High Dimensional Data Enrichment: Interpretable, Fast, and Data-Efficient[*]

Amir Asiaee[†], Samet Oymak[‡], Kevin R. Coombes[§], and Arindam Banerjee[¶]

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 geometric convergence rate and supplement our theoretical analysis with synthetic experiments. Overall, we present a first through statistical and computational analysis of inference in the data enriched model.

Key words. multi-task learning, superposition models, high-dimensional statistics, convergence rate analysis

AMS subject classifications. 62F10, 62J05, 90C25

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 [11, 17, 18]. Such estimators consider a suitable (semi) parametric model of the response: $y = \phi(\mathbf{x}, \boldsymbol{\beta}^*) + w$ based on n samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and the parameter of interest, $\boldsymbol{\beta}^* \in \mathbb{R}^p$. The unique aspect of such high-dimensional regime of $n \ll p$ is that the structure of $\boldsymbol{\beta}^*$ makes the estimation possible for large enough samples $n = m$ known as the sample complexity [9, 10, 34]. 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 [1, 27], broad families of semi-parametric and single-index models [7, 31], non-convex models [5, 22], etc.

In several real world problems, the assumption that one global model parameter $\boldsymbol{\beta}_0^*$ 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 have unique differences. For example, in the context of anti-cancer drug sensitivity prediction where the goal is to predict responses of different tumor cells to a drug, using a same prediction model across cancer types (groups) ignores the unique properties of each cancer and leads to an uninterpretable global model. Alternatively, in such a setting, one can assume a separate model for each group g as $y = \phi(\mathbf{x}, \boldsymbol{\beta}_g^*) + w$ based on a group specific parameter $\boldsymbol{\beta}_g^*$. Such a modeling choice fails to leverage the similarities across the sub-populations, and can only be estimated when sufficient number of samples are available for each group which is not the case in several problems, e.g., anti-cancer drug sensitivity prediction [3, 21].

[*]Submitted to the editors DATE.
Funding: This work was funded by the Fog Research Institute under contract no. FRI-454.
[†]Mathematical Biosciences Institute, The Ohio State University, Columbus, OH (asiaeetaheri.1@osu.edu).
[‡]Department of Electrical and Computer Engineering, UC Riverside, Riverside, CA (oymak@ece.ucr.edu).
[§]Departments of Biomedical Informatics, The Ohio State University, Columbus, OH (coombes.3@osu.edu).
[¶]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN (banerjee@cs.umn.edu).

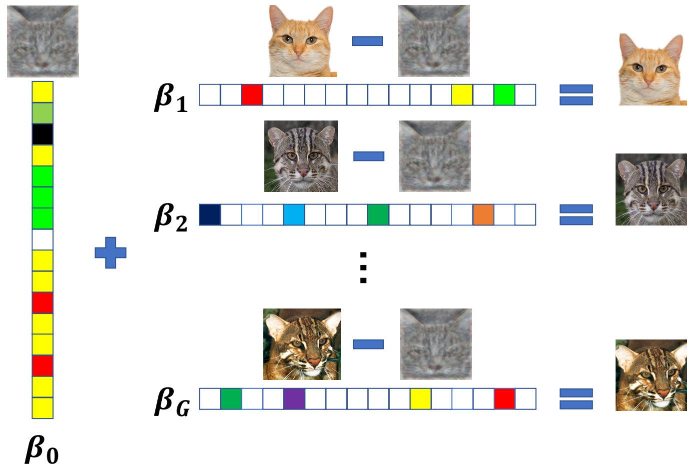


Figure 1: A conceptual illustration of data enrichment model for learning representation of cat species. The common parameter $\boldsymbol{\beta}_0$ captures a *generic cat* which consists of shared features among all cats.

The middle ground model for such a scenario is the *superposition* of common and individual parameters $\boldsymbol{\beta}_0^* + \boldsymbol{\beta}_g^*$ which has been of recent interest [20, 24, 38]. Such a collection of *coupled* superposition models is known by multiple names in the statistical machine learning community. It is a form of multi-task learning [23, 39] when we consider regression in each group as a task. It is also called data sharing [19] since information contained in different groups is shared through the common parameter $\boldsymbol{\beta}_0^*$. And finally, it has been called data enrichment [14] because we enrich our data set with pooling multiple samples from different but related sources.

Following the successful application of such a modeling scheme in recent years [16, 19, 29, 30], we consider the below *data enrichment* (DE) model:

$$y_{gi} = \phi(\mathbf{x}_{gi}, (\boldsymbol{\beta}_0^* + \boldsymbol{\beta}_g^*)) + w_{gi}, \quad g \in \{1, \dots, G\}\ , \tag{1.1}$$

where g and i index the group and samples respectively. DE model (1.1) assumes that there is a *common* parameter $\boldsymbol{\beta}_0^*$ shared between all groups which models similarities between all samples. And there are *individual* per-group parameters $\boldsymbol{\beta}_g^*$s each characterize the deviation of group g, Figure 1.

The setting. Our goal is to design an estimation procedure which consistently recovers all parameters of DE (1.1) fast and with small number of samples. We specifically focus on the high-dimensional small sample regime where the number of samples n_g for each group is much smaller than the ambient dimensionality, i.e., $\forall g : n_g \ll p$. Similar to all other high-dimensional models, we assume that the parameters $\boldsymbol{\beta}_g$ are structured, i.e., for suitable convex functions f_g's, $f_g(\boldsymbol{\beta}_g)$ is small. 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 [26].

1.1. Related Work.

In the context of *Multi-Task Learning* (MTL), similar models have been proposed which have the general form of $y_{gi} = \mathbf{x}_{gi}^T(\boldsymbol{\beta}_{1g}^* + \boldsymbol{\beta}_{2g}^*) + w_{gi}$ where $\mathbf{B}_1 = [\boldsymbol{\beta}_{11}, \dots, \boldsymbol{\beta}_{1G}]$

and $\mathbf{B}_2 = [\boldsymbol{\beta}_{21}, \ldots, \boldsymbol{\beta}_{2G}]$ are two parameter matrices [39]. To capture the relation of tasks, different types of constraints are assumed for parameter matrices. For example, [15] assumes $\mathbf{B}_1$ and $\mathbf{B}_2$ are sparse and low rank respectively. In this parameter matrix decomposition framework for MLT, the most related work to ours is the Dirty Statistical Model (DSM) proposed in [23] where authors regularize the regression with $\|\mathbf{B}_1\|_{1,\infty}$ and $\|\mathbf{B}_2\|_{1,1}$ where norms are p, q-norms on *rows* of matrices, i.e., $\|\cdot\|_{p,q} = \|(\|\cdot\|_q, \ldots, \|\cdot\|_q)\|_p$.

