Subspace Recovery from Heterogeneous Data with Non-isotropic Noise

John Duchi
Stanford University
jduchi@stanford.edu

Vitaly Feldman
Apple
vitaly.edu@gmail.com

Lunjia Hu
Stanford University
lunjia@stanford.edu

Kunal Talwar
Apple
kunal@kunaltalwar.org

Abstract

Recovering linear subspaces from data is a fundamental and important task in statistics and machine learning. Motivated by heterogeneity in Federated Learning settings, we study a basic formulation of this problem: the principal component analysis (PCA), with a focus on dealing with irregular noise. Our data come from n users with user i contributing data samples from a d-dimensional distribution with mean μ_i. Our goal is to recover the linear subspace shared by $\mu_1, \ldots, \mu_n$ using the data points from all users, where every data point from user i is formed by adding an independent mean-zero noise vector to μ_i. If we only have one data point from every user, subspace recovery is information-theoretically impossible when the covariance matrices of the noise vectors can be non-spherical, necessitating additional restrictive assumptions in previous work. We avoid these assumptions by leveraging at least two data points from each user, which allows us to design an efficiently-computable estimator under non-spherical and user-dependent noise. We prove an upper bound for the estimation error of our estimator in general scenarios where the number of data points and amount of noise can vary across users, and prove an information-theoretic error lower bound that not only matches the upper bound up to a constant factor, but also holds even for spherical Gaussian noise. This implies that our estimator does not introduce additional estimation error (up to a constant factor) due to irregularity in the noise. We show additional results for a linear regression problem in a similar setup.

1 Introduction

We study the problem of learning low-dimensional structure amongst data distributions, given multiple samples from each distribution. This problem arises naturally in settings such as federated learning, where we want to learn from data coming from a set of individuals, each of which has samples from their own distribution. These distributions however are related to each other, and in this work, we consider the setting when these distributions have means lying in a low-dimensional subspace. The goal is to learn this subspace, even when the distributions may have different (and potentially non-spherical) variances. This heterogeneity can manifest itself in practice as differing number of samples per user, or the variance differing across individuals, possibly depending on their mean. Recovery of the subspace containing the means can in turn help better estimate individual means. In other words, this can allow for learning good estimator for all individual means, by leveraging information from all the individuals.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

The irregularity of the noise makes this task challenging even when we have sufficiently many individual distributions. For example, suppose we have n individuals and for every $i = 1, \ldots, n$, an unknown $\mu_i \in \mathbb{R}^d$. For simplicity, suppose that $\mu_1, \ldots, \mu_n$ are distributed independently as $N(0, \sigma^2 uu^\mathsf{T})$ for $\sigma \in \mathbb{R}_{\geq 0}$ and an unknown unit vector $u \in \mathbb{R}^d$. In this setting, our goal is to recover the one-dimensional subspace, equivalently the vector u. For every i, we have a data point $x_i = \mu_i + z_i$ where $z_i \in \mathbb{R}^d$ is a mean-zero noise vector. If z_i is drawn independently from a spherical Gaussian $N(0, \alpha^2 I)$, we can recover the unknown subspace with arbitrary accuracy as n grows to infinity because $\frac{1}{n}\sum x_i x_i^\mathsf{T}$ concentrates to $\mathbb{E}[x_i x_i^\mathsf{T}] = \sigma^2 uu^\mathsf{T} + \alpha^2 I$, whose top eigenvector is $\pm u$. However, if the noise z_i is drawn from a non-spherical distribution, the top eigenvector of $\frac{1}{n}\sum x_i x_i^\mathsf{T}$ can deviate from $\pm u$ significantly, and to make things worse, if the noise z_i is drawn independently from a non-spherical Gaussian $N(0, \sigma^2(I - uu^\mathsf{T}) + \alpha^2 I)$, then our data points $x_i = \mu_i + z_i$ distribute independently as $N(0, (\sigma^2 + \alpha^2)I)$, giving no information about the vector u.[1]

The information-theoretic impossibility in this example however disappears as soon as one has at least two samples from each distribution. Indeed, given two data points $x_{i1} = \mu_i + z_{i1}$ and $x_{i2} = \mu_i + z_{i2}$ from user i, as long as the noise z_{i1}, z_{i2} are independent and have zero mean, we always have $\mathbb{E}[x_{i1} x_{i2}^\mathsf{T}] = \sigma^2 uu^\mathsf{T}$ regardless of the specific distributions of z_{i1} and z_{i2}. This allows us to recover the subspace in this example, as long as we have sufficiently many users each contributing at least two examples.

As this is commonly the case in our motivating examples, we make this assumption of multiple data points per user, and show that this intuition extends well beyond this particular example. We design efficiently computable estimators for this subspace recovery problem given samples from multiple heteroscedastic distributions (see Section 1.1 for details). We prove upper bounds on the error of our estimator measured in the *maximum principal angle* (see Section 2 for definition). We also prove an information-theoretic error lower bound, showing that our estimator achieves the optimal error up to a constant factor in general scenarios where the number of data points and the amount of noise can vary across users. Somewhat surprisingly, our lower bound holds even when the noise distributes as spherical Gaussians. Thus non-spherical noise in setting does not lead to increased error.

We then show that our techniques extend beyond the mean estimation problem to a linear regression setting where for each μ_i, we get (at least two) samples $(x_{ij}, x_{ij}^\mathsf{T}\mu_i + z_{ij})$ where z_{ij} is zero-mean noise from some noise distribution that depends on i and x_{ij}. This turns out to be a model that was recently studied in the meta-learning literature under more restrictive assumptions (e.g. z_{ij} is independent of x_{ij}) [Kong et al., 2020, Tripuraneni et al., 2021, Collins et al., 2021, Thekumparampil et al., 2021]. We show a simple estimator achieving an error upper bound matching the ones in prior work without making these restrictive assumptions.

1.1 Our contributions

PCA with heterogeneous and non-isotropic noise: Upper Bounds. In the PCA setting, the data points from each user i are drawn from a user-specific distribution with mean $\mu_i \in \mathbb{R}^d$, and we assume that $\mu_1, \ldots, \mu_n$ lie in a shared k-dimensional subspace that we want to recover. Specifically, we have m_i data points $x_{ij} \in \mathbb{R}^d$ from user i for $j = 1, \ldots, m_i$, and each data point is determined by $x_{ij} = \mu_i + z_{ij}$ where $z_{ij} \in \mathbb{R}^d$ is a noise vector drawn independently from a mean zero distribution. We allow the distribution of z_{ij} to be non-spherical and non-identical across different pairs (i, j). We use $\eta_i \in \mathbb{R}_{\geq 0}$ to quantify the amount of noise in user i's data points by assuming that z_{ij} is an η_i-*sub-Gaussian* random variable.

As mentioned earlier, if we only have a single data point from each user, it is information-theoretically impossible to recover the subspace. Thus, we focus on the case where $m_i \geq 2$ for every $i = 1, \ldots, n$. In this setting, for appropriate weights $w_1, \ldots, w_n \in \mathbb{R}_{\geq 0}$, we compute a matrix A:

$$A = \sum_{i=1}^{n} \frac{w_i}{m_i(m_i - 1)} \sum_{j_1 \neq j_2} x_{ij_1} x_{ij_2}^\mathsf{T}, \tag{1}$$

where the inner summation is over all pairs $j_1, j_2 \in \{1, \ldots, m_i\}$ satisfying $j_1 \neq j_2$. Our estimator is then defined by the subspace spanned by the top-k eigenvectors of A. Although the inner summation

[1] This information-theoretic impossibility naturally extends to recovering k-dimensional subspaces for $k > 1$ by replacing the unit vector $u \in \mathbb{R}^d$ with a matrix $U \in \mathbb{R}^{d \times k}$ with orthonormal columns.

is over $m_i(m_i - 1)$ terms, the time complexity for computing it need not grow quadratically with m_i because of the following equation:

$$\sum_{j_1 \neq j_2} x_{ij_1} x_{ij_2}^\mathsf{T} = \left(\sum_{j=1}^{m_i} x_{ij} \right) \left(\sum_{j=1}^{m_i} x_{ij} \right)^\mathsf{T} - \sum_{j=1}^{m_i} x_{ij} x_{ij}^\mathsf{T}.$$

The flexibility in the weights $w_1, \ldots, w_n$ allows us to deal with variations in m_i and η_i for different users i. In the special case where $\eta_1 = \cdots = \eta_n = \eta$ and $m_1 = \cdots = m_n = m$, we choose $w_1 = \cdots = w_n = 1/n$ and we show that our estimator achieves the following error upper bound with success probability at least $1 - \delta$:

$$\sin\theta = O\left(\left(\frac{\eta\sigma_1}{\sigma_k^2\sqrt{m}} + \frac{\eta^2}{\sigma_k^2 m} \right) \sqrt{\frac{d + \log(1/\delta)}{n}} \right).$$

Here, θ is the maximum principal angle between our estimator and the true subspace shared by $\mu_1, \ldots, \mu_n$, and σ_ℓ^2 is the ℓ-th largest eigenvalue of $\frac{1}{n}\sum_{i=1}^n \mu_i \mu_i^\mathsf{T}$. Our error upper bound for general m_i, η_i, w_i is given in Theorem 3.1.

We instantiate our error upper bound to the case where $\mu_1, \ldots, \mu_n$ are drawn iid from a Gaussian distribution $N(0, \sigma^2 UU^\mathsf{T})$, where the columns of $U \in \mathbb{R}^{d \times k}$ form an orthonormal basis of the subspace containing $\mu_1, \ldots, \mu_n$. By choosing the weights $w_1, \ldots, w_n$ according to $m_1, \ldots, m_n$ and $\eta_1, \ldots, \eta_n$, our estimator achieves the error upper bound

$$\sin\theta \leq O\left(\sqrt{\frac{d + \log(1/\delta)}{\sum_{i=1}^n \gamma_i'}} \right) \tag{2}$$

under a mild assumption (Assumption 3.2), where γ_i' is defined in Definition 3.1 and often equals $\left(\frac{\eta_i^2}{\sigma^2 m_i} + \frac{\eta_i^4}{\sigma^4 m_i^2} \right)^{-1}$.

PCA: Lower Bounds. We show that the error upper bound (2) is optimal up to a constant factor by proving a matching information-theoretic lower bound (Theorem 3.7). Our lower bound holds for general m_i and η_i that can vary among users i, and it holds even when the noise vectors z_{ij} are drawn from spherical Gaussians, showing that our estimator essentially pays no additional cost in error or sample complexity due to non-isotropic noise.

We prove the lower bound using Fano's method on a local packing over the Grassmannian manifold. We carefully select a non-trivial hard distribution so that the strength of our lower bound is not affected by a group of fewer than k users each having a huge amount of data points with little noise.

Linear Models. While the PCA setting is the main focus of our paper, we extend our research to a related linear models setting that has recently been well studied in the meta-learning and federated learning literature [Kong et al., 2020, Tripuraneni et al., 2021, Collins et al., 2021, Thekumparampil et al., 2021]. Here, the user-specific distribution of each user i is parameterized by $\beta_i \in \mathbb{R}^d$, and we again assume that $\beta_1, \ldots, \beta_n$ lie in a linear subspace that we want to recover. From each user i we observe m_i data points $(x_{ij}, y_{ij}) \in \mathbb{R}^d \times \mathbb{R}$ for $j = 1, \ldots, m_i$ drawn from the user-specific distribution satisfying $y_{ij} = x_{ij}^\mathsf{T}\beta_i + z_{ij}$ for an $O(1)$-sub-Gaussian measurement vector $x_{ij} \in \mathbb{R}^d$ with zero mean and identity covariance and an η_i-sub-Gaussian mean-zero noise term $z_{ij} \in \mathbb{R}$. While it may seem that non-isotropic noise is less of a challenge in this setting since each noise term z_{ij} is a scalar, our goal is to handle a challenging scenario where the variances of the noise terms z_{ij} can depend on the *realized* measurements x_{ij}, which is a more general and widely applicable setting compared to those in prior work. Similarly to the PCA setting, our relaxed assumptions on the noise make it information-theoretically impossible to do subspace recovery if we only have one data point from each user (see Section 4), and thus we assume each user contributes at least two data points. We use the subspace spanned by the top-k eigenvectors of the following matrix A as our estimator:

$$A = \sum_{i=1}^n \frac{w_i}{m_i(m_i - 1)} \sum_{j_1 \neq j_2} (x_{ij_1} y_{ij_1})(x_{ij_2} y_{ij_2})^\mathsf{T}. \tag{3}$$

In the special case where $\eta_1 = \cdots = \eta_n = \eta, m_1 = \cdots = m_n = m$, and $\|\beta_i\|_2 \leq r$ for all i, our estimator achieves the following error upper bound:

$$\sin\theta \leq O\left(\log^3(nd/\delta)\sqrt{\frac{d(r^4 + r^2\eta^2 + \eta^4/m)}{mn\sigma_k^4}}\right), \tag{4}$$

where σ_k^2 is the k-th largest eigenvalue of $\frac{1}{n}\sum_{i=1}^n \beta_i\beta_i^\mathsf{T}$ (Corollary L.2). Our error upper bound extends smoothly to more general cases where η_i and m_i vary among users (Theorem L.1). Moreover, our upper bound matches the ones in prior work [e.g. Tripuraneni et al., 2021, Theorem 3] despite requiring less restrictive assumptions.

1.2 Related Work

Principal component analysis under non-isotropic noise has been studied by Vaswani and Narayanamurthy [2017], Zhang et al. [2018] and Narayanamurthy and Vaswani [2020]. When translated to our setting, these papers focus on having only one data point from each user and thus they require additional assumptions—either the level of non-isotropy is low, or the noise is coordinate-wise independent and the subspace is incoherent. The estimation error guarantees in these papers depend crucially on how well these additional assumptions are satisfied. Zhu et al. [2019] and Cai et al. [2021] study PCA with noise and missing data, and Chen et al. [2021] and Cheng et al. [2021] study eigenvalue and eigenvector estimation under heteroscedastic noise. These four papers all assume that the noise is coordinate-wise independent and the subspace/eigenspace is incoherent.

The linear models setting we consider has recently been studied as a basic setting of meta-learning and federated learning by Kong et al. [2020], Tripuraneni et al. [2021], Collins et al. [2021], and Thekumparampil et al. [2021]. These papers all make the assumption that the noise terms z_{ij} are independent of the measurements x_{ij}, an assumption that we relax in this paper. Collins et al. [2021] and Thekumparampil et al. [2021] make improvements in sample complexity and error guarantees compared to earlier work by Kong et al. [2020] and Tripuraneni et al. [2021], but Collins et al. [2021] focus on the noiseless setting ($z_{ij} = 0$) and Thekumparampil et al. [2021] require at least $\Omega(k^2)$ examples per user. Tripuraneni et al. [2021] and Thekumparampil et al. [2021] assume that the measurements x_{ij} are drawn from the standard (multivariate) Gaussian distribution, where as Kong et al. [2020], Collins et al. [2021] and our work make the relaxed assumption that x_{ij} are sub-Gaussian with identity covariance, which, in particular, allows the fourth-order moments of x_{ij} to be non-isotropic. There is a large body of prior work on meta-learning beyond the linear setting [see e.g. Maurer et al., 2016, Tripuraneni et al., 2020, Du et al., 2020].

When collecting data from users, it is often important to ensure that private information about users is not revealed through the release of the learned estimator. Many recent works proposed and analyzed estimators that achieve *user-level differential privacy* in settings including mean estimation [Levy et al., 2021, Esfandiari et al., 2021], meta-learning [Jain et al., 2021] and PAC learning [Ghazi et al., 2021]. Recently, Cummings et al. [2021] study one-dimensional mean estimation in a setting similar to ours, under a differential privacy constraint.

The matrix A we define in (1) is a weighted sum of $A_i := \frac{1}{m_i(m_i-1)}\sum_{j_1\neq j_2} x_{ij_1}x_{ij_2}^\mathsf{T}$ over users $i = 1, \ldots, n$, and each A_i has the form of a U-statistic [Halmos, 1946, Hoeffding, 1948]. U-statistics have been applied to many statistical tasks including tensor completion [Xia and Yuan, 2019] and various testing problems [Zhong and Chen, 2011, He et al., 2021, Schrab et al., 2022]. In our definition of A_i, we do *not* make the assumption that the distributions of $x_{i1}, \ldots, x_{im_i}$ are identical although the assumption is commonly used in applications of U-statistics. The matrix A in (3) is also a weighted sum of U-statistics where we again do not make the assumption of identical distribution.

1.3 Paper Organization

In Section 2, we formally define the maximum principal angle and other notions we use throughout the paper. Our results in the PCA setting and the linear models setting are presented in Sections 3 and 4, respectively. We defer most technical proofs to the appendices.

2 Preliminaries

We use $\|A\|$ to denote the spectral norm of a matrix A, and use $\|u\|_2$ to denote the ℓ_2 norm of a vector u. For positive integers $k \leq d$, we use $\mathbb{O}_{d,k}$ to denote the set of matrices $A \in \mathbb{R}^{d\times k}$ satisfying $A^\mathsf{T} A = I_k$, where I_k is the $k \times k$ identity matrix. We use $\mathbb{O}_d$ to denote $\mathbb{O}_{d,d}$, which is the set of $d \times d$ orthogonal matrices. We use $\mathsf{col}(A)$ to denote the linear subspace spanned by the columns of a matrix A. We use the base-e logarithm throughout the paper.

Maximum Principal Angle. Let $U, \hat{U} \in \mathbb{O}_d$ be two orthogonal matrices. Suppose the columns of U and $\hat{U}$ are partitioned as $U = [U_1\ U_2], \hat{U} = [\hat{U}_1\ \hat{U}_2]$ where $U_1, \hat{U}_1 \in \mathbb{O}_{d,k}$ for an integer k satisfying $0 < k < d$. Let Γ (resp. $\hat{\Gamma}$) be the k-dimensional linear subspace spanned by the columns of U_1 (resp. $\hat{U}_1$). Originating from [Jordan, 1875], the *maximum principal angle* $\theta \in [0, \pi/2]$ between Γ and $\hat{\Gamma}$, denoted by $\angle(\Gamma, \hat{\Gamma})$ or $\angle(U_1, \hat{U}_1)$, is defined by $\sin\theta = \|U_1U_1^\mathsf{T} - \hat{U}_1\hat{U}_1^\mathsf{T}\| = \|U_1^\mathsf{T}\hat{U}_2\| = \|U_2^\mathsf{T}\hat{U}_1\|$. It is not hard to see that the maximum principal angle depend only on the subspaces $\Gamma, \hat{\Gamma}$ and not on the choices of U and $\hat{U}$, and $\sin\angle(\Gamma, \hat{\Gamma})$ is a natural metric between k-dimensional subspaces (see Appendix A for more details where we discuss the definition of principal angles for any two subspaces with possibly different dimensions).

