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

Anonymous Authors[1]

Abstract

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

1. Introduction

Over the past two decades, major advances have been made in estimating structured parameters, e.g., sparse, low-rank, etc., in high-dimensional small sample problems [][][]. Such estimators consider a suitable (semi) parametric model of the response: $y = \phi(\mathbf{x}, \boldsymbol{\beta}^*) + \omega$ based on n samples $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ and $\beta \in \mathbb{R}^p$ is the parameter of interest. The unique aspect of such high-dimensional setup is that the number of samples $n < p$, and the structure in $\boldsymbol{\beta}^*$, e.g., sparsity, low-rank, makes the estimation possible [][][]. While the earlier developments in such high dimensional estimation problems had focused on parametric linear models, the results have been widely extended to non-linear models, e.g., generalized linear models [][], broad families of semi-parametric and single-index models [][], non-convex models [][], etc.

In several real world problems, the assumption that one global model parameter $\boldsymbol{\beta}$ is suitable for the entire population is unrealistic. We consider the more general setting where the population consists of sub-populations (groups) which are similar is many aspects but are different in certain unique aspects. For example, in the context of **(Arindam Says: ... a line on drug sensitivity of cancer cell lines ...)**. In such a setting, one can assume a model for each group based on a group specific parameter $\boldsymbol{\beta}_g^*$, i.e., $y = \phi(\mathbf{x}, \boldsymbol{\beta}_g^*) + \omega$ for group g. Such a modeling approach fails to leverage the similarities across the sub-populations, and can only be estimated when sufficient number of samples are available for each sub-population which is not the case in several problems, e.g., drug sensitivity of cancer cell lines [][].

In this paper, we consider the *data enrichment* strategy recently suggested in the literature (Dondelinger & Mukherjee, 2016; Gross & Tibshirani, 2016; Ollier & Viallon, 2014; 2015) for the above problem setup. **(Arindam Says: I still like the 'data sharing' name – since it has been used in the literature, it will help a reader connect to the existing body of work.)** A data enriched model *enriches* above models by taking their middle ground. It assumes that there is a *common* parameter $\boldsymbol{\beta}_0^*$ shared between all groups and *individual* per-group parameters $\boldsymbol{\beta}_g^*$ which characterize the deviation of group g, i.e.,

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

where g and i index the group and samples respectively. In (1), we have G models coupled by the common parameter $\boldsymbol{\beta}_0^*$ which models similarities between all samples. Individual parameters $\boldsymbol{\beta}_g$s capture unique aspects of each group. We specifically focus on the high-dimensional small sample regime for (1) where the number of samples n_g for each group is much smaller than the ambient dimensionality, i.e.,

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

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

$\forall g : n_g \ll p$. Similar to all other high-dimensional models, we assume that the parameters $\boldsymbol{\beta}_g$ are structured, i.e., for suitable convex functions f_gs, $f_g(\boldsymbol{\beta}_g)$ is small. For example, when the structure is sparsity, f_gs are L_1-norms. Further, for the technical analysis and proofs, we focus on the case of linear models, i.e., $\phi(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}^T\boldsymbol{\beta}$. The results seamlessly extend to more general non-linear models, e.g., generalized linear models [][], broad families of semi-parametric and single-index models [][], non-convex models [][], etc., using existing results, i.e., how models like Lasso have been extended to these settings [][]. **(Arindam Says: Is the last line correct? Lets check this.)**

Note that the models in (1) for each group is a superposition (Gu & Banerjee, 2016)[] or dirty statistical model (Yang & Ravikumar, 2013). **(Arindam Says: cite relevant paper(s) by McCoy-Tropp on convex de-mixing)** Therefore, a *data-enriched* (DE) model is *a system of coupled superposition models*. A related model is proposed by (Jalali et al., 2010) in the context of multi-task learning, where for each task g the output is coming from $y_{gi} = \mathbf{x}_{gi}^T(\boldsymbol{\beta}_{0g}^* + \boldsymbol{\beta}_g^*) + \omega_{gi}$. As emphasized by the subscript of $\boldsymbol{\beta}_{0g}^*$ the common parameters are different in every task but they share a same support (index of non-zero values), i.e., $\boldsymbol{\beta}_{0i}^* \neq \boldsymbol{\beta}_{0j}^*$ but $\text{supp}(\boldsymbol{\beta}_{0i}^*) = \text{supp}(\boldsymbol{\beta}_{0j}^*)$. **(Arindam Says: need a line on what Jalali et al. show, and how our results are (qualitatively) different) (Arindam Says: we don't have a discussion on 'related work' – including hierarchical models, multi-task learning; should we have a sub-section on these, or otherwise discuss these related developments)** At a high level, the DE model can be viewed as a of hierarchical additive model [][], where $\boldsymbol{\beta}_0^*$ corresponds to the root and $\boldsymbol{\beta}_g^*$ correspond to the refinements in the leaves.

The DE model where $\boldsymbol{\beta}_g$s are sparse has recently gained attention because of its application in wide range of domains such as personalized medicine (Dondelinger & Mukherjee, 2016), sentiment analysis, banking strategy (Gross & Tibshirani, 2016), single cell data analysis (Ollier & Viallon, 2015), road safety (Ollier & Viallon, 2014), and disease subtype analysis (Dondelinger & Mukherjee, 2016). More generally, in any high-dimensional problem where the population consists of groups, data enrichment framework has the potential to boost the prediction accuracy and results in a more interpretable set of parameters.

In spite of the recent surge in applying data enrichment framework to different domains, limited advances have been made in understanding the statistical and computational properties of suitable estimators for the data enriched model. In fact, non-asymptotic statistical properties, including sample complexity and statistical rates of convergence, of regularized estimators for the data enriched model is still an open question (Gross & Tibshirani, 2016; Ollier & Viallon, 2014). To the best of our knowledge, the only theoretical guarantee for data enrichment is provided in (Ollier & Viallon, 2015) where authors prove sparsistency of their proposed method under the stringent irrepresentability condition of the design matrix.**(Arindam Says: If they show support recovery, the condition may have been necessary – maybe give them due credit, and show how much more the current results are, though we are doing norm consistency, not support recovery.)** Also beyond sparsity and l_1-norm, no other structure has been investigated for the data enriched model. Moreover, no computational results, such as rates of convergence of the optimization algorithms associated with the proposed estimators, exist in the literature.

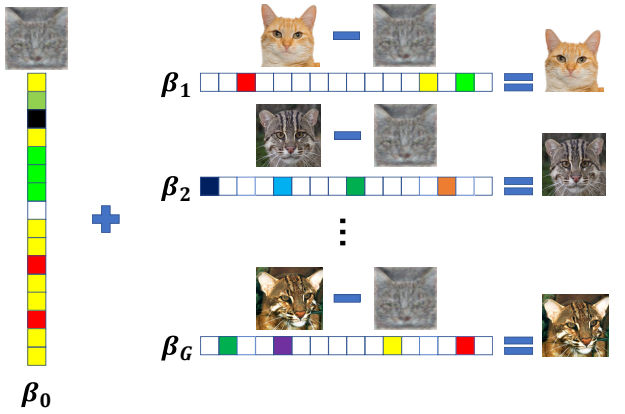


Figure 1. A conceptual illustration of data enrichment model for learning representation of different cat species. The common parameter $\boldsymbol{\beta}_0$ captures a *generic cat* which consists of shared features among all cats.

Notation and Preliminaries: We denote sets by curly $\mathcal{V}$, matrices by bold capital $\mathbf{V}$, random variables by capital V, and vectors by small bold $\mathbf{v}$ letters. We take $[G] = \{0, \ldots, G\}$ and $[G]_\setminus = [G]\setminus\{0\}$. **(Arindam Says: Can we use $[G] = \{1, \ldots, G\}$ and use $G \cup \{0\}$ in places where we need to include $\{0\}$? The notation will be more explicit and easier to follow.)**

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

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

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

A random variable V is sub-Gaussian if its moments satisfies $\forall p \geq 1 : (\mathbb{E}|V|^p)^{1/p} \leq K_2\sqrt{p}$. The minimum value of K_2 is called the sub-Gaussian norm of V, denoted by $\|\|V\|\|_{\psi_2}$ (Vershynin, 2012). A random vector

$\mathbf{v} \in \mathbb{R}^p$ is sub-Gaussian if the one-dimensional marginals $\langle \mathbf{v}, \mathbf{u} \rangle$ are sub-Gaussian random variables for all $\mathbf{u} \in \mathbb{R}^p$. The sub-Gaussian norm of $\mathbf{v}$ is defined (Vershynin, 2012) as $\|\mathbf{v}\|_{\psi_2} = \sup_{\mathbf{u} \in \mathbb{S}^{p-1}} \|\langle \mathbf{v}, \mathbf{u} \rangle\|_{\psi_2}$. For any set $\mathcal{V} \in \mathbb{R}^p$ the Gaussian width of the set $\mathcal{V}$ is defined as $\omega(\mathcal{V}) = \mathbb{E}_{\mathbf{g}} [\sup_{\mathbf{u} \in \mathcal{V}} \langle \mathbf{g}, \mathbf{u} \rangle]$ (Chandrasekaran et al., 2012), **(Arindam Says: cite Vershynin's 2019 book)** where the expectation is over $\mathbf{g} \sim N(\mathbf{0}, \mathbf{I}_{p \times p})$, a vector of independent zero-mean unit-variance Gaussian.

