Data Enrichment: Multi-task Learning in High Dimension with Theoretical Guarantees

Anonymous Authors[1]

Abstract

Given samples from a group of related regression tasks, a data-enriched model describes observations by a common and per-group individual parameters. In high-dimensional regime, each parameter has its own structure such as sparsity or group sparsity. In this paper, we consider the general form of data enrichment where data comes in a fixed but arbitrary number of tasks G and any convex function, e.g., norm, can characterize the structure of both common and individual parameters. We propose an estimator for the high-dimensional data enriched model and investigate its statistical properties. We delineate the sample complexity of our estimator and provide high probability non-asymptotic bound for estimation error of all parameters under a condition weaker than the state-of-the-art. We propose an iterative estimation algorithm with a geometric convergence rate. Overall, we present a first through statistical and computational analysis of inference in the data enriched model.

1. Introduction

Over the past two decades, major advances have been made in estimating structured parameters, e.g., sparse, low-rank, etc., in high-dimensional small sample problems (Donoho, 2006; Candès & Tao, 2010; Friedman et al., 2008). Such estimators consider a suitable (semi) parametric model of the response: $y = \phi(\mathbf{x}, \boldsymbol{\beta}^*) + \omega$ based on n samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is the true parameter of interest. The unique aspect of such high-dimensional setup is that the number of samples $n < p$, and the structure in $\boldsymbol{\beta}^*$, e.g., sparsity, low-rank, makes the estimation possible (Tibshirani, 1996; Candès et al., 2006; Candès & Recht, 2009). In several real world problems, natural grouping among samples arises and learning a single common model $\boldsymbol{\beta}_0$ for all samples or many per group individual models $\boldsymbol{\beta}_g$s are unrealistic. The middle ground model for such a scenario is the *superposition* of common and individual parameters $\boldsymbol{\beta}_0 + \boldsymbol{\beta}_g$ which has been of recent interest in the statistical machine learning community (Gu & Banerjee, 2016) and is known by multiple names. It is a form of multi-task learning (Zhang & Yang, 2017; Jalali et al., 2010) when we consider regression in each group as a task. It is also called data sharing (Gross & Tibshirani, 2016) since information contained in different group is shared through the common parameter $\boldsymbol{\beta}_0$. And finally, it has been called data enrichment (Chen et al., 2015; Asiaee et al., 2018) because we enrich our data set with pooling multiple samples from different but related sources.

In this paper, we consider the following *data enrichment* (DE) model where there is a *common* parameter $\boldsymbol{\beta}_0^*$ shared between all groups plus *individual* per-group parameters $\boldsymbol{\beta}_g^*$ which characterize the deviation of group g:

$$y_{gi} = \phi(\mathbf{x}_{gi}, (\boldsymbol{\beta}_0^* + \boldsymbol{\beta}_g^*)) + \omega_{gi}, \quad g \in \{1, \dots, G\}, \quad (1)$$

where g and i index the group and samples respectively. Note that the DE model is *a system of coupled superposition models*. We specifically focus on the high-dimensional small sample regime for (1) where the number of samples n_g for each group is much smaller than the ambient dimensionality, i.e., $\forall g : n_g \ll p$. Similar to all other high-dimensional models, we assume that the parameters $\boldsymbol{\beta}_g$ are structured, i.e., for suitable convex functions f_g's, $f_g(\boldsymbol{\beta}_g)$ is small. Further, for the technical analysis and proofs, we focus on the case of linear models, i.e., $\phi(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}^T\boldsymbol{\beta}$. The results seamlessly extend to more general non-linear models, e.g., generalized linear models, broad families of semi-parametric and single-index models, non-convex models, etc., using existing results, i.e., how models like LASSO have been extended (e.g. employing ideas such as restricted strong convexity (Negahban & Wainwright, 2012)).

In the context of *Multi-task learning* (MTL), similar models have been proposed which has the general form of $y_{gi} = \mathbf{x}_{gi}^T(\boldsymbol{\beta}_{1g}^* + \boldsymbol{\beta}_{2g}^*) + \omega_{gi}$ where $\mathbf{B}_1 = [\boldsymbol{\beta}_{11}, \dots, \boldsymbol{\beta}_{1G}]$ and $\mathbf{B}_2 = [\boldsymbol{\beta}_{21}, \dots, \boldsymbol{\beta}_{2G}]$ are two parameter matrices (Zhang & Yang, 2017). To capture relation of tasks, different types of constraints are assumed for parameter matrices. For example, (Chen et al., 2012) assumes $\mathbf{B}_1$ and $\mathbf{B}_2$ are sparse

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

and low rank respectively. In this parameter matrix decomposition framework for MLT, the most related work to ours is the one proposed by (Jalali et al., 2010) where authors regularize the regression with $\|\mathbf{B}_1\|_{1,\infty}$ and $\|\mathbf{B}_2\|_{1,1}$ where norms are p,q-norms on rows of matrices. Parameters of $\mathbf{B}_1$ are more general than DE's common parameter when we use $f_0(\boldsymbol{\beta}_0) = \|\boldsymbol{\beta}_0\|_1$. This is because $\|\mathbf{B}_1\|_{1,\infty}$ regularizer enforces shared support of $\boldsymbol{\beta}^*_{1g}$s, i.e., $\text{supp}(\boldsymbol{\beta}^*_{1i}) = \text{supp}(\boldsymbol{\beta}^*_{1j})$ but allows $\boldsymbol{\beta}^*_{1i} \neq \boldsymbol{\beta}^*_{1j}$. Further sparse variation between parameters of different tasks is induced by $\|\mathbf{B}_2\|_{1,1}$ which has an equivalent effect to DE's individual parameters where $f_g(\cdot)$s are l_1-norm. Our analysis of DE framework suggests that it is more data efficient than this setup of (Jalali et al., 2010) because they require every task i to have large enough samples to learn its own common parameters $\boldsymbol{\beta}_i$ while DE shares the common parameter and only requires the *total dataset over all tasks* to be sufficiently large.

The DE model where $\boldsymbol{\beta}_g$'s are sparse has recently gained attention because of its application in wide range of domains such as personalized medicine (Dondelinger & Mukherjee, 2016), sentiment analysis, banking strategy (Gross & Tibshirani, 2016), single cell data analysis (Ollier & Viallon, 2015), road safety (Ollier & Viallon, 2014), and disease subtype analysis (Dondelinger & Mukherjee, 2016). In spite of the recent surge in applying data enrichment framework to different domains, limited advances have been made in understanding the statistical and computational properties of suitable estimators for the data enriched model. In fact, non-asymptotic statistical properties, including sample complexity and statistical rates of convergence, of regularized estimators for the data enriched model is still an open question (Gross & Tibshirani, 2016; Ollier & Viallon, 2014). To the best of our knowledge, the only theoretical guarantee for data enrichment is provided in (Ollier & Viallon, 2015) where authors prove sparsistency of their proposed method under the stringent irrepresentability condition of the design matrix for recovering supports of common and individual parameters. Existing support recovery guarantees (Ollier & Viallon, 2015), sample complexity and l_2 consistency results (Jalali et al., 2010) of related models are restricted to sparsity and l_1-norm, while our estimator and *norm consistency* analysis work for any structure induced by arbitrary convex functions f_g. Moreover, no computational results, such as rates of convergence of the optimization algorithms associated with proposed estimators, exist in the literature.

Notation and Preliminaries: We denote sets by curly $\mathcal{V}$, matrices by bold capital $\mathbf{V}$, random variables by capital V, and vectors by small bold $\mathbf{v}$ letters. We take $[G] = \{0,\dots,G\}$ and $[G]_\setminus = [G] \setminus \{0\}$.

Given G groups and n_g samples in each as $\{\{\mathbf{x}_{gi}, y_{gi}\}_{i=1}^{n_g}\}_{g=1}^{G}$, we can form the per group design matrix $\mathbf{X}_g \in \mathbb{R}^{n_g \times p}$ and output vector $\mathbf{y}_g \in \mathbb{R}^{n_g}$. The total number of samples is $n = \sum_{g=1}^{G} n_g$. The data enriched model takes the following vector form:

$$\mathbf{y}_g = \mathbf{X}_g(\boldsymbol{\beta}^*_0 + \boldsymbol{\beta}^*_g) + \boldsymbol{\omega}_g, \quad \forall g \in [G]_\setminus \tag{2}$$

where each row of $\mathbf{X}_g$ is $\mathbf{x}^T_{gi}$ and $\boldsymbol{\omega}^T_g = (\omega_{g1},\dots,\omega_{gn_g})$ is the noise vector.

A random variable V is sub-Gaussian if its moments satisfies $\forall p \geq 1 : (\mathbb{E}|V|^p)^{1/p} \leq K_2\sqrt{p}$. The minimum value of K_2 is called the sub-Gaussian norm of V, denoted by $\|V\|_{\psi_2}$ (Vershynin, 2012). A random vector $\mathbf{v} \in \mathbb{R}^p$ is sub-Gaussian if the one-dimensional marginals $\langle \mathbf{v}, \mathbf{u}\rangle$ are sub-Gaussian random variables for all $\mathbf{u} \in \mathbb{R}^p$. The sub-Gaussian norm of $\mathbf{v}$ is defined (Vershynin, 2012) as $\|\mathbf{v}\|_{\psi_2} = \sup_{\mathbf{u}\in\mathbb{S}^{p-1}} \|\langle \mathbf{v}, \mathbf{u}\rangle\|_{\psi_2}$. For any set $\mathcal{V} \in \mathbb{R}^p$ the Gaussian width of the set $\mathcal{V}$ is defined as $\omega(\mathcal{V}) = \mathbb{E}_{\mathbf{g}}\left[\sup_{\mathbf{u}\in\mathcal{V}} \langle \mathbf{g}, \mathbf{u}\rangle\right]$ (Vershynin, 2018), where the expectation is over $\mathbf{g} \sim N(\mathbf{0}, \mathbf{I}_{p\times p})$, a vector of independent zero-mean unit-variance Gaussian.

Contributions: We propose the following Data Enrichment (DE) estimator $\hat{\boldsymbol{\beta}}$ for recovering the structured parameters where the structure is induced by *convex* functions $f_g(\cdot)$:

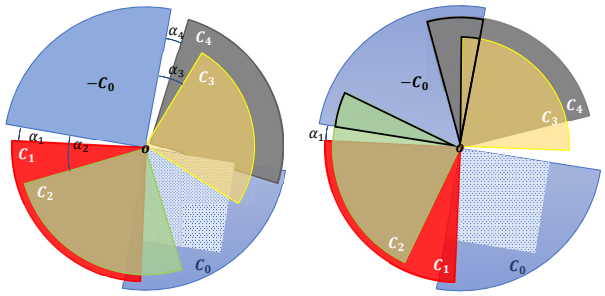
$$\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}^T_0,\dots,\hat{\boldsymbol{\beta}}^T_G) \in \underset{\boldsymbol{\beta}_0,\dots,\boldsymbol{\beta}_G}{\text{argmin}} \frac{1}{n}\sum_{g=1}^{G} \|\mathbf{y}_g - \mathbf{X}_g(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_g)\|_2^2, \quad \text{s.t.} \quad \forall g \in [G] : f_g(\boldsymbol{\beta}_g) \leq f_g(\boldsymbol{\beta}^*_g). \tag{3}$$

We present several statistical and computational results for the DE estimator (3) of the data enriched model:

- The DE estimator (3) succeeds if a geometric condition that we call *Data EnRichment Incoherence Condition* (DERIC) is satisfied, Figure 1b. Compared to other known geometric conditions in the literature such as structural coherence (Gu & Banerjee, 2016) and stable recovery conditions (McCoy & Tropp, 2013), DERIC is a weaker condition, Figure 1a.
- Assuming DERIC holds, we establish a high probability non-asymptotic bound on the weighted sum of parameter-wise estimation error, $\boldsymbol{\delta}_g = \hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}^*_g$ as:

$$\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g\|_2 \leq \gamma O\left(\frac{\max_{g\in[G]} \omega(\mathcal{C}_g \cap \mathbb{S}^{p-1})}{\sqrt{n}}\right), \tag{4}$$

  where $n_0 \triangleq n$ is the total number of samples, $\gamma \triangleq \max_{g\in[G]} \frac{n}{n_g}$ is the *sample condition number*, and $\mathcal{C}_g$ is the error cone corresponding to $\boldsymbol{\beta}^*_g$ exactly defined in Section 2. To the best of our knowledge, this is the first statistical estimation guarantee for the data enrichment.
- We also establish the sample complexity of the DE estimator for all parameters as $\forall g \in [G] : n_g = O(\omega(\mathcal{C}_g \cap \mathbb{S}^{p-1}))^2$. We emphasize that our result proofs that the recovery of the common parameter $\boldsymbol{\beta}_0$ by DE estimator benefits from *all* of the n pooled samples.



(a) Structural Coherence (SC) condition.

(b) Data EnRichment Incoherence Condition (DERIC).

Figure 1. a) State of the art condition for recovering common and individual parameters in superposition models where $\mathcal{C}_g = \text{Cone}(\mathcal{E}_g)$ are error cones and $\mathcal{E}_g = \{\boldsymbol{\delta}_g | f_g(\boldsymbol{\beta}_g^* + \boldsymbol{\delta}_g) \leq f_g(\boldsymbol{\beta}_g^*)\}$ are the error sets for each parameter $\boldsymbol{\beta}_g^* \in [G]$ (Gu & Banerjee, 2016) b) Our more relaxed recovery condition which allows *arbitrary non-zero fraction* of the error cones of individual parameters intersect with $-\mathcal{C}_0$.

- We present an efficient projected block gradient descent algorithm DICER, to solve DE's objective (3) which converges geometrically to the statistical error bound of (4). To the best of our knowledge, this is the first rigorous computational result for the high-dimensional data-enriched regression.

