High Dimensional Data Sharing: Multi-Task Learning with Theoretical Guarantee

Amir Asiaee, Samet Oymak, Kevin R. Coombes, and Arindam Banerjee

***Abstract*—We consider the problem of multi-task learning in high dimension. In particular, we introduce an estimator and investigate its statistical and computational properties for the problem of multiple connected linear regressions known as Data Sharing. The between-tasks connections are captured by a cross-tasks *common parameter* which gets refined by per-task *individual parameters*. Any convex function, e.g., norm, can characterize the structure of both common and individual parameters.We delineate the sample complexity of our estimator and provide high probability non-asymptotic bound for estimation error of all parameters under a geometric condition. We show that the recovery of the common parameter benefits from *all* of the pooled samples. We propose an iterative estimation algorithm with a geometric convergence rate and supplement our theoretical analysis with experiments on synthetic data. Overall, we present a first through statistical and computational analysis of inference in the data sharing model.**

***Index Terms*—Multi-task learning, superposition models, high-dimensional statistics, convergence rate analysis.**

I. 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 [1], [2], [3]. 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 [4], [5], [6]. 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 [7], [8], broad families of semi-parametric and single-index models [9], [10], non-convex models [11], [12], 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 the 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 [13], [14].

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 [15], [16], [17]. Such a collection of *coupled* superposition models is known by multiple names. It is a form of multi-task learning [18], [19] when we consider regression in each group as a task. It is also called data sharing [20] since information contained in different groups is shared through the common parameter $\boldsymbol{\beta}_0^*$. Finally, it has been called data enrichment [21], [22], [23] 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 [24], [20], [25], [26], we consider the below *data sharing* (DS) model:

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

where g and i index the group and samples respectively. The DS model (1) assumes that there is a *common* parameter $\boldsymbol{\beta}_0^*$

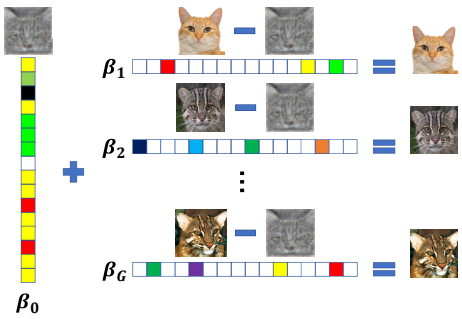


Fig. 1: A conceptual illustration of data sharing model for learning representation of cat species. The common parameter $\boldsymbol{\beta}_0$ captures a *generic cat* which consists of shared features among all cats.

Manuscript received January 25, 2020; revised January 26, 2020. The research was supported in part by NSF grants OAC-1934634, IIS-1908104, IIS-1563950, IIS-1447566, IIS-1447574, IIS-1422557, and CCF-1451986. The work of S. Oymak is partially supported by the NSF award CNS-1932254. The work of A. Asiaee was partially supported by MBI (Mathematical Biosciences Institute) which receives its funding through NSF grant DMS-1440386.

A. Asiaee is with Mathematical Biosciences Institute The Ohio State University, Columbus, OH, 43210 USA (e-mail: asiaeetaheri.1@osu.edu)

S. Oymak is with Department of Electrical and Computer Engineering, UC Riverside, Riverside, CA, 92507 USA (e-mail: oymak@ece.ucr.edu)

K. R. Coombes is with Departments of Biomedical Informatics, The Ohio State University, Columbus, OH, 43210 USA (e-mail: coombes.3@osu.edu)

A. Banerjee is with Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, 55455 USA (e-mail: banerjee@cs.umn.edu)

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, Fig. 1.

Our goal is to design an estimation procedure which consistently recovers all parameters of DS (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 [27], [28], [29], [30], [31].

A. 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 $\mathbf{B}_1 = [\boldsymbol{\beta}_{11}, \ldots, \boldsymbol{\beta}_{1G}]$ and $\mathbf{B}_2 = [\boldsymbol{\beta}_{21}, \ldots, \boldsymbol{\beta}_{2G}]$ are two parameter matrices [19]. To capture the relation between tasks, different types of constraints are assumed for parameter matrices. For example, [32] 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 [18] where authors regularize the regression with $\|\mathbf{B}_1\|_{1,\infty}$ and $\|\mathbf{B}_2\|_{1,1}$ where norms are i,j-norms on *rows*, $\mathbf{b}$, of matrices, i.e., $\|\mathbf{B}\|_{i,j} = \|(\|\mathbf{b}_1\|_j, \ldots, \|\mathbf{b}_p\|_j)\|_i$ and the norms are defined as $\|\mathbf{b}\|_i = \left(\sum_{g=1}^G |b_g|^i\right)^{1/i}$ and $\|\mathbf{b}\|_\infty = \max_{g\in G} |b_g|$.

If in our DS 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 DS 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 DS we have a single common parameter $\boldsymbol{\beta}_0^*$. So one would expect that DS 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 [18], Table I. Mainly, DSM requires every task i to have large enough samples to learn its own common parameters $\boldsymbol{\beta}_i$ but since DS shares the common parameter it only requires the *total dataset over all tasks* to be sufficiently large.