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

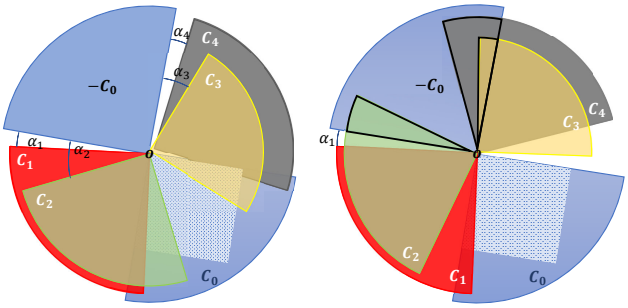
$$\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_0^T, \ldots, \hat{\boldsymbol{\beta}}_G^T) \in \underset{\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_G}{\operatorname{argmin}} \frac{1}{n} \sum_{g=1}^{G} \|\mathbf{y}_g - \mathbf{X}_g(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_g)\|_2^2, \quad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

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

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

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

  where n_g is number of samples per group, $n_0 \triangleq n$ is the total number of samples, $\gamma \triangleq \max_{g \in [G]} \frac{n}{n_g}$ is the *sample condition number*, and $\mathcal{C}_g$ is the error cone corresponding to $\boldsymbol{\beta}_g^*$ exactly defined in Section 2.1. To the best of our knowledge, this is the first statistical estimation guarantee for the data enriched model.
- We also establish the sample complexity of the DE estimator for all parameters as $\forall g \in [G] : n_g = O(\omega(\mathcal{C}_g \cap \mathbb{S}^{p-1}))^2$. We emphasize that our result proofs that the recovery of the common parameter $\boldsymbol{\beta}_0$ by DE estimator benefits from *all* of the n pooled samples.
- We present an efficient Projected Block Gradient Descent (PBGD) algorithm to solve DE's objective (3) which converges geometrically to the statistical error bound of (4). To the best of our knowledge, this is the first rigorous computational result for the high-dimensional data-enriched regression.



(a) Structural Coherence (SC) condition. (b) Data EnRichment Incoherence Condition (DERIC).

Figure 2. a) State of the art condition for recovering common and individual parameters in superposition models where $\mathcal{C}_g = \text{Cone}(\mathcal{E}_g)$ are error cones and $\mathcal{E}_g = \{\boldsymbol{\delta}_g | f_g(\boldsymbol{\beta}_g^* + \boldsymbol{\delta}_g) \leq f_g(\boldsymbol{\beta}_g^*)\}$ are the error sets for each parameter $\boldsymbol{\beta}_g^* \in [G]$ (Gu & Banerjee, 2016) b) Our more relaxed recovery condition which allows *arbitrary non-zero fraction* of the error cones of individual parameters intersect with $-\mathcal{C}_0$.

- We illustrate promising empirical performance of the model on synthetic data as well as on the problem of finding bio-markers associated with drug sensitivity of cell lines from different cancer types, where the support of individual parameters $\hat{\boldsymbol{\beta}}_g^T$ for each cancer g shows a different set of bio-markers per cancer type.

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 sample complexity required for the recovery of the true parameters by our estimator under DERIC condition. We close the statistical analysis in Section 4 by providing non-asymptotic high probability error bound for parameter recovery. 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

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

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

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

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

(Arindam Says: Give at least 2 specific examples, e.g., a sparse+sparse for vector estimation, and a low-rank + sparse for matrix estimation. And we want to carry these examples through the technical results, e.g., see the Chen-Banerjee NIPS 2015 paper.)

Example 1. This is an example

2.1. Error Set and Deterministic Error Bound

(Arindam Says: Unless we need a second subsection, drop this subsection heading – its unusual to have one subsection in a section.)

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

$$\mathcal{E}_g = \left\{\boldsymbol{\delta}_g | f_g(\boldsymbol{\beta}^*_g + \boldsymbol{\delta}_g) \leq f_g(\boldsymbol{\beta}^*_g)\right\}, \quad g \in [G]\,. \tag{7}$$

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

$$\mathcal{H} \triangleq \left\{\boldsymbol{\delta} \in \mathcal{C} \Big| \sum_{g=0}^{G} \frac{n_g}{n} \|\boldsymbol{\delta}_g\|_2 = 1\right\}, \tag{8}$$

$$\bar{\mathcal{H}} \triangleq \left\{\boldsymbol{\delta} \in \mathcal{C} \Big| \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g\|_2 = 1\right\}. \tag{9}$$

Starting from the optimality of $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + \boldsymbol{\delta}$ as $\frac{1}{n}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2 \leq \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2$, we have: $\frac{1}{n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 \leq \frac{1}{n} 2\boldsymbol{\omega}^T\mathbf{X}\boldsymbol{\delta}$ where $\boldsymbol{\omega} = [\boldsymbol{\omega}_1^T, \ldots, \boldsymbol{\omega}_G^T]^T \in \mathbb{R}^n$ is the vector of all noises. Using this basic inequality, we can establish the following deterministic error bound.

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

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

(Arindam Says: Adding a remark – these will make it easier for the reader. This specific remark can be dropped as needed.)

Remark 1. Consider the setting where $n_g = \Theta(\frac{n}{G})$ so that each group has approximately $\frac{1}{G}$ fraction of the samples. Then, $\gamma = \Theta(G)$ and hence

$$\frac{1}{G}\sum_{g=0}^{G} \|\delta_g\|_2 \leq O(G^{1/2}) \frac{\sup_{\mathbf{u}\in\bar{\mathcal{H}}} \boldsymbol{\omega}^T\mathbf{X}\mathbf{u}}{n}\,. \tag{10}$$

3. Restricted Eigenvalue Condition

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

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

(Arindam Says: Minor comment – we are using variants of k for different things, e.g., k, K, κ. While the notation is consistent, this can be a bit hard to follow. If possible, make the ψ_2 norms c, C.)

Unlike standard high-dimensional statistical estimation, for RE condition to be true, parameters of the data enriched model (2) needs to satisfy a geometric condition under which trivial solutions such as $\boldsymbol{\delta}_g = -\boldsymbol{\delta}_0$ for all $g \in [G]_\setminus$ are avoided. To derive this condition, first note that each of the linear models in (2) is a superposition model (Gu & Banerjee, 2016) or dirty statistical model (Yang & Ravikumar, 2013). RE condition of individual superposition models can be established under the so-called Structural Coherence (SC) condition (Gu & Banerjee, 2016; McCoy & Tropp, 2013). However, SC condition on each individual problem fails to utilize the true coupling structure in the data enriched model, where $\boldsymbol{\beta}_0$ is involved in all models. In fact, as we show shortly, using SC on each individual model leads to radically pessimistic estimates of the sample complexity.

Here, we introduce DERIC, a considerably weaker geometric condition compared to SC of (Gu & Banerjee, 2016; McCoy & Tropp, 2013). 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

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

fraction of the $\mathcal{C}_g$ cones. As the number of intersections increases, our bound becomes looser. The rigorous definition of DERIC is provided below.

Definition 2 (Data EnRichment Incoherence Condition (DERIC))**.** *There exists a set $\mathcal{I} \subseteq [G]_\setminus$ 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. $\forall i \in \mathcal{I}$, $\forall \boldsymbol{\delta}_i \in \mathcal{C}_i$, *and* $\boldsymbol{\delta}_0 \in \mathcal{C}_0$*:* $\|\boldsymbol{\delta}_i + \boldsymbol{\delta}_0\|_2 \geq \lambda_{\min}(\|\boldsymbol{\delta}_0\|_2 + \|\boldsymbol{\delta}_i\|_2)$

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

(Arindam Says: why would $\bar{\rho} = 0$ work? also, can $\mathcal{I}$ be empty? pls update as needed)

(Arindam Says: add a remark – so the reader can follow what we are saying)

(Arindam Says: It will be great to have a Figure showing the difference between SC and DERIC.)