2. The Data Enrichment Estimator

A compact form of our proposed DE estimator (3) is:

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}}{\text{argmin}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \forall g \in [G] : f_g(\boldsymbol{\beta}_g) \leq f_g(\boldsymbol{\beta}_g^*), \quad (5)$$

where $\mathbf{y} = (\mathbf{y}_1^T, \dots \mathbf{y}_G^T)^T \in \mathbb{R}^n$, $\boldsymbol{\beta} = ({\boldsymbol{\beta}_0}^T, \dots, {\boldsymbol{\beta}_G}^T)^T \in \mathbb{R}^{(G+1)p}$ and

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_1 & 0 & \cdots & 0 \\ \mathbf{X}_2 & 0 & \mathbf{X}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \mathbf{X}_G & 0 & \cdots & \cdots & \mathbf{X}_G \end{pmatrix} \in \mathbb{R}^{n \times (G+1)p} . \quad (6)$$

Example 1. **(L_1-norm)** When all parameters $\boldsymbol{\beta}_g$s are s_g-sparse, i.e.,$|\text{supp}(\boldsymbol{\beta}_g^*)| = s_g$ by using l_1-norm as the sparsity inducing function, DE (5) instantiates to the *spare DE*:

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}}{\text{argmin}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \forall g \in [G] : \|\boldsymbol{\beta}_g\|_1 \leq \|\boldsymbol{\beta}_g^*\|_1. \quad (7)$$

Consider the group-wise estimation error $\boldsymbol{\delta}_g = \hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g^*$. Since $\hat{\boldsymbol{\beta}}_g = \boldsymbol{\beta}_g^* + \boldsymbol{\delta}_g$ is a feasible point of (5), the error vector $\boldsymbol{\delta}_g$ will belong to the following restricted error set:

$$\mathcal{E}_g = \{\boldsymbol{\delta}_g | f_g(\boldsymbol{\beta}_g^* + \boldsymbol{\delta}_g) \leq f_g(\boldsymbol{\beta}_g^*)\}, \quad g \in [G] . \quad (8)$$

We denote the cone of the error set as $\mathcal{C}_g \triangleq \text{Cone}(\mathcal{E}_g)$ and the spherical cap corresponding to it as $\mathcal{A}_g \triangleq \mathcal{C}_g \cap \mathbb{S}^{p-1}$. Consider the set $\mathcal{C} = \{\boldsymbol{\delta} = (\boldsymbol{\delta}_0^T, \dots, \boldsymbol{\delta}_G^T)^T \big| \boldsymbol{\delta}_g \in \mathcal{C}_g\}$, following two subsets of $\mathcal{C}$ play key roles in our analysis:

$$\mathcal{H} \triangleq \left\{\boldsymbol{\delta} \in \mathcal{C} \big| \sum_{g=0}^{G} \frac{n_g}{n}\|\boldsymbol{\delta}_g\|_2 = 1\right\}, \quad (9)$$

$$\bar{\mathcal{H}} \triangleq \left\{\boldsymbol{\delta} \in \mathcal{C} \big| \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g\|_2 = 1\right\}. \quad (10)$$

Using optimality of $\hat{\boldsymbol{\beta}}$, we can establish the following deterministic error bound.

Theorem 1. *For the proposed estimator (5), assume there exist $0 < \kappa \leq \inf_{\mathbf{u}\in\mathcal{H}} \frac{1}{n}\|\mathbf{X}\mathbf{u}\|_2^2$. Then, for the sample condition number $\gamma = \max_{g\in[G]\setminus} \frac{n}{n_g}$, the following deterministic upper bounds holds:*

$$\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g\|_2 \leq \frac{2\gamma \sup_{\mathbf{u}\in\bar{\mathcal{H}}} \boldsymbol{\omega}^T\mathbf{X}\mathbf{u}}{n\kappa} .$$

3. Restricted Eigenvalue Condition

The main assumptions of Theorem 1 is known as Restricted Eigenvalue (RE) condition in the literature of high dimensional statistics (Banerjee et al., 2014; Negahban et al., 2012; Raskutti et al., 2010): $\inf_{\mathbf{u}\in\mathcal{H}} \frac{1}{n}\|\mathbf{X}\mathbf{u}\|_2^2 \geq \kappa > 0$. Here, we show that for the design matrix $\mathbf{X}$ defined in (6), the RE condition holds with high probability under a suitable geometric condition we call *Data EnRichment Incoherence Condition* (DERIC) and for enough number of samples. For the analysis, similar to existing work (Tropp, 2015; Mendelson, 2014; Gu & Banerjee, 2016), we assume the design matrix to be isotropic sub-Gaussian.[1]

Definition 1. *We assume $\mathbf{x}_{gi}$ are i.i.d. random vectors from a non-degenerate zero-mean, isotropic sub-Gaussian distribution. In other words, $\mathbb{E}[\mathbf{x}] = 0$, $\mathbb{E}[\mathbf{x}^T\mathbf{x}] = \mathbf{I}_{p\times p}$, and $\|\mathbf{x}\|_{\psi_2} \leq k$. As a consequence, $\exists\alpha > 0$ such that $\forall\mathbf{u} \in \mathbb{S}^{p-1}$ we have $\mathbb{E}|\langle\mathbf{x}, \mathbf{u}\rangle| \geq \alpha$. Further, we assume noise $\boldsymbol{\omega}_{gi}$ are i.i.d. zero-mean, unit-variance sub-Gaussian with $\|\|\omega_{gi}\|\|_{\psi_2} \leq K$.*

Definition 2 (Data EnRichment Incoherence Condition (DERIC))**.** *There exists a non-empty set $\mathcal{I} \subseteq [G]\setminus$ of groups where for some scalars $0 < \bar{\rho} \leq 1$ and $\lambda_{\min} > 0$ the following holds:*

1. $\sum_{i\in\mathcal{I}} n_i \geq \lceil\bar{\rho}n\rceil$.
2. $\forall i \in \mathcal{I}$, $\forall\boldsymbol{\delta}_i \in \mathcal{C}_i$, *and* $\boldsymbol{\delta}_0 \in \mathcal{C}_0$: $\|\boldsymbol{\delta}_i + \boldsymbol{\delta}_0\|_2 \geq \lambda_{\min}(\|\boldsymbol{\delta}_0\|_2 + \|\boldsymbol{\delta}_i\|_2)$

Observe that $0 < \lambda_{\min}, \bar{\rho} \leq 1$ by definition.

[1] Extension to an-isotropic sub-Gaussian case is straightforward by techniques developed in (Banerjee et al., 2014; Rudelson & Zhou, 2013).

Using DERIC and the small ball method (Mendelson, 2014), a recent tool from empirical process theory in the following theorem, we elaborate the sample complexity required for satisfying the RE condition:

Theorem 2. *Let $\mathbf{x}_{gi}$s be random vectors defined in Definition 1. Assume DERIC condition of Definition 2 holds for error cones $\mathcal{C}_g$s and $\psi_{\mathcal{I}} = \lambda_{\min}\bar{\rho}/3$. Then, for all $\boldsymbol{\delta} \in \mathcal{H}$, when we have enough number of samples as $\forall g \in [G]_{\setminus}$: $n_g \geq m_g = O(k^6\alpha^{-6}\psi_{\mathcal{I}}^{-2}\omega(\mathcal{A}_g)^2)$, with probability at least $1 - e^{-n\kappa_{\min}/4}$ we have $\inf_{\boldsymbol{\delta}\in\mathcal{H}} \frac{1}{\sqrt{n}}\|\mathbf{X}\boldsymbol{\delta}\|_2 \geq \frac{\kappa_{\min}}{2}$, where $\kappa_{\min} = \min_{g\in[G]_{\setminus}} C\psi_{\mathcal{I}}\frac{\alpha^3}{k^2} - \frac{2c_g k\omega(\mathcal{A}_g)}{\sqrt{n_g}}$ and $\kappa = \frac{\kappa_{\min}^2}{4}$ is the lower bound of the RE condition.*

Example 2. **(L_1-norm)** The Gaussian width of the spherical cap of a p-dimensional s-sparse vector is $\omega(\mathcal{A}) = \Theta(\sqrt{s\log p})$ (Banerjee et al., 2014; Vershynin, 2018). Therefore, the number of samples per group and total required for satisfaction of the RE condition in the sparse DE estimator (7) is $\forall g \in [G] : n_g \geq m_g = \Theta(s_g \log p)$.

4. Estimation Error Bound

Here, we provide a high probability upper bound for the deterministic upper bound of Theorem 1 and derive the final estimation error bound.

Theorem 3. *Assume $\mathbf{x}_{gi}$ and ω_{gi} distributed according to Definition 1 and $\tau > 0$, then with probability at least $1 - \sigma\exp\left(-\min_{g\in[G]}\left[\nu_g n_g - \log(G+1), \frac{\tau^2}{\eta_g^2 k^2}\right]\right)$ we have:*

$$\frac{2}{n}\boldsymbol{\omega}^T\mathbf{X}\boldsymbol{\delta} \leq \sqrt{\frac{8K^2+4}{n}}\max_{g\in[G]}\left(\zeta_g k\omega(\mathcal{A}_g) + \epsilon_g\sqrt{\log(G+1)} + \tau\right)$$

The following corollary characterizes the general error bound and results from the direct combination of Theorem 1, Theorem 2, and Theorem 3.

Corollary 1. *For $\mathbf{x}_{gi}$ and ω_{gi} described in Definition 1 and $\tau > 0$ when we have enough number of samples $\forall g \in [G]$: $n_g > m_g$ which lead to $\kappa > 0$, the following general error bound holds with high probability for estimator (5):*

$$\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g\|_2 \leq \gamma\frac{k\zeta\max_{g\in[G]}\omega(\mathcal{A}_g) + \epsilon\sqrt{\log(G+1)} + \tau}{\kappa_{\min}^2\sqrt{n}}. \quad (11)$$

Example 3. **(L_1-norm)** For the sparse DE estimator of (7), results of Theorem 2 and 3 translates to the following: For enough number of samples as $\forall g \in [G] : n_g \geq m_g = O(s_g \log p)$, the error bound of (11) simplifies to:

$$\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g\|_2 = O\left(\sqrt{\frac{(\max_{g\in[G]} s_g)\log p}{n}}\right) \quad (12)$$

Therefore, individual errors are bounded as $\|\boldsymbol{\delta}_g\|_2 = O(\sqrt{(\max_{g\in[G]} s_g)\log p/n_g})$ which is slightly worse than $O(\sqrt{s_g \log p/n_g})$, the well-known error bound for recovering an s_g-sparse vector from n_g observations using LASSO or similar estimators (Banerjee et al., 2014; Chandrasekaran et al., 2012; Candes et al., 2007; Chatterjee et al., 2014; Bickel et al., 2009). Note that $\max_{g\in[G]} s_g$ (instead of s_g) is the price we pay to recover the common parameter $\boldsymbol{\beta}_0$.

Algorithm 1 DICER

input: $\mathbf{X}, \mathbf{y}$, learning rates $(\mu_0, \ldots, \mu_G)$, initialization $\boldsymbol{\beta}^{(1)} = \mathbf{0}$
output: $\hat{\boldsymbol{\beta}}$
for t = 1 **to** T **do**
  for g=1 **to** G **do**
    $\boldsymbol{\beta}_g^{(t+1)} = \Pi_{\Omega_{f_g}}\left(\boldsymbol{\beta}_g^{(t)} + \mu_g\mathbf{X}_g^T\left(\mathbf{y}_g - \mathbf{X}_g\left(\boldsymbol{\beta}_0^{(t)} + \boldsymbol{\beta}_g^{(t)}\right)\right)\right)$
  end for
  $\boldsymbol{\beta}_0^{(t+1)} = \Pi_{\Omega_{f_0}}\left(\boldsymbol{\beta}_0^{(t)} + \mu_0\mathbf{X}_0^T\left(\mathbf{y} - \mathbf{X}_0\boldsymbol{\beta}_0^{(t)} - \begin{pmatrix}\mathbf{X}_1\boldsymbol{\beta}_1^{(t)}\\ \vdots \\ \mathbf{X}_G\boldsymbol{\beta}_G^{(t)}\end{pmatrix}\right)\right)$
end for

5. Estimation Algorithm

We propose *Data enrIChER* (DICER) a projected block gradient descent algorithm, Algorithm 1, where $\Pi_{\Omega_{f_g}}$ is the Euclidean projection onto the set $\Omega_{f_g}(d_g) = \{f_g(\boldsymbol{\beta}) \leq f_g(\boldsymbol{\beta}_g^*)\}$.To analysis convergence properties of DICER, we should upper bound the error of each iteration.Let's $\boldsymbol{\delta}^{(t)} = \boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*$ be the error of iteration t of DICER, i.e., the distance from the true parameter (not the optimization minimum, $\hat{\boldsymbol{\beta}}$). We show that $\|\boldsymbol{\delta}^{(t)}\|_2$ decreases exponentially fast in t to the statistical error $\|\boldsymbol{\delta}\|_2 = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$.

Theorem 4. *Let $\tau = C\sqrt{\log(G+1)} + b$ for $b > 0$ and $\omega_{0g} = \omega(\mathcal{A}_0) + \omega(\mathcal{A}_g)$. For the step sizes of $\mu_0 = \Theta(\frac{1}{n})$ and $\mu_g = \Theta(\frac{1}{\sqrt{nn_g}})$ and sample complexities of $\forall g \in [G]$: $n_g \geq 2c_g^2(2\omega(\mathcal{A}_g) + \tau)^2$, updates of the Algorithm 1 obey the following with high probability:*

$$\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g^{(t+1)}\|_2 \leq r(\tau)^t\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\beta}_g^*\|_2 + \frac{(G+1)\sqrt{(2K^2+1)}}{\sqrt{n}(1-r(\tau))}\left(\zeta k\max_{g\in[G]}\omega(\mathcal{A}_g) + \tau\right),$$

where $r(\tau) < 1$.

Corollary 2. *For enough number of samples, iterations of DE algorithm with step sizes $\mu_0 = \Theta(\frac{1}{n})$ and $\mu_g = \Theta(\frac{1}{\sqrt{nn_g}})$ geometrically converges to the following with high probability:*

$$\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g^{\infty}\|_2 \leq c\frac{\zeta k\max_{g\in[G]}\omega(\mathcal{A}_g) + C\sqrt{\log(G+1)} + b}{\sqrt{n}(1-r(\tau))} \quad (13)$$

which is a scaled variant of statistical error bound determined in Corollary 1.

References

Asiaee, A., Oymak, S., Coombes, K. R., and Banerjee, A. High dimensional data enrichment: Interpretable, fast, and Data-Efficient. 2018.

Banerjee, A., Chen, S., Fazayeli, F., and Sivakumar, V. Estimation with Norm Regularization. In *Advances in Neural Information Processing Systems*, pp. 1556–1564, 2014.

Bickel, P. J., Ritov, Y., Tsybakov, A. B., et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Candes, E., Tao, T., et al. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

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

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

Candès, E. J., Romberg, J., and Tao, T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

Chatterjee, S., Chen, S., and Banerjee, A. Generalized dantzig selector: Application to the k-support norm. In *Advances in Neural Information Processing Systems*, pp. 1934–1942, 2014.

Chen, A., Owen, A. B., and Shi, M. Data enriched linear regression. *Electronic journal of statistics*, 9(1):1078–1112, 2015.

Chen, J., Liu, J., and Ye, J. Learning incoherent sparse and Low-Rank patterns from multiple tasks. *ACM transactions on knowledge discovery from data*, 5(4):22, 2012.

Dondelinger, F. and Mukherjee, S. High-dimensional regression over disease subgroups. *arXiv preprint arXiv:1611.00953*, 2016.

Donoho, D. L. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Gross, S. M. and Tibshirani, R. Data shared lasso: A novel tool to discover uplift. *Computational Statistics & Data Analysis*, 101:226–235, 2016.

Gu, Q. and Banerjee, A. High dimensional structured superposition models. In *Advances In Neural Information Processing Systems*, pp. 3684–3692, 2016.

Jalali, A., Ravikumar, P., Sanghavi, S., and Ruan, C. A Dirty Model for Multi-task Learning. In *Advances in Neural Information Processing Systems*, pp. 964–972, 2010.

McCoy, M. B. and Tropp, J. A. The achievable performance of convex demixing. *arXiv preprint arXiv:1309.7478*, 2013.

Mendelson, S. Learning Without Concentration. In *Journal of the ACM (JACM)*. To appear, 2014.

Negahban, S. and Wainwright, M. J. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13 (May):1665–1697, 2012.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistical Science*, 27(4):538–557, 2012. ISSN 0883-4237.

Ollier, E. and Viallon, V. Joint estimation of k related regression models with simple l_1-norm penalties. *arXiv preprint arXiv:1411.1594*, 2014.

Ollier, E. and Viallon, V. Regression modeling on stratified data with the lasso. *arXiv preprint arXiv:1508.05476*, 2015.