If in our DE model we pick all structure inducing functions f_g to be l_1-norm, the resulting model is very similar to the DSM where $\|\mathbf{B}_1\|_{1,\infty}$ induces similarity between tasks and $\|\mathbf{B}_2\|_{1,1}$ models their discrepancies. On the other hand, the degree of freedom of DSM model is higher than DE because $\|\mathbf{B}_1\|_{1,\infty}$ regularizer enforces shared support of $\boldsymbol{\beta}^*_{1g}$s, i.e., $\text{supp}(\boldsymbol{\beta}^*_{1i}) = \text{supp}(\boldsymbol{\beta}^*_{1j})$ but allows $\boldsymbol{\beta}^*_{1i} \neq \boldsymbol{\beta}^*_{1j}$ while in DE we have a single common parameter $\boldsymbol{\beta}^*_0$. So one would expect that DE estimators should have smaller sample complexity compared to their DSM counterparts and our analysis confirm that our estimator is more data efficient than DSM estimator of [23], Table 1. Mainly, DSM requires every task i to have large enough samples to learn its own common parameters $\boldsymbol{\beta}_i$ but since DE shares the common parameter it only requires the *total dataset over all tasks* to be sufficiently large.

The linear 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 [16], sentiment analysis, banking strategy [19], single cell data analysis [30], road safety [29], and disease subtype analysis [16]. 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.

Motivation. 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 DE model (1.1). 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 [19, 29]. To the best of our knowledge, the only theoretical guarantee for data enrichment is provided in [30] where authors prove sparsistency of their proposed method under the irrepresentability condition of the design matrix for recovering *supports* of common and individual parameters. Existing support recovery guarantees [30], sample complexity and l_2 consistency results [23] of related MTL models are restricted to sparsity and l_1-norm, while our estimator and *norm consistency* analysis work for *any* structure induced by arbitrary convex functions f_g. Moreover, no computational results, such as rates of convergence of the estimation procedures exist in the literature.

1.2. 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 $[G] = \{1, \ldots, G\}$ and $[G_+] = [G] \cup \{0\}$. Throughout the manuscript c_i and C_i denote positive absolute constants. 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 $n = \sum_{g=1}^{G} n_g$ and the data enriched model takes the following vector form:

$$\mathbf{y}_g = \mathbf{X}_g(\boldsymbol{\beta}^*_0 + \boldsymbol{\beta}^*_g) + \mathbf{w}_g, \quad \forall g \in [G] \tag{1.2}$$

where each row of $\mathbf{X}_g$ is $\mathbf{x}^T_{gi}$ and $\mathbf{w}^T_g = (w_{g1}, \ldots, w_{gn_g})$ is the noise vector. It is useful for indexing to consider the common parameter as the zeroth group and define $n_0 \triangleq n$ and $\mathbf{X}_0 \triangleq [\mathbf{X}^T_1, \ldots, \mathbf{X}^T_G]^T$.

Sub-Gaussian random variable and 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 $\|\|V\|\|_{\psi_2}$ [36]. 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 [36] 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 $\omega(\mathcal{V}) = \mathbb{E}_{\mathbf{g}} [\sup_{\mathbf{u} \in \mathcal{V}} \langle \mathbf{g}, \mathbf{u} \rangle]$ [37], 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. The marginal tail function is defined as $Q_\xi(\mathbf{u}) = \mathbb{P}(|\langle \mathbf{x}, , \mathbf{u} \rangle| > \xi)$ for a fixed vector $\mathbf{u}$, random vector $\mathbf{x}$ and constant $\xi > 0$.

1.3. Our 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 $f_g(\cdot)$:

$$\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_0^T, \dots, \hat{\boldsymbol{\beta}}_G^T) \in \underset{\boldsymbol{\beta}_0, \dots, \boldsymbol{\beta}_G}{\operatorname{argmin}} \frac{1}{n} \sum_{g=1}^{G} \|\mathbf{y}_g - \mathbf{X}_g(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_g)\|_2^2, \text{ s.t. } \forall g \in [G_+] : f_g(\boldsymbol{\beta}_g) \le f_g(\boldsymbol{\beta}_g^*). \tag{1.3}$$

We present several statistical and computational results for the DE estimator (1.3):

- The DE estimator (1.3) succeeds if a geometric condition that we call *Data EnRichment Incoherence Condition* (DERIC) is satisfied, Figure 2b. Compared to other known geometric conditions in the literature such as structural coherence [20] and stable recovery conditions [24], DERIC is a considerably weaker condition, Figure 2a.
- 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:

$$\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g\|_2 \le \gamma O\left(\frac{\max_{g \in [G]} \omega(\mathcal{C}_g \cap \mathbb{S}^{p-1})}{\sqrt{n}}\right), \tag{1.4}$$

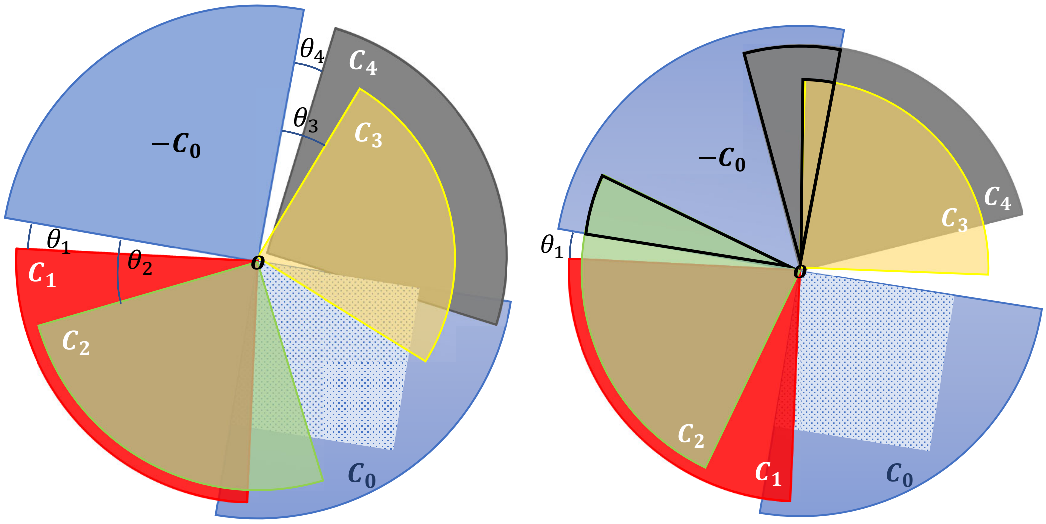
  where $n_0 \triangleq n$ is the total number of samples, $\gamma \triangleq \max_{g \in [G]} \frac{n}{n_g}$ is the *sample condition number*, and $\mathcal{C}_g$ is the error cone corresponding to $\boldsymbol{\beta}_g^*$ exactly defined in section 2. To the best of our knowledge, this is the first statistical estimation guarantee for the data enrichment.
- We also establish the sample complexity of the DE estimator for all parameters as $\forall g \in [G_+] : n_g = O(\omega(\mathcal{C}_g \cap \mathbb{S}^{p-1}))^2$. We emphasize that our result proofs that the recovery of the common parameter $\boldsymbol{\beta}_0$ by DE estimator (1.3) benefits from *all* of the n pooled samples.
- We present an efficient projected block gradient descent algorithm DICER, to solve DE's objective (1.3) which converges geometrically to the statistical error bound of (1.4). To the best of our knowledge, this is the first rigorous computational result for the high-dimensional data-enriched regression.

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 geometrically convergent algorithm, DICER in section 5 and finally supplement our work with synthetic experiments in section 6.

