Collaborative Learning with Shared Linear Representations: Statistical Rates and Optimal Algorithms

Xiaochun Niu[1] **Lili Su**[2] **Jiaming Xu**[1] **Pengkun Yang**[3]

[1]The Fuqua School of Business, Duke University
[2]Department of Electrical and Computer Engineering, Northeastern University
[3]Department of Statistics and Data Science, Tsinghua University
{xn20,jx77}@duke.edu, l.su@northeastern.edu, yangpengkun@tsinghua.edu.cn.

Abstract

Collaborative learning enables multiple clients to learn shared feature representations across local data distributions, with the goal of improving model performance and reducing overall sample complexity. While empirical evidence shows the success of collaborative learning, a theoretical understanding of the optimal statistical rate remains lacking, even in linear settings. In this paper, we identify the optimal statistical rate when clients share a common low-dimensional linear representation. Specifically, we design a spectral estimator with local averaging that approximates the optimal solution to the least squares problem. We establish a minimax lower bound to demonstrate that our estimator achieves the optimal error rate. Notably, the optimal rate reveals two distinct phases. In typical cases, our rate matches the standard rate based on the parameter counting of the linear representation. However, a statistical penalty arises in collaborative learning when there are too many clients or when local datasets are relatively small. Furthermore, our results, unlike existing ones, show that, at a system level, collaboration always reduces overall sample complexity compared to independent client learning. In addition, at an individual level, we provide a more precise characterization of when collaboration benefits a client in transfer learning and private fine-tuning.

1 Introduction

Collaborative learning of shared feature representations across data distributions has become a crucial challenge in machine learning and data science. The goal is to extract common structures from related distributions to improve model performance and reduce overall sample complexity, compared to independently learning each distribution from scratch. Such problems find widespread applications in fields like federated learning [7], multi-task transfer learning [22, 9], and private fine-tuning with public knowledge [29]. For example, in federated learning, clients collaboratively learn a shared model using datasets sampled from their local distributions. In healthcare, federated learning enables doctors to improve disease detections or treatment effect predictions by leveraging medical data stored at multiple hospitals [24]. In addition, multi-task transfer learning enables knowledge transfer across tasks using a pretrained common model. This technique is applicable in few-shot image classification [28], deep reinforcement learning [39], and large language models [19]. Likewise, in private fine-tuning, a shared model is first pretrained on publicly available data and then fine-tuned for specific tasks using private datasets [37]. While empirical evidence shows the success of collaborative

International Workshop on Federated Foundation Models in Conjunction with NeurIPS 2024 (FL@FM-NeurIPS'24). This workshop does not have official proceedings and this paper is non-archival.

learning, these studies often involve large datasets. A theoretical understanding of the optimal statistical rate is still lacking, even in linear settings.

There is extensive literature in linear settings [8, 33, 7, 30, 32, 9, 38, 21], and our work contributes to this line. In such settings, there are M clients (or tasks), where each client i observes n_i data points $\{(x_{ij}, y_{ij})\}_{j=1}^{n_i}$. Here $x_{ij} \in \mathbb{R}^d$ is the covariate and $y_{ij} \in \mathbb{R}$ is the response of the j-th sample. Let $N = \sum_{i=1}^{M} n_i$ be the total number of data across all clients. For client $i \in [M]$ and sample $j \in [n_i]$, the response y_{ij} is represented as

$$y_{ij} = x_{ij}^\intercal \theta_i^\star + \xi_{ij}, \tag{1.1}$$

where $\theta_i^\star \in \mathbb{R}^d$ is the ground-truth parameter and $\xi_{ij} \in \mathbb{R}$ is an additive noise. Suppose that for each i, the covariates $\{x_{ij}\}_{j=1}^{n_i}$ share the same but *unknown* covariance, i.e., $\mathbb{E}[x_{ij}x_{ij}^\intercal] = \Gamma_i$ for all j. We further assume a low-dimensional structure on the parameters, where there exist an orthonormal matrix $B^\star \in \mathcal{O}^{d\times k}$ with $k \le d$ and vectors $\alpha_i^\star \in \mathbb{R}^k$ such that $\Gamma_i\theta_i^\star = B^\star\alpha_i^\star$ for all i. Here $B^\star$ is the shared low-dimensional representation and $\alpha_i^\star$ is the client-specific parameter for client i. The clients aim to collaboratively learn the shared representation $B^\star$ using their observed datasets. Let $\sum_{i=1}^{M} n_i\alpha_i^\star(\alpha_i^\star)^\intercal/N$ be the client diversity matrix, where for now we assume its condition number, the ratio of its largest to smallest eigenvalue, is $\Theta(1)$. This assumption ensures that $\alpha_i^\star$ are not concentrated in certain directions and the data partition n_i is not dominated by a few clients. Consequently, we have sufficient information to accurately estimate all k columns of $B^\star$. In particular, this implies that the client diversity matrix is full rank, so that $k \le M$.[1]

Several previous works [33, 8, 7, 30, 10, 9] have studied the error rate of a consistent estimation for $B^\star$ in this problem. However, existing results exhibit a suboptimal dependence on the subspace dimension k. In particular, there remains a gap between the best-known error bounds, where the upper bound is $O(\sqrt{dk^2/N})$ [33, 8, 10, 9], while the lower bound is $\Omega(\sqrt{dk/N})$ [33].[2] In fact, this suboptimality has been acknowledged in many works [33, 30, 29, 32] as a challenging open problem:

What is the optimal statistical rate to learn the low-dimensional representation $B^\star$?

Identifying the optimal statistical rate is also crucial for understanding the benefits of collaboration compared to clients learning their parameters independently. Aside from the undetermined optimal rate, it remains unclear whether a statistical-computational gap exists. Specifically, we aim to design an efficient estimator to achieve optimal error rates with polynomial runtime. In addition, it is worth noting that when directly observing data points composed of $B^\star$ and noises, rather than through multiple $B^\star\alpha_i^\star$, the standard optimal error rate for estimating $B^\star$ is $\Theta(\sqrt{dk/N})$, based on the parameter counting that $B^\star$ has dk entries. Therefore, we are also interested in comparing the optimal rate for the multi-client problem with this standard rate.

Main Contributions. Our work addresses this challenging open problem by improving both the upper and lower bounds (Corollaries 3.1 and 4.1) and identifying the optimal statistical rate, $\Theta(\sqrt{dk/N} + \sqrt{Mdk^2/N^2})$. To further illustrate our results, we consider a specific regime:

$$n_i \equiv n, \quad N = Mn, \quad n = k^\beta, \quad M = k^{\gamma+1}, \quad d = k^{\delta+1}, \tag{1.2}$$

where $\beta, \gamma, \delta > 0$ are fixed constants since $k \le \min\{d, M\}$. Figure 1 illustrates the phase diagram, with each region indicating whether a consistent estimation of $B^\star$ with vanishing estimation error is possible or not.[3] In particular, a consistent estimation is impossible in light red Region I and possible in light blue Region II, identified by previous works [33, 8, 10]. However, a wide gap remains between these two regions. Our contribution bridges this gap by identifying the optimal sample complexity, which delineates the boundary between Region IV and Regions I and III. We prove that a consistent estimation is impossible in dark red Region III and possible in dark blue Region IV.

[1] Otherwise, if $M < k$, estimating $B^\star$ accurately is impossible. In this case, $\{B^\star\alpha_i^\star\}_{i=1}^{M}$ spans only an M-dimensional subspace of $\mathbb{R}^k$; thus the data $\{x_{ij}\}$ contains information only about that subspace. The remaining $k - M$ columns of $B^\star$ can be any vectors from the $(d - M)$-dimensional complementary space.

[2] For ease of presentation, we will ignore polylogarithmic factors and use standard big O notations in Section 1. In later sections, we will use $\widetilde{O}$ to emphasize the hidden polylogarithmic factors.

[3] For the existing error upper bound $\sqrt{dk^2/N}$ to be $\Theta(1)$, we need $N = \Theta(dk^2)$. This translates to $\beta + \gamma = \delta + 2$ in the regime of (1.2). In a similar vein, the existing lower bound $\sqrt{dk/N}$ corresponds to $\beta + \gamma = \delta + 1$, and the second term in our established error rate $\sqrt{Mdk^2/N^2}$ corresponds to $2\beta + \gamma = \delta + 2$.

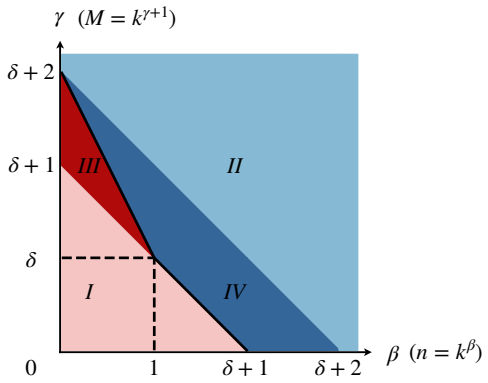

Figure 1: The phase diagram for estimating $B^\star$ in the regime of (1.2). Regions I and II show the impossibility and possibility, respectively, from previous works. Regions III and IV show the impossibility and possibility, respectively, in our work. The boundary between Region I and Regions III and IV is $\beta+\gamma=\delta+1$, between Regions III and IV is $2\beta+\gamma=\delta+2$, and between Regions II and IV is $\beta+\gamma=\delta+2$.

The identified optimal statistical rate, $\Theta(\sqrt{dk/N}+\sqrt{Mdk^2/N^2})$, delineates the boundary between Region IV and Regions I and III in Figure 1, with two distinct phases. Recall that the standard rate $\Theta(\sqrt{dk/N})$ from parameter counting corresponds to $\beta+\gamma=\delta+1$ in this diagram. Our optimal rate matches this standard rate along the boundary between Regions I and IV. However, deviations from the standard rate arise along the boundary between Regions III and IV, when $\gamma>\delta$ or $0<\beta<1$, that is, when $M=\Omega(d)$ or $n=O(k)$. This is when the second term $\sqrt{Mdk^2/N^2}$ becomes dominant. Such a discrepancy highlights a statistical penalty to pay for multi-client collaborative learning, particularly when there are many clients in the system ($M=\Omega(d)$) or when clients have an insufficient number of local data points compared to the subspace dimension ($n=O(k)$).

Our optimal rate quantifies the benefits of collaboration for both the overall system and individual clients. At the system level, unlike previous results, we find that collaboration always reduces overall sample complexity compared to independent client learning (Remark 3.1). However, at the individual level, collaboration is not always beneficial. By applying our optimal rates, we achieve tighter error bounds for transfer learning (Corollary 5.1) and private fine-tuning (Corollary 5.2) at an individual client. This provides a more precise quantification of when collaboration benefits individual clients.

In addition to determining the rate, we introduce the first estimator in the literature that achieves this optimal statistical rate, thereby showing no statistical-computational gap here. Our spectral estimator of $B^\star$ is an optimal solution to an approximated version of the non-convex least squares problem. This estimator leverages two independent replicas of local averages of cross-correlation vectors $\{y_{ij}x_{ij}\}_{j=1}^{n_i}$ at each client. It preserves privacy in federated learning settings since clients can send only their local averages rather than raw data to the server. Furthermore, our estimator only requires at least two data points for each client,[4] which significantly relaxes the strict assumptions that $N/M=\Omega(d)$ imposed by [8, 9, 32].[5]

Finally, our results extend to general ill-conditioned cases, where the analysis allows for any imbalanced data partition and covariate heterogeneity (i.e. covariate shift). Let λ_1 and λ_k be the largest and smallest eigenvalues, respectively, of the client diversity matrix. Our estimator achieves an error upper bound $O(\sqrt{d\lambda_1/(N\lambda_k^2)}+\sqrt{Md/(N^2\lambda_k^2)})$ with $\lambda_1=O(1)$ in Theorem 3.1, improving the best-known error rate $O(\sqrt{d/(N\lambda_k^2)})$ [33, 8, 10]. In addition, Theorem 4.1 establishes a minimax lower bound $\Omega(\sqrt{d/(N\lambda_k)}+\sqrt{Md/(N^2\lambda_k^2)})$, which differs from the upper bound only by a condition number $\sqrt{\lambda_1/\lambda_k}$ in the first term. This lower bound improves the state-of-the-art result in [33], which is $\Omega(\sqrt{1/(N\lambda_k)}+\sqrt{dk/N})$ with $\lambda_k=O(1/k)$. In particular, our lower bound

[4] Note that in the extreme case where every client has only a single data point, i.e., $n_i\equiv 1$, the existing error bound $O(\sqrt{dk^2/N})$ in [33, 10] already matches our improved lower bound in Corollary 4.1. Therefore, assuming $n_i\geq 2$ does not result in a significant loss of generality.

[5] [32] is only able to identify an optimal rate $\Theta(\sqrt{dk/N})$ in the restricted setting $N/M=\Omega(d)$. In contrast, we characterize the optimal rate for the entire region and discover the two distinct phases of the rate.

improves the first term by a factor of $\sqrt{d}$, by leveraging the packing set on the unit sphere to construct multiple problem instances, rather than using Le Cam's two-point method as in [33]. Moreover, the second term in our bound is entirely new, derived by choosing randomly generated α_i's instead of deterministic ones.

Notation and Organization. For positive integers $k \le d$, let $\mathcal{O}^{d\times k}$ be the set of $d \times k$ matrices with k orthogonal unit vectors as columns. Let $\mathbb{S}^{d-1}$ be the unit sphere in $\mathbb{R}^d$. For a matrix M, let $\lambda_r(M)$ denote the r-th largest eigenvalue value of M and $\|M\|$ denote the spectral norm. We use the Bachmann–Landau notations O, Ω, and Θ, and use $\widetilde{O}$ to hide polylogarithmic factors in quantities.

The rest of the paper is organized as follows. Section 2 presents the main model with assumptions. Section 3 introduces our estimator and the error upper bound. Section 4 establishes a minimax lower bound. Section 5 provides several applications of our results. We conclude the paper and discuss future directions in Section 6. In the appendix, Section A provides further related works. Sections B, C, and D prove the upper and lower bounds, and the corollaries of applications, respectively.

2 Model and Assumptions

Our targeting problem is modeled in (1.1). Formally, we impose the following assumptions on the variables, ξ_{ij}, x_{ij}, and $\theta_i^\star$, with a key assumption of the low-dimensional structure for $\theta_i^\star$.

Assumption 2.1 (Sub-gaussian noises)**.** The noise variables ξ_{ij} are independent, zero-mean, sub-gaussian[6] with constant variance proxy $\sigma^2 = \Theta(1)$ and are independent of covariates x_{ij}.

Assumption 2.2 (Sub-gaussian covariates)**.** The covariates x_{ij} are independent, zero-mean, sub-gaussian with variance proxy $\gamma^2 = \Theta(1)$. For each i, x_{ij} share the same but *unknown* covariance, i.e., $\mathbb{E}[x_{ij}x_{ij}^\intercal] = \Gamma_i$ for all j. These covariance matrices are well-conditioned, with $\lambda_1(\Gamma_i)/\lambda_d(\Gamma_i) = \Theta(1)$ for all i.

The sub-gaussian assumptions are standard in statistical learning for deriving tail bounds. Assumption 2.2 generalizes those in [33, 10] by allowing non-identity covariance. We now assume $\Gamma_i\theta_i^\star$ has a common low-dimensional structure.

Assumption 2.3 (Low-dimensional structure)**.** There exist $B^\star \in \mathcal{O}^{d\times k}$ with $k \le d$ and $\alpha_i^\star \in \mathbb{R}^k$ such that $\Gamma_i\theta_i^\star = B^\star\alpha_i^\star$ for $i \in [M]$.

Here $B^\star \in \mathcal{O}^{d\times k}$ is the shared low-dimensional representation and $\alpha_i^\star \in \mathbb{R}^k$ is the client-specific parameter for client i. When $\Gamma_i = I_d$ for all i, this reduces to the standard assumption $\theta_i^\star = B^\star\alpha_i^\star$, imposed by previous works such as [33, 10]. We generalize this standard assumption to the case with non-identity covariance Γ_i, by requiring the cross-correlation vector $\mathbb{E}[y_{ij}x_{ij}] = \Gamma_i\theta_i^\star$ to share a common subspace.[7]

Assumption 2.4 (Client normalization)**.** Each $\alpha_i^\star$ satisfies $\|\alpha_i^\star\| = O(1)$ for $i \in [M]$.

The normalization assumption is standard in the literature. Let $\lambda_r = \lambda_r(\sum_{i=1}^M n_i\alpha_i^\star(\alpha_i^\star)^\intercal)/N$ denote the r-th largest eigenvalue of the client diversity matrix $\sum_{i=1}^M n_i\alpha_i^\star(\alpha_i^\star)^\intercal/N$ for $r \in [k]$, unless otherwise specified. The normalization then gives $\sum_{r=1}^k \lambda_r = \mathrm{Tr}(\sum_{i=1}^M n_i\alpha_i^\star(\alpha_i^\star)^\intercal)/N = \sum_{i=1}^M n_i\|\alpha_i^\star\|^2/N = O(1)$, which further implies that $k\lambda_k \le O(1)$ and $\lambda_1 = O(1)$.

Given the model described in (1.1) and the assumptions introduced, the goal of the clients in these problems is to collectively estimate the shared representation $B^\star$. In particular, we define the following metric to measure the distance between two orthonormal matrices.

Definition 2.1 (Principal angle distance)**.** Let $B, B^\star \in \mathcal{O}^{d\times k}$ be orthonormal matrices. Then the principal angle distance between B and $B^\star$ is

$$\|\sin\Theta(B, B^\star)\| = \|BB^\intercal - B^\star(B^\star)^\intercal\|.$$

[6] A random variable $\xi \in \mathbb{R}$ is sub-gaussian with variance proxy σ^2, denoted by $\xi \sim \mathrm{subG}(\sigma^2)$, if $\mathbb{E}\xi = 0$ and $\mathbb{E}[\exp(t\xi)] \le \exp(\sigma^2t^2/2)$ for any $t \in \mathbb{R}$. A random vector $\xi \in \mathbb{R}^d$ is sub-gaussian with variance proxy σ^2, denoted by $\xi \sim \mathrm{subG}_d(\sigma^2)$, if $u^\intercal\xi \sim \mathrm{subG}(\sigma^2)$ for any $u \in \mathbb{S}^{d-1}$.

[7] Note that since the covariance matrices Γ_i are unknown, one cannot apply the whitening procedure by writing $x_{ij} = \Gamma_i^{1/2}\widetilde{x}_{ij}$ so that $\widetilde{x}_{ij}$ have an identity covariance matrix. Moreover, it is difficult to accurately estimating Γ_i, as the size of the local dataset $n_i \ll d$.

The principal angle distance measures the separation between the column spaces of B and $B^\star$, and is invariant to any rotations of B and $B^\star$. Using this metric, the clients aim to learn an estimator $\widehat{B}$ that minimizes $\|\sin\Theta(B, B^\star)\|$ over $B \in \mathcal{O}^{d\times k}$, ensuring that the column space of $\widehat{B}$ closely aligns with that of $B^\star$. We study the optimal statistical rate of this problem.

3 Estimator and Error Upper Bound

In this section, we propose an estimator of $B^\star$, designed to achieve the optimal error upper bound.

3.1 Optimal Estimator

We review the limitations of existing estimators and introduce the innovations of our proposed one.