Oymak, S., Recht, B., and Soltanolkotabi, M. Sharp time–data tradeoffs for linear inverse problems. *arXiv preprint arXiv:1507.04793*, 2015.

Raskutti, G., Wainwright, M. J., and Yu, B. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

Rudelson, M. and Zhou, S. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447, 2013.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Tropp, J. A. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory - a Renaissance*. To appear, may 2015.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, pp. 210–268. Cambridge University Press, Cambridge, 2012.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Zhang, Y. and Yang, Q. A survey on Multi-Task learning. 2017.

A. Proofs of Theorems

In this Section we present detail proof for each theorem and proposition. To avoid cluttering, during our proofs, we state some needed results as lemmas and provide their proof in the next Section B.

A.1. Proof of Theorem 1

Proof. Starting from the optimality inequality, for the lower bound with the set $\mathcal{H}$ we get:

$$\begin{aligned} \frac{1}{n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 &\geq \frac{1}{n}\inf_{\mathbf{u}\in\mathcal{H}}\|\mathbf{X}\mathbf{u}\|_2^2\left(\sum_{g=0}^{G}\frac{n_g}{n}\|\boldsymbol{\delta}_g\|_2\right)^2 \\ &\geq \kappa\left(\sum_{g=0}^{G}\frac{n_g}{n}\|\boldsymbol{\delta}_g\|_2\right)^2 \\ &\geq \kappa\left(\min_{g\in[G]}\frac{n_g}{n}\right)\left(\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g\|_2\right)^2 \end{aligned} \tag{14}$$

where $0 < \kappa \leq \frac{1}{n}\inf_{\mathbf{u}\in\mathcal{H}}\|\mathbf{X}\mathbf{u}\|_2^2$ is known as Restricted Eigenvalue (RE) condition. The upper bound will factorize as:

$$\frac{2}{n}\boldsymbol{\omega}^T\mathbf{X}\boldsymbol{\delta} \leq \frac{2}{n}\sup_{\mathbf{u}\in\bar{\mathcal{H}}}\boldsymbol{\omega}^T\mathbf{X}\mathbf{u}\left(\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g\|_2\right), \quad \mathbf{u}\in\mathcal{H} \tag{15}$$

Putting together inequalities (14) and (15) completes the proof. ■

A.2. Proof of Proposition 1

Proposition 1. *Assume observations distributed as defined in Definition 1 and pair-wise SC conditions are satisfied. Consider each superposition model (2) in isolation; to recover the common parameter $\boldsymbol{\beta}_0^*$ requires at least one group i to have $n_i = O(\omega^2(\mathcal{A}_0))$. To recover the rest of individual parameters, we need $\forall g \neq i : n_g = O(\omega^2(\mathcal{A}_g))$ samples.*

Proof. Consider only one group for regression in isolation. Note that $\mathbf{y}_g = \mathbf{X}_g(\boldsymbol{\beta}_g^* + \boldsymbol{\beta}_0^*) + \boldsymbol{\omega}_g$ is a superposition model and as shown in (Gu & Banerjee, 2016) the sample complexity required for the RE condition and subsequently recovering $\boldsymbol{\beta}_0^*$ and $\boldsymbol{\beta}_g^*$ is $n_g \geq c(\max_{g\in[G]}\omega(\mathcal{A}_g) + \sqrt{\log 2})^2$. ■

A.3. Proof of Theorem 2

Let's simplify the LHS of the RE condition:

$$\begin{aligned} \frac{1}{\sqrt{n}}\|\mathbf{X}\boldsymbol{\delta}\|_2 &= \left(\frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}|\langle\mathbf{x}_{gi},\boldsymbol{\delta}_0+\boldsymbol{\delta}_g\rangle|^2\right)^{\frac{1}{2}} \\ &\geq \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}|\langle\mathbf{x}_{gi},\boldsymbol{\delta}_0+\boldsymbol{\delta}_g\rangle| \\ &\geq \frac{1}{n}\sum_{g=1}^{G}\xi\|\boldsymbol{\delta}_0+\boldsymbol{\delta}_g\|_2\sum_{i=1}^{n_g}\mathbb{1}\left(|\langle\mathbf{x}_{gi},\boldsymbol{\delta}_0+\boldsymbol{\delta}_g\rangle|\geq\xi\|\boldsymbol{\delta}_0+\boldsymbol{\delta}_g\|_2\right), \end{aligned}$$

where the first inequality is due to Lyapunov's inequality. To avoid cluttering we denote $\boldsymbol{\delta}_{0g} = \boldsymbol{\delta}_0 + \boldsymbol{\delta}_g$ where $\boldsymbol{\delta}_0 \in \mathcal{C}_0$ and $\boldsymbol{\delta}_g \in \mathcal{C}_g$. Now we add and subtract the corresponding per-group marginal tail function, $Q_{\xi_g}(\boldsymbol{\delta}_{0g}) = \mathbb{P}(|\langle\mathbf{x},,\boldsymbol{\delta}_{0g}\rangle| > \xi_g)$

where $\xi_g > 0$. Let $\xi_g = \|\boldsymbol{\delta}_{0g}\|_2 \xi$ then the LHS of the RE condition reduces to:

$$\begin{aligned} \inf_{\boldsymbol{\delta}\in\mathcal{H}} \frac{1}{\sqrt{n}}\|\mathbf{X}\boldsymbol{\delta}\|_2 &\geq \inf_{\boldsymbol{\delta}\in\mathcal{H}} \sum_{g=1}^{G} \frac{n_g}{n}\xi_g Q_{2\xi_g}(\boldsymbol{\delta}_{0g}) \\ &- \sup_{\boldsymbol{\delta}\in\mathcal{H}} \frac{1}{n}\sum_{g=1}^{G}\xi_g \sum_{i=1}^{n_g}\left[Q_{2\xi_g}(\boldsymbol{\delta}_{0g}) - \mathbb{1}(|\langle \mathbf{x}_{gi}, \boldsymbol{\delta}_{0g}\rangle| \geq \xi_g)\right] \\ &= t_1(\mathbf{X}) - t_2(\mathbf{X}) \end{aligned} \tag{16}$$

For the ease of exposition we have written the LHS of (16) as the difference of two terms, i.e., $t_1(\mathbf{X}) - t_2(\mathbf{X})$ and in the followings we lower bound the first term t_1 and upper bound the second term t_2.

A.3.1. Lower Bounding the First Term

Our main result is the following lemma which uses the DERIC condition of the Definition 2 and provides a lower bound for the first term $t_1(\mathbf{X})$:

Lemma 1. *Suppose DERIC holds. Let $\psi_{\mathcal{I}} = \frac{\lambda_{\min}\bar{\rho}}{3}$. For any $\boldsymbol{\delta} \in \mathcal{H}$, we have:*

$$\sum_{g=1}^{G} \frac{n_g}{n}\xi_g Q_{2\xi_g}(\boldsymbol{\delta}_{0g}) \geq \psi_{\mathcal{I}}\xi \frac{(\alpha - 2\xi)^2}{4ck^2}\left(\|\boldsymbol{\delta}_0\|_2 + \sum_{g=1}^{n}\frac{n_g}{n}\|\boldsymbol{\delta}_g\|_2\right), \tag{17}$$

which implies that $t_1(\mathbf{X}) = \inf_{\boldsymbol{\delta}\in\mathcal{H}} \sum_{g=1}^{G} \frac{n_G}{n}\xi_g Q_{2\xi_g}(\boldsymbol{\delta}_{0g})$ satisfies the same RHS bound of (17).

A.3.2. Upper Bounding the Second Term

Let's focus on the second term, i.e., $t_2(\mathbf{X})$. First we want to show that the second term satisfies the bounded difference property defined in Section 3.2. of (Boucheron et al., 2013). In other words, by changing each of $\mathbf{x}_{gi}$ the value of $t_2(\mathbf{X})$ at most change by one. First, we rewrite t_2 as follows:

$$h\left(\mathbf{x}_{11}, \ldots, \mathbf{x}_{jk}, \ldots, \mathbf{x}_{Gn_G}\right) = t_2\left(\mathbf{x}_{11}, \ldots, \mathbf{x}_{jk}, \ldots, \mathbf{x}_{Gn_G}\right) = \sup_{\boldsymbol{\delta}\in\mathcal{H}} g\left(\mathbf{x}_{11}, \ldots, \mathbf{x}_{jk}, \ldots, \mathbf{x}_{Gn_G}\right)$$

where $g\left(\mathbf{x}_{11}, \ldots, \mathbf{x}_{jk}, \ldots, \mathbf{x}_{Gn_G}\right) = \sum_{g=1}^{G}\frac{\xi_g}{n}\sum_{i=1}^{n_g}\left[Q_{2\xi_g}(\boldsymbol{\delta}_{0g}) - \mathbb{1}(|\langle \mathbf{x}_{gi}, \boldsymbol{\delta}_{0g}\rangle| \geq \xi_g)\right]$. To avoid cluttering let's $\mathcal{X} = \{\mathbf{x}_{11}, \ldots, \mathbf{x}_{jk}, \ldots, \mathbf{x}_{Gn_G}\}$. We want to show that t_2 has the bounded difference property, meaning:

$$\sup_{\mathcal{X}, \mathbf{x}'_{jk}} \left|h\left(\mathbf{x}_{11}, \ldots, \mathbf{x}_{jk}, \ldots, \mathbf{x}_{Gn_G}\right) - h\left(\mathbf{x}_{11}, \ldots, \mathbf{x}'_{jk}, \ldots, \mathbf{x}_{Gn_G}\right)\right| \leq c_i$$

for some constant c_i. Note that for bounded functions $f, g : \mathcal{X} \to \mathbb{R}$, we have $|\sup_{\mathcal{X}} f - \sup_{\mathcal{X}} g| \leq \sup_{\mathcal{X}} |f - g|$. Therefore:

$$
\begin{aligned}
& \sup_{\mathcal{X}, \mathbf{x}'_{jk}} \left| h\left(\mathbf{x}_{11}, \ldots, \mathbf{x}_{jk}, \ldots, \mathbf{x}_{Gn_G}\right) - h\left(\mathbf{x}_{11}, \ldots, \mathbf{x}'_{jk}, \ldots, \mathbf{x}_{Gn_G}\right) \right| \\
\leq \; & \sup_{\mathcal{X}, \mathbf{x}'_{jk}} \sup_{\boldsymbol{\delta} \in \mathcal{H}} \left| g\left(\mathbf{x}_{11}, \ldots, \mathbf{x}_{jk}, \ldots, \mathbf{x}_{Gn_G}\right) - g\left(\mathbf{x}_{11}, \ldots, \mathbf{x}'_{jk}, \ldots, \mathbf{x}_{Gn_G}\right) \right| \\
\leq \; & \sup_{\mathcal{X}, \mathbf{x}'_{jk}} \sup_{\boldsymbol{\delta} \in \mathcal{H}} \sup_{\mathbf{x}_{jk}, \mathbf{x}'_{jk}} \frac{\xi_j}{n} \left(\mathbb{1}(|\langle \mathbf{x}'_{jk}, \boldsymbol{\delta}_{0j} \rangle| \geq \xi_j) - \mathbb{1}(|\langle \mathbf{x}_{jk}, \boldsymbol{\delta}_{0j} \rangle| \geq \xi_j) \right) \\
\leq \; & \sup_{\mathcal{X}, \mathbf{x}'_{jk}} \sup_{\boldsymbol{\delta} \in \mathcal{H}} \frac{\xi_j}{n} \\
= \; & \frac{\xi}{n} \sup_{\boldsymbol{\delta} \in \mathcal{H}} \|\boldsymbol{\delta}_0 + \boldsymbol{\delta}_g\|_2 \\
= \; & \frac{\xi}{n} \sup_{\boldsymbol{\delta} \in \mathcal{H}} \|\boldsymbol{\delta}_0\|_2 + \|\boldsymbol{\delta}_g\|_2 \\
(\boldsymbol{\delta} \in \mathcal{H}) \; & = \; \xi \left(\frac{1}{n} + \frac{1}{n_g} \right) \\
\leq \; & \frac{2\xi}{n}
\end{aligned}
$$

Note that for $\boldsymbol{\delta} \in \mathcal{H}$ we have $\|\boldsymbol{\delta}_0\|_2 + \frac{n_g}{n}\|\boldsymbol{\delta}_g\|_2 \leq 1$ which results in $\|\boldsymbol{\delta}_0\|_2 \leq 1$ and $\|\boldsymbol{\delta}_g\|_2 \leq \frac{n}{n_g}$. Now, we can invoke the bounded difference inequality from Theorem 6.2 of (Boucheron et al., 2013) which says that with probability at least $1 - e^{-\tau^2/2}$ we have: $t_2(\mathbf{X}) \leq \mathbb{E}t_2(\mathbf{X}) + \frac{\tau}{\sqrt{n}}$.

Having this concentration bound, it is enough to bound the expectation of the second term. Following lemma provides us with the bound on the expectation.

Lemma 2. *For the random vector* $\mathbf{x}$ *of Definition 1, we have the following bound:*

$$
\frac{2}{n} \mathbb{E} \sup_{\boldsymbol{\delta}_{[G]}} \sum_{g=1}^{G} \xi_g \sum_{i=1}^{n_g} \left[Q_{2\xi_g}(\boldsymbol{\delta}_{0g}) - \mathbb{1}(|\langle \mathbf{x}_{gi}, \boldsymbol{\delta}_{0g} \rangle| \geq \xi_g) \right] \leq \frac{2}{\sqrt{n}} \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} c_g k \omega(\mathcal{A}_g) \|\boldsymbol{\delta}_g\|_2
$$

A.3.3. Continuing the Proof of Theorem 2

Set $n_0 = n$. Putting back bounds of $t_1(\mathbf{X})$ and $t_2(\mathbf{X})$ together from Lemma 1 and 2, with probability at least $1 - e^{-\frac{\tau^2}{2}}$ we have:

$$
\begin{aligned}
\inf_{\boldsymbol{\delta} \in \mathcal{H}} \frac{1}{\sqrt{n}} \|\mathbf{X}\boldsymbol{\delta}\|_2 \; & \geq \; \sum_{g=0}^{G} \frac{n_g}{n} \psi_{\mathcal{I}} \xi \|\boldsymbol{\delta}_g\|_2 \frac{(\alpha - 2\xi)^2}{4ck^2} - \frac{2}{\sqrt{n}} \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} c_g k \omega(\mathcal{A}_g) \|\boldsymbol{\delta}_g\|_2 - \frac{\tau}{\sqrt{n}} \\
\left(q = \frac{(\alpha - 2\xi)^2}{4ck^2} \right) \; & = \; \sum_{g=0}^{G} \frac{n_g}{n} \psi_{\mathcal{I}} \xi \|\boldsymbol{\delta}_g\|_2 q - \frac{2c}{\sqrt{n}} \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} k \omega(\mathcal{A}_g) \|\boldsymbol{\delta}_g\|_2 - \frac{\tau}{\sqrt{n}} \\
& = \; n^{-1} \sum_{g=0}^{G} n_g \|\boldsymbol{\delta}_g\|_2 (\psi_{\mathcal{I}} \xi q - 2ck \frac{\omega(\mathcal{A}_g)}{\sqrt{n_g}}) - \frac{\tau}{\sqrt{n}} \\
(\kappa_g = \psi_{\mathcal{I}} \xi q - \frac{2ck\omega(\mathcal{A}_g)}{\sqrt{n_g}}) \; & = \; \sum_{g=0}^{G} \frac{n_g}{n} \|\boldsymbol{\delta}_g\|_2 \kappa_g - \frac{\tau}{\sqrt{n}} \\
& \geq \; \kappa_{\min} \sum_{g=0}^{G} \frac{n_g}{n} \|\boldsymbol{\delta}_g\|_2 - \frac{\tau}{\sqrt{n}} \\
(\boldsymbol{\delta} \in \mathcal{H}) \; & = \; \kappa_{\min} - \frac{\tau}{\sqrt{n}}
\end{aligned}
$$

where $\kappa_{\min} = \operatorname{argmin}_{g\in[G]} \kappa_g$. Note that all κ_gs should be bounded away from zero. To this end we need the follow sample complexities:

$$\forall g \in [G]: \quad \left(\frac{2ck}{\psi_{\mathcal{I}}\xi q}\right)^2 \omega(\mathcal{A}_g)^2 \quad \leq \quad n_g \tag{18}$$

