Convex and Nonconvex Formulations for Mixed Regression with Two Components: Minimax Optimal Rates

Yudong Chen, Xinyang Yi, and Constantine Caramanis, *Member, IEEE,*

***Abstract*—We consider the mixed regression problem with two components, under adversarial and stochastic noise. We give a convex optimization formulation that provably recovers the true solution, as well as a nonconvex formulation that works under more general settings and remains tractable. Upper bounds are provided on the recovery errors for both arbitrary noise and stochastic noise models. We also give matching minimax lower bounds (up to log factors), showing that our algorithm is information-theoretically optimal in a precise sense. Our results represent the first tractable algorithm guaranteeing successful recovery with tight bounds on recovery errors and sample complexity. Moreover, we pinpoint the statistical cost of mixtures: our minimax-optimal results indicate that the mixture poses a fundamentally more difficult problem in the low-SNR regime, where the learning rate changes.**

I. Introduction

This paper considers the problem of *mixed linear regression*, where each observation of the output variable comes from one of two unknown regression vectors. Formally, we observe n data points $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, which satisfies

$$y_i = \begin{cases} \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1^* \rangle + e_i, & \text{if } z_i = 0, \\ \langle \boldsymbol{x}_i, \boldsymbol{\beta}_2^* \rangle + e_i, & \text{if } z_i = 1, \end{cases} \qquad i = 1, \dots, n,$$

where $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ are two unknown regression vectors in $\mathbb{R}^p$, e_i is the noise, and $z_i \in \{0, 1\}$ can be thought of as a hidden label determining which regression vector generates the i-th data point. Our goal is to estimate the pair $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$. We consider the setting where the covariates $\boldsymbol{x}_i$ and the noise e_i are independent of the labels, and in particular, no information about the labels can be directly inferred from them. This setting means that predicting labels exactly is impossible, with or without knowing $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$.

Manuscript received XXX XX, 20XX; revised XXX XX, 20XX. Y. Chen was supported by NSF grants CCF-1704828, CRII award 1657420, and the School of Operations Research and Information Engineering at Cornell University. X. Yi and C. Caramanis were supported by NSF Grants EECS-1056028, CNS-1302435, CCF-1116955, and the USDOT UTC–D-STOP Center at UT-Austin.

Y. Chen is with the School of Operations Research and Informaiton Engineering, Cornell University, Ithaca, NY 14850 USA (yudong.chen@cornell.edu).

X. Yi and C. Caramanis are with the Department of Electrical and Computer Engineering, the University of Texas at Austin, Austin, TX 78712 USA (yixy@utexas.edu, constantine@utexas.edu).

This work was presented in part at the Conference on Learning Theory, 2014.

If the label of each sample is known, the problem decomposes into two standard linear regression problems, and can be easily solved. Without knowing the labels, however, the problem is significantly more difficult. The main challenge of mixture models, and in particular mixed regression falls in the intersection of the *statistical* and *computational* constraints: the problem is difficult when one cares both about an efficient algorithm, and about near-optimal sample complexity and estimation error. Exponential-effort brute force search (over all possibilities of the labels) typically results in statistically near-optimal estimators. On the other hand, recent tensor-based methods give a polynomial-time algorithm, but at the cost of an $O(p^6)$ sample size (recall $\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^* \in \mathbb{R}^p$) instead of the optimal rate $O(p)$.[1] The Expectation Maximization (EM) algorithm is computationally very efficient, and widely used in practice for mixture problems. However, there has been only limited understanding of its behavior, and in particular, no general theoretical guarantees on global convergence are known.

a) Our contributions and the cost of mixtures.: In this paper, we tackle both statistical and algorithmic objectives at once. The algorithms we give are computationally efficient, specified by solutions of convex optimization problems as well as a tractable nonconvex formulation, which can be solved by polynomial-time, globally convergent procedures. In both the noisy and noiseless settings our results provide better statistical guarantees compared to the best known previous results. In particular, in both the arbitrary noise and stochastic noise regimes, we provide matching estimation error bounds and minimax lower bounds, showing our results are statistically optimal.

An interesting feature of our minimax results is that we pinpoint the statistical cost of dealing with a mixture problem compared to ordinary regression problems. As we detail below in Theorems 3, 4 and 8, we show that in the high SNR regime, there is (up to log factors) no loss, and one can expect to recover the regression parameters at the parametric learning rate (as with ordinary regression). At the low-SNR regime (where, to the best of our knowledge, there are no previous results on mixed regression), the rate changes from $(1/n)^{1/2}$ to $(1/n)^{1/4}$. It is of interest to explore to what we owe this change in rate. On the algorithmic side, our approach is to solve a related low-rank matrix regression problem in the lifted

[1] It may be possible to improve the sample requirement of tensor methods to $O(p^4)$ for the case of Gaussian design.

space, to estimate (something related to) the tensor product of the two regressors. We show that the regression in this lifted matrix space has an error that decays with $(1/n)^{1/2}$. At high SNRs, we prove a perturbation result (Theorem 4) that shows that the top eigenvectors of this matrix inherit the same rate. At low SNRs, however, this comes at a cost, and the rate reduces to $(1/n)^{1/4}$. Our matching lower bounds show that this cost of converting an error bound in the lifted space to an error bound in the regressor space is fundamental and thus encapsulates the crux of the challenge. The lower bounds are established by showing that at low SNR, distinguishing one *mixture* of Gaussians from another is more difficult than distinguishing one Gaussian from another, at exactly the rate change indicated above.

Specifically, our contributions are as follows:

- In the arbitrary noise setting where the noise $\boldsymbol{e} = (e_1, \ldots, e_n)^\top$ can be adversarial, we show that under mild technical conditions, as long as the numbers of observations for each regression vector satisfy $n_1, n_2 \gtrsim p$, our convex algorithm produces an estimator $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ that satisfies
$$\|\hat{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_b^*\|_2 \lesssim \frac{\|\boldsymbol{e}\|_2}{\sqrt{n}}, \quad \text{for } b = 1, 2.$$
This result immediately implies exact recovery in the noiseless case with $O(p)$ samples.
- In the stochastic noise setting with sub-Gaussian noise and balanced labels ($n_1/n_2 \to 1$), if we have $n_1, n_2 \gtrsim p$ and a Gaussian design matrix, our convex estimator satisfies the following error bound (omitting polylog factors):
$$\|\hat{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_b^*\|_2 \lesssim \begin{cases} \sigma\sqrt{\frac{p}{n}}, & \text{if } \gamma \geq \sigma, \\ \frac{\sigma^2}{\gamma}\sqrt{\frac{p}{n}}, & \text{if } \sigma\left(\frac{p}{n}\right)^{\frac{1}{4}} \leq \gamma \leq \sigma, \\ \sigma\left(\frac{p}{n}\right)^{\frac{1}{4}}, & \text{if } \gamma \leq \sigma\left(\frac{p}{n}\right)^{\frac{1}{4}} \end{cases}$$
for $b = 1, 2$, where γ is an lower bound of the signal strength $\|\boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^*\|_2$, and σ^2 is the variance of the noise e_i.
- In the stochastic noise setting with imbalanced labels, we propose a nonconvex optimization based formulation along with a polynomial-time solver with provable guarantees. Specifically, we show that when $\min\{n_1/n_2, n_2/n_1\}$ is lower bounded by any constant and $\gamma/\sigma \gtrsim 1$, our estimator satisfies the bound
$$\|\hat{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_b^*\|_2 \lesssim \sigma\sqrt{\frac{p}{n}}, \quad \text{for } b = 1, 2.$$
- Finally, in both the arbitrary and stochastic noise settings, we provide minimax lower bounds that match the above upper bounds up to at most polylog factors, thereby showing that the results obtained by the estimates produced by our algorithms are information-theoretically optimal. Particularly in the stochastic setting, the situation is a bit more subtle: the minimax rates in fact depend on the signal-to-noise ratio (SNR) γ/σ and exhibit several phases, showing a qualitatively different behavior than in standard regression and many other parametric problems (for which the minimax rate is usually $\sqrt{1/n}$).

II. Related Work

Mixture models and latent variable modeling are very broadly used in a wide array of contexts far beyond regression. Subspace clustering [19, 26, 32], Gaussian mixture models [3, 21] and k-means/medians clustering [13] are popular examples of unsupervised learning for mixture models. Arguably the most popular and broadly implemented approach to mixture problems, including mixed regression, is the EM algorithm [18, 22]. In fact, EM has been used for mixed regression for various application domains [20, 31]. Despite its wide use, still little is known about its performance beyond local convergence [4, 34].

One exception is the work in [37], which studies mixed regression in the noiseless setting. They propose an alternating minimization approach initialized by a grid search, and show that their algorithm recovers the regression vector with a sample complexity of $O(p\log^2 p)$. Extension to the noisy setting is recently considered by the authors of [4]. Focusing on the stochastic noise setting with sufficiently high SNR (that is, when $\gamma \gtrsim \sigma$; cf. Section I), they show that the EM algorithm with good initialization achieves the error bound $\|\hat{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_b^*\|_2 \lesssim \sqrt{\gamma^2 + \sigma^2}\sqrt{\frac{p}{n}}$. In the work [27], EM is adapted to the high-dimensional regression setting, where the regression vectors are known to be sparse and EM is used to solve a penalized (for sparsity) likelihood function. This generalized EM approach achieves support-recovery, though once restricted to that support where the problem becomes a standard mixed regression problem, only convergence to a local optimum can be guaranteed. Very recently for this high-dimensional sparse regression setting, the works in [33] and [36] establish local convergence (that is., assuming that EM is run from a good initialization) of truncated and regularized EM algorithms to a statistically optimal solution.

Mixture problems have been explored using the technology of tensors recently developed in the literature [2, 21]. The authors of [12] consider a tensor-based approach, regressing $\mathbf{x}^{\otimes 3}$ against y^3 and then using the tensor decomposition techniques to efficiently recover each $\boldsymbol{\beta}_b^*$. These methods are not limited to the mixture of only two components, as we are. Yet, even for two components, the tensor approach requires $O(p^6)$ samples, compared to $O(p \cdot \text{polylog}(p))$ that our work requires. As noted in their work, the higher sampling requirement seems to be a common difficulty for algorithms based on high order tensors.

In this work we consider the setting with two mixture components. Binary latent factors are common modeling tools for many applications, as they model on/off-type relationships, among others. We refer to the paper [31] for numerous examples of such problems. While the extension to more than two components is of great interest, much is unknown even for two components. In particular, globally convergent algorithms with even near-optimal sample complexity are, to the best of our knowledge, unknown. In fact, in the low-SNR regime (see below for details) we are unaware of any algorithm able to guarantee non-trivial estimation of the parameters. As explained above, our work shows that there may be a fundamental reason for this: the minimax-optimal rate, and hence

the (statistical) difficulty of the mixed regression problem, is different at the high-SNR and low-SNR regimes.

III. Main Results

In this section we present this paper's main results: our algorithms for mixed regression and matching statistical upper and lower bounds. In addition, we describe the precise setup and assumptions, and introduce the basic notation we use.

A. Problem Set Up

Suppose that $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ are two unknown vectors in $\mathbb{R}^p$. We observe n *noisy* linear measurements $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. Let $\mathcal{I}_1$ and $\mathcal{I}_2$ denote the (unknown) subsets of the measurements corresponding to $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$, respectively, with $\mathcal{I}_1 \cup \mathcal{I}_2 = [n]$, $\mathcal{I}_1 \cap \mathcal{I}_2 = \phi$, $n_1 := |\mathcal{I}_1|$, $n_2 := |\mathcal{I}_2|$ and $n_1 + n_2 = n$. The measurements satisfy the following model: for each $b \in \{1, 2\}$ and $i \in \mathcal{I}_b$,

$$y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta}_b^* \rangle + e_i. \tag{1}$$

Given the data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, the goal is to recover $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$.

To measure the estimation error, we use the following symmetric (semi-)metric: for each pair $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $\boldsymbol{\theta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ in $\mathbb{R}^p \times \mathbb{R}^p$, define

$$\rho(\boldsymbol{\theta}, \boldsymbol{\theta}') := \min \Big\{ \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1'\|_2 + \|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_2'\|_2, \; \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2'\|_2 + \|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1'\|_2 \Big\}, \tag{2}$$

which is the total ℓ_2 distance up to permutation. For an estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ of the true regression vector pair $\boldsymbol{\theta}^* := (\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$, we are interested in bounding the symmetric estimation error $\rho(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$. In the presence of noise, the correct labels (or equivalently, the sets $\mathcal{I}_1$ and $\mathcal{I}_2$) are in general irrecoverable.

