *for all* $i\in[r]$*, and* $N_2 \ge O\left((r+\log 2/\delta_2)\left(\frac{1}{\eta^2}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\right)^2\right)$*, we have*

$$\left\|\widehat{M}_0(\widehat{M}^\top\widehat{M})^{-1}\widehat{M}^\top\mathcal{E}\right\|_2^2 \le O\left(\frac{\|X_0\|_2^2}{\lambda_{\max}(\Sigma)}\frac{1}{N_2}(r+\log 2/\delta_2)\left(\frac{1}{\eta^2}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\right)^2\right),$$

with probability at least $1-\delta_2/2$.

Proof. Note that

$$\left\|\widehat{M}_0(\widehat{M}^\top\widehat{M})^{-1}\widehat{M}^\top\mathcal{E}\right\|_2^2 = \mathcal{E}^\top G\mathcal{E},$$

with $G = \widehat{M}(\widehat{M}^\top\widehat{M})^{-1}\widehat{M}_0^\top\widehat{M}_0(\widehat{M}^\top\widehat{M})^{-1}\widehat{M}^\top$. By the Hanson-Wright inequality ([32], lemma 6.2.1) we have

$$P\left(\left|\boldsymbol{\varepsilon}^\top G\boldsymbol{\varepsilon} - E[\boldsymbol{\varepsilon}^\top G\boldsymbol{\varepsilon}]\right| \ge t\right) \le 2\exp\left(-c\min\left(\frac{t^2}{\|G\|_F^2}, \frac{t}{\|G\|_{\text{op}}}\right)\right).$$

Thus, $\mathcal{E}^\top G\mathcal{E} \le \mathbb{E}[\mathcal{E}^\top G\mathcal{E}] + O\left(\|G\|_F \sqrt{\log(2/\delta_1)}\right) + O\left(\|G\|_{\text{op}} \log(2/\delta_1)\right)$, with probability at least $1 - \delta_1/2$. By Lemma 17 19, 21,

$$\begin{aligned}
\mathbb{E}[\mathcal{E}^\top G\mathcal{E}] &= \text{Tr}(\widehat{M}(\widehat{M}^\top\widehat{M})^{-1}\widehat{M}_0^\top\widehat{M}_0(\widehat{M}^\top\widehat{M})^{-1}\widehat{M}^\top) \\
&\le r\|G\|_{\text{op}} \\
&\le r\left\|\widehat{M}\widehat{M}^\top\right\|_{\text{op}}\left\|(\widehat{M}^\top\widehat{M})^{-1}\right\|_{\text{op}}^2\left\|\widehat{M}_0^\top\widehat{M}_0\right\|_{\text{op}} \\
&\le O\left(\frac{r}{\eta^2}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\|X_0\|_2^2\|Y_0\|_2^2\,\frac{1}{\eta^2\|Y_0\|_2^2\,\lambda_{\min}(\Sigma)N_2}\right) \\
&= O\left(\frac{r}{\eta^4}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}^2(\Sigma)}\|X_0\|_2^2\,\frac{1}{N_2}\right),
\end{aligned}$$

with probability at least $1-\delta_1$, when $N_2 \ge O\left((r+\log 2/\delta_1)\left(\frac{1}{\eta^2}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\right)^2\right)$. In addition, $\|G\|_F\sqrt{\log 2/\delta_1} \le \sqrt{r\log 2/\delta_1}\|G\|_{\text{op}}$. Therefore,

$$\begin{aligned}
\mathcal{E}^\top G\mathcal{E} &\le O\left(r\|G\|_{\text{op}} + (\log 2/\delta_1)\|G\|_{\text{op}} + \sqrt{r\log 2/\delta_1}\|G\|_{\text{op}}\right) \\
&\le O\left((r+\log 2/\delta_1)\|G\|_{\text{op}}\right) \\
&\le O\left(\frac{r+\log 2/\delta_1}{\eta^4}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}^2(\Sigma)}\|X_0\|_2^2\,\frac{1}{N_2}\right).
\end{aligned}$$

□

Lemma 17. *Suppose each covariate x_i is mean-zero, satisfies $\mathbb{E}[xx^\top] = \Sigma$ and Σ-subgaussian, and $\left|Y_0^\top \hat{A}^2{}_i\right| \ge \eta\|Y_0\|_2$ for all $i \in [r]$. When $N_2 \ge O\left((r+\log 1/\delta_2)\left(\frac{1}{\eta^2}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\right)^2\right)$, the matrix $\widehat{M}^\top\widehat{M}$ is invertible and*

$$\left\|(\widehat{M}^\top\widehat{M})^{-1}\right\|_{op} \le O\left(\frac{1}{\eta^2\|Y_0\|_2^2\,\lambda_{\min}(\Sigma)}\frac{1}{N_2}\right),$$

with probability at least $1-\delta_1$.