Taking $\xi = \frac{\alpha}{6}$ we can simplify the sample complexities to the followings:

$$\forall g \in [G]: \quad \left(\frac{Ck^3}{\psi_{\mathcal{I}}\alpha^3}\right)^2 \omega(\mathcal{A}_g)^2 \quad \leq \quad n_g \tag{19}$$

Finally, to conclude, we take $\tau = \sqrt{n}\kappa_{\min}/2$. ■

A.4. Proof of Theorem 3

Proof. From now on, to avoid cluttering the notation assume $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. We massage the equation as follows:

$$\boldsymbol{\omega}^T\mathbf{X}\boldsymbol{\delta} \quad = \quad \sum_{g=0}^{G}\langle \mathbf{X}_g^T\boldsymbol{\omega}_g, \boldsymbol{\delta}_g\rangle = \sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g\|_2\langle \mathbf{X}_g^T\frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \frac{\boldsymbol{\delta}_g}{\|\boldsymbol{\delta}_g\|_2}\rangle\sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2$$

Assume $b_g = \langle \mathbf{X}_g^T\frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \frac{\boldsymbol{\delta}_g}{\|\boldsymbol{\delta}_g\|_2}\rangle\sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2$ and $a_g = \sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g\|_2$. Then the above term is the inner product of two vectors $\mathbf{a} = (a_0, \ldots, a_G)$ and $\mathbf{b} = (b_0, \ldots, b_G)$ for which we have:

$$\begin{aligned}
\sup_{\mathbf{a}\in\mathcal{H}} \mathbf{a}^T\mathbf{b} \quad &= \quad \sup_{\|\mathbf{a}\|_1=1} \mathbf{a}^T\mathbf{b} \\
\text{(definition of the dual norm)} \quad &\leq \quad \|\mathbf{b}\|_\infty \\
&= \quad \max_{g\in[G]} b_g
\end{aligned}$$

Now we can go back to the original form:

$$\begin{aligned}
\sup_{\boldsymbol{\delta}\in\mathcal{H}} \boldsymbol{\omega}^T\mathbf{X}\boldsymbol{\delta} \quad &\leq \quad \max_{g\in[G]}\langle \mathbf{X}_g^T\frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \frac{\boldsymbol{\delta}_g}{\|\boldsymbol{\delta}_g\|_2}\rangle\sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2 \\
&\leq \quad \max_{g\in[G]}\sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2 \sup_{\mathbf{u}_g\in\mathcal{C}_g\cap\mathbb{S}^{p-1}}\langle \mathbf{X}_g^T\frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \mathbf{u}_g\rangle
\end{aligned} \tag{20}$$

To avoid cluttering we name $h_g(\boldsymbol{\omega}_g, \mathbf{X}_g) = \|\boldsymbol{\omega}_g\|_2 \sup_{\mathbf{u}_g\in\mathcal{A}_g}\langle \mathbf{X}_g^T\frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \mathbf{u}_g\rangle$ and $e_g(\tau) = \sqrt{(2K^2+1)n_g}\left(\upsilon_g C_g k\omega(\mathcal{A}_g) + \epsilon_g\sqrt{\log G} + \tau\right)$. Then from (20), we have:

$$\mathbb{P}\left(\frac{2}{n}\sup_{\boldsymbol{\delta}\in\mathcal{H}}\boldsymbol{\omega}^T\mathbf{X}\boldsymbol{\delta} > \frac{2}{n}\max_{g\in[G]}\sqrt{\frac{n}{n_g}}e_g(\tau)\right) \quad \leq \quad \mathbb{P}\left(\frac{2}{n}\max_{g\in[G]}\sqrt{\frac{n}{n_g}}h_g(\boldsymbol{\omega}_g, \mathbf{X}_g) > \frac{2}{n}\max_{g\in[G]}\sqrt{\frac{n}{n_g}}e_g(\tau)\right)$$

To simplify the notation, we drop arguments of h_g for now. From the union bound we have:

$$\begin{aligned}
\mathbb{P}\left(\frac{2}{n}\max_{g\in[G]}\sqrt{\frac{n}{n_g}}h_g > \frac{2}{n}\max_{g\in[G]}\sqrt{\frac{n}{n_g}}e_g(\tau)\right) \quad &\leq \quad \sum_{g=0}^{G}\mathbb{P}\left(h_g > \max_{g\in[G]} e_g(\tau)\right) \\
&\leq \quad \sum_{g=0}^{G}\mathbb{P}\left(h_g > e_g(\tau)\right) \\
&\leq \quad (G+1)\max_{g\in[G]}\mathbb{P}\left(h_g > e_g(\tau)\right) \\
&\leq \quad \sigma\exp\left(-\min_{g\in[G]}\left[\nu_g n_g - \log(G+1), \frac{\tau^2}{\eta_g^2 k^2}\right]\right)
\end{aligned}$$

where $\sigma = \max_{g\in[G]}\sigma_g$ and the last inequality is a result of the following lemma:

Lemma 3. *For $\mathbf{x}_{gi}$ and ω_{gi} defined in Definition 1 and $\tau > 0$, with probability at least $1 - \frac{\sigma_g}{(G+1)} \exp\left(-\min\left[\nu_g n_g - \log(G+1), \frac{\tau^2}{\eta_g^2 k^2}\right]\right)$ we have:*

$$\sqrt{\frac{n}{n_g}} \|\boldsymbol{\omega}_g\|_2 \sup_{\mathbf{u}_g \in \mathcal{A}_g} \langle \mathbf{X}_g^T \frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \mathbf{u}_g \rangle \quad \leq \quad \sqrt{(2K^2+1)n} \left(\zeta_g k \omega(\mathcal{A}_g) + \epsilon_g \sqrt{\log(G+1)} + \tau\right),$$

where $\sigma_g, \eta_g, \zeta_g$ and ϵ_g are group dependent constants.

■

A.5. Proof of Theorem 5

Proof. To analysis convergence properties of DICER, we should upper bound the error of each iteration.Let's $\boldsymbol{\delta}^{(t)} = \boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*$ be the error of iteration t of DICER, i.e., the distance from the true parameter (not the optimization minimum, $\hat{\boldsymbol{\beta}}$). We show that $\|\boldsymbol{\delta}^{(t)}\|_2$ decreases exponentially fast in t to the statistical error $\|\boldsymbol{\delta}\|_2 = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$. We first start with the required definitions for our analysis.

Definition 3. *We define the following positive constants as functions of step sizes $\mu_g > 0$:*

$$\begin{aligned}
\forall g \in [G] \quad &: \quad \rho_g(\mu_g) = \sup_{\mathbf{u},\mathbf{v} \in \mathcal{B}_g} \mathbf{v}^T \left(\mathbf{I}_g - \mu_g \mathbf{X}_g^T \mathbf{X}_g\right) \mathbf{u}, \\
&\qquad \eta_g(\mu_g) = \mu_g \sup_{\mathbf{v} \in \mathcal{B}_g} \mathbf{v}^T \mathbf{X}_g^T \frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \\
\forall g \in [G]_\setminus \quad &: \quad \phi_g(\mu_g) = \mu_g \sup_{\mathbf{v} \in \mathcal{B}_g, \mathbf{u} \in \mathcal{B}_0} -\mathbf{v}^T \mathbf{X}_g^T \mathbf{X}_g \mathbf{u},
\end{aligned}$$

where $\mathcal{B}_g = \mathcal{C}_g \cap \mathbb{B}^p$ is the intersection of the error cone and the unit ball.

In the following theorem, we establish a deterministic bound on iteration errors $\|\boldsymbol{\delta}_g^{(t)}\|_2$ which depends on constants defined in Definition 3.

Theorem 5. *For Algorithm 1 initialized by $\boldsymbol{\beta}^{(1)} = \mathbf{0}$, we have the following deterministic bound for the error at iteration $t+1$:*

$$\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g^{(t+1)}\|_2 \quad \leq \quad \rho^t \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\beta}_g^*\|_2 + \frac{1-\rho^t}{1-\rho} \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \eta_g \|\boldsymbol{\omega}_g\|_2, \tag{21}$$

where $\rho \triangleq \max\left(\rho_0 + \sum_{g=1}^{G} \sqrt{\frac{n_g}{n}} \phi_g, \max_{g \in [G]} \left[\rho_g + \sqrt{\frac{n}{n_g}} \frac{\mu_0}{\mu_g} \phi_g\right]\right)$.

The RHS of (21) consists of two terms. If we keep $\rho < 1$, the first term approaches zero fast, and the second term determines the bound. In the following, we show that for specific choices of step sizes μ_gs, the second term can be upper bounded using the analysis of Section 4. More specifically, the first term corresponds to the optimization error which shrinks in every iteration while the second term is constant times the upper bound of the statistical error characterized in Corollary 1. Therefore, if we keep ρ below one, the estimation error of DE algorithm geometrically converges to the approximate statistical error bound.

One way for having $\rho < 1$ is to keep all arguments of $\max(\cdots)$ defining ρ strictly below 1. To this end, we first establish high probability upper bound for ρ_g, η_g, and ϕ_g (in the Appendix A.6) and then show that with enough number of samples and proper step sizes μ_g, ρ can be kept strictly below one with high probability.

In the following lemma we establish a recursive relation between errors of consecutive iterations which leads to a bound for the tth iteration.

Lemma 4. *We have the following recursive dependency between the error of $t+1$th iteration and tth iteration of DE:*

$$\begin{aligned}\|\boldsymbol{\delta}_g^{(t+1)}\|_2 &\leq \left(\rho_g(\mu_g)\|\boldsymbol{\delta}_g^{(t)}\|_2 + \xi_g(\mu_g)\|\boldsymbol{\omega}_g\|_2 + \phi_g(\mu_g)\|\boldsymbol{\delta}_0^{(t)}\|_2\right)\\ \|\boldsymbol{\delta}_0^{(t+1)}\|_2 &\leq \left(\rho_0(\mu_0)\|\boldsymbol{\delta}_0^{(t)}\|_2 + \xi_0(\mu_0)\|\boldsymbol{\omega}_0\|_2 + \mu_0\sum_{g=1}^{G}\frac{\phi_g(\mu_g)}{\mu_g}\|\boldsymbol{\delta}_g^{(t)}\|_2\right)\end{aligned}$$

By recursively applying the result of Lemma 4, we get the following deterministic bound which depends on constants defined in Definition 3:

$$\begin{aligned}b_{t+1} = \sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g^{(t+1)}\|_2 &\leq \left(\rho_0 + \sum_{g=1}^{G}\sqrt{\frac{n_g}{n}}\phi_g\right)\|\boldsymbol{\delta}_0^{(t)}\|_2 + \sum_{g=1}^{G}\left(\sqrt{\frac{n_g}{n}}\rho_g + \mu_0\frac{\phi_g}{\mu_g}\right)\|\boldsymbol{\delta}_g^{(t)}\|_2 + \sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\xi_g\|\boldsymbol{\omega}_g\|_2\\ &\leq \rho\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g^{(t)}\|_2 + \sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\xi_g\|\boldsymbol{\omega}_g\|_2 \qquad (22)\end{aligned}$$

where $\rho = \max\left(\rho_0 + \sum_{g=1}^{G}\sqrt{\frac{n_g}{n}}\phi_g, \max_{g\in[G]}\left[\rho_g + \sqrt{\frac{n}{n_g}}\frac{\mu_0}{\mu_g}\phi_g\right]\right)$. We have:

$$\begin{aligned}b_{t+1} &\leq \rho b_t + \sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\xi_g\|\boldsymbol{\omega}_g\|_2\\ &\leq (\rho)^2 b_{t-1} + (\rho+1)\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\xi_g\|\boldsymbol{\omega}_g\|_2\\ &\leq (\rho)^t b_1 + \left(\sum_{i=0}^{t-1}(\rho)^i\right)\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\xi_g\|\boldsymbol{\omega}_g\|_2\\ &= (\rho)^t\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\beta}_g^1 - \boldsymbol{\beta}_g^*\|_2 + \left(\sum_{i=0}^{t-1}(\rho)^i\right)\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\xi_g\|\boldsymbol{\omega}_g\|_2\\ (\boldsymbol{\beta}^1 = 0) &\leq (\rho)^t\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\beta}_g^*\|_2 + \frac{1-(\rho)^t}{1-\rho}\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\xi_g\|\boldsymbol{\omega}_g\|_2\end{aligned}$$

■

A.6. Proof of Theorem 4

Proof. First we need following two lemmas which are proved separately in the following sections.

Lemma 5. *Consider $a_g \geq 1$, with probability at least $1 - 6\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right)$ the following upper bound holds:*

$$\rho_g\left(\frac{1}{a_g n_g}\right) \leq \frac{1}{2}\left[\left(1 - \frac{1}{a_g}\right) + \sqrt{2}c_g\frac{2\omega(\mathcal{A}_g)+\tau}{a_g\sqrt{n_g}}\right] \qquad (23)$$

Lemma 6. *Consider $a_g \geq 1$, with probability at least $1 - 4\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right)$ the following upper bound holds:*

$$\phi_g\left(\frac{1}{a_g n_g}\right) \leq \frac{1}{a_g}\left(1 + c_{0g}\frac{\omega(\mathcal{A}_g)+\omega(\mathcal{A}_0)+2\tau}{\sqrt{n_g}}\right) \qquad (24)$$

Note that Lemma 3 readily provides a high probability upper bound for $\eta_g(1/(a_g n_g))$ as $\sqrt{(2K^2+1)}\left(\zeta_g k\omega(\mathcal{A}_g) + \epsilon_g\sqrt{\log G} + \tau\right)/(a_g\sqrt{n_g})$.