The key insight to our optimization formulation is to utilize a particular lifting to the space of $p \times p$ matrices, that then allows recovering $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ with controlled perturbation, without requiring working in the space of 3-tensors. More concretely, defining the true quantities $\boldsymbol{K}^*$ and $\boldsymbol{g}^*$ as

$$\begin{aligned} \boldsymbol{K}^* &:= \frac{1}{2}\left(\boldsymbol{\beta}_1^* \boldsymbol{\beta}_2^{*\top} + \boldsymbol{\beta}_2^* \boldsymbol{\beta}_1^{*\top}\right) \in \mathbb{R}^{p \times p}, \\ \boldsymbol{g}^* &:= \frac{1}{2}\left(\boldsymbol{\beta}_1^* + \boldsymbol{\beta}_2^*\right) \in \mathbb{R}^p, \end{aligned} \tag{3}$$

we solve a linear regression in the matrix and vector variables $\boldsymbol{K}$ and $\boldsymbol{g}$ in order to produce an approximation of $\boldsymbol{K}^*$ and $\boldsymbol{g}^*$. Given the pair $(\boldsymbol{K}^*, \boldsymbol{g}^*)$, the true regression vectors $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ can be recovered exactly, using the identity

$$\boldsymbol{J}^* := \boldsymbol{g}^* \boldsymbol{g}^{*\top} - \boldsymbol{K}^* = \frac{1}{4}\left(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\right)\left(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\right)^\top.$$

Letting λ^* and $\boldsymbol{v}^*$ be the first eigenvalue-eigenvector pair of $\boldsymbol{J}^*$, we have $\sqrt{\lambda^*}\boldsymbol{v}^* := \pm\frac{1}{2}\left(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\right)$ and thus together with $\boldsymbol{g}^*$ we can recover $(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$. Given an approximation, $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ of $(\boldsymbol{K}^*, \boldsymbol{g}^*)$, this procedure is still well defined, and is described fully in Algorithm 1. In fact, this estimation procedure is stable: if $\hat{\boldsymbol{K}}$ and $\hat{\boldsymbol{g}}$ are close to true quantities $\boldsymbol{K}^*$ and $\boldsymbol{g}^*$, then Algorithm 1 outputs a pair $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ that is close to the true regression vectors $(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$.

Algorithm 1 Estimate $(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$ from $(\boldsymbol{K}, \boldsymbol{g})$

Input: $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}}) \in \mathbb{R}^{p \times p} \times \mathbb{R}^p$.
Compute the matrix $\hat{\boldsymbol{J}} = \hat{\boldsymbol{g}}\hat{\boldsymbol{g}}^\top - \hat{\boldsymbol{K}}$, and its first eigenvalue-eigenvector pair $\hat{\lambda}$ and $\hat{\boldsymbol{v}}$.
Compute $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{g}} + \sqrt{\hat{\lambda}}\hat{\boldsymbol{v}}$ and $\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{g}} - \sqrt{\hat{\lambda}}\hat{\boldsymbol{v}}$.
Output: $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$

Below we consider separately the settings with arbitrary noise and stochastic noise in Sections III-B and III-C, and give our algorithms with rigorous sample complexity and estimation error bounds. In Section III-D we further provide matching (up to at most a polylog factor) minimax lower bounds.

a) Notation:: We use lower case bold letters to denote vectors, and capital bold-face letters for matrices. For a vector $\boldsymbol{u}$, the notations u_i and $u(i)$ both denote its i^{th} coordinate. We use standard notation for matrix and vector norms, e.g., $\|\cdot\|_*$ to denote the nuclear norm, $\|\cdot\|_F$ the Frobenius norm, and $\|\cdot\|$ the operator/spectral norm. We define two quantities that we use repeatedly:

$$\alpha := \frac{\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|_2^2}{\|\boldsymbol{\beta}_1^*\|_2^2 + \|\boldsymbol{\beta}_2^*\|^2} \quad \text{and} \quad \gamma := \|\boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^*\|_2. \tag{4}$$

Note that the quantity $\alpha \in [0, 2]$ measures the angle between $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$, and is strictly positive whenever $\boldsymbol{\beta}_1^* \neq \boldsymbol{\beta}_2^*$. For a fixed α, the parameter γ quantifies the signal strength of $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$. We say a number c is a *numerical constant* if c is independent of the dimension p, the number of measurements n, the quantity α and the magnitude/variance of the noise $\boldsymbol{e}$. For ease of parsing, we typically use c to denote a *large* constant, and $\frac{1}{c}$ for a *small* constant.

B. Arbitrary Noise

We consider first mixed regression with arbitrary noise, with the following specific setting. We take the covariate vectors $\{\boldsymbol{x}_i\}$ to have i.i.d. zero-mean and sub-Gaussian entries with sub-Gaussian norm[2] bounded by a numeric constant and satisfy $\mathbb{E}\left[(\boldsymbol{x}_i(l))^2\right] = 1$, $\mathbb{E}\left[(\boldsymbol{x}_i(l))^4\right] = \mu$ for each $i \in [n]$ and $l \in [p]$.[3] We assume that the forth moment μ is a fixed constant and independent of the parameters p, n and α. If the covariates $\{\boldsymbol{x}_i\}$ are standard Gaussian vectors, then these assumptions are satisfied with unit sub-Gaussian norm and $\mu = 3$. The only assumption on the noise vector $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)^\top$ is that it is bounded in ℓ_2 norm. The noise $\boldsymbol{e}$ is otherwise arbitrary, possibly adversarial, and even potentially depending on $\{\boldsymbol{x}_i\}$ and $(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$.

[2]The sub-Gaussian norm of a zero-mean random variable X is defined as $\|X\|_{\psi_2} := \inf\left\{b \geq 0 \,|\, \mathbb{E}\exp(tX) \leq \exp(b^2 t^2/2)\right\}$. The variable X is called sub-Gaussian if $\|X\|_{\psi_2} < \infty$.

[3]Recall that, as shown in the paper [37], the general mixed regression problem with deterministic covariates is NP-hard even in the noiseless setting.

Our algorithm is based on the following convex program:

$$\min_{\boldsymbol{K},\boldsymbol{g}} \quad \|\boldsymbol{K}\|_* \tag{5}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \left| -\langle \boldsymbol{x}_i \boldsymbol{x}_i^\top, \boldsymbol{K} \rangle + 2y_i \langle \boldsymbol{x}_i, \boldsymbol{g} \rangle - y_i^2 \right| \le \eta. \tag{6}$$

The intuition is that in the noiseless setting with $\boldsymbol{e} = \boldsymbol{0}$, if we substitute the desired solution $(\boldsymbol{K}^*, \boldsymbol{g}^*)$ given by equation (3) into the above program, the LHS of the constraint (6) becomes zero; moreover, the rank of $\boldsymbol{K}^*$ is at most 2, and minimizing the nuclear norm term in the objective (5) encourages the optimal solution to have low rank. Our theoretical results give a precise way to set the right hand side, η, of the constraint.

The next two theorems summarize our results for the arbitrary noise setting. Theorem 1 provides guarantees on how close the optimal convex optimization solution $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ is to the true quantities $(\boldsymbol{K}^*, \boldsymbol{g}^*)$. Then the companion result, Theorem 2, provides quality bounds on $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$, produced by applying Algorithm 1 to the solution $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ of the convex program.

Theorem 1 (Arbitrary noise)**.** *There exist numerical positive constants $\{c_i\}_{i=0}^6$ for which the following holds. Assume that $\max\left\{\frac{n_1}{n_2}, \frac{n_2}{n_1}\right\} \le c_0$. Suppose, moreover, that (i) $\mu > 1$ and $\alpha > 0$, (ii) $\min\{n_1, n_2\} \ge c_3 \frac{1}{\alpha} p$, and (iii) the noise satisfies the bound*

$$\|\boldsymbol{e}\|_2 \le \frac{\sqrt{\alpha}}{c_5} \sqrt{n} \left(\|\boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^*\|_2 \right).$$

Suppose further that we choose the tuning parameter η to satisfy

$$\eta \ge c_4 \sqrt{n} \|\boldsymbol{e}\|_2 \|\boldsymbol{\beta}_2^* - \boldsymbol{\beta}_1^*\|_2 .$$

Then, with probability at least $1 - c_1 \exp(-c_2 n)$, any optimal solution $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ to the convex program (5)–(6) satisfies the error bounds

$$\left\| \hat{\boldsymbol{K}} - \boldsymbol{K}^* \right\|_F \le c_6 \frac{1}{\sqrt{\alpha} n} \eta,$$

$$\|\hat{\boldsymbol{g}} - \boldsymbol{g}^*\|_2 \le c_6 \frac{1}{\sqrt{\alpha} n \left(\|\boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^*\|_2 \right)} \eta.$$

Given the solution $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$, we then estimate $\boldsymbol{\theta}^* = (\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$ by using Algorithm 1, which is stable as shown by the theorem below.

Theorem 2 (Estimating $\boldsymbol{\beta}_b^*$, arbitrary noise)**.** *Suppose that the conditions in Theorem 1 hold, and $\eta \asymp \sqrt{n} \|\boldsymbol{e}\|_2 \|\boldsymbol{\beta}_2^* - \boldsymbol{\beta}_1^*\|_2$. Then with probability at least $1 - c_1 \exp(-c_2 n)$, the output $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ of Algorithm 1 satisfies the error bound*

$$\rho(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \le \frac{1}{c_3 \sqrt{\alpha}} \frac{\|\boldsymbol{e}\|_2}{\sqrt{n}}.$$

Theorem 2 immediately implies exact recovery in the noiseless case.

Corollary 1 (Exact recovery)**.** *Suppose that $\boldsymbol{e} = \boldsymbol{0}$, the conditions (i) and (ii) in Theorem 1 hold, and $\eta = 0$. Then with probability at least $1 - c_1 \exp(-c_2 n)$, Algorithm 1 returns the true regression vectors $\{\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*\}$.*

Below we provide several remarks on the above theoretical results.

a) Discussion of assumptions:: Theorem 1 involves several mild technical assumptions.

1) The condition $\mu > 1$ in Theorem 1 is satisfied, for instance, if $\{\boldsymbol{x}_i\}$ is Gaussian (with $\mu = 3$). Moreover, this condition is in general necessary. To see this, suppose that each $\boldsymbol{x}_i(l)$ is a Rademacher ± 1 variable, which has forth moment $\mu = 1$, and the true regression vectors $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ are in $\mathbb{R}^2$. The response variable y_i must have the form

$$y_i = \pm(\boldsymbol{\beta}_b^*)_1 \pm (\boldsymbol{\beta}_b^*)_2.$$

Consider two possibilities: $\boldsymbol{\beta}_1^* = -\boldsymbol{\beta}_2^* = (1, 0)^\top$ or $\boldsymbol{\beta}_1^* = -\boldsymbol{\beta}_2^* = (0, 1)^\top$. In both cases, the observed data $(\boldsymbol{x}_i, y_i)$ takes any one of the values in $\{\pm 1\}^2 \times \{\pm 1\}$ with equal probability, and hence the problem is unidentifiable as it is impossible to distinguish the two possibilities above.

2) The condition $\alpha > 0$ holds if $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ are not equal. Suppose that α is lower-bounded by a constant. The main assumption on the noise, namely, the condition $\|\boldsymbol{e}\|_2 \lesssim \sqrt{n} (\|\boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^*\|_2)$ in Theorem 1, cannot be substantially relaxed if we want a bound on $\|\hat{\boldsymbol{g}} - \boldsymbol{g}^*\|_2$. Indeed, if $|e_i| \gtrsim \|\boldsymbol{\beta}_b^*\|_2$ for all i, then an adversary may choose e_i such that

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_b^* + e_i = 0, \quad \forall i,$$

in which case the convex program (5)–(6) becomes independent of $\boldsymbol{g}$. That said, the case when the noise bound condition is violated can be handled easily. Suppose that $\|\boldsymbol{e}\|_2 \ge c_4 \sqrt{\alpha n} (\|\boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^*\|_2)$ for any positive constant c_4. A standard argument for ordinal linear regression shows that the blind estimator $\hat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta}} \sum_{i \in \mathcal{I}_1 \cup \mathcal{I}_2} \left| \boldsymbol{x}_i^\top \boldsymbol{\beta} - y_i \right|$ satisfies w.h.p. the bound

$$\max\left\{ \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_1^*\|_2, \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_2^*\|_2 \right\} \lesssim \frac{\|\boldsymbol{e}\|_2}{\sqrt{n}},$$

and this bound is optimal (see the minimax lower bound in Section III-D). Therefore, the condition (iv) in Theorem 1 is not really restrictive, in the sense that the case when it holds is precisely the interesting setting.