Proof. Note that $\widehat{M} = (Y_0 \otimes \mathcal{X}^\top)^\top(\hat{A}^2 \odot \hat{A}^1) = (Y_0^\top \otimes \mathcal{X})(\hat{A}^2 \odot \hat{A}^1) = Y_0^\top\hat{A}^2 \odot \mathcal{X}\hat{A}^1$. Thus, by defining matrix $U \in \mathbb{R}^{d_1 \times r}$ with columns $U_i = (Y_0^\top\hat{A}_i^2)\hat{A}_i^1$, it can be written that $\widehat{M}^\top\widehat{M} = U^\top\mathcal{X}^\top\mathcal{X}U$. Note that the columns of U are orthogonal with each other. Since $\mathbb{E}[\frac{1}{N_2}\widehat{M}^\top\widehat{M}] = \mathbb{E}[U^\top(\frac{1}{N_2}\mathcal{X}^\top\mathcal{X})U] =$

$U^\top \Sigma U$, we let $\frac{1}{N_2}\widehat{M}^\top \widehat{M} = \mathcal{E} + U^\top \Sigma U$ with matrix $\mathcal{E}$ satisfying $\mathbb{E}[\mathcal{E}] = 0$. In addition, we have

$$\begin{aligned}
\|U\|_{\text{op}} &= \left\| \hat{A}^1 \text{diag}[(Y_0^\top \hat{A}_i^2)_{i=1}^r] \right\|_{\text{op}} \\
&\le \left\| \hat{A}^1 \right\|_{\text{op}} \max_{i\in[r]} |Y_0^\top \hat{A}_i^2| \\
&\le \left\| \hat{A}^1 \right\|_{\text{op}} \cdot \|Y_0\|_2 \max_{i\in[r]} \left\| \hat{A}_i^2 \right\|_2 \\
&\le \left\| \hat{A}^1 \right\|_{\text{op}} \cdot \|Y_0\|_2 \left(\sum_{i\in[r]} \left\| \hat{A}_i^2 \right\|_2^2 \right)^{\frac{1}{2}} \\
&= 1 \cdot \|Y_0\|_2 \cdot 1 = \|Y_0\|_2 .
\end{aligned}$$

Thus, applying Lemma 12, we conclude that as long as $N_2 \ge O(r + \log(1/\delta_1))$ we have $\|\mathcal{E}\|_{\text{op}} \le O\left(\lambda_{\max}(\Sigma)\|Y_0\|_2^2 \left(\sqrt{\frac{r+\log(1/\delta_1)}{N_2}}\right)\right)$ with probability at least $1-\delta_1$. Besides,

$$\lambda_{\min}(U^\top \Sigma U) = \min_{x\in\mathbb{R}^r} \frac{x^\top U^\top \Sigma U x}{x^\top x} \ge \eta^2 \|Y_0\|_2^2 \min_{\omega\in\mathbb{R}^{d_1}} \frac{\omega^\top \Sigma \omega}{\omega^\top \omega} = \eta^2 \|Y_0\|_2^2 \lambda_{\min}(\Sigma),$$

where the first inequality follows from substituting $\omega = Ux$ and observing

$$\omega^\top \omega = x^\top U^\top U x \ge \min_i |Y_0^\top \hat{A}_i^2|^2 x^\top x \ge \eta^2 \|Y_0\|_2^2 \, x^\top x.$$

Therefore,

$$\begin{aligned}
&\lambda_{\min}(\frac{1}{N_2}\widehat{M}^\top \widehat{M}) \ge \lambda_{\min}(U^\top \Sigma U) - \lambda_{\max}(\mathcal{E}) \\
&\ge \eta^2 \|Y_0\|_2^2 \lambda_{\min}(\Sigma) - \|\mathcal{E}\|_{\text{op}} \\
&\ge O\left(\eta^2 \|Y_0\|_2^2 \lambda_{\min}(\Sigma) - \lambda_{\max}(\Sigma)\|Y_0\|_2^2 \sqrt{\frac{r+\log 1/\delta_1}{N_2}} \right) \\
&= O\left(\|Y_0\|_2^2 \left(\eta^2 \lambda_{\min}(\Sigma) - \lambda_{\max}(\Sigma) \sqrt{\frac{r+\log 1/\delta_1}{N_2}} \right) \right).
\end{aligned}$$

