
High Dimensional Data Enrichment: Interpretable, Fast, and Data-Efficient

Amir Asiaee T.
Ohio State University
asiaeeta@osu.edu

Samet Oymak
UC Riverside
oymak@ece.ucr.edu

Kevin R. Coombes
Ohio State University
coombes.3@osu.edu

Arindam Banerjee
University of Minnesota
banerjee@cs.umn.edu

Abstract

High dimensional structured data enriched model describes groups of observations by shared and per-group individual parameters, each with 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 groups G . Any convex function, e.g., norms, can characterize the structure of both shared and individual parameters. We propose an estimator for high dimensional data enriched model and provide conditions under which it consistently estimates both shared and individual parameters. We also delineate sample complexity of the estimator and present high probability non-asymptotic bound on estimation error of all parameters. Interestingly the sample complexity of our estimator translates to conditions on both per-group sample sizes and the total number of samples. We propose an iterative estimation algorithm with linear convergence rate and supplement our theoretical analysis with synthetic and real experimental results. Particularly, we show the predictive power of data enriched model along with its interpretable results in anticancer drug sensitivity analysis.

1 Introduction

Consider the problem of modeling involving more than one cohort/group in the population which are similar in many ways but have certain unique aspects. Scoping the exposition to linear models, one can assume that for each group, the data comes from a different linear model with distinct parameter β_g^* , i.e., $y_{gi} = \mathbf{x}_{gi}^T \beta_g^* + \omega_{gi}$, where g and i index the group and samples of each group respectively. Such an approach fails to acknowledge the fact the groups are similar in many ways, and will need suitably large sample size for each group for effective modeling. Alternatively, one can ignore the group information and build a global model using the simple linear model $y_i = \mathbf{x}_i^T \beta^* + \omega_i$. While such a model will have the advantage of using a large dataset to build the model, the model will be inaccurate for the groups since it is not modeling their unique aspects.

In this work, we consider the *data enrichment* model recently suggested in the literature [6, 7, 13, 14] for such settings. In particular, a data enriched model assumes that there is a *common* parameter β_0^* shared between all groups which captures the similarity between groups and *individual* per-group parameters β_g that captures the unique aspects of the groups:

$$y_{gi} = \mathbf{x}_{gi}^T (\beta_0^* + \beta_g^*) + \omega_{gi}, \quad g \in \{1, \dots, G\}. \quad (1)$$

In (1), we have G linear regression models that share the parameter β_0^* , which captures the shared characteristics of the different groups. We specifically focus on structured high dimensional data

enriched linear models (1) when the number of samples for each group is much smaller than the ambient dimensionality, i.e., $\forall g : n_g \ll p$ and the parameters β_g are structured, i.e., for suitable convex functions f_g s, $f_g(\beta_g)$ are small. For example, when the structure is sparsity the corresponding function is l_1 -norm.

Note that each of the linear models of (1) is a superposition [8] or dirty statistical model [20]. Therefore, data enriched models are effectively coupled superposition models. A related model is proposed by [9] in the context of multi-task learning, where for each task g the output is coming from $y_{gi} = \mathbf{x}_{gi}^T(\beta_{0g}^* + \beta_g^*) + \omega_{gi}$. As emphasized by the subscript of β_{0g}^* the common parameters are different in every task but they share a same support (index of non-zero values), i.e., $\text{supp}(\beta_{0i}^*) = \text{supp}(\beta_{0j}^*)$.

data enriched model where β_g s are sparse, has recently gained attention because of its application in wide range of domains such as personalized medicine [6], sentiment analysis, banking strategy [7], single cell data analysis [14], road safety [13], and disease subtype analysis [6]. More generally, in any high dimensional domain where the population consists of groups, data enrichment framework has the potential to boost both parameter estimation and prediction.

In spite of the recent surge in applying data enrichment framework to different domains, limited advances have been made in understanding 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 [7, 13]. To the best of our knowledge, the only theoretical guarantee for data enrichment is provided in [14] where authors prove sparsistency of their proposed method under the stringent irrepresentability condition of the design matrix. Also beyond sparsity and l_1 -norm, no other structure has been investigated for these models. Further, no computational results, such as computational rates of convergence of iterative algorithms for the estimators, exist in the literature.

Notation and Preliminaries: We denote sets with 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 $[G] \setminus = [G] - \{0\}$. Given G group and n_g samples in each one 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:

$$\mathbf{y}_g = \mathbf{X}_g(\beta_0^* + \beta_g^*) + \omega_g, \quad \forall g \in [G] \setminus \quad (2)$$

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

A random variable V is sub-Gaussian if the 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 sub-Gaussian norm of V , denoted by $\|V\|_{\psi_2}$ [19]. 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 [19] 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 $\omega(\mathcal{V}) = \mathbb{E}_{\mathbf{g}} [\sup_{\mathbf{u} \in \mathcal{V}} \langle \mathbf{g}, \mathbf{u} \rangle]$ [4], 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 estimator $\hat{\beta}$ for recovering the structured common and individual parameters where the structure is induced by a *convex* functions $f_g(\cdot)$.

$$\hat{\beta} = (\hat{\beta}_0^T, \dots, \hat{\beta}_G^T) \in \underset{\beta_0, \dots, \beta_G}{\text{argmin}} \frac{1}{n} \sum_{g=1}^G \|\mathbf{y}_g - \mathbf{X}_g(\beta_0 + \beta_g)\|_2^2, \text{ s.t. } \forall g \in [G] : f_g(\beta_g) \leq f_g(\beta_g^*).$$

We present several new statistical and computational results for the data enriched model:

- The data enrichment estimator (3) succeeds if a geometric condition that we call *Data Enrichment Incoherence Condition* (DEIC) is satisfied. Compared to other known geometric conditions in the literature such as structural coherence [8] and stable recovery conditions [10], DEIC is a considerably weaker condition.
- Assuming DEIC holds, we establish a high probability non-asymptotic bound on the weighted sum of component-wise estimation error, $\delta_g = \hat{\beta}_g - \beta_g^*$ as:

$$\sum_{g=0}^G \sqrt{\frac{n_g}{n}} \|\delta_g\|_2 \leq C\gamma \frac{\max_{g \in [G]} \omega(\mathcal{C}_g \cap \mathbb{S}^{p-1}) + \sqrt{\log(G+1)}}{\sqrt{n}}, \quad (3)$$

where n_g is number of samples per group, $n = n_0$ is the total number of samples, $\gamma = \max_{g \in [G]} \frac{n}{n_g}$ is the *balance condition number*, and \mathcal{C}_g is the error cone corresponding to β_g^* exactly defined in Section 2.1. To the best of our knowledge, this is the first statistical estimation guarantee for data enriched models.

- We also establish that the required sample complexity for the estimation of parameters for all groups as $n_g = O(\omega(\mathcal{C}_g \cap \mathbb{S}^{p-1}))^2$ and for the common parameter as $n = O(\omega(\mathcal{C}_0 \cap \mathbb{S}^{p-1}))^2$. In other words, enough *total* number of samples n is good enough to recover the common parameter β_0 , illustrating the fact that the common parameter estimation benefits from the shared data.
- We present an efficient Projected Block Gradient Descent (PBGD) algorithm for the estimation problem which converges geometrically to the statistical error bound of (3). To the best of our knowledge, this is the first rigorous computational result for data enrichment models.
- Finally, we apply the data enrichment estimator of (3) to find biological predictor of drug sensitivity of cancer cell lines. Individual components detected by PBGD provides us an interpretable model which detects important drug sensitivity predictors in each cancer.

The rest of this paper is organized as follows: First, we characterize the error set of our estimator and provide a deterministic error bound in Section 2. Then in Section 3, we discuss the restricted eigenvalue condition and calculate the per-group and total sample complexity required for the recovery of the true parameters by our estimator under DEIC condition. We close the statistical analysis in Section 4 by providing high probability error bounds. We delineate our linearly convergent algorithm, PBGD in Section 5 and finally supplement our work with synthetic and real data experiments in Sections 6 and 7.

2 The Data Enrichment Estimator

We write a compact equivalent of our proposed estimator estimator (3) as:

$$\hat{\beta} \in \arg \min_{\beta} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2, \quad \forall g \in [G] : f_g(\beta_g) \leq f_g(\beta_g^*), \quad (4)$$

where $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_G^T)^T \in \mathbb{R}^n$, $\beta = (\beta_0^T, \dots, \beta_G^T)^T \in \mathbb{R}^{(G+1)p}$ and

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_1 & 0 & \cdots & 0 \\ \mathbf{X}_2 & 0 & \mathbf{X}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \mathbf{X}_G & 0 & \cdots & \cdots & \mathbf{X}_G \end{pmatrix} \in \mathbb{R}^{n \times (G+1)p}. \quad (5)$$

2.1 Error Set and Deterministic Error Bound

With componentwise estimation error $\delta_g = \hat{\beta}_g - \beta_g^*$, since $\hat{\beta}_g = \beta_g^* + \delta_g$ is a feasible point of (4), the error vector δ_g will belong to the following restricted error set:

$$\mathcal{E}_g = \{ \delta_g \mid f_g(\beta_g^* + \delta_g) \leq f_g(\beta_g^*) \}, \quad g \in [G].$$

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 $\mathcal{C} = \{ \delta = (\delta_0^T, \dots, \delta_G^T)^T \mid \delta_g \in \mathcal{C}_g \}$, following two subsets of \mathcal{C} play key roles in our analysis:

$$\mathcal{H} \triangleq \left\{ \delta \in \mathcal{C} \mid \sum_{g=0}^G \frac{n_g}{n} \|\delta_g\|_2 = 1 \right\}, \quad \bar{\mathcal{H}} \triangleq \left\{ \delta \in \mathcal{C} \mid \sum_{g=0}^G \sqrt{\frac{n_g}{n}} \|\delta_g\|_2 = 1 \right\}.$$

Starting from the optimality of $\hat{\beta} = \beta^* + \delta$ as $\frac{1}{n} \|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2 \leq \frac{1}{n} \|\mathbf{y} - \mathbf{X}\beta^*\|_2^2$, we have: $\frac{1}{n} \|\mathbf{X}\delta\|_2^2 \leq \frac{1}{n} 2\omega^T \mathbf{X}\delta$ where $\omega = [\omega_1^T, \dots, \omega_G^T]^T \in \mathbb{R}^n$ is the vector of all noises. Using this basic inequality, we can establish the following deterministic error bound.