3) Finally, note that if $n_1/n_2 \to +\infty$, then a single regression vector $\boldsymbol{\beta}_1^*$ explains 100% (asymptotically) of the observed data. Moreover, the standard least squares solution provides an accurate estimator of this $\boldsymbol{\beta}_1^*$.

b) Optimality of sample complexity:: The sample complexity bounds of Theorem 2 and Corollary 1 are optimal. The results require the sample size n_1 and n_2 to be $\Omega(p)$. Since we are estimating two p dimensional vectors without any further structure, this result cannot be improved in general.

C. Stochastic Noise

We now turn to the stochastic noise setting. We assume that the covariates $\{\boldsymbol{x}_i\}$ have i.i.d. Gaussian entries with zero mean and unit variance. For the noise, we assume that $\{e_i\}$ are i.i.d., zero-mean and sub-Gaussian with variance $\mathbb{E}\left[e_i^2\right] = \sigma^2$ and

sub-Gaussian norm $\|e_i\|_{\psi_2} \le c\sigma$ for some absolute constant c, and are independent of $\{\boldsymbol{x}_i\}$.

We discuss two algorithms for consistent estimation of $(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$. First, for Gaussian covariates in the *balanced setting* (that is, $n_1/n_2 \to 1$), we show that by solving a simple convex program, we have asymptotic consistency for any SNR γ/σ. The rates we obtain match information-theoretic lower bounds we give in Section III-D, and hence are minimax optimal. An interesting feature we observe is the rate change in the high- and low-SNR regimes mentioned above — a feature that precisely identifies the cost of solving a mixture problem.

Second, in the general *imbalanced setting* (that is, $n_1/n_2 \not\to 1$), we propose a *nonconvex* yet tractable extension of the above convex program, for which we establish minimax optimal estimation rates under the condition $\gamma/\sigma \gtrsim 1$.

1) Consistent Estimation with Balanced Samples: For the stochastic noise setting, while one can use the same ℓ_1 constraint as we do in arbitrary noise case, it turns out that the analysis is more natural by considering a Lagrangian formulation. In particular, much like in standard regression, the independence assumption on $\{e_i\}$ makes the least-squares objective analytically convenient. We therefore consider the following formulation, regularizing the squared loss objective with the nuclear norm of $\boldsymbol{K}$:

$$\min_{\boldsymbol{K},\boldsymbol{g}} \sum_{i=1}^{n} \left(-\langle \boldsymbol{x}_i\boldsymbol{x}_i^\top, \boldsymbol{K}\rangle + 2y_i \langle \boldsymbol{x}_i, \boldsymbol{g}\rangle - y_i^2 + \sigma^2\right)^2 + \lambda \|\boldsymbol{K}\|_* . \tag{7}$$

We assume that the noise variance σ^2 is known and can be estimated.[4]

As with the arbitrary noise case, we first provide a theorem that bounds the distance between the optimal solution $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ of the above program and the true $(\boldsymbol{K}^*, \boldsymbol{g}^*)$, and then a companion theorem gives error bounds on estimating $(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$.

Theorem 3 (Stochastic noise with nearly balanced samples)**.** *For any constant $0 < c < 2$, there exist numerical positive constants $\{c_i\}_{i=0}^5$, which might depend on c, such that the following hold. Assume that $\max\left\{\frac{n_1}{n_2}, \frac{n_2}{n_1}\right\} \le c_0$. Suppose moreover that (i) $\alpha \ge c$, (ii) $\min\{n_1, n_2\} \ge c_4 p$, and (iii) $\{\boldsymbol{x}_i\}$ are Gaussian. For any tuning parameter λ that satisfies*

$$\lambda \ge c_5\sigma \left(\|\boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^*\|_2 + \sigma\right)\left(\sqrt{np} + |n_1 - n_2|\sqrt{p}\right)\log^3 n,$$

with probability at least $1 - c_1 n^{-c_2}$, any optimal solution $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ to the convex program (7) satisfies the error bounds

$$\left\|\hat{\boldsymbol{K}} - \boldsymbol{K}^*\right\|_F \le c_3 \frac{1}{n}\lambda,$$
$$\|\hat{\boldsymbol{g}} - \boldsymbol{g}^*\|_2 \le c_3 \frac{1}{n\left(\|\boldsymbol{\beta}_1^*\| + \|\boldsymbol{\beta}_2^*\| + \sigma\right)}\lambda.$$

The bounds in the above theorem depend on the sample size difference $|n_1 - n_2|$. This dependence appears as a result of the objective function in the formulation (7) and is not an artifact of our analysis.[5] We later address how to correct for this dependence and handle imbalanced samples. Nevertheless, in the setting where the samples from the two components are approximately balanced in size with $|n_1 - n_2|$ small, the above result implies consistency with optimal convergence rate. In this case, running Algorithm 1 on the optimal solution $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ of the program (7) to estimate $\boldsymbol{\theta}^* = (\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$, we have the following guarantees.

Theorem 4 (Estimating $\boldsymbol{\beta}_b^*$, stochastic noise and nearly balanced samples)**.** *Suppose that $|n_1 - n_2| = O\left(\sqrt{n\log n}\right)$, the conditions (i)–(iii) in Theorem 3 hold, $\lambda \asymp \sigma\left(\|\boldsymbol{\beta}_1^*\| + \|\boldsymbol{\beta}_2^*\| + \sigma\right)\sqrt{np}\log^3 n$, and $n \ge c_3 p \log^8 n$. Then with probability at least $1 - c_1 n^{-c_2}$, the output $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ of Algorithm 1 satisfies the error bound*

$$\rho(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \le c_4\sigma\sqrt{\frac{p}{n}}\log^4 n + c_4 \min\left\{\frac{\sigma^2}{\|\boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^*\|_2}\sqrt{\frac{p}{n}},\ \sigma\left(\frac{p}{n}\right)^{\frac{1}{4}}\right\}\log^4 n.$$

The error bound has three terms, which are proportional to $\sigma\sqrt{\frac{p}{n}}$, $\frac{\sigma^2}{\|\boldsymbol{\beta}_b^*\|_2}\sqrt{\frac{p}{n}}$ and $\sigma\left(\frac{p}{n}\right)^{1/4}$, respectively (ignoring log factors). We show that these three terms match well with the information-theoretic lower bounds given in Section III-D. They represent three phases of the error rate under different SNR; we discuss further in Section III-D.

a) Discussion of Assumptions:: The theoretical results above assume Gaussian covariate distribution. This Gaussianity assumption can be relaxed, but using our analysis, it comes at a cost in terms of convergence rate (and hence sample complexity required for bounded error). It can be shown that $n = \tilde{O}(p\sqrt{p})$ suffices under a general sub-Gaussian assumption on the covariates. We believe that this additional cost is an artifact of our analysis.

2) Consistent Estimation with Imbalanced Labels: Now we turn to the general imbalanced setting, where the sample sizes of the two components n_1 and n_2 may be different. To account for the effect of imbalanced samples, we consider the following "corrected" version of the optimization problem (7):

$$\begin{aligned}\min_{\boldsymbol{K},\boldsymbol{g}} \quad & \sum_{i=1}^{n}\left(-\left\langle \boldsymbol{x}_i\boldsymbol{x}_i^\top, \boldsymbol{K}\right\rangle + 2y_i\langle \boldsymbol{x}_i, \boldsymbol{g}\rangle - y_i^2 + \sigma^2\right)^2 \\ & - \sum_{i=1}^{n} 4\sigma^2\left(y_i - \langle \boldsymbol{x}_i, \boldsymbol{g}\rangle\right)^2 \\ \text{s.t.} \quad & \|\boldsymbol{K}\|_* \le \|\boldsymbol{K}^*\|_*.\end{aligned} \tag{8}$$

Compared to the previous formulation (7), we replace the nuclear norm penalty in the objective function with its constraint version, which again serves as a surrogate for the low rank structure of $\boldsymbol{K}^*$. More crucially, we add a negative term in the objective function of (8), which corrects for the impact of imbalanced samples on the resulting estimate. This formulation is based on the following intuition: Minimizing the first term results in a solution $\hat{\boldsymbol{g}}$ that is biased towards $\boldsymbol{\beta}_1^*$

[4] We note that similar assumptions are made in the paper [12]. It is possible to avoid the dependence on σ by using a symmetrized error term in the objective of (7) (see, e.g., [6]).

[5] Intuitively, if the majority of the observations are generated by one of the $\boldsymbol{\beta}_b^*$, then the objective produces a solution biased toward this $\boldsymbol{\beta}_b^*$ since this solution fits more observations. In Section III-C2, we compensate for such bias by optimizing a modified objective.

when $n_1 > n_2$, whereas the second negative term increases when $\boldsymbol{g}$ gets closer to $\boldsymbol{\beta}_1^*$. Simple calculation shows that in expectation these two effects cancel out. Therefore, jointly minimizing the two terms produces a solution $\hat{\boldsymbol{g}}$ that is a consistent estimator of $(\boldsymbol{\beta}_1^* + \boldsymbol{\beta}_2^*)/2$ even when n_1 and n_2 are imbalanced.

The negative term in the objective function of (8) makes the optimization problem no longer convex. Nevertheless, one can still apply *projected gradient descent* to the nonconvex program (8). To specify the projected gradient descent iteration and ease notation, we use the following shorthand for the objective function:

$$\mathcal{L}_n(\boldsymbol{K}, \boldsymbol{g}) := \sum_{i=1}^{n} \left(-\langle \boldsymbol{x}_i \boldsymbol{x}_i^\top, \boldsymbol{K} \rangle + 2y_i \langle \boldsymbol{x}_i, \boldsymbol{g} \rangle - y_i^2 + \sigma^2\right)^2 - \sum_{i=1}^{n} 4\sigma^2 \left(y_i - \langle \boldsymbol{x}_i, \boldsymbol{g} \rangle\right)^2.$$

Let $\nabla_{\boldsymbol{K}} \mathcal{L}_n$ denote the partial gradient of $\mathcal{L}_n$ over $\boldsymbol{K}$, and $\nabla_{\boldsymbol{g}} \mathcal{L}_n$ the partial gradient of $\mathcal{L}_n$ over $\boldsymbol{g}$. Given any feasible initializer $\left(\boldsymbol{K}^{(0)}, \boldsymbol{g}^{(0)}\right)$ with $\|\boldsymbol{K}^{(0)}\|_* \le \|\boldsymbol{K}^*\|_*$, the projected gradient descent algorithm with step size η is given by the update:

$$\begin{aligned}\left(\boldsymbol{K}^{(t+1)}, \boldsymbol{g}^{(t+1)}\right) \\ \leftarrow \underset{\|\boldsymbol{K}\|_* \le \|\boldsymbol{K}^*\|_*}{\arg\min} \Big\{ &\left\langle \nabla_{\boldsymbol{K}} \mathcal{L}_n^{(t)}, \boldsymbol{K} \right\rangle + \left\langle \nabla_{\boldsymbol{g}} \mathcal{L}_n^{(t)}, \boldsymbol{g} \right\rangle \\ &+ \frac{\eta}{2} \left\|\boldsymbol{K} - \boldsymbol{K}^{(t)}\right\|_F^2 + \frac{\eta\gamma^2}{2} \left\|\boldsymbol{g} - \boldsymbol{g}^{(t)}\right\|_2^2 \Big\},\end{aligned} \tag{9}$$

where $\mathcal{L}_n^{(t)} := \mathcal{L}_n\left(\boldsymbol{K}^{(t)}, \boldsymbol{g}^{(t)}\right)$. Recall that the quantity γ is a measure of the signal strength. We need γ in the algorithm because the smoothness constants of $\mathcal{L}_n(\boldsymbol{K}, \boldsymbol{g})$ with respect to $\boldsymbol{K}$ and $\boldsymbol{g}$ differ by a factor of γ^2. The minimization in (9) can be computed by two simple steps: (1) moving $\boldsymbol{K}$ and $\boldsymbol{g}$ in the negative gradient direction with step sizes $1/\eta$ and $1/(\eta\gamma^2)$, respectively; (2) projecting the new $\boldsymbol{K}$ onto the nuclear norm ball. The second projection step can be done with a singular value decomposition (SVD) followed by shrinking the singular values (see, e.g., [1]).