The linear DS 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 [24], sentiment analysis, banking strategy [20], single cell data analysis [26], road safety [25], and disease subtype analysis [24]. More generally, in any high-dimensional problem where the population consists of groups, data sharing 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 sharing framework to different domains, limited advances have been made in understanding the statistical and computational properties of suitable estimators for the DS model (1). In fact, non-asymptotic statistical properties, including sample complexity and statistical rates of convergence, of regularized estimators for the data sharing model is still an open question [20], [25]. To the best of our knowledge, the only theoretical guarantee for data sharing is provided in [26] 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 [26], sample complexity and l_2 consistency results [18] 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.

B. 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 sharing model takes the following vector form:

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

where each row of $\mathbf{X}_g$ is $\mathbf{x}_{gi}^T$ and $\mathbf{w}_g^T = (w_{g1}, \ldots, w_{gn_g})$ 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 $\mathbf{X}_0 \triangleq [\mathbf{X}_1^T, \ldots, \mathbf{X}_G^T]^T$.

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 $\|V\|_{\psi_2}$ [33]. 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 [33] 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 $\omega(\mathcal{V}) = \mathbb{E}_{\mathbf{g}}[\sup_{\mathbf{u}\in\mathcal{V}}\langle \mathbf{g}, \mathbf{u}\rangle]$ [34], 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 $\xi > 0$.

C. Our Contributions

We propose the following Data Shared (DS) estimator $\hat{\boldsymbol{\beta}}$ for recovering the structured parameters where the structure is

induced by *convex* functions $f_g(\cdot)$:

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

We present several statistical and computational results for the DS estimator (3):

- The DS estimator (3) succeeds if a geometric condition that we call *DAta SHaring Incoherence conditioN* (DASHIN) is satisfied, Fig. 2b. Compared to other known geometric conditions in the literature such as structural coherence [15] and stable recovery conditions [16], DASHIN is a considerably weaker condition, Fig. 2a.
- Assuming DASHIN 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:

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

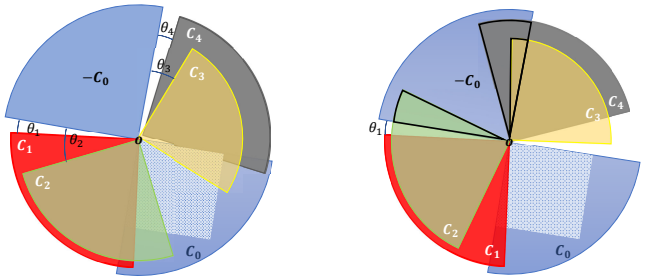
  where $n_0 \triangleq n$ is the total number of samples, $\gamma \triangleq \max_{g\in[G]} \frac{n}{n_g}$ is the *sample condition number*, and $\mathcal{C}_g$ is the error cone corresponding to $\boldsymbol{\beta}_g^*$ exactly defined in Section II. To the best of our knowledge, this is the first statistical estimation guarantee for the data sharing.
- We also establish the sample complexity of the DS 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 DS estimator (3) benefits from *all* of the n pooled samples.
- We present an efficient projected block gradient descent algorithm DASHER, 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-shared 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 II. Then in Section III, we discuss the restricted eigenvalue condition and calculate the sample complexity required for the recovery of the true parameters by our estimator under DASHIN condition. We close the statistical analysis in Section IV by providing non-asymptotic high probability error bound for parameter recovery. We delineate our geometrically convergent algorithm, DASHER in Section V and finally supplement our work with experiments on synthetic data in Section VI.

II. The Data Shared Estimator

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

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \text{s.t. } \forall g \in [G_+] : f_g(\boldsymbol{\beta}_g) \le f_g(\boldsymbol{\beta}_g^*), \tag{5}$$



(a) Structural Coherence (SC) condition.

(b) DAta SHaring Incoherence conditioN (DASHIN).

Fig. 2: Comparison of geometric recovery condition for superposition models known as Structural Coherence (SC) [15] and our DASHIN recovery condition for data sharing model which is a system of coupled superposition models (2). For each parameter $\boldsymbol{\beta}_g^* \in [G]$, $\mathcal{E}_g = \{\boldsymbol{\delta}_g | f_g(\boldsymbol{\beta}_g^* + \boldsymbol{\delta}_g) \le f_g(\boldsymbol{\beta}_g^*)\}$ is the error set and $\mathcal{C}_g = \text{Cone}(\mathcal{E}_g)$ is the error cone. For all i, j, SC requires $-\mathcal{C}_i \cap \mathcal{C}_j = \{0\}$. In panel (a) we only show this condition for $i = 0$, i.e., $-\mathcal{C}_0 \cap \mathcal{C}_j = \{0\}$ where all $\theta_j > 0$. (b) DASHIN only needs one of the $\mathcal{C}_g, g \in [G]$, does not intersect with the inverse of the common parameter error cone $-\mathcal{C}_0$. In panel (b) $-\mathcal{C}_0 \cap \mathcal{C}_1 = \{0\}$ is enough for recovering all parameters.