Theorem 2.1. *For the proposed estimator (4), assume there exist $0 < \kappa \leq \inf_{\mathbf{u} \in \mathcal{H}} \frac{1}{n} \|\mathbf{X}\mathbf{u}\|_2^2$. Then, for $\gamma = \max_{g \in [G]} \frac{n}{n_g}$, the following deterministic upper bounds holds:*

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

3 Restricted Eigenvalue Condition

The main assumption of Theorem 2.1 is known as Restricted Eigenvalue (RE) condition in the literature of high dimensional statistics [1, 12, 16]: $\inf_{\mathbf{u} \in \mathcal{H}} \frac{1}{n} \|\mathbf{X}\mathbf{u}\|_2^2 \geq \kappa > 0$. The RE condition posits that the minimum eigenvalues of the matrix $\mathbf{X}^T \mathbf{X}$ in directions restricted to \mathcal{H} is strictly positive. In this section, we show that for the design matrix \mathbf{X} defined in (5), the RE condition holds with high probability under a suitable geometric condition we call *Data Enrichment Incoherence Condition* (DEIC) along with a precise characterization of group specific sample complexity. For the analysis, similar to existing work [18, 11, 8], we assume the design matrix to be sub-Gaussian.¹

Definition 3.1. *We assume rows \mathbf{x}_{g_i} 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 ω_{g_i} are i.i.d. zero-mean, unit-variance sub-Gaussian with $\|\omega_{g_i}\|_{\psi_2} \leq K$.*

Unlike standard high-dimensional statistical estimation, for RE condition to be true, data enriched model (2) needs to satisfy a geometric condition under which trivial solutions such as $\delta_g = -\delta_0$ for all $g \in [G] \setminus \setminus$ are avoided. To get at this condition, first note that each of the linear models in (2) is a superposition model [8] or dirty statistical model [20]. RE condition of individual superposition models can be established under the so-called Structural Coherence (SC) condition [8, 10]. However, SC conditions on each individual problem fails to utilize the true coupling structure in the data enriched model, where β_0 is involved in all models. In fact, as we show shortly, using SC on each individual models leads to radically pessimistic estimates of the sample complexity.

In this work, we introduce DEIC, a considerably weaker geometric condition compared to SC of [8, 10]. In particular, SC requires that none of the individual error cones \mathcal{C}_g intersect with the inverted error cone $-\mathcal{C}_0$. Instead of this stringent geometric condition, we allow $-\mathcal{C}_0$ to intersect with an arbitrarily large fraction of the \mathcal{C}_g cones. As the number of intersections increases, our bound becomes looser. The rigorous definition of DEIC is provided below.

Definition 3.2 (Data Enrichment Incoherence Condition (DEIC)). *There exists a set $\mathcal{I} \subseteq [G] \setminus \setminus$ of groups where for some scalars $0 \leq \bar{\rho} \leq 1$ and $\lambda_{\min} > 0$ the following holds:*

1. $\sum_{i \in \mathcal{I}} n_i \geq \lceil \bar{\rho} n \rceil$.
2. $\forall i \in \mathcal{I}, \forall \delta_i \in \mathcal{C}_i$, and $\delta_0 \in \mathcal{C}_0$: $\|\delta_i + \delta_0\|_2 \geq \lambda_{\min} (\|\delta_0\|_2 + \|\delta_i\|_2)$

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

In contrast, the existing SC condition [8, 10] applied to each problem requires for $\delta_0 \in \mathcal{C}_0$ and each $\delta_g \in \mathcal{C}_g$ there exist $\lambda > 0$ such that: $\|\delta_0 + \delta_g\|_2 \geq \lambda (\|\delta_0\|_2 + \|\delta_g\|_2)$. Clearly DEIC and SC conditions are satisfied if the error cones \mathcal{C}_g and \mathcal{C}_0 does not have a ray in common, i.e., $\sup \langle \delta_0 / \|\delta_0\|_2, \delta_g / \|\delta_g\|_2 \rangle < 1$ [18, 8]. However, DEIC condition also allows for a large fraction of cones to intersect with \mathcal{C}_0 . Now, we are ready to show that the state-of-the-art estimator of [8] will lead to a considerably pessimistic sample complexity.

Proposition 3.3. *Assume observations distributed as defined in Definition 3.1 and pair-wise SC conditions are satisfied. Consider each superposition model (2) in isolation; to recover the common parameter β_0^* requires at least one group to have $n_g = O(\omega^2(\mathcal{A}_0))$. Recovering the individual parameter β_g^* needs at least $n_g = O((\max_{g \in [G]} \omega(\mathcal{A}_g) + \sqrt{\log 2})^2)$ samples in the group.*

In other words, by separate analysis of superposition estimators neither the estimation of the common parameter β_0 nor the individual parameters β_g benefit from pooling the n samples. But given the nature of coupling in the data enriched model, we hope to be able to get a better sample complexity specifically for the common parameter β_0 . Using DEIC and the small ball method [11], a recent tool from empirical process theory, we get a better sample complexity for satisfying the RE condition.

Theorem 3.4. *Let \mathbf{x}_{g_i} s be random vectors defined in Definition 3.1. Assume DEIC condition of Definition 3.2 holds for \mathcal{C}_g s and $\psi_{\mathcal{I}} = \lambda_{\min} \bar{\rho} / 3$. Then, for all $\delta \in \mathcal{H}$, when we have enough number of samples as $\forall g \in [G] \setminus \setminus$: $n_g \geq m_g = O(k^6 \alpha^{-6} \psi_{\mathcal{I}}^{-2} \omega(\mathcal{A}_g)^2)$, with probability at least*

¹Extension to an-isotropic sub-Gaussian case is straightforward by techniques developed in [1, 17].

$1 - e^{-n\kappa_{\min}/4}$ we have:

$$\inf_{\delta \in \mathcal{H}} \frac{1}{\sqrt{n}} \|\mathbf{X}\delta\|_2 \geq \frac{\kappa_{\min}}{2}$$

where $\kappa_{\min} = \min_{g \in [G]} C\psi_{\mathcal{I}} \frac{\alpha^3}{k^2} - \frac{2c_g k \omega(\mathcal{A}_g)}{\sqrt{n_g}}$ and $\kappa = \frac{\kappa_{\min}^2}{4}$.

4 General Error Bound

In this section, we provide a high probability upper bound for the estimation error of the common and individual components under general convex function $f(\cdot)$. From now on, to avoid cluttering the notation assume $\boldsymbol{\omega} = \boldsymbol{\omega}_0$: We massage the upper bound of Theorem 2.1 as follows:

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

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

$$\sup_{\mathbf{a} \in \mathcal{H}} \mathbf{a}^T \mathbf{b} = \sup_{\|\mathbf{a}\|_1=1} \mathbf{a}^T \mathbf{b} \leq \|\mathbf{b}\|_{\infty} = \max_{g \in [G]} b_g,$$

where the inequality holds because of the definition of the dual norm. Following lemma upper bounds b_g with high probability.

Lemma 4.1. For \mathbf{x}_{gi} and ω_{gi} defined in Definition 3.1, 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:

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

where $\sigma_g, \eta_g, \zeta_g$ and ϵ_g are group dependent constants and $\tau > 0$.

Using Lemma 4.1 below theorem establishes a high probability upper bound for the deterministic bound of Theorem 2.1, i.e., $\frac{2}{n} \boldsymbol{\omega}^T \mathbf{X}\mathbf{u}$.

Theorem 4.2. Assume \mathbf{x}_{gi} to be a sub-Gaussian random variable with $\mathbb{E}[\mathbf{x}_{gi}^T \mathbf{x}_{gi}] = \mathbf{I}_{p \times p}$ and $\|\mathbf{x}_{gi}\|_{\psi_2} \leq k$ and $\boldsymbol{\omega}$ consists of i.i.d. centered unit-variance sub-Gaussian elements with $\|\omega_{gi}\|_{\psi_2} \leq K$, 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:

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

where $\sigma = \max_{g \in [G]} \sigma_g$ and $\tau > 0$.

The following corollary characterizes the general error bound and results from the direct combination of Theorem 2.1, Theorem 3.4, and Theorem 4.2.

Corollary 4.3. For \mathbf{x}_{gi} and ω_{gi} described in Theorem 2.1 and Theorem 4.2 when we have $\forall g \in [G] : n_g > m_g$ which lead to $\kappa > 0$, the following general error bound holds with high probability for estimator (4):

$$\sum_{g=0}^G \sqrt{\frac{n_g}{n}} \|\delta_g\|_2 \leq C\gamma \frac{k\zeta \max_{g \in [G]} \omega(\mathcal{A}_g) + \epsilon \sqrt{\log(G+1)} + \tau}{\kappa_{\min}^2 \sqrt{n}} \quad (6)$$

where $C = 8\sqrt{2K^2 + 1}$, $\zeta = \max_{g \in [G]} \zeta_g$, $\epsilon = \max_{g \in [G]} \epsilon_g$, $\gamma = \max_{g \in [G]} n/n_g$ and $\tau > 0$.

5 Estimation Algorithm

We propose the following Projected Block Gradient Descent algorithm (PBGD), Algorithm 1 where $\Pi_{\Omega_{f_g}}$ is the Euclidean projection onto the set $\Omega_{f_g}(d_g) = \{f_g(\boldsymbol{\beta}) \leq d_g\}$ where $d_g = f_g(\boldsymbol{\beta}_g^*)$ and is dropped to avoid clutter. In practice, d_g can be determined by cross-validation.