Limitations of Existing Estimators. Many recent works have designed estimators with provable error bounds [8, 33, 30, 7, 10]. The method-of-moments estimator in [33] is formed by the top-k eigenvectors of the matrix,

$$Z_1 = \sum_{i=1}^{M}\sum_{j=1}^{n_i} y_{ij}^2 x_{ij}x_{ij}^\intercal\,. \tag{3.1}$$

Their analysis is limited to cases where $x_{ij} \sim N(0, I_d)$, and the corresponding error upper bound is suboptimal compared to the lower bound [33, Theorem 5]. A subsequent work [10] assumes $n_i \geq 2$ and introduces an estimator using the matrix,

$$Z_2 = \sum_{i=1}^{M}\frac{1}{n_i - 1}\sum_{j_1\neq j_2} y_{ij_1}y_{ij_2}x_{ij_1}x_{ij_2}^\intercal\,. \tag{3.2}$$

By excluding the diagonal terms $j_1 = j_2$ in the summation, their estimator is designed to handle scenarios where the noise ξ_{ij} may depend on x_{ij} and shown to achieve the suboptimal error bound of [33]. However, whether this estimator provides any improvement on the estimation error rates remains unclear. Several works [30, 7] study the alternating minimization methods. However, their results rely on initialization via the method-of-moments estimator from [33] and thus still suffer from the suboptimality inherent in the method-of-moments approach. In fact, the suboptimality of the method-of-moments approach has been acknowledged in many works [33, 30, 29, 32] and closing this gap has remained a well-recognized open problem.

A Warm-up Example: Mean Estimation Problems. We will introduce our estimator to address the limitations of these existing ones and improve error upper bounds. To illustrate our ideas, we begin with a simpler mean estimation problem and show that a local averaging estimator is an optimal solution to the least squares problem. Specifically, we consider the scenario where each client i observes n_i data sample vectors, denoted by $u_{ij} \in \mathbb{R}^d$ for $j \in [n_i]$ such that

$$u_{ij} = \theta_i^\star + \xi_{ij} = B^\star\alpha_i^\star + \xi_{ij}.$$

Here $\xi_{ij} \in \mathbb{R}^d$ is an additive noise for the j-th sample and $\theta_i^\star = B^\star\alpha_i^\star$ is the ground-truth parameter at client i, where $B^\star \in \mathcal{O}^{d\times k}$ with $k \leq d$ and $\alpha_i^\star \in \mathbb{R}^k$. The direct approach to solving mean estimation, given the observed datasets, is to minimize the following non-convex least squares loss,

$$\min_{B\in\mathcal{O}^{d\times k},\{\alpha_i\}} \sum_{i=1}^{M}\sum_{j=1}^{n_i}\|u_{ij} - B\alpha_i\|^2. \tag{3.3}$$

Let $\overline{u}_i = (\sum_{j=1}^{n_i} u_{ij})/n_i$ be the local average at client i, and $\widetilde{B}$ be the top-k eigenvectors of the matrix $\sum_{i=1}^{M} n_i\overline{u}_i\overline{u}_i^\intercal$. The following proposition shows that $\widetilde{B}$ solves this least squares problem.

Proposition 3.1. *After first optimizing over α_i, the problem in (3.3) is equivalent to the following*

$$\max_{B\in\mathcal{O}^{d\times k}} \sum_{i=1}^{M} n_i\overline{u}_i^\intercal BB^\intercal\overline{u}_i. \tag{3.4}$$

In addition, the estimator $\widetilde{B}$ formed by the top-k eigenvectors of the matrix $\sum_{i=1}^{M} n_i\overline{u}_i\overline{u}_i^\intercal$ is an optimal solution to problems in (3.3) and (3.4).

Proposition 3.1 demonstrates that $\widetilde{B}$, utilizing local averaging, is an optimal solution to the least squares problem for mean estimation.

Introducing Our Estimator. Now, we return to tackle the original problem by leveraging the idea of local averaging discussed above. Similar to mean estimation, we consider the non-convex least squares problem for linear regression, with $\theta_i = \Gamma_i^{-1} B\alpha_i$,

$$\min_{\{\theta_i\}} \sum_{i=1}^{M}\sum_{j=1}^{n_i} \big(y_{ij} - x_{ij}^\intercal \theta_i\big)^2 = \min_{B,\{\alpha_i\}} \sum_{i=1}^{M}\sum_{j=1}^{n_i} \big(y_{ij} - x_{ij}^\intercal \Gamma_i^{-1} B\alpha_i\big)^2. \tag{3.5}$$

Let $\widehat{z}_i = (\sum_{j=1}^{n_i} y_{ij}x_{ij})/n_i$ be client i's local average, and $A^\dagger$ be the pseudoinverse of a matrix A.

Proposition 3.2. *After first optimizing over α_i, with $\Lambda_i = \Gamma_i^{-1}B(B^\intercal\Gamma_i^{-1}\widehat{\Gamma}_i\Gamma_i^{-1}B)^\dagger B^\intercal\Gamma_i^{-1}$, the problem in (3.5) is equivalent to $\max_{B\in\mathcal{O}^{d\times k}} \sum_{i=1}^{M} n_i \widehat{z}_i^\intercal \Lambda_i \widehat{z}_i$.*

Unfortunately, unlike (3.4) in Proposition 3.1, the problem in Proposition 3.2 lacks a closed-form solution due to the complex form of Λ_i that also involves B. But if we assume $\widehat{\Gamma}_i \approx \Gamma_i \approx I_d$, then $\Lambda_i \approx BB^\intercal$. Thus, we approximate Λ_i using $BB^\intercal$ in Proposition 3.2 and instead solve

$$\max_{B\in\mathcal{O}^{d\times k}} \sum_{i=1}^{M} n_i \widehat{z}_i^\intercal BB^\intercal \widehat{z}_i. \tag{3.6}$$

The problem (3.6) share the same form as (3.4) and therefore the matrix formed by the top-k eigenvectors of $\sum_{i=1}^{M} n_i\widehat{z}_i\widehat{z}_i^\intercal$ is an optimal solution to (3.6). As a result, it is tempting to estimate $B^\star$ based on the top-k eigenvectors of $\sum_{i=1}^{M} n_i\widehat{z}_i\widehat{z}_i^\intercal$. However, since $\{x_{ij}\}$ follows general sub-gaussian distributions with non-identity covariance Γ_i, without additional assumptions on the covariance matrix Γ_i and the fourth-order moments, it is impossible to construct the column space of $B^\star$ solely by using the eigenvectors of $\sum_{i=1}^{M} n_i\widehat{z}_i\widehat{z}_i^\intercal$.[8]

To resolve this issue, we construct two independent replicas, $\overline{z}_i$ and $\widetilde{z}_i$, in replace of the local average $\widehat{z}_i$. For convenience, suppose that $n_i \geq 2$ is an even number. For $i \in [M]$, let $\overline{z}_i = (2/n_i)\cdot\sum_{j=1}^{n_i/2} y_{ij}x_{ij}$ and $\widetilde{z}_i = (2/n_i)\cdot\sum_{j=n_i/2+1}^{n_i} y_{ij}x_{ij}$ be two independent replicas of local averages at client i. We consider the following matrix,

$$Z = \sum_{i=1}^{M} n_i\overline{z}_i\widetilde{z}_i^\intercal = \sum_{i=1}^{M} n_i\Big(\frac{2}{n_i}\sum_{j=1}^{n_i/2} y_{ij}x_{ij}\Big)\Big(\frac{2}{n_i}\sum_{j=n_i/2+1}^{n_i} y_{ij}x_{ij}^\intercal\Big). \tag{3.7}$$

We define $\widehat{B}$ as the matrix formed by the right (or left) top-k singular vectors of Z. Now, with two independent replicas, it is easy to see that

$$\mathbb{E}Z = \mathbb{E}\Big[\sum_{i=1}^{M} n_i\overline{z}_i\widetilde{z}_i^\intercal\Big] = \sum_{i=1}^{M} n_i\mathbb{E}[\overline{z}_i]\mathbb{E}[\widetilde{z}_i^\intercal] = B^\star\Big(\sum_{i=1}^{M} n_i\alpha_i^\star(\alpha_i^\star)^\intercal\Big)(B^\star)^\intercal,$$

and the column space of $\mathbb{E}Z$ recovers that of $B^\star$; thus, $\widehat{B}$, formed by the singular vectors of Z, provides a good estimate for $B^\star$, ensured by the classic perturbation theory for singular vectors [36]. This highlights the benefits of using two replicas. Similar replica ideas have appeared in other related problems of mixed linear regression [22, 27], but with different motivations and results.

Finally, compared to Z_1 in (3.1), our estimator Z applies local averaging of $x_{ij}y_{ij}$ at each client, which leads to a tighter upper bound by effectively reducing noise. For Z_2 in (3.2), while excluding diagonal terms can be viewed as an alternative to our use of independent replicas, our approach provides significant advantages in privacy-sensitive settings, such as federated learning. To compute our estimator, each client can send only vectors of their local averages $\overline{z}_i$ and $\widetilde{z}_i$, or their variants with added noise, to the server, rather than transmitting any raw data $y_{ij}x_{ij}$. Thus, our estimator effectively prevents the leakage of local data. In addition, since our estimator approximates the least squares solution in (3.5), we will show that it achieves the optimal statistical rate without the need for further refinement via alternating minimization.[9]

[8] Interested readers can find a detailed explanation in Proposition B.1.

[9] When the noise variance σ^2 is vanishing, further refinement via alternating minimization may improve the dependence on σ^2 and thus achieve a smaller estimation error, as shown in [30] for sufficiently fast diminishing σ and [7] for $\sigma = 0$.

3.2 Error Upper Bound

The following theorem establishes the error upper bound of our estimator.

Theorem 3.1. *Suppose that Assumptions 2.1-2.4 hold. For the estimator $\widehat{B}$ obtained in (3.7), with probability at least $1 - O((d+N)^{-10})$,*

$$\|\sin\Theta(\widehat{B}, B^\star)\| = O\bigg(\bigg(\sqrt{\frac{d\lambda_1}{N\lambda_k^2}} + \sqrt{\frac{Md}{N^2\lambda_k^2}}\bigg) \cdot \log^3(d+N)\bigg).$$

Here the condition number λ_1/λ_k and the smallest eigenvalue λ_k appear in the numerator and denominator of the rate, respectively. This aligns with our intuition that a larger λ_1/λ_k or a smaller λ_k causes more difficulty in estimating $B^\star$, as the client diversity matrix $\sum_{i=1}^M n_i \alpha_i^\star (\alpha_i^\star)^\intercal / N$ lacks information in certain directions.

Our bound improves over the previously best-known error rates in the literature. Specifically, by further assuming $x_{ij} \sim N(0, I_d)$, [33] shows that the method-of-moments estimator given in (3.1) achieves estimation error rate $\widetilde{O}(\sqrt{d(\sum_{r=1}^k \lambda_r)/(N\lambda_k^2)})$. However, the analysis therein crucially relies on the isotropy property of standard Gaussian vectors (see e.g. the proof of Lemma 4 in [33]). If instead x_{ij}'s were sub-gaussian, the error bound would become $\widetilde{O}(\sqrt{dk\lambda_1/(N\lambda_k^2)})$. The subsequent work [10] shows that a different spectral estimator given in (3.2) achieves a smaller error of $\widetilde{O}(\sqrt{d/(N\lambda_k^2)})$ when x_{ij}'s are sub-gaussian. In comparison, our error bound further improves by a factor of $\sqrt{\min\{\lambda_1, M/N\}}$. Our improvement is particularly significant when the condition number of the client diversity matrix satisfies $\lambda_1/\lambda_k = \Theta(1)$, as shown in the following corollary.

Corollary 3.1. *Suppose Assumptions 2.1-2.4 hold and $\lambda_1 = \Theta(\lambda_k) = \Theta(1/k)$. For the estimator $\widehat{B}$ obtained in (3.7), with probability at least $1 - O((d+N)^{-10})$,*

$$\|\sin\Theta(\widehat{B}, B^\star)\| = O\bigg(\bigg(\sqrt{\frac{dk}{N}} + \sqrt{\frac{Mdk^2}{N^2}}\bigg) \cdot \log^3(d+N)\bigg). \tag{3.8}$$

This rate improves the results given by [33, 10], which are $\widetilde{O}(\sqrt{dk^2/N})$. More strikingly, our rate is order-wise optimal, matching up to a polylogarithmic factor the minimax lower bound shown in the next section. This resolves the challenging open problem of characterizing the optimal estimation error rate, and our estimator is the first in the literature to achieve this optimal rate.

To further illustrate our results, we plot a phase diagram in Figure 1, where the regions indicate whether a consistent estimation of $B^\star$ with vanishing estimation error is possible or not.

Remark 3.1. In contrast to the previous findings, our optimal rate shows that, at a system level, collaboration always reduces overall sample complexity compared to independent client learning. Specifically, collaboration requires only $N = \Theta(\max\{dk, \sqrt{Mdk^2}\})$ data points to learn $B^\star$ with a vanishing error. With a shared estimator of $B^\star$, the M clients can then learn their $\{\alpha_i^\star\} \subset \mathbb{R}^k$, with an additional sample complexity of $\Theta(Mk)$. In contrast, independently learning all parameters $\{\theta_i^\star\} \subset \mathbb{R}^d$ from scratch requires $N = \Theta(Md)$, where $Md \geq \max\{dk, \sqrt{Mdk^2}, Mk\}$ since $k \leq \min\{d, M\}$. Thus, collaboration always reduces sample complexity compared to independent learning. However, the previous best-known results [33] show that collaboration requires $N = \Theta(dk^2)$ data points, where dk^2 is smaller than Md only when $M \geq k^2$.

At an individual level, collaboration is no longer always beneficial. Our optimal rate provides a more precise characterization of when collaboration benefits a new client, as discussed in Section 5.

4 Minimax Lower Bound

This section establishes an information-theoretic lower bound that matches the error upper bound achieved by our estimator up to a polylogarithmic factor in the well-conditioned cases. For fixed M and N, we use the eigenvalues λ_1 and λ_k of the client diversity matrix $\sum_{i=1}^M n_i \alpha_i^\star (\alpha_i^\star)^\intercal / N$ to capture the complexity of the estimation problem. In particular, we analyze the minimax estimation

error against the worst possible choice of the model parameters B, $\{\alpha_i\}_{i=1}^M$, and $\{n_i\}_{i=1}^M$ from a parameter space. The problem of estimating B can then be represented as the Markov chain,

$$\big(B, \{\alpha_i\}_{i=1}^M, \{n_i\}_{i=1}^M\big) \to \{\{(x_{ij}, y_{ij})\}_{j=1}^{n_i}\}_{i=1}^M \to \widehat{B}.$$

Here the data volumes $\{n_i\}_{i=1}^M$, satisfying $\sum_{i=1}^M n_i = N$, can be observed from the data and hence are nuisance parameters.

We now define the parameter space. We take $B \in \mathcal{O}^{d\times k}$ to be any $d \times k$ orthogonal matrix. Let $\alpha = (\alpha_1, \cdots, \alpha_M)$ be the matrix whose columns are the client-specific parameters α_i and $\vec{n} = (n_1, \cdots, n_M)^\intercal$ be the vector with entries n_i. For any $\lambda_1 \geq \lambda_k > 0$, we define $\Psi_{\lambda_1,\lambda_k}$ as the parameter space for all α and $\vec{n}$ that satisfy Assumption 2.4,

$$\Psi_{\lambda_1,\lambda_k} = \Big\{(\alpha, \vec{n}) \in \mathbb{R}_+^{k\times M} \times \mathbb{Z}_+^M : \|\alpha_i\| = O(1)\ \forall i \in [M],\ \sum_{i=1}^M n_i = N,$$
$$\Omega(\lambda_k) I_k \preceq \frac{1}{N}\sum_{i=1}^M n_i \alpha_i \alpha_i^\intercal \preceq O(\lambda_1) I_k\Big\}.$$

We only consider scenarios where $\lambda_k > 0$, as it is otherwise impossible to fully reconstruct $B^\star$. When $\lambda_k = 0$, $\{\alpha_i^\star\}$ spans only an r-dimensional subspace of $\mathbb{R}^k$ with $r < k$. As a result, the parameters $\{\theta_i^\star\}$ and thus the data $\{x_{ij}\}$ will only contain information about an r-dimensional subspace of $B^\star$'s column space. In this case, the remaining $k - r$ columns of $B^\star$ can be any vectors from the $(d - r)$-dimensional complementary space, making it impossible to estimate $B^\star$ accurately. This also implies $M \geq k$. Recall that Assumption 2.4 yields $k\lambda_k = O(1)$ and $\lambda_1 = O(1)$; otherwise, the parameter space is empty. Henceforth, we assume $\lambda_k > 0$, $M \geq k$, $k\lambda_k = O(1)$, and $\lambda_1 = O(1)$.

The following theorem presents the minimax error lower bound. Here $\wedge$ is a shorthand notation for the minimum operation.

Theorem 4.1. *Consider a system with M clients and N data points in total. Assume $x_{ij} \sim N(0, I_d)$ and $\xi_{ij} \sim N(0, 1)$ independently for $i \in [M]$ and $j \in [n]$, and Assumptions 2.3 and 2.4 hold. When $k = \Omega(\log M)$, $d \geq (1 + \rho_1)k$, and $M \geq (1 + \rho_2)k$ for constants $\rho_1, \rho_2 > 0$, we have*

$$\inf_{\widehat{B}\in\mathcal{O}^{d\times k}} \sup_{B\in\mathcal{O}^{d\times k}} \sup_{(\alpha,\vec{n})\in\Psi_{\lambda_1,\lambda_k}} \mathbb{E}\Big[\big\|\sin\Theta(\widehat{B}, B)\big\|\Big] = \Omega\bigg(\bigg(\sqrt{\frac{d}{N\lambda_k}} + \sqrt{\frac{Md}{N^2\lambda_k^2}}\bigg) \wedge 1\bigg).$$

Theorem 4.1 establishes an error lower bound, which improves the state-of-the-art result from [33], which is of order $\Omega(\sqrt{1/(N\lambda_k)} + \sqrt{dk/N})$. Our lower bound matches the upper bound presented in Theorem 3.1, differing only by a condition number $\sqrt{\lambda_1/\lambda_k}$ in the first term and a logarithmic factor. Thus, in the well-conditioned case when $\lambda_1/\lambda_k = \Theta(1)$, we have the following corollary.

Corollary 4.1. *Under the conditions in Theorem 4.1, when $\lambda_1 = \Theta(\lambda_k) = \Theta(1/k)$, we have*

$$\inf_{\widehat{B}\in\mathcal{O}^{d\times k}} \sup_{B\in\mathcal{O}^{d\times k}} \sup_{(\alpha,\vec{n})\in\Psi_{\lambda_1,\lambda_k}} \mathbb{E}\Big[\big\|\sin\Theta(\widehat{B}, B)\big\|\Big] = \Omega\bigg(\bigg(\sqrt{\frac{dk}{N}} + \sqrt{\frac{Mdk^2}{N^2}}\bigg) \wedge 1\bigg).$$

Corollary 4.1 establishes the error lower bound for well-conditioned cases and improves that from [33] of order $\Omega(\sqrt{dk/N})$. This result matches the upper bound in Corollary 3.1 up to a polylogarithmic factor, thereby determining the optimal statistical rate.

5 Applications

Having identified the statistical rate for estimating $B^\star$, we now apply this result to learn the model parameters for a newly joined client or an unseen private task and provide a more precise characterization of when collaboration benefits a new client.

5.1 Transferring Representations to New Clients

We consider a new client $M + 1$, who observes n_{M+1} data points, $\{(x_{M+1,j}, y_{M+1,j})\}_{j=1}^{n_{M+1}}$, generated from the model in (1.1) with the ground-truth parameter $\theta_{M+1}^\star$. Suppose that Assumptions

2.1 and 2.2 hold with $\Gamma_{M+1} = I_d$. We then assume the existence of $\alpha^\star_{M+1} \in \mathbb{R}^k$ such that $\theta^\star_{M+1} = B^\star \alpha^\star_{M+1}$ and $\|\alpha^\star_{M+1}\| = O(1)$. Our goal is to learn $\theta^\star_{M+1}$.