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

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

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

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \text{s.t. } \forall g \in [G_+] : \|\boldsymbol{\beta}_g\|_1 \le \|\boldsymbol{\beta}_g^*\|_1. \tag{7}$$

We call (1) *sparse DS* 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 (5), the error vector $\boldsymbol{\delta}_g$ will belong to the following restricted error set:

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

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

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

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

Starting from the optimality of $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + \boldsymbol{\delta}$ as $\frac{1}{n}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2 \le \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2$, we have: $\frac{1}{n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 \le \frac{1}{n}2\mathbf{w}^T\mathbf{X}\boldsymbol{\delta}$ where $\mathbf{w} = [\mathbf{w}_1^T, \dots, \mathbf{w}_G^T]^T \in \mathbb{R}^n$ is the vector of all noises.

Using this basic inequality, we can establish the following deterministic error bound.

Theorem 1: For the DS 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 $\gamma = \max_{g\in[G]} \frac{n}{n_g}$, 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 $\mathcal{H}$ (9) we have:

$$\begin{aligned}
\frac{1}{n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 &\geq \frac{1}{n} \inf_{\mathbf{u}\in\mathcal{H}} \|\mathbf{X}\mathbf{u}\|_2^2 \left(\sum_{g=0}^{G} \frac{n_g}{n}\|\boldsymbol{\delta}_g\|_2\right)^2 \geq \kappa \left(\sum_{g=0}^{G} \frac{n_g}{n}\|\boldsymbol{\delta}_g\|_2\right)^2 \\
&\geq \kappa \left(\min_{g\in[G]} \sqrt{\frac{n_g}{n}}\right)^2 \left(\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g\|_2\right)^2 \qquad (10) \\
&= \kappa \left(\min_{g\in[G]} \frac{n_g}{n}\right) \left(\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g\|_2\right)^2
\end{aligned}$$

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:

$$\frac{2}{n}\mathbf{w}^T\mathbf{X}\boldsymbol{\delta} \leq \frac{2}{n} \sup_{\mathbf{u}\in\bar{\mathcal{H}}} \mathbf{w}^T\mathbf{X}\mathbf{u} \left(\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g\|_2\right), \mathbf{u} \in \mathcal{H} \quad (11)$$

Putting together inequalities (10) and (11) completes the proof. ∎

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

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

III. Restricted Eigenvalue Condition

The main assumption of Theorem 1 is known as Restricted Eigenvalue (RE) condition in the literature of high-dimensional statistics [35], [36], [37]: $\inf_{\mathbf{u}\in\mathcal{H}} \frac{1}{n}\|\mathbf{X}\mathbf{u}\|_2^2 \geq \kappa > 0$. 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 SHaring Incoherence conditioN* (DASHIN) 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 [15], [38], [39], 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 $\|\mathbf{x}\|_{\psi_2} \leq k_x$. 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 $\|\mathbf{w}_{gi}\|_{\psi_2} \leq k_w$.

[1]Extension to an-isotropic sub-Gaussian case is straightforward by techniques developed in [35], [40].

A. Geometric Condition for 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 the error cones 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 interactions of error cones, 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 [15], [16] which is the generalization of this idea for superposition of multiple parameters as $\sum_{g=0}^{G} \boldsymbol{\beta}_g^*$.

Definition 2 (Structural Coherence (SC) [15], [16]): Consider a superposition model of the form $y = \mathbf{x}^T \sum_{g=0}^{G} \boldsymbol{\beta}_g^* + w$. The SC condition requires that $\forall\boldsymbol{\delta}_g \in \mathcal{C}_g, \exists\lambda$ s.t. $\|\sum_{g=0}^{G} \boldsymbol{\delta}_g\|_2 \geq \lambda \sum_{g=0}^{G} \|\boldsymbol{\delta}_g\|_2$, and leads to the RE condition $\frac{1}{\sqrt{n}}\|\mathbf{X}\sum_{g=1}^{G} \boldsymbol{\delta}_g\|_2 \geq \kappa \sum_{g=1}^{G} \|\boldsymbol{\delta}_g\|_2$.

Remark 2: Note that the SC condition is satisfied if none of the individual error cones $\mathcal{C}_g$ intersect with the inverted error cone $-\mathcal{C}_0$ [15], [39], i.e., $\forall g, \theta_g > 0$ in Fig. 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 DASHIN, a considerably weaker geometric condition compared to SC which leads to recovery of all parameters in the data sharing model.

Definition 3 (DAta SHaring Incoherence conditioN (DASHIN)): 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.

Remark 3: DASHIN is a refinement of SC for the specific problem of data sharing, i.e., system of coupled superposition model each with two components. DASHIN holds even if only one of the $\mathcal{C}_g$s does not intersect with $-\mathcal{C}_0$. More specifically, DASHIN holds if $\exists g, \theta_g > 0$ in Fig. 2b which 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.

B. Sample Complexity

An alternative to our DS estimator (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., $\exists g, -\mathcal{C}_0 \cap \mathcal{C}_g = \{0\}$, one can

	GI-LASSO	Dirty Stat. Model	Plugin Superposition	Sparse DS
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 I: Comparison of the order of per group number of samples (sample complexities) of various methods for recovering sparse DS parameters. Let $s_{0g}=|\text{support}(\boldsymbol{\beta}_0^*+\boldsymbol{\beta}_g^*)|$ be the superimposed support where $s_0,s_g\leq\max(s_0,s_g)\leq s_{0g}$.