In contrast, the existing SC condition (Gu & Banerjee, 2016; McCoy & Tropp, 2013) applied to each superposition model (2) separately requires for $\boldsymbol{\delta}_0 \in \mathcal{C}_0$ and *each* $\boldsymbol{\delta}_g \in \mathcal{C}_g$ there exist $\lambda > 0$ such that: $\|\boldsymbol{\delta}_0 + \boldsymbol{\delta}_g\|_2 \geq \lambda\,(\|\boldsymbol{\delta}_0\|_2 + \|\boldsymbol{\delta}_g\|_2)$. Clearly DERIC 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., $\sup\langle \boldsymbol{\delta}_0/\|\boldsymbol{\delta}_0\|_2, \boldsymbol{\delta}_g/\|\boldsymbol{\delta}_g\|_2\rangle < 1$ (Tropp, 2015; Gu & Banerjee, 2016). However, DERIC 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 (Gu & Banerjee, 2016) will lead to a considerably pessimistic sample complexity.

Proposition 1. *Assume observations distributed as defined in Definition 1 and pair-wise SC conditions are satisfied. Consider each superposition model* (2) *in isolation; to recover the common parameter $\boldsymbol{\beta}_0^*$ requires at least one group to have $n_g = O(\omega^2(\mathcal{A}_0))$. Recovering the individual parameter $\boldsymbol{\beta}_g^*$ needs at least $n_g = O((\max_{g\in[G]} \omega(\mathcal{A}_g) + \sqrt{\log 2})^2)$ samples in the group.*

(Arindam Says: Not sure if I follow this. Assume $g = 1$ has sufficient samples, i.e., $n_1 \geq c\omega^2(\mathcal{A}_0)$. Then, the estimator would proceed in two steps: first, estimate β_0, β_1 from the first problem; then, use the estimated β_0 in the other problems to estimate $\beta_2, \ldots, \beta_G$. With such an estimator, we need $n_g \geq c\omega^2(\mathcal{A}_g), g = 2, \ldots, G$. So only one problem needs to have sufficient samples.)

In other words, by separate analysis of superposition estimators neither the estimation of the common parameter $\boldsymbol{\beta}_0$ nor the individual parameters $\boldsymbol{\beta}_g$ 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 $\boldsymbol{\beta}_0$. Using DERIC and the small ball method (Mendelson, 2014), a recent tool from empirical process theory, we get a better sample complexity for satisfying the RE condition.

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

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

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

4. General Error Bound

In this section, we provide a high probability upper bound for the estimation error of the common and individual parameters.To avoid cluttering the notation, we rename the vector of all noises as $\boldsymbol{\omega}_0 \triangleq \boldsymbol{\omega}$.

(Arindam Says: We are using ω to denote two different things—the Gaussian width and the noise!)

First, we massage the deterministic upper bound of Theorem 1 as follows:

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

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

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

where the inequality holds because of the definition of the dual norm. Next, using the following lemma, we upper bounds b_g with high probability.

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

$$\begin{aligned}&\sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2 \sup_{\mathbf{u}_g\in\mathcal{A}_g}\langle \mathbf{X}_g^T\frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \mathbf{u}_g\rangle \qquad (11)\\ \leq\quad &\sqrt{(2K^2+1)n}\left(\zeta_g k\omega(\mathcal{A}_g) + \epsilon_g\sqrt{\log(G+1)} + \tau\right),\end{aligned}$$

Algorithm 1 PBGD: PROJECTED BLOCK GRADIENT DESCENT

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

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

Using Lemma 1 the below theorem establishes a high probability upper bound for the deterministic bound of Theorem 1, i.e., $\frac{2}{n}\boldsymbol{\omega}^T\mathbf{X}\mathbf{u}$.

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

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

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 1, Theorem 2, and Theorem 3.

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

$$1\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g\|_2 \leq C\gamma\frac{k\zeta\max_{g\in[G]}\omega(\mathcal{A}_g) + \epsilon\sqrt{\log(G+1)} + \tau}{\kappa_{\min}^2\sqrt{n}} \quad (12)$$

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

(Arindam Says: This section will be very hard to follow without examples. Lets use the two examples we (will) introduce earlier.)

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 cluttering. In practice, d_g can be determined by cross-validation.

5.1. Convergence Rate Analysis

(Arindam Says: Unless we need a second subsection, drop this subsection heading – its unusual to have one subsection in a section.)

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

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

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

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

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

Theorem 4. *For Algorithm 1 initialized by $\boldsymbol{\beta}^{(1)} = \mathbf{0}$, we have the following deterministic bound for the error at iteration $t+1$:*

$$\begin{aligned}
&\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g^{(t+1)}\|_2 \qquad (13)\\
&\leq \rho^t\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\beta}_g^*\|_2 + \frac{1-\rho^t}{1-\rho}\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\eta_g\|\boldsymbol{\omega}_g\|_2,
\end{aligned}$$

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

The RHS of (13) 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_gs, the second term can be upper bounded using the analysis of Section 4. More specifically, the first term corresponds to the optimization error which shrinks in every iteration while the second term is constant times the upper bound of the statistical error characterized in Corollary 1. Therefore, if we keep ρ below one, the estimation error of PBGD algorithm linearly converges to the approximate statistical error bound.

One way for having $\rho < 1$ is to keep all arguments of $\max(\cdots)$ defining ρ strictly below 1. To this end, we first

establish high probability upper bound for ρ_g, η_g, and ϕ_g (in the Appendix) 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 3 and the deterministic bound of Theorem 4 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. *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:*

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

and sample complexities of $\forall g \in [G] : n_g \geq 2c_g^2(2\omega(\mathcal{A}_g) + \tau)^2$, updates of the Algorithm 1 obey the following with high probability:

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

where $r(\tau) < 1$.

Corollary 2. *When $t \to \infty$ we have the following with high probability:*

$$\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g^{\infty}\|_2 \leq \frac{(G+1)\sqrt{(2K^2+1)}}{\sqrt{n}(1-r(\tau))} \quad (14)$$
$$\times \quad \left(\zeta k \max_{g\in[G]}\omega(\mathcal{A}_g) + C\sqrt{\log(G+1)} + b\right),$$

It is instructive to compare RHS of (14) with that of (12): $\kappa_{\min}$ defined in Theorem 2 corresponds to $(1 - r(\tau))$ and the extra $G + 1$ factor corresponds to the sample condition number $\gamma = \max_{g\in[G]} \frac{n}{n_g}$. Therefore, Corollary 2 shows that PBGD converges to a scaled variant of statical error bound determined in Corollary 1.

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 individual parameters to be $s = 10$. We generated a dense $\boldsymbol{\beta}_0$ with $\|\boldsymbol{\beta}_0\| = p$ and did not impose any constraint. Iterates $\{\boldsymbol{\beta}_g^{(t)}\}_{g=1}^G$ are obtained by projection onto the ℓ_1 ball $\|\boldsymbol{\beta}_g\|_1$. Nonzero entries of $\boldsymbol{\beta}_g$ 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 $\boldsymbol{\beta}_0$ and $\frac{1}{\sqrt{nn_g}}$ for $\boldsymbol{\beta}_g$, $g \in [G]_\setminus$. Observe that, cones of the individual parameters intersect with that of $\boldsymbol{\beta}_0$ hence this setup actually violates DERIC (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 $\boldsymbol{\beta}_g$'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 $\frac{\|\boldsymbol{\beta}_g^{(t)} - \boldsymbol{\beta}_g\|_2^2}{\|\boldsymbol{\beta}_g\|_2^2}$ and average 10 independent realization for each curve. Figure 3a 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 3b, 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 $\boldsymbol{\beta}_0$ and noise-free groups $\{\boldsymbol{\beta}_g\}_{g=2}^G$ have substantially less estimation error. This implies that noise in one group mostly affects itself rather than the global estimation. In Figure 3c, we increased the sample size to $n_g = 150$ per group. We observe that, in comparison to Figure 3a, rate of convergence receives a boost from the additional samples as predicted by our theory.