2. The Data Enrichment Estimator.

A compact form of our proposed DE estimator (1.3) is:

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \quad \text{s.t. } \forall g \in [G_+] : f_g(\boldsymbol{\beta}_g) \le f_g(\boldsymbol{\beta}_g^*), \tag{2.1}$$



(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 $\mathcal{E}_g = \left\{\boldsymbol{\delta}_g | f_g(\boldsymbol{\beta}_g^* + \boldsymbol{\delta}_g) \leq f_g(\boldsymbol{\beta}_g^*)\right\}$ are the error sets for each parameter $\boldsymbol{\beta}_g^* \in [G]$ [20]. b) Our more relaxed recovery condition which allows *arbitrary non-zero fraction* of the error cones of individual parameters intersect with $-\mathcal{C}_0$.

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

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_1 & 0 & \cdots & 0 \\ \mathbf{X}_2 & 0 & \mathbf{X}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \mathbf{X}_G & 0 & \cdots & \cdots & \mathbf{X}_G \end{pmatrix} \in \mathbb{R}^{n \times (G+1)p} \,. \tag{2.2}$$

Example 2.1. **(l_1-norm)** When all parameters $\boldsymbol{\beta}_g$s are s_g-sparse, i.e.,$|\text{supp}(\boldsymbol{\beta}_g^*)| = s_g$ by using l_1-norm as the sparsity inducing function, our DE estimator of (2.1) becomes:

$$\hat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \quad \text{s.t. } \forall g \in [G_+] : \|\boldsymbol{\beta}_g\|_1 \leq \|\boldsymbol{\beta}_g^*\|_1. \tag{2.3}$$

We call Example 2.1 *sparse DE* estimator and use it as the running example throughout the paper to illustrate outcomes of our analysis.

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 (2.1), the error vector $\boldsymbol{\delta}_g$ will belong to the following restricted error set:

$$\mathcal{E}_g = \left\{\boldsymbol{\delta}_g | f_g(\boldsymbol{\beta}_g^* + \boldsymbol{\delta}_g) \leq f_g(\boldsymbol{\beta}_g^*)\right\}, \quad g \in [G_+]. \tag{2.4}$$

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

play key roles in our analysis:

$$\mathcal{H} \triangleq \left\{ \boldsymbol{\delta} \in \mathcal{C} \middle| \sum_{g=0}^{G} \frac{n_g}{n} \|\boldsymbol{\delta}_g\|_2 = 1 \right\}, \quad \bar{\mathcal{H}} \triangleq \left\{ \boldsymbol{\delta} \in \mathcal{C} \middle| \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g\|_2 = 1 \right\}. \tag{2.5}$$

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: $\frac{1}{n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 \leq \frac{1}{n} 2\mathbf{w}^T\mathbf{X}\boldsymbol{\delta}$ where $\mathbf{w} = [\mathbf{w}_1^T, \ldots, \mathbf{w}_G^T]^T \in \mathbb{R}^n$ is the vector of all noises. Using this basic inequality, we can establish the following deterministic error bound.

Theorem 2.2. *For the DE estimator (2.1), 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 $\gamma = \max_{g\in[G]} \frac{n}{n_g}$, the following deterministic upper bounds holds:*

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

Proof. We lower bound the LHS and upper bound the RHS of the optimality inequality $\frac{1}{n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 \leq \frac{1}{n}2\mathbf{w}^T\mathbf{X}\boldsymbol{\delta}$ using the definition of the sets $\mathcal{H}$ and $\bar{\mathcal{H}}$ respectively. Starting with the lower bound using the definition of set $\mathcal{H}$ (2.5) we have:

$$\begin{aligned} \frac{1}{n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 &\geq \frac{1}{n}\inf_{\mathbf{u}\in\mathcal{H}} \|\mathbf{X}\mathbf{u}\|_2^2 \left(\sum_{g=0}^{G} \frac{n_g}{n} \|\boldsymbol{\delta}_g\|_2 \right)^2 \geq \kappa \left(\sum_{g=0}^{G} \frac{n_g}{n} \|\boldsymbol{\delta}_g\|_2 \right)^2 \\ &\geq \kappa \left(\min_{g\in[G]} \sqrt{\frac{n_g}{n}} \right)^2 \left(\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g\|_2 \right)^2 = \kappa \left(\min_{g\in[G]} \frac{n_g}{n} \right) \left(\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g\|_2 \right)^2 \end{aligned} \tag{2.6}$$

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 factorizes as:

$$\frac{2}{n}\mathbf{w}^T\mathbf{X}\boldsymbol{\delta} \leq \frac{2}{n}\sup_{\mathbf{u}\in\bar{\mathcal{H}}} \mathbf{w}^T\mathbf{X}\mathbf{u} \left(\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g\|_2 \right), \quad \mathbf{u} \in \mathcal{H} \tag{2.7}$$

Putting together inequalities (2.6) and (2.7) completes the proof. ■

Remark 2.3. 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

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

3. Restricted Eigenvalue Condition.

The main assumptions of Theorem 2.2 is known as Restricted Eigenvalue (RE) condition in the literature of high-dimensional statistics [2, 28, 32]: $\inf_{\mathbf{u}\in\mathcal{H}} \frac{1}{n}\|\mathbf{X}\mathbf{u}\|_2^2 \geq \kappa > 0$. 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 (2.2), the RE condition holds with high probability under a suitable geometric condition we call *Data EnRichment Incoherence Condition* (DERIC) and for enough number of samples. We precisely characterize total and per-group sample complexities required for successful parameter recovery. For the analysis, similar to existing work [20, 25, 35], we assume the design matrix to be isotropic sub-Gaussian.[1]

Definition 3.1. *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 $|||\mathbf{x}|||_{\psi_2} \leq k_x$. 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 $\mathbf{w}_{gi}$ are i.i.d. zero-mean, unit-variance sub-Gaussian with $|||\mathbf{w}_{gi}|||_{\psi_2} \leq k_w$.*

3.1. Geometric Condition of Recovery.

Unlike standard high-dimensional statistical estimation, for RE condition to be true, parameters of superposition models need to satisfy geometric conditions which limits the interaction of parameters with each other to make sure that recovery is possible. In this section, we elaborate our sufficient geometric condition for recovery and compare it with state-of-the-art condition for recovery of superposition models.

To intuitively illustrate the necessity of such a geometric condition, consider the simplest superposition model i.e., $\boldsymbol{\beta}_0^* + \boldsymbol{\beta}_g^*$. Without any restriction on parameter interactions, any estimates such that $\hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_g = \boldsymbol{\beta}_0^* + \boldsymbol{\beta}_g^*$ are valid ones. To avoid such trivial solutions two error cones need to satisfy $\boldsymbol{\delta}_g \neq -\boldsymbol{\delta}_0$. In general, the RE condition of individual superposition models can be established under the so-called Structural Coherence (SC) condition [20, 24] which is the generalization of this idea for superposition of multiple parameters as $\sum_{g=0}^G \boldsymbol{\beta}_g^*$.

Definition 3.2 (Structural Coherence (SC) [20, 24]). *Consider a superposition model of the form $y = \mathbf{x}^T \sum_{g=0}^G \boldsymbol{\beta}_g^* + w$. The SC condition requires that*

$$\forall \boldsymbol{\delta}_g \in \mathcal{C}_g, \exists\lambda \quad s.t. \quad \|\sum_{g=0}^G \boldsymbol{\delta}_g\|_2 \geq \lambda \sum_{g=0}^G \|\boldsymbol{\delta}_g\|_2,$$

and leads to the corresponding RE condition for the superposition model.