Algorithm 1 PBGD: PROJECTED BLOCK GRADIENT DESCENT

1: **input:** \mathbf{X}, \mathbf{y} , learning rates (μ_0, \dots, μ_G) , initialization $\beta^{(1)} = \mathbf{0}$
 2: **output:** $\hat{\beta}$
 3: **for** $t = 1$ **to** T **do**
 4: **for** $g=1$ **to** G **do**
 5: $\beta_g^{(t+1)} = \Pi_{\Omega_{f_g}} \left(\beta_g^{(t)} + \mu_g \mathbf{X}_g^T \left(\mathbf{y}_g - \mathbf{X}_g \left(\beta_0^{(t)} + \beta_g^{(t)} \right) \right) \right)$
 6: **end for**
 7: $\beta_0^{(t+1)} = \Pi_{\Omega_{f_0}} \left(\beta_0^{(t)} + \mu_0 \mathbf{X}_0^T \left(\mathbf{y} - \mathbf{X}_0 \beta_0^{(t)} - \begin{pmatrix} \mathbf{X}_1 \beta_1^{(t)} \\ \vdots \\ \mathbf{X}_G \beta_G^{(t)} \end{pmatrix} \right) \right)$
 8: **end for**

5.1 Convergence Rate Analysis

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

Definition 5.1. We define the following positive constants for $\mu_g > 0$:

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

where $\mathcal{B}_g = \mathcal{C}_g \cap \mathbb{B}^p$ is the intersection of the error cone and the unit ball and $\boldsymbol{\omega}_0 := \boldsymbol{\omega}$.

In the following theorem, we establish a deterministic bound on iteration errors which depends on constants defined in Definition 5.1.

Theorem 5.2. For the PBGD algorithm 1 initialized by $\beta^{(1)} = \mathbf{0}$, we have the following deterministic bound for the error at iteration $t + 1$:

$$\sum_{g=0}^G \sqrt{\frac{n_g}{n}} \|\delta_g^{(t+1)}\|_2 \leq \rho^t \sum_{g=0}^G \sqrt{\frac{n_g}{n}} \|\beta_g^*\|_2 + \frac{1 - \rho^t}{1 - \rho} \sum_{g=0}^G \sqrt{\frac{n_g}{n}} \eta_g \|\boldsymbol{\omega}_g\|_2, \quad (7)$$

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

The RHS of (7) consists of two terms. If we keep $\rho < 1$, the first term approaches zero exponentially fast, i.e., with linear rate, and the second term determines the bound. In the following we show that for specific choices of step sizes μ_g s, 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 4.3. Therefore, if we can keep ρ below one, the estimation error of PBGD algorithm geometrically converges to the approximate statistical error bound.

One way for having $\rho < 1$ is to keep all arguments of $\max(\dots)$ determining 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) 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 Section 6, we empirically illustrate such geometric convergence. The high probability bounds for constants in Definition 5.1 and the deterministic bound of Theorem 5.2 leads to the following theorem which shows for enough number of samples, of the same order as the statistical sample complexity, we can keep ρ below one and have geometric convergence.

Theorem 5.3. Let $\tau = C \sqrt{\log(G+1)} + b$ for $b > 0$ and $\omega_{0g} = \omega(\mathcal{A}_0) + \omega(\mathcal{A}_g)$. For the per group step sizes of:

$$\mu_0 = \frac{1}{4n} \times \min_{g \in [G] \setminus} \left(1 + c_{0g} \frac{\omega_{0g} + \tau}{\sqrt{n_g}} \right)^{-2}, \quad \text{and} \quad \mu_g = \frac{1}{2\sqrt{nn_g}} \left(1 + c_{0g} \frac{\omega_{0g} + \tau}{\sqrt{n_g}} \right)^{-1}$$

and sample complexities of:

$$n > 2c_0^2(2\omega(\mathcal{A}_0) + \tau)^2, \quad \text{and} \quad \forall g \in [G] \setminus \{0\} : n_g \geq 2c_g^2(2\omega(\mathcal{A}_g) + \tau)^2$$

updates of the Algorithm 1 obey the following with high probability:

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

where $r(\tau) < 1$.

Corollary 5.4. When $t \rightarrow \infty$ we have the following with high probability:

$$\sum_{g=0}^G \sqrt{\frac{n_g}{n}} \|\delta_g^\infty\|_2 \leq \frac{(G+1)\sqrt{(2K^2+1)}}{\sqrt{n}(1-r(\tau))} \left(\zeta k \max_{g \in [G]} \omega(\mathcal{A}_g) + C\sqrt{\log(G+1)} + b \right), \quad (8)$$

It is instructive to compare RHS of (8) with that of (6): κ_{\min} defined in Theorem 3.4 corresponds to $(1-r(\tau))$ and the extra $G+1$ factor corresponds to $\gamma = \max_{g \in [G]} \frac{n}{n_g}$. Therefore, Corollary 5.4 shows that PBGD converges to a scaled variant of statical error bound determined in Corollary 4.3.

6 Synthetic Experiments

We considered sparsity based simulations with varying G and sparsity levels. In our first set of simulations, we set $p = 100$, $G = 10$ and sparsity of the private parameteres to be $s = 10$. We generated a dense β_0 with $\|\beta_0\| = p$ and did not impose any constraint. Iterates $\{\beta_g^{(t)}\}_{g=1}^G$ are obtained by projection onto the ℓ_1 ball $\|\beta_g\|_1$. Nonzero entries of β_g are generated with $\mathcal{N}(0, 1)$ and nonzero supports are picked uniformly at random. Inspired from our theoretical step size choices, in all experiments, we used simplified learning rates of $\frac{1}{n}$ for β_0 and $\frac{1}{\sqrt{nn_g}}$ for β_g , $g \geq 1$. Observe that, cones of the individual parameters intersect with that of β_0 hence this setup actually violates DEIC (which requires an arbitrarily small constant fraction of groups to be non-intersecting). Our intuition is that the individual parameters are mostly incoherent with each other and the existence of a nonzero perturbation over β_g 's that keeps all measurements intact is unlikely. Remarkably, experimental results still show successful learning of all parameters from small amount of samples. We picked $n_g = 60$ for each group. Hence, in total, we have $11p = 1100$ unknowns, $200 = G \times 10 + 100$ degrees of freedom and $G \times 60 = 600$ samples. In all figures, we study the normalized squared error $\frac{\|\beta_g^{(t)} - \beta_g\|_2^2}{\|\beta_g\|_2^2}$ and average 10 independent realization for each curve. Figure 1a shows the estimation performance as a function of iteration number t . While each group might behave slightly different, we do observe that all parameters are linear converging to ground truth.

In Figure 1b, we test the noise robustness of our algorithm. We add a $\mathcal{N}(0, 1)$ noise to the $n_1 = 60$ measurements of the first group *only*. The other groups are left untouched. While all parameters suffer nonzero estimation error, we observe that, the global parameter β_0 and noise-free groups $\{\beta_g\}_{g=2}^G$ have substantially less estimation error. This implies that noise in one group mostly affects itself rather than the global estimation. In Figure 1c, we increased the sample size to $n_g = 150$ per group. We observe that, in comparison to Figure 1a, rate of convergence receives a boost from the additional samples as predicted by our theory.

Finally, Figure 1d considers a very high-dimensional problem where $p = 1000$, $G = 100$, individual parameters are 10 sparse, β_0 is 100 sparse and $n_g = 150$. The total degrees of freedom is 1100, number of unknowns are 101000 and total number of datapoints are $150 \times 100 = 15000$. While individual parameters have substantial variation in terms of convergence rate, at the end of 1000 iteration, all parameters have relative reconstruction error below 10^{-6} .

7 Drug Sensitivity Analysis for Cancer Cell Lines

In this section we investigate the application of DS in analyzing the response of patients with cancer to different doses of various drugs. Each cancer type (lung, blood, etc.) is a group g in our DS model and the respond of patient i with cancer g to the drug is our output y_{gi} . The set of features for each patient \mathbf{x}_{gi} consists of gene expressions, copy number variation, and mutations and y_{gi} is the ‘‘activity

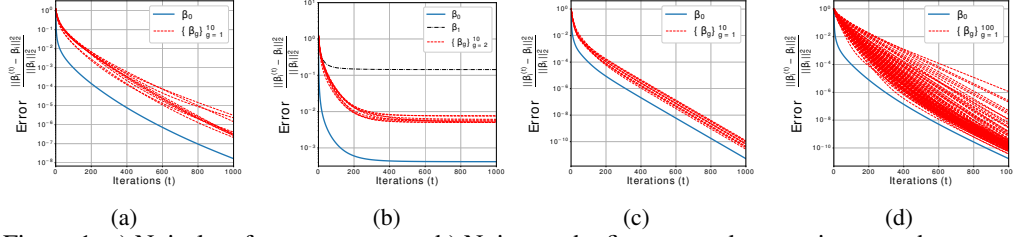


Figure 1: a) Noiseless fast convergence. b) Noise on the first group does not impact other groups as much. c) Increasing sample size improves rate of convergence. d) Our algorithm converges fast even with a large number of groups $G = 100$.

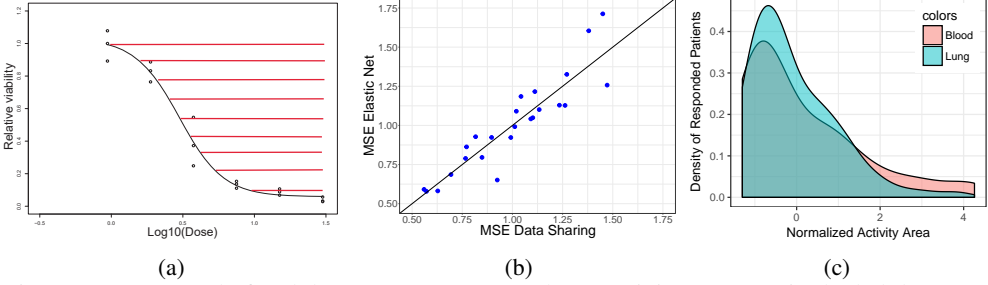


Figure 2: a) A sample fitted dose-response curve where Activity Area y_{gi} is shaded. b) Comparison of Mean Square Error of elastic net and data enrichment in predicting the response to 24 drugs for lung and blood cancer cell lines. Each dot represents an experiment for a drug. Prediction accuracy of both algorithms are very close. c) Distribution of responses to Saracatinib.

area” above the dose-response curve, Figure 2a. Given x_{gi} and a drug, we have two goals: accurately predict a patient’s response to the drug and identifying genetic predictors of drug sensitivity.