The optimization formulation (8) is non-convex. Nevertheless, the projected gradient descent algorithm still converges to a statistically accurate solution, as we show below. The intuition is that the gradient descent iterates will stay in the directions along which the objective function is convex-like. In particular, in addition to the statistical error (how close the optimal solution of (8) is to the ground-truth quantity), we can control the optimization error — how close we can get to the optimal solution of (8). Accordingly, Theorem 5 below consists of two parts. Part (a) is an analogue to Theorem 3, and bounds the *statistical error*: the distance between the global optimum $\left(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}}\right)$ of (8) and the true pair $\left(\boldsymbol{K}^*, \boldsymbol{g}^*\right)$. Part (b) bounds the *optimization error* of the gradient descent iteration (9), that is, the distance between the iterates $\left(\boldsymbol{K}^{(t)}, \boldsymbol{g}^{(t)}\right)$ and the global optimum $\left(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}}\right)$.

Theorem 5 (Stochastic noise with imbalanced samples)**.** *Suppose the conditions (i)–(iii) in Theorem 3 hold for some suitable constants. The following results hold for any (n_1, n_2) that satisfies $n_1/n_2 = \Theta(1)$.*

(a) *(Statistical error) There exist positive constants $\{c_i\}_{i=1}^4$ such that if $\gamma/\sigma \ge c_1$, then with probability at least $1 - c_3 n^{-c_4}$, the global optimum $\left(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}}\right)$ of program (8) satisfies the bounds*

$$\left\|\hat{\boldsymbol{K}} - \boldsymbol{K}^*\right\|_F \le c_2 \gamma\sigma \sqrt{\frac{p}{n}} \log^3 n,$$
$$\|\hat{\boldsymbol{g}} - \boldsymbol{g}^*\|_2 \le c_2 \sigma \sqrt{\frac{p}{n}} \log^3 n.$$

(b) *(Optimization error) Let $\boldsymbol{H}^{(t)} := \boldsymbol{K}^{(t)} - \hat{\boldsymbol{K}}$ and $\boldsymbol{h}^{(t)} := \boldsymbol{g}^{(t)} - \hat{\boldsymbol{g}}$ be the optimization error terms at step t of the projected gradient descent algorithm (9) for the program (8). There exist positive constants $\{c_i\}_{i=1}^4$ such that if $\gamma/\sigma \ge c_1$ and $\eta \ge c_2 n (\sqrt{p})^3 \log n$, then with probability at least $1 - c_3 n^{-c_4}$, there holds the bound*

$$\begin{aligned}&\|\boldsymbol{H}^{(t)}\|_F^2 + \gamma^2 \|\boldsymbol{h}^{(t)}\|_2^2 \\ &\le \left(1 - \frac{c_3 n}{\eta}\right)^t \left(\|\boldsymbol{H}^{(0)}\|_F^2 + \gamma^2 \|\boldsymbol{h}^{(0)}\|_2^2\right) + c_4 \|\hat{\boldsymbol{K}} - \boldsymbol{K}^*\|_F^2\end{aligned}$$

for every $t = 1, 2, \ldots$

Part (b) establishes geometric convergence of the optimization error. Thus, the iterate $(\boldsymbol{K}^{(t)}, \boldsymbol{g}^{(t)})$ quickly converges to some solution whose distance to $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ is of the same order as the statistical error. Therefore, the distance between $(\boldsymbol{K}^{(t)}, \boldsymbol{g}^{(t)})$ and $(\boldsymbol{K}^*, \boldsymbol{g}^*)$ satisfies the bound in part (a) when t is sufficiently large by the triangle inequality. If we choose the step size $\eta \asymp n(\sqrt{p})^3 \log n$, the contractive factor is roughly $1 - 1/((\sqrt{p})^3 \log n)$, hence $T \asymp (\sqrt{p})^3 \log n$ iterations of projected gradient descent suffice. We can then apply Algorithm 1 to the T^{th} iterate $(\boldsymbol{K}^{(T)}, \boldsymbol{g}^{(T)})$ to obtain an accurate estimate of the true regression vectors $\boldsymbol{\theta}^* = (\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$, as shown in the theorem below.

Theorem 6 (Estimating $\boldsymbol{\beta}_b^*$, stochastic noise with imbalanced samples)**.** *Suppose that the projected gradient descent in (9) is initialized with $\boldsymbol{K}^{(0)} = \boldsymbol{0}$ and $\boldsymbol{g}^{(0)} = \boldsymbol{0}$. Under the setting of Theorem 5, there exist positive constants $\{c_i\}_{i=1}^7$ such that if $n \ge c_1 p \log^6 n$, $\eta \ge c_2 n (\sqrt{p})^3 \log n$ and*

$$T \ge c_3 \eta n^{-1} \max\left\{1 - c_4 \sigma^2 \gamma^{-2} \frac{p}{n} \log^6 n,\ 0\right\},$$

then with probability at least $1 - c_5 n^{-c_6}$, the output $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ of Algorithm 1 applied to the input $(\boldsymbol{K}^{(T)}, \boldsymbol{g}^{(T)})$ satisfies the error bound

$$\rho(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \le c_7 \sigma \sqrt{\frac{p}{n}} \log^3 n.$$

Remark 1 (Scalability)**.** *Both the formulations (5)–(6) and (7) can be cast as Semidefinite Programs (SDP). In the arbitrary noise setting, the constraint in the convex program (5)–(6) can be rewritten as a collection of linear constraints through the standard transformation of convex ℓ_1 constraints. The Lagrangian formulation (7) in the setting of stochastic noise,*

involves minimizing the sum of a trace norm term and a smooth quadratic term. The computational complexity of solving this regularized quadratic optimization in the matrix space has similar complexity to problems such as matrix completion [9] and PhaseLift [10], and various first order methods can easily be adapted, thus allowing solution of large scale instances of the mixed regression problem.

D. Minimax Lower Bounds

We now derive minimax lower bounds on the estimation errors for both the arbitrary and stochastic noise settings, and show that these match our upper bounds. Recall the error (semi)-metric $\rho\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right)$ defined in equation (2) for a pair $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $\boldsymbol{\theta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ in $\mathbb{R}^p \times \mathbb{R}^p$. It is straightforward to verify that the metric $\rho(\cdot, \cdot)$ satisfies the triangle inequality. An estimator

$$\hat{\boldsymbol{\theta}} \equiv \hat{\boldsymbol{\theta}}\left(\boldsymbol{X}, \boldsymbol{y}\right) = \left(\hat{\boldsymbol{\beta}}_1(\boldsymbol{X}, \boldsymbol{y}), \hat{\boldsymbol{\beta}}_2(\boldsymbol{X}, \boldsymbol{y})\right).$$

of the true regression vectors $\boldsymbol{\theta}^* = (\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$ is any measurable function of the observed data $(\boldsymbol{X}, \boldsymbol{y})$. For each number $\underline{\gamma} > 0$, we consider the following parameter class

$$\Theta(\underline{\gamma}) := \Big\{\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \in \mathbb{R}^p \times \mathbb{R}^p : \\ 2\left\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\right\|_2 \geq \left\|\boldsymbol{\beta}_1\right\|_2 + \left\|\boldsymbol{\beta}_2\right\|_2 \geq \underline{\gamma}\Big\}, \tag{10}$$

that is, pairs of regression vectors whose norms and separation are lower bounded. Recall that $z_i \in \{0, 1\}$ is the hidden label for the i-th observation, that is, $z_i = 0$ if and only if $i \in \mathcal{I}_1$, for each $i = 1, 2, \ldots, n$.

1) Lower Bounds for Arbitrary Noise: In the arbitrary noise setting, the noise vector $\boldsymbol{e}$ is assumed to lie in the ℓ_2-ball $\mathbb{B}(\epsilon) := \{\boldsymbol{u} \in \mathbb{R}^n : \|\boldsymbol{u}\|_2 \leq \epsilon\}$ and otherwise arbitrary. For this setting we have the following minimax lower bounds.

Theorem 7 (Lower bound, arbitrary noise)**.** *There exist universal positive constants c_0, c_1 for which the following is true. If $n \geq c_1 p$, then for any $\underline{\gamma} > 0$ and any hidden label vector $\boldsymbol{z} \in \{0, 1\}^n$, there holds the bound*

$$\inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}^* \in \Theta(\underline{\gamma})} \sup_{\boldsymbol{e} \in \mathbb{B}(\epsilon)} \rho(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \geq c_0 \frac{\epsilon}{\sqrt{n}} \tag{11}$$

with probability at least $1 - n^{-10}$, where the probability is with respect to the randomness in $\boldsymbol{X}$.

The lower bound above matches the upper bound given in Theorem 2, thus showing that our convex formulation is minimax optimal and order-wise unimprovable. Informally, Theorems 2 and 7 together establish the following minimax rate of the arbitrary noise setting

$$\rho(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \asymp \frac{\|\boldsymbol{e}\|_2}{\sqrt{n}},$$

which is valid for $n \gtrsim p$.

2) Lower Bounds for Stochastic Noise: For the stochastic setting where the noise is i.i.d. Gaussian, we further assume that the two components have equal mixing weights: $\mathbb{P}(z_i = 0) = \mathbb{P}(z_i = 1) = 1/2$ for each $i = 1, 2, \ldots, n$. For this setting we have the following minimax lower bounds.

Theorem 8 (Lower bound, stochastic noise)**.** *Suppose that $n \geq p \geq 64$, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ has i.i.d. standard Gaussian entries, $\boldsymbol{e}$ has i.i.d. zero-mean Gaussian entries with variance σ^2, and $z_i \overset{iid}{\sim}$ Bernoulli$(1/2)$. The following statements hold for some absolute constants $0 < c_0, c_1 < 1$.*

1) *For each $\underline{\gamma} > \sigma$, we have*

$$\inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}^* \in \Theta(\underline{\gamma})} \mathbb{E}_{\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{e}} \left[\rho(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\right] \geq c_0 \sigma \sqrt{\frac{p}{n}}. \tag{12}$$

2) *For each $c_1 \sigma \left(\frac{p}{n}\right)^{1/4} \leq \underline{\gamma} \leq \sigma$, we have*

$$\inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}^* \in \Theta(\underline{\gamma})} \mathbb{E}_{\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{e}} \left[\rho(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\right] \geq c_0 \frac{\sigma^2}{\underline{\gamma}} \sqrt{\frac{p}{n}}. \tag{13}$$

3) *For each $0 < \underline{\gamma} \leq c_1 \sigma \left(\frac{p}{n}\right)^{1/4}$, we have*

$$\inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}^* \in \Theta(\underline{\gamma})} \mathbb{E}_{\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{e}} \left[\rho(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\right] \geq c_0 \sigma \left(\frac{p}{n}\right)^{1/4}. \tag{14}$$

Here $\mathbb{E}_{\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{e}}[\cdot]$ denotes the expectation with respect to the covariate matrix $\boldsymbol{X}$, the hidden label vector $\boldsymbol{z}$ and the noise vector $\boldsymbol{e}$.

We see that the three lower bounds in the above theorem match each of the three terms in the upper bound given in Theorem 4 up to at most polylog factors, proving the minimax optimality of the error bounds of our convex formulation. Informally, Theorems 4 and 8 together establish the following minimax error rate (up to a polylog factor) in the stochastic noise setting:

$$\rho(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) \asymp \begin{cases} \sigma\sqrt{\frac{p}{n}}, & \text{if } \underline{\gamma} \gtrsim \sigma, \\ \frac{\sigma^2}{\underline{\gamma}}\sqrt{\frac{p}{n}}, & \text{if } \sigma\left(\frac{p}{n}\right)^{\frac{1}{4}} \lesssim \underline{\gamma} \lesssim \sigma, \\ \sigma\left(\frac{p}{n}\right)^{\frac{1}{4}}, & \text{if } \underline{\gamma} \lesssim \sigma\left(\frac{p}{n}\right)^{\frac{1}{4}}. \end{cases}$$

Here, $\underline{\gamma}$ is any lower bound on $\|\boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^*\|_2$ and represents the signal strength (recall the definition of the parameter class in equation (10)). Notice how the scaling of the minimax error rate exhibits three phases depending on the SNR $\underline{\gamma}/\sigma$: (i) In the high SNR regime with $\underline{\gamma} \gtrsim \sigma$, we see a fast rate — proportional to $1/\sqrt{n}$ — that is dominated by the error of estimating a single $\boldsymbol{\beta}_b^*$ and is the same as the rate for standard linear regression. (ii) In the low SNR regime with $\underline{\gamma} \lesssim \sigma\left(\frac{p}{n}\right)^{\frac{1}{4}}$, we have a slow rate that is proportional to $1/n^{\frac{1}{4}}$, which is associated with the demixing of the two components $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$. (iii) In the medium SNR regime, the error rate transitions between the fast and slow phases, and depends in a precise way on the SNR. For related phenomena, see the work in [3, 14].