Remark 3.3. Note that the SC condition is satisfied if none of the individual error cones $\mathcal{C}_g$ intersect with the inverted error cone $-\mathcal{C}_0$ [20, 35], i.e., $\forall g, \theta_g > 0$ in Figure 2a where

$$\cos(\theta_g) = \sup_{\boldsymbol{\delta}_0\in\mathcal{C}_0, \boldsymbol{\delta}_g\in\mathcal{C}_g} -\langle \boldsymbol{\delta}_0/\|\boldsymbol{\delta}_0\|_2, \boldsymbol{\delta}_g/\|\boldsymbol{\delta}_g\|_2\rangle.$$

Next, we introduce DERIC, a considerably weaker geometric condition compared to SC which leads to recovery of all parameters in the data enriched model.

Definition 3.4 (Data EnRichment Incoherence Condition (DERIC)). *There exists a non-empty set $\mathcal{I} \subseteq [G]$ of groups where for some scalars $0 < \bar{\rho} \leq 1$ and $\lambda_{\min} > 0$ the following holds:*

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

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

[1] Extension to an-isotropic sub-Gaussian case is straightforward by techniques developed in [2, 33].

Remark 3.5. Comparing to the SC conditions, DERIC holds even if only one of the $\mathcal{C}_g$s does not intersect with $-\mathcal{C}_0$. More specifically, DERIC holds if $\exists g, \theta_g > 0$ in Figure 2b. Therefore, instead of SC stringent geometric condition, DERIC allows $-\mathcal{C}_0$ to intersect with an arbitrarily large fraction of the $\mathcal{C}_g$ cones and as the number of intersections increases, our final error bound becomes looser.

3.2. Sample Complexity. An alternative to our DE estimator (1.3) may be based on G *isolated* superposition model $\mathbf{y}_g = \mathbf{X}_g(\boldsymbol{\beta}_0^* + \boldsymbol{\beta}_g^*) + \mathbf{w}_g$ each with two components. Now, if SC holds for at least one of the superposition models, i.e., $\exists g, -\mathcal{C}_0 \cap \mathcal{C}_g = \emptyset$, one can recover $\hat{\boldsymbol{\beta}}_0$ and plug it in to the remaining $G-1$ superposition estimators to estimate the corresponding $\hat{\boldsymbol{\beta}}_g$s. We call such an estimator, *plugin superposition* estimator. For such a trivial (based on existing literature) estimator, it seems that DERIC has no advantage over SC. But the downfall of such estimator is that it fails to utilize the true coupling structure in the data enriched model, where $\boldsymbol{\beta}_0^*$ is involved in all groups. In fact, below we show, the above combination of superposition and plug-in estimators using SC leads to a pessimistic estimates of the sample complexity for $\boldsymbol{\beta}_0^*$ recovery.

Proposition 3.6. *Assume observations distributed as defined in Definition 3.1 and pair-wise SC conditions are satisfied. Consider each superposition model (1.2) in isolation; to recover the common parameter $\boldsymbol{\beta}_0^*$ plugin superposition requires at least one group i to have $n_i = O(\max(\omega^2(\mathcal{A}_0), \omega^2(\mathcal{A}_i)))$. To recover the rest of individual parameters, it needs $\forall g \neq i : n_g = O(\omega^2(\mathcal{A}_g))$ samples.*

In other words, by separate analysis of superposition estimators at least one problem needs to have sufficient samples for recovering the common parameter $\boldsymbol{\beta}_0$ and therefore the common parameter recovery does not benefit from the pooled 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 [25], a tool from empirical process theory in the following theorem, we get a better sample complexity required for satisfying the RE condition:

Theorem 3.7. *Let $\mathbf{x}_{gi}$s be random vectors defined in Definition 3.1. Assume DERIC condition of Definition 3.4 holds for error cones $\mathcal{C}_g$s and $\psi_{\mathcal{I}} = \min\{1/2, \lambda_{\min}\bar{\rho}/3\}$. Then, for all $\boldsymbol{\delta} \in \mathcal{H}$, when we have enough number of samples as $\forall g \in [G_+] : n_g \geq m_g = O(k_x^6 \alpha^{-6} \psi_{\mathcal{I}}^{-2} \omega(\mathcal{A}_g)^2)$, with probability at least $1 - e^{-n\kappa_{\min}/4}$ we have:*

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

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

Remark 3.8. Note that $\kappa = \frac{\kappa_{\min}^2}{4}$ is the lower bound of the RE condition of Theorem 2.2, i.e., $0 < \kappa \leq \inf_{\mathbf{u} \in \mathcal{H}} \frac{1}{n} \|\mathbf{X}\mathbf{u}\|_2^2$ and is determined by the group with the worst individual RE condition.

Example 3.9. **(l_1-norm)** The Gaussian width of the spherical cap of a p-dimensional s-sparse vector is $\omega(\mathcal{A}) = \Theta(\sqrt{s \log p})$ [2, 37]. Therefore, the number of samples per group and total required for satisfaction of the RE condition in the sparse DE estimator Example 2.1 is $\forall g \in [G] : n_g \geq m_g = \Theta(s_g \log p)$. Table 1 compares sample complexities of sparse-DE estimator with three baselines: plugin superposition estimator of Proposition 3.6, G Independent LASSO (GI-LASSO), and Jalali's Dirty Statistical Model (DSM) [23]. Note that GI-LASSO does not recover the common parameter and DSM needs all groups have same number of samples.

	GI-LASSO	Dirty Stat. Model	Plugin Superposition	Sparse DE
m_g	$s_{0g} \log p$	$G \max_{g \in [G]} s_{0g} \log(p)$	$\exists i \in [G] : \max(s_0, s_i) \log p$ $\forall g \neq i : s_g \log p$	$s_g \log p$

Table 1: Comparison of the order of per group number of samples (sample complexities) of various methods for recovering sparse DE parameters. Let $s_{0g} = |\text{support}(\beta_0^* + \beta_g^*)|$ be the superimposed support where $s_0, s_g \leq \max(s_0, s_g) \leq s_{0g}$.

3.3. Proof of Theorem 3.7.

Let's simplify the LHS of the RE condition:

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

$$\text{(Lyapunov's inequality)} \geq \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}|\langle \mathbf{x}_{gi}, \boldsymbol{\delta}_0 + \boldsymbol{\delta}_g\rangle|$$

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

$$= \frac{1}{n}\sum_{g=1}^{G}\xi_g\sum_{i=1}^{n_g}\mathbb{1}\left(|\langle \mathbf{x}_{gi}, \boldsymbol{\delta}_{0g}\rangle| \geq \xi_g\right),$$

where to avoid cluttering we denoted $\boldsymbol{\delta}_{0g} = \boldsymbol{\delta}_0 + \boldsymbol{\delta}_g$ and $\xi_g = \xi\|\boldsymbol{\delta}_{0g}\|_2 > 0$. Now we add and subtract the corresponding per-group marginal tail function, $Q_{\xi_g}(\boldsymbol{\delta}_{0g}) = \mathbb{P}(|\langle \mathbf{x}, , \boldsymbol{\delta}_{0g}\rangle| > \xi_g)$ and take inf:

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

$$= t_1(\mathbf{X}) - t_2(\mathbf{X}) \tag{3.1}$$

For the ease of exposition we consider the LHS of (3.1) 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.