Finally, Figure 3d considers a very high-dimensional problem where $p = 1000$, $G = 100$, individual parameters are 10 sparse, $\boldsymbol{\beta}_0$ 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. Anti-Cancer Drug Sensitivity Prediction

In this section, we investigate the application of DE 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 DE 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 area" above the dose-response curve, Figure 4a. Given $\mathbf{x}_{gi}$ and a drug, we have two goals: accurately predict a patient's response to the drug and identifying genetic predictors of drug sensitivity. We present more details about and results of the drug

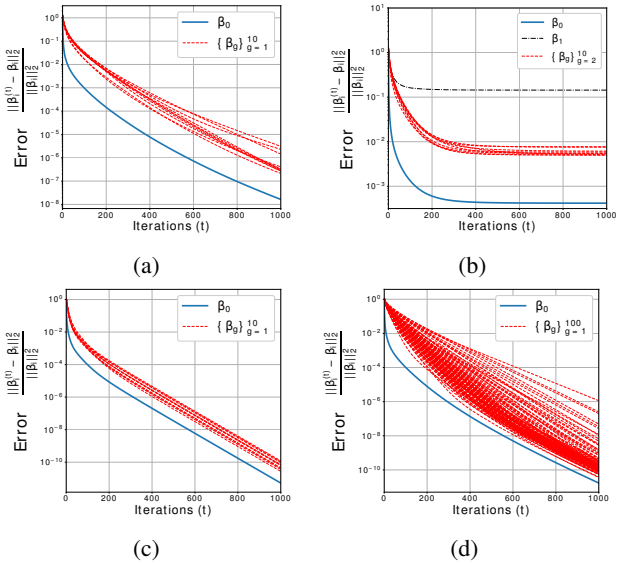


Figure 3. a) Noiseless fast convergence. b) Noise on the first group does not impact other groups as much. c) Increasing sample size improves rate of convergence. d) Our algorithm convergences fast even with a large number of groups $G = 100$.

sensitivity experiment in Appendix **??**.

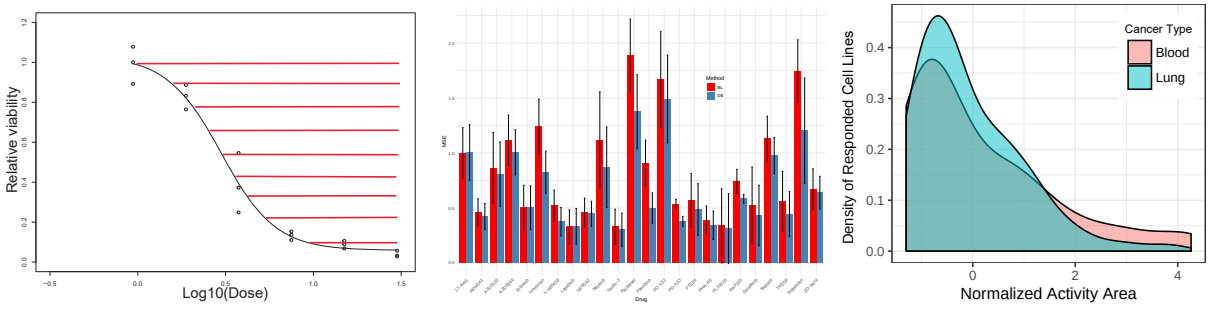
We use Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012) which is a compilation ∼500 human cancer cell lines where their responses to 24 anticancer drugs have been measured.We perform two *experiments* where the number of cancers in each data set are $G = 2$ or 3 and we name them TWO and THREE experiments, respectively. We consider lung and blood[2] for TWO while for THREE we predict the drug sensitivity of skin, breast and ovary cancer cell lines. Beyond these five cancer types, others have less than 50 samples, so we remove them from consideration. Each experiment consists of 24 *problems* each corresponds to a drug. Not all of the 500 cell lines have been treated with all of the drugs. Therefore each problem has a different number of samples n where $n \in [70, 130]$ for TWO and $n \in [?]$ for THREE experiments. We perform a standard preprocessing (Barretina et al., 2012) where we remove features with less than .2 absolute correlation with the response of interest.Note that the features that get removed vary by problem, therefore the dimension p is reduced from from $> 30,000$ to $p \in [1000, 15000]$.

Prediction: In each TWO and THREE experiments, we predict the drug sensitivity for 24 different drugs. Since the values of d_g in constraint sets $\Omega_{f_g}(d_g)$ are unknown, we tune them by 5-fold cross-validation and report the mean squared error (MSE) of DE and a baseline method. Our *baseline* method BL is the LASSO (Tibshirani, 1996) equivalent of DE where we set $\forall g \in [G]_{\setminus} d_g = 0$ and only estimate the common parameter $\boldsymbol{\beta}_0$. Figure 4b and 4b illustrates the performance of DE and BL for TWO and THREE experiments respectively. Note that DE outperforms BL in ? and ? out of 24 problems in TWO and THREE experiments, respectively.

To ensure that the prediction improvement of DE over the baseline is significant, we supplement our analysis with the bootstrapped error of both methods for the TWO experiment. For each problem in the TWO experiment, we generate 100 bootstrapped data sets by sampling with replacement as $\{(\mathbf{X}^{(i)}_{\text{TWO}}, \mathbf{y}^{(i)}_{\text{TWO}})\}_{i=1}^{100}$. Then, we fix d_gs hyperparameters to values determined by cross-validation in the last stage and run both methods and compute pairs of MSEs as $\{(\text{MSE}^{(i)}_{\text{DE}}, \text{MSE}^{(i)}_{\text{BL}})\}$. We perform paired t-test to determine if mean difference between MSEs is significant. In ? out of 24 problems DE's MSE is lesser than BL's with significance level of $\alpha = 0.05$. A representative set of results is demonstrated in Figure **??**.

Interpretation We select Saracatinib, a drug which shows activity on both lung and blood cancers, Figure 4c. 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 (Chen et al., 2009) 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.

[2]By blood cancer, we mean any cancer originate from haematopoietic and lymphoid tissues.



(a) (b) (c)

Figure 4. 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. Note that some cell lines of both lung and blood cancers have responded to Saracatinib which makes it a good candidate for interpretability analysis.

(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 (Chen et al., 2009).

References

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

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603, 2012.

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

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

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. Toppgene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*, 37 (suppl_2):W305–W311, 2009.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Yang, E. and Ravikumar, P. Dirty statistical models. In *Advances in Neural Information Processing Systems*, pp. 611–619, 2013.

A. Proofs

A.1. Proof of Theorem 1

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

$$\begin{aligned}\frac{1}{n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 &\geq \frac{1}{n}\inf_{\mathbf{u}\in\mathcal{H}}\|\mathbf{X}\mathbf{u}\|_2^2\left(\sum_{g=0}^{G}\frac{n_g}{n}\|\boldsymbol{\delta}_g\|_2\right)^2 \\ &\geq \kappa\left(\sum_{g=0}^{G}\frac{n_g}{n}\|\boldsymbol{\delta}_g\|_2\right)^2 \\ &\geq \kappa\left(\min_{g\in[G]}\frac{n_g}{n}\right)\left(\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g\|_2\right)^2\end{aligned} \tag{15}$$

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

$$\frac{2}{n}\boldsymbol{\omega}^T\mathbf{X}\boldsymbol{\delta} \leq \frac{2}{n}\sup_{\mathbf{u}\in\bar{\mathcal{H}}}\boldsymbol{\omega}^T\mathbf{X}\mathbf{u}\left(\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g\|_2\right), \quad \mathbf{u}\in\mathcal{H} \tag{16}$$

Putting together inequalities (15) and (16) completes the proof. ■

A.2. Proof of Proposition 1

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

A.3. Proof of Theorem 2

Let's simplify the LHS of the RE condition:

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

where the first inequality is due to Lyapunov's inequality. To avoid cluttering we denote $\boldsymbol{\delta}_{0g}=\boldsymbol{\delta}_0+\boldsymbol{\delta}_g$ where $\boldsymbol{\delta}_0\in\mathcal{C}_0$ and $\boldsymbol{\delta}_g\in\mathcal{C}_g$. Now we add and subtract the corresponding per-group marginal tail function, $Q_{\xi_g}(\boldsymbol{\delta}_{0g})=\mathbb{P}(|\langle\mathbf{x},,\boldsymbol{\delta}_{0g}\rangle|>\xi_g)$ where $\xi_g>0$. Let $\xi_g=\|\boldsymbol{\delta}_{0g}\|_2\xi$ then the LHS of the RE condition reduces to:

$$\begin{aligned}\inf_{\boldsymbol{\delta}\in\mathcal{H}}\frac{1}{\sqrt{n}}\|\mathbf{X}\boldsymbol{\delta}\|_2 &\geq \inf_{\boldsymbol{\delta}\in\mathcal{H}}\sum_{g=1}^{G}\frac{n_g}{n}\xi_g Q_{2\xi_g}(\boldsymbol{\delta}_{0g}) \\ &- \sup_{\boldsymbol{\delta}\in\mathcal{H}}\frac{1}{n}\sum_{g=1}^{G}\xi_g\sum_{i=1}^{n_g}\left[Q_{2\xi_g}(\boldsymbol{\delta}_{0g})-\mathbb{1}(|\langle\mathbf{x}_{gi},\boldsymbol{\delta}_{0g}\rangle|\geq\xi_g)\right] \\ &= t_1(\mathbf{X})-t_2(\mathbf{X})\end{aligned} \tag{17}$$

For the ease of exposition we have written the LHS of (17) as the difference of two terms, i.e., $t_1(\mathbf{X})-t_2(\mathbf{X})$ and in the followings we lower bound the first term t_1 and upper bound the second term t_2.