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 which it seems that DASHIN has no advantage over SC. But the disadvantage of plugin superposition estimator is that it fails to utilize the true coupling structure in the data sharing model, where $\boldsymbol{\beta}_0^*$ is involved in all groups. In fact, below we show that the plugin superposition estimator under SC condition leads to a pessimistic sample complexity for $\boldsymbol{\beta}_0^*$ recovery.

Proposition 2: Assume observations distributed as defined in 1 and pair-wise SC conditions are satisfied. Consider each superposition model (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 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 sharing model, we hope to be able to get a better sample complexity specifically for the common parameter $\boldsymbol{\beta}_0$. Using DASHIN and the small ball method [38], a tool from empirical process theory in the following theorem, we get a better sample complexity required for satisfying the RE condition:

Theorem 3: Let $\mathbf{x}_{gi}$s be random vectors defined in Definition 1. Assume DASHIN condition of Definition 3 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 $\kappa_{\min}=\min_{g\in[G_+]}C\psi_{\mathcal{I}}\frac{\alpha^3}{k_x^2}-\frac{2c_gk_x\omega(\mathcal{A}_g)}{\sqrt{n_g}}$.

Remark 4: Note that $\kappa=\frac{\kappa_{\min}^2}{4}$ is the lower bound of the RE condition of Theorem 1, 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 RE condition.

Example 2: **(l_1-norm)** The Gaussian width of the spherical cap of a p-dimensional s-sparse vector is $\omega(\mathcal{A})=\Theta(\sqrt{s\log p})$ [35], [34]. Therefore, the number of samples per group and total required for satisfaction of the RE condition in the sparse DS estimator Example 1 is $\forall g\in[G]: n_g\geq m_g=\Theta(s_g\log p)$. Table I compares sample complexities of the sparse DS estimator with three baselines: plugin superposition estimator of Proposition 2, G Independent LASSO (GI-LASSO), and Jalali's Dirty Statistical Model (DSM) [18]. Note that GI-LASSO does not recover the common parameter and DSM needs all groups have same number of samples.

C. Proof of Theorem 3

Let's simplify the LHS of the RE condition:

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\|\mathbf{X}\boldsymbol{\delta}\|_2=\left(\frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}|\langle\mathbf{x}_{gi},\boldsymbol{\delta}_0+\boldsymbol{\delta}_g\rangle|^2\right)^{\frac{1}{2}}\\
&\geq\frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}|\langle\mathbf{x}_{gi},\boldsymbol{\delta}_0+\boldsymbol{\delta}_g\rangle|\\
&\geq\frac{1}{n}\sum_{g=1}^{G}\xi\|\boldsymbol{\delta}_0+\boldsymbol{\delta}_g\|_2\sum_{i=1}^{n_g}\mathbb{1}\left(|\langle\mathbf{x}_{gi},\boldsymbol{\delta}_0+\boldsymbol{\delta}_g\rangle|\geq\xi\|\boldsymbol{\delta}_0+\boldsymbol{\delta}_g\|_2\right)\\
&=\frac{1}{n}\sum_{g=1}^{G}\xi_g\sum_{i=1}^{n_g}\mathbb{1}\left(|\langle\mathbf{x}_{gi},\boldsymbol{\delta}_{0g}\rangle|\geq\xi_g\right),
\end{aligned}$$

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:

$$\begin{aligned}
&\inf_{\boldsymbol{\delta}\in\mathcal{H}}\frac{1}{\sqrt{n}}\|\mathbf{X}\boldsymbol{\delta}\|_2\geq\inf_{\boldsymbol{\delta}\in\mathcal{H}}\sum_{g=1}^{G}\frac{n_g}{n}\xi_gQ_{2\xi_g}(\boldsymbol{\delta}_{0g})\\
&-\sup_{\boldsymbol{\delta}\in\mathcal{H}}\frac{1}{n}\sum_{g=1}^{G}\xi_g\sum_{i=1}^{n_g}\left[Q_{2\xi_g}(\boldsymbol{\delta}_{0g})-\mathbb{1}(|\langle\mathbf{x}_{gi},\boldsymbol{\delta}_{0g}\rangle|\geq\xi_g)\right]\\
&=t_1(\mathbf{X})-t_2(\mathbf{X})
\end{aligned}\tag{12}$$

For the ease of exposition we consider the LHS of (12) as the difference of two terms, i.e., $t_1(\mathbf{X})-t_2(\mathbf{X})$ and in the followings we lower bound t_1 and upper bound t_2.

1) Lower Bounding the First Term $t_1(\mathbf{X})$: Our main result is the following lemma which uses the DASHIN condition of 3 and provides a lower bound for the first term $t_1(\mathbf{X})$:

Lemma 4: Suppose DASHIN holds. Let $\psi_{\mathcal{I}}=\frac{\lambda_{\min}\bar{\rho}}{3}$. For any $\boldsymbol{\delta}\in\mathcal{H}$, we have:

$$\sum_{g=1}^{G}\frac{n_g}{n}\xi_gQ_{2\xi_g}(\boldsymbol{\delta}_{0g})\geq\psi_{\mathcal{I}}\xi\frac{(\alpha-2\xi)^2}{4ck_x^2}\sum_{g=0}^{n}\frac{n_g}{n}\|\boldsymbol{\delta}_g\|_2,\tag{13}$$

Lemma 4 implies that $t_1(\mathbf{X})$ is lower bounded by the same RHS bound of (13).