3.3.1. Lower Bounding the First Term.

Our main result is the following lemma which uses the DERIC condition of the Definition 3.4 and provides a lower bound for the first term $t_1(\mathbf{X})$:

Lemma 3.10. *Suppose DERIC holds. Let $\psi_{\mathcal{I}} = \frac{\lambda_{\min}\bar{\rho}}{3}$. For any $\boldsymbol{\delta} \in \mathcal{H}$, we have:*

$$\sum_{g=1}^{G}\frac{n_g}{n}\xi_g Q_{2\xi_g}(\boldsymbol{\delta}_{0g}) \geq \psi_{\mathcal{I}}\xi\frac{(\alpha - 2\xi)^2}{4ck_x^2}\sum_{g=0}^{n}\frac{n_g}{n}\|\boldsymbol{\delta}_g\|_2, \tag{3.2}$$

Lemma 3.10 implies that $t_1(\mathbf{X})$ is lower bounded by the same RHS bound of (3.2).

3.3.2. Upper Bounding the Second Term. First we show $t_2(\mathbf{X})$ satisfies the bounded difference property defined in Section 3.2. of [6], i.e., by changing each of $\mathbf{x}_{gi}$ the value of $t_2(\mathbf{X})$ at most change by one. We rewrite t_2 as $t_2(\mathbf{X}) = \sup_{\delta \in \mathcal{H}} g_\delta(\mathbf{X})$ where $g_\delta(\mathbf{X})$ is the argument of sup in (3.1). Now we denote the design matrix resulted from replacement of kth sample from jth group $\mathbf{x}_{jk}$ with another sample $\mathbf{x}'_{jk}$ by $\mathbf{X}'_{jk}$. Then our goal is to show $\forall j \in [G], k \in [n_j], \sup_{\mathbf{X}, \mathbf{x}'_{jk}} |t_2(\mathbf{X}) - t_2(\mathbf{X}'_{jk})| \leq c_i$ 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:

$$
\begin{aligned}
\sup_{\mathbf{X}, \mathbf{x}'_{jk}} \left|t_2(\mathbf{X}) - t_2\left(\mathbf{X}'_{jk}\right)\right| &\leq \sup_{\mathbf{X}, \mathbf{x}'_{jk}} \sup_{\boldsymbol{\delta} \in \mathcal{H}} \left|g(\mathbf{X}) - g\left(\mathbf{X}'_{jk}\right)\right| \\
&\leq \sup_{\mathbf{x}_{jk}, \mathbf{x}'_{jk}} \sup_{\boldsymbol{\delta} \in \mathcal{H}} \frac{\xi_j}{n} \left|\mathbb{1}(|\langle \mathbf{x}'_{jk}, \boldsymbol{\delta}_{0j} \rangle| \geq \xi_j) - \mathbb{1}(|\langle \mathbf{x}_{jk}, \boldsymbol{\delta}_{0j} \rangle| \geq \xi_j)\right| \\
&\leq \sup_j \sup_{\boldsymbol{\delta} \in \mathcal{H}} \frac{\xi_j}{n} = \frac{\xi}{n} \sup_j \sup_{\boldsymbol{\delta} \in \mathcal{H}} \|\boldsymbol{\delta}_0 + \boldsymbol{\delta}_j\|_2 \\
&\leq \frac{\xi}{n} \sup_j \sup_{\boldsymbol{\delta} \in \mathcal{H}} \|\boldsymbol{\delta}_0\|_2 + \|\boldsymbol{\delta}_j\|_2 \\
(\boldsymbol{\delta} \in \mathcal{H}) &\leq \xi \left(\frac{1}{n} + \frac{1}{n_j}\right) \leq \frac{2\xi}{n}
\end{aligned}
$$

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 $\|\boldsymbol{\delta}_g\|_2 \leq \frac{n}{n_g}$ which justifies the last inequality. Now, we can invoke the bounded difference inequality from Theorem 6.2 of [6] 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 $t_2(\mathbf{X})$ using the following lemma:

Lemma 3.11. *For the random vector $\mathbf{x}$ of Definition 3.1, we have the following bound:*

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

3.3.3. Continuing the Proof of Theorem 3.7. Define $q \triangleq \frac{(\alpha - 2\xi)^2}{4ck^2}$. Putting back bounds of $t_1(\mathbf{X})$ and $t_2(\mathbf{X})$ together from Lemmas 3.10 and 3.11, with probability at least $1 - e^{-\frac{\tau^2}{2}}$ we have:

$$\inf_{\delta \in \mathcal{H}} \frac{1}{\sqrt{n}} \|\mathbf{X}\boldsymbol{\delta}\|_2 \leq \sum_{g=0}^{G} \frac{n_g}{n} \psi_{\mathcal{I}} \xi \|\boldsymbol{\delta}_g\|_2 q - \frac{2}{\sqrt{n}} \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} k_x c_g \omega(\mathcal{A}_g) \|\boldsymbol{\delta}_g\|_2 - \frac{\tau}{\sqrt{n}}$$

$$= n^{-1} \sum_{g=0}^{G} n_g \|\boldsymbol{\delta}_g\|_2 (\psi_{\mathcal{I}} \xi q - 2 c_g k_x \frac{\omega(\mathcal{A}_g)}{\sqrt{n_g}}) - \frac{\tau}{\sqrt{n}}$$

$$(\kappa_g = \psi_{\mathcal{I}} \xi q - \frac{2 c_g k_x \omega(\mathcal{A}_g)}{\sqrt{n_g}}) = \sum_{g=0}^{G} \frac{n_g}{n} \|\boldsymbol{\delta}_g\|_2 \kappa_g - \frac{\tau}{\sqrt{n}}$$

$$\geq \kappa_{\min} \sum_{g=0}^{G} \frac{n_g}{n} \|\boldsymbol{\delta}_g\|_2 - \frac{\tau}{\sqrt{n}}$$

$$(\boldsymbol{\delta} \in \mathcal{H}) = \kappa_{\min} - \frac{\tau}{\sqrt{n}}$$

where $\kappa_{\min} = \min_{g \in [G]} \kappa_g$. To conclude the proof, take $\tau = \sqrt{n}\kappa_{\min}/2$.

To satisfy the RE condition all κ_gs should be bounded away from zero. To this end we need the following sample complexities $\forall g \in [G_+] : \left(\frac{2c_g k}{\psi_{\mathcal{I}} \xi q}\right)^2 \omega(\mathcal{A}_g)^2 \leq n_g$ where by taking $\xi = \frac{\alpha}{6}$ simplifies to:

$$\forall g \in [G_+] : \quad O\left(k^6 \psi_{\mathcal{I}}^{-2} \alpha^{-6} \omega(\mathcal{A}_g)^2\right) \leq n_g$$

4. General Error Bound.

In this section, we present our main statistical result which is a non-asymptotic high probability upper bound for the estimation error of the common and individual parameters.