If we substitute an estimator $\widehat{B}$, learned from clients 1 to M, in place of the shared $B^\star$, the problem reduces to learning $\widehat{\alpha}_{M+1}$ as follows,

$$\widehat{\alpha}_{M+1} = \underset{\widehat{\alpha}}{\operatorname{argmin}} \sum_{j=1}^{n_{M+1}} \|x_{M+1,j}^\intercal \widehat{B}\widehat{\alpha} - y_{M+1,j}\|^2. \tag{5.1}$$

There is an explicit solution for $\widehat{\alpha}_{M+1}$ in (5.1). Notably, [33, Theorem 4] provides an error upper bound for $\widehat{B}\widehat{\alpha}_{M+1}$ when $\|\sin\Theta(\widehat{B}, B^\star)\| \le \delta^2$ for any $\delta > 0$. Recall that Corollary 3.1 establishes an error bound for our estimator $\widehat{B}$ in well-conditioned cases. Thus, as a direct corollary using our estimator and its error bound, we derive the following result.

Corollary 5.1 (Transfer learning)**.** *Suppose that Assumptions 2.1-2.4 hold and $\lambda_1 = \Theta(\lambda_k) = \Theta(1/k)$. For $\widehat{B}$ given by (3.7) and then $\widehat{\alpha}_{M+1}$ given by (5.1), with high probability,*

$$\|\widehat{B}\widehat{\alpha}_{M+1} - B^\star\alpha^\star_{M+1}\|^2 = \widetilde{O}\left(\frac{dk}{N} + \frac{Mdk^2}{N^2} + \frac{k}{n_{M+1}}\right).$$

Corollary 5.1 decomposes the estimation error into two parts, where the first part $\widetilde{O}(dk/N + Mdk^2/N^2)$ captures the error for estimating $B^\star$, and the second part $\widetilde{O}(k/n_{M+1})$ evaluates the error for estimating $\alpha^\star_{M+1}$ given $\widehat{B}$.

Compared to the previous literature [33], our optimal rate provides a more precise characterization of when collaboration benefits the new client. If client $M+1$ estimates its parameter $\theta^\star_{M+1} \in \mathbb{R}^d$ from scratch based on the local data, the resulting error rate will be $\widetilde{O}(d/n_{M+1})$. Therefore, it is advantageous to first learn the shared representation when $dk/N \ll d/n_{M+1}$ and $Mdk^2/N^2 \ll d/n_{M+1}$. Such conditions are satisfied when $n_{M+1} \ll \min\{N/k, N^2/(Mk^2)\}$. Conversely, if $n_{M+1} \gg \min\{N/k, N^2/(Mk^2)\}$, then the collaboration is unhelpful and the new client would be better off learning individually.

5.2 Private Fine-tuning for New Clients

In addition, [29] studies a differentially private variant of learning $\alpha^\star_{M+1}$ in the same setting further with $x_{ij} \sim N(0, I_d)$. We present an additional corollary under (ϵ, δ)-differential privacy (see the formal definition in [11, 12]), building upon [29, Theorem 5.4] and derived using our estimator.

Corollary 5.2 (Private transfer learning)**.** *Suppose that Assumptions 2.1-2.4 hold, $\lambda_1 = \Theta(\lambda_k) = \Theta(1/k)$, and $x_{ij} \sim N(0, I_d)$. Given $\widehat{B}$ obtained in (3.7), there exists an (ε, δ)-differentially private estimator $\widehat{\alpha}^{\varepsilon,\delta}_{M+1}$ such that, with high probability,*

$$\left\|\widehat{B}\widehat{\alpha}^{\varepsilon,\delta}_{M+1} - B^\star\alpha^\star_{M+1}\right\|^2 = \widetilde{O}\left(\frac{dk}{N} + \frac{Mdk^2}{N^2} + \frac{k}{n_{M+1}} + \frac{k^2\log(1/\delta)}{n^2_{M+1}\varepsilon^2}\right).$$

For comparison, if the client privately estimates its d-dimensional parameter $\theta^\star_{M+1}$ from scratch, the resulting tight error rate is $\widetilde{O}(d/n_{M+1} + d^2\log(1/\delta)/(n^2_{M+1}\varepsilon^2))$ [34, 3]. Thus, when the estimation error of $\widehat{B}$, $\widetilde{O}(dk/N + Mdk^2/N^2)$, is smaller, learning $\widehat{B}$ first will effectively reduce its error rates.

6 Discussion and Future Work

In this work, we introduce a spectral estimator with local averaging and analyze its performance with an improved error upper bound. In addition, we sharpen the existing minimax lower bound. Our results together settle the optimal statistical rate in well-conditioned cases. Our optimal rate shows that collaboration among clients always reduces overall sample complexity compared to independent local learning and further quantifies the benefits of transfer learning and private fine-tuning for new clients or tasks. Furthermore, the optimal rate reveals two distinct phases, where a statistical penalty arises for collaborative learning, especially with many clients or relatively small local datasets.

An open problem is whether we can eliminate the dependency on condition numbers and achieve the optimal rate in ill-conditioned cases. Moreover, when the noise variance is rapidly diminishing, can we obtain the optimal error rates? Finally, is it possible to extend our analysis to generalized linear models and non-linear regression models while maintaining similar guarantees of optimal rates?

Acknowledgement

J. Xu is deeply grateful to Laurent Massoulié for highlighting the gaps between the existing upper and lower bounds in estimating the shared linear representation $B^\star$, and for the insightful discussions on how to address these gaps. X. Niu and J. Xu are supported in part by an NSF CAREER award CCF-2144593. P. Yang is supported in part by an NSFC grant 12101353.

References

[1] Rie Kubota Ando, Tong Zhang, and Peter Bartlett. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of machine learning research*, 6(11), 2005.

[2] Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

[3] T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

[4] T Tony Cai and Hongji Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1), 2021.

[5] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

[6] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.

[7] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pages 2089–2099. PMLR, 2021.

[8] Simon Shaolei Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021.

[9] Yaqi Duan and Kaizheng Wang. Adaptive and robust multi-task learning. *The Annals of Statistics*, 51(5):2015–2039, 2023.

[10] John C Duchi, Vitaly Feldman, Lunjia Hu, and Kunal Talwar. Subspace recovery from heterogeneous data with non-isotropic noise. *Advances in Neural Information Processing Systems*, 35:5854–5866, 2022.

[11] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25*, pages 486–503. Springer, 2006.

[12] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

[13] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in neural information processing systems*, 33:3557–3568, 2020.

[14] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.

[15] Avishek Ghosh, Justin Hong, Dong Yin, and Kannan Ramchandran. Robust federated learning in a heterogeneous environment. *arXiv preprint arXiv:1906.06629*, 2019.

[16] Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[17] Steve Hanneke and Samory Kpotufe. A no-free-lunch theorem for multitask learning. *The Annals of Statistics*, 50(6):3119–3143, 2022.

[18] Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 178–191, 2016.

[19] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[20] Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. Adaptive gradient-based meta-learning methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[21] Parker Knight and Rui Duan. Multi-task learning with summary statistics. *Advances in Neural Information Processing Systems*, 36, 2024.

[22] Weihao Kong, Raghav Somani, Zhao Song, Sham Kakade, and Sewoong Oh. Meta-learning for mixed linear regression. In *International Conference on Machine Learning*, pages 5394–5404. PMLR, 2020.

[23] Yoonho Lee and Seungjin Choi. Gradient-based meta-learning with learned layerwise metric and subspace. In *International Conference on Machine Learning*, pages 2927–2936. PMLR, 2018.

[24] Sai Li, Tianxi Cai, and Rui Duan. Targeting underrepresented populations in precision medicine: A federated transfer learning approach. *The Annals of Applied Statistics*, 17(4):2970–2992, 2023.

[25] Youming Liu and Xinyu Qi. Optimal estimation for lower bound of the packing number. *Statistics & Probability Letters*, 186:109487, 2022.

[26] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

[27] Lili Su, Jiaming Xu, and Pengkun Yang. Global convergence of federated learning for mixed regression. *IEEE Transactions on Information Theory*, 2024.

[28] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 403–412, 2019.

[29] Pratiksha Thaker, Amrith Setlur, Zhiwei Steven Wu, and Virginia Smith. On the benefits of public representations for private transfer learning under distribution shift. *arXiv preprint arXiv:2312.15551*, 2023.

[30] Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Statistically and computationally efficient linear meta-representation learning. *Advances in Neural Information Processing Systems*, 34:18487–18500, 2021.

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

[32] Ye Tian, Yuqi Gu, and Yang Feng. Learning from similar linear representations: Adaptivity, minimaxity, and robustness. *arXiv preprint arXiv:2303.17765*, 2023.

[33] Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR, 2021.

[34] Prateek Varshney, Abhradeep Thakurta, and Prateek Jain. (nearly) optimal private linear regression for sub-gaussian data via adaptive clipping. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, pages 1126–1166. PMLR, 02–05 Jul 2022.

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

[36] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.

[37] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2022.

[38] Thomas TCK Zhang, Leonardo Felipe Toso, James Anderson, and Nikolai Matni. Sample-efficient linear representation learning from non-iid non-isotropic data. In *The Twelfth International Conference on Learning Representations*, 2024.

[39] Zhuangdi Zhu, Kaixiang Lin, Anil K Jain, and Jiayu Zhou. Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

A Further Related Work

Our work contributes to the growing literature studying collaborative learning of shared feature representations, which dates back to [5, 31, 2, 1, 26]. This problem has a broad range of applications, including federated learning, multi-task transfer learning, meta-learning, and private fine-tuning with public knowledge [33, 8, 7, 29, 9].

Most relevant recent works, including [33, 8, 7, 30, 10], focus on the linear model described in (1.1). However, none of them identifies the optimal statistical rate, even in well-conditioned cases. In particular, [33] introduces a method-of-moments estimator for standard Gaussian covariates but obtains upper and lower bounds suboptimal to the subspace dimension k. Concurrently, [8] provides purely statistical guarantees by directly analyzing the optimal solution to the nonconvex empirical risk minimization, yet still obtains a suboptimal error upper bound. In addition, [10] presents a spectral estimator that achieves an optimal rate for simpler mean estimation problems, but its error upper bound for linear regression settings remains suboptimal. Furthermore, [7, 30] study alternating minimization initialized at the method-of-moments estimator, which reduces the error rates when the noise level of ξ_{ij} rapidly diminishing or zero but the rates remain suboptimal with respect to the subspace dimension k.

Several studies have explored variants of the model in (1.1). For example, [29] studies differentially private fine-tuning given the method-of-moments estimator. Moreover, [9] studies adaptive and robust multi-task learning, where, in a specific low-rank scenario, they assume $\theta_i^\star = B^\star \alpha_i^\star + v_i^\star$ with a bounded offset term $v_i^\star$. In addition, [32] considers the case when clients share similar representations such that $\theta_i^\star = B_i^\star \alpha_i^\star$, with the subspaces $B_i^\star$ constrained to certain angles, while also allowing outliers. However, the statistical rates in [29] and [9] are suboptimal with respect to k, and both [9] and [32] impose much stricter assumptions that $N/M = \Omega(d)$.

Other related works including gradient-based meta-learning [23, 20, 13], non-parametric transfer learning [16, 4], and the hardness of multi-task learning [17]. It is also worth noting that the shared linear representation model in (1.1) provides an effective approach for addressing data heterogeneity with concept shift in federated learning [7], which includes the clustered federated learning framework [15, 14, 27] as a special case.

B Proof of the Upper Bound

In this section, we prove the results in Section 3.

B.1 Proofs of Results in Section 3.1

We prove Proposition 3.1 by first fixing B and optimizing for α_i.

Proof of Proposition 3.1. We first partially optimize for α_i given a fixed $B \in \mathcal{O}^{d\times k}$. By taking partial derivatives and solving $\sum_{j=1}^{n_i} B^\intercal(u_{ij} - B\widehat{\alpha}_i) = 0$ with $B^\intercal B = I_k$, we have

$$\widehat{\alpha}_i = B^\intercal \overline{u}_i.$$

Substituting the optimal $\widehat{\alpha}_i$ into the original problem, this leaves us to find $B \in \mathcal{O}^{d\times k}$ to minimize:

$$\begin{aligned}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\left\|u_{ij} - BB^\intercal\overline{u}_i\right\|^2 &= \sum_{i=1}^{M}\sum_{j=1}^{n_i}\left\|u_{ij} - \overline{u}_i + \overline{u}_i - BB^\intercal\overline{u}_i\right\|^2\\ &= \sum_{i=1}^{M}\sum_{j=1}^{n_i}\left\|u_{ij} - \overline{u}_i\right\|^2 + \sum_{i=1}^{M} n_i\left\|\overline{u}_i - BB^\intercal\overline{u}_i\right\|^2\\ &= \sum_{i=1}^{M}\sum_{j=1}^{n_i}\left\|u_{ij} - \overline{u}_i\right\|^2 + \sum_{i=1}^{M} n_i\left\|\overline{u}_i\right\|^2 - \sum_{i=1}^{M} n_i(\overline{u}_i)^\intercal BB^\intercal\overline{u}_i,\end{aligned}$$

where the second equality holds since $\sum_{j=1}^{n_i}(u_{ij} - \overline{u}_i)^\intercal(\overline{u}_i - BB^\intercal\overline{u}_i) = 0$ for $i \in [M]$ and the last equality holds due to $B^\intercal B = I_k$. Thus, the least-squares problem is equivalent to the following one,

$$\max_{B\in\mathcal{O}^{d\times k}} \sum_{i=1}^{M} n_i\overline{u}_i^\intercal BB^\intercal\overline{u}_i.$$

In addition, we have

$$\sum_{i=1}^{M} n_i\overline{u}_i^\intercal BB^\intercal\overline{u}_i = \sum_{i=1}^{M} n_i\mathrm{Tr}\big(B^\intercal\overline{u}_i\overline{u}_i^\intercal B\big) = \mathrm{Tr}\Big(B^\intercal\Big(\sum_{i=1}^{M} n_i\overline{u}_i\overline{u}_i^\intercal\Big)B\Big).$$

We define $Z = \sum_{i=1}^{M} n_i\overline{u}_i\overline{u}_i^\intercal$. Solving the PCA problem $\max_{B\in\mathbb{R}^{d\times k}}\mathrm{Tr}(BZB)$ s.t. $B^\intercal B = I_k$, we obtain the optimal $\widehat{B}$ as the top-k eigenvectors of Z. □

Similarly, we prove Proposition 3.2.

Proof of Proposition 3.2. We first partially optimize for α_i for a fixed $B \in \mathcal{O}^{d\times k}$. Let $\widehat{\Gamma}_i = (\sum_{j=1}^{n_i} x_{ij}x_{ij}^\intercal)/n_i$. By taking partial derivatives and solving $\sum_{j=1}^{n_i} B^\intercal\Gamma_i^{-1}x_{ij}(y_{ij} - x_{ij}^\intercal\Gamma_i^{-1}B\alpha_i) = 0$ with $B^\intercal B = I_k$, we have

$$\widehat{\alpha}_i = \big(B^\intercal\Gamma_i^{-1}\widehat{\Gamma}_i\Gamma_i^{-1}B\big)^\dagger B^\intercal\Gamma_i^{-1}\widehat{z}_i.$$

Let $\Lambda_i = \Gamma_i^{-1}B\big(B^\intercal\Gamma_i^{-1}\widehat{\Gamma}_i\Gamma_i^{-1}B\big)^\dagger B^\intercal\Gamma_i^{-1}$. Substituting the optimal $\widehat{\alpha}_i$ into the original problem, this leaves us to find $B \in \mathcal{O}^{d\times k}$ to minimize:

$$\begin{aligned}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\big(y_{ij} - x_{ij}^\intercal\Gamma_i^{-1}B\widehat{\alpha}_i\big)^2 &= \sum_{i=1}^{M}\sum_{j=1}^{n_i}\big(y_{ij} - x_{ij}^\intercal\Lambda_i\widehat{z}_i\big)^2\\ &= \sum_{i=1}^{M}\Big(\sum_{j=1}^{n_i} y_{ij}^2 - 2n_i\widehat{z}_i^\intercal\Lambda_i\widehat{z}_i + n_i\widehat{z}_i^\intercal\Lambda_i\widehat{\Gamma}_i\Lambda_i\widehat{z}_i\Big)\\ &= \sum_{i=1}^{M}\Big(\sum_{j=1}^{n_i} y_{ij}^2 - n_i\widehat{z}_i^\intercal\Lambda_i\widehat{z}_i\Big),\end{aligned}$$

where the last equality holds since it is easy to compute that $\Lambda_i \widehat{\Gamma}_i \Lambda_i = \Lambda_i$. Thus, the least squares problem in (3.5) is equivalent to the following one,

$$\max_{B \in \mathcal{O}^{d\times k}} \sum_{i=1}^{M} n_i \widehat{z}_i^{\intercal} \Lambda_i \widehat{z}_i.$$

□

We now present the following proposition with the formal form of $\mathbb{E}[\sum_{i=1}^{M} n_i \widehat{z}_i \widehat{z}_i^{\intercal}]$.

Proposition B.1. *Under Assumptions 2.1-2.3, the matrix $\sum_{i=1}^{M} n_i \widehat{z}_i \widehat{z}_i^{\intercal}$ satisfies*

$$\mathbb{E}\Big[\sum_{i=1}^{M} n_i \widehat{z}_i \widehat{z}_i^{\intercal}\Big] = B^{\star}\Big(\sum_{i=1}^{M}(n_i - 1)\alpha_i^{\star}(\alpha_i^{\star})^{\intercal}\Big)(B^{\star})^{\intercal} + \sum_{i=1}^{M}\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{E}[x_{ij}^{\intercal}\theta_i^{\star}(\theta_i^{\star})^{\intercal}x_{ij}x_{ij}x_{ij}^{\intercal}] + \sum_{i=1}^{M}\mathbb{E}[\xi_{ij}^2]\Gamma_i.$$

Proof of Proposition B.1. We begin with $\mathbb{E}[\sum_{i=1}^{M} n_i \widehat{z}_i \widehat{z}_i^{\intercal}]$. By the definition of $\widehat{z}_i$, we have

$$\begin{aligned}\mathbb{E}\Big[\sum_{i=1}^{M} n_i \widehat{z}_i \widehat{z}_i^{\intercal}\Big] &= \mathbb{E}\Big[\sum_{i=1}^{M} n_i \Big(\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}x_{ij}\Big)\Big(\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}x_{ij}^{\intercal}\Big)\Big] \\ &= \sum_{i=1}^{M}\frac{1}{n_i}\Big(\sum_{j_1\neq j_2}\mathbb{E}[y_{ij_1}x_{ij_1}]\mathbb{E}[y_{ij_2}x_{ij_2}^{\intercal}] + \sum_{j=1}^{n_i}\mathbb{E}[y_{ij}^2 x_{ij}x_{ij}^{\intercal}]\Big). \qquad \text{(B.1)}\end{aligned}$$

For the second term above, since ξ_{ij} and x_{ij} are independent, we have

$$\begin{aligned}\sum_{i=1}^{M}\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{E}[y_{ij}^2 x_{ij}x_{ij}^{\intercal}] &= \sum_{i=1}^{M}\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{E}[(x_{ij}^{\intercal}\theta_i^{\star} + \xi_{ij})^2 x_{ij}x_{ij}^{\intercal}] \\ &= \sum_{i=1}^{M}\frac{1}{n_i}\sum_{j=1}^{n_i}\Big(\mathbb{E}[x_{ij}^{\intercal}\theta_i^{\star}(\theta_i^{\star})^{\intercal}x_{ij}x_{ij}x_{ij}^{\intercal}] + \mathbb{E}[\xi_{ij}^2]\mathbb{E}[x_{ij}x_{ij}^{\intercal}]\Big) \\ &= \sum_{i=1}^{M}\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{E}[x_{ij}^{\intercal}\theta_i^{\star}(\theta_i^{\star})^{\intercal}x_{ij}x_{ij}x_{ij}^{\intercal}] + \sum_{i=1}^{M}\mathbb{E}[\xi_{ij}^2]\Gamma_i.\end{aligned}$$

In addition, we have $\mathbb{E}[y_{ij_1}x_{ij_1}] = \mathbb{E}[x_{ij_1}x_{ij_1}^{\intercal}]\theta_i^{\star} = B^{\star}\alpha_i^{\star}$. We conclude the proof by substituting these results into (B.1). □

B.2 Proof of Theorem 3.1

In this section, we prove the error upper bound of our estimator. We first introduce Wedin's $\sin\Theta$ theorem, which generalizes Davis–Kahan theorem to singular subspaces.