We use Cancer Cell Line Encyclopedia (CCLE) [2] which is a compilation ~ 500 human cancer cell lines where responses of them to 24 anticancer drugs have been measured. From the 36 cancer type available in CCLE, we focus on lung and blood. Not all of the 500 lines have been treated with all of the drugs. Therefore we end up with a different number of samples n for each drug where the range is $n \in [150, 200]$. Also, we perform a standard preprocessing [2] where we remove features with less than .1 absolute correlation with the response of interest which reduce the dimension to $p \in [1500, 5000]$ range.

Prediction: Here we run, 24 different experiments, each for one drug. Since the values of d_g in constraint sets $\Omega_{f_g}(d_g)$ are unknown, we tune them by 10-fold cross-validation and report the mean squared error (MSE) of the Elastic Net [21] (method used in the original CCLE paper[2]) and the data enrichment in Figure 2b. Both methods have very close prediction performance.

Interpretation We select Saracatinib, a drug that works on both lung and blood cancers, Figure 2c. Fixing the d_g parameters, we select the genes which have non-zero coefficient 40 times across 50 runs of PBGD on bootstrapped samples. Now, we have three lists of genes based on the supports of shared, lung, and blood parameters. We perform gene enrichment analysis using ToppGene [5] to see where in functional/disease/drug databases these genes have been observed together with statistical significance. Table 1 summarizes a highlight of our findings which shows lung and blood parameters are correctly capturing a meaningful set of genes.

(Blood, 512)		(Lung, 500)	
Highlights	p-Val	Highlights	p-Val
Regulation of immune response	2.1E-8	Secondary malignant neoplasm of Lung	8.9E-6
T cell activation	5.0E-8	Lung cancer	2.9E-5
Leukocyte activation	1.0E-6	Adenosquamous cell lung cancer	3.9E-5

Table 1: Each column is (Cancer Type, Number of significant genes) and highlights show where the set of genes have been observed together. p-Values are computed by Fisher’s exact test [5].

References

- [1] A. Banerjee, S. Chen, F. Fazayeli, and V. Sivakumar. Estimation with Norm Regularization. In *Advances in Neural Information Processing Systems*, pages 1556–1564, 2014.
- [2] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603, 2012.
- [3] Stephane Boucheron, Gabor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.
- [4] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.
- [5] Jing Chen, Eric E Bardes, Bruce J Aronow, and Anil G Jegga. Toppgene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*, 37(suppl_2):W305–W311, 2009.
- [6] F. Dondelinger and S. Mukherjee. High-dimensional regression over disease subgroups. *arXiv preprint arXiv:1611.00953*, 2016.
- [7] S. M. Gross and R. Tibshirani. Data shared lasso: A novel tool to discover uplift. *Computational Statistics & Data Analysis*, 101:226–235, 2016.
- [8] Q. Gu and A. Banerjee. High dimensional structured superposition models. In *Advances In Neural Information Processing Systems*, pages 3684–3692, 2016.
- [9] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A Dirty Model for Multi-task Learning. In *Advances in Neural Information Processing Systems*, pages 964–972, 2010.
- [10] Michael B McCoy and Joel A Tropp. The achievable performance of convex demixing. *arXiv preprint arXiv:1309.7478*, 2013.
- [11] Shahar Mendelson. Learning Without Concentration. In *Journal of the ACM (JACM)*. To appear, 2014.
- [12] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A Unified Framework for High-Dimensional Analysis of ℓ_1 -Estimators with Decomposable Regularizers. *Statistical Science*, 27(4):538–557, 2012.
- [13] E. Ollier and V. Viallon. Joint estimation of k related regression models with simple ℓ_1 -norm penalties. *arXiv preprint arXiv:1411.1594*, 2014.
- [14] E. Ollier and V. Viallon. Regression modeling on stratified data with the lasso. *arXiv preprint arXiv:1508.05476*, 2015.
- [15] S. Oymak, B. Recht, and M. Soltanolkotabi. Sharp time–data tradeoffs for linear inverse problems. *arXiv preprint arXiv:1507.04793*, 2015.
- [16] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.
- [17] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447, 2013.
- [18] Joel A. Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory - a Renaissance*. To appear, may 2015.
- [19] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, pages 210–268. Cambridge University Press, Cambridge, 2012.
- [20] E. Yang and P. Ravikumar. Dirty statistical models. In *Advances in Neural Information Processing Systems*, pages 611–619, 2013.
- [21] H. Zou and T. Hastie. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society*, 67(2):301–320, 2005.

A Proofs

A.1 Proof of Theorem 2.1

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

$$\begin{aligned} \frac{1}{n} \|\mathbf{X}\boldsymbol{\delta}\|_2^2 &\geq \frac{1}{n} \inf_{\mathbf{u} \in \mathcal{H}} \|\mathbf{X}\mathbf{u}\|_2^2 \left(\sum_{g=0}^G \frac{n_g}{n} \|\boldsymbol{\delta}_g\|_2 \right)^2 \\ &\geq \kappa \left(\sum_{g=0}^G \frac{n_g}{n} \|\boldsymbol{\delta}_g\|_2 \right)^2 \\ &\geq \kappa \left(\min_{g \in [G]} \frac{n_g}{n} \right) \left(\sum_{g=0}^G \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g\|_2 \right)^2 \end{aligned} \quad (9)$$

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:

$$\frac{2}{n} \boldsymbol{\omega}^T \mathbf{X}\boldsymbol{\delta} \leq \frac{2}{n} \sup_{\mathbf{u} \in \mathcal{H}} \boldsymbol{\omega}^T \mathbf{X}\mathbf{u} \left(\sum_{g=0}^G \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g\|_2 \right), \quad \mathbf{u} \in \mathcal{H} \quad (10)$$

Putting together inequalities (9) and (10) completes the proof. \blacksquare

A.2 Proof of Proposition 3.3

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 [8] the sample complexity required for the RE condition and subsequently recovering $\boldsymbol{\beta}_0^*$ and $\boldsymbol{\beta}_g^*$ is $n_g \geq c(\max_{g \in [G]} \omega(\mathcal{A}_g) + \sqrt{\log 2})^2$. \blacksquare

A.3 Proof of Theorem 3.4

Let's simplify the LHS of the RE condition:

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

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:

$$\begin{aligned} \inf_{\boldsymbol{\delta} \in \mathcal{H}} \frac{1}{\sqrt{n}} \|\mathbf{X}\boldsymbol{\delta}\|_2 &\geq \inf_{\boldsymbol{\delta} \in \mathcal{H}} \sum_{g=1}^G \frac{n_g}{n} \xi_g Q_{2\xi_g}(\boldsymbol{\delta}_{0g}) \\ &\quad - \sup_{\boldsymbol{\delta} \in \mathcal{H}} \frac{1}{n} \sum_{g=1}^G \xi_g \sum_{i=1}^{n_g} [Q_{2\xi_g}(\boldsymbol{\delta}_{0g}) - \mathbb{1}(|\langle \mathbf{x}_{gi}, \boldsymbol{\delta}_{0g} \rangle| \geq \xi_g)] \\ &= t_1(\mathbf{X}) - t_2(\mathbf{X}) \end{aligned} \quad (11)$$

For the ease of exposition we have written the LHS of (11) 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 DEIC condition of the Definition 3.2 and provides a lower bound for the first term $t_1(\mathbf{X})$:

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

$$\sum_{g=1}^G \frac{n_g}{n} \xi_g Q_{2\xi_g}(\delta_{0g}) \geq \psi_{\mathcal{I}} \xi \frac{(\alpha - 2\xi)^2}{4ck^2} \left(\|\delta_0\|_2 + \sum_{g=1}^n \frac{n_g}{n} \|\delta_g\|_2 \right), \quad (12)$$

which implies that $t_1(\mathbf{X}) = \inf_{\delta \in \mathcal{H}} \sum_{g=1}^G \frac{n_g}{n} \xi_g Q_{2\xi_g}(\delta_{0g})$ satisfies the same RHS bound of (12).

Proof. LHS of (12) is the weighted summation of $\xi_g Q_{2\xi_g}(\delta_{0g}) = \|\delta_{0g}\|_2 \xi \mathbb{P}(|\langle \mathbf{x}_g, \delta_{0g} / \|\delta_{0g}\|_2 \rangle| > 2\xi) = \|\delta_{0g}\|_2 \xi Q_{2\xi}(\mathbf{u})$ where $\xi > 0$ and $\mathbf{u} = \delta_{0g} / \|\delta_{0g}\|_2$ is a unit length vector. So we can rewrite the LHS of (12) as:

$$\sum_{g=1}^G \frac{n_g}{n} \xi_g Q_{2\xi_g}(\delta_{0g}) = \sum_{g=1}^G \frac{n_g}{n} \|\delta_0 + \delta_g\|_2 \xi Q_{2\xi}(\mathbf{u})$$

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

Lemma A.2. *Let \mathbf{u} be any unit length vector and suppose \mathbf{x} obeys Definition 3.1. Then for any \mathbf{u} , we have*

$$Q_{2\xi}(\mathbf{u}) \geq \frac{(\alpha - 2\xi)^2}{4ck^2}. \quad (13)$$

Lemma A.3. *Suppose Definition 3.2 holds. Then, we have:*

$$\sum_{i=1}^G n_i \|\delta_0 + \delta_i\|_2 \geq \frac{\bar{\rho} \lambda_{\min}}{3} \left(Gn \|\delta_0\|_2 + \sum_{i=1}^G n_i \|\delta_i\|_2 \right), \quad \forall i \in [G] : \delta_i \in \mathcal{C}_i. \quad (14)$$

■

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 [3]. 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:

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

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

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

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

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

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 [3, Theorem 6.2] 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 A.4. *For the random vector \mathbf{x} of Definition 3.1, we have the following bound:*

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

A.3.3 Continuing the Proof of Theorem 3.4

Set $n_0 = n$. Putting back bounds of $t_1(\mathbf{X})$ and $t_2(\mathbf{X})$ together from Lemma A.1 and A.4, 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_g s should be bounded away from zero. To this end we need the follow sample complexities:

$$\forall g \in [G]: \quad \left(\frac{2ck}{\psi_{\mathcal{I}} \xi q} \right)^2 \omega(\mathcal{A}_g)^2 \leq n_g \quad (15)$$