With the definition of the maximum principal angle, we can now state a variant of the Davis–Kahan $\sin\theta$ theorem [Davis and Kahan, 1970] that will be useful in our analysis (see Appendix E for proof):

Theorem 2.1 (Variant of Davis–Kahan $\sin\theta$ theorem)**.** *Let $A, \hat{A} \in \mathbb{R}^{d\times d}$ be symmetric matrices. Let λ_i denote the i-th largest eigenvalue of A. For a positive integer k smaller than d, let θ denote the maximum principal angle between the subspaces spanned by the top-k eigenvectors of A and $\hat{A}$. Assuming $\lambda_k > \lambda_{k+1}$,*

$$\sin\theta \leq \frac{2\|A - \hat{A}\|}{\lambda_k - \lambda_{k+1}}.$$

Sub-Gaussian and sub-exponential distributions. We say a random variable $x \in \mathbb{R}$ with expectation $\mathbb{E}[x] \in \mathbb{R}$ has *sub-Gaussian* constant $b \in \mathbb{R}_{\geq 0}$ if $\mathbb{E}[|x - \mathbb{E}[x]|^p]^{1/p} \leq b\sqrt{p}$ for every $p \geq 1$. We say x has *sub-exponential* constant $b \in \mathbb{R}_{\geq 0}$ if $\mathbb{E}[|x - \mathbb{E}[x]|^p]^{1/p} \leq bp$ for every $p \geq 1$. We say a random vector $y \in \mathbb{R}^d$ has sub-Gaussian (resp. sub-exponential) constant $b \in \mathbb{R}_{\geq 0}$ if for every unit vector $u \in \mathbb{R}^d$ (i.e., $\|u\|_2 = 1$), the random variable $u^\mathsf{T} y \in \mathbb{R}$ has sub-Gaussian (resp. sub-exponential) constant b. We say y is *b-sub-Gaussian* (resp. *b-sub-exponential*) if it has sub-Gaussian (resp. sub-exponential) constant b.

3 Principal Component Analysis

In the principal component analysis (PCA) setting, our goal is to recover the k-dimensional subspace Γ spanned by the user-specific means $\mu_1, \ldots, \mu_n \in \mathbb{R}^d$ of the n users. From each user i, we have $m_i \geq 2$ data points

$$x_{ij} = \mu_i + z_{ij} \quad \text{for } j = 1, \ldots, m_i. \tag{5}$$

We assume the noise $z_{ij} \in \mathbb{R}^d$ is drawn independently from a mean zero distribution with sub-Gaussian constant η_i. We do *not* assume that the variance of z_{ij} is the same along every direction, *nor* do we assume that the distribution of z_{ij} is the same for different (i, j). We first show an error upper bound for our estimator when the user-specific means $\mu_1, \ldots, \mu_n$ are deterministic vectors (Section 3.1) and then apply this result to the case where $\mu_1, \ldots, \mu_n$ are drawn from a sub-Gaussian distribution (Section 3.2). In Section 3.3 we prove an information-theoretic error lower bound matching our upper bound.

3.1 Fixed User-Specific Means

We first focus on the case where $\mu_1, \ldots, \mu_n$ are deterministic vectors. In this case, all the randomness in the data comes from the noise z_{ij}. Our estimator is the subspace $\hat{\Gamma}$ spanned by the top-k eigenvectors of A defined in (1). For $\ell = 1, \ldots, d$, we define σ_ℓ^2 to be the ℓ-th largest eigenvalue of

$\sum_{i=1}^n w_i \mu_i \mu_i^\top$. Since $\mu_1, \ldots, \mu_n$ share a k-dimensional subspace, $\sigma_\ell = 0$ for $\ell > k$. We prove the following general theorem on the error guarantee of our estimator:

Theorem 3.1. *Define $\xi^2 = \| \sum_{i=1}^n w_i^2 \mu_i \mu_i^\top \eta_i^2 / m_i \|$ and let θ denote the maximum principal angle between our estimator $\hat{\Gamma}$ and the true subspace Γ spanned by $\mu_1, \ldots, \mu_n$. For any $\delta \in (0, 1/2)$, with probability at least $1 - \delta$,*

$$\sin\theta = O\left(\sigma_k^{-2}\sqrt{(d+\log(1/\delta))\left(\xi^2 + \sum_{i=1}^n \frac{w_i^2\eta_i^4}{m_i^2}\right)} + \sigma_k^{-2}(d+\log(1/\delta))\max_i \frac{w_i\eta_i^2}{m_i}\right). \quad (6)$$

We can simplify the bound in Theorem 3.1 by considering special cases:

Corollary 3.2. *Assume $\max\{\eta_1/\sqrt{m_1}, \ldots, \eta_n/\sqrt{m_n}\} = t$ and we choose $w_1 = \cdots = w_n = 1/n$. For any $\delta \in (0, 1/2)$, with probability at least $1 - \delta$,*

$$\sin\theta = O\left(\frac{t\sigma_1 + t^2}{\sigma_k^2}\sqrt{\frac{d+\log(1/\delta)}{n}}\right). \quad (7)$$

In particular, when $\eta_1 = \cdots = \eta_n = \eta$, and $m_1 = \cdots = m_n = m$, error bound (7) *becomes*

$$\sin\theta = O\left(\left(\frac{\eta\sigma_1}{\sigma_k^2\sqrt{m}} + \frac{\eta^2}{\sigma_k^2 m}\right)\sqrt{\frac{d+\log(1/\delta)}{n}}\right).$$

We defer the complete proof of Theorem 3.1 and Corollary 3.2 to Appendices F and G. Our proof is based on the Davis-Kahan $\sin\theta$ theorem (Theorem 2.1). Since $\sigma_{k+1}^2 = 0$, Theorem 2.1 implies

$$\sin\theta \le \frac{2\|A - \sum_{i=1}^n w_i\mu_i\mu_i^\top\|}{\sigma_k^2}. \quad (8)$$

This reduces our goal to proving an upper bound on the spectral norm of $A - \sum_{i=1}^n w_i\mu_i\mu_i^\top$. Since for distinct j_1 and j_2 in $\{1, \ldots, m_i\}$ we have $\mathbb{E}[x_{ij_1}x_{ij_2}^\top] = \mu_i\mu_i^\top$, our construction of A in (1) guarantees $\mathbb{E}[A] = \sum_{i=1}^n w_i\mu_i\mu_i^\top$. Therefore, our goal becomes controlling the deviation of A from its expectation, and we achieve this goal using techniques for matrix concentration inequalities.

3.2 Sub-Gaussian User-Specific Means

We apply our error upper bound in Theorem 3.1 to the case where $\mu_1, \ldots, \mu_n \in \mathbb{R}^d$ are drawn iid from $N(0, \sigma^2 UU^\top)$ for an unknown $U \in \mathbb{O}_{d,k}$. We still assume that each data point $x_{ij} \in \mathbb{R}^d$ is generated by adding a noise vector $z_{ij} \in \mathbb{R}^d$ to the user-specific mean μ_i as in (5). We do *not* assume that the noise vectors $(z_{ij})_{1\le i\le n, 1\le j\le m_i}$ are independent of the user-specific means $(\mu_i)_{1\le i\le n}$, but we assume that when conditioned on $(\mu_i)_{1\le i\le n}$, every noise vector z_{ij} independently follows a distribution with mean zero and sub-Gaussian constant η_i. We use the same estimator $\hat{\Gamma}$ as before: $\hat{\Gamma}$ is the subspace spanned by the top-k eigenvectors of A defined in (1). We determine the optimal weights $w_1, \ldots, w_n$ in (1) as long as $m_1, \ldots, m_n$ and $\eta_1, \ldots, \eta_n$ satisfy a mild assumption (Assumption 3.2), achieving an error upper bound in Theorem 3.4. In the next subsection, we prove an error lower bound (Theorem 3.7) that matches our upper bound (Theorem 3.4) up to a constant factor, assuming $d \ge (1 + \Omega(1))k$ and $\delta = \Theta(1)$.

We prove our error upper bound in a slightly more general setting than $\mu_1, \ldots, \mu_n$ drawn iid from $N(0, \sigma^2 UU^\top)$. Specifically, we make the following assumption on the distribution of $\mu_1, \ldots, \mu_n$:

Assumption 3.1. *The user-specific means $\mu_1, \ldots, \mu_n \in \mathbb{R}^d$ are mean-zero independent random vectors supported on an unknown k-dimensional subspace Γ. Moreover, for a parameter $\sigma > 0$, for every $i = 1, \ldots, n$, μ_i has sub-Gaussian constant $O(\sigma)$, and the k-th largest eigenvalue of $\mathbb{E}[\mu_i\mu_i^\top]$ is at least σ^2.*

Under this assumption, we have the following lower bound on the σ_k^2 in Theorem 3.1 (see Appendix H for proof):

Claim 3.3. *Under Assumption 3.1, let $w_1, \dots, w_n \in \mathbb{R}_{\geq 0}$ be user weights satisfying $w_1 + \cdots + w_n = 1$ and σ_k^2 be the k-th largest eigenvalue of $\sum_{i=1}^n w_i \mu_i \mu_i^\mathsf{T}$. There exists an absolute constant $C_* > 1$ such that for any $\delta \in (0, 1/2)$, as long as $\max_{1 \leq i \leq n} w_i \leq 1/C_*(k + \log(1/\delta))$, then $\sigma_k^2 \geq \sigma^2/2$ with probability at least $1 - \delta/2$.*

The following definition is important for us to choose the weights $w_1, \dots, w_n$ in (1) optimally:

Definition 3.1. *Define $\gamma_i = \left(\frac{\eta_i^2}{\sigma^2 m_i} + \frac{\eta_i^4}{\sigma^4 m_i^2}\right)^{-1}$ and assume w.l.o.g. that $\gamma_1 \geq \cdots \geq \gamma_n$. Define $\gamma_i' = \gamma_i$ if $i \geq k$, and $\gamma_i' = \gamma_k$ if $i < k$.*

Intuitively, we can view γ_i as measuring the "amount of information" provided by the data points from user i. This is consistent with the fact that γ_i increases as the number m_i of data points from user i increases, and γ_i decreases as the noise magnitude η_i from user i increases. With the users sorted so that $\gamma_1 \geq \cdots \geq \gamma_n$, the quantity γ_i' is then defined to be γ_k for the k most "informative" users $i = 1, \dots, k$, and $\gamma_i' = \gamma_i$ for other users. We make the following mild assumption on γ_i' under which we achieve optimal estimation error:

Assumption 3.2. $\sum_{i=1}^n \gamma_i' \geq C_*(k + \log(1/\delta))\gamma_1'$ *for C_* defined in Claim 3.3.*

By the definition of γ_i', it is easy to show that Assumption 3.2 is equivalent to $\sum_{i=k+1}^n \gamma_i \geq ((C_* - 1)k + C_* \log(1/\delta))\gamma_k$. Therefore, if we view γ_i as the "amount of information" from user i, Assumption 3.2 intuitively requires that a significant contribution to the total "information" comes from outside the k most "informative" users. This assumption allows us to avoid the case where we only have exactly $n = k$ users: in that case, we would have $\sigma_k^2 \approx \sigma^2/k^2$ for uniform weights $w_1 = \cdots = w_n$ (see [Rudelson and Vershynin, 2008] and references therein), as opposed to the desired $\sigma_k^2 \geq \sigma^2/2$ in Claim 3.3.

Assumption 3.2 is a mild assumption. For example, when $\gamma_k = \cdots = \gamma_n$, Assumption 3.2 holds as long as $n \geq C_*(k + \log(1/\delta))$. Also, since $\gamma_1' = \cdots = \gamma_k' \geq \gamma_{k+1}' \geq \cdots \geq \gamma_n' \geq 0$, it trivially holds that $\sum_{i=1}^n \gamma_i' \geq k\gamma_1'$. Assumption 3.2 is relatively mild when compared to this trivial inequality.

Under Assumption 3.2, we show that it is optimal to choose the weights $w_1, \dots, w_n$ as

$$w_i = \frac{\gamma_i'}{\sum_{\ell=1}^n \gamma_\ell'}. \tag{9}$$

Specifically, if we plug (9) into Theorem 3.1 and bound ξ and σ_k based on the distribution of $\mu_1, \dots, \mu_n$, we get the following error upper bound which matches our lower bound (Theorem 3.7) in Section 3.3. We defer its proof to Appendix I.

Theorem 3.4. *Under Assumptions 3.1 and 3.2, if we choose $w_1, \dots, w_n$ as in* (9) *and define $\theta = \angle(\Gamma, \hat{\Gamma})$, for $\delta \in (0, 1/2)$, with probability at least $1 - \delta$,*

$$\sin\theta \leq O\left(\sqrt{\frac{d + \log(1/\delta)}{\sum_{i=1}^n \gamma_i'}}\right). \tag{10}$$

For comparison, consider the setting when $\sigma = \eta_i = 1$ for every $i = 1, \dots, n$. The result then says that $\sin\theta$ is bounded by approximately $\sqrt{\frac{d}{\sum_{i=1}^n m_i}}$. This is the same rate as we would get if we have $\sum_{i=1}^n m_i$ users each contributing a single independent data point with homogeneous spherical noise. Thus as long as the data points are not too concentrated on fewer than k users, the heterogeneity comes at no additional cost.

3.3 Lower Bound

We prove a lower bound matching the upper bound in Theorem 3.4 up to constant in the setting where $\delta = \Theta(1)$, $d \geq (1 + \Omega(1))k$.

For every positive integer d, there is a natural "uniform" distribution over $\mathbb{O}_d$ given by Haar's theorem [Haar, 1933] (see e.g. [Diestel and Spalsbury, 2014] for a textbook). We denote this distribution by $\mathsf{Haar}(\mathbb{O}_d)$. A random matrix A drawn from $\mathsf{Haar}(\mathbb{O}_d)$ has the following invariance property: for any deterministic matrix $B \in \mathbb{O}_d$, the random matrices A, AB and BA all have the same distribution.

For an integer $k \le d$, we can construct a random matrix $A_1 \in \mathbb{O}_{d,k}$ by first drawing $A \in \mathbb{R}^{d\times d}$ from $\mathsf{Haar}(\mathbb{O}_d)$ and then take the first k columns of A. We denote the distribution of A_1 by $\mathsf{Haar}(\mathbb{O}_{d,k})$. The invariance property of $\mathsf{Haar}(\mathbb{O}_d)$ immediately implies the following claims:

Claim 3.5. *Let $A \in \mathbb{O}_d$ be a random matrix drawn from $\mathsf{Haar}(\mathbb{O}_d)$ and let $B \in \mathbb{O}_{d,k}$ be a fixed matrix. Then AB distributes as $\mathsf{Haar}(\mathbb{O}_{d,k})$.*

Proof. The matrix B can be written as the first k columns of a matrix $C \in \mathbb{O}_d$. Now AB is the first k columns of AC, where AC distributes as $\mathsf{Haar}(\mathbb{O}_d)$ by the invariance property. This implies that AB distributes as $\mathsf{Haar}(\mathbb{O}_{d,k})$. □

Claim 3.6. *Let $B \in \mathbb{O}_{d,k}$ be a random matrix. Assume for every fixed matrix $A \in \mathbb{O}_d$, the random matrices B and AB have the same distribution. Then $B \sim \mathsf{Haar}(\mathbb{O}_{d,k})$.*

Proof. If we draw A independently from $\mathsf{Haar}(\mathbb{O}_d)$, the random matrices B and AB still have the same distribution. By Claim 3.5, AB distributes as $\mathsf{Haar}(\mathbb{O}_{d,k})$, so B must also distribute as $\mathsf{Haar}(\mathbb{O}_{d,k})$. □

With the definition of $\mathsf{Haar}(\mathbb{O}_{d,k})$, we state our lower bound in the following theorem:

Theorem 3.7. *Let k, d, n be positive integers satisfying $k < d$ and $k \le n$. Let $m_1, \dots, m_n$ be positive integers and $\sigma, \eta_1, \dots, \eta_n$ be positive real numbers. Suppose we draw $U \in \mathbb{O}_{d,k}$ from $\mathsf{Haar}(\mathbb{O}_{d,k})$ and then draw $\mu_1, \dots, \mu_n$ independently from $N(0, \sigma^2 UU^\mathsf{T})$. For every $i = 1, \dots, n$, we draw m_i data points x_{ij} for $j = 1, \dots, m_i$ as $x_{ij} = \mu_i + z_{ij}$, where each z_{ij} is drawn independently from the spherical Gaussian $N(0, \eta_i^2 I)$. Let $\hat{\Gamma}$ be any estimator mapping $(x_{ij})_{1\le i\le n, 1\le j\le m_i}$ to a (possibly randomized) k-dimensional subspace of $\mathbb{R}^d$. Let θ denote the maximum principal angle between $\hat{\Gamma}((x_{ij})_{1\le i\le n, 1\le j\le m_i})$ and the true subspace $\Gamma = \mathsf{col}(U)$. If real numbers $t \ge 0$ and $\delta \in [0, 1/2)$ satisfy $\Pr[\sin\theta \le t] \ge 1 - \delta$, then*

$$t \ge \Omega\left(\min\left\{1, \sqrt{\frac{(d-k)(1-\delta)}{\sum_{i=k}^n \gamma_i}}\right\}\right), \tag{11}$$

where $\gamma_1, \dots, \gamma_n$ are defined in Definition 3.1.

Note that $\gamma_i' = \gamma_i$ for $i \ge k$, so our upper bound in (10) matches the lower bound (11) up to a constant factor assuming $\delta = \Theta(1)$ and $d \ge (1 + \Omega(1))k$.

We use the local Fano method to prove the lower bound using the technical lemmas in Appendix D. In particular, we reduce our goal to proving an upper bound on the KL divergence between Gaussian distributions whose covariance matrices are defined based on matrices $U, \hat{U} \in \mathbb{O}_{d,k}$ with $\|UU^\mathsf{T} - \hat{U}\hat{U}^\mathsf{T}\|_F$ bounded. We prove the following lemma in Appendix J that upper bounds the KL divergence using $\|UU^\mathsf{T} - \hat{U}\hat{U}^\mathsf{T}\|_F$:

Lemma 3.8. *For $\sigma \in \mathbb{R}_{\ge 0}, \eta \in \mathbb{R}_{>0}, U, \hat{U} \in \mathbb{O}_{d,k}$, define $\Sigma = \sigma^2 UU^\mathsf{T} + \eta^2 I$ and $\hat{\Sigma} = \sigma^2 \hat{U}\hat{U}^\mathsf{T} + \eta^2 I$. Then,*

$$D_{\mathrm{kl}}(N(0, \hat{\Sigma}) \| N(0, \Sigma)) = \frac{\sigma^4 \|UU^\mathsf{T} - \hat{U}\hat{U}^\mathsf{T}\|_F^2}{4(\sigma^2\eta^2 + \eta^4)}.$$

Lemma 3.8 and the results in Appendix D allow us to prove a version of (11) in which the sum in the demoninator is over $i = 1, \dots, n$. This, however, is weaker and less useful than (11) in which the sum in the denominator is over $i = k, k+1, \dots, n$. To prove Theorem 3.7, we extract a hard distribution in which the data points from users $1, \dots, k-1$ are "useless" in terms of subspace recovery.