Theorem 4.1. *For $\mathbf{x}_{gi}$ and w_{gi} described in Definition 3.1 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 for estimator (2.1) with probability at least $1 - \sigma \exp\left(-\min\left[\nu \min_{g \in [G]} n_g - \log(G+1), \tau^2\right]\right)$:*

(4.1)
$$\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g\|_2 \leq C\gamma \frac{\max_{g \in [G_+]} \omega(\mathcal{A}_g) + \sqrt{\log(G+1)} + \tau}{\kappa_{\min}^2 \sqrt{n}}$$

where $\gamma = \max_{g \in [G]} n/n_g$, $\tau > 0$, and σ, ν, and C are constants.

Corollary 4.2. *From Theorem 4.1 one can immediately entail the error bound for estimation of the common and individual parameters as follows:*

$$\forall g \in [G_+] : \quad \|\boldsymbol{\delta}_g\|_2 = O\left(\gamma \frac{\max_{g \in [G_+]} \omega(\mathcal{A}_g) + \sqrt{\log(G+1)}}{\sqrt{n_g}}\right)$$

Example 4.3. **(l_1-norm)** For sparse DE estimator of Example 2.1, results of Theorems 3.7 and 4.1 translates to the following: For enough number of samples as $\forall g \in [G_+] : n_g \geq m_g = O(s_g \log p)$, the upper bound of error simplifies to:

(4.2)
$$\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g\|_2 = O\left(\gamma \sqrt{\frac{(\max_{g \in [G_+]} s_g) \log p}{n}}\right)$$

Therefore, individual errors are bounded as $\|\boldsymbol{\delta}_g\|_2 = O(\gamma\sqrt{(\max_{g\in[G]} s_g)\log p/n_g})$ which is slightly worse than $O(\sqrt{s_g \log p/n_g})$, the well-known error bound for recovering an s_g-sparse vector from n_g observations using LASSO or similar estimators [2, 4, 8, 12, 13].

4.1. Proof of Theorem 4.1. To avoid cluttering the notation, we rename the vector of all noises as $\mathbf{w}_0 \triangleq \mathbf{w}$. First, we massage the deterministic upper bound of Theorem 2.2 as follows:

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

Assume $q_g = \langle \mathbf{X}_g^T\frac{\mathbf{w}_g}{\|\mathbf{w}_g\|_2}, \frac{\boldsymbol{\delta}_g}{\|\boldsymbol{\delta}_g\|_2}\rangle\sqrt{\frac{n}{n_g}}\|\mathbf{w}_g\|_2$ and $p_g = \sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g\|_2$. Then the above term is the inner product of two vectors $\mathbf{p} = (p_0, \dots, p_G)$ and $\mathbf{q} = (q_0, \dots, q_G)$ for which we have:

$$\sup_{\mathbf{p}\in\bar{\mathcal{H}}}\mathbf{p}^T\mathbf{q} = \sup_{\|\mathbf{p}\|_1=1}\mathbf{p}^T\mathbf{q} \le \|\mathbf{q}\|_\infty = \max_{g\in[G_+]} q_g,$$

where the inequality holds because of the definition of the dual norm. Now we can go back to the original form:

$$\begin{aligned}
\text{(4.3)}\qquad \sup_{\boldsymbol{\delta}\in\mathcal{H}}\mathbf{w}^T\mathbf{X}\boldsymbol{\delta} &\le \max_{g\in[G]}\langle \mathbf{X}_g^T\frac{\mathbf{w}_g}{\|\mathbf{w}_g\|_2}, \frac{\boldsymbol{\delta}_g}{\|\boldsymbol{\delta}_g\|_2}\rangle\sqrt{\frac{n}{n_g}}\|\mathbf{w}_g\|_2 \\
&\le \max_{g\in[G]}\sqrt{\frac{n}{n_g}}\|\mathbf{w}_g\|_2 \sup_{\mathbf{u}_g\in\mathcal{C}_g\cap\mathbb{S}^{p-1}}\langle \mathbf{X}_g^T\frac{\mathbf{w}_g}{\|\mathbf{w}_g\|_2}, \mathbf{u}_g\rangle
\end{aligned}$$

To avoid cluttering we define a random quantity $h_g(\mathbf{w}_g, \mathbf{X}_g) \triangleq \|\mathbf{w}_g\|_2\sup_{\mathbf{u}_g\in\mathcal{A}_g}\langle \mathbf{X}_g^T\frac{\mathbf{w}_g}{\|\mathbf{w}_g\|_2}, \mathbf{u}_g\rangle$ and a corresponding constant $e_g(\tau) \triangleq c_g\sqrt{(2k_w^2+1)k_x^2 n_g}\left(\omega(\mathcal{A}_g) + \sqrt{\log(G+1)} + \tau\right)$. Then from (4.3), we have:

$$\begin{aligned}
\mathbb{P}\left(\sup_{\boldsymbol{\delta}\in\mathcal{H}}\mathbf{w}^T\mathbf{X}\boldsymbol{\delta} > \max_{g\in[G]}\sqrt{\frac{n}{n_g}}e_g(\tau)\right) &\le \mathbb{P}\left(\max_{g\in[G]}\sqrt{\frac{n}{n_g}}h_g(\mathbf{w}_g, \mathbf{X}_g) > \max_{g\in[G]}\sqrt{\frac{n}{n_g}}e_g(\tau)\right) \\
\text{(Union Bound)} &\le \sum_{g=0}^{G}\mathbb{P}\left(\sqrt{\frac{n}{n_g}}h_g(\mathbf{w}_g, \mathbf{X}_g) > \max_{g\in[G]}\sqrt{\frac{n}{n_g}}e_g(\tau)\right) \\
&\le \sum_{g=0}^{G}\mathbb{P}\left(h_g(\mathbf{w}_g, \mathbf{X}_g) > e_g(\tau)\right) \\
&\le (G+1)\max_{g\in[G_+]}\mathbb{P}\left(h_g(\mathbf{w}_g, \mathbf{X}_g) > e_g(\tau)\right) \\
&\le \sigma\exp\left(-\min\left[\nu\min_{g\in[G]} n_g - \log(G+1), \tau^2\right]\right)
\end{aligned}$$

where the last inequality is the result of the following lemma:

Algorithm 5.1 DICER

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

Lemma 4.4. *For $\mathbf{x}_{gi}$ and ω_{gi} defined in Definition 3.1 and $\tau > 0$, with probability at least $1 - \frac{\sigma_g}{(G+1)} \exp\left(-\min\left[\nu n_g - \log(G+1), \tau^2\right]\right)$ we have:*

$$\|\mathbf{w}_g\|_2 \sup_{\mathbf{u}_g \in \mathcal{A}_g} \langle \mathbf{X}_g^T \frac{\mathbf{w}_g}{\|\mathbf{w}_g\|_2}, \mathbf{u}_g \rangle \leq c_g \sqrt{(2k_w^2 + 1) k_x^2 n_g} \left(\omega(\mathcal{A}_g) + \sqrt{\log(G+1)} + \tau \right),$$

where σ_g, ν and c_g are constants.

The proof completes by replacing $\max_{g \in [G]} \sqrt{\frac{n}{n_g}} e_g(\tau)$ as the upper bound of $\sup_{\boldsymbol{\delta} \in \mathcal{H}} \mathbf{w}^T \mathbf{X} \boldsymbol{\delta}$ and $\kappa_{\min}^2/4$ as the lower bound of κ (from Theorem 3.7) both into the deterministic bound of Theorem 2.2 .