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

$$\forall g \in [G]: \quad \left(\frac{Ck^3}{\psi_{\mathcal{I}} \alpha^3} \right)^2 \omega(\mathcal{A}_g)^2 \leq n_g \quad (16)$$

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

A.4 Proof of Lemma 4.1

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 $s_g = \sqrt{\frac{n}{n_g}} \sqrt{(2K^2 + 1)n_g}$.

$$\begin{aligned} \mathbb{P}(h_g(\boldsymbol{\omega}_g, \mathbf{X}_g) > e_g s_g) &= \mathbb{P}\left(h_g(\boldsymbol{\omega}_g, \mathbf{X}_g) > e_g s_g \mid \sqrt{\frac{n}{n_g}} \|\boldsymbol{\omega}_g\|_2 > s_g\right) \mathbb{P}\left(\sqrt{\frac{n}{n_g}} \|\boldsymbol{\omega}_g\|_2 > s_g\right) \\ &+ \mathbb{P}\left(h_g(\boldsymbol{\omega}_g, \mathbf{X}_g) > e_g s_g \mid \sqrt{\frac{n}{n_g}} \|\boldsymbol{\omega}_g\|_2 < s_g\right) \mathbb{P}\left(\sqrt{\frac{n}{n_g}} \|\boldsymbol{\omega}_g\|_2 < s_g\right) \\ &\leq \mathbb{P}\left(\sqrt{\frac{n}{n_g}} \|\boldsymbol{\omega}_g\|_2 > s_g\right) + \mathbb{P}\left(h_g(\boldsymbol{\omega}_g, \mathbf{X}_g) > e_g s_g \mid \sqrt{\frac{n}{n_g}} \|\boldsymbol{\omega}_g\|_2 < s_g\right) \\ &\leq \mathbb{P}\left(\|\boldsymbol{\omega}_g\|_2 > \sqrt{(2K^2 + 1)n_g}\right) + \mathbb{P}\left(\sup_{\mathbf{u}_g \in \mathcal{C}_g \cap \mathbb{S}^{p-1}} \langle \mathbf{X}_g^T \frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \mathbf{u}_g \rangle > e_g\right) \\ &\leq \mathbb{P}\left(\|\boldsymbol{\omega}_g\|_2 > \sqrt{(2K^2 + 1)n_g}\right) + \sup_{\mathbf{v} \in \mathbb{S}^{p-1}} \mathbb{P}\left(\sup_{\mathbf{u}_g \in \mathcal{C}_g \cap \mathbb{S}^{p-1}} \langle \mathbf{X}_g^T \mathbf{v}, \mathbf{u}_g \rangle > e_g\right) \end{aligned} \quad (17)$$

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 $\mathbb{E}\|\boldsymbol{\omega}_g\|_2^2 \leq n_g$ [1] 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 $\tau = 2K^2 n_g$:

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

Now we upper bound the second term of (17). Given any fixed $\mathbf{v} \in \mathbb{S}^{p-1}$, $\mathbf{X}_g \mathbf{v}$ is a sub-Gaussian random vector with $\|\mathbf{X}_g^T \mathbf{v}\|_{\psi_2} \leq C_g k$ [1]. From Theorem 9 of [1] for any $\mathbf{v} \in \mathbb{S}^{p-1}$ we have:

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

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 $\epsilon_g = \theta_g C_g k$.

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

Now we put back results to the original inequality (17):

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

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

A.5 Proof of Theorem 4.2

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

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

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 $\mathbf{a} = (a_0, \dots, a_G)$ and $\mathbf{b} = (b_0, \dots, b_G)$ for which we have:

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

Now we can go back to the original form:

$$\begin{aligned} \sup_{\boldsymbol{\delta} \in \mathcal{H}} \boldsymbol{\omega}^T \mathbf{X} \boldsymbol{\delta} &\leq \max_{g \in [G]} \langle \mathbf{X}_g^T \frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \frac{\boldsymbol{\delta}_g}{\|\boldsymbol{\delta}_g\|_2} \rangle \sqrt{\frac{n}{n_g}} \|\boldsymbol{\omega}_g\|_2 \\ &\leq \max_{g \in [G]} \sqrt{\frac{n}{n_g}} \|\boldsymbol{\omega}_g\|_2 \sup_{\mathbf{u}_g \in \mathcal{C}_g \cap \mathbb{S}^{p-1}} \langle \mathbf{X}_g^T \frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \mathbf{u}_g \rangle \end{aligned} \quad (18)$$

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 $e_g(\tau) = \sqrt{(2K^2 + 1)n_g} (v_g C_g k \omega(\mathcal{A}_g) + \epsilon_g \sqrt{\log G} + \tau)$. Then from (18), we have:

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

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

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

where $\sigma = \max_{g \in [G]} \sigma_g$. ■

A.6 Proof of Lemma A.5

Proof. We upper bound the individual error $\|\delta_g^{(t+1)}\|_2$ and the common one $\|\delta_0^{(t+1)}\|_2$ in the followings:

$$\begin{aligned}
\|\delta_g^{(t+1)}\|_2 &= \|\beta_g^{(t+1)} - \beta_g^*\|_2 \\
&= \left\| \Pi_{\Omega_{f_g}} \left(\beta_g^{(t)} + \mu_g \mathbf{X}_g^T (\mathbf{y}_g - \mathbf{X}_g (\beta_0^{(t)} + \beta_g^{(t)})) \right) - \beta_g^* \right\|_2 \\
(\text{Lemma 6.3 of [15]}) &= \left\| \Pi_{\Omega_{f_g} - \{\beta_g^*\}} \left(\beta_g^{(t)} + \mu_g \mathbf{X}_g^T (\mathbf{y}_g - \mathbf{X}_g (\beta_0^{(t)} + \beta_g^{(t)})) - \beta_g^* \right) \right\|_2 \\
&= \left\| \Pi_{\mathcal{E}_g} \left(\delta_g^{(t)} + \mu_g \mathbf{X}_g^T (\mathbf{y}_g - \mathbf{X}_g (\beta_0^{(t)} + \beta_g^{(t)}) - \mathbf{X}_g (\beta_0^* + \beta_g^*) + \mathbf{X}_g (\beta_0^* + \beta_g^*)) \right) \right\|_2 \\
&= \left\| \Pi_{\mathcal{E}_g} \left(\delta_g^{(t)} + \mu_g \mathbf{X}_g^T (\boldsymbol{\omega}_g - \mathbf{X}_g (\delta_0^{(t)} + \delta_g^{(t)})) \right) \right\|_2 \\
(\text{Lemma 6.4 of [15]}) &\leq \left\| \Pi_{\mathcal{C}_g} \left(\delta_g^{(t)} + \mu_g \mathbf{X}_g^T (\boldsymbol{\omega}_g - \mathbf{X}_g (\delta_0^{(t)} + \delta_g^{(t)})) \right) \right\|_2 \\
(\text{Lemma 6.2 of [15]}) &\leq \sup_{\mathbf{v} \in \mathcal{C}_g \cap \mathbb{B}^p} \mathbf{v}^T \left(\delta_g^{(t)} + \mu_g \mathbf{X}_g^T (\boldsymbol{\omega}_g - \mathbf{X}_g (\delta_0^{(t)} + \delta_g^{(t)})) \right) \\
(\mathcal{B}_g = \mathcal{C}_g \cap \mathbb{B}^p) &= \sup_{\mathbf{v} \in \mathcal{B}_g} \mathbf{v}^T \left(\delta_g^{(t)} + \mu_g \mathbf{X}_g^T (\boldsymbol{\omega}_g - \mathbf{X}_g (\delta_0^{(t)} + \delta_g^{(t)})) \right) \\
&\leq \sup_{\mathbf{v} \in \mathcal{B}_g} \mathbf{v}^T (\mathbf{I}_g - \mu_g \mathbf{X}_g^T \mathbf{X}_g) \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 \delta_0^{(t)} \\
&\leq \left\| \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} \\
&\quad + \mu_g \|\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) \|\delta_g^{(t)}\|_2 + \xi_g(\mu_g) \|\boldsymbol{\omega}_g\|_2 + \phi_g(\mu_g) \|\delta_0^{(t)}\|_2
\end{aligned}$$

So the final bound becomes:

$$\|\delta_g^{(t+1)}\|_2 \leq \rho_g(\mu_g) \|\delta_g^{(t)}\|_2 + \xi_g(\mu_g) \|\boldsymbol{\omega}_g\|_2 + \phi_g(\mu_g) \|\delta_0^{(t)}\|_2 \quad (19)$$

Now we upper bound the error of common parameter. Remember common parameter's update:

$$\begin{aligned}
\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). \\
\|\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 (\mathbf{y}_g - \mathbf{X}_g(\boldsymbol{\beta}_0^{(t)} + \boldsymbol{\beta}_g^{(t)})) \right) - \boldsymbol{\beta}_0^* \right\|_2 \\
\text{(Lemma 6.3 of [15])} &= \left\| \Pi_{\Omega_{f_0} - \{\boldsymbol{\beta}_0^*\}} \left(\boldsymbol{\beta}_0^{(t)} + \mu_0 \sum_{g=1}^G \mathbf{X}_g^T (\mathbf{y}_g - \mathbf{X}_g(\boldsymbol{\beta}_0^{(t)} + \boldsymbol{\beta}_g^{(t)})) - \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 (\mathbf{y}_g - \mathbf{X}_g(\boldsymbol{\beta}_0^{(t)} + \boldsymbol{\beta}_g^{(t)})) \right) \right\|_2 \\
\text{(Lemma 6.4 of [15])} &\leq \left\| \Pi_{\mathcal{C}_0} \left(\boldsymbol{\delta}_0^{(t)} + \mu_0 \sum_{g=1}^G \mathbf{X}_g^T (\boldsymbol{\omega}_g - \mathbf{X}_g(\boldsymbol{\delta}_0^{(t)} + \boldsymbol{\delta}_g^{(t)})) \right) \right\|_2 \\
\text{(Lemma 6.2 of [15])} &\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 (\boldsymbol{\omega}_g - \mathbf{X}_g(\boldsymbol{\delta}_0^{(t)} + \boldsymbol{\delta}_g^{(t)})) \right) \\
&\leq \sup_{\mathbf{v} \in \mathcal{B}_0} \mathbf{v}^T (\mathbf{I} - \mu_0 \sum_{g=1}^G \mathbf{X}_g^T \mathbf{X}_g) \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 \\
&\quad + \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 (\mathbf{I} - \mu_0 \mathbf{X}_0^T \mathbf{X}_0) \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 \\
&\quad + \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 \tag{20}
\end{aligned}$$