Let Γ_1 be the $(k-1)$-dimensional subspace spanned by $\mu_1, \dots, \mu_{k-1}$. We let $v_1, \dots, v_{k-1}$ be a random orthonormal basis of Γ_1, and we append another vector $v_k \in \Gamma$ to form an orthonormal basis $v_1, \dots, v_k$ of Γ. We define $V_1 = [v_1 \ \cdots \ v_{k-1}] \in \mathbb{O}_{d,k-1}$ and $V = [v_1 \ \cdots \ v_k] \in \mathbb{O}_{d,k}$. In Figure 1 we show a graphical model demonstrating the dependency among the random objects we defined.

Let us focus on the joint distribution of $(V_1, V, (\mu_1, \dots, \mu_{k-1}))$. By the invariance property, for any matrices $\tilde{V}_1 \in \mathbb{O}_{d,k-1}, \tilde{V} \in \mathbb{O}_{d,k}$, measurable set $S \subseteq (\mathbb{R}^d)^{k-1}$, and orthogonal matrix $G \in \mathbb{O}_d$,

$$\Pr[(\mu_1, \dots, \mu_{k-1}) \in S | V = \tilde{V}, V_1 = \tilde{V}_1] = \Pr[(\mu_1, \dots, \mu_{k-1}) \in S_G | V = G\tilde{V}, V_1 = G\tilde{V}_1],$$

$$
\begin{tikzcd}
 & (\mu_1, \ldots, \mu_{k-1}) \arrow[r] & \Gamma_1 \arrow[r] & V_1 \arrow[dr] & \\
U \arrow[ur] \arrow[rrr] \arrow[dr] & & & \Gamma \arrow[r] & V \\
 & (\mu_k, \ldots, \mu_n) & & &
\end{tikzcd}
$$

Figure 1: Graphical model A.

where $S_G = \{(G\tilde{\mu}_1, \ldots, G\tilde{\mu}_{k-1}) : (\tilde{\mu}_1, \ldots, \tilde{\mu}_{k-1}) \in S\}$. For any $\tilde{V}, \tilde{V}' \in \mathbb{O}_{d,k}$ whose first $k-1$ columns are both $\tilde{V}_1$, there exists $G \in \mathbb{O}_d$ such that $\tilde{V}' = G\tilde{V}$ and thus $\tilde{V}_1 = G\tilde{V}_1$. This implies that for any $\tilde{\mu} \in \mathsf{col}(\tilde{V}_1)$, we have $G\tilde{\mu} = \tilde{\mu}$, and thus $(S \cap \mathsf{col}(\tilde{V}_1)^{k-1})_G = S \cap \mathsf{col}(\tilde{V}_1)^{k-1}$ for any measurable $S \subseteq (\mathbb{R}^d)^{k-1}$. Here, $\mathsf{col}(\tilde{V}_1)^{k-1} = \{(\tilde{\mu}_1, \ldots, \tilde{\mu}_{k-1}) : \tilde{\mu}_i \in \mathsf{col}(\tilde{V}_1) \text{ for } i = 1, \ldots, k-1\} \subseteq (\mathbb{R}^d)^{k-1}$. When conditioned on $V_1 = \tilde{V}_1$, for every $i = 1, \ldots, k-1$ we have $\mu_i \in \Gamma_1 = \mathsf{col}(V_1) = \mathsf{col}(\tilde{V}_1)$, which implies that $(\mu_1, \ldots, \mu_{k-1}) \in \mathsf{col}(\tilde{V}_1)^{k-1}$. Therefore,

$$
\begin{aligned}
& \Pr[(\mu_1, \ldots, \mu_{k-1}) \in S | V = \tilde{V}, V_1 = \tilde{V}_1] \\
= \; & \Pr[(\mu_1, \ldots, \mu_{k-1}) \in S \cap \mathsf{col}(\tilde{V}_1)^{k-1} | V = \tilde{V}, V_1 = \tilde{V}_1] \\
= \; & \Pr[(\mu_1, \ldots, \mu_{k-1}) \in (S \cap \mathsf{col}(\tilde{V}_1)^{k-1})_G | V = G\tilde{V}, V_1 = G\tilde{V}_1] \\
= \; & \Pr[(\mu_1, \ldots, \mu_{k-1}) \in S \cap \mathsf{col}(\tilde{V}_1)^{k-1} | V = \tilde{V}', V_1 = \tilde{V}_1] \\
= \; & \Pr[(\mu_1, \ldots, \mu_{k-1}) \in S | V = \tilde{V}', V_1 = \tilde{V}_1].
\end{aligned}
$$

This implies that $(\mu_1, \ldots, \mu_{k-1})$ and V are conditionally independent given V_1. Therefore, the joint distribution of $(V_1, V, (\mu_1, \ldots, \mu_{k-1}))$ can be formed by first drawing V and V_1, and then drawing $\mu_1, \ldots, \mu_{k-1}$ based only on V_1 and not on V. Since $\mu_k, \ldots, \mu_n$ are drawn iid from $N(0, \sigma^2 UU^\mathsf{T}) = N(0, \sigma^2 VV^\mathsf{T})$, we have the graphical model shown in Figure 2.

$$
\begin{tikzcd}
V \arrow[r] \arrow[dr] & V_1 \arrow[r] & (\mu_1, \ldots, \mu_{k-1}) \\
 & (\mu_k, \ldots, \mu_n) &
\end{tikzcd}
$$

Figure 2: Graphical model B.

By Claim 3.6, the marginal distribution of V is $\mathsf{Haar}(\mathbb{O}_{d,k})$. By Claim 3.5, we can implement this distribution by first drawing $W \sim \mathsf{Haar}(\mathbb{O}_d)$ and then drawing E independently from any distribution over $\mathbb{O}_{d,k}$ and let $V = WE$. We choose the distribution of E later, where we ensure that the first $k-1$ columms of E is always $\begin{bmatrix} I_{k-1} \\ 0 \end{bmatrix}$. This guarantees that the first $k-1$ columns of W and V are the same, and thus V_1 is exactly the first $k-1$ columns of W, resulting in the graphical model shown in Figure 3.

$$
\begin{tikzcd}
W \arrow[r] \arrow[dr] & V_1 \arrow[r] & (\mu_1, \ldots, \mu_{k-1}) \\
E \arrow[r] & V \arrow[r] & (\mu_k, \ldots, \mu_n)
\end{tikzcd}
$$

Figure 3: Graphical model C.

Note that in Figure 3 there is no directed path from E to $(\mu_1, \ldots, \mu_{k-1})$. Intuitively, this means that knowing $(\mu_1, \ldots, \mu_{k-1})$ gives us no information about E. Now by choosing the distribution of E appropriately, we can prove (11) in which the denominator does not contain $\gamma_1, \ldots, \gamma_{k-1}$. We defer the complete proof of Theorem 3.7 to Appendix K.

4 Linear Models

In the linear models setting, the data distribution of user i is parameterized by an unknown vector $\beta_i \in \mathbb{R}^d$. As before, we assume that the vectors $\beta_1, \ldots, \beta_n$ from the n users lie in an unknown k-dimensional subspace Γ. Our goal is to recover the subspace using the following data. For every

$i = 1, \ldots, n$, we have m_i data points from user i: $(x_{i1}, y_{i1}), \ldots, (x_{im_i}, y_{im_i}) \in \mathbb{R}^d \times \mathbb{R}$. For every $j = 1, \ldots, m_i$, we assume the *measurement* $x_{ij} \in \mathbb{R}^d$ is a random vector drawn independently from an $O(1)$-sub-Gaussian distribution with zero mean and identity covariance matrix. The measurement *outcome* y_{ij} is determined by $y_{ij} = x_{ij}^\mathsf{T}\beta_i + z_{ij}$, where the random *noise* $z_{ij} \in \mathbb{R}$ can depend on the measurements $x_{i1}, \ldots, x_{im_i}$. When conditioned on $x_{i1}, \ldots, x_{im_i}$, we assume every z_{ij} for $j = 1, \ldots, m_i$ is independently drawn from an η_i-sub-Gaussian distribution with zero mean, but we do *not* assume that the conditional distribution of z_{ij} is the same for every $j = 1, \ldots, m_i$. The (in)dependence among x_{ij} and z_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$ can be summarized by the example graphical model in Figure 4.

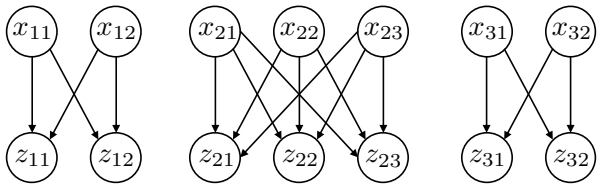

Figure 4: An example for $n = 3, m_1 = 2, m_2 = 3, m_3 = 2$.

Since we allow the noise z_{ij} to depend on the measurements x_{ij}, it is information-theoretically impossible to recover the subspace if we only have one data point from every user. Consider the scenario where every β_i is drawn independently from $N(0, \sigma^2 uu^\mathsf{T})$ for an unknown unit vector $u \in \mathbb{R}^d$ and every x_{ij} is drawn independently and uniformly from $\{-1, 1\}^d$. If we set z_{ij} to be $z_{ij} = x_{ij}^\mathsf{T}\nu_{ij}$ where ν_{ij} is independently drawn from $N(0, \sigma^2(I - uu^\mathsf{T}))$, then every y_{ij} satisfies $y_{ij} = x_{ij}^\mathsf{T}(\beta_i + \nu_{ij})$ where $\beta_i + \nu_{ij}$ distributes as $N(0, \sigma^2 I)$ independently from x_{ij}. This implies that the joint distribution of $((x_{i1}, y_{i1}))_{i=1,\ldots,n}$ does not change with u, i.e., we get no information about u from one data point per user.

Thus, we assume $m_i \geq 2$ for every user i. In this case, we achieve error upper bounds that match the ones in [Tripuraneni et al., 2021] despite our relaxed assumptions on the noise. Our estimator is the subspace $\hat{\Gamma}$ spanned by the top-k eigenvectors of A defined in (3). We defer the analysis of our estimator to Appendix L.

Acknowledgments and Disclosure of Funding

Part of this work was performed while LH was interning at Apple.

References

Changxiao Cai, Gen Li, Yuejie Chi, H. Vincent Poor, and Yuxin Chen. Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *Ann. Statist.*, 49(2):944–967, 2021. ISSN 0090-5364. doi: 10.1214/20-aos1986. URL `https://doi.org/10.1214/20-aos1986`.

T. Tony Cai, Zongming Ma, and Yihong Wu. Sparse PCA: optimal rates and adaptive estimation. *Ann. Statist.*, 41(6):3074–3110, 2013. ISSN 0090-5364. doi: 10.1214/13-AOS1178. URL `https://doi.org/10.1214/13-AOS1178`.

Yuxin Chen, Chen Cheng, and Jianqing Fan. Asymmetry helps: eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices. *Ann. Statist.*, 49(1):435–458, 2021. ISSN 0090-5364. doi: 10.1214/20-AOS1963. URL `https://doi.org/10.1214/20-AOS1963`.

Chen Cheng, Yuting Wei, and Yuxin Chen. Tackling small eigen-gaps: fine-grained eigenvector estimation and inference under heteroscedastic noise. *IEEE Trans. Inform. Theory*, 67(11):7380–7419, 2021. ISSN 0018-9448. doi: 10.1109/TIT.2021.3111828. URL `https://doi.org/10.1109/TIT.2021.3111828`.

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.

Rachel Cummings, Vitaly Feldman, Audra McMillan, and Kunal Talwar. Mean estimation with user-level privacy under data heterogeneity, 2021. TPDP.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Joe Diestel and Angela Spalsbury. *The joys of Haar measure*, volume 150 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2014. ISBN 978-1-4704-0935-7. doi: 10.1090/gsm/150. URL `https://doi.org/10.1090/gsm/150`.

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*, 2020.

Hossein Esfandiari, Vahab Mirrokni, and Shyam Narayanan. Tight and robust private mean estimation with few users. *arXiv preprint arXiv:2110.11876*, 2021.

Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. User-level private learning via correlated sampling. *Advances in Neural Information Processing Systems*, 34, 2021.

Alfred Haar. Der Massbegriff in der Theorie der kontinuierlichen Gruppen. *Ann. of Math. (2)*, 34(1):147–169, 1933. ISSN 0003-486X. doi: 10.2307/1968346. URL `https://doi.org/10.2307/1968346`.

Paul R. Halmos. The theory of unbiased estimation. *Ann. Math. Statistics*, 17:34–43, 1946. ISSN 0003-4851. doi: 10.1214/aoms/1177731020. URL `https://doi.org/10.1214/aoms/1177731020`.

Yinqiu He, Gongjun Xu, Chong Wu, and Wei Pan. Asymptotically independent U-statistics in high-dimensional testing. *Ann. Statist.*, 49(1):154–181, 2021. ISSN 0090-5364. doi: 10.1214/20-AOS1951. URL `https://doi.org/10.1214/20-AOS1951`.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statistics*, 19: 293–325, 1948. ISSN 0003-4851. doi: 10.1214/aoms/1177730196. URL `https://doi.org/10.1214/aoms/1177730196`.

Prateek Jain, John Rush, Adam Smith, Shuang Song, and Abhradeep Guha Thakurta. Differentially private model personalization. *Advances in Neural Information Processing Systems*, 34, 2021.

Camille Jordan. Essai sur la géométrie à n dimensions. *Bull. Soc. Math. France*, 3:103–174, 1875. ISSN 0037-9484. URL `http://www.numdam.org/item?id=BSMF_1875__3__103_2`.

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.

Daniel Levy, Ziteng Sun, Kareem Amin, Satyen Kale, Alex Kulesza, Mehryar Mohri, and Ananda Theertha Suresh. Learning with user-level privacy. *Advances in Neural Information Processing Systems*, 34, 2021.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *J. Mach. Learn. Res.*, 17:Paper No. 81, 32, 2016. ISSN 1532-4435.

Praneeth Narayanamurthy and Namrata Vaswani. Fast robust subspace tracking via PCA in sparse data-dependent noise. *IEEE Journal on Selected Areas in Information Theory*, 1(3):723–744, 2020. doi: 10.1109/JSAIT.2020.3040365.

Mark Rudelson and Roman Vershynin. The Littlewood-Offord problem and invertibility of random matrices. *Adv. Math.*, 218(2):600–633, 2008. ISSN 0001-8708. doi: 10.1016/j.aim.2008.01.010. URL `https://doi.org/10.1016/j.aim.2008.01.010`.

Antonin Schrab, Ilmun Kim, Benjamin Guedj, and Arthur Gretton. Efficient aggregated kernel tests using incomplete u-statistics. *arXiv preprint arXiv:2206.09194*, 2022.

G. W. Stewart and Ji Guang Sun. *Matrix perturbation theory*. Computer Science and Scientific Computing. Academic Press, Inc., Boston, MA, 1990. ISBN 0-12-670230-6.

Kiran Koshy Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Sample efficient linear meta-learning by alternating minimization. *arXiv preprint arXiv:2105.08306*, 2021.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33:7852–7862, 2020.

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

Namrata Vaswani and Praneeth Narayanamurthy. Finite sample guarantees for PCA in non-isotropic and data-dependent noise. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 783–789. IEEE, 2017.

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

Dong Xia and Ming Yuan. On polynomial time methods for exact low-rank tensor completion. *Found. Comput. Math.*, 19(6):1265–1313, 2019. ISSN 1615-3375. doi: 10.1007/s10208-018-09408-6. URL `https://doi.org/10.1007/s10208-018-09408-6`.

Anru R Zhang, T Tony Cai, and Yihong Wu. Heteroskedastic PCA: Algorithm, optimality, and applications. *arXiv preprint arXiv:1810.08316*, 2018.

Ping-Shou Zhong and Song Xi Chen. Tests for high-dimensional regression coefficients with factorial designs. *J. Amer. Statist. Assoc.*, 106(493):260–274, 2011. ISSN 0162-1459. doi: 10.1198/jasa.2011.tm10284. URL `https://doi.org/10.1198/jasa.2011.tm10284`.

Ziwei Zhu, Tengyao Wang, and Richard J Samworth. High-dimensional principal component analysis with heterogeneous missingness. *arXiv preprint arXiv:1906.12125*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work aims to develop better subspace recovery methods for a diverse population of users with heterogeneous data distributions. Researchers and practitioners working on transfer learning, federated learning, data analysis, and statistics will benefit from our work. We are not aware of any negative societal impacts of our work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Basics about Principal Angles

We provide a formal definition of the principal angles introduced by Jordan [1875] and briefly discuss some basic properties of this notion (see e.g. [Stewart and Sun, 1990] for a textbook).

Let $U, \hat{U} \in \mathbb{O}_d$ be two orthogonal matrices. Suppose the columns of U and $\hat{U}$ are partitioned as $U = [U_1\ U_2], \hat{U} = [\hat{U}_1\ \hat{U}_2]$ where $U_1 \in \mathbb{O}_{d,k}$ and $\hat{U}_1 \in \mathbb{O}_{d,\hat{k}}$ for integers $k, \hat{k}$ satisfying $0 < k \leq \hat{k} < d$. Starting from a singular value decomposition of $U_1^\mathsf{T}\hat{U}_1$ as $R_1^\mathsf{T}\Sigma\hat{R}_1$, one can find an integer $\ell \in \mathbb{Z}_{\geq 0}$ and angles $0 < \theta_1 \leq \cdots \leq \theta_\ell \leq \pi/2$ such that there exist orthogonal matrices $P, \hat{P} \in \mathbb{O}_d, R_1 \in \mathbb{O}_k, R_2 \in \mathbb{O}_{d-k}, \hat{R}_1 \in \mathbb{O}_{\hat{k}}, \hat{R}_2 \in \mathbb{O}_{d-\hat{k}}$ satisfying

$$U = P \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix}, \quad \hat{U} = \hat{P} \begin{bmatrix} \hat{R}_1 & 0 \\ 0 & \hat{R}_2 \end{bmatrix}, \quad \text{and} \tag{12}$$

$$P^\mathsf{T}\hat{P} = (\hat{P}^\mathsf{T}P)^\mathsf{T} = \begin{bmatrix} I_{k-\ell} & 0 & 0 & 0 & 0 \\ 0 & \cos\Theta & 0 & \sin\Theta & 0 \\ 0 & 0 & I_{\hat{k}-k} & 0 & 0 \\ 0 & -\sin\Theta & 0 & \cos\Theta & 0 \\ 0 & 0 & 0 & 0 & I_{d-\hat{k}-\ell} \end{bmatrix}, \tag{13}$$

where $\cos\Theta = \mathrm{diag}(\cos\theta_1, \ldots, \cos\theta_\ell)$ and $\sin\Theta = \mathrm{diag}(\sin\theta_1, \ldots, \sin\theta_\ell)$. It is easy to see that $\cos\theta_1, \ldots, \cos\theta_\ell$ are exactly the singular values of $U_1^\mathsf{T}\hat{U}_1$ that are not equal to 1, so $\ell, \theta_1, \ldots, \theta_\ell$ are unique. Let Γ (resp. $\hat{\Gamma}$) be the k-dimensional (resp. $\hat{k}$-dimensional) linear subspace spanned by the columns of U_1 (resp. $\hat{U}_1$). The (non-zero) *principal angles* between Γ and $\hat{\Gamma}$ are defined to be $\theta_1, \ldots, \theta_\ell$. It is not hard to see that the principal angles depend only on the subspaces $\Gamma, \hat{\Gamma}$ and not on the choices of U and $\hat{U}$. The *maximum principal angle* between Γ and $\hat{\Gamma}$, denoted by $\angle(\Gamma, \hat{\Gamma})$ or $\angle(U_1, \hat{U}_1)$, is defined to be θ_ℓ (or 0 when $\ell = 0$). When $k = \hat{k}$, using (12) and (13), it is not hard to show that the non-zero eigenvalues of $U_1U_1^\mathsf{T} - \hat{U}_1\hat{U}_1^\mathsf{T}$ are $\pm\sin\theta_1, \ldots, \pm\sin\theta_\ell$. This implies $\|U_1U_1^\mathsf{T} - \hat{U}_1\hat{U}_1^\mathsf{T}\| = \sin\angle(\Gamma, \hat{\Gamma})$ and $\|U_1U_1^\mathsf{T} - \hat{U}_1\hat{U}_1^\mathsf{T}\|_F = \sqrt{2\sum_{i=1}^\ell \sin^2\theta_i}$. In particular, $\sin\angle(\Gamma, \hat{\Gamma})$ is a natural metric between k-dimensional subspaces.