5. Estimation Algorithm. We propose *Data enrIChER* (DICER) a projected block gradient descent algorithm, Algorithm 5.1, 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.

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

Definition 5.1. *We define the following positive constants as functions of step sizes $\mu_g > 0$:*

$$\forall g \in [G_+] : \rho_g(\mu_g) = \sup_{\mathbf{u}, \mathbf{v} \in \mathcal{B}_g} \mathbf{v}^T \big(\mathbf{I}_g - \mu_g \mathbf{X}_g^T \mathbf{X}_g \big) \mathbf{u},$$

$$\eta_g(\mu_g) = \mu_g \sup_{\mathbf{v} \in \mathcal{B}_g} \mathbf{v}^T \mathbf{X}_g^T \frac{\mathbf{w}_g}{\|\mathbf{w}_g\|_2},$$

$$\forall g \in [G] : \phi_g(\mu_g) = \mu_g \sup_{\mathbf{v} \in \mathcal{B}_g, \mathbf{u} \in \mathcal{B}_0} -\mathbf{v}^T \mathbf{X}_g^T \mathbf{X}_g \mathbf{u},$$

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 5.1 where to simplify the notation we drop μ_g arguments.

Theorem 5.2. *For Algorithm 5.1 initialized by $\boldsymbol{\beta}^{(1)} = \mathbf{0}$, we have the following deterministic bound for the error at iteration $t+1$:*

$$\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g^{(t+1)}\|_2 \leq \rho^t \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\beta}_g^*\|_2 + \frac{1-\rho^t}{1-\rho} \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \eta_g \|\boldsymbol{\omega}_g\|_2, \tag{5.1}$$

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

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

Lemma 5.3. *We have the following recursive dependency between the error of $t+1$th iteration and tth iteration of* DICER*:*

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

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

By recursively applying results of Lemma 5.3, we get the following deterministic bound which depends on constants defined in Definition 5.1:

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

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:

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

$$(\boldsymbol{\beta}^1 = 0) \leq \rho^t\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\|\boldsymbol{\beta}_g^*\|_2 + \frac{1-\rho^t}{1-\rho}\sum_{g=0}^{G}\sqrt{\frac{n_g}{n}}\xi_g\|\boldsymbol{\omega}_g\|_2 \tag{5.2}$$

∎

The RHS of (5.2) consists of two terms. If we keep $\rho < 1$, the first term approaches zero fast, and the second term determines the bound. In the following, we show that for specific choices of step sizes μ_gs we can keep $\rho < 1$ with high probability and 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 of the same order of the upper bound of the statistical error characterized in Theorem 4.1.

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 subsection SM1.2) 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. The high probability bounds for constants in Definition 5.1 and the deterministic bound of Theorem 5.2 leads to the following theorem which shows that for enough number of samples, of the same order as the statistical sample complexity of Theorem 3.7, we can keep ρ below one and have geometric convergence.

Theorem 5.4. *Let $\tau = \sqrt{\log(G+1)}/\zeta + \epsilon$ for $\epsilon, \zeta > 0$. For the step sizes of:*

$$\mu_0 = \frac{\min_{g\in[G]} h_g(\tau)^{-2}}{4n}, \forall \in [G] : \mu_g = \frac{h_g(\tau)^{-1}}{2\sqrt{nn_g}}$$

where $h_g(\tau) = \left(1 + c_{0g}\frac{\omega(\mathcal{A}_g)+\omega(\mathcal{A}_0)+2\tau}{\sqrt{n_g}}\right)$ and sample complexities of $\forall g \in [G_+] : n_g \geq C_g(\omega(\mathcal{A}_g) + \tau)^2$, with probability at least $1 - \sigma\exp(-\min(\nu \min_{g\in[G]} n_g - \log(G+1), \zeta\epsilon^2))$ updates of Algorithm 5.1 obey the following:

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

where $r(\tau) < 1$ and υ, ζ, and σ are constants.

Corollary 5.5. *For enough number of samples, iterations of DE algorithm with step sizes $\mu_0 = \Theta(\frac{1}{n})$ and $\mu_g = \Theta(\frac{1}{\sqrt{nn_g}})$ geometrically converges to the following with high probability:*

$$\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g^{\infty}\|_2 \leq c\frac{\max_{g\in[G_+]} \omega(\mathcal{A}_g) + \sqrt{\log(G+1)}/\zeta + \theta}{\sqrt{n}(1-r(\tau))} \tag{5.3}$$

where $c = C(G+1)\sqrt{(2k_w^2+1)k_x^2}$.

It is instructive to compare RHS of (5.3) with that of Theorem 4.1: $\kappa_{\min}$ defined in Theorem Theorem 3.7 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 5.5 shows that with the number of samples in the order of sample complexity determined in Theorem 3.7 DICER converges to the statistical error bound determined in Theorem 4.1.

5.1. Proof Sketch of Theorem 5.4.

We want to determine $r(\tau) < 1$ such that $\rho < r(\tau)$ with high probability. Here, we provide a proof sketch using the below probabilistic bounds on constants of Definition 5.1 while ignoring detailed computation of subsequent probabilities in finding $r(\tau)$. The full probabilistic proof is provided in subsection SM1.2. First we need the following lemma to upper bound constants of Definition 5.1:

Lemma 5.6. *Consider $a_g \geq 1$ the following upper bounds hold:*

$$\rho_g\left(\frac{1}{a_g n_g}\right) \leq \frac{1}{2}\left[\left(1-\frac{1}{a_g}\right)+\sqrt{2}c_g\frac{2\omega_g+\tau}{a_g\sqrt{n_g}}\right], \quad \text{w.p. at least} \quad 1-2\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right)$$

$$\eta_g\left(\frac{1}{a_g n_g}\right) \leq \frac{c_g k_x(\omega_g+\tau)}{a_g n_g}, \quad \text{w.p. at least} \quad 1-\pi_g\exp\left(-\tau^2\right)$$

$$\phi_g\left(\frac{1}{a_g n_g}\right) \leq \frac{1}{a_g}\left(1+c_{0g}\frac{\omega_{0g}+2\tau}{\sqrt{n_g}}\right), \quad \text{w.p. at least} \quad 1-2\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right)$$

where $\omega_g = \omega(\mathcal{A}_g)$ and $\omega_{0g} = \omega(\mathcal{A}_g)+\omega(\mathcal{A}_0)$.