Starting from the deterministic form of the bound in Theorem 5 and putting in the step sizes as $\mu_g = \frac{1}{n_g a_g}$:

$$\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g^{(t+1)}\|_2 \leq (\rho)^t \sum_{g=0}^{G} \|\boldsymbol{\beta}_g^*\|_2 + \frac{1-(\rho)^t}{1-\rho} \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \eta_g \left(\frac{1}{n_g a_g}\right) \|\boldsymbol{\omega}_g\|_2, \tag{25}$$

where

$$\rho(a_0, \cdots, a_G) = \max\left(\rho_0\left(\frac{1}{na_0}\right) + \sum_{g=1}^{G} \sqrt{\frac{n_g}{n}} \phi_g\left(\frac{1}{n_g a_g}\right), \max_{g\in[G]} \rho_g\left(\frac{1}{n_g a_g}\right) + \sqrt{\frac{n}{n_g}} \frac{\mu_0}{\mu_g} \phi_g\left(\frac{1}{n_g a_g}\right)\right) \tag{26}$$

Remember the following two results to upper bound ρ_gs and ϕ_gs from Lemmas 5 and 6:

$$\begin{aligned}
\rho_g\left(\frac{1}{a_g n_g}\right) &\leq \frac{1}{2}\left[\left(1-\frac{1}{a_g}\right) + \sqrt{2} c_g \frac{2\omega(\mathcal{A}_g)+\tau}{a_g\sqrt{n_g}}\right], \quad \text{w.p.} \quad 1 - 6\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right) \\
\phi_g\left(\frac{1}{a_g n_g}\right) &\leq \frac{1}{a_g}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right), \quad \text{w.p.} \quad 1 - 4\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right)
\end{aligned}$$

First we want to keep $\rho_0 + \sum_{g=1}^{G} \sqrt{\frac{n_g}{n}} \phi_g$ of (26) strictly below 1.

$$\begin{aligned}
\rho_0\left(\frac{1}{a_0 n}\right) + \sum_{g=1}^{G} \sqrt{\frac{n_g}{n}} \phi_g\left(\frac{1}{a_g n_g}\right) &\leq \frac{1}{2}\left[\left(1-\frac{1}{a_0}\right) + \sqrt{2} c_0 \frac{2\omega_0+\tau}{a_0\sqrt{n}}\right] \\
&+ \frac{1}{2}\sum_{g=1}^{G} \frac{2}{a_g}\sqrt{\frac{n_g}{n}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)
\end{aligned}$$

Remember that $a_g \geq 1$ was arbitrary. So we pick it as $a_g = 2\sqrt{\frac{n}{n_g}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)/b_g$ where $b_g \leq 2\sqrt{\frac{n}{n_g}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)$ (because we need $a_g \geq 1$) and the condition becomes:

$$\rho_0\left(\frac{1}{a_0 n}\right) + \sum_{g=1}^{G} \sqrt{\frac{n_g}{n}} \phi_g\left(\frac{1}{a_g n_g}\right) \leq \frac{1}{2}\left[\left(1-\frac{1}{a_0}\right) + \sqrt{2} c_0 \frac{2\omega(\mathcal{A}_0)+\tau}{a_0\sqrt{n}}\right] + \frac{1}{2}\sum_{g=1}^{G} \frac{n_g}{n} b_g \leq 1$$

We want to upper bound the RHS by $1/\theta_f$ which will determine the sample complexity for the shared component:

$$\sqrt{2} c_0 \frac{2\omega(\mathcal{A}_0)+\tau}{\sqrt{n}} \leq a_0\left(1 - \sum_{g=1}^{G} \frac{n_g}{n} b_g\right) + 1 \tag{27}$$

Note that any lower bound on the RHS of (27) will lead to the correct sample complexity for which the coefficient of $\|\boldsymbol{\delta}_0^{(t)}\|_2$ (determined in (26)) will be below one. Since $a_0 \geq 1$ we can ignore the first term by assuming $\max_{g\in[G]_\setminus} b_g \leq 1$ and the condition becomes:

$$\begin{aligned}
&n > 2c_0^2(2\omega(\mathcal{A}_0)+\tau)^2, \forall g \in [G]_\setminus : a_g = 2b_g^{-1}\sqrt{\frac{n}{n_g}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right), \\
&a_0 \geq 1, 0 < b_g \leq 2\sqrt{\frac{n}{n_g}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right), \max_{g\in[G]_\setminus} b_g \leq 1,
\end{aligned}$$

which can be simplified to:

$$\begin{aligned}
&n > 2c_0^2(2\omega(\mathcal{A}_0)+\tau)^2, a_0 \geq 1, \\
&\forall g \in [G]_\setminus : a_g = 2b_g^{-1}\sqrt{\frac{n}{n_g}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right), 0 < b_g \leq 1
\end{aligned} \tag{28}$$

Secondly, we want to bound all of $\rho_g + \mu_0\sqrt{\frac{n}{n_g}}\frac{\phi_g}{\mu_g}$ terms of (26) for $\mu_g = \frac{1}{a_g n_g}$ by 1:

$$\begin{aligned}\rho_g\left(\frac{1}{n_g a_g}\right) + \sqrt{\frac{n}{n_g}}\frac{\mu_0}{\mu_g}\phi_g\left(\frac{1}{n_g a_g}\right) &= \rho_g\left(\frac{1}{n_g a_g}\right) + \sqrt{\frac{n_g}{n}}\frac{a_g}{a_0}\phi_g\left(\frac{1}{n_g a_g}\right) \\ &= \frac{1}{2}\Bigg[\left[\left(1-\frac{1}{a_g}\right) + \sqrt{2}c_g\frac{2\omega_g+\tau}{a_g\sqrt{n_g}}\right] \\ &+ \frac{2}{a_0}\sqrt{\frac{n_g}{n}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)\Bigg] \\ &\leq 1\end{aligned} \tag{29}$$

The condition becomes:

$$\sqrt{2}c_g\frac{2\omega_g+\tau}{\sqrt{n_g}} \leq a_g + 1 - \sqrt{\frac{n_g}{n}}\frac{2a_g}{a_0}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right) \tag{30}$$

Remember that we chose $a_g = 2b_g^{-1}\sqrt{\frac{n}{n_g}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)$. We substitute the value of a_g by keeping in mind the constraints for the b_g and the condition reduces to:

$$\sqrt{2}c_g\frac{2\omega_g+\tau}{d_g} \leq \sqrt{n_g}, \quad d_g := a_g + 1 - \frac{4}{b_g a_0}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)^2 \tag{31}$$

for $d_g > 0$. Note that any positive lower bound of the d_g will satisfy the condition in (31) and the result is a valid sample complexity. In the following we show that $d_g > 1$.We have $a_0 \geq 1$ condition from (28), so we take $a_0 = 4\max_{g\in[G]_\setminus}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)^2$ and look for a lower bound for d_g:

$$d_g \geq a_g + 1 - {b_g}^{-1} \tag{32}$$

$$\begin{aligned}(a_g \text{ from } (28)) &= 2b_g^{-1}\sqrt{\frac{n}{n_g}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right) + 1 - {b_g}^{-1} \\ &= 1 + b_g^{-1}\left[2\sqrt{\frac{n}{n_g}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right) - 1\right]\end{aligned} \tag{33}$$

The term inside of the last bracket (33) is always positive and therefore a lower bound is one, i.e., $d_g \geq 1$. From the condition (31) we get the following sample complexity:

$$n_g > 2c_g^2(2\omega_g+\tau)^2 \tag{34}$$

Now we need to determine b_g from previous conditions (28), knowing that $a_0 = 4\max_{g\in[G]_\setminus}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)^2$. We have $0 < b_g \leq 1$ in (28) and we take the largest step by setting $b_g = 1$.

Here we summarize the setting under which we have the linear convergence:

$$\begin{aligned}&n > 2c_0^2\left(2\omega(\mathcal{A}_0)+\tau\right)^2, \forall g \in [G]_\setminus : n_g \geq 2c_g^2(2\omega(\mathcal{A}_g)+\tau)^2 \\ &a_0 = 4\max_{g\in[G]_\setminus}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)^2, a_g = 2\sqrt{\frac{n}{n_g}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right) \\ &\mu_0 = \frac{1}{4n} \times \frac{1}{\max_{g\in[G]_\setminus}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)^2}, \mu_g = \frac{1}{2\sqrt{nn_g}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)^{-1}\end{aligned} \tag{35}$$

Now we rewrite the same analysis using the tail bounds for the coefficients to clarify the probabilities. To simplify the notation, let $r_{g1} = \frac{1}{2}\left[\left(1-\frac{1}{a_g}\right) + \sqrt{2}c_g\frac{2\omega(\mathcal{A}_g)+\tau}{a_g\sqrt{n_g}}\right]$ and $r_{g2} = \frac{1}{a_g}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)$ and $r_0(\tau) = r_{01} + \sum_{g=1}^{G}\sqrt{\frac{n_g}{n}}r_{g2}$

and $r_g(\tau) = r_{g1} + \sqrt{\frac{n_g}{n}}\frac{a_g}{a_0} r_{g2}, \forall g \in [G]_\setminus$, and $r(\tau) = \max_{g\in[G]} r_g$. All of which are computed using a_gs specified in (35). Basically r is an instantiation of an upper bound of the ρ defined in (26) using a_gs in (35).

We are interested to upper bound the following probability:

$$
\begin{aligned}
& \mathbb{P}\left(\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g^{(t+1)}\|_2 \geq r(\tau)^t \sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\beta}_g^*\|_2 + \frac{(G+1)\sqrt{(2K^2+1)}}{(1-r(\tau))\sqrt{n}}\left(\zeta k \max_{g\in[G]}\omega(\mathcal{A}_g)+\tau\right)\right) \\
\leq\ & \mathbb{P}\Bigg((\rho)^t \sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\beta}_g^*\|_2 + \frac{1-(\rho)^t}{1-\rho}\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\eta_g\left(\frac{1}{n_g a_g}\right)\|\boldsymbol{\omega}_g\|_2 \\
\geq\ & r(\tau)^t \sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\beta}_g^*\|_2 + \frac{(G+1)\sqrt{(2K^2+1)}}{(1-r(\tau))\sqrt{n}}\left(\zeta k \max_{g\in[G]}\omega(\mathcal{A}_g)+\tau\right)\Bigg) \\
\leq\ & \mathbb{P}\left(\rho \geq r(\tau)\right) \\
+\ & \mathbb{P}\left(\frac{1}{1-\rho}\sum_{g=0}^{G}\sqrt{n_g}\eta_g\left(\frac{1}{n_g a_g}\right)\|\boldsymbol{\omega}_g\|_2 \geq \frac{(G+1)\sqrt{(2K^2+1)}}{(1-r(\tau))}\left(\zeta k \max_{g\in[G]}\omega(\mathcal{A}_g)+\tau\right)\right)
\end{aligned} \tag{36}
$$

where the first inequality comes from the deterministic bound of (25), We first focus on bounding the first term $\mathbb{P}\left(\rho \geq r(\tau)\right)$:

$$
\begin{aligned}
& \mathbb{P}\left(\rho \geq r(\tau)\right) \\
=\ & \mathbb{P}\left(\max\left(\rho_0\left(\frac{1}{na_0}\right) + \sum_{g=1}^{G}\sqrt{\frac{n_g}{n}}\phi_g\left(\frac{1}{n_g a_g}\right), \max_{g\in[G]}\rho_g\left(\frac{1}{n_g a_g}\right) + \sqrt{\frac{n}{n_g}}\frac{\mu_0}{\mu_g}\phi_g\left(\frac{1}{n_g a_g}\right)\right) \geq \max_{g\in[G]} r(\tau)\right) \\
\leq\ & \mathbb{P}\left(\rho_0\left(\frac{1}{na_0}\right) + \sum_{g=1}^{G}\sqrt{\frac{n_g}{n}}\phi_g\left(\frac{1}{n_g a_g}\right) \geq r_0\right) + \sum_{g=1}^{G}\mathbb{P}\left(\rho_g\left(\frac{1}{n_g a_g}\right) + \sqrt{\frac{n}{n_g}}\frac{\mu_0}{\mu_g}\phi_g\left(\frac{1}{n_g a_g}\right) \geq r_g\right) \\
\leq\ & \mathbb{P}\left(\rho_0\left(\frac{1}{na_0}\right) \geq r_{01}\right) + \sum_{g=1}^{G}\mathbb{P}\left(\phi_g\left(\frac{1}{n_g a_g}\right) \geq r_{g2}\right) + \sum_{g=1}^{G}\left[\mathbb{P}\left(\rho_g\left(\frac{1}{n_g a_g}\right) \geq r_{g1}\right) + \mathbb{P}\left(\phi_g\left(\frac{1}{n_g a_g}\right) \geq r_{g2}\right)\right] \\
\leq\ & \sum_{g=0}^{G}\mathbb{P}\left(\rho_g\left(\frac{1}{n_g a_g}\right) \geq r_{g1}\right) + 2\sum_{g=1}^{G}\mathbb{P}\left(\phi_g\left(\frac{1}{n_g a_g}\right) \geq r_{g2}\right) \\
\leq\ & \sum_{g=0}^{G} 6\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right) + 2\sum_{g=1}^{G} 4\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right) \\
\leq\ & 6(G+1)\exp\left(-\gamma \min_{g\in[G]}(\omega(\mathcal{A}_g)+\tau)^2\right) + 8G\exp\left(-\gamma \min_{g\in[G]_\setminus}(\omega(\mathcal{A}_g)+\tau)^2\right) \\
\leq\ & 14(G+1)\exp\left(-\gamma \min_{g\in[G]}(\omega(\mathcal{A}_g)+\tau)^2\right)
\end{aligned} \tag{37}
$$

Now we focus on bounding the second term:

$$
\begin{aligned}
& \mathbb{P}\left(\frac{1}{1-\rho}\sum_{g=0}^{G}\sqrt{n_g}\eta_g\left(\frac{1}{n_g a_g}\right)\|\boldsymbol{\omega}_g\|_2 \geq \frac{(G+1)\sqrt{(2K^2+1)}}{(1-r(\tau))}\left(\zeta k \max_{g\in[G]}\omega(\mathcal{A}_g)+\tau\right)\right) \\
\leq\ & \mathbb{P}\left(\frac{1}{1-\rho}\sum_{g=0}^{G}\sqrt{n_g}\eta_g\left(\frac{1}{n_g a_g}\right)\|\boldsymbol{\omega}_g\|_2 \geq \frac{1}{(1-r(\tau))}\sum_{g=0}^{G}\sqrt{(2K^2+1)}\left(\zeta_g k\omega(\mathcal{A}_g)+\tau\right)\right) \\
\leq\ & \mathbb{P}\left(\sum_{g=0}^{G}\sqrt{n_g}\eta_g\left(\frac{1}{n_g a_g}\right)\|\boldsymbol{\omega}_g\|_2 \geq \sum_{g=0}^{G}\sqrt{(2K^2+1)}\left(\zeta_g k\omega(\mathcal{A}_g)+\tau\right)\right) + \mathbb{P}\left(\rho \geq r(\tau)\right) \\
\leq\ & \sum_{g=0}^{G}\mathbb{P}\left(\sqrt{n_g}\eta_g\left(\frac{1}{n_g a_g}\right)\|\boldsymbol{\omega}_g\|_2 \geq \sqrt{(2K^2+1)}\left(\zeta_g k\omega(\mathcal{A}_g)+\tau\right)\right) + \mathbb{P}\left(\rho \geq r(\tau)\right)
\end{aligned} \tag{38}
$$