B Basic Facts and Concentration Inequalities

Lemma B.1 ([See Vershynin, 2018, Proposition 2.6.1])**.** *If $X_1, \ldots, X_n \in \mathbb{R}$ are independent random variables with sub-Gaussian constants $b_1, \ldots, b_n \in \mathbb{R}_{\geq 0}$ respectively, then $\sum_{i=1}^n X_i$ has sub-Gaussian constant $O\left(\sqrt{\sum_{i=1}^n b_i^2}\right)$.*

Lemma B.2 ([See Vershynin, 2018, Lemma 2.7.6])**.** *If $X \in \mathbb{R}$ is a mean-zero random variable with sub-Gaussian constant $b \in \mathbb{R}_{\geq 0}$, then X^2 has sub-exponential constant $O(b^2)$.*

Lemma B.3 ([See Vershynin, 2018, Theorem 2.6.2])**.** *Let $x_1, \ldots, x_n \in \mathbb{R}$ be independent random variables. For every $i = 1, \ldots, n$, assume x_i has mean zero and sub-Gaussian constant $b_i \in \mathbb{R}_{\geq 0}$. Then for every $\delta \in (0, 1/2)$, with probability at least $1 - \delta$,*

$$\left|\sum_{i=1}^n x_i\right| \leq O\left(\sqrt{\log(1/\delta)\sum_{i=1}^n b_i^2}\right).$$

Lemma B.4 ([See Vershynin, 2018, Theorem 2.8.1])**.** *Let $x_1, \ldots, x_n \in \mathbb{R}$ be independent random variables. For every $i = 1, \ldots, n$, assume x_i has mean zero and sub-exponential constant $b_i \in \mathbb{R}_{\geq 0}$. Then for every $\delta \in (0, 1/2)$, with probability at least $1 - \delta$,*

$$\left|\sum_{i=1}^n x_i\right| \leq O\left(\sqrt{\log(1/\delta)\sum_{i=1}^n b_i^2} + \log(1/\delta)\max_{i=1,\ldots,n} b_i\right).$$

C Vector and Matrix Concentration Inequalities

Lemma C.1 ([See Vershynin, 2018, Corollary 4.2.13 and Exercise 4.4.3])**.** *There exists an absolute constant $C > 1$ and a set $O'_d \subseteq \mathbb{O}_{d,1}$ for every positive integer d such that*

1. $|O'_d| \le 2^{Cd}$ *for every* $d \in \mathbb{Z}_{>0}$.

2. *for every* $d \in \mathbb{Z}_{>0}$ *and every symmetric matrix* $A \in \mathbb{R}^{d\times d}$,
$$\|A\| \le C \sup_{u\in O'} |u^\mathsf{T} A u|.$$

3. *for every* $d_1, d_2 \in \mathbb{Z}_{>0}$ *and every matrix* $A \in \mathbb{R}^{d_1\times d_2}$,
$$\|A\| \le C \sup_{u_1\in O'_{d_1}} \sup_{u_2\in O'_{d_2}} |u_1^\mathsf{T} A u_2|.$$

Lemma C.2. *Suppose* $x_1, \dots, x_n \in \mathbb{R}^d$ *are independent random vectors, and each* x_i *is mean-zero and* 1*-sub-Gaussian. Suppose* $w_1, \dots, w_n \in \mathbb{R}$ *are fixed real numbers. Define* $E = \sum_{i=1}^n w_i x_i x_i^\mathsf{T}$ *and* $\bar{E} = \mathbb{E}[E]$. *Then for any* $\delta \in (0, 1/2)$, *with probability at least* $1-\delta$,
$$\|E - \bar{E}\| \le O\left(\sqrt{(d+\log(1/\delta))\sum_{i=1}^n w_i^2} + (d+\log(1/\delta))\max_i |w_i|\right).$$

Proof. Let O' denote the set $O'_d \subseteq \mathbb{O}_{d,1}$ guaranteed by Lemma C.1. For any fixed $u \in O'$, by Lemma B.2, $(x_i^\mathsf{T} u)^2$ has sub-exponential constant $O(1)$, and thus $w_i(x_i^\mathsf{T} u)^2$ has sub-exponential constant $|w_i|$. Define $\delta' := \delta/|O'|$. By Lemma B.4, with probability at least $1-\delta'$,
$$\left|\sum_{i=1}^n w_i (x_i^\mathsf{T} u)^2 - \mathbb{E}\left[\sum_{i=1}^n w_i (x_i^\mathsf{T} u)^2\right]\right| \le O\left(\sqrt{\log(1/\delta')\sum_{i=1}^n w_i^2} + \log(1/\delta')\max_i |w_i|\right). \quad (14)$$
By a union bound, with probability at least $1-\delta$, the above inequality holds for every $u \in O'$. By the definition of O',
$$\|E - \bar{E}\| \le O\left(\sup_{u\in O'} |u(E-\bar{E})u|\right) = O\left(\sup_{u\in O'}\left|\sum_{i=1}^n w_i (x_i^\mathsf{T} u)^2 - \mathbb{E}\left[\sum_{i=1}^n w_i (x_i^\mathsf{T} u)^2\right]\right|\right). \quad (15)$$
Combining (14) and (15) and noting that $\log(1/\delta') = O(d+\log(1/\delta))$ completes the proof. □

Lemma C.3. *Suppose* $x_1, \dots, x_n \in \mathbb{R}^d$ *are independent random vectors, and each* x_i *is mean-zero and* 1*-sub-Gaussian. Suppose* $b_1, \dots, b_n \in \mathbb{R}$ *are fixed real numbers. Then for any* $\delta \in (0, 1/2)$, *with probability at least* $1-\delta$,
$$\|\,[x_1 \quad \cdots \quad x_n]\,\mathrm{diag}(b_1, \dots, b_n)\| \le O\left(\sqrt{\sum_{i=1}^n b_i^2 + (d+\log(1/\delta))\max_i b_i^2}\right).$$

Proof. Define $F = [x_1 \quad \cdots \quad x_n]\,\mathrm{diag}(b_1, \dots, b_n)$ and $E = FF^\mathsf{T} = \sum_{i=1}^n b_i^2 x_i x_i^\mathsf{T}$. By Lemma C.2, with probability at least $1-\delta$,
$$\|E - \mathbb{E}[E]\| \le O\left(\sqrt{(d+\log(1/\delta))\sum_{i=1}^n b_i^4} + (d+\log(1/\delta))\max_i b_i^2\right). \quad (16)$$
For every unit vector $u \in \mathbb{R}^d$, $x_i^\mathsf{T} u$ is 1-sub-Gaussian, and thus $\mathbb{E}[(x_i^\mathsf{T} u)^2] \le O(1)$. This implies that $u\mathbb{E}[x_i x_i^\mathsf{T}]u \le O(1)$ for every unit vector $u \in \mathbb{R}^d$ and thus $\|\mathbb{E}[x_i x_i^\mathsf{T}]\| \le O(1)$. Now we have
$$\|\mathbb{E}[E]\| \le \sum_{i=1}^n b_i^2 \|x_i x_i^\mathsf{T}\| \le O\left(\sum_{i=1}^n b_i^2\right). \quad (17)$$
Combining (16) and (17) and using the fact that $\sum_{i=1}^n b_i^4 \le \left(\sum_{i=1}^n b_i^2\right)\max_i b_i^2$, with probability at least $1-\delta$,
$$\|E\| \le \|E - \mathbb{E}[E]\| + \|\mathbb{E}[E]\| \le O\left(\sum_{i=1}^n b_i^2 + (d+\log(1/\delta))\max_i b_i^2\right).$$
The lemma is proved by $\|F\| = \sqrt{\|E\|}$. □

Lemma C.4. *Let $X \in \mathbb{R}^{d\times m}$ be a fixed matrix and let $v \in \mathbb{R}^m$ be a mean zero random vector with sub-Gaussian constant 1. For $\delta \in (0, 1/2)$, with probability at least $1-\delta$,*

$$\|Xv\|_2 \le O(\|X\|\sqrt{r + \log(1/\delta)}),$$

where r is the rank of X.

Proof. By the singular value decomposition, there exists $P \in \mathbb{O}_d$ such that all but the first r rows of PX are zeros. Let $Y \in \mathbb{R}^{r\times m}$ denote the first r rows of PX. We have

$$\|Xv\|_2 = \|PXv\|_2 = \|Yv\|_2. \tag{18}$$

Let O' be the O'_r in Lemma C.1. For every $u \in O'$, $u^\mathsf{T} Yv$ has sub-Gaussian constant $O(\|u^\mathsf{T} Y\|_2)$. By Lemma B.3, for $\delta' = \delta/|O'|$, with probability at least $1-\delta'$,

$$|u^\mathsf{T} Yv| \le O\left(\|u^\mathsf{T} Y\|_2\sqrt{\log(1/\delta')}\right) \le O\left(\|Y\|\sqrt{\log(1/\delta')}\right).$$

By a union bound, with probability at least $1-\delta$,

$$\sup_{u\in O'} |u^\mathsf{T} Yv| \le O\left(\|u^\mathsf{T} Y\|_2\sqrt{\log(1/\delta')}\right) \le O\left(\|Y\|\sqrt{\log(1/\delta')}\right). \tag{19}$$

By the definition of O',

$$\|Yv\|_2 \le O\left(\sup_{u\in O'} |u^\mathsf{T} Yv|\right). \tag{20}$$

The lemma is proved by combining (18), (19) and (20) and noting that $\|Y\| = \|X\|$ and $\log(1/\delta') = O(r + \log(1/\delta))$. □

Lemma C.5. *Let $Z_1, \dots, Z_n \in \mathbb{R}^{d_1\times d_2}$ be mean-zero independent random matrices. Suppose for real numbers $R, p_1, \dots, p_n \in \mathbb{R}_{\ge 0}$, we have $\Pr[\|Z_i\| \ge R] \le p_i$. Moreover,*

$$\max\{\|\mathbb{E}[Z_i Z_i^\mathsf{T}]\|, \|\mathbb{E}[Z_i^\mathsf{T} Z_i]\|\} \le \sigma_i^2.$$

Then with probability at least $1 - 2(d_1+d_2)e^{-t} - \sum_{i=1}^n p_i$,

$$\left\|\sum_{i=1}^n Z_i\right\| \le O\left(\sqrt{\sum_{i=1}^n \sigma_i^2 t} + Rt\right).$$

Proof. Since probabilities are always nonnegative, the lemma is trivial if $\sum_{i=1}^n p_i \ge 1$ or if $t \le c$ for a sufficiently small positive constant c. We thus assume $\sum_{i=1}^n p_i \le 1$ and $t \ge \Omega(1)$.

Define $Z'_i = Z_i \mathbf{1}(\|Z_i\| \le R)$. For every unit vector $u \in \mathbb{R}^d$, by Cauchy-Schwarz,

$$\mathbb{E}[\|Z_i u\|_2 \mathbf{1}(\|Z_i\| \ge R)]^2 \le \mathbb{E}[\|Z_i u\|_2^2]\mathbb{E}[\mathbf{1}(\|Z_i\| \ge R)^2] \le \sigma_i^2 \Pr[\|Z_i\| \ge R] \le \sigma_i^2 p_i.$$

By Jenson's inequality,

$$\mathbb{E}[\|Z_i u\|_2 \mathbf{1}(\|Z_i\| \ge R)] \ge \|\mathbb{E}[Z_i u \mathbf{1}(\|Z_i\| \ge R)]\|_2 = \|\mathbb{E}[Z_i \mathbf{1}(\|Z_i\| \ge R)]u\|_2.$$

Combining,

$$\|\mathbb{E}[Z_i \mathbf{1}(\|Z_i\| \ge R)]u\|_2 \le \sigma_i\sqrt{p_i}, \quad \text{for every unit vector } u \in \mathbb{R}^{d_2},$$

which implies that $\|\mathbb{E}[Z'_i]\| = \|\mathbb{E}[Z_i] - \mathbb{E}[Z'_i]\| = \|\mathbb{E}[Z_i \mathbf{1}(\|Z_i\| > R)]\| \le \sigma_i\sqrt{p_i}$.

We apply the matrix Bernstein inequality [see Vershynin, 2018, Exercise 5.4.15] to Z'_i. For every i, $\|Z'_i - \mathbb{E}[Z'_i]\| \le \|Z'_i\| + \|\mathbb{E}[Z'_i]\| \le \|Z'_i\| + \mathbb{E}[\|Z'_i\|] \le 2R$, and

$$\mathbb{E}[(Z'_i - \mathbb{E}[Z'_i])(Z'_i - \mathbb{E}[Z'_i])^\mathsf{T}] = \mathbb{E}[Z'_i (Z'_i)^\mathsf{T}] - \mathbb{E}[Z'_i]\mathbb{E}[Z'_i]^\mathsf{T} \preceq \mathbb{E}[Z'_i (Z'_i)^\mathsf{T}] \preceq \sigma_i^2 I.$$

The matrix Bernstein inequality implies that with probability at least $1 - 2(d_1+d_2)e^{-t}$,

$$\left\|\sum_{i=1}^n Z'_i - \sum_{i=1}^n \mathbb{E}[Z'_i]\right\| \le O\left(\sqrt{\sum_{i=1}^n \sigma_i^2 t} + Rt\right).$$

By the union bound, with probability at least $1 - 2(d_1 + d_2)e^{-t} - \sum_{i=1}^n p_i$,

$$\left\| \sum_{i=1}^n Z_i - \sum_{i=1}^n \mathbb{E}[Z_i'] \right\| \le O\left(\sqrt{\sum_{i=1}^n \sigma_i^2 t} + Rt \right),$$

in which case,

$$\begin{aligned}
\left\| \sum_{i=1}^n Z_i \right\| &\le \left\| \sum_{i=1}^n \mathbb{E}[Z_i'] \right\| + O\left(\sqrt{\sum_{i=1}^n \sigma_i^2 t} + Rt \right) \\
&\le \sum_{i=1}^n \sigma_i \sqrt{p_i} + O\left(\sqrt{\sum_{i=1}^n \sigma_i^2 t} + Rt \right) \\
&\le \sqrt{\left(\sum_{i=1}^n \sigma_i^2 \right) \left(\sum_{i=1}^n p_i \right)} + O\left(\sqrt{\sum_{i=1}^n \sigma_i^2 t} + Rt \right). \qquad \text{(by Cauchy-Schwarz)}
\end{aligned}$$

The lemma is proved by our assumption that $\sum_{i=1}^n p_i \le 1$ and $t \ge \Omega(1)$. □

D Fano's Inequality and Local Packing Numbers

For two probability distributions p_1, p_2 over the same set $\mathcal{X}$ with probability densities $p_1(x)$ and $p_2(x)$, their KL divergence is defined to be

$$D_{\mathrm{kl}}(p_1 \| p_2) = \mathbb{E}_{X \sim p_1} \left[\log \frac{p_1(X)}{p_2(X)} \right].$$

For a joint distribution p over random variables X and Y, we define the mutual information between X and Y to be

$$I(X;Y) = D_{\mathrm{kl}}(p \| p_X \times p_Y),$$

where p_X (resp. p_Y) is the marginal distribution of X (resp. Y), and $p_X \times p_Y$ is the product distribution of p_X and p_Y. The following claim is standard:

Claim D.1. *For random variables X, Y, Z, if Y and Z are conditionally independent given X, then*

$$I(X;Y,Z) \le I(X;Y) + I(X;Z), \tag{21}$$

$$I(Y;X,Z) = I(Y;X). \tag{22}$$

Theorem D.2 (Fano's Inequality)**.** *Let $X, Y, \hat{X}$ be random objects. Assume their joint distribution forms a Markov chain $X \to Y \to \hat{X}$. Assume that the marginal distribution of X is uniform over a finite set $\mathcal{X}$. Then,*

$$\Pr[\hat{X} \ne X] \ge 1 - \frac{I(X;Y) + \log 2}{\log |\mathcal{X}|}.$$

Lemma D.3. *In the setting of Theorem D.2, let P_x denote the conditional distribution of Y given $X = x$. Then,*

$$I(X;Y) \le \max_{x, x' \in \mathcal{X}} D_{\mathrm{kl}}(P_{x'} \| P_x).$$

Proof. Let $P_x(y)$ denote $\Pr[Y = y | X = x]$. Since X distributes uniformly over $\mathcal{X}$, the marginal distribution of Y is given by $\Pr[Y = y] = \sum_{x \in \mathcal{X}} P_x(y)/|\mathcal{X}|$. By the definition of $I(X;Y)$,

$$\begin{aligned}
I(X;Y) &= \mathbb{E}_X \mathbb{E}_{Y \sim P_X} \left[\log \frac{P_X(Y)/|\mathcal{X}|}{\sum_{x \in \mathcal{X}} P_x(Y)/|\mathcal{X}|^2} \right] \\
&= \mathbb{E}_X \mathbb{E}_{Y \sim P_X} \left[\log \frac{P_X(Y)}{\sum_{x \in \mathcal{X}} P_x(Y)/|\mathcal{X}|} \right] \\
&\le \mathbb{E}_X \mathbb{E}_{Y \sim P_X} \left[\frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \log \frac{P_X(Y)}{P_x(Y)} \right] \qquad \text{(by Jensen's Inequality)}
\end{aligned}$$

$$= \frac{1}{|\mathcal{X}|} \sum_{x' \in \mathcal{X}} \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathbb{E}_{Y \sim P_{x'}} \left[\log \frac{P_{x'}(Y)}{P_x(Y)} \right]$$
$$= \frac{1}{|\mathcal{X}|^2} \sum_{x \in \mathcal{X}} \sum_{x' \in \mathcal{X}} D_{\mathrm{kl}}(P_{x'} \| P_x)$$
$$\leq \max_{x, x' \in \mathcal{X}} D_{\mathrm{kl}}(P_{x'} \| P_x). \qquad \square$$

Lemma D.4 ([see Cai et al., 2013, Lemma 1]). *There exists an absolute constant $C > 0$ with the following property. For any positive integers $k \leq d$ and real number $t \in (0, 1/C]$, there exists a subset $O' \subseteq \mathbb{O}_{d,k}$ with size at least $10^{k(d-k)}$ such that*

1. *For distinct $U, V \in O'$, $\|UU^\mathsf{T} - VV^\mathsf{T}\|_F > t$.*
2. *For any $U, V \in O'$, $\|UU^\mathsf{T} - VV^\mathsf{T}\|_F \leq Ct$.*

E Proof of Theorem 2.1

Theorem 2.1 (Variant of Davis–Kahan $\sin\theta$ theorem). *Let $A, \hat{A} \in \mathbb{R}^{d \times d}$ be symmetric matrices. Let λ_i denote the i-th largest eigenvalue of A. For a positive integer k smaller than d, let θ denote the maximum principal angle between the subspaces spanned by the top-k eigenvectors of A and $\hat{A}$. Assuming $\lambda_k > \lambda_{k+1}$,*

$$\sin\theta \leq \frac{2\|A - \hat{A}\|}{\lambda_k - \lambda_{k+1}}.$$

Proof. We assume $\frac{2\|A-\hat{A}\|}{\lambda_k - \lambda_{k+1}} < 1$ because otherwise the theorem is trivial.