A.3.1. Lower Bounding the First Term

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

Lemma 2. *Suppose DERIC holds. Let $\psi_{\mathcal{I}} = \frac{\lambda_{\min}\bar{\rho}}{3}$. For any $\boldsymbol{\delta} \in \mathcal{H}$, we have:*

$$\sum_{g=1}^{G} \frac{n_g}{n} \xi_g Q_{2\xi_g}(\boldsymbol{\delta}_{0g}) \geq \psi_{\mathcal{I}} \xi \frac{(\alpha - 2\xi)^2}{4ck^2} \left(\|\boldsymbol{\delta}_0\|_2 + \sum_{g=1}^{n} \frac{n_g}{n} \|\boldsymbol{\delta}_g\|_2 \right), \tag{18}$$

which implies that $t_1(\mathbf{X}) = \inf_{\boldsymbol{\delta} \in \mathcal{H}} \sum_{g=1}^{G} \frac{n_G}{n} \xi_g Q_{2\xi_g}(\boldsymbol{\delta}_{0g})$ satisfies the same RHS bound of (18)*.*

Proof. LHS of (18) is the weighted summation of $\xi_g Q_{2\xi_g}(\boldsymbol{\delta}_{0g}) = \|\boldsymbol{\delta}_{0g}\|_2 \xi \mathbb{P}(|\langle \mathbf{x}, , \boldsymbol{\delta}_{0g}/\|\boldsymbol{\delta}_{0g}\|_2 \rangle| > 2\xi) = \|\boldsymbol{\delta}_{0g}\|_2 \xi Q_{2\xi}(\mathbf{u})$ where $\xi > 0$ and $\mathbf{u} = \boldsymbol{\delta}_{0g}/\|\boldsymbol{\delta}_{0g}\|_2$ is a unit length vector. So we can rewrite the LHS of (18) as:

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

With this observation, the lower bound of the Lemma 2 is a direct consequence of the following two results:

Lemma 3. *Let $\mathbf{u}$ be any unit length vector and suppose $\mathbf{x}$ obeys Definiton 1. Then for any $\mathbf{u}$, we have*

$$Q_{2\xi}(\mathbf{u}) \geq \frac{(\alpha - 2\xi)^2}{4ck^2}. \tag{19}$$

Lemma 4. *Suppose Definition 2 holds. Then, we have:*

$$\sum_{i=1}^{G} n_i \|\boldsymbol{\delta}_0 + \boldsymbol{\delta}_i\|_2 \geq \frac{\bar{\rho}\lambda_{\min}}{3} \left(Gn\|\boldsymbol{\delta}_0\|_2 + \sum_{i=1}^{G} n_i \|\boldsymbol{\delta}_i\|_2 \right), \quad \forall i \in [G] : \boldsymbol{\delta}_i \in \mathcal{C}_i. \tag{20}$$

■

A.3.2. Upper Bounding the Second Term

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

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

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

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

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

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

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 (**?**)Theorem 6.2]boucheron13 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 5. *For the random vector* $\mathbf{x}$ *of Definition 1, we have the following bound:*

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

A.3.3. Continuing the Proof of Theorem 2

Set $n_0 = n$. Putting back bounds of $t_1(\mathbf{X})$ and $t_2(\mathbf{X})$ together from Lemma 2 and 5, with probability at least $1 - e^{-\frac{\tau^2}{2}}$ we have:

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

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

$$\forall g \in [G]: \quad \left(\frac{2ck}{\psi_{\mathcal{I}}\xi q}\right)^2 \omega(\mathcal{A}_g)^2 \quad \leq \quad n_g \tag{21}$$

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

$$\forall g \in [G]: \quad \left(\frac{Ck^3}{\psi_{\mathcal{I}}\alpha^3}\right)^2 \omega(\mathcal{A}_g)^2 \quad \leq \quad n_g \tag{22}$$

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

A.4. Proof of Lemma 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 $s_g = \sqrt{\frac{n}{n_g}}\sqrt{(2K^2+1)n_g}$.

$$\begin{aligned}
\mathbb{P}\left(h_g(\boldsymbol{\omega}_g, \mathbf{X}_g) > e_g s_g\right) &= \mathbb{P}\left(h_g(\boldsymbol{\omega}_g, \mathbf{X}_g) > e_g s_g \middle| \sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2 > s_g\right)\mathbb{P}\left(\sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2 > s_g\right) \\
&+ \mathbb{P}\left(h_g(\boldsymbol{\omega}_g, \mathbf{X}_g) > e_g s_g \middle| \sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2 < s_g\right)\mathbb{P}\left(\sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2 < s_g\right) \\
&\leq \mathbb{P}\left(\sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2 > s_g\right) + \mathbb{P}\left(h_g(\boldsymbol{\omega}_g, \mathbf{X}_g) > e_g s_g \middle| \sqrt{\frac{n}{n_g}}\|\boldsymbol{\omega}_g\|_2 < s_g\right) \\
&\leq \mathbb{P}\left(\|\boldsymbol{\omega}_g\|_2 > \sqrt{(2K^2+1)n_g}\right) + \mathbb{P}\left(\sup_{\mathbf{u}_g\in\mathcal{C}_g\cap\mathbb{S}^{p-1}}\langle \mathbf{X}_g^T\frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \mathbf{u}_g\rangle > e_g\right) \\
&\leq \mathbb{P}\left(\|\boldsymbol{\omega}_g\|_2 > \sqrt{(2K^2+1)n_g}\right) + \sup_{\mathbf{v}\in\mathbb{S}^{p-1}}\mathbb{P}\left(\sup_{\mathbf{u}_g\in\mathcal{C}_g\cap\mathbb{S}^{p-1}}\langle \mathbf{X}_g^T\mathbf{v}, \mathbf{u}_g\rangle > e_g\right)
\end{aligned} \tag{23}$$

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

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

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

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

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

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

Now we upper bound the second term of (23). Given any fixed $\mathbf{v} \in \mathbb{S}^{p-1}$, $\mathbf{X}_g\mathbf{v}$ is a sub-Gaussian random vector with $\left\|\mathbf{X}_g^T\mathbf{v}\right\|_{\psi_2} \leq C_g k$ (Banerjee et al., 2014). From Theorem 9 of (Banerjee et al., 2014) for any $\mathbf{v} \in \mathbb{S}^{p-1}$ we have:

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

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

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

Now we put back results to the original inequality (23):

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

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

A.5. Proof of Theorem 3

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

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

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

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

Now we can go back to the original form:

$$\begin{aligned}
\sup_{\boldsymbol{\delta} \in \mathcal{H}} \boldsymbol{\omega}^T \mathbf{X} \boldsymbol{\delta} &\leq \max_{g \in [G]} \langle \mathbf{X}_g^T \frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \frac{\boldsymbol{\delta}_g}{\|\boldsymbol{\delta}_g\|_2} \rangle \sqrt{\frac{n}{n_g}} \|\boldsymbol{\omega}_g\|_2 \\
&\leq \max_{g \in [G]} \sqrt{\frac{n}{n_g}} \|\boldsymbol{\omega}_g\|_2 \sup_{\mathbf{u}_g \in \mathcal{C}_g \cap \mathbb{S}^{p-1}} \langle \mathbf{X}_g^T \frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \mathbf{u}_g \rangle
\end{aligned} \tag{24}$$

To avoid cluttering we name $h_g(\boldsymbol{\omega}_g, \mathbf{X}_g) = \|\boldsymbol{\omega}_g\|_2 \sup_{\mathbf{u}_g \in \mathcal{A}_g} \langle \mathbf{X}_g^T \frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2}, \mathbf{u}_g \rangle$ and $e_g(\tau) = \sqrt{(2K^2+1)n_g} \left(\upsilon_g C_g k \omega(\mathcal{A}_g) + \epsilon_g \sqrt{\log G} + \tau \right)$. Then from (24), 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:

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

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

A.6. Proof of Lemma 6

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

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

So the final bound becomes:

$$\|\boldsymbol{\delta}_g^{(t+1)}\|_2 \leq \rho_g(\mu_g)\|\boldsymbol{\delta}_g^{(t)}\|_2 + \xi_g(\mu_g)\|\boldsymbol{\omega}_g\|_2 + \phi_g(\mu_g)\|\boldsymbol{\delta}_0^{(t)}\|_2 \tag{25}$$

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

$$
\begin{aligned}
\|\boldsymbol{\delta}_0^{(t+1)}\|_2 &= \|\boldsymbol{\beta}_0^{(t+1)} - \boldsymbol{\beta}_0^*\|_2 \\
&= \left\| \Pi_{\Omega_{f_0}}\left(\boldsymbol{\beta}_0^{(t)} + \mu_0 \sum_{g=1}^{G} \mathbf{X}_g^T\left(\mathbf{y}_g - \mathbf{X}_g(\boldsymbol{\beta}_0^{(t)} + \boldsymbol{\beta}_g^{(t)})\right)\right) - \boldsymbol{\beta}_0^* \right\|_2 \\
\text{(Lemma 6.3 of (Oymak et al., 2015))} &= \left\| \Pi_{\Omega_{f_0} - \{\boldsymbol{\beta}_0^*\}}\left(\boldsymbol{\beta}_0^{(t)} + \mu_0 \sum_{g=1}^{G} \mathbf{X}_g^T\left(\mathbf{y}_g - \mathbf{X}_g(\boldsymbol{\beta}_0^{(t)} + \boldsymbol{\beta}_g^{(t)})\right) - \boldsymbol{\beta}_0^*\right) \right\|_2 \\
&= \left\| \Pi_{\mathcal{E}_0}\left(\boldsymbol{\delta}_0^{(t)} + \mu_0 \sum_{g=1}^{G} \mathbf{X}_g^T\left(\mathbf{y}_g - \mathbf{X}_g(\boldsymbol{\beta}_0^{(t)} + \boldsymbol{\beta}_g^{(t)})\right)\right) \right\|_2 \\
\text{(Lemma 6.4 of (Oymak et al., 2015))} &\leq \left\| \Pi_{\mathcal{C}_0}\left(\boldsymbol{\delta}_0^{(t)} + \mu_0 \sum_{g=1}^{G} \mathbf{X}_g^T\left(\boldsymbol{\omega}_g - \mathbf{X}_g(\boldsymbol{\delta}_0^{(t)} + \boldsymbol{\delta}_g^{(t)})\right)\right) \right\|_2 \\
\text{(Lemma 6.2 of (Oymak et al., 2015))} &\leq \sup_{\mathbf{v}\in\mathcal{B}_0} \mathbf{v}^T\left(\boldsymbol{\delta}_0^{(t)} + \mu_0 \sum_{g=1}^{G} \mathbf{X}_g^T\left(\boldsymbol{\omega}_g - \mathbf{X}_g(\boldsymbol{\delta}_0^{(t)} + \boldsymbol{\delta}_g^{(t)})\right)\right) \\
&\leq \sup_{\mathbf{v}\in\mathcal{B}_0} \mathbf{v}^T\big(\mathbf{I} - \mu_0 \sum_{g=1}^{G} \mathbf{X}_g^T\mathbf{X}_g\big)\boldsymbol{\delta}_0^{(t)} + \mu_0 \sup_{\mathbf{v}\in\mathcal{B}_0} \mathbf{v}^T \sum_{g=1}^{G} \mathbf{X}_g^T\boldsymbol{\omega}_g \\
&+ \mu_0 \sup_{\mathbf{v}\in\mathcal{B}_0} -\mathbf{v}^T \sum_{g=1}^{G} \mathbf{X}_g^T\mathbf{X}_g\boldsymbol{\delta}_g^{(t)} \\
&\leq \|\boldsymbol{\delta}_0^{(t)}\|_2 \sup_{\mathbf{u},\mathbf{v}\in\mathcal{B}_0} \mathbf{v}^T\big(\mathbf{I} - \mu_0\mathbf{X}_0^T\mathbf{X}_0\big)\mathbf{u} + \mu_0 \sup_{\mathbf{v}\in\mathcal{B}_0} \mathbf{v}^T\mathbf{X}_0^T \frac{\boldsymbol{\omega}_0}{\|\boldsymbol{\omega}_0\|_2}\|\boldsymbol{\omega}_0\|_2 \\
&+ \mu_0 \sum_{g=1}^{G} \sup_{\mathbf{v}_g\in\mathcal{B}_0, \mathbf{u}_g\in\mathcal{B}_g} -\mathbf{v}_g^T\mathbf{X}_g^T\mathbf{X}_g\mathbf{u}_g\|\boldsymbol{\delta}_g^{(t)}\|_2 \\
&\leq \rho_0(\mu_0)\|\boldsymbol{\delta}_0^{(t)}\|_2 + \xi_0(\mu_0)\|\boldsymbol{\omega}_0\|_2 + \mu_0 \sum_{g=1}^{G} \frac{\phi_g(\mu_g)}{\mu_g}\|\boldsymbol{\delta}_g^{(t)}\|_2 \qquad (26)
\end{aligned}
$$

To avoid cluttering we drop μ_g as the arguments. Putting together (25) and (26) 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 4

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

Lemma 6. *We have the following recursive dependency between the error of $t+1$th iteration and tth iteration of PBGD:*

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

By recursively applying the result of Lemma 6, we get the following deterministic bound which depends on constants defined in Definition 3:

$$\begin{aligned}b_{t+1} = \sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g^{(t+1)}\|_2 &\leq \left(\rho_0 + \sum_{g=1}^{G}\sqrt{\frac{n_g}{n}}\phi_g\right)\|\boldsymbol{\delta}_0^{(t)}\|_2 + \sum_{g=1}^{G}\left(\sqrt{\frac{n_g}{n}}\rho_g + \mu_0\frac{\phi_g}{\mu_g}\right)\|\boldsymbol{\delta}_g^{(t)}\|_2 + \sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\xi_g\|\boldsymbol{\omega}_g\|_2\\ &\leq \rho\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g^{(t)}\|_2 + \sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\xi_g\|\boldsymbol{\omega}_g\|_2 \qquad (27)\end{aligned}$$

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

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

■

A.8. Proof of Theorem 5

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

Lemma 7. *Consider $a_g \geq 1$, with probability at least $1 - 6\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right)$ the following upper bound holds:*

$$\rho_g\left(\frac{1}{a_g n_g}\right) \leq \frac{1}{2}\left[\left(1 - \frac{1}{a_g}\right) + \sqrt{2}c_g\frac{2\omega(\mathcal{A}_g)+\tau}{a_g\sqrt{n_g}}\right] \qquad (28)$$

Lemma 8. *Consider $a_g \geq 1$, with probability at least $1 - 4\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right)$ the following upper bound holds:*

$$\phi_g\left(\frac{1}{a_g n_g}\right) \leq \frac{1}{a_g}\left(1 + c_{0g}\frac{\omega(\mathcal{A}_g)+\omega(\mathcal{A}_0)+2\tau}{\sqrt{n_g}}\right) \qquad (29)$$

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

Starting from the deterministic form of the bound in Theorem 4 and putting in the step sizes as $\mu_g = \frac{1}{n_g a_g}$:

$$\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g^{(t+1)}\|_2 \leq (\rho)^t \sum_{g=0}^{G} \|\boldsymbol{\beta}_g^*\|_2 + \frac{1-(\rho)^t}{1-\rho} \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \eta_g \left(\frac{1}{n_g a_g}\right) \|\boldsymbol{\omega}_g\|_2, \tag{30}$$

where

$$\rho(a_0, \cdots, a_G) = \max\left(\rho_0\left(\frac{1}{na_0}\right) + \sum_{g=1}^{G} \sqrt{\frac{n_g}{n}} \phi_g\left(\frac{1}{n_g a_g}\right), \max_{g\in[G]} \rho_g\left(\frac{1}{n_g a_g}\right) + \sqrt{\frac{n}{n_g}} \frac{\mu_0}{\mu_g} \phi_g\left(\frac{1}{n_g a_g}\right)\right) \tag{31}$$

Remember the following two results to upper bound ρ_gs and ϕ_gs from Lemmas 7 and 8:

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

First we want to keep $\rho_0 + \sum_{g=1}^{G} \sqrt{\frac{n_g}{n}} \phi_g$ of (31) strictly below 1.