The lower bounds in Theorem 8 apply to the balanced sample setting. The error upper bound in Theorem 6 for our nonconvex approach in the imbalanced setting, is also near-optimal in the high SNR regime, as this bound matches (up

to logarithmic factors) the minimax lower bound of standard linear regression for estimating a single $\boldsymbol{\beta}_b^*$ (that is, when the labels are known).

E. Implications for Phase Retrieval

As an illustration of the power of our results, we discuss an application to the *Phase Retrieval* problem, which has recently received much attention (see, e.g., the work in [6, 7, 11, 16, 23, 15]). Recall that in the real value setting, the phase retrieval problem is essentially a regression problem without sign information. Recent work has mostly focused on the noiseless case. Here, the problem is as follows: we observe $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, 2, \dots n$, where

$$y_i = \left|\boldsymbol{x}_i^\top \boldsymbol{\beta}^*\right|.$$

The goal is to recover the unknown vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$. The stability of recovery algorithms has also been considered. Most work has focused on the setting where noise is added to the phase-less measurements, that is,

$$y_i = \left|\boldsymbol{x}_i^\top \boldsymbol{\beta}^*\right| + e_i. \tag{15}$$

In many applications, however, it is also natural to consider the setting where the measurement noise is added before the phase is lost. This corresponds to the model

$$y_i = \left|\boldsymbol{x}_i^\top \boldsymbol{\beta}^* + e_i\right|. \tag{16}$$

We may call formulation (16) the *Noisy Phase Model*, as opposed to the *Noisy Magnitude Model* in (15) considered by previous work. This problem can be reduced to a mixed regression problem and solved by our algorithm. The reduction is as follows. We generate n independent Rademacher random variables $\epsilon_i, i = 1, \dots, n$. For each i, we set $y_i' = \epsilon_i y_i$. Let $s_i := \operatorname{sign}\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}^* + e_i\right)$ and $e_i' := \epsilon_i s_i e_i$, where we use the convention that $\operatorname{sign}(0) = 1$. Under the noisy phase model (16), we then have

$$y_i' = \epsilon_i \cdot \left|\boldsymbol{x}_i^\top \boldsymbol{\beta}^* + e_i\right| = \epsilon_i s_i \left(\boldsymbol{x}_i^\top \boldsymbol{\beta}^* + e_i\right) = \boldsymbol{x}_i^\top \left(\epsilon_i s_i \boldsymbol{\beta}^*\right) + e_i'.$$

If we let $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}^*$, $\boldsymbol{\beta}_2^* = -\boldsymbol{\beta}^*$, $\mathcal{I}_1 = \{i : \epsilon_i s_i = 1\}$ and $\mathcal{I}_2 = \{i : \epsilon_i s_i = -1\}$, then the model becomes

$$y_i' = \boldsymbol{x}_i^\top \boldsymbol{\beta}_b^* + e_i', \quad \forall i \in \mathcal{I}_b,$$

which is precisely the mixed regression model we considered.

Note that with probability at least $1 - n^{-3}$, $\frac{n}{2} - \sqrt{10 n \log n} \le n_b = |\mathcal{I}_b| \le \frac{n}{2} + \sqrt{10 n \log n}$ for $b = 1, 2$, so $|n_1 - n_2| = O\left(\sqrt{n \log n}\right)$. Also note that $\|\boldsymbol{e}'\|_2 = \|\boldsymbol{e}\|_2$. Conditioned on $\{\mathcal{I}_b\}$, the distribution of $\{\boldsymbol{x}_i\}$ is the same as its unconditional distribution. Therefore, applying our arbitrary-noise result from Theorem 2, we immediately get the following guarantees for phase retrieval under the Noisy Phase Model.

Corollary 2 (Phase retrieval, arbitrary noise)**.** *Consider the Noisy Phase Model in (16). Suppose that the $\{\boldsymbol{x}_i\}$ are i.i.d., zero-mean sub-Gaussian with bounded sub-Gaussian norm, unit variance and fourth moment $\mu > 1$, and that $n \gtrsim p$, $\eta \asymp c_4 \sqrt{n} \|\boldsymbol{e}\|_2 \|\boldsymbol{\beta}^*\|_2$ and the noise is arbitrary but bounded in magnitude: $\|\boldsymbol{e}\|_2 \lesssim \sqrt{n} \|\boldsymbol{\beta}^*\|_2$. Then using the reduction described above, the output of the program (5)–(6) followed by Algorithm 1 satisfies the error bound*

$$\min_{b=1,2} \left\|\hat{\boldsymbol{\beta}}_b - \boldsymbol{\beta}^*\right\|_2 \lesssim \frac{\|\boldsymbol{e}\|_2}{\sqrt{n}}$$

with probability at least $1 - n^{-2}$.

In the corollary above we assumed $\|\boldsymbol{e}\|_2 \lesssim \sqrt{n} \|\boldsymbol{\beta}^*\|_2$. Similarly as before, the large noise case with $\|\boldsymbol{e}\|_2 \ge c_4 \sqrt{n} \|\boldsymbol{\beta}^*\|_2$ can be handled easily, using the blind estimator $\hat{\boldsymbol{\beta}} := \min_{\boldsymbol{\beta}} \sum_{i \in [n]} \left|\boldsymbol{x}_i^\top \boldsymbol{\beta} - y_i\right|$, which in this case again satisfies the optimal error bound $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \lesssim \frac{\|\boldsymbol{e}\|_2}{\sqrt{n}}$.

Next, consider the stochastic noise case where e_i is i.i.d., zero-mean symmetric sub-Gaussian with variance σ^2. Conditioned on $\{\mathcal{I}_b\}$, the conditional distributions of $\{e_i'\}$ and $\{\boldsymbol{x}_i\}$ inherit the properties of e_i and the unconditional $\boldsymbol{x}_i$, and are independent of each other. Applying Theorem 4, we have the following guarantee.

Corollary 3 (Phase retrieval, stochastic noise)**.** *Consider the Noisy Phase Model in (16). Suppose that the $\{\boldsymbol{x}_i\}$ are i.i.d., zero-mean Gaussian with unit variance, and that the noise e_i is i.i.d., zero-mean symmetric sub-Gaussian with sub-Gaussian norm bounded by $c_3\sigma$ and variance equal to σ^2. Suppose further that $n \gtrsim p$ and $\lambda \asymp \sigma\left(\|\boldsymbol{\beta}^*\|_2 + \sigma\right)\sqrt{np}\log^4 n$. Then using the reduction described above, the output of the program (7) followed by Algorithm 1 satisfies the error bound*

$$\begin{aligned}&\min_{b=1,2} \left\|\hat{\boldsymbol{\beta}}_b - \boldsymbol{\beta}^*\right\|_2 \\ &\lesssim \sigma\sqrt{\frac{p}{n}}\log^4 n + \min\left\{\frac{\sigma^2\sqrt{\frac{p}{n}}}{\|\boldsymbol{\beta}^*\|_2}, \sigma\left(\frac{p}{n}\right)^{\frac{1}{4}}\right\}\log^4 n\end{aligned}$$

with probability at least $1 - n^{-2}$.

As a passing observation, we note that the error bounds above for phase retrieval are both order-wise optimal. For the deterministic noise setting considered in Corollary 2, we cannot achieve a smaller error even if the phase is not lost. For the stochastic noise setting, note that our corresponding minimax lower bounds for mixed regression in Section III-D2 is in fact derived under the symmetric setting $\boldsymbol{\beta}_1^* = -\boldsymbol{\beta}_2^*$. In this case one can reduce a mixed regression problem to a phase retrieval problem by dropping the signs in $\{y_i\}$, hence our minimax lower bounds certify the near-optimality of Corollary 3 for phase retrieval.

IV. Proofs

We now provide the proofs of the main Theorems. Conceptually, there are three parts to our results. *i) Regression error* (Theorems 1, 3, 5). We prove that the lifted optimizations in the matrix space (for deterministic noise, and balanced and unbalanced stochastic noise) recover good approximations to $\boldsymbol{K}^*$ and $\boldsymbol{g}^*$. We note that these results have no dependence on SNR; that is, alone, they do not reveal a change in the rate of convergence. *ii) Decomposition and perturbation error* (Theorems 2, 4, 6). We prove a matrix perturbation result that shows that a good approximation of $\boldsymbol{K}^*$ and $\boldsymbol{g}^*$ results in a

good approximation of $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$. The change in rate in the low-SNR setting, emerges from these decomposition results. *iii) Minimax lower bounds* (Theorems 7 and 8). We prove minimax (information theoretic) lower bounds matching the upper bounds. For the lower bounds, the change in rate comes from a detection problem that involves distinguishing mixtures of Gaussians with different separation of the component centers.

We prove these results here, and in the interest of readability, defer the proofs of technical lemmas to the appendix.

A. Notation and Preliminaries

We use $\boldsymbol{\beta}_{-b}^*$ to denote $\boldsymbol{\beta}_2^*$ if $b = 1$ and $\boldsymbol{\beta}_1^*$ if $b = 2$. Let $\boldsymbol{\delta}_b^* := \boldsymbol{\beta}_b^* - \boldsymbol{\beta}_{-b}^*$. Without loss of generality, we assume $\mathcal{I}_1 = \{1, \ldots, n_1\}$ and $\mathcal{I}_2 = \{n_1 + 1, \ldots, n\}$. For $i = 1, \ldots, n_1$, we define $\boldsymbol{x}_{1,i} := \boldsymbol{x}_i$, $y_{1,i} = y_i$ and $e_{1,i} = e_i$; correspondingly, for $i = 1, \ldots, n_2$, we define $\boldsymbol{x}_{2,i} := \boldsymbol{x}_{n_1+i}$, $y_{2,i} := y_{n_1+i}$ and $e_{2,n+i}$. For each $b = 1, 2$, let $\boldsymbol{X}_b \in \mathbb{R}^{n_b \times p}$ be the matrix with rows $\{\boldsymbol{x}_{b,i}^\top, i = 1, \ldots, n_b\}$. Also let $\boldsymbol{e}_b := [e_{b,1} \ \cdots \ e_{b,n_b}]^\top \in \mathbb{R}^{n_b}$.