Let $\hat{\lambda}_i$ denote the i-th largest eigenvalue of $\hat{A}$. By Weyl's inequality,

$$\hat{\lambda}_{k+1} - \lambda_{k+1} \leq \|A - \hat{A}\| \leq \frac{\lambda_k - \lambda_{k+1}}{2},$$

where we used our assumption $\frac{2\|A-\hat{A}\|}{\lambda_k - \lambda_{k+1}} < 1$ in the last inequality. This implies

$$\lambda_k - \hat{\lambda}_{k+1} \geq \frac{\lambda_k - \lambda_{k+1}}{2} > 0. \tag{23}$$

By the Davis-Kahan $\sin\theta$ theorem [Davis and Kahan, 1970],

$$\sin\theta \leq \frac{\|A - \hat{A}\|}{\lambda_k - \hat{\lambda}_{k+1}}. \tag{24}$$

Combining (23) and (24) completes the proof. $\square$

F Proof of Theorem 3.1

We recall:

Theorem 3.1. *Define $\xi^2 = \| \sum_{i=1}^n w_i^2 \mu_i \mu_i^\mathsf{T} \eta_i^2 / m_i \|$ and let θ denote the maximum principal angle between our estimator $\hat{\Gamma}$ and the true subspace Γ spanned by $\mu_1, \ldots, \mu_n$. For any $\delta \in (0, 1/2)$, with probability at least $1 - \delta$,*

$$\sin\theta = O\left(\sigma_k^{-2} \sqrt{(d + \log(1/\delta)) \left(\xi^2 + \sum_{i=1}^n \frac{w_i^2 \eta_i^4}{m_i^2} \right)} + \sigma_k^{-2} (d + \log(1/\delta)) \max_i \frac{w_i \eta_i^2}{m_i} \right). \tag{6}$$

By (8), it suffices to upper bound the spectral norm of $A - \sum_{i=1}^n w_i \mu_i \mu_i^\mathsf{T}$. We achieve this goal by proving Lemmas F.1 to F.3 in which we bound the spectral norm of each term on the right-hand-side of the following decomposition:

$$A - \sum_{i=1}^n w_i \mu_i \mu_i^\mathsf{T} = \sum_{i=1}^n w_i \mu_i \bar{z}_i^\mathsf{T} + \sum_{i=1}^n w_i \bar{z}_i \mu_i^\mathsf{T} + \sum_{i=1}^n \frac{w_i}{m_i(m_i - 1)} \sum_{j_1 \neq j_2} z_{ij_1} z_{ij_2}, \tag{25}$$

where $\bar{z}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} z_{ij}$.

Lemma F.1. *In the setting of Theorem 3.1, for any $\delta \in (0, 1/2)$, with probability at least $1-\delta$,*

$$\|\sum_{i=1}^{n} w_i \mu_i \bar{z}_i^\top\| \le O\left(\xi\sqrt{d+\log(1/\delta)}\right).$$

Proof. Let O' denote the set $O'_d \subseteq \mathbb{O}_{d,1}$ guaranteed by Lemma C.1. For any fixed $u, v \in O'$, by Lemma B.1, $\bar{z}_i^\top v$ is $\eta_i/\sqrt{m_i}$-sub-Gaussian, and thus $w_i(\mu_i^\top u)(\bar{z}_i^\top v)$ is $w_i|\mu_i^\top u|\eta_i/\sqrt{m_i}$-sub-Gaussian. By Lemma B.3, for any $\delta' \in (0, 1/2)$, with probability at least $1-\delta'$,

$$\left|\sum_{i=1}^{n} w_i(\mu_i^\top u)(\bar{z}_i^\top v)\right| \le O\left(\sqrt{\log(1/\delta')\sum_{i=1}^{n} w_i^2(\mu_i^\top u)^2\eta_i^2/m_i}\right). \tag{26}$$

By the definition of ξ,

$$\sum_{i=1}^{n} w_i^2(\mu_i^\top u)^2\eta_i^2/m_i = u^\top\left(\sum_{i=1}^{n} w_i^2\mu_i\mu_i^\top\eta_i^2/m_i\right)u \le \left\|\sum_{i=1}^{n} w_i^2\mu_i\mu_i^\top\eta_i^2/m_i\right\| = \xi^2.$$

Plugging this into (26), and setting $\delta' = \delta/|O'|^2$, with probability at least $1-\delta/|O'|^2$,

$$\left|\sum_{i=1}^{n} w_i(\mu_i^\top u)(\bar{z}_i^\top v)\right| \le O\left(\xi\sqrt{\log(1/\delta')}\right) \le O\left(\xi\sqrt{d+\log(1/\delta)}\right),$$

where we used the fact that $\log|O'| = O(d)$ to obtain the last inequality. By a union bound over $u, v \in O'$, with probability at least $1-\delta$, $\sup_{u\in O'}\sup_{v\in O'}|u^\top\left(\sum_{i=1}^{n} w_i\mu_i\bar{z}_i^\top\right)v| \le O(\xi\sqrt{d+\log(1/\delta)}$. The lemma is proved by combining this with $\left\|\sum_{i=1}^{n} w_i\mu_i\bar{z}_i^\top\right\| \le \sup_{u\in O'}\sup_{v\in O'}|u^\top\left(\sum_{i=1}^{n} w_i\mu_i\bar{z}_i^\top\right)v|$ by Lemma C.1. □

Lemma F.2. *In the setting of Theorem 3.1, define $Z = \sum_{i=1}^{n}\frac{w_i}{m_i(m_i-1)}(\sum_{j=1}^{m_i} z_{ij})(\sum_{j=1}^{m_i} z_{ij})^\top$. For any $\delta \in (0, 1/2)$, with probability at least $1-\delta$,*

$$\|Z - \mathbb{E}[Z]\| \le O\left(\sqrt{(d+\log(1/\delta))\sum_{i=1}^{n} w_i^2\eta_i^4/m_i^2} + (d+\log(1/\delta))\max_i w_i\eta_i^2/m_i\right). \tag{27}$$

Proof. By Lemma B.1, $\sum_{j=1}^{m_i} z_{ij}$ has sub-Gaussian constant b_i for some $b_i \in \mathbb{R}_{\ge 0}$ satisfying $b_i = O(\sqrt{m_i}\eta_i)$. Define $\hat{z}_i = \frac{1}{b_i}\sum_{j=1}^{m_i} z_{ij}$ and now $\hat{z}_i$ has sub-Gaussian constant 1. We write Z in terms of $\hat{z}_i$:

$$Z = \sum_{i=1}^{n}\frac{w_i b_i^2}{m_i(m_i-1)}\hat{z}_i\hat{z}_i^\top.$$

Applying Lemma C.2 completes the proof. □

Lemma F.3. *In the setting of Theorem 3.1, define $Z = \sum_{i=1}^{n}\frac{w_i}{m_i(m_i-1)}\sum_{j=1}^{m_i} z_{ij}z_{ij}^\top$. For any $\delta \in (0, 1/2)$, with probability at least $1-\delta$,*

$$\|Z - \mathbb{E}[Z]\| \le O\left(\sqrt{(d+\log(1/\delta))\sum_{i=1}^{n} w_i^2\eta_i^4/m_i^3} + (d+\log(1/\delta))\max_i w_i\eta_i^2/m_i^2\right). \tag{28}$$

Proof. Define $\hat{z}_{ij} = \frac{1}{\eta_i}z_{ij}$ and now $\hat{z}_{ij}$ has sub-Gaussian constant 1. We write Z in terms of $\hat{z}_{ij}$:

$$Z = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{w_i\eta_i^2}{m_i(m_i-1)}\hat{z}_{ij}\hat{z}_{ij}^\top.$$

Applying Lemma C.2 completes the proof. □

We are now ready to finish the proof of Theorem 3.1.

Proof of Theorem 3.1. We start with the last term in (25):

$$\sum_{i=1}^{n} \frac{w_i}{m_i(m_i-1)} \sum_{j_1 \neq j_2} z_{ij_1} z_{ij_2}$$
$$= \sum_{i=1}^{n} \frac{w_i}{m_i(m_i-1)} \left(\sum_{j=1}^{m_i} z_{ij} \right) \left(\sum_{j=1}^{m_i} z_{ij} \right)^{\mathsf{T}} - \sum_{i=1}^{n} \frac{w_i}{m_i(m_i-1)} \sum_{j=1}^{m_i} z_{ij} z_{ij}^{\mathsf{T}}.$$

It is clear that $\mathbb{E}\left[\sum_{i=1}^{n} \frac{w_i}{m_i(m_i-1)} \sum_{j_1 \neq j_2} z_{ij_1} z_{ij_2}\right] = 0$, so combining Lemmas F.2 and F.3, with probability at least $1-\delta$,

$$\left\| \sum_{i=1}^{n} \frac{w_i}{m_i(m_i-1)} \sum_{j_1 \neq j_2} z_{ij_1} z_{ij_2} \right\| \leq O\left(\sqrt{(d+\log(1/\delta)) \sum_{i=1}^{n} w_i^2 \eta_i^4 / m_i^2} + (d+\log(1/\delta)) \max_i w_i \eta_i^2 / m_i \right), \tag{29}$$

where we used the fact that the right-hand-side of (27) is at most the right-hand-side of (28) (up to a constant factor).

By Lemma F.1,

$$\left\| \sum_{i=1}^{n} w_i \mu_i \bar{z}_i^{\mathsf{T}} + \sum_{i=1}^{n} w_i \bar{z}_i \mu_i^{\mathsf{T}} \right\| \leq 2 \left\| \sum_{i=1}^{n} w_i \mu_i \bar{z}_i^{\mathsf{T}} \right\| \leq O\left(\xi \sqrt{d+\log(1/\delta)} \right). \tag{30}$$

Plugging (29) and (30) into (25) and then into (8) proves the theorem. □

G Proof of Corollary 3.2

Proof. By the definition of ξ^2 in Theorem 3.1, $\xi^2 = \| \sum_{i=1}^{n} \frac{\mu_i \mu_i^{\mathsf{T}} t^2}{n^2} \| \leq \frac{t^2 \sigma_1^2}{n}$. Plugging this into (6), we know that the following inequality holds with probability at least $1-\delta$:

$$\sin\theta \leq O\left(\sigma_k^{-2} \sqrt{(d+\log(1/\delta)) \left(\frac{t^2\sigma_1^2}{n} + \frac{t^4}{n} \right)} + \sigma_k^{-2}(d+\log(1/\delta)) \frac{t^2}{n} \right)$$
$$= O\left(\frac{t\sigma_1 + t^2}{\sigma_k^2} \sqrt{\frac{d+\log(1/\delta)}{n}} + \sigma_k^{-2}(d+\log(1/\delta)) \frac{t^2}{n} \right). \tag{31}$$

Since $\sin\theta \leq 1$ always holds, inequality (31) implies

$$\sin\theta \leq O\left(\frac{t\sigma_1 + t^2}{\sigma_k^2} \sqrt{\frac{d+\log(1/\delta)}{n}} + \min\left\{ 1, \sigma_k^{-2}(d+\log(1/\delta)) \frac{t^2}{n} \right\} \right) \tag{32}$$

Since $\min\{1, y\} \leq \sqrt{y}$ for every $y \in \mathbb{R}_{\geq 0}$, we have

$$\min\left\{ 1, \sigma_k^{-2}(d+\log(1/\delta)) \frac{t^2}{n} \right\} \leq \frac{t}{\sigma_k} \sqrt{\frac{(d+\log(1/\delta)}{n}} \leq \frac{t\sigma_1}{\sigma_k^2} \sqrt{\frac{(d+\log(1/\delta)}{n}}. \tag{33}$$

Plugging (33) into (32) proves the corollary. □

H Proof of Claim 3.3

Proof. Without loss of generality, we can assume that the subspace Γ in Assumption 3.1 is the subspace containing all vectors with all but the first k coordinates being zeros. We let $\hat{\mu}_i \in \mathbb{R}^k$ denote the first k coordinates of μ_i. By Lemma C.2, with probability at least $1-\delta/2$,

$$\left\| \sum_{i=1}^{n} w_i \hat{\mu}_i \hat{\mu}_i^{\mathsf{T}} - \mathbb{E}\left[\sum_{i=1}^{n} w_i \hat{\mu}_i \hat{\mu}_i^{\mathsf{T}} \right] \right\|$$

$$\leq O\left(\sigma^2\left(\sqrt{(k+\log(1/\delta))\sum_{i=1}^n w_i^2}+(k+\log(1/\delta))\max_i |w_i|\right)\right)$$
$$\leq O\left(\sigma^2\left(\sqrt{\frac{1}{C_*}}+\frac{1}{C_*}\right)\right), \tag{34}$$

where we used the fact that $\sum_{i=1}^n w_i^2 \leq \max_i w_i \leq 1/C^*(k+\log(1/\delta))$. Choosing C_* large enough, (34) implies

$$\left\|\sum_{i=1}^n w_i\hat{\mu}_i\hat{\mu}_i^\mathsf{T} - \mathbb{E}\left[\sum_{i=1}^n w_i\hat{\mu}_i\hat{\mu}_i^\mathsf{T}\right]\right\| \leq \sigma^2/2. \tag{35}$$

By Assumption 3.1, the k-th largest eigenvalue of $\mathbb{E}[\mu_i\mu_i^\mathsf{T}]$ is at least σ^2. This implies that the k-th largest eigenvalue of $\mathbb{E}[\sum_{i=1}^n w_i\hat{\mu}_i\hat{\mu}_i^\mathsf{T}]$ is at least σ^2. The claim is proved by combining this with (35) and noting that $\sum_{i=1}^n w_i\hat{\mu}_i\hat{\mu}_i^\mathsf{T}$ and $\sum_{i=1}^n w_i\mu_i\mu_i^\mathsf{T}$ have the same k-th largest eigenvalue. □

I Proof of Theorem 3.4

Theorem 3.4. *Under Assumptions 3.1 and 3.2, if we choose $w_1, \ldots, w_n$ as in* (9) *and define $\theta = \angle(\Gamma, \hat{\Gamma})$, for $\delta \in (0, 1/2)$, with probability at least $1-\delta$,*

$$\sin\theta \leq O\left(\sqrt{\frac{d+\log(1/\delta)}{\sum_{i=1}^n \gamma_i'}}\right). \tag{10}$$

Proof. By Assumption 3.2 and our choice of w_i, we have $w_i \leq 1/C_*(k+\log(1/\delta))$. By the definition of C_*, with probability at least $1-\delta/2$,

$$\sigma_k^2 \geq \sigma^2/2. \tag{36}$$

By Lemma C.3, with probability at least $1-\delta/4$,

$$\xi \leq O\left(\sqrt{\sum_{i=1}^n \frac{w_i^2\eta_i^2\sigma^2}{m_i}+(k+\log(1/\delta))\max_i \frac{w_i^2\eta_i^2\sigma^2}{m_i}}\right). \tag{37}$$

Replacing δ in Theorem 3.1 by $\delta/4$ and plugging (36) and (37) into (6), by the union bound, with probability at least $1-\delta$,

$$\sin\theta \leq O\Bigg(\sqrt{(d+\log(1/\delta))\sum_{i=1}^n\left(\frac{w_i^2\eta_i^2}{\sigma^2 m_i}+\frac{w_i^2\eta_i^4}{\sigma^4 m_i^2}\right)}$$
$$+\sqrt{(d+\log(1/\delta))(k+\log(1/\delta))\max_i \frac{w_i^2\eta_i^2}{\sigma^2 m_i}}+(d+\log(1/\delta))\max_i \frac{w_i\eta_i^2}{\sigma^2 m_i}.\Bigg) \tag{38}$$

By definition, $\frac{\eta_i^2}{\sigma^2 m_i}+\frac{\eta_i^4}{\sigma^4 m_i^2} = 1/\gamma_i \leq 1/\gamma_i'$. Therefore, inequality (38) implies

$$\sin\theta$$
$$\leq O\left(\sqrt{(d+\log(1/\delta))\sum_{i=1}^n \frac{w_i^2}{\gamma_i'}}+\sqrt{(d+\log(1/\delta))(k+\log(1/\delta))\max_i \frac{w_i^2}{\gamma_i'}}+(d+\log(1/\delta))\max_i \frac{w_i}{\gamma_i'}\right).$$

Plugging $w_i = \frac{\gamma_i'}{\sum_{\ell=1}^n \gamma_\ell'}$ into the inequality above,

$$\sin\theta \leq O\left(\sqrt{\frac{d+\log(1/\delta)}{\sum_{i=1}^n \gamma_i'}}+\sqrt{\frac{(d+\log(1/\delta))(k+\log(1/\delta))\max_i \gamma_i'}{\left(\sum_{i=1}^n \gamma_i'\right)^2}}+\frac{d+\log(1/\delta)}{\sum_{i=1}^n \gamma_i'}\right). \tag{39}$$

Since $\sin\theta \le 1$ always holds, inequality (39) implies

$$\sin\theta \le O\left(\sqrt{\frac{d+\log(1/\delta)}{\sum_{i=1}^n \gamma_i'}} + \sqrt{\frac{(d+\log(1/\delta))(k+\log(1/\delta))\max_i \gamma_i'}{(\sum_{i=1}^n \gamma_i')^2}} + \min\left\{1, \frac{d+\log(1/\delta)}{\sum_{i=1}^n \gamma_i'}\right\}\right).$$

The theorem is proved by the following facts:

$$(k+\log(1/\delta))\max_i \gamma_i' \le O\left(\sum_{i=1}^n \gamma_i'\right), \qquad \text{(by Assumption 3.2)}$$

$$\min\left\{1, \frac{d+\log(1/\delta)}{\sum_{i=1}^n \gamma_i'}\right\} \le \sqrt{\frac{d+\log(1/\delta)}{\sum_{i=1}^n \gamma_i'}}. \qquad \square$$

J Proof of Lemma 3.8

Proof. The KL divergence between mean-zero Gaussians can be computed by the following formula:

$$D_{\mathrm{kl}}(N(0,\hat\Sigma)\|N(0,\Sigma)) = \frac{1}{2}\left(\mathrm{tr}(\Sigma^{-1}\hat\Sigma) - d + \log\frac{\det\Sigma}{\det\hat\Sigma}\right). \tag{40}$$

It is clear that

$$\det\Sigma = \det\hat\Sigma = (\sigma^2+\eta^2)^k(\eta^2)^{d-k}. \tag{41}$$

It remains to compute $\mathrm{tr}(\Sigma^{-1}\hat\Sigma) - d$.