To avoid cluttering we drop μ_g as the arguments. Putting together (19) and (20) inequalities we reach to the followings:

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

■

A.7 Proof of Theorem 5.2

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

Lemma A.5. *We have the following recursive dependency between the error of $t + 1$ th iteration and t th iteration of PBGD:*

$$\begin{aligned}\|\delta_g^{(t+1)}\|_2 &\leq \left(\rho_g(\mu_g)\|\delta_g^{(t)}\|_2 + \xi_g(\mu_g)\|\omega_g\|_2 + \phi_g(\mu_g)\|\delta_0^{(t)}\|_2\right) \\ \|\delta_0^{(t+1)}\|_2 &\leq \left(\rho_0(\mu_0)\|\delta_0^{(t)}\|_2 + \xi_0(\mu_0)\|\omega_0\|_2 + \mu_0 \sum_{g=1}^G \frac{\phi_g(\mu_g)}{\mu_g} \|\delta_g^{(t)}\|_2\right)\end{aligned}$$

By recursively applying the result of Lemma A.5, we get the following deterministic bound which depends on constants defined in Definition 5.1:

$$\begin{aligned}b_{t+1} = \sum_{g=0}^G \sqrt{\frac{n_g}{n}} \|\delta_g^{(t+1)}\|_2 &\leq \left(\rho_0 + \sum_{g=1}^G \sqrt{\frac{n_g}{n}} \phi_g\right) \|\delta_0^{(t)}\|_2 + \sum_{g=1}^G \left(\sqrt{\frac{n_g}{n}} \rho_g + \mu_0 \frac{\phi_g}{\mu_g}\right) \|\delta_g^{(t)}\|_2 + \sum_{g=0}^G \sqrt{\frac{n_g}{n}} \xi_g \|\omega_g\|_2 \\ &\leq \rho \sum_{g=0}^G \sqrt{\frac{n_g}{n}} \|\delta_g^{(t)}\|_2 + \sum_{g=0}^G \sqrt{\frac{n_g}{n}} \xi_g \|\omega_g\|_2\end{aligned}\quad (21)$$

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

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

■

A.8 Proof of Theorem 5.3

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

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

$$\rho_g \left(\frac{1}{a_g n_g}\right) \leq \frac{1}{2} \left[\left(1 - \frac{1}{a_g}\right) + \sqrt{2} c_g \frac{2\omega(\mathcal{A}_g) + \tau}{a_g \sqrt{n_g}}\right] \quad (22)$$

Lemma A.7. *Consider $a_g \geq 1$, with probability at least $1 - 4 \exp(-\gamma_g(\omega(\mathcal{A}_g) + \tau)^2)$ the following upper bound holds:*

$$\phi_g \left(\frac{1}{a_g n_g}\right) \leq \frac{1}{a_g} \left(1 + c_{0g} \frac{\omega(\mathcal{A}_g) + \omega(\mathcal{A}_0) + 2\tau}{\sqrt{n_g}}\right) \quad (23)$$

Note that Lemma 4.1 readily provides a high probability upper bound for $\eta_g(1/(a_g n_g))$ as $\sqrt{(2K^2 + 1)} (\zeta_g k \omega(\mathcal{A}_g) + \epsilon_g \sqrt{\log G} + \tau) / (a_g \sqrt{n_g})$.

Starting from the deterministic form of the bound in Theorem 5.2 and putting in the step sizes as $\mu_g = \frac{1}{n_g a_g}$:

$$\sum_{g=0}^G \sqrt{\frac{n_g}{n}} \|\delta_g^{(t+1)}\|_2 \leq (\rho)^t \sum_{g=0}^G \|\beta_g^*\|_2 + \frac{1 - (\rho)^t}{1 - \rho} \sum_{g=0}^G \sqrt{\frac{n_g}{n}} \eta_g \left(\frac{1}{n_g a_g} \right) \|\omega_g\|_2, \quad (24)$$

where

$$\rho(a_0, \dots, a_G) = \max \left(\rho_0 \left(\frac{1}{n a_0} \right) + \sum_{g=1}^G \sqrt{\frac{n_g}{n}} \phi_g \left(\frac{1}{n_g a_g} \right), \max_{g \in [G]} \rho_g \left(\frac{1}{n_g a_g} \right) + \sqrt{\frac{n}{n_g}} \frac{\mu_0}{\mu_g} \phi_g \left(\frac{1}{n_g a_g} \right) \right) \quad (25)$$

Remember the following two results to upper bound ρ_g s and ϕ_g s from Lemmas A.6 and A.7:

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

First we want to keep $\rho_0 + \sum_{g=1}^G \sqrt{\frac{n_g}{n}} \phi_g$ of (25) strictly below 1.

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

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:

$$\sqrt{2} c_0 \frac{2\omega(\mathcal{A}_0) + \tau}{\sqrt{n}} \leq a_0 \left(1 - \sum_{g=1}^G \frac{n_g}{n} b_g \right) + 1 \quad (26)$$

Note that any lower bound on the RHS of (26) will lead to the correct sample complexity for which the coefficient of $\|\delta_0^{(t)}\|_2$ (determined in (25)) will be below one. Since $a_0 \geq 1$ we can ignore the first term by assuming $\max_{g \in [G] \setminus \{0\}} b_g \leq 1$ and the condition becomes:

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

which can be simplified to:

$$\begin{aligned} n &> 2c_0^2 (2\omega(\mathcal{A}_0) + \tau)^2, a_0 \geq 1, \quad (27) \\ \forall g \in [G] \setminus \{0\} : a_g &= 2b_g^{-1} \sqrt{\frac{n}{n_g}} \left(1 + c_{0g} \frac{\omega_{0g} + \tau}{\sqrt{n_g}} \right), 0 < b_g \leq 1 \end{aligned}$$

Secondly, we want to bound all of $\rho_g + \mu_0 \sqrt{\frac{n}{n_g} \frac{\phi_g}{\mu_g}}$ terms of (25) for $\mu_g = \frac{1}{a_g n_g}$ by 1:

$$\begin{aligned} \rho_g \left(\frac{1}{n_g a_g} \right) + \sqrt{\frac{n}{n_g} \frac{\mu_0}{\mu_g}} \phi_g \left(\frac{1}{n_g a_g} \right) &= \rho_g \left(\frac{1}{n_g a_g} \right) + \sqrt{\frac{n_g}{n} \frac{a_g}{a_0}} \phi_g \left(\frac{1}{n_g a_g} \right) \\ &= \frac{1}{2} \left[\left[\left(1 - \frac{1}{a_g} \right) + \sqrt{2c_g} \frac{2\omega_g + \tau}{a_g \sqrt{n_g}} \right] \right. \\ &\quad \left. + \frac{2}{a_0} \sqrt{\frac{n_g}{n}} \left(1 + c_{0g} \frac{\omega_{0g} + \tau}{\sqrt{n_g}} \right) \right] \\ &\leq 1 \end{aligned} \quad (28)$$

The condition becomes:

$$\sqrt{2c_g} \frac{2\omega_g + \tau}{\sqrt{n_g}} \leq a_g + 1 - \sqrt{\frac{n_g}{n} \frac{2a_g}{a_0}} \left(1 + c_{0g} \frac{\omega_{0g} + \tau}{\sqrt{n_g}} \right) \quad (29)$$

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:

$$\sqrt{2c_g} \frac{2\omega_g + \tau}{d_g} \leq \sqrt{n_g}, \quad d_g := a_g + 1 - \frac{4}{b_g a_0} \left(1 + c_{0g} \frac{\omega_{0g} + \tau}{\sqrt{n_g}} \right)^2 \quad (30)$$

for $d_g > 0$. Note that any positive lower bound of the d_g will satisfy the condition in (30) 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 (27), so we take $a_0 = 4 \max_{g \in [G] \setminus \{1\}} \left(1 + c_{0g} \frac{\omega_{0g} + \tau}{\sqrt{n_g}} \right)^2$ and look for a lower bound for d_g :

$$d_g \geq a_g + 1 - b_g^{-1} \quad (31)$$

$$\begin{aligned} (a_g \text{ from (27)}) &= 2b_g^{-1} \sqrt{\frac{n}{n_g}} \left(1 + c_{0g} \frac{\omega_{0g} + \tau}{\sqrt{n_g}} \right) + 1 - b_g^{-1} \\ &= 1 + b_g^{-1} \left[2\sqrt{\frac{n}{n_g}} \left(1 + c_{0g} \frac{\omega_{0g} + \tau}{\sqrt{n_g}} \right) - 1 \right] \end{aligned} \quad (32)$$

The term inside of the last bracket (32) is always positive and therefore a lower bound is one, i.e., $d_g \geq 1$. From the condition (30) we get the following sample complexity:

$$n_g > 2c_g^2 (2\omega_g + \tau)^2 \quad (33)$$

Now we need to determine b_g from previous conditions (27), knowing that $a_0 = 4 \max_{g \in [G] \setminus \{1\}} \left(1 + c_{0g} \frac{\omega_{0g} + \tau}{\sqrt{n_g}} \right)^2$. We have $0 < b_g \leq 1$ in (27) and we take the largest step by setting $b_g = 1$.