While the measurements in the original model are given by $\boldsymbol{X}$, in the lifted space, one can regard the measurements as given by rank-one matrices that are quadratic in $\boldsymbol{x}_{b,i}$. Thus it is natural to define the matrices $\boldsymbol{A}_{b,i} := \boldsymbol{x}_{b,i}\boldsymbol{x}_{b,i}^\top$, $i \in [n_b]$ and the mapping $\mathcal{A}_b : \mathbb{R}^{p\times p} \mapsto \mathbb{R}^{n_b}$ given by

$$(\mathcal{A}_b \boldsymbol{Z})_i = \frac{1}{n_b} \langle \boldsymbol{A}_{b,i}, \boldsymbol{Z} \rangle, \quad \text{for each } i \in [n_b].$$

Because of their quadratic nature, these measurements are not mean-zero. As we detail below, our proofs rely on establishing a restricted isometry-like property of the measurements, but as the measurements are not zero mean, this does not hold. It is convenient, therefore, to define matrices and a mapping that are related to $\boldsymbol{A}_{b,i}$ and $\mathcal{A}_b$, but with zero mean. To this end, for $b = 1, 2$ and $j = 1, \ldots, \lfloor n_b/2 \rfloor$, define the matrix $\boldsymbol{B}_{b,j} := \boldsymbol{x}_{b,2j}\boldsymbol{x}_{b,2j}^\top - \boldsymbol{x}_{b,2j-1}\boldsymbol{x}_{b,2j-1}^\top$, as well as the vector $\boldsymbol{d}_{b,j} = e_{b,2j}\boldsymbol{x}_{b,2j} - e_{b,2j-1}\boldsymbol{x}_{b,2j-1}$. For $b \in \{1, 2\}$, define the mapping $\mathcal{B}_b : \mathbb{R}^{p\times p} \mapsto \mathbb{R}^{\lfloor n_b/2 \rfloor}$ by

$$(\mathcal{B}_b \boldsymbol{Z})_j = \frac{1}{\lfloor n_b/2 \rfloor} \langle \boldsymbol{B}_{b,j}, \boldsymbol{Z} \rangle, \quad \text{for each } j = 1, \ldots, \lfloor n_b \rfloor.$$

Since $y_{b,i} = \boldsymbol{x}_{b,i}^\top \boldsymbol{\beta}_b^* + e_{b,i}$, $i \in [n_b]$, we have for any $\boldsymbol{Z} \in \mathbb{R}^{p\times p}$, $\boldsymbol{z} \in \mathbb{R}^p$ and for all $j = 1, \ldots, \lfloor n_b \rfloor$,

$$\begin{aligned}
&\frac{1}{\lfloor n_b/2 \rfloor} \left(\langle \boldsymbol{B}_{b,j}, \boldsymbol{Z} \rangle - 2\boldsymbol{d}_{b,j}^\top \boldsymbol{z} \right) \\
=&\frac{1}{\lfloor n_b/2 \rfloor} \left\langle \boldsymbol{B}_{b,j}, \boldsymbol{Z} - 2\boldsymbol{\beta}_b^* \boldsymbol{z}^\top \right\rangle + (e_{b,2j}\boldsymbol{x}_{b,2j} - e_{b,2j-1}\boldsymbol{x}_{b,2j})^\top \boldsymbol{z} \\
=&\left(\mathcal{B}_b \left(\boldsymbol{Z} - 2\boldsymbol{\beta}_b^* \boldsymbol{z}^\top\right)\right)_j + (e_{b,2j}\boldsymbol{x}_{b,2j} - e_{b,2j-1}\boldsymbol{x}_{b,2j})^\top \boldsymbol{z}.
\end{aligned}$$

A key part of the proof is in expressing the error in terms of the operators $\mathcal{A}_b$ and then $\mathcal{B}_b$, and then showing that $\mathcal{B}_b$ satisfies a restricted isometry property. Also key and common to all the proofs, is to show that the optimization formulations recover a near low-rank matrix $\hat{\boldsymbol{K}}$. For this, we need to control the part of $\hat{\boldsymbol{K}}$ that has different column and row space from $\boldsymbol{K}^*$. The following notation and definitions are standard. Let the rank-2 SVD of $\boldsymbol{K}^*$ be $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. Note that $\boldsymbol{U}$ and $\boldsymbol{V}$ have the same column space, which equals $\text{span}(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$. Define the projection matrix $\boldsymbol{P}_{\boldsymbol{U}} := \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{V}\boldsymbol{V}^\top$ and the subspace $T := \{\boldsymbol{P}_{\boldsymbol{U}}\boldsymbol{Z} + \boldsymbol{Y}\boldsymbol{P}_{\boldsymbol{U}} : \boldsymbol{Z}, \boldsymbol{Y} \in \mathbb{R}^{p\times p}\}$. Let $T^\perp$ be the orthogonal subspace of T. The projections to T and $T^\perp$ are given by

$$\mathcal{P}_T \boldsymbol{Z} := \boldsymbol{P}_{\boldsymbol{U}}\boldsymbol{Z} + \boldsymbol{Z}\boldsymbol{P}_{\boldsymbol{U}} - \boldsymbol{P}_{\boldsymbol{U}}\boldsymbol{Z}\boldsymbol{P}_{\boldsymbol{U}}, \quad \mathcal{P}_{T^\perp} \boldsymbol{Z} := \boldsymbol{Z} - \mathcal{P}_T \boldsymbol{Z}.$$

Denote the optimal solution to the optimization problem of interest (either (5) or (7)) as $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}}) = (\boldsymbol{K}^* + \hat{\boldsymbol{H}}, \boldsymbol{g}^* + \hat{\boldsymbol{h}})$. Let $\hat{\boldsymbol{H}}_T := \mathcal{P}_T \hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{H}}_{T^\perp} := \mathcal{P}_{T^\perp} \hat{\boldsymbol{H}}$.

The optimization proofs follow a similar spirit and conceptual flow. The first part of the proof asserts that the error (compared to $(\boldsymbol{K}^*, \boldsymbol{g}^*)$ must satisfy certain conditions controlled by the operators $\mathcal{B}_b$, $b = 1, 2$. For Theorem 1 this is a consequence of the constraints; for Theorem 3 this is a consequence of the objective function. Essentially these results say that the errors must lie in certain directions away from $(\boldsymbol{K}^*, \boldsymbol{g}^*)$.

The next step comes in using properties of $\mathcal{B}_b$. In particular, we show that these operators satisfy a restricted isometry-like (RIP) condition. Together with the characterization of how $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ and $(\boldsymbol{K}^*, \boldsymbol{g}^*)$ can differ, we conclude that along all those directions, the objective function has strong convexity, that is, curvature bounded from below). This allows us to provide bounds on how far $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ can be from $(\boldsymbol{K}^*, \boldsymbol{g}^*)$.

B. Proof of Theorem 1

As outlined above, the proof follows from three main steps. First, the ℓ_1 error term that in this formulation appears in the LHS of the constraint (6) in the optimization, is naturally related to the operators $\mathcal{A}_b$. Using the definitions of $\mathcal{A}_b$, $\mathcal{B}_b$, we establish a relation between $\boldsymbol{e}, \eta$ and the feasibility of the optimal solution $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}}) = (\boldsymbol{K}^* + \hat{\boldsymbol{H}}, \boldsymbol{g}^* + \hat{\boldsymbol{h}})$. Second, we show that the operator $\mathcal{B}$ is an approximate isometry on low-rank matrices. Finally, this allows us to obtain good upper and lower bounds on the error terms, and hence the accuracy of the solution.

Now for the details. Recall that $\hat{\boldsymbol{H}}_T := \mathcal{P}_T \hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{H}}_{T^\perp} := \mathcal{P}_{T^\perp} \hat{\boldsymbol{H}}$. Note that $\hat{\boldsymbol{H}}_T$ has rank at most 4 and $\hat{\boldsymbol{H}}_{T^\perp}$ has rank at most $p - 4$. We have

$$\begin{aligned}
\left\|\hat{\boldsymbol{K}}\right\|_* - \|\boldsymbol{K}^*\|_* &\geq \left\|\boldsymbol{K}^* + \hat{\boldsymbol{H}}_{T^\perp}\right\|_* - \left\|\hat{\boldsymbol{H}}_T\right\|_* - \|\boldsymbol{K}^*\|_* \\
&= \left\|\hat{\boldsymbol{H}}_{T^\perp}\right\|_* - \left\|\hat{\boldsymbol{H}}_T\right\|_*. \qquad (17)
\end{aligned}$$

1) Step (1): Consequence of Feasibility: For any $(\boldsymbol{K}, \boldsymbol{g}) = (\boldsymbol{K}^* + \boldsymbol{H}, \boldsymbol{g}^* + \boldsymbol{h})$, it is easy to check that

$$\begin{aligned}
&- \left\langle \boldsymbol{x}_{b,i}\boldsymbol{x}_{b,i}^\top, \boldsymbol{K} \right\rangle + 2y_{b,i} \langle \boldsymbol{x}_{b,i}, \boldsymbol{g} \rangle - y_{b,i}^2 \\
=&- \left\langle \boldsymbol{x}_{b,i}\boldsymbol{x}_{b,i}^\top, \boldsymbol{H} \right\rangle + 2y_{b,i} \langle \boldsymbol{x}_{b,i}, \boldsymbol{h} \rangle - e_{b,i}\boldsymbol{x}_{b,i}^\top \boldsymbol{\delta}_b^* - e_{b,i}^2. \qquad (18)
\end{aligned}$$

Therefore, the constraint (6) is equivalent to

$$\begin{aligned}
&\sum_{b=1}^{2} \sum_{i=1}^{n_b} \Big| - \left\langle \boldsymbol{x}_{b,i}\boldsymbol{x}_{b,i}^\top, \boldsymbol{H} \right\rangle + 2\left(\boldsymbol{x}_{b,i}^\top \boldsymbol{\beta}_b^* + e_{b,i}\right) \langle \boldsymbol{x}_{b,i}, \boldsymbol{h} \rangle \\
&- e_{b,i}\boldsymbol{x}_{b,i}^\top \boldsymbol{\delta}_b^* - e_{b,i}^2 \Big| \leq \eta.
\end{aligned}$$

Using the notation from Section IV-A, this can be rewritten as

$$\sum_b \left\| n_b \mathcal{A}_b \left(-\boldsymbol{H} + 2\boldsymbol{\beta}_b^* \boldsymbol{h}^\top \right) + 2\boldsymbol{e}_b \circ (\boldsymbol{X}_b \boldsymbol{h}) - \boldsymbol{e}_b \circ (\boldsymbol{X}_b \boldsymbol{\delta}_b^*) - \boldsymbol{e}_b^2 \right\|_1 \leq \eta, \tag{19}$$

where $\circ$ denotes the element-wise product and $\boldsymbol{e}_b^2 = \boldsymbol{e}_b \circ \boldsymbol{e}_b$.

First, note that $\boldsymbol{K}^*$ and $\boldsymbol{g}^*$ are feasible. By standard bounds on the spectral norm of random matrices [30], we know that with probability at least $1 - 2\exp(-cn_b)$,

$$\|\boldsymbol{X}_b \boldsymbol{z}\|_2 \lesssim \sqrt{n_b} \|\boldsymbol{z}\|_2, \forall \boldsymbol{z} \in \mathbb{R}^p.$$

We thus have

$$\begin{aligned} \left\| -\boldsymbol{e}_b \circ (\boldsymbol{X}_b \boldsymbol{\delta}_b^*) - \boldsymbol{e}_b^2 \right\|_1 &\leq c_1 \left(\sqrt{n_b} \|\boldsymbol{e}_b\|_2 \|\boldsymbol{\delta}_b^*\|_2 + \|\boldsymbol{e}\|_2^2 \right) \\ &\overset{(a)}{\leq} c_1 \sqrt{n_b} \|\boldsymbol{e}\|_2 \|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|_2 \\ &\overset{(b)}{\leq} \eta, \end{aligned}$$

where we use the assumptions on $\boldsymbol{e}$ and η in the steps (a) and (b), respectively. The inequality above implies that (19) holds with $\boldsymbol{H} = \boldsymbol{0}$ and $\boldsymbol{h} = \boldsymbol{0}$, thus showing the feasibility of $(\boldsymbol{K}^*, \boldsymbol{g}^*)$.

Since $\left(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}}\right)$ is feasible by assumption, combining the last two displayed equations and (19), we further have

$$\begin{aligned} &\sum_b \left\| n_b \mathcal{A}_b \left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right) \right\|_1 \\ \leq &\sum_b \left\| 2\boldsymbol{e}_b \circ \left(\boldsymbol{X}_b \hat{\boldsymbol{h}} \right) \right\|_1 + \sum_b \left\| -2\boldsymbol{e}_b \circ (\boldsymbol{X}_b \boldsymbol{\delta}_b^*) - \boldsymbol{e}_b^2 \right\|_1 + \eta \\ \leq &c_2 \sum_b \sqrt{n_b} \|\boldsymbol{e}_b\|_2 \|\hat{\boldsymbol{h}}\|_2 + 2\eta. \end{aligned} \tag{20}$$

Now from the definition of $\mathcal{A}_b$ and $\mathcal{B}_b$, we have

$$\begin{aligned} &\lfloor n_b/2 \rfloor \left\| \mathcal{B}_b \left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right) \right\|_1 \\ \leq &\sum_{j=1}^{\lfloor n_b/2 \rfloor} \left\| \left\langle \boldsymbol{A}_{b,2j}, -\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right\rangle \right\|_1 \\ &+ \left\| \left\langle \boldsymbol{A}_{b,2j-1}, -\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right\rangle \right\|_1 \\ \leq &n_b \left\| \mathcal{A}_b \left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right) \right\|_1. \end{aligned}$$

It follows from (20) and $n_1 \asymp n_2 \asymp n$ that

$$\sum_b n \left\| \mathcal{B}_b \left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{h}^\top \right) \right\|_1 - c_2 \sum_b \sqrt{n} \|\boldsymbol{e}_b\|_2 \|\hat{\boldsymbol{h}}\|_2 \leq 2\eta. \tag{21}$$

This concludes Step (1) of the proof.