2) Upper Bounding the Second Term $t_2(\mathbf{X})$: First we show $t_2(\mathbf{X})$ satisfies the bounded difference property defined in Section 3.2. of [41], 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 $t_2(\mathbf{X})=\sup_{\boldsymbol{\delta}\in\mathcal{H}}g_\delta(\mathbf{X})$ where $g_\delta(\mathbf{X})$ is the argument of sup in (12). 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:

$$\sup_{\mathbf{X},\mathbf{x}'_{jk}} |t_2(\mathbf{X}) - t_2(\mathbf{X}'_{jk})| \leq \sup_{\mathbf{X},\mathbf{x}'_{jk}} \sup_{\boldsymbol{\delta}\in\mathcal{H}} \left|g(\mathbf{X}) - g(\mathbf{X}'_{jk})\right|$$

$$\leq \sup_{\mathbf{x}_{jk},\mathbf{x}'_{jk}} \sup_{\boldsymbol{\delta}\in\mathcal{H}} \frac{\xi_j}{n} \left|\mathbb{1}(|\langle \mathbf{x}'_{jk}, \boldsymbol{\delta}_{0j}\rangle| \geq \xi_j) - \mathbb{1}(|\langle \mathbf{x}_{jk}, \boldsymbol{\delta}_{0j}\rangle| \geq \xi_j)\right|$$

$$\leq \sup_j \sup_{\boldsymbol{\delta}\in\mathcal{H}} \frac{\xi_j}{n} = \frac{\xi}{n} \sup_j \sup_{\boldsymbol{\delta}\in\mathcal{H}} \|\boldsymbol{\delta}_0 + \boldsymbol{\delta}_j\|_2$$

$$\leq \frac{\xi}{n} \sup_j \sup_{\boldsymbol{\delta}\in\mathcal{H}} \|\boldsymbol{\delta}_0\|_2 + \|\boldsymbol{\delta}_j\|_2 \leq \xi\left(\frac{1}{n} + \frac{1}{n_j}\right) \leq \frac{2\xi}{n}$$

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 [41] 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 5: For the random vector $\mathbf{x}$ of Definition 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) Continuing the Proof of Theorem 3: 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 4 and 5, with probability at least $1 - e^{-\frac{\tau^2}{2}}$ we have:

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

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

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

$$(i) = \sum_{g=0}^{G} \frac{n_g}{n}\|\boldsymbol{\delta}_g\|_2 \kappa_g - \frac{\tau}{\sqrt{n}}$$

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

where $\kappa_g \triangleq \psi_{\mathcal{I}}\xi q - \frac{2c_g k_x \omega(\mathcal{A}_g)}{\sqrt{n_g}}$ and $\kappa_{\min} \triangleq \min_{g\in[G]} \kappa_g$ in steps (i) and (ii), and the last step follows from the fact that $\boldsymbol{\delta} \in \mathcal{H}$. 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: $\forall g \in [G_+] : O\left(k^6 \psi_{\mathcal{I}}^{-2} \alpha^{-6} \omega(\mathcal{A}_g)^2\right) \leq n_g$. ■

IV. 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 6: For $\mathbf{x}_{gi}$ and w_{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 for estimator (5) with probability at least $1 - \sigma\exp\left(-\min\left[\nu \min_{g\in[G]} n_g - \log(G+1), \tau^2\right]\right)$:

$$\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g\|_2 \leq C\gamma \frac{\max_{g\in[G_+]} \omega(\mathcal{A}_g) + \sqrt{\log(G+1)} + \tau}{\kappa_{\min}^2 \sqrt{n}} \quad (14)$$

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

Corollary 7: From (14) one can immediately entail the error bound for estimation of all 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 3: For the balanced sample condition number $\gamma = \Theta(G)$ discussed in Remark 1 we have the following error bound for all parameters:

$$\forall g \in [G_+] : \|\boldsymbol{\delta}_g\|_2 = O\left(G^{3/2} \frac{\max_{g\in[G_+]} \omega(\mathcal{A}_g) + \sqrt{\log(G+1)}}{\sqrt{n}}\right) \quad (15)$$

where the upper bound of error scales as $\frac{1}{\sqrt{n}}$ for all parameters.

Example 4: **(l_1-norm)** For the sparse DS estimator of Example 1, results of Theorems 3 and 6 translates to the following. For enough samples as $\forall g \in [G_+] : n_g \geq m_g = O(s_g \log p)$, the upper bound of error simplifies to:

$$\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}}\|\boldsymbol{\delta}_g\|_2 = O\left(\gamma\sqrt{\frac{(\max_{g\in[G_+]} s_g)\log p}{n}}\right), \quad (16)$$

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 [35], [42], [43], [44], [45].