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

$$\begin{aligned} n &> 2c_0^2 (2\omega(\mathcal{A}_0) + \tau)^2, \forall g \in [G] \setminus \{1\} : n_g \geq 2c_g^2 (2\omega(\mathcal{A}_g) + \tau)^2 \\ a_0 &= 4 \max_{g \in [G] \setminus \{1\}} \left(1 + c_{0g} \frac{\omega_{0g} + \tau}{\sqrt{n_g}} \right)^2, a_g = 2\sqrt{\frac{n}{n_g}} \left(1 + c_{0g} \frac{\omega_{0g} + \tau}{\sqrt{n_g}} \right) \\ \mu_0 &= \frac{1}{4n} \times \frac{1}{\max_{g \in [G] \setminus \{1\}} \left(1 + c_{0g} \frac{\omega_{0g} + \tau}{\sqrt{n_g}} \right)^2}, \mu_g = \frac{1}{2\sqrt{nn_g}} \left(1 + c_{0g} \frac{\omega_{0g} + \tau}{\sqrt{n_g}} \right)^{-1} \end{aligned} \quad (34)$$

Now we rewrite the same analysis using the tail bounds for the coefficients to clarify the probabilities. To simplify the notation, let $r_{g1} = \frac{1}{2} \left[\left(1 - \frac{1}{a_g} \right) + \sqrt{2c_g} \frac{2\omega(\mathcal{A}_g) + \tau}{a_g \sqrt{n_g}} \right]$ 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 $r_g(\tau) = r_{g1} + \sqrt{\frac{n_g}{n} \frac{a_g}{a_0}} r_{g2}, \forall g \in [G] \setminus \{1\}$, and $r(\tau) =$

$\max_{g \in [G]} r_g$. All of which are computed using a_g s specified in (34). Basically r is an instantiation of an upper bound of the ρ defined in (25) using a_g s in (34).

We are interested to upper bound the following probability:

$$\begin{aligned}
& \mathbb{P} \left(\sum_{g=0}^G \sqrt{\frac{n_g}{n}} \|\delta_g^{(t+1)}\|_2 \geq r(\tau)^t \sum_{g=0}^G \sqrt{\frac{n_g}{n}} \|\beta_g^*\|_2 + \frac{(G+1)\sqrt{(2K^2+1)}}{(1-r(\tau))\sqrt{n}} \left(\zeta k \max_{g \in [G]} \omega(\mathcal{A}_g) + \tau \right) \right) \\
& \leq \mathbb{P} \left((\rho)^t \sum_{g=0}^G \sqrt{\frac{n_g}{n}} \|\beta_g^*\|_2 + \frac{1-(\rho)^t}{1-\rho} \sum_{g=0}^G \sqrt{\frac{n_g}{n}} \eta_g \left(\frac{1}{n_g a_g} \right) \|\omega_g\|_2 \right) \\
& \geq r(\tau)^t \sum_{g=0}^G \sqrt{\frac{n_g}{n}} \|\beta_g^*\|_2 + \frac{(G+1)\sqrt{(2K^2+1)}}{(1-r(\tau))\sqrt{n}} \left(\zeta k \max_{g \in [G]} \omega(\mathcal{A}_g) + \tau \right) \\
& \leq \mathbb{P}(\rho \geq r(\tau)) \\
& + \mathbb{P} \left(\frac{1}{1-\rho} \sum_{g=0}^G \sqrt{n_g} \eta_g \left(\frac{1}{n_g a_g} \right) \|\omega_g\|_2 \geq \frac{(G+1)\sqrt{(2K^2+1)}}{(1-r(\tau))} \left(\zeta k \max_{g \in [G]} \omega(\mathcal{A}_g) + \tau \right) \right) \quad (35)
\end{aligned}$$

where the first inequality comes from the deterministic bound of (24), We first focus on bounding the first term $\mathbb{P}(\rho \geq r(\tau))$:

$$\begin{aligned}
& \mathbb{P}(\rho \geq r(\tau)) \\
& = \mathbb{P} \left(\max \left(\rho_0 \left(\frac{1}{na_0} \right) + \sum_{g=1}^G \sqrt{\frac{n_g}{n}} \phi_g \left(\frac{1}{n_g a_g} \right), \max_{g \in [G]} \rho_g \left(\frac{1}{n_g a_g} \right) + \sqrt{\frac{n}{n_g}} \frac{\mu_0}{\mu_g} \phi_g \left(\frac{1}{n_g a_g} \right) \right) \geq \max_{g \in [G]} r(\tau) \right) \\
& \leq \mathbb{P} \left(\rho_0 \left(\frac{1}{na_0} \right) + \sum_{g=1}^G \sqrt{\frac{n_g}{n}} \phi_g \left(\frac{1}{n_g a_g} \right) \geq r_0 \right) + \sum_{g=1}^G \mathbb{P} \left(\rho_g \left(\frac{1}{n_g a_g} \right) + \sqrt{\frac{n}{n_g}} \frac{\mu_0}{\mu_g} \phi_g \left(\frac{1}{n_g a_g} \right) \geq r_g \right) \\
& \leq \mathbb{P} \left(\rho_0 \left(\frac{1}{na_0} \right) \geq r_{01} \right) + \sum_{g=1}^G \mathbb{P} \left(\phi_g \left(\frac{1}{n_g a_g} \right) \geq r_{g2} \right) + \sum_{g=1}^G \left[\mathbb{P} \left(\rho_g \left(\frac{1}{n_g a_g} \right) \geq r_{g1} \right) + \mathbb{P} \left(\phi_g \left(\frac{1}{n_g a_g} \right) \geq r_{g2} \right) \right] \\
& \leq \sum_{g=0}^G \mathbb{P} \left(\rho_g \left(\frac{1}{n_g a_g} \right) \geq r_{g1} \right) + 2 \sum_{g=1}^G \mathbb{P} \left(\phi_g \left(\frac{1}{n_g a_g} \right) \geq r_{g2} \right) \\
& \leq \sum_{g=0}^G 6 \exp(-\gamma_g(\omega(\mathcal{A}_g) + \tau)^2) + 2 \sum_{g=1}^G 4 \exp(-\gamma_g(\omega(\mathcal{A}_g) + \tau)^2) \\
& \leq 6(G+1) \exp \left(-\gamma \min_{g \in [G]} (\omega(\mathcal{A}_g) + \tau)^2 \right) + 8G \exp \left(-\gamma \min_{g \in [G] \setminus \{0\}} (\omega(\mathcal{A}_g) + \tau)^2 \right) \\
& \leq 14(G+1) \exp \left(-\gamma \min_{g \in [G]} (\omega(\mathcal{A}_g) + \tau)^2 \right) \quad (36)
\end{aligned}$$

Now we focus on bounding the second term:

$$\begin{aligned}
& \mathbb{P} \left(\frac{1}{1-\rho} \sum_{g=0}^G \sqrt{n_g} \eta_g \left(\frac{1}{n_g a_g} \right) \|\omega_g\|_2 \geq \frac{(G+1)\sqrt{(2K^2+1)}}{(1-r(\tau))} \left(\zeta k \max_{g \in [G]} \omega(\mathcal{A}_g) + \tau \right) \right) \\
& \leq \mathbb{P} \left(\frac{1}{1-\rho} \sum_{g=0}^G \sqrt{n_g} \eta_g \left(\frac{1}{n_g a_g} \right) \|\omega_g\|_2 \geq \frac{1}{(1-r(\tau))} \sum_{g=0}^G \sqrt{(2K^2+1)} (\zeta_g k \omega(\mathcal{A}_g) + \tau) \right) \\
& \leq \mathbb{P} \left(\sum_{g=0}^G \sqrt{n_g} \eta_g \left(\frac{1}{n_g a_g} \right) \|\omega_g\|_2 \geq \sum_{g=0}^G \sqrt{(2K^2+1)} (\zeta_g k \omega(\mathcal{A}_g) + \tau) \right) + \mathbb{P}(\rho \geq r(\tau)) \\
& \leq \sum_{g=0}^G \mathbb{P} \left(\sqrt{n_g} \eta_g \left(\frac{1}{n_g a_g} \right) \|\omega_g\|_2 \geq \sqrt{(2K^2+1)} (\zeta_g k \omega(\mathcal{A}_g) + \tau) \right) + \mathbb{P}(\rho \geq r(\tau)) \quad (37)
\end{aligned}$$

Focusing on the summand of the first term, remember from Definition 5.1 that $\eta_g(\mu_g) = \frac{1}{a_g n_g} \sup_{\mathbf{v} \in \mathcal{B}_g} \mathbf{v}^T \mathbf{X}_g^T \frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}$, $g \in [G]$ and $a_g \geq 1$:

$$\mathbb{P} \left(\|\boldsymbol{\omega}_g\|_2 \sup_{\mathbf{v} \in \mathcal{B}_g} \mathbf{v}^T \mathbf{X}_g^T \frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2} \geq a_g \sqrt{(2K^2 + 1)n_g} (\zeta_g k \omega(\mathcal{A}_g) + \tau) \right) \leq \sigma_g \exp \left(- \min \left[\nu_g n_g, \frac{\tau^2}{\eta_g^2 k^2} \right] \right) \quad (38)$$

where we used the intermediate form of Lemma 4.1 for $\tau > 0$. Putting all of the bounds (36), (37), and (38) back into the (35):

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

where $v = \max(28, \sigma)$ 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:

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

■

A.9 Proof of Lemma A.2

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 $\|\mathbf{x}\|_{\psi_2} \leq k$ [18].