2) Step (2): RIP and Lower Bounds: The bound in (21) relates the ℓ_1-norm of $\mathcal{B}$ and η. Since we want a bound on the ℓ_2 and Frobenius norms of $\hat{\boldsymbol{h}}$ and $\hat{\boldsymbol{H}}$ respectively, a major step is the proof of an RIP-like property for $\mathcal{B}$:

Lemma 1. *The following holds for some numerical constants $c, \underline{\delta}, \bar{\delta}$. For $b = 1, 2$, if $\mu > 1$ and $n_b \geq c\rho p$, then with probability $1 - \exp(-n_b)$, we have*

$$\underline{\delta} \|\boldsymbol{Z}\|_F \leq \|\mathcal{B}_b \boldsymbol{Z}\|_1 \leq \bar{\delta} \|\boldsymbol{Z}\|_F$$

simultaneously for all $\boldsymbol{Z} \in \mathbb{R}^{p \times p}$ with rank$(\boldsymbol{Z}) \leq \rho$.

This lemma follows from the more general Lemma 5 that appears in the proof of Theorem 3, by setting $\sigma = 0$.

We now turn to the implications of Lemma 1, in order to get lower bounds on the term $\left\| \mathcal{B}_b \left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \boldsymbol{h}^\top \right) \right\|_1$ from the first term in (21), in terms of $\|\hat{\boldsymbol{h}}\|_2$ and $\|\hat{\boldsymbol{H}}\|_F$.

Since we have proved that $(\boldsymbol{K}^*, \boldsymbol{g}^*)$ is feasible, we have $\left\| \hat{\boldsymbol{K}} \right\|_* \leq \|\boldsymbol{K}^*\|_*$ by optimality. It follows from (17) that

$$\left\| \hat{\boldsymbol{H}}_{T^\perp} \right\|_* \leq \left\| \hat{\boldsymbol{H}}_T \right\|_*. \tag{22}$$

Let $K = c\frac{1}{\alpha}$ for c some numeric constant to be chosen later. We can partition $\hat{\boldsymbol{H}}_{T^\perp}$ into a sum of $M := \frac{p-4}{K}$ matrices $\hat{\boldsymbol{H}}_1, \ldots, \hat{\boldsymbol{H}}_M$ according to the SVD of $\hat{\boldsymbol{H}}_{T^\perp}$, such that rank$(\hat{\boldsymbol{H}}_i) \leq K$ and the smallest singular value of $\hat{\boldsymbol{H}}_i$ is larger than the largest singular value of $\hat{\boldsymbol{H}}_{i+1}$ (cf. [24]). By Lemma 1, we get that for each $b = 1, 2$,

$$\begin{aligned} \sum_{i=2}^M \left\| \mathcal{B}_b(\hat{\boldsymbol{H}}_i) \right\|_1 \leq &\bar{\delta} \sum_{i=2}^M \left\| \hat{\boldsymbol{H}}_i \right\|_F \leq \bar{\delta} \sum_{i=2}^M \frac{1}{\sqrt{K}} \left\| \hat{\boldsymbol{H}}_{i-1} \right\|_* \\ \leq &\frac{\bar{\delta}}{\sqrt{K}} \left\| \hat{\boldsymbol{H}}_{T^\perp} \right\|_* \overset{(a)}{\leq} \frac{\bar{\delta}}{\sqrt{K}} \sqrt{4} \left\| \hat{\boldsymbol{H}}_T \right\|_F, \end{aligned} \tag{23}$$

where (a) follows from (22) and the rank of $\hat{\boldsymbol{H}}_T$. It follows that for $b = 1, 2$,

$$\begin{aligned} &\left\| \mathcal{B}_b \left(\hat{\boldsymbol{H}} - 2\boldsymbol{\beta}_b^* \hat{h}^\top \right) \right\|_1 \\ \overset{(a)}{\geq} &\left\| \mathcal{B}_b \left(\hat{\boldsymbol{H}}_T + \hat{\boldsymbol{H}}_1 - 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right) \right\|_1 - \sum_{i=2}^M \left\| \mathcal{B}(\hat{\boldsymbol{H}}_i) \right\|_1 \\ \overset{(b)}{\geq} &\underline{\delta} \left\| \hat{\boldsymbol{H}}_T + \hat{\boldsymbol{H}}_1 - 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right\|_F - 2\bar{\delta} \sqrt{\frac{1}{K}} \left\| \hat{\boldsymbol{H}}_T \right\|_F \\ \overset{(c)}{\geq} &\underline{\delta} \left\| \hat{\boldsymbol{H}}_T - 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right\|_F + \left\| \hat{\boldsymbol{H}}_1 \right\|_F - 2\bar{\delta} \sqrt{\frac{1}{K}} \left\| \hat{\boldsymbol{H}}_T \right\|_F \\ \geq &\underline{\delta} \left\| \hat{\boldsymbol{H}}_T - 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right\|_F - 2\bar{\delta} \sqrt{\frac{1}{K}} \left\| \hat{\boldsymbol{H}}_T \right\|_F, \end{aligned}$$

where (a) follows from the triangle inequality, (b) follows from Lemma 1 and (23), and (c) follows from the fact that $\hat{\boldsymbol{H}}_T - \beta_b \hat{\boldsymbol{h}}^\top \in T$ and $\hat{\boldsymbol{H}}_1 \in T^\perp$. Summing the above inequality for $b = 1, 2$, we obtain

$$\begin{aligned} &\sum_b \left\| \mathcal{B}_b \left(\hat{\boldsymbol{H}} - 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right) \right\|_1 \\ \geq &\underline{\delta} \sum_b \left\| \hat{\boldsymbol{H}}_T - 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right\|_F - 4\bar{\delta} \sqrt{\frac{1}{K}} \left\| \hat{\boldsymbol{H}}_T \right\|_F. \end{aligned} \tag{24}$$

The first term in the RHS of (24) can be bounded using the following lemma, whose proof is deferred to Appendix F-A.

Lemma 2. *We have*

$$\begin{aligned} &\sum_b \left\| \hat{\boldsymbol{H}}_T - 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right\|_F \geq \sqrt{\alpha} \left\| \hat{\boldsymbol{H}}_T \right\|_F, \\ &\sum_b \left\| \hat{\boldsymbol{H}}_T - 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right\|_F \geq \sqrt{\alpha} \left(\|\boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^*\|_2 \right) \|\hat{\boldsymbol{h}}\|_2. \end{aligned}$$

Combining (24) and the lemma, we obtain

$$\sum_b \left\|\mathcal{B}_b\left(\hat{\boldsymbol{H}}-2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top\right)\right\|_1 \geq \left(\underline{\delta}\sqrt{\alpha}-4\bar{\delta}\sqrt{\frac{1}{K}}\right)\left\|\hat{\boldsymbol{H}}_T\right\|_F$$

and

$$\begin{aligned}&\sum_b \left\|\mathcal{B}_b\left(\hat{\boldsymbol{H}}-2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top\right)\right\|_1\\ &\geq \left(\underline{\delta}-4\bar{\delta}\sqrt{\frac{1}{\alpha K}}\right)\sum_b\left\|\hat{\boldsymbol{H}}_T-\boldsymbol{\beta}_b\hat{\boldsymbol{h}}^\top\right\|_F\\ &\geq \left(\underline{\delta}-4\bar{\delta}\sqrt{\frac{1}{\alpha K}}\right)\sqrt{\alpha}\left(\|\boldsymbol{\beta}_1^*\|_2+\|\boldsymbol{\beta}_2^*\|_2\right)\|\hat{\boldsymbol{h}}\|_2.\end{aligned}$$

Recall that $K=c\frac{1}{\alpha}$. When c is sufficiently large, the above inequalities imply that for some numeric constant c',

$$\sum_b \left\|\mathcal{B}_b\left(\hat{\boldsymbol{H}}-2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top\right)\right\|_1 \geq \frac{\sqrt{\alpha}}{c''}\left\|\hat{\boldsymbol{H}}_T\right\|_F \overset{(d)}{\geq} \frac{\sqrt{\alpha}}{c'}\left\|\hat{\boldsymbol{H}}\right\|_F, \tag{25}$$

$$\sum_b \left\|\mathcal{B}_b\left(\hat{\boldsymbol{H}}-2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top\right)\right\|_1 \geq \frac{\sqrt{\alpha}}{c'}\left(\|\boldsymbol{\beta}_1^*\|_2+\|\boldsymbol{\beta}_2^*\|_2\right)\|\hat{\boldsymbol{h}}\|_2, \tag{26}$$

where the inequality (d) follows from (22) and $\mathrm{rank}(\hat{\boldsymbol{H}}_T)\leq 4$. This concludes the proof of Step (2).

3) Step (3): Producing Error Bounds: We now combine the results from the above steps, in order to obtain bounds on $\|\hat{\boldsymbol{h}}\|_2$ and $\|\hat{\boldsymbol{H}}\|_F$ in terms of η, and the other parameters of the problem, hence concluding the proof of Theorem 1.

From Step (1), we concluded the bound (21), which we reproduce:

$$\sum_b n\left\|\mathcal{B}_b\left(-\hat{\boldsymbol{H}}+2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top\right)\right\|_1 - c_2\sum_b\sqrt{n}\,\|\boldsymbol{e}_b\|_2\,\|\hat{\boldsymbol{h}}\|_2 \leq 2\eta.$$

Applying (26) to the LHS above, we get

$$\sqrt{n}\sum_b\left(\sqrt{\alpha}\sqrt{n}\,\|\boldsymbol{\beta}_b^*\|_2-\|\boldsymbol{e}_b\|_2\right)\|\hat{\boldsymbol{h}}\|_2 \lesssim 2\eta.$$

Under the assumption $\|\boldsymbol{e}\|_2\leq\frac{1}{c_5}\sqrt{\alpha}\sqrt{n}\left(\|\boldsymbol{\beta}_1^*\|_2+\|\boldsymbol{\beta}_2^*\|_2\right)$ for some c_5 sufficiently large, we obtain the following bound for $\|\hat{\boldsymbol{h}}\|_2$:

$$\|\hat{\boldsymbol{h}}\|_2 \lesssim \frac{1}{\sqrt{\alpha}n\left(\|\boldsymbol{\beta}_1^*\|_2+\|\boldsymbol{\beta}_2^*\|_2\right)}\eta.$$

To obtain a bound on $\left\|\hat{\boldsymbol{H}}\right\|_F$, we note that

$$\begin{aligned}\sum_b\|\boldsymbol{e}_b\|_2\,\|\hat{\boldsymbol{h}}\|_2 &\leq \frac{1}{c_5}\sqrt{n}\sum_b\sqrt{\alpha}\,\|\boldsymbol{\beta}_b^*\|_2\,\|\hat{\boldsymbol{h}}\|_2\\ &\leq \frac{c'}{c_5}\sqrt{n}\sum_b\left\|\mathcal{B}_b\left(\hat{\boldsymbol{H}}-2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top\right)\right\|_1,\end{aligned}$$

where we use the assumption on $\|\boldsymbol{e}\|$ and (26) in the two inequalities, respectively. When c_5 is large, we combine the last displayed equation with (21) to obtain

$$n\sqrt{\alpha}\left\|\hat{\boldsymbol{H}}\right\|_F \lesssim n\sum_b\left\|\mathcal{B}_b\left(\hat{\boldsymbol{H}}_T-2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top\right)\right\|_1 \lesssim 2\eta,$$

where we use (25) in the last inequality. This implies

$$\left\|\hat{\boldsymbol{H}}\right\|_F \lesssim \frac{1}{n\sqrt{\alpha}}\eta,$$

completing the proof of Step (3) and thus Theorem 1.

C. Proof of Theorem 3

We now turn to Theorem 3 for the stochastic noise setting. The main conceptual flow of the proof is quite similar to the deterministic noise case, though some significant additional steps are required. For the deterministic case, the starting point is the constraint, which allows us to bound $\mathcal{A}_b$ and $\mathcal{B}_b$ in terms of η using feasibility of $(\boldsymbol{K}^*,\boldsymbol{g}^*)$ and $(\boldsymbol{K}^*+\hat{\boldsymbol{H}},\boldsymbol{g}^*+\hat{\boldsymbol{h}})$. In the stochastic setup we have a Lagrangian (regularized) formulation, and hence we obtain the analogous result from optimality. Thus, the first step here involves showing that as a consequence of optimality, the solution $(\hat{\boldsymbol{K}},\hat{\boldsymbol{g}})=(\boldsymbol{K}^*+\hat{\boldsymbol{H}},\boldsymbol{g}^*+\hat{\boldsymbol{h}})$ satisfies inequality (29) below, which implies that $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{h}}$ must live in a certain cone. The RIP-like condition for $\mathcal{B}_b$ in the stochastic case is more demanding. We prove a second RIP-like condition (Lemma 5). We then bound $\mathcal{A}$ by terms involving $\mathcal{B}$, and then invoke the RIP condition and the cone constraint.