Define $J = \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{d\times d}$. By the definition of principal angles in Section 2, there exists $P, \hat P \in \mathbb{O}_d$ such that $UU^\top = PJP^\top, \hat U\hat U^\top = \hat P J \hat P^\top$ and

$$P^\top \hat P = \begin{bmatrix} I_{k-\ell} & 0 & 0 & 0 \\ 0 & \cos\Theta & \sin\Theta & 0 \\ 0 & -\sin\Theta & \cos\Theta & 0 \\ 0 & 0 & 0 & I_{d-\hat k-\ell} \end{bmatrix},$$

where $\theta_1,\ldots,\theta_\ell$ are the principal angles between $\mathsf{col}(U)$ and $\mathsf{col}(\hat U)$, $\cos\Theta = \mathrm{diag}(\cos\theta_1,\ldots,\cos\theta_\ell)$, and $\sin\Theta = \mathrm{diag}(\sin\theta_1,\ldots,\sin\theta_\ell)$. Now we have $\hat U\hat U^\top = \hat P J \hat P^\top = P(P^\top\hat P)J(P^\top\hat P)^\top P^\top = P\hat J P^\top$ where

$$\hat J = (P^\top\hat P)J(P^\top\hat P)^\top = \begin{bmatrix} I_{k-\ell} & 0 & 0 & 0 \\ 0 & (\cos\Theta)^2 & -\cos\Theta\sin\Theta & 0 \\ 0 & -\cos\Theta\sin\Theta & (\sin\Theta)^2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Therefore, $\Sigma = P(\sigma^2 J + \eta^2 I)P^\top$ and $\hat\Sigma = P(\sigma^2\hat J + \eta^2 I)P^\top$, which implies

$$\Sigma^{-1}\hat\Sigma = P(\sigma^2 J + \eta^2 I)^{-1}(\sigma^2\hat J + \eta^2 I)P^\top.$$

Since $\sigma^2 J + \eta^2 I = \mathrm{diag}(\underbrace{\sigma^2+\eta^2,\ldots,\sigma^2+\eta^2}_{k}, \underbrace{\eta^2,\ldots,\eta^2}_{d-k})$ and the diagonal entries of $\sigma^2\hat J + \eta^2 I$ are

$$\underbrace{\sigma^2+\eta^2,\ldots,\sigma^2+\eta^2}_{k-\ell}, \underbrace{\sigma^2\cos^2\theta_1+\eta^2,\ldots,\sigma^2\cos^2\theta_\ell+\eta^2}_{\ell},$$
$$\underbrace{\sigma^2\sin^2\theta_1+\eta^2,\ldots,\sigma^2\sin^2\theta_\ell+\eta^2}_{\ell}, \underbrace{\eta^2,\ldots,\eta^2}_{d-k-\ell},$$

we have

$$\mathrm{tr}(\Sigma^{-1}\hat\Sigma) - d = \mathrm{tr}((\sigma^2 J + \eta^2 I)^{-1}(\sigma^2\hat J + \eta^2 I)) - d$$

$$
\begin{aligned}
&= \sum_{i=1}^{\ell} \left(\frac{\sigma^2 \cos^2 \theta_i + \eta^2}{\sigma^2 + \eta^2} - 1 \right) + \sum_{i=1}^{\ell} \left(\frac{\sigma^2 \sin^2 \theta_i + \eta^2}{\eta^2} - 1 \right) \\
&= \sum_{i=1}^{\ell} \sin^2 \theta_i \left(\frac{\sigma^2}{\eta^2} - \frac{\sigma^2}{\sigma^2 + \eta^2} \right) \\
&= \frac{\sigma^4 \|UU^\mathsf{T} - \hat{U}\hat{U}^\mathsf{T}\|_F^2}{2(\eta^2\sigma^2 + \eta^4)}. \qquad (42)
\end{aligned}
$$

Plugging (41) and (42) into (40) completes the proof. □

K Proof of Theorem 3.7

We recall Theorem 3.7:

Theorem 3.7. *Let k, d, n be positive integers satisfying $k < d$ and $k \leq n$. Let $m_1, \ldots, m_n$ be positive integers and $\sigma, \eta_1, \ldots, \eta_n$ be positive real numbers. Suppose we draw $U \in \mathbb{O}_{d,k}$ from $\mathsf{Haar}(\mathbb{O}_{d,k})$ and then draw $\mu_1, \ldots, \mu_n$ independently from $N(0, \sigma^2 UU^\mathsf{T})$. For every $i = 1, \ldots, n$, we draw m_i data points x_{ij} for $j = 1, \ldots, m_i$ as $x_{ij} = \mu_i + z_{ij}$, where each z_{ij} is drawn independently from the spherical Gaussian $N(0, \eta_i^2 I)$. Let $\hat{\Gamma}$ be any estimator mapping $(x_{ij})_{1 \leq i \leq n, 1 \leq j \leq m_i}$ to a (possibly randomized) k-dimensional subspace of $\mathbb{R}^d$. Let θ denote the maximum principal angle between $\hat{\Gamma}((x_{ij})_{1 \leq i \leq n, 1 \leq j \leq m_i})$ and the true subspace $\Gamma = \mathsf{col}(U)$. If real numbers $t \geq 0$ and $\delta \in [0, 1/2)$ satisfy $\Pr[\sin\theta \leq t] \geq 1 - \delta$, then*

$$
t \geq \Omega \left(\min \left\{ 1, \sqrt{\frac{(d-k)(1-\delta)}{\sum_{i=k}^{n} \gamma_i}} \right\} \right), \qquad (11)
$$

where $\gamma_1, \ldots, \gamma_n$ are defined in Definition 3.1.

Proof. The theorem is trivial if t is lower bounded by a positive absolute constant, so without loss of generality, we can assume that $t \leq 1/(2\sqrt{2}C)$ for the constant C in Lemma D.4.

Now we describe the distribution of E on which we prove the error lower bound (11) that does not depend on $\gamma_1, \ldots, \gamma_{k-1}$. As we mentioned in Section 3.3, the first $k-1$ columns of E are fixed to be $\begin{bmatrix} I_{k-1} \\ 0 \end{bmatrix}$, so we only need to describe the distribution of the k-th column of E. By Lemma D.4, there exists $O' \subseteq \mathbb{O}_{d-k+1,1}$ with size at least 10^{d-k} such that

1. for distinct $u, v \in O'$, $\|uu^\mathsf{T} - vv^\mathsf{T}\|_F > 2\sqrt{2}t$;
2. for any $u, v \in O'$, $\|uu^\mathsf{T} - vv^\mathsf{T}\|_F \leq O(t)$.

The k-th column of E is chosen by first drawing a uniform random $u \in O'$ and then prepend zeros to it. That is

$$
E = \begin{bmatrix} I_{k-1} & 0 \\ 0 & u \end{bmatrix}. \qquad (43)
$$

We let $\mathcal{E}$ denote the support of the E so that E is chosen uniformly at random from $\mathcal{E}$.

By the law of total expectation, conditioned on $W, V_1, (\mu_1, \ldots, \mu_{k-1})$ and $(x_{ij})_{1 \leq i < k, 1 \leq j \leq m_i}$ being fixed to some specific value, we still have $\Pr[\sin\theta \geq t] \leq \delta$, where now $\Pr[\cdot]$ represents the conditional probability. We will abuse notation and omit explicitly writing out the conditioning throughout the proof.

After the conditioning, the randomness in V comes only from the randomness in $E \in \mathcal{E}$, and the distribution of V is the uniform distribution over the set $\mathcal{V} := \{WE : E \in \mathcal{E}\}$. For any two matrices $V', V'' \in \mathcal{V}$, there exists $E', E'' \in \mathcal{E}$ such that $V' = WE'$ and $V'' = WE''$. Moreover, there exists $u, v \in O'$ such that $E' = \begin{bmatrix} I_{k-1} & 0 \\ 0 & u \end{bmatrix}$ and $E'' = \begin{bmatrix} I_{k-1} & 0 \\ 0 & v \end{bmatrix}$. This implies that

$$
\|V'(V')^\mathsf{T} - V''(V'')^\mathsf{T}\|_F = \|E'(E')^\mathsf{T} - E''(E'')^\mathsf{T}\|_F = \|uu^\mathsf{T} - vv^\mathsf{T}\|_F \leq O(t). \qquad (44)
$$

Moreover, when $V' \neq V''$, we have $u \neq v$ and thus

$$\|V'(V')^\mathsf{T} - V''(V'')^\mathsf{T}\| = \|E'(E')^\mathsf{T} - E''(E'')^\mathsf{T}\| = \|uu^\mathsf{T} - vv^\mathsf{T}\| > 2t, \tag{45}$$

where we used the fact that $\|uu^\mathsf{T} - vv^\mathsf{T}\| = \sin\angle(u,v)$ and $2\sqrt{2}t < \|uu^\mathsf{T} - vv^\mathsf{T}\|_F = \sqrt{2\sin^2\angle(u,v)}$. We define $\hat{V} \in \mathcal{V}$ so that the maximum principal angle between $\mathsf{col}(\hat{V})$ and $\hat{\Gamma}$ is minimized. When $\sin\theta \leq t$ (i.e., $\sin\angle(\mathsf{col}(V),\Gamma) \leq t$), for any $V' \in \mathcal{V}$ different from V,

$$\sin\angle(\mathsf{col}(V'),\Gamma) \geq \sin\angle(V,V') - \sin\angle(\mathsf{col}(V),\Gamma) > 2t - t = t \geq \sin\angle(\mathsf{col}(V),\Gamma),$$

where we used $\sin\angle(V,V') = \|VV^\mathsf{T} - V'(V')^\mathsf{T}\| > 2t$ by (45). Therefore, $\sin\theta \leq t$ implies $\hat{V} = V$ and thus $\Pr[\sin\theta > t] \geq \Pr[\hat{V} \neq V]$.

Now we apply Theorem D.2 to the following Markov chain

$$V \to (x_{ij})_{k\leq i\leq n, 1\leq j\leq m_i} \to \hat{V}$$

to get

$$\delta \geq \Pr[\sin\theta > t] \geq \Pr[\hat{V} \neq V] \geq 1 - \frac{I(V;(x_{ij})_{k\leq i\leq n,1\leq j\leq m_i}) + \log 2}{\log|\mathcal{V}|}. \tag{46}$$

Now we bound $I(V;(x_{ij})_{k\leq i\leq n,1\leq j\leq m_i})$. According to the graphical model in Figure 3, $\mu_k,\ldots,\mu_n$ are conditionally independent given V, and thus $(x_{ij})_{1\leq j\leq m_i}$ are conditionally independent for $i = k,\ldots,n$ given V. By (21),

$$I(V;(x_{ij})_{k\leq i\leq n,1\leq j\leq m_i}) \leq \sum_{i=k}^{n} I(V;(x_{ij})_{1\leq j\leq m_i}). \tag{47}$$

Let $\bar{x}_i$ denote $\frac{1}{m_i}\sum_{j=1}^{m_i} x_{ij}$. Since x_{ij} for $j = 1,\ldots,m_i$ are drawn iid from $N(\mu_i,\eta_i^2 I)$, it is a standard fact that $\bar{x}_i$ is a sufficient statistic for μ_i, i.e., $(x_{ij})_{1\leq j\leq m_i}$ and μ_i are conditionally independent given $\bar{x}_i$. Therefore, for any measurable set $S \in (\mathbb{R}^d)^{m_i}$,

$$\Pr[(x_{ij})_{1\leq j\leq m_i} \in S|\mu_i,\bar{x}_i] = \Pr[(x_{ij})_{1\leq j\leq m_i} \in S|\bar{x}_i]. \tag{48}$$

Since $(x_{ij})_{j=1,\ldots,m_i}$ and V are conditionally independent given μ_i,

$$\Pr[(x_{ij})_{1\leq j\leq m_i} \in S|V,\mu_i,\bar{x}_i] = \Pr[(x_{ij})_{1\leq j\leq m_i} \in S|\mu_i,\bar{x}_i]. \tag{49}$$

Combining (48) and (49), $(x_{ij})_{1\leq j\leq m_i}$ and V are conditionally independent given $\bar{x}_i$. By (22),

$$I(V;(x_{ij})_{1\leq j\leq m_i}) = I(V;(x_{ij})_{1\leq j\leq m_i},\bar{x}_i) = I(V;\bar{x}_i). \tag{50}$$

Since $\bar{x}_i = \mu_i + \frac{1}{m_i}\sum_{j=1}^{m_i} z_j$, the conditional distribution of $\bar{x}_i$ given V is $N(0,\sigma^2 VV^\mathsf{T} + \frac{\eta_i^2}{m_i}I)$. By Lemma 3.8 and inequality (44),

$$I(V;\bar{x}_i) \leq \sup_{V',V''\in\mathcal{V}} D_{\mathrm{kl}}\left(N\left(0,\sigma^2 V'(V')^\mathsf{T} + \frac{\eta_i^2}{m_i}I\right)\middle\|N\left(0,\sigma^2 V''(V'')^\mathsf{T} + \frac{\eta_i^2}{m_i}I\right)\right) \leq O(t^2\gamma_i). \tag{51}$$

Combining (47), (50), and (51),

$$I(V;(x_{ij})_{k\leq i\leq n,1\leq j\leq m_i}) \leq O\left(t^2\sum_{i=k}^{n}\gamma_i\right).$$

The theorem is proved by plugging this into (46) and noting that $\log|\mathcal{V}| \geq 2(d-k)$. □

L Our Results in the Linear Models Setting

In the linear models setting (see Section 4), our estimator is the subspace $\hat{\Gamma}$ spanned by the top-k eigenvectors of A defined in (3). We use θ to denote the maximum principal angle between our estimator $\hat{\Gamma}$ and the true subspace Γ. To describe our guarantee on θ, it is convenient to first define the following quantities for every $i = 1,\ldots,n$:

$$p_i = \frac{(d+\log(n/\delta))(m_i+\log(n/\delta))}{m_i^2}, \quad q_i = \sqrt{\frac{d+\log(n/\delta)}{m_i}} + p_i.$$

We prove the following guarantees for our estimator in Appendices L.1 and L.2:

Theorem L.1. *There exists an absolute constant $C > 0$ such that for any $\delta \in (0, 1/2)$, with probability at least $1 - \delta$,*

$$\sin\theta \le C\sigma_k^{-2}\sqrt{\sum_{i=1}^n w_i^2 d\log(d/\delta)\left(\frac{\|\beta_i\|_2^4 + \eta_i^2\|\beta_i\|_2^2}{m_i} + \frac{\eta_i^4}{m_i^2}\right)} + C\sigma_k^{-2}\max_i w_i \log(d/\delta)(\|\beta_i\|_2^2 q_i + \eta_i\|\beta_i\|_2 q_i + \eta_i^2 p_i).$$

Corollary L.2. *Suppose $m \in \mathbb{Z}_{>0}$ and $\eta \in \mathbb{R}_{>0}$ satisfy $2 \le m \le \min_{1\le i\le n} m_i$ and $\max_{1\le i\le n}\eta_i \le \eta$. Choosing $w_1 = \cdots = w_n = 1/n$, for any $\delta \in (0, 1/2)$, with probability at least $1 - \delta$,*

$$\sin\theta \le O\left(\log^3(nd/\delta)\left(\sqrt{\frac{d(\frac{1}{n}\sum_{i=1}^n \|\beta_i\|_2^4 + \frac{1}{n}\sum_{i=1}^n \eta^2\|\beta_i\|_2^2 + \eta^4/m)}{mn\sigma_k^4}} + \frac{d\max_i\|\beta_i\|_2^2}{mn\sigma_k^2}\right)\right). \tag{52}$$

If we further assume that $\max_{1\le i\le n}\|\beta_i\|_2 \le r$, then (52) implies

$$\sin\theta \le O\left(\log^3(nd/\delta)\left(\sqrt{\frac{d(r^4 + r^2\eta^2 + \eta^4/m)}{mn\sigma_k^4}}\right)\right). \tag{53}$$

In [Tripuraneni et al., 2021], it is further assumed that $\eta = \Theta(1), r = \Theta(1), \|\beta_i\|_2 = \Theta(1), \delta = (mn)^{-100}$. Defining $\kappa = \frac{1}{n}\sum_{i=1}^n \|\beta_i\|_2^2/k\sigma_k^2 = \Theta(1/k\sigma_k^2)$ as in [Tripuraneni et al., 2021] and using the fact $\sum_{i=1}^n \|\beta_i\|_2^4 \le \left(\max_i \|\beta_i\|_2^2\right)\sum_{i=1}^n \|\beta_i\|_2^2$, our bound becomes

$$\sin\theta \le O\left(\sqrt{\frac{d}{mn\sigma_k^4}}\log^3(mn)\right) = O\left(\sqrt{\frac{\kappa}{\sigma_k^2}\frac{dk}{mn}}\log^3(mn)\right),$$

matching the bound in [Tripuraneni et al., 2021, Theorem 3].

As in the PCA setting, our proof of Theorem L.1 is based on the Davis-Kahan $\sin\theta$ theorem (Theorem 2.1) and a bound on the spectral norm of the difference between the matrix A and its expectation. We prove and make crucial use of a generalization of the matrix Bernstein inequality (Lemma C.5) which turns out to be slightly more convenient than a similar inequality used in [Tripuraneni et al., 2021].