Therefore, as long as $N_2 \ge O\left((r+\log 1/\delta_1)\left(\frac{\lambda_{\max}(\Sigma)}{\eta^2 \lambda_{\min}(\Sigma)}\right)^2\right)$, $\frac{1}{N_2}\widehat{M}^\top \widehat{M}$ is invertible and so is $\widehat{M}^\top \widehat{M}$. Now, $(\widehat{M}^\top \widehat{M})^{-1} = \frac{1}{N_2}(\frac{1}{N_2}\widehat{M}^\top \widehat{M})^{-1} = \frac{1}{N_2}(\mathcal{E} + U^\top \Sigma U)^{-1}$. Moreover,

$$\left\| (U^\top \Sigma U)^{-1} \mathcal{E} \right\|_{\text{op}} \le \left\| (U^\top \Sigma U)^{-1} \right\|_{\text{op}} \|\mathcal{E}\|_{\text{op}} \le \frac{\|\mathcal{E}\|_{\text{op}}}{\lambda_{\min}(U^\top \Sigma U)} \le O\left(\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} \frac{1}{\eta^2} \sqrt{\frac{r+\log 1/\delta_1}{N_2}} \right).$$

Therefore, as long as $N_2 \ge O\left((r+\log 1/\delta_2)\left(\frac{1}{\eta^2}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\right)^2\right)$, we have $\left\| (U^\top \Sigma U)^{-1} \mathcal{E} \right\|_{\text{op}} \le 1/4$. Finally, applying Lemma 14 we have $(\frac{1}{N_2}\widehat{M}^\top \widehat{M})^{-1} = (U^\top \Sigma U)^{-1} + F$, where $\|F\|_{\text{op}} \le \frac{1}{3}\left\| (U^\top \Sigma U)^{-1} \right\|_{\text{op}}$,

and

$$\left\|(\widehat{M}^\top \widehat{M})^{-1}\right\| \le \frac{4}{3N_2}\left\|(U^\top \Sigma U)^{-1}\right\|_{\text{op}} \le \frac{4}{3N_2}\frac{1}{\eta^2\|Y_0\|_2^2\,\lambda_{\min}(\Sigma)}.$$

□

Lemma 18. *Suppose each covariate x_i is mean-zero, satisfies $\mathbb{E}[xx^\top] = \Sigma$ and Σ-subgaussian, and $\max\{\sin\theta(\hat{A}^1, A^1), \sin\theta(\hat{A}^2, A^2)\} \le \delta$, then if $N_2 \ge O(r + \log 1/\delta_1)$, we have*

$$\left\|\widehat{M} - M\right\|_{op} \le O\left(\|Y_0\|_2\,\delta\sqrt{N_2 r\lambda_{max}(\Sigma)}\right),$$

with probability at least $1 - \delta_1$.

Proof. Write $\hat{A}^1 = A^1 + E^1$, $\hat{A}^2 = A^2 + E^2$. Note that $\hat{A}^1_\perp \hat{A}^{1\top}_\perp + \hat{A}^1\hat{A}^{1\top} = I_{d_1}$, we have

$$\begin{aligned}\left\|A^1 - \hat{A}^1\right\|_{\text{op}} &= \left\|\hat{A}^1_\perp \hat{A}^{1\top}_\perp A^1 - \hat{A}^1(I_r - \hat{A}^{1\top}A^1)\right\|_{\text{op}} \le \left\|\hat{A}^1_\perp\right\|_{\text{op}}\left\|\hat{A}^{1\top}_\perp A^1\right\|_{\text{op}} + \left\|\hat{A}^1\right\|_{\text{op}}\left\|I_r - \hat{A}^{1\top}A^1\right\|_{\text{op}} \\ &\le 1\cdot\sin\theta(\hat{A}^1, A^1) + 1\cdot\sin^2\theta(\hat{A}^1, A^1) = O(\delta),\end{aligned}$$

where the last inequality is due to

$$\begin{aligned}\left\|I_r - \hat{A}^{1\top}A^1\right\|_{\text{op}} &= \lambda_{\max}(I_r - \hat{A}^{1\top}A^1) = 1 - \lambda_{\min}(\hat{A}^{1\top}A^1) = 1 - \cos\theta(\hat{A}^1, A^1) \\ &= \frac{1 - \cos^2\theta(\hat{A}^1, A^1)}{1 + \cos\theta(\hat{A}^1, A^1)} = \frac{\sin^2\theta(\hat{A}^1, A^1)}{1 + \cos\theta(\hat{A}^1, A^1)} \le \sin^2\theta(\hat{A}^1, A^1) = \delta^2,\end{aligned}$$