Focusing on the summand of the first term, remember from Definition 3 that $\eta_g(\mu_g) = \frac{1}{a_g n_g} \sup_{\mathbf{v}\in\mathcal{B}_g} \mathbf{v}^T \mathbf{X}_g^T \frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \quad g \in [G]$ and $a_g \geq 1$:

$$\mathbb{P}\left(\|\boldsymbol{\omega}_g\|_2 \sup_{\mathbf{v}\in\mathcal{B}_g} \mathbf{v}^T \mathbf{X}_g^T \frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2} \geq a_g \sqrt{(2K^2+1)n_g}\,(\zeta_g k \omega(\mathcal{A}_g) + \tau) \right) \leq \sigma_g \exp\left(-\min\left[\nu_g n_g, \frac{\tau^2}{\eta_g^2 k^2} \right] \right) \tag{39}$$

where we used the intermediate form of Lemma 3 for $\tau > 0$. Putting all of the bounds (37), (38), and (39) back into the (36):

$$\begin{aligned} & \sigma_g (G+1) \exp\left(-\min_{g\in[G]} \left(\min\left[\nu_g n_g, \frac{\tau^2}{\eta_g^2 k^2} \right] \right) \right) + 28(G+1) \exp\left(-\gamma \min_{g\in[G]} (\omega(\mathcal{A}_g) + \tau)^2 \right) \\ \leq \quad & \upsilon \exp\left[\min_{g\in[G]} \left(-\min\left[\nu_g n_g - \log G, \gamma(\omega(\mathcal{A}_g) + t)^2, \frac{t^2}{\eta_g^2 k^2} \right] \right) \right] \end{aligned}$$

where $\upsilon = \max(28, \sigma)$ and $\gamma = \min_{g\in[G]} \gamma_g$ and $\tau = t + \max(\epsilon, \gamma^{-1/2})\sqrt{\log(G+1)}$ where $\epsilon = k \max_{g\in[G]} \eta_g$. Note that $\tau = t + C\sqrt{\log(G+1)}$ increases the sample complexities to the followings:

$$n > 2c_0^2 \left(2\omega(\mathcal{A}_0) + C\sqrt{\log(G+1)} + t \right)^2, \forall g \in [G]_\setminus : n_g \geq 2c_g^2 (2\omega(\mathcal{A}_g) + C\sqrt{\log(G+1)} + t)^2$$

and it also affects step sizes as follows:

$$\mu_0 = \frac{1}{4n} \times \min_{g\in[G]_\setminus} \left(1 + c_{0g} \frac{\omega_{0g} + C\sqrt{\log(G+1)} + t}{\sqrt{n_g}} \right)^{-2}, \mu_g = \frac{1}{2\sqrt{n n_g}} \left(1 + c_{0g} \frac{\omega_{0g} + C\sqrt{\log(G+1)} + t}{\sqrt{n_g}} \right)^{-1}$$

■

B. Proofs of Lemmas

Here, we present proofs of each lemma used during the proofs of theorems in Section A.

B.1. Proof of Lemma 1

Proof. LHS of (17) is the weighted summation of $\xi_g Q_{2\xi_g}(\boldsymbol{\delta}_{0g}) = \|\boldsymbol{\delta}_{0g}\|_2 \xi \mathbb{P}(|\langle \mathbf{x}, , \boldsymbol{\delta}_{0g}/\|\boldsymbol{\delta}_{0g}\|_2 \rangle| > 2\xi) = \|\boldsymbol{\delta}_{0g}\|_2 \xi Q_{2\xi}(\mathbf{u})$ where $\xi > 0$ and $\mathbf{u} = \boldsymbol{\delta}_{0g}/\|\boldsymbol{\delta}_{0g}\|_2$ is a unit length vector. So we can rewrite the LHS of (17) as:

$$\sum_{g=1}^{G} \frac{n_g}{n} \xi_g Q_{2\xi_g}(\boldsymbol{\delta}_{0g}) = \sum_{g=1}^{G} \frac{n_g}{n} \|\boldsymbol{\delta}_0 + \boldsymbol{\delta}_g\|_2 \xi Q_{2\xi}(\mathbf{u})$$

With this observation, the lower bound of the Lemma 1 is a direct consequence of the following two results:

Lemma 7. *Let* $\mathbf{u}$ *be any unit length vector and suppose* $\mathbf{x}$ *obeys Definiton 1. Then for any* $\mathbf{u}$*, we have*

$$Q_{2\xi}(\mathbf{u}) \geq \frac{(\alpha - 2\xi)^2}{4ck^2}. \tag{40}$$

Lemma 8. *Suppose Definition 2 holds. Then, we have:*

$$\sum_{i=1}^{G} n_i \|\boldsymbol{\delta}_0 + \boldsymbol{\delta}_i\|_2 \geq \frac{\bar{\rho}\lambda_{\min}}{3} \left(Gn\|\boldsymbol{\delta}_0\|_2 + \sum_{i=1}^{G} n_i \|\boldsymbol{\delta}_i\|_2 \right), \quad \forall i \in [G] : \boldsymbol{\delta}_i \in \mathcal{C}_i. \tag{41}$$

■

B.2. Proof of Lemma 2

Proof. Consider the following soft indicator function which we use in our derivation:

$$\psi_a(s) = \begin{cases} 0, & |s| \leq a \\ (|s| - a)/a, & a \leq |s| \leq 2a \\ 1, & 2a < |s| \end{cases}$$

Now:

$$\begin{aligned}
& \mathbb{E} \sup_{\boldsymbol{\delta}_{[G]}} \sum_{g=1}^{G} \xi_g \sum_{i=1}^{n_g} \left[Q_{2\xi_g}(\boldsymbol{\delta}_{0g}) - \mathbb{1}(|\langle \mathbf{x}_{gi}, \boldsymbol{\delta}_{0g} \rangle| \geq \xi_g) \right] \\
= \; & \mathbb{E} \sup_{\boldsymbol{\delta}_{[G]}} \sum_{g=1}^{G} \xi_g \sum_{i=1}^{n_g} \left[\mathbb{E}\mathbb{1}(|\langle \mathbf{x}_{gi}, \boldsymbol{\delta}_{0g} \rangle| \geq 2\xi_g) - \mathbb{1}(|\langle \mathbf{x}_{gi}, \boldsymbol{\delta}_{0g} \rangle| \geq \xi_g) \right] \\
\leq \; & \mathbb{E} \sup_{\boldsymbol{\delta}_{[G]}} \sum_{g=1}^{G} \xi_g \sum_{i=1}^{n_g} \left[\mathbb{E}\psi_{\xi_g}(\langle \mathbf{x}, \boldsymbol{\delta}_{0g} \rangle) - \psi_{\xi_g}(\langle \mathbf{x}_{gi}, \boldsymbol{\delta}_{0g} \rangle) \right] \\
\leq \; & 2\mathbb{E} \sup_{\boldsymbol{\delta}_{[G]}} \sum_{g=1}^{G} \xi_g \sum_{i=1}^{n_g} \epsilon_{gi} \psi_{\xi_g}(\langle \mathbf{x}_{gi}, \boldsymbol{\delta}_{0g} \rangle) \\
\leq \; & 2\mathbb{E} \sup_{\boldsymbol{\delta}_{[G]}} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \epsilon_{gi} \langle \mathbf{x}_{gi}, \boldsymbol{\delta}_{0g} \rangle
\end{aligned}$$

where ϵ_{gi} are iid copies of Rademacher random variable which are independent of every other random variables and themselves. Now we add back $\frac{1}{n}$ and expand $\boldsymbol{\delta}_{0g} = \boldsymbol{\delta}_0 + \boldsymbol{\delta}_g$:

$$\begin{aligned}
\frac{2}{n}\mathbb{E} \sup_{\boldsymbol{\delta}_{[G]} \in \mathcal{C}_{[G]}} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \epsilon_{gi} \langle \mathbf{x}_{gi}, \boldsymbol{\delta}_{0g} \rangle &= \frac{2}{n}\mathbb{E} \sup_{\boldsymbol{\delta}_0 \in \mathcal{C}_0} \sum_{i=1}^{n} \epsilon_i \langle \mathbf{x}_i, \boldsymbol{\delta}_0 \rangle + \frac{2}{n}\mathbb{E} \sup_{\boldsymbol{\delta}_{[G]\setminus} \in \mathcal{C}_{[G]\setminus}} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \epsilon_{gi} \langle \mathbf{x}_{gi}, \boldsymbol{\delta}_g \rangle \\
&= \frac{2}{\sqrt{n}}\mathbb{E} \sup_{\boldsymbol{\delta}_0 \in \mathcal{C}_0} \sum_{i=1}^{n} \langle \frac{1}{\sqrt{n}} \epsilon_i \mathbf{x}_i, \boldsymbol{\delta}_0 \rangle + \frac{2}{\sqrt{n}}\mathbb{E} \sup_{\boldsymbol{\delta}_{[G]\setminus} \in \mathcal{C}_{[G]\setminus}} \sum_{g=1}^{G} \sqrt{\frac{n_g}{n}} \sum_{i=1}^{n_g} \langle \frac{1}{\sqrt{n_g}} \epsilon_{gi} \mathbf{x}_{gi}, \boldsymbol{\delta}_g \rangle \\
(n_0 := n, \epsilon_{0i} := \epsilon_0, \mathbf{x}_{0i} := \mathbf{x}_i) \quad &= \frac{2}{\sqrt{n}}\mathbb{E} \sup_{\boldsymbol{\delta}_{[G]} \in \mathcal{C}_{[G]}} \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \sum_{i=1}^{n_g} \langle \frac{1}{\sqrt{n_g}} \epsilon_{gi} \mathbf{x}_{gi}, \boldsymbol{\delta}_g \rangle \\
(\mathbf{h}_g := \frac{1}{\sqrt{n_g}} \sum_{i=1}^{n_g} \epsilon_{gi} \mathbf{x}_{gi}) \quad &= \frac{2}{\sqrt{n}}\mathbb{E} \sup_{\boldsymbol{\delta}_{[G]} \in \mathcal{C}_{[G]}} \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \langle \mathbf{h}_g, \boldsymbol{\delta}_g \rangle \\
(\mathcal{A}_g \in \mathcal{C}_g \cap \mathbb{S}^{p-1}) \quad &\leq \frac{2}{\sqrt{n}}\mathbb{E} \sup_{\boldsymbol{\delta}_{[G]} \in \mathcal{A}_{[G]}} \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \langle \mathbf{h}_g, \boldsymbol{\delta}_g \rangle \|\boldsymbol{\delta}_g\|_2 \\
&\leq \frac{2}{\sqrt{n}} \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \mathbb{E}_{\mathbf{h}_g} \sup_{\boldsymbol{\delta}_g \in \mathcal{A}_g} \langle \mathbf{h}_g, \boldsymbol{\delta}_g \rangle \|\boldsymbol{\delta}_g\|_2 \\
&\leq \frac{2}{\sqrt{n}} \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} c_g k \omega(\mathcal{A}_g) \|\boldsymbol{\delta}_g\|_2
\end{aligned}$$

Note that the $\mathbf{h}_{gi}$ is a sub-Gaussian random vector which let us bound the $\mathbb{E}\sup$ using the Gaussian width (Tropp, 2015) in the last step. ■

B.3. Proof of Lemma 3

Proof. To avoid cluttering let $h_g(\boldsymbol{\omega}_g, \mathbf{X}_g) = \sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2 \sup_{\mathbf{u}_g \in \mathcal{A}_g} \langle \mathbf{X}_g^T \frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \mathbf{u}_g \rangle$, $e_g = \zeta_g k \omega(\mathcal{A}_g) + \epsilon_g \sqrt{\log G} + \tau$, where $s_g = \sqrt{\frac{n}{n_g}}\sqrt{(2K^2+1)n_g}$.

$$
\begin{aligned}
\mathbb{P}\left(h_g(\boldsymbol{\omega}_g, \mathbf{X}_g) > e_g s_g\right) &= \mathbb{P}\left(h_g(\boldsymbol{\omega}_g, \mathbf{X}_g) > e_g s_g \middle| \sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2 > s_g\right)\mathbb{P}\left(\sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2 > s_g\right) \\
&+ \mathbb{P}\left(h_g(\boldsymbol{\omega}_g, \mathbf{X}_g) > e_g s_g \middle| \sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2 < s_g\right)\mathbb{P}\left(\sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2 < s_g\right) \\
&\leq \mathbb{P}\left(\sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2 > s_g\right) + \mathbb{P}\left(h_g(\boldsymbol{\omega}_g, \mathbf{X}_g) > e_g s_g \middle| \sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2 < s_g\right) \\
&\leq \mathbb{P}\left(\|\boldsymbol{\omega}_g\|_2 > \sqrt{(2K^2+1)n_g}\right) + \mathbb{P}\left(\sup_{\mathbf{u}_g \in \mathcal{C}_g \cap \mathbb{S}^{p-1}} \langle \mathbf{X}_g^T \frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \mathbf{u}_g \rangle > e_g\right) \\
&\leq \mathbb{P}\left(\|\boldsymbol{\omega}_g\|_2 > \sqrt{(2K^2+1)n_g}\right) + \sup_{\mathbf{v} \in \mathbb{S}^{p-1}} \mathbb{P}\left(\sup_{\mathbf{u}_g \in \mathcal{C}_g \cap \mathbb{S}^{p-1}} \langle \mathbf{X}_g^T \mathbf{v}, \mathbf{u}_g \rangle > e_g\right)
\end{aligned} \tag{42}
$$

Let's focus on the first term. Since $\boldsymbol{\omega}_g$ consists of i.i.d. centered unit-variance sub-Gaussian elements with $|||\omega_{gi}|||_{\psi_2} < K$, ω_{gi}^2 is sub-exponential with $|||\omega_{gi}|||_{\psi_1} < 2K^2$. Let's apply the Bernstein's inequality to $\|\boldsymbol{\omega}_g\|_2^2 = \sum_{i=1}^{n_g} \omega_{gi}^2$:

$$
\mathbb{P}\left(\left|\|\boldsymbol{\omega}_g\|_2^2 - \mathbb{E}\|\boldsymbol{\omega}_g\|_2^2\right| > \tau\right) \leq 2\exp\left(-\nu_g \min\left[\frac{\tau^2}{4K^4 n_g}, \frac{\tau}{2K^2}\right]\right)
$$

We also know that $\mathbb{E}\|\omega_g\|_2^2 \leq n_g$ (Banerjee et al., 2014) which gives us:

$$
\mathbb{P}\left(\|\boldsymbol{\omega}_g\|_2 > \sqrt{n_g + \tau}\right) \leq 2\exp\left(-\nu_g \min\left[\frac{\tau^2}{4K^4 n_g}, \frac{\tau}{2K^2}\right]\right)
$$

Finally, we set $\tau = 2K^2 n_g$:

$$
\mathbb{P}\left(\|\boldsymbol{\omega}_g\|_2 > \sqrt{(2K^2+1)n_g}\right) \leq 2\exp\left(-\nu_g n_g\right) = \frac{2}{(G+1)}\exp\left(-\nu_g n_g + \log(G+1)\right)
$$

Now we upper bound the second term of (42). Given any fixed $\mathbf{v} \in \mathbb{S}^{p-1}$, $\mathbf{X}_g \mathbf{v}$ is a sub-Gaussian random vector with $|||\mathbf{X}_g^T \mathbf{v}|||_{\psi_2} \leq C_g k$ (Banerjee et al., 2014). From Theorem 9 of (Banerjee et al., 2014) for any $\mathbf{v} \in \mathbb{S}^{p-1}$ we have:

$$
\mathbb{P}\left(\sup_{\mathbf{u}_g \in \mathcal{A}_g} \langle \mathbf{X}_g^T \mathbf{v}, \mathbf{u}_g \rangle > \upsilon_g C_g k \omega(\mathcal{A}_g) + t\right) \leq \pi_g \exp\left(-\left(\frac{t}{\theta_g C_g k \phi_g}\right)^2\right)
$$

where $\phi_g = \sup_{\mathbf{u}_g \in \mathcal{A}_g} \|\mathbf{u}_g\|_2$ and in our problem $\phi_g = 1$. We now substitute $t = \tau + \epsilon_g \sqrt{\log(G+1)}$ where $\epsilon_g = \theta_g C_g k$.