Theorem B.1 (Wedin's $\sin\Theta$ theorem [36])**.** *Consider two matrices $M^{\star}$ and $M = M^{\star} + E$ in $\mathbb{R}^{d\times d}$, with singular values $\sigma_1^{\star} \geq \cdots \geq \sigma_d^{\star}$ and $\sigma_1 \geq \cdots \geq \sigma_d$ respectively. Let $u_i^{\star}$ (resp. u_i) and $v_i^{\star}$ (resp. v_i) be the left and right singular vectors associated with $\sigma_i^{\star}$ (resp. σ_i). For $r \leq d$, we define matrices $\Sigma^{\star} = \mathsf{diag}\,(\sigma_1^{\star}, \cdots, \sigma_r^{\star})$, $\Sigma_{\perp}^{\star} = \mathsf{diag}\,(\sigma_{r+1}^{\star}, \cdots, \sigma_d^{\star})$, $U^{\star} = (u_1^{\star}, \cdots, u_r^{\star}) \in \mathbb{R}^{d\times r}$, $U_{\perp}^{\star} = (u_{r+1}^{\star}, \cdots, u_d^{\star}) \in \mathbb{R}^{d\times(d-r)}$, $V^{\star} = (v_1^{\star}, \cdots, v_r^{\star}) \in \mathbb{R}^{d\times r}$, and $V_{\perp}^{\star} = (v_{r+1}^{\star}, \cdots, v_d^{\star}) \in \mathbb{R}^{d\times(d-r)}$. Similar matrices Σ, $\Sigma_{\perp}$, U, $U_{\perp}$, V, $V_{\perp}$ are defined for M. Then their singular value decompositions are given by*

$$M^{\star} = (U^{\star} \quad U_{\perp}^{\star})\begin{pmatrix}\Sigma^{\star} & 0\\ 0 & \Sigma_{\perp}^{\star}\end{pmatrix}\begin{pmatrix}(V^{\star})^{\intercal}\\ (V^{\star})_{\perp}^{\intercal}\end{pmatrix}, \quad M = (U \quad U_{\perp})\begin{pmatrix}\Sigma & 0\\ 0 & \Sigma_{\perp}\end{pmatrix}\begin{pmatrix}V^{\intercal}\\ V_{\perp}^{\intercal}\end{pmatrix},$$

If $\|E\| < (\sigma_r^{\star} - \sigma_{r+1}^{\star})/2$, we have

$$\max\big\{\|\sin\Theta(U, U^{\star})\|, \|\sin\Theta(V, V^{\star})\|\big\} \leq \frac{2\|E\|}{\sigma_r^{\star} - \sigma_{r+1}^{\star}}.$$

We now prove Theorem 3.1 using Wedin's theorem.

Proof of Theorem 3.1. We first compute $\mathbb{E}Z$. The independence of x_{ij} and ξ_{ij} yields that

$$\mathbb{E}\overline{z}_i = \frac{2}{n_i}\sum_{i=1}^{n_i/2}\mathbb{E}[y_{ij}x_{ij}] = \mathbb{E}[x_{i1}x_{i1}^\intercal]\theta_i^\star = B^\star\alpha_i^\star.$$

Similarly, we have $\mathbb{E}\widetilde{z}_i = B^\star\alpha_i^\star$ and thus

$$\mathbb{E}Z = \sum_{i=1}^{M} n_i(\mathbb{E}\overline{z}_i)(\mathbb{E}\widetilde{z}_i^\intercal) = B^\star\Big(\sum_{i=1}^{M} n_i\alpha_i^\star(\alpha_i^\star)^\intercal\Big)(B^\star)^\intercal.$$

Let $\overline{\Gamma}_i = 2/n_i\cdot\sum_{j=1}^{n_i/2}x_{ij}x_{ij}^\intercal$, $\widetilde{\Gamma}_i = 2/n_i\cdot\sum_{j=n_i/2+1}^{n_i}x_{ij}x_{ij}^\intercal$, $\overline{E}_i = 2/n_i\cdot\sum_{j=1}^{n_i/2}\xi_{ij}x_{ij}$, and $\widetilde{E}_i = 2/n_i\cdot\sum_{j=n_i/2+1}^{n_i}\xi_{ij}x_{ij}$. Then, by applying the model (1.1), we have $\overline{z}_i = \overline{\Gamma}_i\theta_i^\star + \overline{E}_i$ and $\widetilde{z}_i = \widetilde{\Gamma}_i\theta_i^\star + \widetilde{E}_i$. By substituting $\overline{z}_i = \overline{\Gamma}_i\theta_i^\star + \overline{E}_i$ and $\widetilde{z}_i = \widetilde{\Gamma}_i\theta_i^\star + \widetilde{E}_i$ into the estimator Z, we have

$$\begin{aligned}
Z - \mathbb{E}Z &= \sum_{i=1}^{M} n_i\overline{z}_i\widetilde{z}_i^\intercal - B^\star\Big(\sum_{i=1}^{M} n_i\alpha_i^\star(\alpha_i^\star)^\intercal\Big)(B^\star)^\intercal \\
&= \sum_{i=1}^{M} n_i(\overline{\Gamma}_i\theta_i + \overline{E}_i)(\widetilde{\Gamma}_i\theta_i + \widetilde{E}_i)^\intercal - B^\star\Big(\sum_{i=1}^{M} n_i\alpha_i^\star(\alpha_i^\star)^\intercal\Big)(B^\star)^\intercal \\
&= \sum_{i=1}^{M} n_i\overline{\Gamma}_i\theta_i\theta_i^\intercal\widetilde{\Gamma}_i^\intercal - B^\star\Big(\sum_{i=1}^{M} n_i\alpha_i^\star(\alpha_i^\star)^\intercal\Big)(B^\star)^\intercal + \sum_{i=1}^{M} n_i\overline{\Gamma}_i\theta_i\widetilde{E}_i^\intercal + \sum_{i=1}^{M} n_i\overline{E}_i\big(\widetilde{\Gamma}_i\theta_i\big)^\intercal \\
&\qquad + \sum_{i=1}^{M} n_i\overline{E}_i\widetilde{E}_i^\intercal.
\end{aligned}$$

Lemmas B.5, B.6, and B.7 will bound the fluctuations of each term in spectral norm using random matrix tools. Thus, by substituting these results, we have, with probability at least $1 - O((d+N)^{-10})$,

$$\begin{aligned}
\|Z - \mathbb{E}Z\| &\le \Big\|\sum_{i=1}^{M} n_i\overline{\Gamma}_i\theta_i\theta_i^\intercal\widetilde{\Gamma}_i^\intercal - B^\star\Big(\sum_{i=1}^{M} n_i\alpha_i^\star(\alpha_i^\star)^\intercal\Big)(B^\star)^\intercal\Big\| + \Big\|\sum_{i=1}^{M} n_i\overline{\Gamma}_i\theta_i\widetilde{E}_i^\intercal\Big\| \\
&\qquad + \Big\|\sum_{i=1}^{M} n_i\overline{E}_i\big(\widetilde{\Gamma}_i\theta_i\big)^\intercal\Big\| + \Big\|\sum_{i=1}^{M} n_i\overline{E}_i\widetilde{E}_i^\intercal\Big\| \qquad \text{(B.2)} \\
&= O\big(\big(\sqrt{Md} + \sqrt{Nd\lambda_1} + d\big)\cdot\log^3(d+N)\big).
\end{aligned}$$

Finally, we apply Wedin's $\sin\Theta$ theorem, noting $\mathbb{E}Z = B^\star\Big(\sum_{i=1}^{M} n_i\alpha_i^\star(\alpha_i^\star)^\intercal\Big)(B^\star)^\intercal$ is rank-k with $\lambda_k(\mathbb{E}Z) = N\lambda_k$ and $\lambda_{k+1}(\mathbb{E}Z) = 0$. Therefore, we have

$$\begin{aligned}
\|\sin\Theta(\widehat{B}, B^\star)\| &\le \frac{2\|Z - \mathbb{E}Z\|}{\lambda_k(\mathbb{E}Z)} = \frac{O((\sqrt{Md} + \sqrt{Nd\lambda_1} + d)\cdot\log^3(d+N))}{N\lambda_k} \\
&= O\bigg(\bigg(\sqrt{\frac{d\lambda_1}{N\lambda_k^2}} + \sqrt{\frac{Md}{N^2\lambda_k^2}} + \frac{d}{N\lambda_k}\bigg)\cdot\log^3(d+N)\bigg) \\
&= O\bigg(\bigg(\sqrt{\frac{d\lambda_1}{N\lambda_k^2}} + \sqrt{\frac{Md}{N^2\lambda_k^2}}\bigg)\cdot\log^3(d+N)\bigg),
\end{aligned}$$

where the last line follows from $d/(N\lambda_k) \le \sqrt{d/(N\lambda_k)} \le \sqrt{d\lambda_1/(N\lambda_k^2)}$ given $N\lambda_k = \Omega(d)$ and $\lambda_1/\lambda_k \ge 1$. □

Now it remains to bound each error term using the truncated matrix Bernstein inequality. Interested readers can find a proof of this truncated variant in [18, Section A.2.2].

Theorem B.2 (Truncated matrix Bernstein's inequality). *Let $Z_1, \cdots, Z_M \in \mathbb{R}^{d_1 \times d_2}$ be independent random matrices. Suppose there exist positive constants β, q, and $\delta \leq 1$, such that for any $i \in [M]$,*

$$\mathbb{P}\big(\|Z_i - \mathbb{E}Z_i\| \geq \beta\big) \leq \delta$$
$$\big\|\mathbb{E}Z_i - \mathbb{E}[Z_i \mathbb{1}\{\|Z_i\| < \beta\}]\big\| \leq q.$$

In addition, let v be the matrix variance statistic defined as

$$v = \max\Big\{\Big\|\sum_{i=1}^{M}\big(\mathbb{E}[Z_i Z_i^\intercal] - (\mathbb{E}Z_i)(\mathbb{E}Z_i^\intercal)\big)\Big\|, \Big\|\sum_{i=1}^{M}\big(\mathbb{E}[Z_i^\intercal Z_i] - (\mathbb{E}Z_i^\intercal)(\mathbb{E}Z_i)\big)\Big\|\Big\}.$$

Then for any $t \geq Mq$, we have

$$\mathbb{P}\bigg(\Big\|\sum_{i=1}^{M}\big(Z_i - \mathbb{E}Z_i\big)\Big\| \geq t\bigg) \leq (d_1 + d_2)\exp\Big(-\frac{(t - Mq)^2/2}{v + 2\beta(t - Mq)/3}\Big) + M\delta.$$

We present a user-friendly corollary of Theorem B.2 as follows [6].

Corollary B.1. *Suppose the conditions of Theorem B.2 hold, and set $d = \max\{d_1, d_2\}$. For any $c \geq 2$, with probability at least $1 - 2d^{-c+1} - M\delta$, we have*

$$\Big\|\sum_{i=1}^{M}\big(Z_i - \mathbb{E}Z_i\big)\Big\| \leq \sqrt{2cv\log d} + \frac{2c}{3}\beta\log d + Mq.$$

We now begin our analysis of the error terms. The next lemma applies Bernstein's inequality on a normalized sum of sub-exponential variables and helps identify the truncation level of our targeted random matrices.

Lemma B.1. *Let $\zeta_1, \cdots, \zeta_n \in \mathbb{R}$ and $x_1, \cdots, x_n \in \mathbb{R}^d$ be a sequence of sub-gaussian random variables and random vectors respectively, with constant variance proxies. Assume (ζ_j, x_j) are mutually independent pairs across different $j \in [n]$, while ζ_j and x_j may be dependent on each other. Let η be the normalized sum of their products, that is, $\eta = (\sum_{j=1}^{n}\zeta_j x_j)/\sqrt{n}$. Then there exist positive constants c_1 and c_2 such that for any $t > 0$,*

$$\mathbb{P}\big(\|\eta - \mathbb{E}\eta\| \geq t\big) \leq 2d\exp\big(-\min\big\{c_1 t^2/d, c_2 t\sqrt{n/d}\big\}\big).$$

Proof. Let $x_j^{(r)}$ denote the r-th entry of vector x_j, and define $\eta^{(r)} = (\sum_{j=1}^{n}\zeta_j x_j^{(r)})/\sqrt{n}$ as the r-th entry of η. Then we have $\|\eta - \mathbb{E}\eta\|^2 = \sum_{r=1}^{d}(\eta^{(r)} - \mathbb{E}\eta^{(r)})^2$ and thus

$$\mathbb{P}\big(\|\eta - \mathbb{E}\eta\| \geq t\big) = \mathbb{P}\Big(\sum_{r=1}^{d}\big(\eta^{(r)} - \mathbb{E}\eta^{(r)}\big)^2 \geq t^2\Big) \leq \sum_{r=1}^{d}\mathbb{P}\Big(\big|\eta^{(r)} - \mathbb{E}\eta^{(r)}\big| \geq t/\sqrt{d}\Big).$$

For a fixed r, the product of sub-gaussians, $\zeta_j x_j^{(r)}$, is sub-exponential; $\{\zeta_j x_j^{(r)} - \mathbb{E}\zeta_j x_j^{(r)}\}_{j\in[n]}$ is a sequence of independent, mean zero, sub-exponential random variables. Thus, Bernstein's inequality [35, Theorem 2.8.1] yields the existence of positive constants c_1 and c_2 such that,

$$\begin{aligned}\mathbb{P}\Big(\big|\eta^{(r)} - \mathbb{E}\eta^{(r)}\big| \geq t/\sqrt{d}\Big) &= \mathbb{P}\bigg(\Big|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\Big(\zeta_j x_j^{(r)} - \mathbb{E}\zeta_j x_j^{(r)}\Big)\Big| \geq t/\sqrt{d}\bigg)\\ &\leq 2\exp\big(-\min\big\{c_1 t^2/d, c_2 t\sqrt{n/d}\big\}\big).\end{aligned}$$

We conclude the proof by combining the above two equations. □

The following lemma assists in bounding the mean shift after truncation.

Lemma B.2. *Let $a_i, b_i \in \mathbb{R}^d$ be independent random vectors and $Z_i = a_i b_i^\intercal$. Then for any $\beta > 0$,*

$$\big\|\mathbb{E}Z_i - \mathbb{E}[Z_i\mathbb{1}\{\|Z_i\| < \beta\}]\big\| \leq \sqrt{\mathbb{E}\big[\|a_i\|^2\big]\mathbb{E}\big[\|b_i\|^2\big]\mathbb{P}\big(\|Z_i\| \geq \beta\big)}.$$

Proof. We bound the mean shift after truncation as follows,

$$\begin{aligned}\big\|\mathbb{E}Z_i - \mathbb{E}[Z_i\mathbb{1}\{\|Z_i\| < \beta\}]\big\| &= \big\|\mathbb{E}[Z_i\mathbb{1}\{\|Z_i\| \geq \beta\}]\big\| \\ &\leq \mathbb{E}\big[\|Z_i\|\mathbb{1}\{\|Z_i\| \geq \beta\}\big] \\ &\leq \sqrt{\mathbb{E}\big[\|Z_i\|^2\big]\mathbb{E}\big[\mathbb{1}\{\|Z_i\| \geq \beta\}\big]},\end{aligned}$$

where Cauchy–Schwarz inequality gives the last line. Since a_i and b_i are independent, we have $\mathbb{E}[\|Z_i\|^2] = \mathbb{E}[\|a_i\|^2\|b_i\|^2] = \mathbb{E}[\|a_i\|^2]\mathbb{E}[\|b_i\|^2]$. We further note $\mathbb{E}[\mathbb{1}\{\|Z_i\| \geq \beta\}] = \mathbb{P}(\|Z_i\| \geq \beta)$ and complete the proof. □

The following two lemmas bound the variance of random vectors $\sqrt{n_i}\overline{\Gamma}_i\theta_i$ and $\sqrt{n_i}\overline{E}_i$.

Lemma B.3. *Suppose Assumptions 2.1-2.4 hold. Let* $a_i = \sqrt{n_i}\overline{\Gamma}_i\theta_i = 2/\sqrt{n_i}\cdot\sum_{j=1}^{n_i/2} x_{ij}x_{ij}^\intercal\theta_i$. *We have*

$$\|\mathbb{E}a_i\| \leq \sqrt{N\lambda_1}, \quad \mathbb{E}\big[\|a_i\|^2\big] = O(d\|\alpha_i^\star\|^2) + n_i\|\alpha_i^\star\|^2.$$

In addition, for any $s \in \mathbb{S}^{d-1}$, *we have*

$$s^\intercal\mathbb{E}[a_ia_i^\intercal]s = O(\|\alpha_i^\star\|^2) + n_is^\intercal B^\star\alpha_i^\star(\alpha_i^\star)^\intercal(B^\star)^\intercal s.$$

Proof. We first note that $\mathbb{E}a_i = \sqrt{n_i}B^\star\alpha_i^\star$ and $\mathbb{E}a_i\mathbb{E}a_i^\intercal = n_iB^\star\alpha_i^\star(\alpha_i^\star)^\intercal(B^\star)^\intercal$. For any $s \in \mathbb{S}^{d-1}$, we have $\|(B^\star)^\intercal s\| \leq 1$ and thus

$$\begin{aligned}s^\intercal\mathbb{E}a_i\mathbb{E}a_i^\intercal s = n_is^\intercal B^\star\alpha_i^\star(\alpha_i^\star)^\intercal(B^\star)^\intercal s &\leq s^\intercal B^\star\Big(\sum_{i=1}^M n_i\alpha_i^\star(\alpha_i^\star)^\intercal\Big)(B^\star)^\intercal s \\ &\leq N\lambda_1\|(B^\star)^\intercal s\| \leq N\lambda_1.\end{aligned}$$

Then we can bound the vector norm as follows, and conclude the first statement,

$$\|\mathbb{E}a_i\|^2 = \|\mathbb{E}a_i\mathbb{E}a_i^\intercal\| \leq N\lambda_1.$$

Next, we compute $\mathbb{E}[\|a_i\|^2]$ using its definition,

$$\begin{aligned}\mathbb{E}\big[\|a_i\|^2\big] = \mathbb{E}\big[n_i\theta_i^\intercal\overline{\Gamma}_i^\intercal\overline{\Gamma}_i\theta_i\big] &= \mathbb{E}\Big[n_i\theta_i^\intercal\Big(\frac{2}{n_i}\sum_{j=1}^{n_i/2}x_{ij}x_{ij}^\intercal\Big)\Big(\frac{2}{n_i}\sum_{j=1}^{n_i/2}x_{ij}x_{ij}^\intercal\Big)\theta_i\Big] \\ &= \frac{4}{n_i}\Big[\mathbb{E}\Big(\sum_{j=1}^{n_i/2}\theta_i^\intercal x_{ij}x_{ij}^\intercal x_{ij}x_{ij}^\intercal\theta_i\Big) + \mathbb{E}\Big(\sum_{j=1}^{n_i/2}\sum_{r\neq j}\theta_i^\intercal x_{ij}x_{ij}^\intercal x_{ir}x_{ir}^\intercal\theta_i\Big)\Big] \\ &= 2\mathbb{E}\big[(\theta_i^\intercal x_{i1})^2\|x_{i1}\|^2\big] + (n_i-2)\theta_i^\intercal\mathbb{E}\big[x_{i1}x_{i1}^\intercal\big]\mathbb{E}\big[x_{i2}x_{i2}^\intercal\big]\theta_i \\ &= 2\mathbb{E}\big[((\theta_i/\|\theta_i\|)^\intercal x_{i1})^2\|x_{i1}\|^2\big]\cdot\|\theta_i\|^2 + (n_i-2)\|B^\star\alpha_i^\star\|^2. \qquad \text{(B.3)}\end{aligned}$$

We bound the first term using Cauchy–Schwarz inequality,

$$\mathbb{E}\big[((\theta_i/\|\theta_i\|)^\intercal x_{i1})^2\|x_{i1}\|^2\big]\cdot\|\theta_i\|^2 \leq \sqrt{\mathbb{E}\big[((\theta_i/\|\theta_i\|)^\intercal x_{i1})^4\big]\mathbb{E}\big[\|x_{i1}\|^4\big]}\cdot\|\theta_i\|^2.$$

The moments of sub-gaussian variables are bounded by constants that rely on the variance proxies, thus $\mathbb{E}[((\theta_i/\|\theta_i\|)^\intercal x_{i1})^4] = O(1)$. Given bounded moments $\mathbb{E}[(x_{i1}^{(r)})^4]$, we have $\mathbb{E}[\|x_{i1}\|^4] = \mathbb{E}[(\sum_{r=1}^d(x_{i1}^{(r)})^2)^2] \leq d\sum_{r=1}^d\mathbb{E}[(x_{i1}^{(r)})^4] = O(d^2)$. In addition, $\lambda_d(\Gamma_i) = \Theta(1)$ from Assumption 2.2 implies $\|\theta_i\|^2 = \|\Gamma_i^{-1}B^\star\alpha_i^\star\|^2 = \Theta(\|\alpha_i^\star\|^2)$. As a summary, the first term of (B.3) is of the following order:

$$2\mathbb{E}\big[((\theta_i/\|\theta_i\|)^\intercal x_{i1})^2\|x_{i1}\|^2\big]\cdot\|\theta_i\|^2 = O(d\|\alpha_i^\star\|^2).$$

Thus, following from (B.3) and $(B^\star)^\intercal B^\star = I_k$, we conclude $\mathbb{E}[\|a_i\|^2] = O(d\|\alpha_i^\star\|^2) + n_i\|\alpha_i^\star\|^2$.