L.1 Proof of Theorem L.1

Plugging $y_{ij} = x_{ij}^\top\beta_i + z_{ij}$ into the definition of A,

$$A = \sum_{i=1}^n \frac{w_i}{m_i(m_i-1)}\sum_{j_1\ne j_2} x_{ij_1}(x_{ij_1}^\top\beta_i + z_{ij_1})(\beta_i^\top x_{ij_2} + z_{ij_2})x_{ij_2}^\top.$$

When $j_1 \ne j_2$, conditioned on x_{ij_1} and x_{ij_2}, the noise terms z_{ij_1} and z_{ij_2} are independent and have zero mean. Therefore,

$$\mathbb{E}[x_{ij_1}(x_{ij_1}^\top\beta_i + z_{ij_1})(\beta_i^\top x_{ij_2} + z_{ij_2})x_{ij_2}^\top] = \mathbb{E}[x_{ij_1}x_{ij_1}^\top\beta_i\beta_i^\top x_{ij_2}x_{ij_2}^\top] = \mathbb{E}[\beta_i\beta_i^\top],$$

where we used the fact that x_{ij_1} and x_{ij_2} are independent and have zero mean and identity covariance matrix. Thus, the expectation of A is $\bar{A} = \sum_{i=1}^n w_i\beta_i\beta_i^\top$, and as in the PCA setting, our goal is to bound the spectral norm of the difference $A - \bar{A}$. We decompose A as

$$A = E + F + F^\top + G, \text{ where} \tag{54}$$

$$E = \sum_{i=1}^n \frac{w_i}{m_i(m_i-1)}\sum_{j_1\ne j_2} x_{ij_1}x_{ij_1}^\top\beta_i\beta_i^\top x_{ij_2}x_{ij_2}^\top,$$

$$F = \sum_{i=1}^n \frac{w_i}{m_i(m_i-1)}\sum_{j_1\ne j_2} x_{ij_1}x_{ij_1}^\top\beta_i z_{ij_2}x_{ij_2}^\top,$$

$$G = \sum_{i=1}^{n} \frac{w_i}{m_i(m_i-1)} \sum_{j_1 \neq j_2} x_{ij_1} z_{ij_1} z_{ij_2} x_{ij_2}^{\mathsf{T}}.$$

For every $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, define $b_{ij} = z_{ij} x_{ij}$ and $h_{ij} = x_{ij} x_{ij}^{\mathsf{T}} \beta_i$. Now we can rewrite E, F, G as $E = E_1 - E_2, F = F_1 - F_2, G = G_1 - G_2$, where

$$E_1 = \sum_{i=1}^{n} \frac{w_i}{m_i(m_i-1)} \left(\sum_{j=1}^{m_i} h_{ij} \right) \left(\sum_{j=1}^{m_i} h_{ij} \right)^{\mathsf{T}}, \qquad E_2 = \sum_{i=1}^{n} \frac{w_i}{m_i(m_i-1)} \sum_{j=1}^{m_i} h_{ij} h_{ij}^{\mathsf{T}},$$

$$F_1 = \sum_{i=1}^{n} \frac{w_i}{m_i(m_i-1)} \left(\sum_{j=1}^{m_i} h_{ij} \right) \left(\sum_{j=1}^{m_i} b_{ij} \right)^{\mathsf{T}}, \qquad F_2 = \sum_{i=1}^{n} \frac{w_i}{m_i(m_i-1)} \sum_{j=1}^{m_i} h_{ij} b_{ij}^{\mathsf{T}},$$

$$G_1 = \sum_{i=1}^{n} \frac{w_i}{m_i(m_i-1)} \left(\sum_{j=1}^{m_i} b_{ij} \right) \left(\sum_{j=1}^{m_i} b_{ij} \right)^{\mathsf{T}}, \qquad G_2 = \sum_{i=1}^{n} \frac{w_i}{m_i(m_i-1)} \sum_{j=1}^{m_i} b_{ij} b_{ij}^{\mathsf{T}}.$$

Define

$$r_i = \frac{(d + \log(m_i n/\delta)) \log(m_i n/\delta)}{m_i^2}, \quad s_i = \sqrt{\frac{d + \log(m_i n/\delta)}{m_i^4}} + r_i.$$

The following lemma controls the deviation of $E_1, E_2, F_1, F_2, G_1, G_2$ from their expectations $\bar{E}_1$, $\bar{E}_2, \bar{F}_1, \bar{F}_2, \bar{G}_1, \bar{G}_2$ in spectral norm:

Lemma L.3. *There exists an absolute constant $C > 0$ such that for any $\delta \in (0, 1/2)$, each of the following inequality holds with probability at least $1 - \delta$:*

$$\|E_1 - \bar{E}_1\| \leq C \sqrt{\sum_{i=1}^{n} \frac{w_i^2 \|\beta_i\|_2^4 d \log(d/\delta)}{m_i}} + C \max_i w_i \|\beta_i\|_2^2 q_i \log(d/\delta), \tag{55}$$

$$\|E_2 - \bar{E}_2\| \leq C \sqrt{\sum_{i=1}^{n} \frac{w_i^2 \|\beta_i\|_2^4 d \log(d/\delta)}{m_i^3}} + C \max_i w_i \|\beta_i\|_2^2 s_i \log(d/\delta), \tag{56}$$

$$\|F_1 - \bar{F}_1\| \leq C \sqrt{\sum_{i=1}^{n} \frac{w_i^2 \eta_i^2 \|\beta_i\|_2^2 d \log(d/\delta)}{m_i}} + C \max_i w_i \eta_i \|\beta_i\|_2 q_i \log(d/\delta), \tag{57}$$

$$\|F_2 - \bar{F}_2\| \leq C \sqrt{\sum_{i=1}^{n} \frac{w_i^2 \eta_i^2 \|\beta_i\|_2^2 d \log(d/\delta)}{m_i^3}} + C \max_i w_i \eta_i \|\beta_i\|_2 s_i \log(d/\delta), \tag{58}$$

$$\|G_1 - \bar{G}_1\| \leq C \sqrt{\sum_{i=1}^{n} \frac{w_i^2 \eta_i^4 d \log(d/\delta)}{m_i^2}} + C \max_i w_i \eta_i^2 p_i \log(d/\delta), \tag{59}$$

$$\|G_2 - \bar{G}_2\| \leq C \sqrt{\sum_{i=1}^{n} \frac{w_i^2 \eta_i^4 d \log(d/\delta)}{m_i^3}} + C \max_i w_i \eta_i^2 r_i \log(d/\delta). \tag{60}$$

Before we prove Lemma L.3, we first use it to prove Theorem L.1:

Proof of Theorem L.1. By (54),

$$\|A - \bar{A}\| \leq \|E_1 - \bar{E}_1\| + \|E_2 - \bar{E}_2\| + \|F_1 - \bar{F}_1\| + \|F_2 - \bar{F}_2\| + \|G_1 - \bar{G}_1\| + \|G_2 - \bar{G}_2\|.$$

Setting δ in Lemma L.3 to be $\delta/6$ and using the union bound, for an absolute constant $C > 0$, with probability at least $1 - \delta$,

$$\|A - \bar{A}\| \leq C \sqrt{\sum_{i=1}^{n} w_i^2 d \log(d/\delta) \left(\frac{\|\beta_i\|_2^4 + \eta_i^2 \|\beta_i\|_2^2}{m_i} + \frac{\eta_i^4}{m_i^2} \right)}$$

$$+ C \max_i w_i \log(d/\delta)(\|\beta_i\|_2^2 q_i + \eta_i \|\beta_i\|_2 q_i + \eta_i^2 p_i),$$

where we used the fact that the right-hand-sides of (56), (58) and (60) are upper bounded by a constant times the right-hand-sides of (55), (57), (59), respectively. Since $\bar{A} = \sum_{i=1}^n w_i \beta_i \beta_i^\mathsf{T}$, the theorem is proved by

$$\sin\theta \le \frac{2\|A - \bar{A}\|}{\sigma_k^2}$$

due to Theorem 2.1. □

In the lemmas below, we prove helper inequalities that we use to prove Lemma L.3. In these lemmas, we focus on a single user i and thus omit the subscript i.

Lemma L.4. *Let $x_1, \ldots, x_m \in \mathbb{R}^d$ are independent random vectors. Assume every x_i has zero mean and is $O(1)$-sub-Gaussian. Conditioned on $x_1, \ldots, x_m$, let $z_1, \ldots, z_m \in \mathbb{R}$ be η-sub-Gaussian independent random variables with zero mean. Define $b = \sum_{j=1}^m z_j x_j$. For any $\delta \in (0, 1/2)$, with probability at least $1 - \delta$,*

$$\|b\|_2 \le \eta \sqrt{(d + \log(1/\delta))(m + \log(1/\delta))}. \tag{61}$$

Moreover,

$$\|\mathbb{E}[bb^\mathsf{T}]\| \le O(\eta^2 m), \tag{62}$$

$$\|\mathbb{E}[(bb^\mathsf{T})^2]\| \le O(\eta^4 m^2 d). \tag{63}$$

Proof. Define $X = [x_1 \ \cdots \ x_m] \in \mathbb{R}^{d \times m}$ and $z = [z_1 \ \cdots \ z_m]^\mathsf{T} \in \mathbb{R}^m$. Now $b = Xz$. By Lemma C.3, with probability at least $1 - \delta/2$,

$$\|X\| \le O\left(\sqrt{d + m + \log(1/\delta)}\right). \tag{64}$$

Let r denote the rank of X. By Lemma C.4, with probability at least $1 - \delta/2$,

$$\|b\|_2 = \|Xz\|_2 \le O\left(\eta \|X\| \sqrt{r + \log(1/\delta)}\right) \le O\left(\eta \|X\| \sqrt{\min\{d, m\} + \log(1/\delta)}\right). \tag{65}$$

Combining (64) and (65) using the union bound, with probability at least $1 - \delta$,

$$\|b\|_2 \le O\left(\eta \sqrt{(d + m + \log(1/\delta))(\min\{d, m\} + \log(1/\delta))}\right).$$

This proves (61).

For every fixed unit vector $u \in \mathbb{R}^d$,

$$u^\mathsf{T} \mathbb{E}[bb^\mathsf{T}] u = \sum_{j=1}^m u^\mathsf{T} x_j z_j^2 x_j^\mathsf{T} u = \sum_{j=1}^m z_j^2 (x_j^\mathsf{T} u)^2 \le \sum_{j=1}^m O(\eta^2) = O(\eta^2 m).$$

This proves (62).

$$u^\mathsf{T} \mathbb{E}[(bb^\mathsf{T})^2] u = \sum_{j_1, j_2, j_3, j_4} \mathbb{E}[u^\mathsf{T} z_{j_1} x_{j_1} x_{j_2}^\mathsf{T} z_{j_2} z_{j_3} x_{j_3} x_{j_4}^\mathsf{T} z_{j_4} u]$$

Since $z_1, \ldots, z_m$ are independent and have zero mean when conditioned on $x_1, \ldots, x_m$, the term $\mathbb{E}[u^\mathsf{T} z_{j_1} x_{j_1} x_{j_2}^\mathsf{T} z_{j_2} z_{j_3} x_{j_3} x_{j_4}^\mathsf{T} z_{j_4} u]$ is zero unless (j_1, j_2, j_3, j_4) belongs to

$$S := \{(j_1, j_2, j_3, j_4) \in \{1, \ldots, m\}^4 : (j_1 = j_2 \wedge j_3 = j_4) \vee (j_1 = j_3 \wedge j_2 = j_4) \vee (j_1 = j_4 \wedge j_2 = j_3)\}.$$

Letting $e_1, \ldots, e_d$ be an orthonormal bases of $\mathbb{R}^d$, we have

$$u^\mathsf{T} \mathbb{E}[(bb^\mathsf{T})^2] u = \sum_{(j_1, j_2, j_3, j_4) \in S} \sum_{\ell=1}^d \mathbb{E}[z_{j_1} z_{j_2} z_{j_3} z_{j_4} (u^\mathsf{T} x_{j_1})(x_{j_2}^\mathsf{T} e_\ell)(e_\ell^\mathsf{T} x_{j_3})(x_{j_4}^\mathsf{T} u)]$$

$$
\begin{aligned}
&\leq \sum_{(j_1,j_2,j_3,j_4)\in S} \sum_{\ell=1}^{d} \frac{\eta^4}{8} \mathbb{E}[(z_{j_1}/\eta)^8 + (z_{j_2}/\eta)^8 + (z_{j_3}/\eta)^8 + (z_{j_4}/\eta)^8 \\
&\quad + (u^\mathsf{T} x_{j_1})^8 + (x_{j_2}^\mathsf{T} e_\ell)^8 + (e_\ell^\mathsf{T} x_{j_3})^8 + (x_{j_4}^\mathsf{T} u)^8] \\
&\leq \sum_{(j_1,j_2,j_3,j_4)\in S} \sum_{\ell=1}^{d} O(\eta^4) \\
&\leq O(\eta^4 m^2 d).
\end{aligned}
$$

This proves (63). □

Lemma L.5. *In the same setting as the lemma above, define $h = \sum_{j=1}^m x_j x_j^\mathsf{T} \beta$ for a vector $\beta \in \mathbb{R}^d$. For any $\delta \in (0, 1/2)$, with probability at least $1-\delta$,*

$$
\|h\|_2 \leq O\left(\|\beta\|_2 \sqrt{(d+m+\log(1/\delta))(m+\log(1/\delta))}\right). \tag{66}
$$

Also, with probability at least $1-\delta$,

$$
\|hh^\mathsf{T} - \mathbb{E}[hh^\mathsf{T}]\| \leq O\left(\|\beta\|_2^2 \left(m\sqrt{m(d+\log(1/\delta)} + (d+\log(1/\delta))(m+\log(1/\delta))\right)\right) \tag{67}
$$

Moreover,

$$
\|\mathbb{E}[(hh^\mathsf{T} - \mathbb{E}[hh^\mathsf{T}])^2]\| \leq O(\|\beta\|_2^4 m^3 d). \tag{68}
$$

Proof. Define $X = [x_1 \ \cdots \ x_m] = \mathbb{R}^{d\times m}$. We have $h = XX^\mathsf{T}\beta$. Since each of the m entries in $X^\mathsf{T}\beta$ has sub-Gaussian constant $O(\|\beta\|_2)$, by Lemma C.3, with probability at least $1-\delta/2$, $\|X^\mathsf{T}\beta\|_2 \leq O(\|\beta\|_2\sqrt{m+\log(1/\delta)})$. By Lemma C.3, with probability at least $1-\delta/2$, $\|X\| \leq O(\sqrt{d+m+\log(1/\delta)})$. This proves (66) by a union bound.

We show that with probability at least $1-\delta$,

$$
\|h - \mathbb{E}[h]\|_2 \leq O\left(\|\beta\|_2 \sqrt{(d+\log(1/\delta))(m+\log(1/\delta))}\right). \tag{69}
$$

If $d \geq m$, this follows from (66) and

$$
\|\mathbb{E}[h]\|_2 = O(m\|\beta\|_2). \tag{70}
$$

If $d \leq m$, with probability at least $1-\delta$,

$$
\begin{aligned}
\|h - \mathbb{E}[h]\|_2 &= \|(XX^\mathsf{T} - \mathbb{E}[XX^\mathsf{T}])\beta\|_2 \\
&\leq \|XX^\mathsf{T} - \mathbb{E}[XX^\mathsf{T}]\| \cdot \|\beta\|_2 \\
&\leq O\left(\|\beta\|_2 \sqrt{(d+\log(1/\delta))(m+d+\log(1/\delta))}\right) && \text{(by Lemma C.2)} \\
&\leq O\left(\|\beta\|_2 \sqrt{(d+\log(1/\delta))(m+\log(1/\delta))}\right).
\end{aligned}
$$

Therefore, inequality (69) holds with probability at least $1-\delta$. Now we show that $\|\mathbb{E}[h]\mathbb{E}[h]^\mathsf{T} - \mathbb{E}[hh^\mathsf{T}]\| \leq O(m\|\beta\|_2^2)$. Indeed,

$$
\mathbb{E}[h]\mathbb{E}[h]^\mathsf{T} - \mathbb{E}[hh^\mathsf{T}] = \sum_{j_1,j_2} \left(\mathbb{E}[x_{j_1}x_{j_1}^\mathsf{T}\beta]\mathbb{E}[x_{j_2}x_{j_2}^\mathsf{T}\beta]^\mathsf{T} - \mathbb{E}[(x_{j_1}x_{j_1}^\mathsf{T}\beta)(x_{j_2}x_{j_2}^\mathsf{T}\beta)^\mathsf{T}]\right).
$$

Since x_{j_1} and x_{j_2} are independent when $j_1 \neq j_2$,

$$
\mathbb{E}[h]\mathbb{E}[h]^\mathsf{T} - \mathbb{E}[hh^\mathsf{T}] = \sum_{j=1}^{m} \left(\mathbb{E}[x_j x_j^\mathsf{T}\beta]\mathbb{E}[x_j x_j^\mathsf{T}\beta]^\mathsf{T} - \mathbb{E}[(x_j x_j^\mathsf{T}\beta)(x_j x_j^\mathsf{T}\beta)^\mathsf{T}]\right). \tag{71}
$$

It is clear that $\mathbb{E}[x_j x_j^\mathsf{T}\beta] = \beta$, and for every unit vector $u \in \mathbb{R}^d$,

$$
\begin{aligned}
&u^\mathsf{T}\mathbb{E}[(x_j x_j^\mathsf{T}\beta)(x_j x_j^\mathsf{T}\beta)^\mathsf{T}]u \\
&= \mathbb{E}[(u^\mathsf{T} x_j)(x_j^\mathsf{T}\beta)(\beta^\mathsf{T} x_j)(x_j^\mathsf{T} u)]
\end{aligned}
$$

$$\leq \frac{\|\beta\|_2^2}{4}\mathbb{E}[(u^\mathsf{T} x_j)^4 + (x_j^\mathsf{T}(\beta/\|\beta\|_2))^4 + ((\beta/\|\beta\|_2)^\mathsf{T} x_j)^4 + (x_j^\mathsf{T} u)^4]$$
$$\leq O(\|\beta\|_2^2).$$

This implies that $\|\mathbb{E}[(x_j x_j^\mathsf{T}\beta)(x_j x_j^\mathsf{T}\beta)^\mathsf{T}]\| \leq O(\|\beta\|_2^2)$. Now it is clear from (71) that $\|\mathbb{E}[h]\mathbb{E}[h]^\mathsf{T} - \mathbb{E}[hh^\mathsf{T}]\| \leq O(m\|\beta\|_2^2)$, which implies

$$\begin{aligned}\|hh^\mathsf{T} - \mathbb{E}[hh^\mathsf{T}]\| &\leq O(m\|\beta\|_2^2) + \|hh^\mathsf{T} - \mathbb{E}[h]\mathbb{E}[h]^\mathsf{T}\| \\ &\leq O(m\|\beta\|_2^2) + \|(h-\mathbb{E}[h])h^\mathsf{T}\| + \|\mathbb{E}[h](h-\mathbb{E}[h])^\mathsf{T}\| \\ &\leq O(m\|\beta\|_2^2) + \|h-\mathbb{E}[h]\|_2(\|h\|_2 + \|\mathbb{E}[h]\|_2).\end{aligned}$$

Plugging (69), (66), and (70) into the inequality above, we get (67).