where $\theta(\hat{A}^1, A^1)$ is the principal angle between column subspaces of $\hat{A}^1$ and A^1. Therefore, $\left\|E^1\right\|_{\text{op}} \le O(\delta)$, in the same way, $\left\|E^2\right\|_{\text{op}} \le O(\delta)$. Now,

$$\begin{aligned}\left\|\widehat{M} - M\right\|_{\text{op}} &= \left\|(Y_0 \otimes \mathcal{X}^\top)^\top(\hat{A}^2 \odot \hat{A}^1) - (Y_0 \otimes \mathcal{X}^\top)^\top(A^2 \odot A^1)\right\|_{\text{op}} \\ &= \left\|(Y_0^\top\hat{A}^2 \odot \mathcal{X}\hat{A}^1) - (Y_0^\top A^2 \odot \mathcal{X}A^1)\right\|_{\text{op}} \\ &= \left\|(Y_0^\top(A^2 + E^2) \odot \mathcal{X}(A^1 + E^1)) - (Y_0^\top A^2 \odot \mathcal{X}A^1)\right\|_{\text{op}} \\ &\le \left\|Y_0^\top A^2 \odot \mathcal{X}E^1\right\|_{\text{op}} + \left\|Y_0^\top E^2 \odot \mathcal{X}A^1\right\|_{\text{op}} + \left\|Y_0^\top E^2 \odot \mathcal{X}E^1\right\|_{\text{op}}.\end{aligned}$$

Consider the first term,

$$\begin{aligned}\left\|Y_0^\top A^2 \odot \mathcal{X}E^1\right\|_{\text{op}}^2 &= \lambda_{\max}((Y_0^\top A^2 \odot \mathcal{X}E^1)^\top(Y_0^\top A^2 \odot \mathcal{X}E^1)) \\ &= \lambda_{\max}(U^\top\mathcal{X}^\top\mathcal{X}U) = N_2\lambda_{\max}(U^\top(\frac{1}{N_2}\mathcal{X}^\top\mathcal{X})U),\end{aligned}$$

where $U \in \mathbb{R}^{d_1\times r}$ has columns $U_i = (Y_0^\top A_i^2)E_i^1$ with A_i^2 and E_i^1 being columns of A^2 and E^1. Note that $\|U\|_{\text{op}} \le \|Y_0\|_2\left\|E^1\right\|_F \le \|Y_0\|_2\sqrt{r}\left\|E^1\right\|_{\text{op}} = O(\sqrt{r}\|Y_0\|_2\,\delta)$, by Lemma 12, $\lambda_{\max}(U^\top(\frac{1}{N_2}\mathcal{X}^\top\mathcal{X})U)$ is uppper bounded by $O\left(\left\|U^\top\Sigma U\right\|_{\text{op}} + \lambda_{\max}(\Sigma)r\|Y_0\|_2^2\,\delta^2\sqrt{\frac{r+\log 1/\delta_1}{N_2}}\right)$. Moreover, $\left\|U^\top\Sigma U\right\|_{\text{op}} \le$

$\lambda_{\max}(\Sigma)\|U\|_{\text{op}}^2 \leq \lambda(\Sigma) r \|Y_0\|_2^2 \delta^2$. Thus, when $N_2 \geq O(r + \log 1/2\delta_1)$, $\lambda_{max}(U^\top(\frac{1}{N_2}\mathcal{X}^\top\mathcal{X})U) \leq O(\lambda_{\max}(\Sigma) r \|Y_0\|_2^2 \delta^2)$. Therefore, $\left\|Y_0^\top A^2 \odot \mathcal{X}E^1\right\|_{\text{op}} \leq O\left(\delta \|Y_0\|_2 \sqrt{N_2 r \lambda_{\max}(\Sigma)}\right)$.

The second term $\left\|Y_0^\top E^2 \odot \mathcal{X}A^1\right\|_{\text{op}}$ and the third term $\left\|Y_0^\top E^2 \odot \mathcal{X}E^1\right\|_{\text{op}}$ can be shown in the same way having an upper bound of the same magnitude. □

Lemma 19. *Suppose each covariate x_i is mean-zero, satisfies $\mathbb{E}[xx^\top] = \Sigma$ and Σ-subgaussian, and $N_2 \geq O(r + \log 1/\delta_1)$. Then*

$$\left\|\widehat{M}\widehat{M}^\top\right\|_{op} \leq O\left(N_2 \lambda_{\max}(\Sigma)\|Y_0\|_2^2\right),$$

with probability at least $1 - \delta_1$.