Similarly, we compute $\mathbb{E}[a_ia_i^\intercal]$ as follows,

$$\mathbb{E}[a_ia_i^\intercal] = \mathbb{E}\big[n_i\overline{\Gamma}_i\theta_i\theta_i^\intercal\overline{\Gamma}_i^\intercal\big] = \mathbb{E}\Big[n_i\Big(\frac{2}{n_i}\sum_{j=1}^{n_i/2}x_{ij}x_{ij}^\intercal\Big)\theta_i\theta_i^\intercal\Big(\frac{2}{n_i}\sum_{j=1}^{n_i/2}x_{ij}x_{ij}^\intercal\Big)\Big]$$

$$= \frac{4}{n_i}\Big[\mathbb{E}\Big(\sum_{j=1}^{n_i/2} x_{ij}x_{ij}^\intercal\theta_i\theta_i^\intercal x_{ij}x_{ij}^\intercal\Big) + \mathbb{E}\Big(\sum_{j=1}^{n_i/2}\sum_{r\neq j} x_{ij}x_{ij}^\intercal\theta_i\theta_i^\intercal x_{ir}x_{ir}^\intercal\Big)\Big]$$
$$= 2\mathbb{E}\big[(\theta_i^\intercal x_{i1})^2 x_{i1}x_{i1}^\intercal\big] + (n_i-2)B^\star\alpha_i^\star(\alpha_i^\star)^\intercal(B^\star)^\intercal. \tag{B.4}$$

Then for the first term and any $s\in\mathbb{S}^{d-1}$, Cauchy–Schwarz inequality gives that

$$\begin{aligned} s^\intercal\mathbb{E}\big[(\theta_i^\intercal x_{i1})^2 x_{i1}x_{i1}^\intercal\big]s &= \mathbb{E}\big[((\theta_i/\|\theta_i\|)^\intercal x_{i1})^2(s^\intercal x_{i1})^2\big]\cdot\|\theta_i\|^2 \\ &\le \sqrt{\mathbb{E}\big[((\theta_i/\|\theta_i\|)^\intercal x_{i1})^4\big]\mathbb{E}\big[(s^\intercal x_{i1})^4\big]}\cdot\|\theta_i\|^2 \\ &= O(\|\alpha_i^\star\|^2), \end{aligned}$$

where last line holds since $(\theta_i/\|\theta_i\|)^\intercal x_{i1}$ and $u^\intercal x_{i1}$ are sub-gaussian variables with bounded constant moments, and $\|\theta_i\|^2 = \Theta(\|\alpha_i^\star\|^2)$ due to Assumption 2.2. Thus, following from (B.4), for any $s\in\mathbb{S}^{d-1}$, we conclude that

$$\begin{aligned} s^\intercal\mathbb{E}[a_ia_i^\intercal]s &= 2s^\intercal\mathbb{E}\big[(\theta_i^\intercal x_{i1})^2x_{i1}x_{i1}^\intercal\big]s + (n_i-2)s^\intercal B^\star\alpha_i^\star(\alpha_i^\star)^\intercal(B^\star)^\intercal s \\ &= O(\|\alpha_i^\star\|^2) + n_i s^\intercal B^\star\alpha_i^\star(\alpha_i^\star)^\intercal(B^\star)^\intercal s. \end{aligned}$$

□

Lemma B.4. *Suppose Assumptions 2.1-2.4 hold. Let* $b_i = \sqrt{n_i}\overline{E}_i = 2/\sqrt{n_i}\cdot\sum_{j=1}^{n_i/2}\xi_{ij}x_{ij}$. *We have*

$$\mathbb{E}\big[\|b_i\|^2\big] = O(d), \quad \big\|\mathbb{E}[b_ib_i^\intercal]\big\| = O(1).$$

Proof. By definition, we have

$$\begin{aligned} \mathbb{E}\big[\|b_i\|^2\big] &= \mathbb{E}\big[n_i\overline{E}_i^\intercal\overline{E}_i\big] = \mathbb{E}\Big[n_i\Big(\frac{2}{n_i}\sum_{j=1}^{n_i/2}\xi_{ij}x_{ij}\Big)^\intercal\Big(\frac{2}{n_i}\sum_{j=1}^{n_i/2}\xi_{ij}x_{ij}\Big)\Big] \\ &= \frac{4}{n_i}\mathbb{E}\Big[\sum_{j=1}^{n_i/2}\sum_{r=1}^{n_i/2}\xi_{ij}\xi_{ir}x_{ij}^\intercal x_{ir}\Big] = \frac{4}{n_i}\sum_{j=1}^{n_i/2}\mathbb{E}[\xi_{ij}^2]\mathbb{E}[\|x_{ij}\|^2] = 2\mathbb{E}[\xi_{i1}^2]\mathbb{E}\big[\|x_{i1}\|^2\big]. \end{aligned}$$

Since $\mathbb{E}[\|x_{i1}\|^2] = \sum_{r=1}^d\mathbb{E}(x_{i1}^{(r)})^2 = d\mathbb{E}(x_{i1}^{(1)})^2$, and the variances of sub-gaussian random variables, $\mathbb{E}[\xi_{i1}^2]$ and $\mathbb{E}(x_{i1}^{(1)})^2$, are bounded by constants dependent on the variance proxies, we have $\mathbb{E}[\|b_i\|^2] = O(d)$. Similarly, the straightforward computation gives

$$\mathbb{E}[b_ib_i^\intercal] = 2\mathbb{E}[\xi_{i1}^2]\mathbb{E}[x_{i1}x_{i1}^\intercal],$$

where the variance $\mathbb{E}[\xi_{i1}^2]$ is bounded. For any $u\in\mathbb{S}^{d-1}$, since $u^\intercal x_{i1}$ is sub-gaussian, its variance $u^\intercal\mathbb{E}[x_{i1}x_{i1}^\intercal]u = \mathbb{E}[\|u^\intercal x_{i1}\|^2]$ is bounded. Thus, we have $\|\mathbb{E}[x_{i1}x_{i1}^\intercal]\| = O(1)$ and conclude that $\|\mathbb{E}[b_ib_i^\intercal]\| \le 2\mathbb{E}[\xi_{i1}^2]\|\mathbb{E}[x_{i1}x_{i1}^\intercal]\| = O(1)$. □

We now bound the three error terms in the proof of Theorem 3.1 in the following lemmas.

Bounding the first error term in (B.2).

Lemma B.5. *Suppose Assumptions 2.1-2.4 hold. With probability at least* $1-O((d+N)^{-10})$, *we have*

$$\Big\|\sum_{i=1}^M n_i\overline{\Gamma}_i\theta_i\theta_i^\intercal\widetilde{\Gamma}_i^\intercal - B^\star\Big(\sum_{i=1}^M n_i\alpha_i^\star(\alpha_i^\star)^\intercal\Big)(B^\star)^\intercal\Big\| = O\big(\big(\sqrt{Md}+\sqrt{Nd\lambda_1}+d\big)\cdot\log^3(d+N)\big).$$

Proof. Let $a_i = \sqrt{n_i}\overline{\Gamma}_i\theta_i$, $u_i = \sqrt{n_i}\widetilde{\Gamma}_i\theta_i$, and $Z_i = a_iu_i^\intercal$. Then a_i and u_i are independent and identically distributed, and we aim to bound $\|\sum_{i=1}^M(n_i\overline{\Gamma}_i\theta_i\theta_i^\intercal\widetilde{\Gamma}_i^\intercal - n_iB^\star\alpha_i^\star(\alpha_i^\star)^\intercal(B^\star)^\intercal)\| = \|\sum_{i=1}^M(Z_i-\mathbb{E}Z_i)\|$ by the truncated matrix Bernstein's inequality in Corollary B.1. Since $\|Z_i-\mathbb{E}Z_i\|$ might be unbounded, we first identify an appropriate truncation level. By adding and subtracting terms, we have

$$\big\|Z_i-\mathbb{E}Z_i\big\| = \big\|a_iu_i^\intercal - \mathbb{E}a_i\mathbb{E}u_i^\intercal\big\|$$

$$\leq \big\|(a_i - \mathbb{E}a_i)(u_i - \mathbb{E}u_i)^\intercal\big\| + \big\|(a_i - \mathbb{E}a_i)\mathbb{E}u_i^\intercal\big\| + \big\|\mathbb{E}a_i(u_i - \mathbb{E}u_i)^\intercal\big\|.$$

Since a_i and u_i share identical distributions, the above equation yields that for any $\beta > 0$,

$$\begin{aligned}
\mathbb{P}\big(\|Z_i - \mathbb{E}Z_i\| \geq \beta\big) &\leq \mathbb{P}\big(\|(a_i - \mathbb{E}a_i)(u_i - \mathbb{E}u_i)^\intercal\| \geq \beta/3\big) + 2\mathbb{P}\big(\|(a_i - \mathbb{E}a_i)\mathbb{E}u_i^\intercal\| \geq \beta/3\big) \\
&\leq 2\mathbb{P}\big(\|a_i - \mathbb{E}a_i\| \geq \sqrt{\beta/3}\big) + 2\mathbb{P}\big(\|\mathbb{E}u_i\| \cdot \|a_i - \mathbb{E}a_i\| \geq \beta/3\big) \\
&\leq 2\mathbb{P}\big(\|a_i - \mathbb{E}a_i\| \geq \sqrt{\beta/3}\big) + 2\mathbb{P}\big(\|a_i - \mathbb{E}a_i\| \geq \beta/\big(3\sqrt{N\lambda_1}\big)\big),
\end{aligned}$$

where the last inequality holds since $\|\mathbb{E}u_i\| \leq \sqrt{N\lambda_1}$ by Lemma B.3. Note $a_i = 2/\sqrt{n_i} \cdot \sum_{j=1}^{n_i/2}(x_{ij}^\intercal\theta_i)x_{ij}$, where $x_{ij}^\intercal\theta_i$ is a sub-gaussian variable and x_{ij} is a sub-gaussian vector. Thus, following from the above equation, we apply the bound from Lemma B.1 to $\|a_i - \mathbb{E}a_i\|$, and have the existence of constants c_1, c_2, c_3, c_4 such that for any $\beta > 0$,

$$\begin{aligned}
\mathbb{P}\big(\|Z_i - \mathbb{E}Z_i\| \geq \beta\big) \leq\ & 4d\exp\big(-\min\big\{c_1\beta/d, c_2\sqrt{n_i\beta/d}\big\}\big) \\
&+ 4d\exp\big(-\min\big\{c_3\beta^2/(Nd\lambda_1), c_4\beta\sqrt{n_i/(Nd\lambda_1)}\big\}\big).
\end{aligned}$$

For any $\delta > 0$, we take a large enough C and set $\beta = C\max\{d\log^2(d/\delta), \sqrt{Nd\lambda_1}\log(d/\delta)\}$. Thus, using $n_i \geq 1$, we obtain from the above equation ,

$$\mathbb{P}\big(\|Z_i - \mathbb{E}Z_i\| \geq \beta\big) \leq \delta. \tag{B.5}$$

Next, we bound the mean shift after truncation using Lemma B.2. For β defined above, we have

$$\big\|\mathbb{E}Z_i - \mathbb{E}[Z_i\mathbb{1}\{\|Z_i\| < \beta\}]\big\| \leq \sqrt{\mathbb{E}\big[\|a_i\|^2\big]\mathbb{E}\big[\|u_i\|^2\big]}\sqrt{\delta}.$$

Lemma B.3 provides the bound that $\mathbb{E}[\|a_i\|^2] = \mathbb{E}[\|u_i\|^2] = O(d + N)$. Thus, we have

$$\big\|\mathbb{E}Z_i - \mathbb{E}[Z_i\mathbb{1}\{\|Z_i\| < \beta\}]\big\| = O\big((d+N)\sqrt{\delta}\big) \coloneqq q. \tag{B.6}$$

Then it remains to determine the variance statistic. We first have

$$\mathbb{E}[Z_iZ_i^\intercal] = \mathbb{E}[a_iu_i^\intercal u_ia_i^\intercal] = \mathbb{E}[\|u_i\|^2]\mathbb{E}[a_ia_i^\intercal].$$

Note that $\mathbb{E}Z_i = n_iB^\star\alpha_i^\star(\alpha_i^\star)^\intercal(B^\star)^\intercal$. Then Lemma B.3 yields that, for any $s \in \mathbb{S}^{d-1}$,

$$\begin{aligned}
&s^\intercal\big(\mathbb{E}[Z_iZ_i^\intercal] - (\mathbb{E}Z_i)(\mathbb{E}Z_i^\intercal)\big)s \\
&\quad= \mathbb{E}[\|u_i\|^2]\cdot s^\intercal\mathbb{E}[a_ia_i^\intercal]s - n_i^2\|B^\star\alpha_i^\star\|^2\cdot s^\intercal B^\star\alpha_i^\star(\alpha_i^\star)^\intercal(B^\star)^\intercal s \\
&\quad= \big(O(d\|\alpha_i^\star\|^2) + n_i\|\alpha_i^\star\|^2\big)\cdot\big(O(\|\alpha_i^\star\|^2) + n_is^\intercal B^\star\alpha_i^\star(\alpha_i^\star)^\intercal(B^\star)^\intercal s\big) \\
&\qquad\quad - n_i^2\|\alpha_i^\star\|^2\cdot s^\intercal B^\star\alpha_i^\star(\alpha_i^\star)^\intercal(B^\star)^\intercal s \\
&\quad= O\big(d\|\alpha_i^\star\|^4 + n_i\|\alpha_i^\star\|^4 + dn_i\|\alpha_i^\star\|^2 s^\intercal B^\star\alpha_i^\star(\alpha_i^\star)^\intercal(B^\star)^\intercal s\big).
\end{aligned}$$

Since $\sum_{i=1}^M n_i\|\alpha_i^\star\|^2 = \mathrm{Tr}(\sum_{i=1}^M n_i\alpha_i^\star(\alpha_i^\star)^\intercal) = O(Nk\lambda_1)$, when summing over all $i \in [M]$, we have

$$\begin{aligned}
&s^\intercal\Big[\sum_{i=1}^M\big(\mathbb{E}[Z_iZ_i^\intercal] - (\mathbb{E}Z_i)(\mathbb{E}Z_i^\intercal)\big)\Big]s \\
&= O\Big(d\sum_{i=1}^M\|\alpha_i^\star\|^4 + \max_i\|\alpha_i^\star\|^2\cdot\sum_{i=1}^M n_i\|\alpha_i^\star\|^2 + d\max_i\|\alpha_i^\star\|^2\cdot s^\intercal B^\star\big(\sum_{i=1}^M n_i\alpha_i^\star(\alpha_i^\star)^\intercal\big)(B^\star)^\intercal s\Big) \\
&= O\big(Md\cdot\max_i\|\alpha_i^\star\|^4 + Nk\lambda_1\cdot\max_i\|\alpha_i^\star\|^2 + Nd\lambda_1\cdot\max_i\|\alpha_i^\star\|^2\cdot\|(B^\star)^\intercal s\|\big) \\
&= O(Md + Nd\lambda_1).
\end{aligned} \tag{B.7}$$

where the third line holds since $\lambda_1(\sum_{i=1}^M n_i\alpha_i^\star(\alpha_i^\star)^\intercal) \leq N\lambda_1$ and the last line follows from $d \geq k$, $\max_i\|\alpha_i^\star\|^2 = O(1)$, and $\|(B^\star)^\intercal s\| \leq 1$. In addition, note that $\mathbb{E}[Z_i^\intercal Z_i] = \mathbb{E}[u_ia_i^\intercal a_iu_i^\intercal] = \mathbb{E}[\|a_i\|^2]\mathbb{E}[u_iu_i^\intercal]$. Since a_i and u_i share identical distributions, and $\mathbb{E}Z_i = \mathbb{E}Z_i^\intercal$, we observe

$\mathbb{E}[Z_i Z_i^\intercal] - (\mathbb{E}Z_i)(\mathbb{E}Z_i^\intercal) = \mathbb{E}[Z_i^\intercal Z_i] - (\mathbb{E}Z_i^\intercal)(\mathbb{E}Z_i)$. Thus, following from (B.7), the variance statistic is bounded such that

$$\begin{aligned} v &= \max\Big\{\Big\|\sum_{i=1}^{M}\big(\mathbb{E}[Z_i Z_i^\intercal] - (\mathbb{E}Z_i)(\mathbb{E}Z_i^\intercal)\big)\Big\|, \Big\|\sum_{i=1}^{M}\big(\mathbb{E}[Z_i^\intercal Z_i] - (\mathbb{E}Z_i^\intercal)(\mathbb{E}Z_i)\big)\Big\|\Big\} \\ &= O(Md + Nd\lambda_1). \end{aligned} \tag{B.8}$$

We apply Corollary B.1 with $c = 11$, and β, q, and v discussed in (B.5), (B.6), and (B.8). Then with probability at least $1 - 2d^{-10} - M\delta$, we have

$$\begin{aligned} \Big\|\sum_{i=1}^{M}\big(Z_i - \mathbb{E}Z_i\big)\Big\| &\le \sqrt{2cv\log d} + 2c\beta\log d/3 + Mq \\ &= O\big(\sqrt{(Md + Nd\lambda_1)\log d} + \big(d\log^2(d/\delta) + \sqrt{Nd\lambda_1}\log(d/\delta)\big)\log d + \sqrt{\delta}M(d+N)\big). \end{aligned}$$

Note that $M \le N$. Taking $\delta = (d+N)^{-11}$, it holds with probability at least $1 - O((d+N)^{-10})$ that,

$$\Big\|\sum_{i=1}^{M}\big(Z_i - \mathbb{E}Z_i\big)\Big\| = O\big(\big(\sqrt{Md} + \sqrt{Nd\lambda_1} + d\big)\cdot\log^3(d+N)\big). \qquad \square$$

Bounding the second and third error terms in (B.2).