A. Proof of Theorem 6

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 1 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 $\mathbf{p} = (p_0, \ldots, p_G)$ and $\mathbf{q} = (q_0, \ldots, q_G)$ for which we have: $\sup_{\mathbf{p}\in\overline{\mathcal{H}}} \mathbf{p}^T\mathbf{q} = \sup_{\|\mathbf{p}\|_1=1} \mathbf{p}^T\mathbf{q} \leq \|\mathbf{q}\|_\infty = \max_{g\in[G_+]} q_g$,where the inequality holds because of the definition of the dual norm. Going back to the original form:

$$\sup_{\boldsymbol{\delta}\in\mathcal{H}} \mathbf{w}^T\mathbf{X}\boldsymbol{\delta} \leq \max_{g\in[G]} \langle \mathbf{X}_g^T \frac{\mathbf{w}_g}{\|\mathbf{w}_g\|_2}, \frac{\boldsymbol{\delta}_g}{\|\boldsymbol{\delta}_g\|_2}\rangle \sqrt{\frac{n}{n_g}}\|\mathbf{w}_g\|_2 \quad (17)$$

$$\leq \max_{g\in[G]} \sqrt{\frac{n}{n_g}}\|\mathbf{w}_g\|_2 \sup_{\mathbf{u}_g \in \mathcal{C}_g \cap \mathbb{S}^{p-1}} \langle \mathbf{X}_g^T \frac{\mathbf{w}_g}{\|\mathbf{w}_g\|_2}, \mathbf{u}_g\rangle$$

Algorithm 1 DASHER

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

To avoid cluttering we define a random quantity $h_g(\mathbf{w}_g, \mathbf{X}_g)$ and a corresponding constant $e_g(\tau)$ as:

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

Then from (17), we have:

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

where the first inequality follows from the Union Bound and the last one is the result of the following lemma:

Lemma 8: For $\mathbf{x}_{gi}$ and ω_{gi} defined in Definition 1 and $\tau > 0$, with probability at least $1 - \frac{\sigma_g}{(G+1)} \exp\left(- \min\left[\nu n_g - \log(G+1), \tau^2 \right] \right)$ we have:

$$\begin{aligned}
&\|\mathbf{w}_g\|_2 \sup_{\mathbf{u}_g \in \mathcal{A}_g} \langle \mathbf{X}_g^T \frac{\mathbf{w}_g}{\|\mathbf{w}_g\|_2}, \mathbf{u}_g \rangle \qquad (18) \\
&\quad \le c_g \sqrt{(2k_w^2+1)k_x^2 n_g} \left(\omega(\mathcal{A}_g) + \sqrt{\log(G+1)} + \tau \right),
\end{aligned}$$

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) both into the bound of Theorem 1 . ■

V. Estimation Algorithm

We propose *DAta SHarER* (DASHER) a projected block gradient descent algorithm, Algorithm 1, where $\Pi_{\Omega_{f_g}}$ is the Euclidean projection onto the set $\Omega_{f_g}(d_g) = \{f_g(\boldsymbol{\beta}) \le d_g\}$ where $d_g = f_g(\boldsymbol{\beta}_g^*)$ and is dropped to avoid cluttering.

To analysis convergence properties of DASHER, 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 DASHER, 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 4: We define the following positive constants as functions of step sizes $\mu_g > 0$:

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

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

Below, we establish a deterministic bound on iteration errors $\|\boldsymbol{\delta}_g^{(t)}\|_2$ which depends on constants of Definition 4 where to simplify the notation μ_g arguments are dropped.

Theorem 9: The following deterministic bound for the error at iteration $t+1$ of Algorithm 1, initialized by $\boldsymbol{\beta}^{(1)} = \mathbf{0}$, holds:

$$\begin{aligned}
&\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g^{(t+1)}\|_2 \qquad (19) \\
&\quad \le \rho^t \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\beta}_g^*\|_2 + \frac{1-\rho^t}{1-\rho} \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \eta_g \|\boldsymbol{\omega}_g\|_2,
\end{aligned}$$

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

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 10: The following recursive dependency holds between the error of $t+1$th and tth iterations of DASHER:

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

By recursively applying results of Lemma 10, we get the following deterministic bound:

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

We have:

$$\begin{aligned} b_{t+1} &\le \rho b_t + \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \xi_g \|\boldsymbol{\omega}_g\|_2 \\ &\le \rho^2 b_{t-1} + (\rho+1) \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \xi_g \|\boldsymbol{\omega}_g\|_2 \\ &\le \rho^t b_1 + \left(\sum_{i=0}^{t-1} \rho^i \right) \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \xi_g \|\boldsymbol{\omega}_g\|_2 \\ &= \rho^t \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\beta}_g^1 - \boldsymbol{\beta}_g^*\|_2 + \left(\sum_{i=0}^{t-1} \rho^i \right) \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \xi_g \|\boldsymbol{\omega}_g\|_2 \\ &\le \rho^t \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\beta}_g^*\|_2 + \frac{1-\rho^t}{1-\rho} \sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \xi_g \|\boldsymbol{\omega}_g\|_2 \end{aligned} \tag{20}$$

where the last inequality follows from $\boldsymbol{\beta}^1 = 0$. ∎

The RHS of (20) 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 IV. 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 6.