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

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

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

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

$$\sqrt{2} c_0 \frac{2\omega(\mathcal{A}_0)+\tau}{\sqrt{n}} \leq a_0\left(1 - \sum_{g=1}^{G} \frac{n_g}{n} b_g\right) + 1 \tag{32}$$

Note that any lower bound on the RHS of (32) will lead to the correct sample complexity for which the coefficient of $\|\boldsymbol{\delta}_0^{(t)}\|_2$ (determined in (31)) will be below one. Since $a_0 \geq 1$ we can ignore the first term by assuming $\max_{g\in[G]_\setminus} b_g \leq 1$ and the condition becomes:

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

which can be simplified to:

$$\begin{aligned}
&n > 2c_0^2(2\omega(\mathcal{A}_0)+\tau)^2, a_0 \geq 1, \\
&\forall g \in [G]_\setminus : a_g = 2b_g^{-1}\sqrt{\frac{n}{n_g}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right), 0 < b_g \leq 1
\end{aligned} \tag{33}$$

Secondly, we want to bound all of $\rho_g + \mu_0\sqrt{\frac{n}{n_g}}\frac{\phi_g}{\mu_g}$ terms of (31) for $\mu_g = \frac{1}{a_g n_g}$ by 1:

$$\begin{aligned}\rho_g\left(\frac{1}{n_g a_g}\right) + \sqrt{\frac{n}{n_g}}\frac{\mu_0}{\mu_g}\phi_g\left(\frac{1}{n_g a_g}\right) &= \rho_g\left(\frac{1}{n_g a_g}\right) + \sqrt{\frac{n_g}{n}}\frac{a_g}{a_0}\phi_g\left(\frac{1}{n_g a_g}\right) \\ &= \frac{1}{2}\Bigg[\left[\left(1-\frac{1}{a_g}\right) + \sqrt{2}c_g\frac{2\omega_g+\tau}{a_g\sqrt{n_g}}\right] \\ &+ \frac{2}{a_0}\sqrt{\frac{n_g}{n}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)\Bigg] \\ &\leq 1\end{aligned} \tag{34}$$

The condition becomes:

$$\sqrt{2}c_g\frac{2\omega_g+\tau}{\sqrt{n_g}} \leq a_g + 1 - \sqrt{\frac{n_g}{n}}\frac{2a_g}{a_0}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right) \tag{35}$$

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

$$\sqrt{2}c_g\frac{2\omega_g+\tau}{d_g} \leq \sqrt{n_g}, \quad d_g := a_g + 1 - \frac{4}{b_g a_0}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)^2 \tag{36}$$

for $d_g > 0$. Note that any positive lower bound of the d_g will satisfy the condition in (36) 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 (33), so we take $a_0 = 4\max_{g\in[G]_\setminus}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)^2$ and look for a lower bound for d_g:

$$d_g \geq a_g + 1 - {b_g}^{-1} \tag{37}$$

$$\begin{aligned}(a_g \text{ from (33)}) &= 2b_g^{-1}\sqrt{\frac{n}{n_g}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right) + 1 - {b_g}^{-1} \\ &= 1 + b_g^{-1}\left[2\sqrt{\frac{n}{n_g}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right) - 1\right]\end{aligned} \tag{38}$$

The term inside of the last bracket (38) is always positive and therefore a lower bound is one, i.e., $d_g \geq 1$. From the condition (36) we get the following sample complexity:

$$n_g > 2c_g^2(2\omega_g+\tau)^2 \tag{39}$$

Now we need to determine b_g from previous conditions (33), knowing that $a_0 = 4\max_{g\in[G]_\setminus}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)^2$. We have $0 < b_g \leq 1$ in (33) and we take the largest step by setting $b_g = 1$.

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

$$\begin{aligned}&n > 2c_0^2\left(2\omega(\mathcal{A}_0)+\tau\right)^2, \forall g \in [G]_\setminus : n_g \geq 2c_g^2(2\omega(\mathcal{A}_g)+\tau)^2 \\ &a_0 = 4\max_{g\in[G]_\setminus}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)^2, a_g = 2\sqrt{\frac{n}{n_g}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right) \\ &\mu_0 = \frac{1}{4n} \times \frac{1}{\max_{g\in[G]_\setminus}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)^2}, \mu_g = \frac{1}{2\sqrt{nn_g}}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)^{-1}\end{aligned} \tag{40}$$

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

and $r_g(\tau) = r_{g1} + \sqrt{\frac{n_g}{n}}\frac{a_g}{a_0} r_{g2}, \forall g \in [G]_\setminus$, and $r(\tau) = \max_{g\in[G]} r_g$. All of which are computed using a_gs specified in (40). Basically r is an instantiation of an upper bound of the ρ defined in (31) using a_gs in (40).

We are interested to upper bound the following probability:

$$
\begin{aligned}
& \mathbb{P}\left(\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g^{(t+1)}\|_2 \geq r(\tau)^t \sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\beta}_g^*\|_2 + \frac{(G+1)\sqrt{(2K^2+1)}}{(1-r(\tau))\sqrt{n}}\left(\zeta k \max_{g\in[G]}\omega(\mathcal{A}_g)+\tau\right)\right) \\
\leq\ & \mathbb{P}\Bigg((\rho)^t \sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\beta}_g^*\|_2 + \frac{1-(\rho)^t}{1-\rho}\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\eta_g\left(\frac{1}{n_g a_g}\right)\|\boldsymbol{\omega}_g\|_2 \\
\geq\ & r(\tau)^t \sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\beta}_g^*\|_2 + \frac{(G+1)\sqrt{(2K^2+1)}}{(1-r(\tau))\sqrt{n}}\left(\zeta k \max_{g\in[G]}\omega(\mathcal{A}_g)+\tau\right)\Bigg) \\
\leq\ & \mathbb{P}\left(\rho \geq r(\tau)\right) \\
+\ & \mathbb{P}\left(\frac{1}{1-\rho}\sum_{g=0}^{G}\sqrt{n_g}\eta_g\left(\frac{1}{n_g a_g}\right)\|\boldsymbol{\omega}_g\|_2 \geq \frac{(G+1)\sqrt{(2K^2+1)}}{(1-r(\tau))}\left(\zeta k \max_{g\in[G]}\omega(\mathcal{A}_g)+\tau\right)\right) \qquad (41)
\end{aligned}
$$

where the first inequality comes from the deterministic bound of (30), We first focus on bounding the first term $\mathbb{P}\left(\rho \geq r(\tau)\right)$:

$$
\begin{aligned}
& \mathbb{P}\left(\rho \geq r(\tau)\right) \\
=\ & \mathbb{P}\left(\max\left(\rho_0\left(\frac{1}{na_0}\right)+\sum_{g=1}^{G}\sqrt{\frac{n_g}{n}}\phi_g\left(\frac{1}{n_g a_g}\right), \max_{g\in[G]}\rho_g\left(\frac{1}{n_g a_g}\right)+\sqrt{\frac{n}{n_g}}\frac{\mu_0}{\mu_g}\phi_g\left(\frac{1}{n_g a_g}\right)\right) \geq \max_{g\in[G]} r(\tau)\right) \\
\leq\ & \mathbb{P}\left(\rho_0\left(\frac{1}{na_0}\right)+\sum_{g=1}^{G}\sqrt{\frac{n_g}{n}}\phi_g\left(\frac{1}{n_g a_g}\right) \geq r_0\right) + \sum_{g=1}^{G}\mathbb{P}\left(\rho_g\left(\frac{1}{n_g a_g}\right)+\sqrt{\frac{n}{n_g}}\frac{\mu_0}{\mu_g}\phi_g\left(\frac{1}{n_g a_g}\right) \geq r_g\right) \\
\leq\ & \mathbb{P}\left(\rho_0\left(\frac{1}{na_0}\right) \geq r_{01}\right) + \sum_{g=1}^{G}\mathbb{P}\left(\phi_g\left(\frac{1}{n_g a_g}\right) \geq r_{g2}\right) + \sum_{g=1}^{G}\left[\mathbb{P}\left(\rho_g\left(\frac{1}{n_g a_g}\right) \geq r_{g1}\right) + \mathbb{P}\left(\phi_g\left(\frac{1}{n_g a_g}\right) \geq r_{g2}\right)\right] \\
\leq\ & \sum_{g=0}^{G}\mathbb{P}\left(\rho_g\left(\frac{1}{n_g a_g}\right) \geq r_{g1}\right) + 2\sum_{g=1}^{G}\mathbb{P}\left(\phi_g\left(\frac{1}{n_g a_g}\right) \geq r_{g2}\right) \\
\leq\ & \sum_{g=0}^{G} 6\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right) + 2\sum_{g=1}^{G} 4\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right) \\
\leq\ & 6(G+1)\exp\left(-\gamma \min_{g\in[G]}(\omega(\mathcal{A}_g)+\tau)^2\right) + 8G\exp\left(-\gamma \min_{g\in[G]_\setminus}(\omega(\mathcal{A}_g)+\tau)^2\right) \\
\leq\ & 14(G+1)\exp\left(-\gamma \min_{g\in[G]}(\omega(\mathcal{A}_g)+\tau)^2\right) \qquad (42)
\end{aligned}
$$