$$
\begin{aligned}
\mathbb{P}\left(\sup_{\mathbf{u}_g \in \mathcal{A}_g} \langle \mathbf{X}_g^T \mathbf{v}, \mathbf{u}_g \rangle > \upsilon_g C_g k \omega(\mathcal{A}_g) + \epsilon_g \sqrt{\log(G+1)} + \tau\right) &\leq \pi_g \exp\left(-\left(\frac{\tau + \epsilon_g \sqrt{\log(G+1)}}{\epsilon_g}\right)^2\right) \\
&\leq \pi_g \exp\left(-\log G - \left(\frac{\tau}{\theta_g C_g k}\right)^2\right) \\
&\leq \frac{\pi_g}{(G+1)} \exp\left(-\left(\frac{\tau}{\theta_g C_g k}\right)^2\right)
\end{aligned}
$$

Now we put back results to the original inequality (42):

$$
\begin{aligned}
& \mathbb{P}\left(h_g(\boldsymbol{\omega}_g, \mathbf{X}_g) > \sqrt{\frac{n}{n_g}}\sqrt{(2K^2+1)n_g} \times \left(\upsilon_g C_g k \omega(\mathcal{A}_g) + \epsilon_g \sqrt{\log(G+1)} + \tau\right)\right) \\
\leq \quad & \frac{\sigma_g}{(G+1)} \exp\left(-\min\left[\nu_g n_g - \log(G+1), \frac{\tau^2}{\theta_g^2 C_g^2 k^2}\right]\right) \\
\leq \quad & \frac{\sigma_g}{(G+1)} \exp\left(-\min\left[\nu_g n_g - \log(G+1), \frac{\tau^2}{\eta_g^2 k^2}\right]\right)
\end{aligned}
$$

where $\sigma_g = \pi_g + 2$, $\zeta_g = \upsilon_g C_g$, $\eta_g = \theta_g C_g$. ■

B.4. Proof of Lemma 4

Proof. We upper bound the individual error $\|\boldsymbol{\delta}_g^{(t+1)}\|_2$ and the common one $\|\boldsymbol{\delta}_0^{(t+1)}\|_2$ in the followings:

$$
\begin{aligned}
\|\boldsymbol{\delta}_g^{(t+1)}\|_2 &= \|\boldsymbol{\beta}_g^{(t+1)} - \boldsymbol{\beta}_g^*\|_2 \\
&= \left\|\Pi_{\Omega_{f_g}}\left(\boldsymbol{\beta}_g^{(t)} + \mu_g \mathbf{X}_g^T\left(\mathbf{y}_g - \mathbf{X}_g(\boldsymbol{\beta}_0^{(t)} + \boldsymbol{\beta}_g^{(t)})\right)\right) - \boldsymbol{\beta}_g^*\right\|_2 \\
\text{(Lemma 6.3 of (Oymak et al., 2015))} &= \left\|\Pi_{\Omega_{f_g} - \{\boldsymbol{\beta}_g^*\}}\left(\boldsymbol{\beta}_g^{(t)} + \mu_g \mathbf{X}_g^T\left(\mathbf{y}_g - \mathbf{X}_g(\boldsymbol{\beta}_0^{(t)} + \boldsymbol{\beta}_g^{(t)})\right) - \boldsymbol{\beta}_g^*\right)\right\|_2 \\
&= \left\|\Pi_{\mathcal{E}_g}\left(\boldsymbol{\delta}_g^{(t)} + \mu_g \mathbf{X}_g^T\left(\mathbf{y}_g - \mathbf{X}_g(\boldsymbol{\beta}_0^{(t)} + \boldsymbol{\beta}_g^{(t)}) - \mathbf{X}_g(\boldsymbol{\beta}_0^* + \boldsymbol{\beta}_g^*) + \mathbf{X}_g(\boldsymbol{\beta}_0^* + \boldsymbol{\beta}_g^*)\right)\right)\right\|_2 \\
&= \left\|\Pi_{\mathcal{E}_g}\left(\boldsymbol{\delta}_g^{(t)} + \mu_g \mathbf{X}_g^T\left(\boldsymbol{\omega}_g - \mathbf{X}_g(\boldsymbol{\delta}_0^{(t)} + \boldsymbol{\delta}_g^{(t)})\right)\right)\right\|_2 \\
\text{(Lemma 6.4 of (Oymak et al., 2015))} &\leq \left\|\Pi_{\mathcal{C}_g}\left(\boldsymbol{\delta}_g^{(t)} + \mu_g \mathbf{X}_g^T\left(\boldsymbol{\omega}_g - \mathbf{X}_g(\boldsymbol{\delta}_0^{(t)} + \boldsymbol{\delta}_g^{(t)})\right)\right)\right\|_2 \\
\text{(Lemma 6.2 of (Oymak et al., 2015))} &\leq \sup_{\mathbf{v} \in \mathcal{C}_g \cap \mathbb{B}^p} \mathbf{v}^T\left(\boldsymbol{\delta}_g^{(t)} + \mu_g \mathbf{X}_g^T\left(\boldsymbol{\omega}_g - \mathbf{X}_g(\boldsymbol{\delta}_0^{(t)} + \boldsymbol{\delta}_g^{(t)})\right)\right) \\
(\mathcal{B}_g = \mathcal{C}_g \cap \mathbb{B}^p) &= \sup_{\mathbf{v} \in \mathcal{B}_g} \mathbf{v}^T\left(\boldsymbol{\delta}_g^{(t)} + \mu_g \mathbf{X}_g^T\left(\boldsymbol{\omega}_g - \mathbf{X}_g(\boldsymbol{\delta}_0^{(t)} + \boldsymbol{\delta}_g^{(t)})\right)\right) \\
&\leq \sup_{\mathbf{v} \in \mathcal{B}_g} \mathbf{v}^T(\mathbf{I}_g - \mu_g \mathbf{X}_g^T \mathbf{X}_g)\boldsymbol{\delta}_g^{(t)} + \mu_g \sup_{\mathbf{v} \in \mathcal{B}_g} \mathbf{v}^T \mathbf{X}_g^T \boldsymbol{\omega}_g + \mu_g \sup_{\mathbf{v} \in \mathcal{B}_g} -\mathbf{v}^T \mathbf{X}_g^T \mathbf{X}_g \boldsymbol{\delta}_0^{(t)} \\
&\leq \left\|\boldsymbol{\delta}_g^{(t)}\right\|_2 \sup_{\mathbf{u}, \mathbf{v} \in \mathcal{B}_g} \mathbf{v}^T(\mathbf{I}_g - \mu_g \mathbf{X}_g^T \mathbf{X}_g)\mathbf{u} + \mu_g \|\boldsymbol{\omega}_g\|_2 \sup_{\mathbf{v} \in \mathcal{B}_g} \mathbf{v}^T \mathbf{X}_g^T \frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2} \\
&+ \mu_g \|\boldsymbol{\delta}_0^{(t)}\|_2 \sup_{\mathbf{v} \in \mathcal{B}_g, \mathbf{u} \in \mathcal{B}_0} -\mathbf{v}^T \mathbf{X}_g^T \mathbf{X}_g \mathbf{u} \\
&= \rho_g(\mu_g)\|\boldsymbol{\delta}_g^{(t)}\|_2 + \xi_g(\mu_g)\|\boldsymbol{\omega}_g\|_2 + \phi_g(\mu_g)\|\boldsymbol{\delta}_0^{(t)}\|_2
\end{aligned}
$$

So the final bound becomes:

$$
\|\boldsymbol{\delta}_g^{(t+1)}\|_2 \leq \rho_g(\mu_g)\|\boldsymbol{\delta}_g^{(t)}\|_2 + \xi_g(\mu_g)\|\boldsymbol{\omega}_g\|_2 + \phi_g(\mu_g)\|\boldsymbol{\delta}_0^{(t)}\|_2 \tag{43}
$$

Now we upper bound the error of common parameter. Remember common parameter's update: $\boldsymbol{\beta}_0^{(t+1)} = \Pi_{\Omega_{f_0}}\left(\boldsymbol{\beta}_0^{(t)} + \mu_0 \mathbf{X}_0^T \begin{pmatrix} (\mathbf{y}_1 - \mathbf{X}_1(\boldsymbol{\beta}_0^{(t)} + \boldsymbol{\beta}_1^{(t)})) \\ \vdots \\ (\mathbf{y}_G - \mathbf{X}_G(\boldsymbol{\beta}_0^{(t)} + \boldsymbol{\beta}_G^{(t)})) \end{pmatrix}\right)$.

$$
\begin{aligned}
\|\boldsymbol{\delta}_0^{(t+1)}\|_2 &= \|\boldsymbol{\beta}_0^{(t+1)} - \boldsymbol{\beta}_0^*\|_2 \\
&= \left\|\Pi_{\Omega_{f_0}}\left(\boldsymbol{\beta}_0^{(t)} + \mu_0 \sum_{g=1}^{G} \mathbf{X}_g^T\left(\mathbf{y}_g - \mathbf{X}_g(\boldsymbol{\beta}_0^{(t)} + \boldsymbol{\beta}_g^{(t)})\right)\right) - \boldsymbol{\beta}_0^*\right\|_2 \\
\text{(Lemma 6.3 of (Oymak et al., 2015))} &= \left\|\Pi_{\Omega_{f_0}-\{\boldsymbol{\beta}_0^*\}}\left(\boldsymbol{\beta}_0^{(t)} + \mu_0 \sum_{g=1}^{G} \mathbf{X}_g^T\left(\mathbf{y}_g - \mathbf{X}_g(\boldsymbol{\beta}_0^{(t)} + \boldsymbol{\beta}_g^{(t)})\right) - \boldsymbol{\beta}_0^*\right)\right\|_2 \\
&= \left\|\Pi_{\mathcal{E}_0}\left(\boldsymbol{\delta}_0^{(t)} + \mu_0 \sum_{g=1}^{G} \mathbf{X}_g^T\left(\mathbf{y}_g - \mathbf{X}_g(\boldsymbol{\beta}_0^{(t)} + \boldsymbol{\beta}_g^{(t)})\right)\right)\right\|_2 \\
\text{(Lemma 6.4 of (Oymak et al., 2015))} &\leq \left\|\Pi_{\mathcal{C}_0}\left(\boldsymbol{\delta}_0^{(t)} + \mu_0 \sum_{g=1}^{G} \mathbf{X}_g^T\left(\boldsymbol{\omega}_g - \mathbf{X}_g(\boldsymbol{\delta}_0^{(t)} + \boldsymbol{\delta}_g^{(t)})\right)\right)\right\|_2 \\
\text{(Lemma 6.2 of (Oymak et al., 2015))} &\leq \sup_{\mathbf{v}\in\mathcal{B}_0} \mathbf{v}^T\left(\boldsymbol{\delta}_0^{(t)} + \mu_0 \sum_{g=1}^{G} \mathbf{X}_g^T\left(\boldsymbol{\omega}_g - \mathbf{X}_g(\boldsymbol{\delta}_0^{(t)} + \boldsymbol{\delta}_g^{(t)})\right)\right) \\
&\leq \sup_{\mathbf{v}\in\mathcal{B}_0} \mathbf{v}^T\big(\mathbf{I} - \mu_0 \sum_{g=1}^{G} \mathbf{X}_g^T\mathbf{X}_g\big)\boldsymbol{\delta}_0^{(t)} + \mu_0 \sup_{\mathbf{v}\in\mathcal{B}_0} \mathbf{v}^T \sum_{g=1}^{G} \mathbf{X}_g^T\boldsymbol{\omega}_g \\
&+ \mu_0 \sup_{\mathbf{v}\in\mathcal{B}_0} -\mathbf{v}^T \sum_{g=1}^{G} \mathbf{X}_g^T\mathbf{X}_g\boldsymbol{\delta}_g^{(t)} \\
&\leq \|\boldsymbol{\delta}_0^{(t)}\|_2 \sup_{\mathbf{u},\mathbf{v}\in\mathcal{B}_0} \mathbf{v}^T\big(\mathbf{I} - \mu_0\mathbf{X}_0^T\mathbf{X}_0\big)\mathbf{u} + \mu_0 \sup_{\mathbf{v}\in\mathcal{B}_0} \mathbf{v}^T\mathbf{X}_0^T \frac{\boldsymbol{\omega}_0}{\|\boldsymbol{\omega}_0\|_2}\|\boldsymbol{\omega}_0\|_2 \\
&+ \mu_0 \sum_{g=1}^{G} \sup_{\mathbf{v}_g\in\mathcal{B}_0,\mathbf{u}_g\in\mathcal{B}_g} -\mathbf{v}_g^T\mathbf{X}_g^T\mathbf{X}_g\mathbf{u}_g\|\boldsymbol{\delta}_g^{(t)}\|_2 \\
&\leq \rho_0(\mu_0)\|\boldsymbol{\delta}_0^{(t)}\|_2 + \xi_0(\mu_0)\|\boldsymbol{\omega}_0\|_2 + \mu_0 \sum_{g=1}^{G} \frac{\phi_g(\mu_g)}{\mu_g}\|\boldsymbol{\delta}_g^{(t)}\|_2 \qquad (44)
\end{aligned}
$$

To avoid cluttering we drop μ_g as the arguments. Putting together (43) and (44) inequalities we reach to the followings:

$$
\begin{aligned}
\|\boldsymbol{\delta}_g^{(t+1)}\|_2 &\leq \rho_g\|\boldsymbol{\delta}_g^{(t)}\|_2 + \xi_g\|\boldsymbol{\omega}_g\|_2 + \phi_g\|\boldsymbol{\delta}_0^{(t)}\|_2 \\
\|\boldsymbol{\delta}_0^{(t+1)}\|_2 &\leq \rho_0\|\boldsymbol{\delta}_0^{(t)}\|_2 + \xi_0\|\boldsymbol{\omega}_0\|_2 + \mu_0 \sum_{g=1}^{G} \frac{\phi_g}{\mu_g}\|\boldsymbol{\delta}_g^{(t)}\|_2
\end{aligned}
$$

■

B.5. Proof of Lemma 5

We will need the following lemma in our proof. It establishes the RE condition for individual isotropic sub-Gaussian designs and provides us with the essential tool for proving high probability bounds.