Lemma B.6. *Suppose Assumptions 2.1-2.4 hold. With probability at least $1 - O((d+N)^{-10})$, we have*

$$\Big\|\sum_{i=1}^{M} n_i\overline{\Gamma}_i\theta_i\widetilde{E}_i^\intercal\Big\| = O\big(\big(\sqrt{Md} + \sqrt{Nd\lambda_1} + d\big)\cdot\log^3(d+N)\big).$$

Proof. Let $a_i = \sqrt{n_i}\overline{\Gamma}_i\theta_i$, $b_i = \sqrt{n_i}\widetilde{E}_i$, and $Z_i = a_i b_i^\intercal$ with $\mathbb{E}b_i = 0$ and $\mathbb{E}Z_i = 0$. We first identify a truncation level of $\|Z_i\|$. Since $\|Z_i\| = \|a_i b_i^\intercal\| \le \|(a_i - \mathbb{E}a_i)b_i^\intercal\| + \|(\mathbb{E}a_i)b_i^\intercal\|$, we have for any $\beta \ge 0$,

$$\begin{aligned} \mathbb{P}\big(\|Z_i\| \ge \beta\big) &\le \mathbb{P}\big(\|(a_i - \mathbb{E}a_i)b_i^\intercal\| \ge \beta/2\big) + \mathbb{P}\big(\|(\mathbb{E}a_i)b_i^\intercal\| \ge \beta/2\big) \\ &\le \mathbb{P}\big(\|a_i - \mathbb{E}a_i\| \ge \sqrt{\beta/2}\big) + \mathbb{P}\big(\|b_i\| \ge \sqrt{\beta/2}\big) + \mathbb{P}\big(\|b_i\| \ge \beta/2\big(\sqrt{N\lambda_1}\big)\big), \end{aligned}$$

where the last inequality holds since $\|\mathbb{E}a_i\| \le \sqrt{N\lambda_1}$ by Lemma B.3. Note that $a_i = 2/\sqrt{n_i}\cdot\sum_{j=1}^{n_i/2}(x_{ij}^\intercal\theta_i)x_{ij}$ with sub-gaussian variable $x_{ij}^\intercal\theta_i$ and vector x_{ij} and $b_i = 2/\sqrt{n_i}\cdot\sum_{j=n_i/2+1}^{n_i}\xi_{ij}x_{ij}$. Applying the bounds on $\|a_i - \mathbb{E}a_i\|$ and $\|b_i\|$ from Lemma B.1 to the above equation, there are constants c_1, c_2, c_3, c_4 such that for any $\beta > 0$,

$$\begin{aligned} \mathbb{P}\big(\|Z_i\| \ge \beta\big) &\le 4d\exp\big(-\min\big\{c_1\beta/d, c_2\sqrt{n_i\beta/d}\big\}\big) \\ &\quad + 2d\exp\big(-\min\big\{c_3\beta^2/(Nd\lambda_1), c_4\beta\sqrt{n_i/(Nd\lambda_1)}\big\}\big). \end{aligned}$$

For any $\delta > 0$, we take a large enough C and set $\beta = C\max\{d\log^2(d/\delta), \sqrt{Nd\lambda_1}\log(d/\delta)\}$ thereby obtaining (noting $n_i \ge 1$)

$$\mathbb{P}\big(\|Z_i\| \ge \beta\big) \le \delta. \tag{B.9}$$

Next, we bound the mean shift after truncation using Lemma B.2. For β defined above, we have

$$\big\|\mathbb{E}Z_i - \mathbb{E}[Z_i\mathbb{1}\{\|Z_i\| < \beta\}]\big\| \le \sqrt{\mathbb{E}\big[\|a_i\|^2\big]\mathbb{E}\big[\|b_i\|^2\big]}\sqrt{\delta}.$$

Lemmas B.3 and B.4 provide the bounds that $\mathbb{E}[\|a_i\|^2] = O(d+N)$ and $\mathbb{E}[\|b_i\|^2] = O(d)$. Thus, we have

$$\big\|\mathbb{E}Z_i - \mathbb{E}[Z_i\mathbb{1}\{\|Z_i\| < \beta\}]\big\| = O\big(\sqrt{d(d+N)}\sqrt{\delta}\big) := q. \tag{B.10}$$

Then it remains to determine the variance statistic. We first have

$$\mathbb{E}[Z_i Z_i^\intercal] = \mathbb{E}[a_i b_i^\intercal b_i a_i^\intercal] = \mathbb{E}[\|b_i\|^2]\mathbb{E}[a_i a_i^\intercal].$$

Then Lemmas B.3 and B.4 yield that, for any $s \in \mathbb{S}^{d-1}$,

$$\begin{aligned} s^\intercal \mathbb{E}[Z_i Z_i^\intercal] s &= \mathbb{E}[\|b_i\|^2] \cdot s^\intercal \mathbb{E}[a_i a_i^\intercal] s \\ &= O(d\|\alpha_i^\star\|^2) + O(d) \cdot n_i s^\intercal B^\star \alpha_i^\star (\alpha_i^\star)^\intercal (B^\star)^\intercal s. \end{aligned}$$

Therefore, when summing over $i \in [M]$, we have

$$\begin{aligned} s^\intercal \Big(\sum_{i=1}^M \mathbb{E}[Z_i Z_i^\intercal] \Big) s &= O(d) \cdot \sum_{i=1}^M \|\alpha_i^\star\|^2 + O(d) \cdot s^\intercal B^\star \Big(\sum_{i=1}^M n_i \alpha_i^\star (\alpha_i^\star)^\intercal \Big) (B^\star)^\intercal s \\ &= O(Md) \cdot \max_i \|\alpha_i^\star\|^2 + O(Nd\lambda_1) \cdot \|(B^\star)^\intercal s\| \\ &= O(Md + Nd\lambda_1), \end{aligned} \tag{B.11}$$

where the second line holds since $\lambda_1(\sum_{i=1}^M n_i \alpha_i^\star (\alpha_i^\star)^\intercal) \le N\lambda_1$ and the last line holds since $\max_i \|\alpha_i^\star\|^2 = O(1)$ and $\|(B^\star)^\intercal s\| \le 1$. Next, we have $\mathbb{E}[Z_i^\intercal Z_i] = \mathbb{E}[b_i a_i^\intercal a_i b_i^\intercal] = \mathbb{E}[\|a_i\|^2]\mathbb{E}[b_i b_i^\intercal]$. Then Lemmas B.3 and B.4 give that

$$\big\|\mathbb{E}[Z_i^\intercal Z_i]\big\| \le \mathbb{E}[\|a_i\|^2] \cdot \big\|\mathbb{E}[b_i b_i^\intercal]\big\| = \big(O(d\|\alpha_i^\star\|^2) + n_i\|\alpha_i^\star\|^2\big) \cdot O(1) = O(d\|\alpha_i^\star\|^2 + n_i\|\alpha_i^\star\|^2).$$

Thus, since $\sum_{i=1}^M n_i \|\alpha_i^\star\|^2 = \mathrm{Tr}(\sum_{i=1}^M n_i \alpha_i \alpha_i^\intercal) = O(Nk\lambda_1)$, we have

$$\begin{aligned} \Big\| \sum_{i=1}^M \mathbb{E}[Z_i^\intercal Z_i] \Big\| &\le \sum_{i=1}^M \big\|\mathbb{E}[Z_i^\intercal Z_i]\big\| \\ &= O(Md) \cdot \max_i \|\alpha_i^\star\|^2 + O\Big(\sum_{i=1}^M n_i \|\alpha_i^\star\|^2 \Big) \\ &= O(Md + Nk\lambda_1). \end{aligned} \tag{B.12}$$

Following from (B.11) and (B.12) and recalling $\mathbb{E}Z_i = 0$, we bound the variance statistic as follows,

$$\begin{aligned} v &= \max\Big\{ \Big\| \sum_{i=1}^M \mathbb{E}[Z_i Z_i^\intercal] \Big\|, \Big\| \sum_{i=1}^M \mathbb{E}[Z_i^\intercal Z_i] \Big\| \Big\} \\ &= \max\big\{ O(Md + Nd\lambda_1), O(Md + Nk\lambda_1) \big\} \\ &= O(Md + Nd\lambda_1). \end{aligned} \tag{B.13}$$

We apply Corollary B.1 with $c = 11$, and β, q, and v discussed in (B.9), (B.10), and (B.13). Then with probability at least $1 - 2d^{-10} - M\delta$, we have

$$\begin{aligned} &\Big\| \sum_{i=1}^M (Z_i - \mathbb{E}Z_i) \Big\| \le \sqrt{2cv \log d} + 2c\beta \log d/3 + Mq \\ &= O\big(\sqrt{(Md + Nd\lambda_1)\log d} + \big(d\log^2(d/\delta) + \sqrt{Nd\lambda_1}\log(d/\delta) \big) \log d + \sqrt{\delta}\sqrt{d(d+N)}M \big). \end{aligned}$$

Note that $M \le N$. Taking $\delta = (d+N)^{-11}$, it holds with probability at least $1 - O((d+N)^{-10})$ that,

$$\Big\| \sum_{i=1}^M (Z_i - \mathbb{E}Z_i) \Big\| = O\big(\big(\sqrt{Md} + \sqrt{Nd\lambda_1} + d \big) \cdot \log^3(d+N) \big).$$

□

Bounding the last error term in (B.2).

Lemma B.7. *Suppose Assumptions 2.1-2.4 hold. With probability at least $1 - O((d+N)^{-10})$, we have*

$$\Big\| \sum_{i=1}^M n_i \overline{E}_i \widetilde{E}_i^\intercal \Big\| = O\big(\big(\sqrt{Md} + d \big) \cdot \log^3(d+N) \big).$$

Proof. Let $b_i = \sqrt{n_i}\overline{E}_i$, $v_i = \sqrt{n_i}\widetilde{E}_i$, and $Z_i = b_i v_i^\intercal = n_i \overline{E}_i \widetilde{E}_i^\intercal$, where $\mathbb{E}b_i = \mathbb{E}v_i = 0$ and $\mathbb{E}Z_i = 0$. We first identify a truncation level of $\|Z_i\|$. By applying Lemma B.1 to bound the norms of $b_i = 2/\sqrt{n_i} \cdot \sum_{j=1}^{n_i/2} \xi_{ij} x_{ij}$ and $v_i = 2/\sqrt{n_i} \cdot \sum_{j=n_i/2+1}^{n_i} \xi_{ij} x_{ij}$, there are c_1 and c_2 such that for any $\beta \geq 0$,

$$\begin{aligned}\mathbb{P}\big(\|Z_i\| \geq \beta\big) &\leq \mathbb{P}\big(\|a_i\|\|b_i\| \geq \beta\big) \leq \mathbb{P}\big(\|a_i\| \geq \sqrt{\beta}\big) + \mathbb{P}\big(\|b_i\| \geq \sqrt{\beta}\big) \\ &\leq 4d \exp\big(-\min\big\{c_1\beta/d, c_2\sqrt{n_i\beta/d}\big\}\big).\end{aligned}$$

For any $\delta > 0$, we take $\beta = Cd\log^2(d/\delta)$ with a large enough C in the above equation such that $\beta \geq C\max\{d\log(d/\delta), d\log^2(d/\delta)/\min_i\{n_i\}\}$ and thus obtain

$$\mathbb{P}\big(\|Z_i\| \geq \beta\big) \leq \delta. \tag{B.14}$$

Next, for β defined above, we establish bounds on the mean shift after truncation using Lemma B.2. Here we substitute $\mathbb{E}[\|b_i\|^2] = \mathbb{E}[\|v_i\|^2] = O(d)$ from Lemma B.4 to obtain,

$$\big\|\mathbb{E}Z_i - \mathbb{E}[Z_i \mathbb{1}\{\|Z_i\| < \beta\}]\big\| \leq \sqrt{\mathbb{E}\big[\|b_i\|^2\big]\mathbb{E}\big[\|v_i\|^2\big]}\sqrt{\delta} \leq O\big(\sqrt{\delta}d\big) \coloneqq q. \tag{B.15}$$

Then it remains to determine the variance statistic. By definition, we have

$$\begin{aligned}\mathbb{E}[Z_i Z_i^\intercal] &= \mathbb{E}[b_i v_i^\intercal v_i b_i^\intercal] = \mathbb{E}[\|v_i\|^2]\mathbb{E}[b_i b_i^\intercal], \\ \mathbb{E}[Z_i^\intercal Z_i] &= \mathbb{E}[v_i b_i^\intercal b_i v_i^\intercal] = \mathbb{E}[\|b_i\|^2]\mathbb{E}[v_i v_i^\intercal].\end{aligned}$$

Since b_i and v_i share identical distributions, here we have $\mathbb{E}[Z_i Z_i^\intercal] = \mathbb{E}[Z_i^\intercal Z_i]$. Lemma B.4 yields that

$$\big\|\mathbb{E}[Z_i Z_i^\intercal]\big\| = \big\|\mathbb{E}[Z_i^\intercal Z_i]\big\| \leq \mathbb{E}[\|v_i\|^2] \cdot \big\|\mathbb{E}[b_i b_i^\intercal]\big\| = O(d).$$

Since Z_i is mean zero, we further have

$$v = \max\Big\{\Big\|\sum_{i=1}^M \mathbb{E}[Z_i Z_i^\intercal]\Big\|, \Big\|\sum_{i=1}^M \mathbb{E}[Z_i^\intercal Z_i]\Big\|\Big\} = O(Md). \tag{B.16}$$

Applying Corollary B.1 with $c = 11$, and β, q, and v discussed in (B.14), (B.15), and (B.16), we obtain, with probability at least $1 - 2d^{-10} - M\delta$,

$$\begin{aligned}\Big\|\sum_{i=1}^M \big(Z_i - \mathbb{E}Z_i\big)\Big\| &\leq \sqrt{2cv\log d} + 2c\beta\log d/3 + Mq \\ &= O\big(\sqrt{Md\log d} + d\log^2(d/\delta)\log d + \sqrt{\delta}Md\big).\end{aligned}$$

Note that $M \leq N$. Taking $\delta = (d+N)^{-11}$, it holds with probability at least $1 - O((d+N)^{-10})$ that,

$$\Big\|\sum_{i=1}^M \big(Z_i - \mathbb{E}Z_i\big)\Big\| = O\big(\big(\sqrt{Md} + d\big) \cdot \log^3(d+N)\big).$$

□

C Proof of the Lower Bound

This section proves the lower bound by an information-theoretic argument via Fano's method. In particular, we establish the two terms in Theorem 4.1 separately in the subsequent theorems. For convenience, we define the parameter space of α for a fixed $\vec{n} = (n_1, \ldots, n_M)$ as follows:

$$\Psi_{\lambda_1,\lambda_k}^{n_1,\cdots,n_M} = \Big\{\alpha \in \mathbb{R}^{k\times M} \colon \|\alpha_i\| = O(1)\ \forall i \in [M], \Omega(\lambda_k)I_k \preceq \frac{1}{N}\sum_{i=1}^M n_i \alpha_i \alpha_i^\intercal \preceq O(\lambda_1)I_k\Big\}. \tag{C.1}$$

We assume $\lambda_k > 0$, $k\lambda_k = O(1)$, $\lambda_1 = O(1)$, $M \geq k$; otherwise, the parameter space is empty.

Theorem C.1. *Consider a system with M clients and N data points in total. Assume $x_{ij} \sim N(0, I_d)$ and $\xi_{ij} \sim N(0,1)$ independently for $i \in [M]$ and $j \in [n_i]$. Then for the model in (1.1), when $d \geq (1+\rho_1)k$ for a constant $\rho_1 > 0$, we have*

$$\inf_{\widehat{B} \in \mathcal{O}^{d\times k}} \sup_{B \in \mathcal{O}^{d\times k}} \sup_{\substack{n_1,\cdots,n_M \\ \sum_{i=1}^M n_i = N}} \sup_{\alpha \in \Psi_{\lambda_1,\lambda_k}^{n_1,\cdots,n_M}} \mathbb{E}\Big[\big\|\sin\Theta(\widehat{B}, B)\big\|\Big] = \Omega\bigg(\sqrt{\frac{d}{N\lambda_k}} \wedge 1\bigg).$$

Theorem C.2. *Consider a system with M clients and N data points in total. Suppose $x_{ij} \sim N(0, I_d)$ and $\xi_{ij} \sim N(0,1)$ independently for $i \in [M]$ and $j \in [n]$. For the model in (1.1), when $k = \Omega(\log M)$, $d \geq (1+\rho_1)k$, and $M \geq (1+\rho_2)k$ for constants $\rho_1, \rho_2 > 0$, we have*

$$\inf_{\widehat{B} \in \mathcal{O}^{d\times k}} \sup_{B \in \mathcal{O}^{d\times k}} \sup_{\substack{n_1,\cdots,n_M \\ \sum_{i=1}^M n_i = N}} \sup_{\alpha \in \Psi_{\lambda_1,\lambda_k}^{n_1,\cdots,n_M}} \mathbb{E}\Big[\big\|\sin\Theta(\widehat{B}, B)\big\|\Big] = \Omega\bigg(\sqrt{\frac{Md}{N^2\lambda_k^2}} \wedge 1\bigg).$$

Then Theorem 4.1 follows by combining these two theorems. We now prove the theorems, beginning with some preliminaries, including Fano's inequality.

Lemma C.1 (Fano's inequality)**.** *Let $X \to Y \to \widehat{X}$ be a Markov chain. Suppose that X is uniform over a finite set $\mathcal{X}$. Then we have*

$$\mathbb{P}(\widehat{X} \neq X) \geq 1 - \frac{I(X;Y) + \log 2}{\log|\mathcal{X}|},$$

where $I(X;Y) \triangleq \mathbb{E}_X\left[D_{\mathrm{KL}}(P_{Y|X} \| P_Y)\right]$ is the mutual information between X and Y.

The following lemma presents the KL-divergence between multivariate Gaussian distributions.

Lemma C.2. *Suppose P and Q are d-dimensional multivariate Gaussian distributions, where $P = N(\mu_1, \Sigma_1)$ and $Q = N(\mu_2, \Sigma_2)$, with $\mu_1, \mu_2 \in \mathbb{R}^d$ and nonsingular $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$. Then we have*

$$D_{\mathrm{KL}}(P\|Q) = \frac{1}{2}\bigg[\log\frac{|\Sigma_2|}{|\Sigma_1|} + \mathrm{Tr}\big(\Sigma_2^{-1}\Sigma_1 - I_d\big) + (\mu_2-\mu_1)^\intercal\Sigma_2^{-1}(\mu_2-\mu_1)\bigg].$$

The following lemma constructs $\mathcal{B}$ as a $(c\sqrt{\varepsilon})$-separated packing set of $\mathcal{O}^{d\times k}$ for a constant $c > 0$ and any $0 < \varepsilon < 1/2$. Let $\{e_r\}_{r=1}^k$ be the standard basis in $\mathbb{R}^k$.

Lemma C.3 (Packing set)**.** *There exists a constant $c > 0$ such that one can construct a packing set $\widetilde{\mathcal{B}} = \{b_1, \cdots, b_K\} \subset \mathcal{O}^{(d-k)\times 1}$ with $K \geq 10^{(d-k)}$ that is $(\sqrt{2}c)$-separated, i.e., $\|\sin\Theta(b_r, b_s)\| \geq \sqrt{2}c$ for any $r \neq s$.*

Moreover, given any $0 < \varepsilon < 1/2$, for $r \in [K]$, we define $B_r \in \mathbb{R}^{d\times k}$ as

$$B_r = \begin{pmatrix} e_1, & \cdots, & e_{k-1}, & \sqrt{1-\varepsilon}e_k \\ 0, & \cdots, & 0, & \sqrt{\varepsilon}b_r \end{pmatrix}.$$

Then the set $\mathcal{B} = \{B_1, \cdots, B_K\}$ forms a packing set of $\mathcal{O}^{d\times k}$ such that $\|\sin\Theta(B_r, B_s)\| \geq c\sqrt{\varepsilon}$ for any $r \neq s$.