Proof. Write

$$\begin{aligned}\left\|\widehat{M}\widehat{M}^\top\right\|_{\text{op}} &= \lambda_{\max}(\widehat{M}^\top\widehat{M}) = \lambda_{\max}((Y_0^\top\hat{A}^2 \odot \mathcal{X}\hat{A}^1)^\top(Y_0^\top\hat{A}^2 \odot \mathcal{X}\hat{A}^1)) \\ &= \lambda_{\max}(U^\top(\mathcal{X}^\top\mathcal{X})U) = N_2\lambda_{\max}(U^\top(\frac{1}{N_2}\mathcal{X}^\top\mathcal{X})U),\end{aligned}$$

where U has orthogonal columns $U_i = (Y_0^\top\hat{A}_i^2)\hat{A}_i^1$. Since $\|U\|_{\text{op}} \leq \|Y_0\|_2 \max_{i\in[r]}\left\|\hat{A}_i^2\right\|_2 \leq \|Y_0\|_2$. By Lemma 12, when $N_2 \geq O(r + \log 1/\delta_1)$,

$$\lambda_{\max}(U^\top(\frac{1}{N_2}\mathcal{X}^\top\mathcal{X})U) \leq O\left(\left\|U^\top\Sigma U\right\|_{\text{op}} + \lambda_{\max}(\Sigma)\|Y_0\|_2^2\sqrt{\frac{r + \log 1/\delta_1}{N_2}}\right) \leq O(\lambda_{\max}(\Sigma)\|Y_0\|_2^2).$$

Therefore, $\left\|\widehat{M}\widehat{M}^\top\right\|_{\text{op}} \leq O(N_2\lambda_{\max}(\Sigma)\|Y_0\|_2^2)$ with probability at least $1 - \delta_1$. □

Lemma 20. *If $\max\{\sin\theta(\hat{A}^1, A^1), \sin\theta(\hat{A}^2, A^2)\} \leq \delta$, then*

$$\left\|\widehat{M}_0 - M_0\right\|_{op} \leq O\left(\sqrt{r}\delta\|X_0\|_2\|Y_0\|_2\right).$$

Proof. Let $\hat{A}^2 = A^2 + E^2$, $\hat{A}^1 = A^1 + E^1$, then

$$\left\|\widehat{M}_0 - M_0\right\|_{\text{op}} \leq \left\|Y_0^\top A^2 \odot X_0^\top E^1\right\|_{\text{op}} + \left\|Y_0^\top E^2 \odot X_0^\top A^1\right\|_{\text{op}} + \left\|Y_0^\top E^2 \odot X_0^\top E^1\right\|_{\text{op}}.$$

The first term $\left\|Y_0^\top A^2 \odot X_0^\top E^1\right\|_{\text{op}}^2 = \lambda_{\max}(U^\top X_0 X_0^\top U)$, where U has columns $U_i = (Y_0^\top A_i^2)E_i^1$. Using the upper bound of $\left\|E^1\right\|_{\text{op}}$ in Lemma 18 we have $\|U\|_{\text{op}} \leq \|Y_0\|_2\left\|E^1\right\|_F \leq O(\sqrt{r}\|Y_0\|_2\,\delta)$. Therefore,

$$\left\|Y_0^\top A^2 \odot X_0^\top E^1\right\|_2 \leq \sqrt{\|U\|_{\text{op}}^2\|X_0\|_2^2} \leq O\left(\sqrt{r}\|X_0\|_2\|Y_0\|_2\,\delta\right).$$

Thus, $\left\|\widehat{M}_0 - M_0\right\|_{\text{op}} \leq O(\sqrt{r}\delta\|X_0\|_2\|Y_0\|_2)$. □

Lemma 21.

$$\left\|\widehat{M}_0\right\|_{op} \leq \|X_0\|_2 \|Y_0\|_2 \,.$$

Proof.

$$\left\|\widehat{M}_0\right\|_{\mathrm{op}} = \left\|Y_0^\top \hat{A}^2 \odot X_0^\top \hat{A}^1\right\|_{\mathrm{op}} = \left\|X_0^\top U\right\|_{\mathrm{op}} \leq \|X_0\|_2 \|U\|_{\mathrm{op}} \leq \|X_0\|_2 \|Y_0\|_2 \,,$$

where U has columns $U_i = (Y_0^\top \hat{A}_i^2)\hat{A}_i^1$. □