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 Section VIII-B of Supplement) 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 4 and the deterministic bound of Theorem 9 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, we can keep ρ below one and have geometric convergence.

Theorem 11: Let $\tau = \sqrt{\log(G+1)}/\zeta + \epsilon$ for $\epsilon, \zeta > 0$. For the step sizes of:

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

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 $\forall g \in [G_+] : n_g \ge C_g(\omega(\mathcal{A}_g)+\tau)^2$, with probability at least $1 - \sigma\exp(-\min(\nu \min_{g\in[G]} n_g - \log(G+1), \zeta\epsilon^2))$ updates of Algorithm 1 obey the following:

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

where $r(\tau) < 1$ is a constant depending on τ defined in (22) and υ, ζ, and σ are constants.

Corollary 12: For enough number of samples, iterations of DS algorithm with step sizes $\mu_0 = \Theta(\frac{1}{n})$ and $\mu_g = \Theta(\frac{1}{\sqrt{n n_g}})$ geometrically converges to the following with high probability:

$$\sum_{g=0}^{G} \sqrt{\frac{n_g}{n}} \|\boldsymbol{\delta}_g^{\infty}\|_2 \le c \frac{\max_{g\in[G_+]} \omega(\mathcal{A}_g) + \sqrt{\log(G+1)}/\zeta + \theta}{\sqrt{n}(1-r(\tau))} \tag{21}$$

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

It is instructive to compare RHS of (21) with that of (14): $\kappa_{\min}$ defined in Theorem 3 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 12 shows that with the number of samples in the order of sample complexity determined in Theorem 3 DASHER converges to the statistical error bound determined in Theorem 6.

A. Proof Sketch of Theorem 11

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 4 while ignoring detailed computation of subsequent probabilities in finding $r(\tau)$. The full probabilistic proof is provided in Section VIII-B of Supplement. First we need the following lemma to upper bound constants introduced in Definition 4:

Lemma 13: For $a_g \ge 1$ the following upper bounds hold:

$$\begin{aligned} &\rho_g\left(\frac{1}{a_g n_g}\right) \le \frac{1}{2}\left[\left(1-\frac{1}{a_g}\right) + \sqrt{2} c_g \frac{2\omega_g + \tau}{a_g\sqrt{n_g}}\right], \\ &\qquad \text{with probability at least} 1 - 2\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right). \\ &\eta_g\left(\frac{1}{a_g n_g}\right) \le \frac{c_g k_x(\omega_g+\tau)}{a_g n_g}, \\ &\qquad \text{with probability at least} 1 - \pi_g\exp\left(-\tau^2\right). \\ &\phi_g\left(\frac{1}{a_g n_g}\right) \le \frac{1}{a_g}\left(1 + c_{0g}\frac{\omega_{0g}+2\tau}{\sqrt{n_g}}\right), \\ &\qquad \text{with probability at least} 1 - 2\exp\left(-\gamma_g(\omega(\mathcal{A}_g)+\tau)^2\right). \end{aligned}$$

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 9 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:

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

where we substitute a_g value in step (i). 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:

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

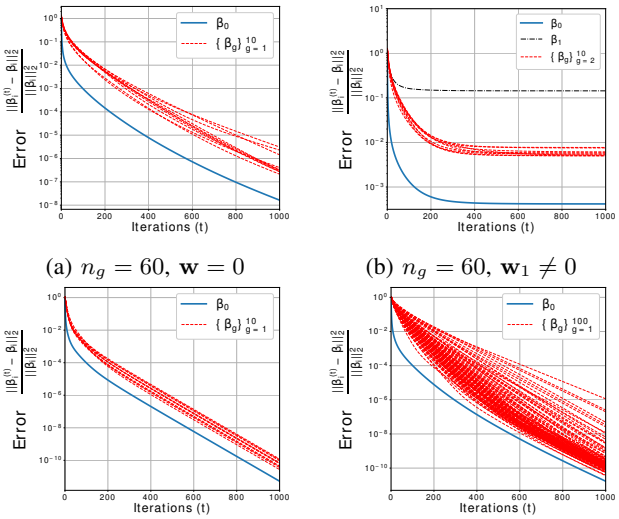
Condition 2 becomes:

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

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 $n_g>8c_g(\omega(\mathcal{A}_g)+\tau)^2$. ■

VI. Experiments on Synthetic Data

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 $\boldsymbol{\beta}_g$, $g\in[G]$. Observe that, cones of the individual parameters intersect with that of $\boldsymbol{\beta}_0$ hence this setup actually violates DASHIN (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.



(a) $n_g=60$, $\mathbf{w}=0$ (b) $n_g=60$, $\mathbf{w}_1\neq 0$

(c) $n_g=150$, $\mathbf{w}=0$ (d) $n_g=150$, $\mathbf{w}=0$

Fig. 3: In (a), (b), and (c) experiments $p=100$, $G=10$, $\forall g\in[G]: s_g=10$, and $s_0=p$. For (d) $p=1000$, $G=100$, $\forall g\in[G]: s_g=10$, and $s_0=100$. (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) DASHER convergences fast even with a large number of groups $G=100$.