Proof. First, [25] ensures the existence of a $(\sqrt{2}c)$-separated packing set $\widetilde{\mathcal{B}} \subset \mathcal{O}^{(d-k)\times 1}$ with cardinality $K = |\widetilde{\mathcal{B}}| \geq 10^{(d-k)}$. This verifies the first statement. We now study the properties of $\mathcal{B}$. For any $r \in [K]$, we have $B_r^\intercal B_r = I_k$; thus $\mathcal{B} \subset \mathcal{O}^{d\times k}$. In addition, we compute that

$$B_rB_r^\intercal = \begin{pmatrix} I_k - \varepsilon e_ke_k^\intercal & \sqrt{\varepsilon(1-\varepsilon)}e_kb_r^\intercal \\ \sqrt{\varepsilon(1-\varepsilon)}b_re_k^\intercal & \varepsilon b_rb_r^\intercal \end{pmatrix}.$$

Fix any $r \neq s$ and recall $\|\sin\Theta(B_r, B_s)\| = \|B_rB_r^\intercal - B_sB_s^\intercal\|$. Then by the definition of spectral norm, we have

$$\begin{aligned}\|\sin\Theta(B_r, B_s)\| &= \|B_rB_r^\intercal - B_sB_s^\intercal\| \\ &\geq \sqrt{\varepsilon(1-\varepsilon)}\|e_k(b_r-b_s)^\intercal\|\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{\varepsilon(1-\varepsilon)}\|b_r - b_s\| \\
&\geq \sqrt{\varepsilon(1-\varepsilon)}\|\sin\Theta(b_r, b_s)\| \\
&\geq \sqrt{\varepsilon(1-\varepsilon)} \times \sqrt{2}c \geq c\sqrt{\epsilon},
\end{aligned}
$$

where the first inequality holds because the spectral norm of a matrix is no smaller than that of its submatrix; the second equality holds because $e_k(b_r - b_s)^\intercal$ is rank-one; the second inequality holds because $\|b_r - b_s\| = 2\|\sin(\Theta(b_r, b_s)/2)\| \geq \|\sin\Theta(b_r, b_s)\|$; the last inequality holds because $0 < \varepsilon < 1/2$. □

C.1 Proof of Theorem C.1

We first prove Theorem C.1 by considering deterministic $\{\alpha_i\}$.

Proof of Theorem C.1. For any $\widehat{B}$, B, $\{n_i\}$, and α, Markov's inequality gives that

$$
\mathbb{E}\big[\big\|\sin\Theta(\widehat{B}, B)\big\|\big] \geq \frac{c\sqrt{\varepsilon}}{2}\mathbb{P}\Big(\big\|\sin\Theta(\widehat{B}, B)\big\| \geq \frac{c\sqrt{\varepsilon}}{2}\Big).
$$

Thus, to conclude the proof, it suffices to show that when $\varepsilon = \Theta(d/(N\lambda_k))$, the following holds,

$$
\inf_{\widehat{B}\in\mathcal{O}^{d\times k}} \sup_{B\in\mathcal{O}^{d\times k}} \sup_{\substack{n_1,\cdots,n_M \\ \sum_{i=1}^M n_i = N}} \sup_{\alpha\in\Psi_{\lambda_1,\lambda_k}^{n_1,\cdots,n_M}} \mathbb{P}\big(\big\|\sin\Theta(\widehat{B}, B)\big\| \geq c\sqrt{\varepsilon}/2\big) \geq 1/2. \tag{C.2}
$$

We will prove this for the remainder of the analysis. We take $\mathcal{B} = \{B_1, \cdots, B_K\} \subset \mathcal{O}^{d\times k}$ as the $(c\sqrt{\varepsilon})$-separated packing set given by Lemma C.3, where $K \geq 10^{(d-k)}$. Then we have $\log K \geq c_1 d$ for a constant c_1 since $d \geq (1+\rho_1)k$ for a constant $\rho_1 > 0$. We sample $B \sim \mathrm{Unif}(\mathcal{B})$. Then we take $n_i = n = N/M$ for all $i \in [M]$. The choice of α_i will be specified later.

Given a shared subspace $B \in \mathcal{O}^{d\times k}$, Assumption 2.3 implies $\theta_i^\star = B\alpha_i$ since $\Gamma_i = I_d$ for $i \in [M]$. Each client i observes n data points from the model in (1.1). Let $y_i = (y_{i1}; \cdots; y_{in}) \in \mathbb{R}^n$, $x_i = (x_{i1}, \cdots, x_{in}) \in \mathbb{R}^{d\times n}$, and $\xi_i = (\xi_{i1}; \cdots; \xi_{in}) \in \mathbb{R}^n$ be the concatenation of local variables at client i, and $Y = (y_1, \cdots, y_M)$ and $X = (x_1, \cdots, x_M)$ be the entire dataset. We have $y_i = x_i^\intercal B\alpha_i + \xi_i$.

Let $\mathbb{P}_{B,(X,Y)}(\cdot)$ denote the joint distribution of $(B, (X, Y))$, where $B \sim \mathrm{Unif}(\mathcal{B})$, $x_{ij} \sim N(0, I_d)$ independently for i and j, and Y is generated by the model in (1.1) given B, α, and X. For any $\widehat{B}$ and $\varepsilon > 0$, we lower-bound the supremum by an average and obtain

$$
\begin{aligned}
&\sup_{B\in\mathcal{O}^{d\times k}} \sup_{\substack{n_1,\cdots,n_M \\ \sum_{i=1}^M n_i = N}} \sup_{\alpha\in\Psi_{\lambda_1,\lambda_k}^{n_1,\cdots,n_M}} \mathbb{P}\big(\big\|\sin\Theta(\widehat{B}, B)\big\| \geq c\sqrt{\varepsilon}/2\big) \\
&\qquad \geq \mathbb{P}_{B,(X,Y)}\big(\big\|\sin\Theta(\widehat{B}, B)\big\| \geq c\sqrt{\varepsilon}/2\big).
\end{aligned} \tag{C.3}
$$

Let $\phi\colon \mathcal{O}^{d\times k} \to \mathcal{B}$ be a quantizer that maps any $B \in \mathcal{O}^{d\times k}$ to the closet point in $\mathcal{B}$. Recall that $\|\sin\Theta(B_r, B_s)\| \geq c\sqrt{\varepsilon}$ for any $r \neq s$. For any $B_i \in \mathcal{B}$, if $\phi(\widehat{B}) \neq B_i$, then we have $\|\sin\Theta(\widehat{B}, B_i)\| \geq c\sqrt{\varepsilon}/2$. Thus, we obtain

$$
\begin{aligned}
\mathbb{P}_{B,(X,Y)}\big(\big\|\sin\Theta(\widehat{B}, B)\big\| \geq c\sqrt{\varepsilon}/2\big) &\geq \mathbb{P}_{B,(X,Y)}\big(\phi(\widehat{B}) \neq B\big) \\
&\geq 1 - \frac{I(B; (X, Y)) + \log 2}{\log K},
\end{aligned} \tag{C.4}
$$

where the last inequality follows from Fano's inequality stated in Lemma C.1.

Next, we establish an upper bound for $I(B; (X, Y))$. Since $(x_1, y_1), \cdots, (x_M, y_M)$ are independent conditioned on B, we have

$$
I(B; (X, Y)) = I\big(B; (x_1, y_1), \cdots, (x_M, y_M)\big) \leq \sum_{i=1}^M I\big(B; (x_i, y_i)\big).
$$

It remains to choose α_i and bound the mutual information $I(B;(x_i,y_i))$ for a fixed i. We fix α_i for $i \in [M]$ such that $\alpha_i = O(1)$ and $\sum_{i=1}^M n\alpha_i\alpha_i^\intercal = N\lambda_k I_k$. Specifically, for each $i \in [M]$, we write i as $i = rk + s$, where $r, s \in \mathbb{Z}^+$, $r \le \lfloor M/k \rfloor$ and $s < k$. Let $\alpha_i = \sqrt{M\lambda_k/\lfloor M/k \rfloor} e_{s+1}$ if $r \le \lfloor M/k \rfloor$, and $\alpha_i = 0$ otherwise. In this way, since $0 < k\lambda_k = O(1)$ and $M \ge k$ by assumption, $\|\alpha_i\| = O(1)$ satisfies Assumption 2.4 and $\sum_{i=1}^M n\alpha_i\alpha_i^\intercal = N\lambda_k I_k$. Since α_i is assumed to be deterministic and $y_i = x_i^\intercal B\alpha_i + \xi_i$, we have

$$P_{y_i \,|\, x_i, B} = N(x_i^\intercal B\alpha_i, I_n).$$

Since B and x_i are independent, we have

$$\begin{aligned} I\big(B;(x_i,y_i)\big) &= \mathbb{E}_B \mathbb{E}_{x_i}\big[D_{\mathrm{KL}}(P_{y_i \,|\, x_i,B} \| P_{y_i \,|\, x_i})\big] \\ &= \mathbb{E}_B \mathbb{E}_{x_i}\big[D_{\mathrm{KL}}(P_{y_i \,|\, x_i,B} \| \mathbb{E}_{B'}[P_{y_i \,|\, x_i,B'}])\big] \\ &\le \mathbb{E}_{B'} \mathbb{E}_B \mathbb{E}_{x_i}\big[D_{\mathrm{KL}}(P_{y_i \,|\, x_i,B} \| P_{y_i \,|\, x_i,B'})\big], \end{aligned}$$

where the inequality follows from the convexity of KL-divergence. Combining the above two inequalities, we obtain

$$\begin{aligned} I\big(B;(X,Y)\big) &\le \mathbb{E}_{B'} \mathbb{E}_B \Big(\sum_{i=1}^M \mathbb{E}_{x_i}\big[D_{\mathrm{KL}}(P_{y_i \,|\, x_i,B} \| P_{y_i \,|\, x_i,B'})\big] \Big) \\ &\le \max_{B_r, B_s \in \mathcal{B}} \sum_{i=1}^M \mathbb{E}_{x_i}\big[D_{\mathrm{KL}}(P_{y_i \,|\, x_i,B_r} \| P_{y_i \,|\, x_i,B_s})\big]. \end{aligned} \tag{C.5}$$

We now compute the divergence for fixed $B_r \ne B_s$. Recall that $P_{y_i \,|\, x_i,B_r} = N(x_i^\intercal B_r\alpha_i, I_n)$. Since $\mathbb{E}[x_i x_i^\intercal] = \sum_{j=1}^n \mathbb{E}[x_{ij}x_{ij}^\intercal] = nI_d$, Lemma C.2 yields that

$$\begin{aligned} \mathbb{E}_{x_i}\big[D_{\mathrm{KL}}(P_{y_i \,|\, x_i,B_r} \| P_{y_i \,|\, x_i,B_s})\big] &= \frac{1}{2}\alpha_i^\intercal (B_r - B_s)^\intercal \mathbb{E}[x_i x_i^\intercal](B_r - B_s)\alpha_i \\ &= \frac{1}{2} n\alpha_i^\intercal (B_r - B_s)^\intercal (B_r - B_s)\alpha_i \\ &= \frac{1}{2}\mathrm{Tr}\big((B_r - B_s)^\intercal (B_r - B_s) \cdot n\alpha_i\alpha_i^\intercal\big). \end{aligned}$$

Let $\Delta_{r,s} = (B_r - B_s)^\intercal (B_r - B_s)/2 = I_k - B_r^\intercal B_s/2 - B_s^\intercal B_r/2$. For B_r, B_s defined in Lemma C.3, we compute that $B_r^\intercal B_s = B_s^\intercal B_r = \mathsf{diag}\,(1, \cdots, 1, 1 - \varepsilon + \varepsilon b_r^\intercal b_s)$. Thus, $\Delta_{r,s} = \mathsf{diag}\,(0, \cdots, 0, \varepsilon(1 - b_r^\intercal b_s))$. Substituting this into the above equation and recalling that $\sum_{i=1}^M n\alpha_i\alpha_i^\intercal = N\lambda_k I_k$, we have

$$\begin{aligned} \sum_{i=1}^M \mathbb{E}_{x_i}\big[D_{\mathrm{KL}}(P_{y_i \,|\, x_i,B_r} \| P_{y_i \,|\, x_i,B_s})\big] &= \mathrm{Tr}\Big(\Delta_{r,s} \cdot \Big(\sum_{i=1}^M n\alpha_i\alpha_i^\intercal\Big)\Big) \\ &= N\lambda_k \cdot \mathrm{Tr}\big(\mathsf{diag}\,(0, \cdots, 0, \varepsilon(1 - b_r^\intercal b_s))\big) \\ &= \varepsilon(1 - b_r^\intercal b_s) N\lambda_k \\ &\le 2\varepsilon N\lambda_k, \end{aligned}$$

where the last inequality holds since $-b_r^\intercal b_s \le \|b_r\|\|b_s\| = 1$. Substituting the above into (C.5), we obtain $I\big(B;(X,Y)\big) \le 2\varepsilon N\lambda_k$. Thus, when $\varepsilon > 0$ satisfies

$$\varepsilon = \frac{c_1 d}{6N\lambda_k} \wedge 1 = \Theta\Big(\frac{d}{N\lambda_k} \wedge 1\Big),$$

we have $I(B;(X,Y)) \le 2\varepsilon N\lambda_k = c_1 d/3$. Thus, following from (C.4), it holds that $\mathbb{P}_{B,\alpha,X}(\|\sin\Theta(\widehat{B}, B)\| \ge c\sqrt{\varepsilon}/2) \ge 1/2$ since $\log K \ge c_1 d$. Finally, substituting (C.4) into (C.3), we conclude the proof of (C.2) and thus the lemma. □

C.2 Proof of Theorem C.2

The proof of Theorem C.2 is similar to that of Theorem C.1 except for the choice of α_i and the way to bound the mutual information $I(B;(x_i,y_i))$. In particular, rather than a deterministic choice of $\{\alpha_i\}$, we will consider random $\{\alpha_i\}$. The following lemma shows the concentration of α for Gaussian-generated columns.

Lemma C.4. *Assume $k = \Omega(\log M)$ and $M \geq (1+\rho_2)k$ for a constant $\rho_2 > 0$, and fix $n_i = n = N/M$ for $i \in [M]$. When generating $\alpha_i \sim N(0, \lambda_k I_k)$ independently, we have*

$$\mathbb{P}\big(\alpha \in \Psi_{\lambda_1,\lambda_k}^{n,\cdots,n}\big) \geq 3/4.$$

Proof. By the definition of $\Psi_{\lambda_1,\lambda_k}^{n,\cdots,n}$ in (C.1), we have

$$\mathbb{P}\big(\alpha \in \Psi_{\lambda_1,\lambda_k}^{n,\cdots,n}\big) \geq \mathbb{P}\Big(\|\alpha_i\| = O(1), \forall i\Big) + \mathbb{P}\Big(\Omega(\lambda_k) I_k \preceq \frac{1}{N}\sum_{i=1}^{M} n\alpha_i\alpha_i^\intercal \preceq O(\lambda_1) I_k\Big) - 1.$$

By union bound,

$$\begin{aligned} 1 - \mathbb{P}\Big(\|\alpha_i\| = O(1), \forall i\Big) &= \mathbb{P}\Big(\exists i \in [M] : \|\alpha_i\| \neq O(1)\Big) \\ &\leq \sum_{i=1}^{M} \mathbb{P}\Big(\|\alpha_i\| \neq O(1)\Big) = M\mathbb{P}\Big(\|\alpha_1\| \neq O(1)\Big) \end{aligned}$$

Thus, to ensure $\mathbb{P}\Big(\|\alpha_i\| = O(1), \forall i\Big) \geq 7/8$, it suffices to show that $\mathbb{P}(\|\alpha_1\| \neq O(1)) \leq 1/(8M)$. Since $\alpha_1/\sqrt{\lambda_k}$ is a $k \times 1$ standard Gaussian matrix, using the concentration inequality [35, Theorem 7.3.3], we have for any $t \geq 0$, with probability at least $1 - 2\exp(-t^2/2)$,

$$\|\alpha_1\| \leq \sqrt{\lambda_k}(\sqrt{k} + 1 + t)$$

Thus, taking $t = \sqrt{2\log(16M)}$ we have with probability at least $1 - 1/(8M)$,

$$\|\alpha_1\| \leq \sqrt{\lambda_k}\left(\sqrt{k} + 1 + \sqrt{2\log(16M)}\right) \leq O(1),$$

where the last inequality holds due to the assumptions that $k\lambda_k = O(1)$ and $k = \Omega(\log M)$, so that $\lambda_k \log M = O(\log(M)/k) = O(1)$. Thus, we have

$$\mathbb{P}\Big(\|\alpha_1\| \neq O(1)\Big) \leq 1/(8M).$$

Now it remains to show that

$$\mathbb{P}\left(\Omega(\lambda_k) I_k \preceq \frac{1}{N}\sum_{i=1}^{M} n\alpha_i\alpha_i^\intercal \preceq O(\lambda_1) I_k\right) \geq 7/8.$$

Let $\sigma_r(\cdot)$ be the r-th largest singular value of a matrix. Since $\lambda_r(\sum_{i=1}^{M} n\alpha_i\alpha_i^\intercal) = n\sigma_r^2(\alpha)$, it reduces to proving that

$$\mathbb{P}\left(\Omega(\sqrt{M\lambda_k}) \leq \sigma_k(\alpha) \leq \sigma_1(\alpha) \leq O(\sqrt{M\lambda_1})\right) \geq 7/8.$$

We proceed by bounding $\sigma_1(\alpha)$ and $\sigma_k(\alpha)$. Using the concentration properties of the standard Gaussian matrix $\alpha/\sqrt{\lambda_k}$ [35, Theorem 7.3.3], we have for any $t \geq 0$, with probability at least $1 - 2\exp(-t^2/2)$,

$$\sqrt{\lambda_k}(\sqrt{M} - \sqrt{k} - t) \leq \sigma_k(\alpha) \leq \sigma_1(\alpha) \leq \sqrt{\lambda_k}(\sqrt{M} + \sqrt{k} + t).$$

Thus, by picking $t = \sqrt{2\log(16)}$, we have, with probability at least $7/8$,

$$\sqrt{\lambda_k}(\sqrt{M} - \sqrt{k} - \sqrt{2\log(16)}) \leq \sigma_k(\alpha) \leq \sigma_1(\alpha) \leq \sqrt{\lambda_k}(\sqrt{M} + \sqrt{k} + \sqrt{2\log(16)}).$$

Finally, since $M \geq (1+\rho_2)k$ for a constant $\rho_2 > 0$,

$$\begin{aligned} \sqrt{M} - \sqrt{k} - \sqrt{2\log(16)} &\geq \Omega(\sqrt{M}) \\ \sqrt{M} + \sqrt{k} + \sqrt{2\log(16)} &\leq O(\sqrt{M}). \end{aligned}$$

This concludes the proof of the lemma. □

We now prove Theorem C.2 by following similar steps as in the proof of Theorem C.1, but with Gaussian-generated α.