Lemma 9 (Theorem 11 of (Banerjee et al., 2014)). *For all $g \in [G]$, for the matrix $\mathbf{X}_g \in \mathbb{R}^{n_g\times p}$ with independent isotropic sub-Gaussian rows, i.e., $\|\|\mathbf{x}_{gi}\|\|_{\psi_2} \leq k$ and $\mathbb{E}[\mathbf{x}_{gi}\mathbf{x}_{gi}^T] = \mathbf{I}$, the following result holds with probability at least*

$1-2\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right)$ *for* $\tau>0$*:*

$$\forall \mathbf{u}_g \in \mathcal{C}_g : n_g\left(1-c_g\frac{\omega(\mathcal{A}_g)+\tau}{\sqrt{n_g}}\right)\|\mathbf{u}_g\|_2^2 \le \|\mathbf{X}_g\mathbf{u}_g\|_2^2 \le n_g\left(1+c_g\frac{\omega(\mathcal{A}_g)+\tau}{\sqrt{n_g}}\right)\|\mathbf{u}_g\|_2^2$$

where $c_g>0$ *is constant.*

The statement of Lemma 9 characterizes the distortion in the Euclidean distance between points $\mathbf{u}_g \in \mathcal{C}_g$ when the matrix $\mathbf{X}_g/n_g$ is applied to them and states that any sub-Gaussian design matrix is approximately isometry, with high probability:

$$(1-\alpha)\|\mathbf{u}_g\|_2^2 \le \frac{1}{n_g}\|\mathbf{X}_g\mathbf{u}_g\|_2^2 \le (1+\alpha)\|\mathbf{u}_g\|_2^2$$

where $\alpha = c_g\frac{\omega(\mathcal{A}_g)}{\sqrt{n_g}}$.

Now the proof for Lemma 5:

Proof. First we upper bound each of the coefficients $\forall g \in [G]$:

$$\rho_g(\mu_g) \quad = \quad \sup_{\mathbf{u},\mathbf{v}\in\mathcal{B}_g} \mathbf{v}^T\big(\mathbf{I}_g-\mu_g\mathbf{X}_g^T\mathbf{X}_g\big)\mathbf{u}$$

We upper bound the argument of the sup as follows:

$$\begin{aligned}
\mathbf{v}^T\big(\mathbf{I}_g-\mu_g\mathbf{X}_g^T\mathbf{X}_g\big)\mathbf{u} &= \frac{1}{4}\left[(\mathbf{u}+\mathbf{v})^T(\mathbf{I}-\mu_g\mathbf{X}_g^T\mathbf{X}_g)(\mathbf{u}+\mathbf{v})-(\mathbf{u}-\mathbf{v})^T(\mathbf{I}-\mu_g\mathbf{X}_g^T\mathbf{X}_g)(\mathbf{u}-\mathbf{v})\right]\\
&= \frac{1}{4}\left[\|\mathbf{u}+\mathbf{v}\|_2^2-\mu_g\|\mathbf{X}_g(\mathbf{u}+\mathbf{v})\|_2^2-\|\mathbf{u}-\mathbf{v}\|_2^2+\mu_g\|\mathbf{X}_g(\mathbf{u}-\mathbf{v})\|_2^2\right]\\
\text{(Lemma 9)} &\le \frac{1}{4}\Bigg[\left(1-\mu_g n_g\left(1-c_g\frac{2\omega(\mathcal{A}_g)+\tau}{\sqrt{n_g}}\right)\right)\|\mathbf{u}+\mathbf{v}\|_2\\
&- \left(1-\mu_g n_g\left(1+c_g\frac{2\omega(\mathcal{A}_g)+\tau}{\sqrt{n_g}}\right)\right)\|\mathbf{u}-\mathbf{v}\|_2\Bigg]\\
\left(\mu_g=\frac{1}{a_g n_g}\right) &\le \frac{1}{4}\left[\left(1-\frac{1}{a_g}\right)(\|\mathbf{u}+\mathbf{v}\|_2-\|\mathbf{u}-\mathbf{v}\|_2)+c_g\frac{2\omega(\mathcal{A}_g)+\tau}{a_g\sqrt{n_g}}(\|\mathbf{u}+\mathbf{v}\|_2+\|\mathbf{u}-\mathbf{v}\|_2)\right]\\
&\le \frac{1}{4}\left[\left(1-\frac{1}{a_g}\right)2\|\mathbf{v}\|_2+c_g\frac{2\omega(\mathcal{A}_g)+\tau}{a_g\sqrt{n_g}}2\sqrt{2}\right]
\end{aligned}$$

where the last line follows from the triangle inequality and the fact that $\|\mathbf{u}+\mathbf{v}\|_2+\|\mathbf{u}-\mathbf{v}\|_2 \le 2\sqrt{2}$ which itself follows from $\|\mathbf{u}+\mathbf{v}\|_2^2+\|\mathbf{u}-\mathbf{v}\|_2^2 \le 4$. Note that we applied the Lemma 9 for bigger sets of $\mathcal{A}_g+\mathcal{A}_g$ and $\mathcal{A}_g-\mathcal{A}_g$ where Gaussian width of both of them are upper bounded by $2\omega(\mathcal{A}_g)$. The above holds with high probability (computed below). Now we set :

$$\mathbf{v}^T\big(\mathbf{I}_g-\frac{1}{a_g n_g}\mathbf{X}_g^T\mathbf{X}_g\big)\mathbf{u} \quad \le \quad \frac{1}{2}\left[\left(1-\frac{1}{a_g}\right)+\sqrt{2}c_g\frac{2\omega(\mathcal{A}_g)+\tau}{a_g\sqrt{n_g}}\right] \tag{45}$$

To keep the upper bound of ρ_g in (45) below any arbitrary $\frac{1}{b}<1$ we need $n_g = O(b^2(\omega(\mathcal{A}_g)+\tau)^2)$ samples.

Now we rewrite the same analysis using the tail bounds for the coefficients to clarify the probabilities. Let's set $\mu_g = \frac{1}{a_g n_g}$, $d_g := \frac{1}{2}\left(1-\frac{1}{a_g}\right)+\sqrt{2}c_g\frac{\omega(\mathcal{A}_g)+\tau/2}{a_g\sqrt{n_g}}$ and name the bad events of $\|\mathbf{X}_g(\mathbf{u}+\mathbf{v})\|_2^2 < n_g\left(1-c_g\frac{2\omega(\mathcal{A}_g)+\tau}{\sqrt{n_g}}\right)$ and $\|\mathbf{X}_g(\mathbf{u}-\mathbf{v})\|_2^2 > n_g\left(1+c_g\frac{2\omega(\mathcal{A}_g)+\tau}{\sqrt{n_g}}\right)$ as $\mathcal{E}_1$ and $\mathcal{E}_2$ respectively:

$$\begin{aligned}
\mathbb{P}(\rho_g \ge d_g) &\le \mathbb{P}(\rho_g \ge d_g|\neg\mathcal{E}_1,\neg\mathcal{E}_2)+2\mathbb{P}(\mathcal{E}_1)+\mathbb{P}(\mathcal{E}_2)\\
\text{Lemma 9} &\le 0+6\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right)
\end{aligned}$$

which concludes the proof. ■

B.6. Proof of Lemma 6

Proof. The following holds for any $\mathbf{u}$ and $\mathbf{v}$ because of $\|\mathbf{X}_g(\mathbf{u}+\mathbf{v})\|_2^2 \geq 0$:

$$-\mathbf{v}^T\mathbf{X}_g^T\mathbf{X}_g\mathbf{u} \leq \frac{1}{2}\left(\|\mathbf{X}_g\mathbf{u}\|_2^2 + \|\mathbf{X}_g\mathbf{v}\|_2^2\right) \tag{46}$$

Now we can bound ϕ_g as follows:

$$\phi_g(\mu_g) = \mu_g \sup_{\mathbf{v}\in\mathcal{B}_g,\mathbf{u}\in\mathcal{B}_0} -\mathbf{v}^T\mathbf{X}_g^T\mathbf{X}_g\mathbf{u} \quad \leq \quad \frac{\mu_g}{2}\left(\sup_{\mathbf{u}\in\mathcal{B}_0}\|\mathbf{X}_g\mathbf{u}\|_2^2 + \sup_{\mathbf{v}\in\mathcal{B}_g}\|\mathbf{X}_g\mathbf{v}\|_2^2\right) \tag{47}$$

So we have:

$$\begin{aligned}
\phi_g\left(\frac{1}{a_g n_g}\right) &\leq \frac{1}{2a_g}\left(\frac{1}{n_g}\sup_{\mathbf{u}\in\mathcal{B}_0}\|\mathbf{X}_g\mathbf{u}\|_2^2 + \frac{1}{n_g}\sup_{\mathbf{v}\in\mathcal{B}_g}\|\mathbf{X}_g\mathbf{v}\|_2^2\right) \qquad (48)\\
\text{(Lemma 9)} &\leq \frac{1}{a_g}\left(1 + c_{0g}\frac{\omega(\mathcal{A}_g)+\omega(\mathcal{A}_0)+2\tau}{2\sqrt{n_g}}\right)\\
(\omega_{0g} = \max(\omega(\mathcal{A}_0),\omega(\mathcal{A}_g)) &\leq \frac{1}{a_g}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)
\end{aligned}$$

where $c_{0g} = \max(c_0, c_g)$.

To compute the exact probabilities lets define $s_g := \frac{1}{a_g}\left(1 + c_{0g}\frac{\omega(\mathcal{A}_g)+\omega(\mathcal{A}_0)+2\tau}{2\sqrt{n_g}}\right)$ and name the bad events of $\frac{1}{n_g}\sup_{\mathbf{u}\in\mathcal{B}_0}\|\mathbf{X}_g\mathbf{u}\|_2^2 > 1 + c_0\frac{\omega(\mathcal{A}_0)+\tau}{\sqrt{n_g}}$ and $\frac{1}{n_g}\sup_{\mathbf{v}\in\mathcal{B}_g}\|\mathbf{X}_g\mathbf{v}\|_2^2 > 1 + c_g\frac{\omega(\mathcal{A}_g)+\tau}{\sqrt{n_g}}$ as $\mathcal{E}_1$ and $\mathcal{E}_2$ respectively.

$$\begin{aligned}
\mathbb{P}(\phi_g > s_g) &\leq \mathbb{P}(\phi_g > s_g|\neg\mathcal{E}_1)\mathbb{P}(\neg\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_1) \qquad (49)\\
&\leq \mathbb{P}(\mathcal{E}_2) + \mathbb{P}(\mathcal{E}_1)\\
&\leq 4\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right)
\end{aligned}$$

■

B.7. Proof of Lemma 7

Proof. To obtain lower bound, we use the Paley–Zygmund inequality for the zero-mean, non-degenerate ($0 < \alpha \leq \mathbb{E}|\langle\mathbf{x},\mathbf{u}\rangle|, \mathbf{u}\in\mathbb{S}^{p-1}$) sub-Gaussian random vector $\mathbf{x}$ with $\|\mathbf{x}\|_{\psi_2} \leq k$ (Tropp, 2015).

$$Q_{2\xi}(\mathbf{u}) \geq \frac{(\alpha-2\xi)^2}{4ck^2}.$$

■

B.8. Proof of Lemma 8

Proof. We split $[G]_\setminus - \mathcal{I}$ into two groups $\mathcal{J}, \mathcal{K}$. $\mathcal{J}$ consists of $\boldsymbol{\delta}_i$'s with $\|\boldsymbol{\delta}_i\|_2 \geq 2\|\boldsymbol{\delta}_0\|_2$ and $\mathcal{K} = [G]_\setminus - \mathcal{I} - \mathcal{J}$. We use the bounds

$$\|\boldsymbol{\delta}_0 + \boldsymbol{\delta}_i\|_2 \geq \begin{cases} \lambda_{\min}(\|\boldsymbol{\delta}_i\|_2 + \|\boldsymbol{\delta}_0\|_2) & \text{if } i\in\mathcal{I}\\ \|\boldsymbol{\delta}_i\|_2/2 & \text{if } i\in\mathcal{J}\\ 0 & \text{if } i\in\mathcal{K}\end{cases} \tag{50}$$

This implies

$$\sum_{i=1}^{G} n_i\|\boldsymbol{\delta}_0 + \boldsymbol{\delta}_i\|_2 \geq \sum_{i\in\mathcal{J}}\frac{n_i}{2}\|\boldsymbol{\delta}_i\|_2 + \lambda_{\min}\sum_{i\in\mathcal{I}} n_i(\|\boldsymbol{\delta}_i\|_2 + \|\boldsymbol{\delta}_0\|_2).$$

Let $S_{\mathcal{S}} = \sum_{i\in\mathcal{S}} n_i \|\boldsymbol{\delta}_i\|_2$ for $\mathcal{S} = \mathcal{I}, \mathcal{J}, \mathcal{K}$. We know that over $\mathcal{K}$, $\|\boldsymbol{\delta}_i\|_2 \leq 2\|\boldsymbol{\delta}_0\|_2$ which implies $S_{\mathcal{K}} = \sum_{i\in\mathcal{K}} n_i \|\boldsymbol{\delta}_i\|_2 \leq 2\sum_{i\in\mathcal{K}} n_i \|\boldsymbol{\delta}_0\|_2 \leq 2n\|\boldsymbol{\delta}_0\|_2$. Set $\psi_{\mathcal{I}} = \min\{1/2, \lambda_{\min}\bar{\rho}/3\} = \lambda_{\min}\bar{\rho}/3$. Using $1/2 \geq \psi_{\mathcal{I}}$, we write:

$$
\begin{aligned}
\sum_{i=1}^{G} n_i \|\boldsymbol{\delta}_0 + \boldsymbol{\delta}_i\|_2 &\geq \psi_{\mathcal{I}} S_{\mathcal{J}} + \lambda_{\min} \sum_{i\in\mathcal{I}} n_i (\|\boldsymbol{\delta}_i\|_2 + \|\boldsymbol{\delta}_0\|_2) \\
(S_{\mathcal{K}} \leq 2n\|\boldsymbol{\delta}_0\|_2) &\geq \psi_{\mathcal{I}} S_{\mathcal{J}} + \psi_{\mathcal{I}} S_{\mathcal{K}} - 2\psi_{\mathcal{I}} n \|\boldsymbol{\delta}_0\|_2 + \left(\sum_{i\in\mathcal{I}} n_i\right) \lambda_{\min} \|\boldsymbol{\delta}_0\|_2 + \lambda_{\min} S_{\mathcal{I}} \\
(\lambda_{\min} \geq \psi_{\mathcal{I}}) &\geq \psi_{\mathcal{I}}(S_{\mathcal{I}} + S_{\mathcal{J}} + S_{\mathcal{K}}) + \left(\left(\sum_{i\in\mathcal{I}} n_i\right) \lambda_{\min} - 2\psi_{\mathcal{I}} n\right) \|\boldsymbol{\delta}_0\|_2.
\end{aligned}
$$

Now, observe that, assumption of the Definition 2, $\sum_{i\in\mathcal{I}} n_i \geq \bar{\rho} n$ implies:

$$
\left(\sum_{i\in\mathcal{I}} n_i\right) \lambda_{\min} - 2\psi_{\mathcal{I}} n \geq (\bar{\rho}\lambda_{\min} - 2\psi_{\mathcal{I}}) n \geq \psi_{\mathcal{I}} n.
$$

Combining all, we obtain:

$$
\sum_{i=1}^{G} n_i \|\boldsymbol{\delta}_0 + \boldsymbol{\delta}_i\|_2 \geq \psi_{\mathcal{I}}(S_{\mathcal{I}} + S_{\mathcal{J}} + S_{\mathcal{K}} + \|\boldsymbol{\delta}_0\|_2) = \psi_{\mathcal{I}}(n\|\boldsymbol{\delta}_0\|_2 + \sum_{i=1}^{G} n_i \|\boldsymbol{\delta}_i\|_2).
$$

■