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. Fig. 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 Fig. 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 Fig. 3c, we increased the sample size to $n_g=150$ per group. We observe that, in comparison to Fig. 3a, rate of convergence receives a boost from the additional samples as predicted by our theory.

Finally, Fig. 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}.

VII. Conclusion

We presented an estimator for the joint estimation of common and individual parameters of the data sharing model. We showed that the sample complexity for estimation of the shared

parameter depends on the total number of sample n. In addition, the shared parameter error rate decays as $1/\sqrt{n}$. These results indicate that our estimator benefits from the pooled data in estimating the common parameters. Both sample complexity and upper bound of error depend on the *maximum* Gaussian width among the spherical caps induced by the error cones of all parameters. We provided a projected gradient descent algorithm for estimation of the parameters and analyzed its convergence rate and showed geometric convergence for a carefully selected step size. Finally, we complemented the theoretical results presented with simulation results.

REFERENCES

[1] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[2] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[4] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, p. 717, 2009.

[5] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on information theory*, vol. 52, no. 2, pp. 489–509, 2006.

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

[7] F. Bach, R. Jenatton, J. Mairal, G. Obozinski *et al.*, "Optimization with sparsity-inducing penalties," *Foundations and Trends® in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2012.

[8] 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.

[9] 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.

[10] Y. Plan, R. Vershynin, and E. Yudovina, "High-dimensional estimation with geometric constraints," *Information and Inference: A Journal of the IMA*, vol. 6, no. 1, pp. 1–40, 2017.

[11] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and computational harmonic analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[12] 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.

[13] 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*, vol. 483, no. 7391, p. 603, 2012.

[14] 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*, vol. 166, no. 3, pp. 740–754, 2016.

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

[16] M. B. McCoy and J. A. Tropp, "The achievable performance of convex demixing," Sep. 2013.

[17] 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.

[18] 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.

[19] Y. Zhang and Q. Yang, "A survey on Multi-Task learning," 2017.

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

[21] A. Asiaee, S. Oymak, K. R. Coombes, and A. Banerjee, "High dimensional data enrichment: Interpretable, fast, and data-efficient," *arXiv preprint arXiv:1806.04047*, 2018.

[22] ——, "Data enrichment: Multi-task learning in high dimension with theoretical guarantees," in *Adaptive and Multitask Learning Workshop at ICML*, 2019.

[23] A. Chen, A. B. Owen, and M. Shi, "Data enriched linear regression," *Electronic journal of statistics*, vol. 9, no. 1, pp. 1078–1112, 2015.

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

[25] E. Ollier and V. Viallon, "Joint estimation of K related regression models with simple L_1-norm penalties," Nov. 2014.

[26] ——, "Regression modeling on stratified data with the lasso," Aug. 2015.

[27] S. Kakade, O. Shamir, K. Sindharan, and A. Tewari, "Learning exponential families in High-Dimensions: Strong convexity and sparsity," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 2010, pp. 381–388.

[28] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *Journal of Machine Learning Research*, vol. 13, no. May, pp. 1665–1697, 2012.

[29] Y. Plan and R. Vershynin, "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach," *IEEE transactions on information theory / Professional Technical Group on Information Theory*, vol. 59, no. 1, pp. 482–494, 2013.

[30] ——, "The generalized lasso with Non-Linear observations," *IEEE transactions on information theory / Professional Technical Group on Information Theory*, vol. 62, no. 3, pp. 1528–1537, Mar. 2016.

[31] Z. Yang, Z. Wang, H. Liu, Y. Eldar, and T. Zhang, "Sparse nonlinear regression: Parameter estimation under nonconvexity," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 2016, pp. 2472–2481.

[32] J. Chen, J. Liu, and J. Ye, "Learning incoherent sparse and Low-Rank patterns from multiple tasks," *ACM transactions on knowledge discovery from data*, vol. 5, no. 4, p. 22, 2012.

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

[34] ——, *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018, vol. 47.

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

[36] 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*, vol. 27, no. 4, pp. 538–557, 2012.

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

[38] S. Mendelson, "Learning without concentration," *Journal of the ACM*, vol. 62, no. 3, pp. 21:1–21:25, 2015.

[39] J. A. Tropp, "Convex recovery of a structured signal from independent random linear measurements," in *Sampling Theory, a Renaissance*. Springer, 2015, pp. 67–101.

[40] M. Rudelson and S. Zhou, "Reconstruction from anisotropic random measurements," *IEEE Transactions on Information Theory*, vol. 59, no. 6, pp. 3434–3447, 2013.

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

[42] P. J. Bickel, Y. Ritov, A. B. Tsybakov *et al.*, "Simultaneous analysis of lasso and dantzig selector," *The Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 2009.

[43] E. Candes, T. Tao *et al.*, "The dantzig selector: Statistical estimation when p is much larger than n," *The Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[44] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational Mathematics*, vol. 12, no. 6, pp. 805–849, 2012.

[45] 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.
[46] A. W. v. d. Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, NY, 1996.
[47] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, Berlin, Heidelberg, 1991.
[48] S. Oymak, B. Recht, and M. Soltanolkotabi, "Sharp time–data tradeoffs for linear inverse problems," *IEEE Transactions on Information Theory*, vol. 64, no. 6, pp. 4129–4158, 2017.