Finally,

$$\mathbb{E}[(hh^\mathsf{T} - \mathbb{E}[hh^\mathsf{T}])^2] = \mathbb{E}[(hh^\mathsf{T})^2] - \mathbb{E}[hh^\mathsf{T}]^2 = \sum_{j_1,j_2,j_3,j_4} H_{j_1 j_2 j_3 j_4}, \tag{72}$$

where

$$H_{j_1 j_2 j_3 j_4} = \mathbb{E}[x_{j_1}x_{j_1}^\mathsf{T}\beta\beta^\mathsf{T}x_{j_2}x_{j_2}^\mathsf{T}x_{j_3}x_{j_3}^\mathsf{T}\beta\beta^\mathsf{T}x_{j_4}x_{j_4}^\mathsf{T}] - \mathbb{E}[x_{j_1}x_{j_1}^\mathsf{T}\beta\beta^\mathsf{T}x_{j_2}x_{j_2}^\mathsf{T}]\mathbb{E}[x_{j_3}x_{j_3}^\mathsf{T}\beta\beta^\mathsf{T}x_{j_4}x_{j_4}^\mathsf{T}].$$

When j_1, j_2, j_3, j_4 are distinct, $H_{j_1 j_2 j_3 j_4}$ is zero. When j_1, j_2, j_3, j_4 are not distinct, letting $e_1, \ldots, e_d$ be an orthonormal basis for $\mathbb{R}^d$, for every fixed unit vector $u \in \mathbb{R}^d$, we have

$$\begin{aligned}&u^\mathsf{T}\mathbb{E}[x_{j_1}x_{j_1}^\mathsf{T}\beta\beta^\mathsf{T}x_{j_2}x_{j_2}^\mathsf{T}x_{j_3}x_{j_3}^\mathsf{T}\beta\beta^\mathsf{T}x_{j_4}x_{j_4}^\mathsf{T}]u \\ =& \sum_{\ell=1}^d \mathbb{E}[(u^\mathsf{T}x_{j_1})(x_{j_1}^\mathsf{T}\beta)(\beta^\mathsf{T}x_{j_2})(x_{j_2}^\mathsf{T}e_\ell)(e_\ell^\mathsf{T}x_{j_3})(x_{j_3}^\mathsf{T}\beta)(\beta^\mathsf{T}x_{j_4})(x_{j_4}^\mathsf{T}u)] \\ \leq& \sum_{\ell=1}^d \frac{\|\beta\|_2^4}{8}\mathbb{E}[(u^\mathsf{T}x_{j_1})^8 + (x_{j_1}^\mathsf{T}(\beta/\|\beta\|_2))^8 + ((\beta/\|\beta\|_2)^\mathsf{T}x_{j_2})^8 + (x_{j_2}^\mathsf{T}e_\ell)^8 \\ &+ (e_\ell^\mathsf{T}x_{j_3})^8 + (x_{j_3}^\mathsf{T}(\beta/\|\beta\|_2))^8 + ((\beta/\|\beta\|_2)^\mathsf{T}x_{j_4})^8 + (x_{j_4}^\mathsf{T}u)^8] \\ =& O(\|\beta\|_2^4 d).\end{aligned}$$

This implies $\|\mathbb{E}[x_{j_1}x_{j_1}^\mathsf{T}\beta\beta^\mathsf{T}x_{j_2}x_{j_2}^\mathsf{T}x_{j_3}x_{j_3}^\mathsf{T}\beta\beta^\mathsf{T}x_{j_4}x_{j_4}^\mathsf{T}]\| = O(\|\beta\|_2^4 d)$. Similarly, $\|\mathbb{E}[x_{j_1}x_{j_1}^\mathsf{T}\beta\beta^\mathsf{T}x_{j_2}x_{j_2}^\mathsf{T}]\| = O(\|\beta\|_2^2)$ because

$$\begin{aligned}u^\mathsf{T}\mathbb{E}[x_{j_1}x_{j_1}^\mathsf{T}\beta\beta^\mathsf{T}x_{j_2}x_{j_2}^\mathsf{T}]u &= \mathbb{E}[(u^\mathsf{T}x_{j_1})(x_{j_1}^\mathsf{T}\beta)(\beta^\mathsf{T}x_{j_2})(x_{j_2}^\mathsf{T}u)] \\ &\leq \frac{\|\beta\|_2^2}{4}\mathbb{E}[(u^\mathsf{T}x_{j_1})^4 + (x_{j_1}^\mathsf{T}(\beta/\|\beta\|_2))^4 + ((\beta/\|\beta\|_2)^\mathsf{T}x_{j_2})^4 + (x_{j_2}^\mathsf{T}u)^4] \\ &\leq O(\|\beta\|_2^2).\end{aligned}$$

Therefore, $\|H_{j_1 j_2 j_3 j_4}\| \leq O(\|\beta\|_2^4 d)$ when j_1, j_2, j_3, j_4 are not distinct. Now (68) follows from (72). □

Lemma L.6. *In the lemma above,*

$$\|\mathbb{E}[hb^\mathsf{T}bh^\mathsf{T}]\| \leq O(\eta^2\|\beta\|_2^2 m^3 d), \text{ and } \|\mathbb{E}[bh^\mathsf{T}hb^\mathsf{T}]\| \leq O(\eta^2\|\beta\|_2^2 m^3 d). \tag{73}$$

Proof. For every unit vector u,

$$u^\mathsf{T}\mathbb{E}[hb^\mathsf{T}bh^\mathsf{T}]u = \sum_{j_1,j_2,j_3,j_4} \mathbb{E}[u^\mathsf{T}x_{j_1}x_{j_1}^\mathsf{T}\beta z_{j_2}x_{j_2}^\mathsf{T}x_{j_3}z_{j_3}\beta x_{j_4}x_{j_4}^\mathsf{T}u] \tag{74}$$

If $j_2 \neq j_3$, we have $\mathbb{E}[z_{j_2}z_{j_3}|x_{j_1}, x_{j_2}, x_{j_3}, x_{j_4}] = 0$ and thus $\mathbb{E}[u^\mathsf{T}x_{j_1}x_{j_1}^\mathsf{T}\beta z_{j_2}x_{j_2}^\mathsf{T}x_{j_3}z_{j_3}\beta x_{j_4}x_{j_4}^\mathsf{T}u] = 0$. If $j_2 = j_3$, letting $e_1, \ldots, e_d$ be an orthonormal basis for $\mathbb{R}^d$,

$$\begin{aligned}&\mathbb{E}[u^\mathsf{T}x_{j_1}x_{j_1}^\mathsf{T}\beta z_{j_2}x_{j_2}^\mathsf{T}x_{j_3}z_{j_3}\beta x_{j_4}x_{j_4}^\mathsf{T}u] \\ =& \sum_{\ell=1}^d \mathbb{E}[z_{j_2}^2(u^\mathsf{T}x_{j_1})(x_{j_1}^\mathsf{T}\beta)(x_{j_2}^\mathsf{T}e_\ell)(e_\ell^\mathsf{T}x_{j_3})(\beta^\mathsf{T}x_{j_4})(x_{j_4}^\mathsf{T}u)]\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\|\beta\|_2^2}{6} \sum_{\ell=1}^{d} \mathbb{E}[z_{j_2}^2((u^\mathsf{T} x_{j_1})^6 + (x_{j_1}^\mathsf{T}(\beta/\|\beta\|_2))^6 + (x_{j_2}^\mathsf{T} e_\ell)^6 \\
&\quad + (e_\ell^\mathsf{T} x_{j_3})^6 + ((\beta/\|\beta\|_2)^\mathsf{T} x_{j_4})^6 + (x_{j_4}^\mathsf{T} u)^6)] \\
&\leq O(d\eta^2\|\beta\|_2^2).
\end{aligned}
$$

Plugging this into (74) and noting that the size of $\{(j_1, j_2, j_3, j_4) \in \{1, \ldots, m\}^4 : j_2 = j_3\}$ is $O(m^3)$, we get $\|\mathbb{E}[hb^\mathsf{T} bh^\mathsf{T}]\| \leq O(\eta^2\|\beta\|_2^2 m^3 d)$.

Similarly,

$$
u^\mathsf{T}\mathbb{E}[bh^\mathsf{T} hb^\mathsf{T}]u = \sum_{j_1,j_2,j_3,j_4} \mathbb{E}[u^\mathsf{T} z_{j_1} x_{j_1} \beta^\mathsf{T} x_{j_2} x_{j_2}^\mathsf{T} x_{j_3} x_{j_3}^\mathsf{T} \beta z_{j_4} x_{j_4}^\mathsf{T} u]. \tag{75}
$$

If $j_1 \neq j_4$, then $\mathbb{E}[u^\mathsf{T} z_{j_1} x_{j_1} \beta^\mathsf{T} x_{j_2} x_{j_2}^\mathsf{T} x_{j_3} x_{j_3}^\mathsf{T} \beta z_{j_4} x_{j_4}^\mathsf{T} u] = 0$. When $j_1 = j_4$,

$$
\begin{aligned}
&\mathbb{E}[u^\mathsf{T} z_{j_1} x_{j_1} \beta^\mathsf{T} x_{j_2} x_{j_2}^\mathsf{T} x_{j_3} x_{j_3}^\mathsf{T} \beta z_{j_4} x_{j_4}^\mathsf{T} u] \\
&= \sum_{\ell=1}^{d} \mathbb{E}[z_{j_1}^2 (u^\mathsf{T} x_{j_1})(\beta^\mathsf{T} x_{j_2})(x_{j_2}^\mathsf{T} e_\ell)(e_\ell^\mathsf{T} x_{j_3})(x_{j_3}^\mathsf{T} \beta)(x_{j_4}^\mathsf{T} u)] \\
&\leq \frac{\|\beta\|_2^2}{6} \sum_{\ell=1}^{d} \mathbb{E}[z_{j_1}^2((u^\mathsf{T} x_{j_1})^6 + ((\beta/\|\beta\|_2)^\mathsf{T} x_{j_2})^6 + (x_{j_2}^\mathsf{T} e_\ell)^6 \\
&\quad + (e_\ell^\mathsf{T} x_{j_3})^6 + (x_{j_3}^\mathsf{T}(\beta/\|\beta\|_2))^6 + (x_{j_4}^\mathsf{T} u)^6)] \\
&= O(d\eta^2\|\beta\|_2^2).
\end{aligned}
$$

Plugging this into (75) proves $\|\mathbb{E}[bh^\mathsf{T} hb^\mathsf{T}]\| \leq O(\eta^2\|\beta\|_2^2 m^3 d)$. □

We can now finish the proof of Lemma L.3.

Proof of Lemma L.3. To bound $\|E_1 - \bar{E}_1\|$, we set $Z_i' = \frac{w_i}{m_i(m_i-1)} \left(\sum_{j=1}^{m_i} h_{ij}\right) \left(\sum_{j=1}^{m_i} h_{ij}\right)^\mathsf{T}$ and $Z_i = Z_i' - \mathbb{E}[Z_i']$. By (67), with probability at least $1 - \delta/(2n)$,

$$
\|Z_i\| \leq O\left(w_i \|\beta_i\|_2^2 q_i\right).
$$

Also, by (68),

$$
\|\mathbb{E}[Z_i Z_i^\mathsf{T}]\| = \|\mathbb{E}[Z_i^\mathsf{T} Z_i]\| = O\left(\frac{w_i^2 \|\beta_i\|_2^4 d}{m_i}\right).
$$

Lemma C.5 proves that (55) holds with probability at least $1 - \delta$.

To bound $\|E_2 - \bar{E}_2\|$, we set $Z_{ij}' = \frac{w_i}{m_i(m_i-1)} h_{ij} h_{ij}^\mathsf{T}$ and $Z_{ij} = Z_{ij}' - \mathbb{E}[Z_{ij}']$. By (67), with probability at least $1 - \delta/(2nm_i)$,

$$
\|Z_{ij}\| \leq O(w_i \|\beta_i\|_2^2 s_i).
$$

Also, by (68),

$$
\|\mathbb{E}[Z_{ij} Z_{ij}^\mathsf{T}]\| = \|\mathbb{E}[Z_{ij}^\mathsf{T} Z_{ij}]\| = O\left(\frac{w_i^2 \|\beta_i\|_2^4 d}{m_i^4}\right).
$$

Lemma C.5 proves that (56) holds with probability at least $1 - \delta$.

To bound $\|F_1 - \bar{F}_1\|$, we set $Z_i = \frac{w_i}{m_i(m_i-1)} \left(\sum_{j=1}^{m_i} h_{ij}\right) \left(\sum_{j=1}^{m_i} b_{ij}\right)^\mathsf{T}$. Note that $\mathbb{E}[Z_i] = 0$. By (61) and (66), with probability at least $1 - \delta/(2n)$,

$$
\|Z_i\| \leq O\left(w_i \eta_i \|\beta_i\|_2 q_i\right).
$$

Also, by (73),

$$
\max\left\{\|\mathbb{E}[Z_i Z_i^\mathsf{T}]\|, \|\mathbb{E}[Z_i^\mathsf{T} Z_i]\|\right\} = O\left(\frac{w_i^2 \eta_i^2 \|\beta_i\|_2^2 d}{m_i}\right).
$$

Lemma C.5 proves that (57) holds with probability at least $1 - \delta$.

To bound $\|F_2 - \bar{F}_2\|$, we set $Z_{ij} = \frac{w_i}{m_i(m_i-1)} h_{ij} b_{ij}^\mathsf{T}$. Note that $\mathbb{E}[Z_{ij}] = 0$. By (61) and (66), with probability at least $1 - \delta/(2nm_i)$,

$$\|Z_{ij}\| \le O(w_i \eta_i \|\beta_i\|_2 s_i).$$

Also, by (73),

$$\max\left\{\|\mathbb{E}[Z_{ij} Z_{ij}^\mathsf{T}]\|, \|\mathbb{E}[Z_{ij}^\mathsf{T} Z_{ij}]\|\right\} = O\left(\frac{w_i^2 \eta_i^2 \|\beta_i\|_2^2 d}{m_i^4}\right).$$

Lemma C.5 proves that (58) holds with probability at least $1 - \delta$.

To bound $\|G_1 - \bar{G}_1\|$, we set $Z_i' = \frac{w_i}{m_i(m_i-1)} \left(\sum_{j=1}^{m_i} b_{ij}\right) \left(\sum_{j=1}^{m_i} b_{ij}\right)^\mathsf{T}$ and $Z_i = Z_i' - \mathbb{E}[Z_i']$. By (61) and (62), with probability at least $1 - \delta/(2n)$,

$$\|Z_i\| \le O\left(w_i \eta_i^2 p_i\right).$$

Also, by (63),

$$\|\mathbb{E}[Z_i Z_i^\mathsf{T}]\| = \|\mathbb{E}[Z_i^\mathsf{T} Z_i]\| = \|\mathbb{E}[Z_i^2]\| \le \|\mathbb{E}[(Z_i')^2]\| = O\left(\frac{w_i^2 \eta^4 d}{m_i^2}\right).$$

Lemma C.5 proves that (59) holds with probability at least $1 - \delta$.

To bound $\|G_2 - \bar{G}_2\|$, we set $Z_{ij}' = \frac{w_i}{m_i(m_i-1)} b_{ij} b_{ij}^\mathsf{T}$ and $Z_{ij} = Z_{ij}' - \mathbb{E}[Z_{ij}']$. By (61) and (62), with probability at least $1 - \delta/(2nm_i)$,

$$\|Z_{ij}\| \le O(w_i \eta_i^2 r_i).$$

Also, by (63),

$$\|\mathbb{E}[Z_{ij} Z_{ij}^\mathsf{T}]\| = \|\mathbb{E}[Z_{ij}^\mathsf{T} Z_{ij}]\| = \|\mathbb{E}[Z_{ij}^2]\| \le \|\mathbb{E}[(Z_{ij}')^2]\| = O\left(\frac{w_i^2 \eta_i^4 d}{m_i^4}\right).$$

Lemma C.5 proves that (60) holds with probability at least $1 - \delta$. □

L.2 Proof of Corollary L.2

Proof. We note that

$$p_i \le O\left(\frac{d \log^2(n/\delta)}{m}\right), \quad q_i \le p_i + O\left(\sqrt{\frac{d \log(n/\delta)}{m}}\right).$$

Therefore, by Theorem L.1, with probability at least $1 - \delta$,

$$\begin{aligned}\sin\theta \le\ & O\left(\log(d/\delta)\sqrt{\log(n/\delta)}\sqrt{\frac{d(\frac{1}{n}\sum_{i=1}^n \|\beta_i\|_2^4 + \frac{1}{n}\sum_{i=1}^n \eta^2 \|\beta_i\|_2^2 + \eta^4/m)}{mn\sigma_k^2}}\right) \\ & + O\left(\log(d/\delta)\log^2(n/\delta)\frac{d(\eta^2 + \max_i \|\beta_i\|_2^2)}{mn\sigma_k^2}\right).\end{aligned}$$

Since $\sin\theta \le 1$ always holds, the inequality above implies

$$\begin{aligned}\sin\theta \le\ & O\left(\log^3(nd/\delta)\left(\sqrt{\frac{d(\frac{1}{n}\sum_{i=1}^n \|\beta_i\|_2^4 + \frac{1}{n}\sum_{i=1}^n \eta^2 \|\beta_i\|_2^2 + \eta^4/m)}{mn\sigma_k^4}} + \frac{d \max_i \|\beta_i\|_2^2}{mn\sigma_k^2}\right)\right) \\ & + O\left(\min\left\{1, \frac{d\eta^2 \log(d/\delta)\log^2(n/\delta)}{mn\sigma_k^2}\right\}\right). \qquad (76)\end{aligned}$$

Using $\frac{1}{n}\sum_{i=1}^n \|\beta_i\|_2^2 = \mathrm{tr}(\frac{1}{n}\sum_{i=1}^n \beta_i \beta_i^\mathsf{T}) = \mathrm{tr}(\sum_{i=1}^n w_i \beta_i \beta_i^\mathsf{T}) \ge \sum_{\ell=1}^k \sigma_\ell^2 \ge k\sigma_k^2$, we have

$$\min\left\{1, \frac{d\eta^2 \log(d/\delta)\log^2(n/\delta)}{mn\sigma_k^2}\right\} \le \sqrt{\frac{d\eta^2 \log(d/\delta)\log^2(n/\delta)}{mn\sigma_k^2}}$$

$$\leq \sqrt{\frac{d(\frac{1}{n}\sum_{i=1}^{n}\eta^2\|\beta_i\|_2^2)\log(d/\delta)\log^2(n/\delta)}{mn\sigma_k^4}}.$$

Plugging this into (76) proves (52). When $\|\beta_i\|_2 \leq r$ for every $i = 1, \ldots, n$, (52) implies

$$\sin\theta \leq O\left(\log^3(nd/\delta)\left(\sqrt{\frac{d(r^4 + r^2\eta^2 + \eta^4/m)}{mn\sigma_k^4}} + \frac{dr^2}{mn\sigma_k^2}\right)\right).$$

Since $\sin\theta \leq 1$, the inequality above implies

$$\sin\theta \leq O\left(\log^3(nd/\delta)\left(\sqrt{\frac{d(r^4 + r^2\eta^2 + \eta^4/m)}{mn\sigma_k^4}} + \min\left\{1, \frac{dr^2}{mn\sigma_k^2}\right\}\right)\right). \tag{77}$$

Using $r^2 \geq \frac{1}{n}\sum_{i=1}^{n}\|\beta_i\|_2^2 \geq k\sigma_k^2$,

$$\min\left\{1, \frac{dr^2}{mn\sigma_k^2}\right\} \leq \sqrt{\frac{dr^2}{mn\sigma_k^2}} \leq \sqrt{\frac{dr^4}{mn\sigma_k^4}}.$$

Plugging this into (77) proves (53). □