Now we focus on bounding the second term:

$$
\begin{aligned}
& \mathbb{P}\left(\frac{1}{1-\rho}\sum_{g=0}^{G}\sqrt{n_g}\eta_g\left(\frac{1}{n_g a_g}\right)\|\boldsymbol{\omega}_g\|_2 \geq \frac{(G+1)\sqrt{(2K^2+1)}}{(1-r(\tau))}\left(\zeta k \max_{g\in[G]}\omega(\mathcal{A}_g)+\tau\right)\right) \\
\leq\ & \mathbb{P}\left(\frac{1}{1-\rho}\sum_{g=0}^{G}\sqrt{n_g}\eta_g\left(\frac{1}{n_g a_g}\right)\|\boldsymbol{\omega}_g\|_2 \geq \frac{1}{(1-r(\tau))}\sum_{g=0}^{G}\sqrt{(2K^2+1)}\left(\zeta_g k\omega(\mathcal{A}_g)+\tau\right)\right) \\
\leq\ & \mathbb{P}\left(\sum_{g=0}^{G}\sqrt{n_g}\eta_g\left(\frac{1}{n_g a_g}\right)\|\boldsymbol{\omega}_g\|_2 \geq \sum_{g=0}^{G}\sqrt{(2K^2+1)}\left(\zeta_g k\omega(\mathcal{A}_g)+\tau\right)\right) + \mathbb{P}\left(\rho \geq r(\tau)\right) \\
\leq\ & \sum_{g=0}^{G}\mathbb{P}\left(\sqrt{n_g}\eta_g\left(\frac{1}{n_g a_g}\right)\|\boldsymbol{\omega}_g\|_2 \geq \sqrt{(2K^2+1)}\left(\zeta_g k\omega(\mathcal{A}_g)+\tau\right)\right) + \mathbb{P}\left(\rho \geq r(\tau)\right) \qquad (43)
\end{aligned}
$$

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

$$\mathbb{P}\left(\|\boldsymbol{\omega}_g\|_2 \sup_{\mathbf{v}\in\mathcal{B}_g} \mathbf{v}^T \mathbf{X}_g^T \frac{\boldsymbol{\omega}_g}{\|\boldsymbol{\omega}_g\|_2} \geq a_g \sqrt{(2K^2+1)n_g}\,(\zeta_g k \omega(\mathcal{A}_g) + \tau) \right) \leq \sigma_g \exp\left(-\min\left[\nu_g n_g, \frac{\tau^2}{\eta_g^2 k^2} \right] \right) \tag{44}$$

where we used the intermediate form of Lemma 1 for $\tau > 0$. Putting all of the bounds (42), (43), and (44) back into the (41):

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

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

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

and it also affects step sizes as follows:

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

■

A.9. Proof of Lemma 3

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

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

■

A.10. Proof of Lemma 4

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

$$\|\boldsymbol{\delta}_0 + \boldsymbol{\delta}_i\|_2 \geq \begin{cases} \lambda_{\min}(\|\boldsymbol{\delta}_i\|_2 + \|\boldsymbol{\delta}_0\|_2) & \text{if } i \in \mathcal{I} \\ \|\boldsymbol{\delta}_i\|_2/2 & \text{if } i \in \mathcal{J} \\ 0 & \text{if } i \in \mathcal{K} \end{cases} \tag{45}$$

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 $\mathcal{K}$, $\|\boldsymbol{\delta}_i\|_2 \le 2\|\boldsymbol{\delta}_0\|_2$ which implies $S_{\mathcal{K}} = \sum_{i\in\mathcal{K}} n_i\|\boldsymbol{\delta}_i\|_2 \le 2\sum_{i\in\mathcal{K}} n_i\|\boldsymbol{\delta}_0\|_2 \le 2n\|\boldsymbol{\delta}_0\|_2$. Set $\psi_{\mathcal{I}} = \min\{1/2, \lambda_{\min}\bar{\rho}/3\} = \lambda_{\min}\bar{\rho}/3$. Using $1/2 \ge \psi_{\mathcal{I}}$, we write:

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

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

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

Combining all, we obtain:

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

■

A.11. Proof of Lemma 5

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

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

Now:

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

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

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

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

A.12. Proof of Lemma 7

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

Lemma 9 (Theorem 11 of (Banerjee et al., 2014)). *For all $g \in [G]$, for the matrix $\mathbf{X}_g \in \mathbb{R}^{n_g\times p}$ with independent isotropic sub-Gaussian rows, i.e., $\|\mathbf{x}_{gi}\|_{\psi_2} \leq k$ and $\mathbb{E}[\mathbf{x}_{gi}\mathbf{x}_{gi}^T] = \mathbf{I}$, the following result holds with probability at least $1 - 2\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right)$ for $\tau > 0$:*

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

where $c_g > 0$ is constant.

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

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

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

Now the proof for Lemma 7:

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

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

We upper bound the argument of the sup as follows:

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

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 9 for bigger sets of $\mathcal{A}_g + \mathcal{A}_g$ and $\mathcal{A}_g - \mathcal{A}_g$ where Gaussian width of both of them are upper bounded by $2\omega(\mathcal{A}_g)$. The above holds with high probability (computed below). Now we set :

$$\mathbf{v}^T\big(\mathbf{I}_g - \frac{1}{a_g n_g}\mathbf{X}_g^T\mathbf{X}_g\big)\mathbf{u} \leq \frac{1}{2}\left[\left(1-\frac{1}{a_g}\right) + \sqrt{2}c_g\frac{2\omega(\mathcal{A}_g)+\tau}{a_g\sqrt{n_g}}\right] \tag{46}$$

To keep the upper bound of ρ_g in (46) 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:

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

which concludes the proof. ■

A.13. Proof of Lemma 8

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

$$-\mathbf{v}^T\mathbf{X}_g^T\mathbf{X}_g\mathbf{u} \leq \frac{1}{2}\left(\|\mathbf{X}_g\mathbf{u}\|_2^2 + \|\mathbf{X}_g\mathbf{v}\|_2^2\right) \tag{47}$$

Now we can bound ϕ_g as follows:

$$\phi_g(\mu_g) = \mu_g \sup_{\mathbf{v}\in\mathcal{B}_g, \mathbf{u}\in\mathcal{B}_0} -\mathbf{v}^T\mathbf{X}_g^T\mathbf{X}_g\mathbf{u} \leq \frac{\mu_g}{2}\left(\sup_{\mathbf{u}\in\mathcal{B}_0}\|\mathbf{X}_g\mathbf{u}\|_2^2 + \sup_{\mathbf{v}\in\mathcal{B}_g}\|\mathbf{X}_g\mathbf{v}\|_2^2\right) \tag{48}$$

So we have:

$$\begin{aligned}
\phi_g\left(\frac{1}{a_g n_g}\right) &\leq \frac{1}{2a_g}\left(\frac{1}{n_g}\sup_{\mathbf{u}\in\mathcal{B}_0}\|\mathbf{X}_g\mathbf{u}\|_2^2 + \frac{1}{n_g}\sup_{\mathbf{v}\in\mathcal{B}_g}\|\mathbf{X}_g\mathbf{v}\|_2^2\right) \qquad (49) \\
\text{(Lemma 9)} &\leq \frac{1}{a_g}\left(1 + c_{0g}\frac{\omega(\mathcal{A}_g)+\omega(\mathcal{A}_0)+2\tau}{2\sqrt{n_g}}\right) \\
(\omega_{0g} = \max(\omega(\mathcal{A}_0),\omega(\mathcal{A}_g)) &\leq \frac{1}{a_g}\left(1 + c_{0g}\frac{\omega_{0g}+\tau}{\sqrt{n_g}}\right)
\end{aligned}$$

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

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

$$\begin{aligned}\mathbb{P}(\phi_g > s_g) &\leq \mathbb{P}(\phi_g > s_g|\neg\mathcal{E}_1)\mathbb{P}(\neg\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_1) \\ &\leq \mathbb{P}(\mathcal{E}_2) + \mathbb{P}(\mathcal{E}_1) \\ &\leq 4\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right)\end{aligned} \tag{50}$$

■