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

■

A.10 Proof of Lemma A.3

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

$$\|\boldsymbol{\delta}_0 + \boldsymbol{\delta}_i\|_2 \geq \begin{cases} \lambda_{\min}(\|\boldsymbol{\delta}_i\|_2 + \|\boldsymbol{\delta}_0\|_2) & \text{if } i \in \mathcal{I} \\ \|\boldsymbol{\delta}_i\|_2/2 & \text{if } i \in \mathcal{J} \\ 0 & \text{if } i \in \mathcal{K} \end{cases} \quad (39)$$

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 3.2, $\sum_{i \in \mathcal{I}} n_i \geq \bar{\rho} n$ 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:

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

■

A.11 Proof of Lemma A.4

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:

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

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 $\delta_{0g} = \delta_0 + \delta_g$:

$$\begin{aligned}
\frac{2}{n} \mathbb{E} \sup_{\delta_{[G]} \in \mathcal{C}_{[G]}} \sum_{g=1}^G \sum_{i=1}^{n_g} \epsilon_{gi} \langle \mathbf{x}_{gi}, \delta_{0g} \rangle &= \frac{2}{n} \mathbb{E} \sup_{\delta_0 \in \mathcal{C}_0} \sum_{i=1}^n \epsilon_i \langle \mathbf{x}_i, \delta_0 \rangle + \frac{2}{n} \mathbb{E} \sup_{\delta_{[G] \setminus \in \mathcal{C}_{[G] \setminus}} \sum_{g=1}^G \sum_{i=1}^{n_g} \epsilon_{gi} \langle \mathbf{x}_{gi}, \delta_g \rangle \\
&= \frac{2}{\sqrt{n}} \mathbb{E} \sup_{\delta_0 \in \mathcal{C}_0} \sum_{i=1}^n \left\langle \frac{1}{\sqrt{n}} \epsilon_i \mathbf{x}_i, \delta_0 \right\rangle + \frac{2}{\sqrt{n}} \mathbb{E} \sup_{\delta_{[G] \setminus \in \mathcal{C}_{[G] \setminus}} \sum_{g=1}^G \sqrt{\frac{n_g}{n}} \sum_{i=1}^{n_g} \left\langle \frac{1}{\sqrt{n_g}} \epsilon_{gi} \mathbf{x}_{gi}, \delta_g \right\rangle \\
(n_0 := n, \epsilon_{0i} := \epsilon_0, \mathbf{x}_{0i} := \mathbf{x}_i) &= \frac{2}{\sqrt{n}} \mathbb{E} \sup_{\delta_{[G]} \in \mathcal{C}_{[G]}} \sum_{g=0}^G \sqrt{\frac{n_g}{n}} \sum_{i=1}^{n_g} \left\langle \frac{1}{\sqrt{n_g}} \epsilon_{gi} \mathbf{x}_{gi}, \delta_g \right\rangle \\
(\mathbf{h}_g := \frac{1}{\sqrt{n_g}} \sum_{i=1}^{n_g} \epsilon_{gi} \mathbf{x}_{gi}) &= \frac{2}{\sqrt{n}} \mathbb{E} \sup_{\delta_{[G]} \in \mathcal{C}_{[G]}} \sum_{g=0}^G \sqrt{\frac{n_g}{n}} \langle \mathbf{h}_g, \delta_g \rangle \\
(\mathcal{A}_g \in \mathcal{C}_g \cap \mathbb{S}^{p-1}) &\leq \frac{2}{\sqrt{n}} \mathbb{E} \sup_{\delta_{[G]} \in \mathcal{A}_{[G]}} \sum_{g=0}^G \sqrt{\frac{n_g}{n}} \langle \mathbf{h}_g, \delta_g \rangle \|\delta_g\|_2 \\
&\leq \frac{2}{\sqrt{n}} \sum_{g=0}^G \sqrt{\frac{n_g}{n}} \mathbb{E}_{\mathbf{h}_g} \sup_{\delta_g \in \mathcal{A}_g} \langle \mathbf{h}_g, \delta_g \rangle \|\delta_g\|_2 \\
&\leq \frac{2}{\sqrt{n}} \sum_{g=0}^G \sqrt{\frac{n_g}{n}} c_g k \omega(\mathcal{A}_g) \|\delta_g\|_2
\end{aligned}$$

Note that the \mathbf{h}_{gi} is a sub-Gaussian random vector which let us bound the $\mathbb{E} \sup$ using the Gaussian width [18] in the last step. \blacksquare

A.12 Proof of Lemma A.6

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 A.8 (Theorem 11 of [1]). *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 $1 - 2 \exp(-\gamma_g(\omega(\mathcal{A}_g) + \tau)^2)$ for $\tau > 0$:*

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

where $c_g > 0$ is constant.

The statement of Lemma A.8 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:

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

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

Now the proof for Lemma A.6:

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

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

We upper bound the argument of the sup as follows:

$$\begin{aligned}
\mathbf{v}^T (\mathbf{I}_g - \mu_g \mathbf{X}_g^T \mathbf{X}_g) \mathbf{u} &= \frac{1}{4} [(\mathbf{u} + \mathbf{v})^T (\mathbf{I} - \mu_g \mathbf{X}_g^T \mathbf{X}_g) (\mathbf{u} + \mathbf{v}) - (\mathbf{u} - \mathbf{v})^T (\mathbf{I} - \mu_g \mathbf{X}_g^T \mathbf{X}_g) (\mathbf{u} - \mathbf{v})] \\
&= \frac{1}{4} [\|\mathbf{u} + \mathbf{v}\|_2^2 - \mu_g \|\mathbf{X}_g (\mathbf{u} + \mathbf{v})\|_2^2 - \|\mathbf{u} - \mathbf{v}\|_2^2 + \mu_g \|\mathbf{X}_g (\mathbf{u} - \mathbf{v})\|_2^2] \\
(\text{Lemma A.8}) &\leq \frac{1}{4} \left[\left(1 - \mu_g n_g \left(1 - c_g \frac{2\omega(\mathcal{A}_g) + \tau}{\sqrt{n_g}} \right) \right) \|\mathbf{u} + \mathbf{v}\|_2 \right. \\
&\quad \left. - \left(1 - \mu_g n_g \left(1 + c_g \frac{2\omega(\mathcal{A}_g) + \tau}{\sqrt{n_g}} \right) \right) \|\mathbf{u} - \mathbf{v}\|_2 \right] \\
\left(\mu_g = \frac{1}{a_g n_g} \right) &\leq \frac{1}{4} \left[\left(1 - \frac{1}{a_g} \right) (\|\mathbf{u} + \mathbf{v}\|_2 - \|\mathbf{u} - \mathbf{v}\|_2) + c_g \frac{2\omega(\mathcal{A}_g) + \tau}{a_g \sqrt{n_g}} (\|\mathbf{u} + \mathbf{v}\|_2 + \|\mathbf{u} - \mathbf{v}\|_2) \right] \\
&\leq \frac{1}{4} \left[\left(1 - \frac{1}{a_g} \right) 2\|\mathbf{v}\|_2 + c_g \frac{2\omega(\mathcal{A}_g) + \tau}{a_g \sqrt{n_g}} 2\sqrt{2} \right]
\end{aligned}$$

where the last line follows from the triangle inequality and the fact that $\|\mathbf{u} + \mathbf{v}\|_2 + \|\mathbf{u} - \mathbf{v}\|_2 \leq 2\sqrt{2}$ which itself follows from $\|\mathbf{u} + \mathbf{v}\|_2^2 + \|\mathbf{u} - \mathbf{v}\|_2^2 \leq 4$. Note that we applied the Lemma A.8 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 :

$$\mathbf{v}^T (\mathbf{I}_g - \frac{1}{a_g n_g} \mathbf{X}_g^T \mathbf{X}_g) \mathbf{u} \leq \frac{1}{2} \left[\left(1 - \frac{1}{a_g} \right) + \sqrt{2} c_g \frac{2\omega(\mathcal{A}_g) + \tau}{a_g \sqrt{n_g}} \right] \quad (40)$$

To keep the upper bound of ρ_g in (40) 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:

$$\begin{aligned}
\mathbb{P}(\rho_g \geq d_g) &\leq \mathbb{P}(\rho_g \geq d_g | \neg \mathcal{E}_1, \neg \mathcal{E}_2) + 2\mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_2) \\
(\text{Lemma A.8}) &\leq 0 + 6 \exp(-\gamma_g(\omega(\mathcal{A}_g) + \tau)^2)
\end{aligned}$$

which concludes the proof. ■

A.13 Proof of Lemma A.7

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

$$-\mathbf{v}^T \mathbf{X}_g^T \mathbf{X}_g \mathbf{u} \leq \frac{1}{2} (\|\mathbf{X}_g \mathbf{u}\|_2^2 + \|\mathbf{X}_g \mathbf{v}\|_2^2) \quad (41)$$

Now we can bound ϕ_g as follows:

$$\phi_g(\mu_g) = \mu_g \sup_{\mathbf{v} \in \mathcal{B}_g, \mathbf{u} \in \mathcal{B}_0} -\mathbf{v}^T \mathbf{X}_g^T \mathbf{X}_g \mathbf{u} \leq \frac{\mu_g}{2} \left(\sup_{\mathbf{u} \in \mathcal{B}_0} \|\mathbf{X}_g \mathbf{u}\|_2^2 + \sup_{\mathbf{v} \in \mathcal{B}_g} \|\mathbf{X}_g \mathbf{v}\|_2^2 \right) \quad (42)$$

So we have:

$$\begin{aligned}
\phi_g \left(\frac{1}{a_g n_g} \right) &\leq \frac{1}{2a_g} \left(\frac{1}{n_g} \sup_{\mathbf{u} \in \mathcal{B}_0} \|\mathbf{X}_g \mathbf{u}\|_2^2 + \frac{1}{n_g} \sup_{\mathbf{v} \in \mathcal{B}_g} \|\mathbf{X}_g \mathbf{v}\|_2^2 \right) \quad (43) \\
(\text{Lemma A.8}) &\leq \frac{1}{a_g} \left(1 + c_{0g} \frac{\omega(\mathcal{A}_g) + \omega(\mathcal{A}_0) + 2\tau}{2\sqrt{n_g}} \right) \\
(\omega_{0g} = \max(\omega(\mathcal{A}_0), \omega(\mathcal{A}_g))) &\leq \frac{1}{a_g} \left(1 + c_{0g} \frac{\omega_{0g} + \tau}{\sqrt{n_g}} \right)
\end{aligned}$$

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.

$$\begin{aligned}
\mathbb{P}(\phi_g > s_g) &\leq \mathbb{P}(\phi_g > s_g | \neg \mathcal{E}_1) \mathbb{P}(\neg \mathcal{E}_1) + \mathbb{P}(\mathcal{E}_1) \\
&\leq \mathbb{P}(\mathcal{E}_2) + \mathbb{P}(\mathcal{E}_1) \\
&\leq 4 \exp(-\gamma_g (\omega(\mathcal{A}_g) + \tau)^2)
\end{aligned} \tag{44}$$

■