To keep $\rho < 1$ in the deterministic bound of Theorem 5.2 with the step sizes $\mu_g = \frac{1}{n_g a_g}$ we need to find the number of samples which satisfy the following conditions:

- Condition 1: $\rho_0(\mu_0) + \sum_{g=1}^G \sqrt{\frac{n_g}{n}}\phi_g(\mu_g) < 1$
- Condition 2: $\forall g \in [G] : \rho_g(\mu_g) + \sqrt{\frac{n}{n_g}\frac{\mu_0}{\mu_g}}\phi_g(\mu_g) < 1$

where according to the step sizes determine in the Theorem $a_0 \triangleq (4n\max_{g\in[G]}(1+c_{0g}\frac{\omega_{0g}+2\tau}{\sqrt{n_g}})^2)^{-1}$ and $a_g \triangleq (2\sqrt{n/n_g}(1+c_{0g}\frac{\omega_{0g}+2\tau}{\sqrt{n_g}}))^{-1}$. Condition 1 requires $\rho_0 + \sum_{g=1}^G \sqrt{\frac{n_g}{n}}\phi_g$ to be strictly below 1 which is equivalent to:

$$\begin{aligned}
\rho_0(\mu_0) + \sum_{g=1}^G \sqrt{\frac{n_g}{n}}\phi_g(\mu_g) &\leq \frac{1}{2}\left[\left(1-\frac{1}{a_0}\right)+\sqrt{2}c_0\frac{2\omega_0+\tau}{a_0\sqrt{n}}\right]+\frac{1}{2}\sum_{g=1}^G\frac{2}{a_g}\sqrt{\frac{n_g}{n}}\left(1+c_{0g}\frac{\omega_{0g}+2\tau}{\sqrt{n_g}}\right)\\
(\text{Substitute } a_g) &= \frac{1}{2}\left[\left(1-\frac{1}{a_0}\right)+\sqrt{2}c_0\frac{2\omega_0+\tau}{a_0\sqrt{n}}\right]+\frac{1}{2}\sum_{g=1}^G\frac{n_g}{n}\\
&= \frac{1}{2}\left[\left(2-\frac{1}{a_0}\right)+\sqrt{2}c_0\frac{2\omega_0+\tau}{a_0\sqrt{n}}\right] < 1
\end{aligned}$$

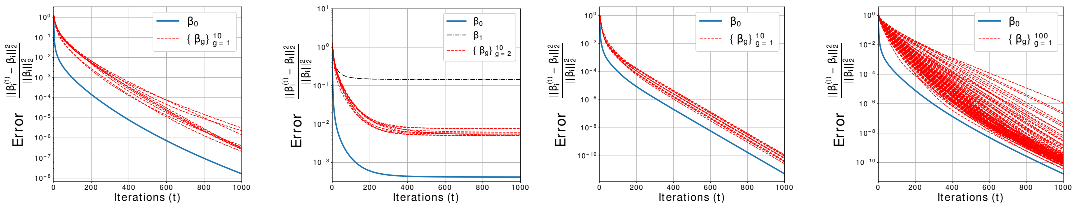
So Condition 1 reduces to $n > 8c_0^2(\omega(\mathcal{A}_0)+\tau)^2$.

Secondly in Condition 2, we want to bound all of $\rho_g + \mu_0\sqrt{\frac{n}{n_g}}\frac{\phi_g}{\mu_g}$ terms for $\mu_g = \frac{1}{a_g n_g}$ by 1:

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

Condition 2 becomes:

$$\begin{aligned}
\sqrt{2}c_g\frac{2\omega_g+\tau}{\sqrt{n_g}} &\leq 1+a_g-\sqrt{\frac{n_g}{n}}\frac{2a_g}{a_0}\left(1+c_{0g}\frac{\omega_{0g}+2\tau}{\sqrt{n_g}}\right)\\
(\text{Substitute } a_g) &= 1+a_g-\frac{4}{a_0}\left(1+c_{0g}\frac{\omega_{0g}+2\tau}{\sqrt{n_g}}\right)^2\\
(\text{Substitute } a_0) &\leq 1+a_g
\end{aligned}$$



(a) (b) (c) (d)

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 $G = 100$.

So the sample complexity should be $\sqrt{n_g} > \frac{\sqrt{2}c_g(2\omega_g+2\tau)}{1+a_g}$ and since $a_g > 1$, the final per group sample complexity should be $n_g > 8c_g(\omega(\mathcal{A}_g) + \tau)^2$.

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}.

Acknowledgments.

REFERENCES

[1] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, et al., *Optimization with sparsity-inducing penalties*, Foundations and Trends® in Machine Learning, 4 (2012), pp. 1–106.

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

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

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

[5] T. Blumensath and M. E. Davies, *Iterative hard thresholding for compressed sensing*, Applied and computational harmonic analysis, 27 (2009), pp. 265–274.

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

[7] P. T. Boufounos and R. G. Baraniuk, *1-bit compressive sensing*, in Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on, IEEE, 2008, pp. 16–21.

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

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

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

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

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

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

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

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

[16] F. Dondelinger, S. Mukherjee, and Alzheimer's Disease Neuroimaging Initiative, *The joint lasso: high-dimensional regression for group structured data*, Biostatistics, (2018).

[17] D. L. Donoho, *Compressed sensing*, IEEE Transactions on information theory, 52 (2006), pp. 1289–1306.

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

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

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

[21] F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. Goncalves, S. Barthorpe, H. Lightfoot, et al., *A landscape of pharmacogenomic interactions in cancer*, Cell, 166 (2016), pp. 740–754.

[22] P. Jain, P. Netrapalli, and S. Sanghavi, *Low-rank matrix completion using alternating minimization*, in Proceedings of the forty-fifth annual ACM symposium on Theory of computing, ACM, 2013, pp. 665–674.

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

[24] M. B. McCoy and J. A. Tropp, *The achievable performance of convex demixing*, (2013), https://arxiv.org/abs/1309.

7478.
[25] S. Mendelson, *Learning without concentration*, Journal of the ACM, 62 (2015), pp. 21:1–21:25.
[26] S. Negahban and M. J. Wainwright, *Restricted strong convexity and weighted matrix completion: Optimal bounds with noise*, Journal of Machine Learning Research, 13 (2012), pp. 1665–1697.
[27] S. Negahban, B. Yu, M. J. Wainwright, and P. K. Ravikumar, *A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers*, in Advances in Neural Information Processing Systems, 2009, pp. 1348–1356.
[28] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, *A Unified Framework for High-Dimensional Analysis of \$M\$-Estimators with Decomposable Regularizers*, Statistical Science, 27 (2012), pp. 538–557.
[29] E. Ollier and V. Viallon, *Joint estimation of K related regression models with simple L_1-norm penalties*, (2014), https://arxiv.org/abs/1411.1594.
[30] E. Ollier and V. Viallon, *Regression modeling on stratified data with the lasso*, (2015), https://arxiv.org/abs/1508.05476.
[31] Y. Plan, R. Vershynin, and E. Yudovina, *High-dimensional estimation with geometric constraints*, Information and Inference: A Journal of the IMA, 6 (2017), pp. 1–40.
[32] G. Raskutti, M. J. Wainwright, and B. Yu, *Restricted eigenvalue properties for correlated gaussian designs*, Journal of Machine Learning Research, 11 (2010), pp. 2241–2259.
[33] M. Rudelson and S. Zhou, *Reconstruction from anisotropic random measurements*, IEEE Transactions on Information Theory, 59 (2013), pp. 3434–3447.
[34] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological), (1996), pp. 267–288.
[35] J. A. Tropp, *Convex recovery of a structured signal from independent random linear measurements*, in Sampling Theory, a Renaissance, Springer, 2015, pp. 67–101.
[36] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, in Compressed Sensing, Cambridge University Press, Cambridge, 2012, pp. 210–268.
[37] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge University Press, 2018.
[38] E. Yang and P. K. Ravikumar, *Dirty statistical models*, in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds., Curran Associates, Inc., 2013, pp. 611–619.
[39] Y. Zhang and Q. Yang, *A survey on Multi-Task learning*, (2017), https://arxiv.org/abs/1707.08114.