Proof of Theorem C.2. Similar to the proof of Theorem C.1, we only need to show that when $\varepsilon = \Theta(Md/(N^2\lambda_k^2))$, the following holds:

$$\inf_{\widehat{B}\in\mathcal{O}^{d\times k}} \sup_{B\in\mathcal{O}^{d\times k}} \sup_{\substack{n_1,\cdots,n_M \\ \sum_{i=1}^M n_i = N}} \sup_{\alpha\in\Psi_{\lambda_1,\lambda_k}^{n_1,\cdots,n_M}} \mathbb{P}\big(\big\|\sin\Theta(\widehat{B},B)\big\| \geq c\sqrt{\varepsilon}/2\big) \geq 1/2. \tag{C.6}$$

We generate $\alpha_i \sim N(0,\lambda_k I_k)$ independently for $i\in[M]$. Let $\mathcal{B} = \{B_1,\cdots,B_K\} \subset \mathcal{O}^{d\times k}$ be the $(c\sqrt{\varepsilon})$-separated packing set given by Lemma C.3, where $\log K \geq c_1 d$ for a constant c_1. We sample $B\sim \mathrm{Unif}(\mathcal{B})$. Let $\mathbb{P}_{B,\alpha,(X,Y)}(\cdot)$ denote the joint distribution of $(B,\alpha,(X,Y))$ where $B\sim\mathrm{Unif}(\mathcal{B})$, $\alpha_i\sim N(0,\lambda_k I_k)$ independently for i, $x_{ij}\sim N(0,I_d)$ independently for i and j, and Y is generated by the model in (1.1) given B, α, and X. For any $\widehat{B}$ and $\varepsilon>0$, we lower-bound the supremum and obtain

$$\begin{aligned}
&\sup_{B\in\mathcal{O}^{d\times k}} \sup_{\substack{n_1,\cdots,n_M \\ \sum_{i=1}^M n_i = N}} \sup_{\alpha\in\Psi_{\lambda_1,\lambda_k}^{n_1,\cdots,n_M}} \mathbb{P}\big(\big\|\sin\Theta(\widehat{B},B)\big\| \geq c\sqrt{\varepsilon}/2\big) \\
&\quad\geq \mathbb{P}_{B,\alpha,(X,Y)}\big(\big\|\sin\Theta(\widehat{B},B)\big\| \geq c\sqrt{\varepsilon}/2 \,\big|\, \alpha\in\Omega_{\lambda_1,\lambda_k}^{n,\cdots,n}\big) \\
&\quad\geq \mathbb{P}_{B,\alpha,(X,Y)}\big(\big\|\sin\Theta(\widehat{B},B)\big\| \geq c\sqrt{\varepsilon}/2, \alpha\in\Omega_{\lambda_1,\lambda_k}^{n,\cdots,n}\big) \\
&\quad\geq \mathbb{P}_{B,\alpha,(X,Y)}\big(\big\|\sin\Theta(\widehat{B},B)\big\| \geq c\sqrt{\varepsilon}/2\big) + \mathbb{P}\big(\alpha\in\Omega_{\lambda_1,\lambda_k}^{n,\cdots,n}\big) - 1,
\end{aligned} \tag{C.7}$$

where the inequalities hold since for any events $\mathcal{E}$ and $\mathcal{A}$, we have $\mathbb{P}(\mathcal{E}\,|\,\mathcal{A}) = \mathbb{P}(\mathcal{E}\cap\mathcal{A})/\mathbb{P}(\mathcal{A}) \geq \mathbb{P}(\mathcal{E}\cap\mathcal{A})$ and $\mathbb{P}(\mathcal{E}\cap\mathcal{A}) \geq \mathbb{P}(\mathcal{E}) + \mathbb{P}(\mathcal{A}) - 1$. Lemma C.4 gives that $\mathbb{P}\big(\alpha\in\Omega_{\lambda_1,\lambda_k}^{n,\cdots,n}\big) - 1 \geq -1/4$. Thus, it remains to lower bound the first term of (C.7). Similar to the proof of Theorem C.1, using Fano's inequality, we have

$$\mathbb{P}_{B,\alpha,(X,Y)}\big(\big\|\sin\Theta(\widehat{B},B)\big\| \geq c\sqrt{\varepsilon}/2\big) \geq 1 - \frac{I(B;(X,Y)) + \log 2}{\log K}, \tag{C.8}$$

where the independence of $(x_1,y_1),\cdots,(x_M,y_M)$ gives that $I(B;(X,Y)) \leq \sum_{i=1}^M I(B;(x_i,y_i))$. We now bound $I\big(B;(x_i,y_i)\big)$ for a fixed i. Given any $B_r\in\mathcal{B}$ and recalling $\alpha_i\sim N(0,\lambda_k I_k)$, the model $y_i = x_i^\intercal B_r\alpha_i + \xi_i$ implies that $P_{y_i\,|\,x_i,B_r} = N(0,\Sigma_{ir})$ with

$$\Sigma_{ir} = \lambda_k x_i^\intercal B_r B_r^\intercal x_i + I_n, \quad \text{where } B_r B_r^\intercal = \begin{pmatrix} I_k - \varepsilon e_k e_k^\intercal & \sqrt{\varepsilon(1-\varepsilon)} e_k b_r^\intercal \\ \sqrt{\varepsilon(1-\varepsilon)} b_r e_k^\intercal & \varepsilon b_r b_r^\intercal \end{pmatrix}.$$

We define $Q_{\cdot\,|\,x_i} = N(0,\Sigma_{Q_i})$ with Σ_{Q_i} shown as follows

$$\Sigma_{Q_i} = \lambda_k x_i^\intercal \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix}_{d\times d} x_i + I_n.$$

Since B and x_i are independent, we can bound $I(B;(x_i,y_i))$ as follows,

$$\begin{aligned}
I\big(B;(x_i,y_i)\big) &= \mathbb{E}_B\mathbb{E}_{x_i}\big[D_{\mathrm{KL}}(P_{y_i\,|\,x_i,B}\|P_{y_i\,|\,x_i})\big] \\
&= \mathbb{E}_B\mathbb{E}_{x_i}\big[D_{\mathrm{KL}}(P_{y_i\,|\,x_i,B}\|Q_{\cdot\,|\,x_i})\big] - D_{\mathrm{KL}}(P_{y_i\,|\,x_i}\|Q_{\cdot\,|\,x_i}) \\
&\leq \mathbb{E}_B\mathbb{E}_{x_i}\big[D_{\mathrm{KL}}(P_{y_i\,|\,x_i,B}\|Q_{\cdot\,|\,x_i})\big] \\
&\leq \max_{B_r\in\mathcal{B}} \mathbb{E}_{x_i}\big[D_{\mathrm{KL}}(P_{y_i\,|\,x_i,B_r}\|Q_{\cdot\,|\,x_i})\big].
\end{aligned} \tag{C.9}$$

We now compute $\mathbb{E}_{x_i}[D_{\mathrm{KL}}(P_{y_i\,|\,x_i,B_r}\|Q_{\cdot\,|\,x_i})]$ for a fixed B_r. Lemma C.2 yields that

$$\mathbb{E}_{x_i}\big[D_{\mathrm{KL}}(P_{y_i\,|\,x_i,B_r}\|Q_{\cdot\,|\,x_i})\big] = \frac{1}{2}\mathbb{E}_{x_i}\Big[\log\frac{|\Sigma_{Q_i}|}{|\Sigma_{ir}|} + \mathrm{Tr}\big(\Sigma_{Q_i}^{-1}\Sigma_{ir} - I_n\big)\Big].$$

Because of the non-negativity of the KL divergence: $2D_{\mathrm{KL}}(Q_{\cdot\,|\,x_i}\|P_{y_i\,|\,x_i,B_r}) = \log\frac{|\Sigma_{ir}|}{|\Sigma_{Q_i}|} + \mathrm{Tr}\big(\Sigma_{ir}^{-1}\Sigma_{Q_i} - I_n\big) \geq 0$, we have $\log\frac{|\Sigma_{Q_i}|}{|\Sigma_{ir}|} \leq \mathrm{Tr}\big(\Sigma_{ir}^{-1}\Sigma_{Q_i} - I_n\big)$. Thus, we can bound the above equation as

$$\mathbb{E}_{x_i}\big[D_{\mathrm{KL}}(P_{y_i\,|\,x_i,B_r}\|Q_{\cdot\,|\,x_i})\big] \leq \frac{1}{2}\mathbb{E}_{x_i}\Big[\mathrm{Tr}\big(\Sigma_{ir}^{-1}\Sigma_{Q_i} + \Sigma_{Q_i}^{-1}\Sigma_{ir} - 2I_n\big)\Big].$$

Let $\Delta = \Sigma_{ir} - \Sigma_{Q_i}$. Since $\Sigma_{ir} \succeq I_n$ and $\Sigma_{Q_i} \succeq I_n$, we have

$$\begin{aligned}\operatorname{Tr}\big(\Sigma_{ir}^{-1}\Sigma_{Q_i} + \Sigma_{Q_i}^{-1}\Sigma_{ir} - 2I_n\big) &= \operatorname{Tr}\Big(\Sigma_{ir}^{-1}\big(\Sigma_{Q_i} + \Sigma_{ir}\Sigma_{Q_i}^{-1}\Sigma_{ir} - 2\Sigma_{ir}\big)\Big)\\ &\le \operatorname{Tr}\big(\Sigma_{Q_i} + \Sigma_{ir}\Sigma_{Q_i}^{-1}\Sigma_{ir} - 2\Sigma_{ir}\big)\\ &= \operatorname{Tr}\big(\Delta\Sigma_{Q_i}^{-1}\Delta\big) = \operatorname{Tr}\big(\Sigma_{Q_i}^{-1}\Delta^2\big) \le \operatorname{Tr}(\Delta^2).\end{aligned}$$

Combining the above two equations, we have

$$\mathbb{E}_{x_i}\big[D_{\mathrm{KL}}(P_{y_i\,|\,x_i,B_r}\|Q_{\cdot\,|\,x_i})\big] \le \frac{1}{2}\mathbb{E}_{x_i}\big[\operatorname{Tr}(\Delta^2)\big] = \frac{1}{2}\operatorname{Tr}\big(\mathbb{E}_{x_i}(\Delta^2)\big). \tag{C.10}$$

We split $x_i = (a_i; u_i)$ with $a_i = (a_{i1}, \cdots, a_{in}) \in \mathbb{R}^{k\times n}$ and $u_i = (u_{i1}, \cdots, u_{in}) \in \mathbb{R}^{(d-k)\times n}$. Then $a_{ij} \sim N(0, I_k)$, $u_{ij} \sim N(0, I_{d-k})$, and all $\{a_{ij}\}$ and $\{u_{ij}\}$ are mutually independent. Note that

$$\begin{aligned}\Delta &= \lambda_k \begin{pmatrix} a_i^\intercal & u_i^\intercal\end{pmatrix}\begin{pmatrix} -\varepsilon e_k e_k^\intercal & \sqrt{\varepsilon(1-\varepsilon)}e_k b_r^\intercal \\ \sqrt{\varepsilon(1-\varepsilon)}b_r e_k^\intercal & \varepsilon b_r b_r^\intercal\end{pmatrix}\begin{pmatrix} a_i^\intercal \\ u_i^\intercal\end{pmatrix}\\ &= \lambda_k(-\varepsilon a_i^\intercal e_k e_k^\intercal a_i + \sqrt{\varepsilon(1-\varepsilon)}u_i^\intercal b_r e_k^\intercal a_i + \sqrt{\varepsilon(1-\varepsilon)}a_i^\intercal e_k b_r^\intercal u_i + \varepsilon u_i^\intercal b_r b_r^\intercal u_i).\end{aligned}$$

Let $\widetilde{a}_i = a_i^\intercal e_k$ and $\widetilde{u}_i = u_i^\intercal b_r$. We have $\widetilde{a}_i \sim N(0, I_n)$ and $\widetilde{u}_i \sim N(0, I_n)$. By the symmetry and independence, we obtain

$$\mathbb{E}_{x_i}(\Delta^2) = 2\lambda_k^2\mathbb{E}_{x_i}\big[\varepsilon^2\widetilde{a}_i\widetilde{a}_i^\intercal\widetilde{a}_i\widetilde{a}_i^\intercal - \varepsilon^2\widetilde{a}_i\widetilde{a}_i^\intercal\widetilde{u}_i\widetilde{u}_i^\intercal + (1-\varepsilon)\varepsilon\widetilde{u}_i\widetilde{a}_i^\intercal\widetilde{u}_i\widetilde{a}_i^\intercal + (1-\varepsilon)\varepsilon\widetilde{u}_i\widetilde{a}_i^\intercal\widetilde{a}_i\widetilde{u}_i^\intercal\big]. \tag{C.11}$$

By the linearity, we compute the trace of each term above. Note that $\widetilde{a}_i^\intercal\widetilde{a}_i \sim \chi^2(n)$. We first have

$$\operatorname{Tr}\big(\mathbb{E}_{x_i}\big(\widetilde{a}_i\widetilde{a}_i^\intercal\widetilde{a}_i\widetilde{a}_i^\intercal\big)\big) = \mathbb{E}_{a_i}\big[(\widetilde{a}_i^\intercal\widetilde{a}_i)^2\big] = \operatorname{Var}\big(\widetilde{a}_i^\intercal\widetilde{a}_i\big) + \big(\mathbb{E}_{a_i}\big[\widetilde{a}_i^\intercal\widetilde{a}_i\big]\big)^2 = n^2 + 2n.$$

For the second term, we have

$$\operatorname{Tr}\big(\mathbb{E}_{x_i}\big(\widetilde{a}_i\widetilde{a}_i^\intercal\widetilde{u}_i\widetilde{u}_i^\intercal\big)\big) = \operatorname{Tr}\big([\mathbb{E}_{a_i}(\widetilde{a}_i\widetilde{a}_i^\intercal)]\cdot[\mathbb{E}_{u_i}(\widetilde{u}_i\widetilde{u}_i^\intercal)]\big) = \operatorname{Tr}(I_n) = n.$$

For the third term, we have

$$\begin{aligned}\operatorname{Tr}\big(\mathbb{E}_{x_i}\big(\widetilde{u}_i\widetilde{a}_i^\intercal\widetilde{u}_i\widetilde{a}_i^\intercal\big)\big) &= \operatorname{Tr}\big(\mathbb{E}_{a_i}\big(\mathbb{E}_{u_i}\big(\widetilde{a}_i^\intercal\widetilde{u}_i\widetilde{a}_i^\intercal\widetilde{u}_i \,\big|\, \widetilde{a}_i\big)\big)\big)\\ &= \operatorname{Tr}\big(\mathbb{E}_{a_i}\big(\operatorname{Var}_{u_i}\big(\widetilde{a}_i^\intercal\widetilde{u}_i \,\big|\, \widetilde{a}_i\big)\big)\big) = \operatorname{Tr}\big(\mathbb{E}_{a_i}\big(\widetilde{a}_i^\intercal\widetilde{a}_i\big)\big) = n.\end{aligned}$$

Then we compute the last term as follows,

$$\operatorname{Tr}\big(\mathbb{E}_{x_i}\big(\widetilde{u}_i\widetilde{a}_i^\intercal\widetilde{a}_i\widetilde{u}_i^\intercal\big)\big) = \operatorname{Tr}\big(\mathbb{E}_{u_i}\big(\mathbb{E}_{a_i}\big(\widetilde{u}_i\widetilde{a}_i^\intercal\widetilde{a}_i\widetilde{u}_i^\intercal \,\big|\, \widetilde{u}_i\big)\big)\big) = n\operatorname{Tr}\big(\mathbb{E}_{u_i}\big(\widetilde{u}_i\widetilde{u}_i^\intercal\big)\big) = n\operatorname{Tr}(I_n) = n^2.$$

Substituting the above four terms into (C.11), we have

$$\operatorname{Tr}\big(\mathbb{E}_{x_i}(\Delta^2)\big) = 2\lambda_k^2\big[\varepsilon^2(n^2+2n) - \varepsilon^2 n + (1-\varepsilon)\varepsilon n + (1-\varepsilon)\varepsilon n^2\big] = 2\varepsilon\lambda_k^2(n^2+n).$$

Combining the above with (C.9) and (C.10), we have

$$I\big(B; (x_i, y_i)\big) \le \varepsilon\lambda_k^2(n^2+n).$$

Recall that $N = Mn$. When $\varepsilon > 0$ satisfies

$$\varepsilon = \frac{c_1 d}{5M\lambda_k^2(n^2+n)} \wedge 1 = \Theta\Big(\frac{Md}{N^2\lambda_k^2} \wedge 1\Big),$$

we bound the mutual information by $I(B; (X, Y)) \le \sum_{i=1}^M I(B; (x_i, y_i)) \le \varepsilon M\lambda_k^2(n^2+n) = c_1 d/5$. Recall that $\log K \ge c_1 d$. Thus, (C.8) yields that

$$\mathbb{P}_{B,\alpha,X}(\|\sin\Theta(\widehat{B}, B)\| \ge c\sqrt{\varepsilon}/2) \ge 3/4. \tag{C.12}$$

Finally, by substituting Lemma C.4 and (C.12) into (C.7), we conclude the proof of (C.6) and the theorem. □

D Proofs of Corollaries in Section 5

Theorem 4 from [33] shows the error rate of learning α_{M+1}.

Theorem D.1 (Theorem 4 from [33]). *Suppose that Assumptions 2.1-2.4 hold and* $\lambda_1 = \Theta(\lambda_k) = \Theta(1/k)$. *If* $\|\sin\Theta(\widehat{B}, B^\star)\| \le \delta$ *and* $n_{M+1} \ge k \log n_{M+1}$, *then with probability at least* $1 - O(n_{M+1}^{-100})$, *the output* $\widehat{\alpha}_{M+1}$ *given by (5.1) satisfies*

$$\|\widehat{B}\widehat{\alpha}_{M+1} - B^\star \alpha^\star_{M+1}\|^2 = \widetilde{O}\bigg(\delta^2 + \frac{k}{n_{M+1}}\bigg).$$

Proof of Corollary 5.1. By substituting $\delta = \widetilde{O}(\sqrt{dk/N} + \sqrt{Mdk^2/N^2})$ given by Corollary 3.1 into Theorem D.1, we conclude the proof of Corollary 5.1. □

Theorem 5.4 from [29] gives the error rate for learning a differentially private estimator of α_{M+1}. In particular, let $L(\theta) = \mathbb{E}_{(x,y)}[(x^\intercal\theta - y)^2]/2$ be the population risk at client $M+1$, where (x, y) is from the model in (1.1) with $\theta^\star_{M+1}$.

Theorem D.2 (Theorem 5.4 from [29]). *Suppose that Assumptions 2.1-2.4 hold and* $\lambda_1 = \Theta(\lambda_k) = \Theta(1/k)$. *In addition,* $x_{ij} \sim N(0, I_d)$. *If* $\|\sin\Theta(\widehat{B}, B^\star)\| \le \delta$, *then there exists an* (ε, δ)*-differentially private estimator* $\widehat{\alpha}^\varepsilon_{M+1}$ *such that, with high probability,*

$$L(\widehat{B}\widehat{\alpha}^\varepsilon_{M+1}) - L(B^\star\alpha^\star_{M+1}) = \widetilde{O}\bigg(\frac{k}{n_{M+1}} + \frac{k^2\log(1/\delta)}{n_{M+1}^2\varepsilon^2}\bigg) + \delta^2.$$

Proof of Corollary 5.2. We first show that $L(\theta) - L(\theta^\star_{M+1}) = \|\theta - \theta^\star_{M+1}\|^2/2$ for any $\theta \in \mathbb{R}^d$. By substituting $y = x^\intercal\theta^\star_{M+1} + \xi$ into $L(\theta)$, since ξ and x are independent, and $\mathbb{E}[xx^\intercal] = I_d$, we have

$$\begin{aligned} L(\theta) &= \mathbb{E}_{(x,y)}[(x^\intercal\theta - y)^2]/2 = \mathbb{E}_{(x,y)}[(x^\intercal\theta - x^\intercal\theta^\star_{M+1} - \xi)^2]/2 \\ &= (\theta - \theta^\star_{M+1})^\intercal\mathbb{E}[xx^\intercal](\theta - \theta^\star_{M+1})/2 + \mathbb{E}[\xi^2]/2 \\ &= \|\theta - \theta^\star_{M+1}\|^2/2 + \mathbb{E}[\xi^2]/2. \end{aligned}$$

We further have $L(\theta^\star_{M+1}) = \mathbb{E}[\xi^2]/2$, and thus,

$$L(\theta) - L(\theta^\star_{M+1}) = \|\theta - \theta^\star_{M+1}\|^2/2.$$

Therefore, by substituting $\delta = \widetilde{O}(\sqrt{dk/N} + \sqrt{Mdk^2/N^2})$ given by Corollary 3.1 into Theorem D.2, we conclude the proof of Corollary 5.2. □