Now we turn to each step of the proof. We continue to use the notation given in Section IV-A. We let $\boldsymbol{D}_b := (\lfloor n_b/2\rfloor)^{-1}\left[\boldsymbol{d}_{b,1},\ldots,\boldsymbol{d}_{b,\lfloor n_b/2\rfloor}\right]^\top \in \mathbb{R}^{\lfloor n_b/2\rfloor\times p}$. Recall that we defined

$$\gamma := \|\boldsymbol{\beta}_1^*\|_2+\|\boldsymbol{\beta}_2^*\|_2.$$

Since the $\{\boldsymbol{x}_i\}$ are assumed to be Gaussian with i.i.d. entries, the statement of the theorem is invariant under rotation of the $\boldsymbol{\beta}_b^*$'s. Therefore, it suffices to prove the theorem assuming $\boldsymbol{\beta}_1^*-\boldsymbol{\beta}_2^*$ is supported on the first coordinate. The follow lemma shows that we can further assume $\{\boldsymbol{x}_i\}$ and $\boldsymbol{e}$ have bounded entries, since we are interested in results that hold with high probability. This simplifies the subsequent analysis.

Lemma 3. *There exists an absolute constant $c>0$ such that, if the conclusion of Theorem 3 holds w.h.p. with the additional assumption that*

$$\begin{aligned}\boldsymbol{x}_i(l) &\leq c\sqrt{\log n}, \forall i\in[n], l\in[p],\\ e_i &\leq c\sigma\sqrt{\log n}, \forall i\in[n],\end{aligned}$$

then it also holds w.h.p. without this assumption.

We prove this lemma in Appendix F-B. In the sequel, we therefore assume $\mathrm{support}(\boldsymbol{\beta}_1^*-\boldsymbol{\beta}_2^*)=\{1\}$, and the $\{\boldsymbol{x}_i\}$ and $\{\boldsymbol{e}_i\}$ satisfy the bounds in the above lemma.

1) Step (1): Consequence of Optimality: This step uses optimality of the solution $(\hat{\boldsymbol{K}},\hat{\boldsymbol{g}})=(\boldsymbol{K}^*+\hat{\boldsymbol{H}},\boldsymbol{g}^*+\hat{\boldsymbol{h}})$, to get a bound on $\mathcal{A}$. By optimality, we have

$$\begin{aligned}&\sum_b\sum_{i\in\mathcal{I}_b}\left(-\left\langle\boldsymbol{x}_i\boldsymbol{x}_i^\top,\hat{\boldsymbol{K}}\right\rangle+2y_i\left\langle\boldsymbol{x}_i,\hat{\boldsymbol{g}}\right\rangle-y_i^2+\sigma^2\right)^2+\lambda\left\|\hat{\boldsymbol{K}}\right\|_*\\ &\leq\sum_b\sum_{i\in\mathcal{I}_b}\left(-\left\langle\boldsymbol{x}_i\boldsymbol{x}_i^\top,\boldsymbol{K}^*\right\rangle+2y_i\left\langle\boldsymbol{x}_i,\boldsymbol{g}^*\right\rangle-y_i^2+\sigma^2\right)^2+\lambda\left\|\boldsymbol{K}^*\right\|_*.\end{aligned}$$

Using the expression (18), we have

$$\sum_b \sum_{i\in\mathcal{I}_b} \Big(-\left\langle \boldsymbol{x}_i\boldsymbol{x}_i^\top, \hat{\boldsymbol{H}}\right\rangle + 2(\boldsymbol{x}_i^\top\boldsymbol{\beta}_b^* + e_i)\left\langle \boldsymbol{x}_i, \hat{\boldsymbol{h}}\right\rangle - e_i\boldsymbol{x}_i^\top\boldsymbol{\delta}_b^* - (e_i^2-\sigma^2)\Big)^2 + \lambda\left\|\hat{\boldsymbol{K}}\right\|_* \le \sum_b\sum_{i\in\mathcal{I}_b}\left(-e_i\boldsymbol{x}_i^\top\boldsymbol{\delta}_b^* - (e_i^2-\sigma^2)\right)^2 + \lambda\left\|\boldsymbol{K}^*\right\|_*.$$

Defining the noise vectors $\boldsymbol{w}_{1,b} := -\boldsymbol{e}_b \circ (\boldsymbol{X}_b\boldsymbol{\delta}_b^*)$, $\boldsymbol{w}_{2,b} := -\left(\boldsymbol{e}_b^2 - \sigma^2\mathbf{1}\right)$ and $\boldsymbol{w}_b = \boldsymbol{w}_{1,b} - \boldsymbol{w}_{2,b}$, we can rewrite the inequality above as

$$\sum_b \left\|n_b\mathcal{A}_b\left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top\right) + 2\boldsymbol{e}_b\circ(\boldsymbol{X}_b\hat{\boldsymbol{h}}) + \boldsymbol{w}_b\right\|_2^2 + \lambda\left\|\hat{\boldsymbol{K}}\right\|_* \le \sum_{b=1,2}\|\boldsymbol{w}_b\|_2^2 + \lambda\left\|\boldsymbol{K}^*\right\|_*.$$

Expanding the squares and rearranging terms, we obtain

$$\begin{aligned}
&\sum_b \left\|n_b\mathcal{A}_b\left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top\right) + 2\boldsymbol{e}_b\circ(\boldsymbol{X}_b\hat{\boldsymbol{h}})\right\|_2^2 \\
&\le \sum_b 2\left\langle \hat{\boldsymbol{H}} - 2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top, n_b\mathcal{A}_b^*\boldsymbol{w}_b\right\rangle - \sum_b\left\langle \hat{\boldsymbol{h}}, 4\boldsymbol{X}_b^\top\mathrm{diag}(\boldsymbol{e}_b)\boldsymbol{w}_b\right\rangle \\
&\quad + \lambda\left(\left\|\boldsymbol{K}^*\right\|_* - \left\|\hat{\boldsymbol{K}}\right\|_*\right) \qquad (27)\\
&\overset{(a)}{\le} \left(\left\|\hat{\boldsymbol{H}}_T\right\|_* + \left\|\hat{\boldsymbol{H}}_{T^\perp}\right\|_*\right)\cdot P + \|\hat{\boldsymbol{h}}\|_2\cdot Q \\
&\quad + \lambda\left(\left\|\boldsymbol{K}^*\right\|_* - \left\|\hat{\boldsymbol{K}}\right\|_*\right) \\
&\overset{(b)}{\le} \left(\left\|\hat{\boldsymbol{H}}_T\right\|_* + \left\|\hat{\boldsymbol{H}}_{T^\perp}\right\|_*\right)\cdot P + \|\hat{\boldsymbol{h}}\|_2\cdot Q \\
&\quad + \lambda\left(\left\|\hat{\boldsymbol{H}}_T\right\|_* - \left\|\hat{\boldsymbol{H}}_{T^\perp}\right\|_*\right),
\end{aligned}$$

where $\mathcal{A}_b^*$ is the adjoint operator of $\mathcal{A}_b$, in (a) we have defined

$$P := 2\sum_b \|n_b\mathcal{A}_b^*\boldsymbol{w}_b\|, \qquad (28)$$

$$Q := 4\sum_b \|\boldsymbol{\beta}_b^*\|_2\|n_b\mathcal{A}_b^*\boldsymbol{w}_b\| + \sqrt{p}\left\|\sum_b 4\boldsymbol{X}_b^\top\mathrm{diag}(\boldsymbol{e}_b)\boldsymbol{w}_b\right\|_\infty,$$

and (b) follows from (17). We need the following lemma, which bounds the noise terms P and Q. Its proof is a substantial part of the proof to the main result, but quite lengthy. We therefore defer it to Appendix A.

Lemma 4. *Under the assumptions of the theorem, we have $\lambda \ge 2P$ and $\lambda \ge \frac{1}{\sigma+\gamma}Q$ with high probability.*

Applying the lemma, we get

$$\begin{aligned}
&\sum_b \left\|n_b\mathcal{A}_b\left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top\right) + 2\boldsymbol{e}_b\circ(\boldsymbol{X}_b\hat{\boldsymbol{h}})\right\|_2^2 \\
&\le \lambda\left(\frac{3}{2}\left\|\hat{\boldsymbol{H}}_T\right\|_* - \frac{1}{2}\left\|\hat{\boldsymbol{H}}_{T^\perp}\right\|_*\right) + \lambda(\gamma+\sigma)\|\hat{\boldsymbol{h}}\|_2. \qquad (29)
\end{aligned}$$

Since the left hand side of (29) is non-negative, we obtain the following cone constraint for the optimal solution:

$$\left\|\hat{\boldsymbol{H}}_{T^\perp}\right\|_* \le \frac{5}{2}\left\|\hat{\boldsymbol{H}}_T\right\|_* + (\gamma+\sigma)\|\hat{\boldsymbol{h}}\|_2. \qquad (30)$$

This concludes the proof of Step (1) of the proof.

2) Step (2): RIP and Lower Bounds: We can get a lower bound to the expression in the LHS of (29) using $\mathcal{B}$, as follows. Similarly as before, let K be some numeric constant to be chosen later. We partition $\hat{\boldsymbol{H}}_{T^\perp}$ into a sum of $M := \frac{p-4}{K}$ matrices $\hat{\boldsymbol{H}}_1, \ldots, \hat{\boldsymbol{H}}_M$ according to the SVD of $\hat{\boldsymbol{H}}_{T^\perp}$, such that $\mathrm{rank}(\hat{\boldsymbol{H}}_i) \le K$ and the smallest singular value of $\hat{\boldsymbol{H}}_i$ is larger than the largest singular value of $\hat{\boldsymbol{H}}_{i+1}$. Then we have the following chain of inequalities:

$$\begin{aligned}
&\sum_b \left\|n_b\mathcal{A}_b\left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top\right) + 2\boldsymbol{e}_b\circ\left(\boldsymbol{X}_b\hat{\boldsymbol{h}}\right)\right\|_2^2 \\
&\overset{(a)}{\ge} \sum_b \left\|n_b\mathcal{B}_b\left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top\right) + 2n_b\boldsymbol{D}_b\hat{\boldsymbol{h}}\right\|_2^2 \\
&\overset{(b)}{\ge} \sum_b n_b\left\|\mathcal{B}_b\left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top\right) + 2\boldsymbol{D}_b\hat{\boldsymbol{h}}\right\|_1^2 \\
&\overset{(c)}{\gtrsim} n\left(\sum_b \left\|\mathcal{B}_b\left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top\right) + 2\boldsymbol{D}_b\hat{\boldsymbol{h}}\right\|_1\right)^2 \\
&\overset{(d)}{\ge} n\Bigg(\sum_b \left\|\mathcal{B}_b\left(-\hat{\boldsymbol{H}}_T + 2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top + \hat{\boldsymbol{H}}_1\right) + 2\boldsymbol{D}_b\hat{\boldsymbol{h}}\right\|_1 \\
&\qquad - \sum_b\sum_{i=2}^M \left\|\mathcal{B}_b(\hat{\boldsymbol{H}}_i)\right\|_1\Bigg)^2. \qquad (31)
\end{aligned}$$

Here (a) follows from the definitions of $\mathcal{A}_b$ and $\mathcal{B}_b$ and the triangle inequality, (b) follows from $\|\boldsymbol{u}\|_2^2 \ge \frac{1}{n_b}\|\boldsymbol{u}\|_1^2$ for all $\boldsymbol{u}\in\mathbb{R}^{n_b}$, (c) follows from $n_1 \approx n_2$, and (d) follows from the triangle inequality.

We see that in order to obtain lower bounds on (31) in terms of $\|\hat{\boldsymbol{h}}\|_2$ and $\|\hat{\boldsymbol{H}}\|_F$, we need an extension of the previous RIP-like result from Lemma 1, in order to deal with the first term in (31). The following lemma is proved in Appendix B.

Lemma 5. *The following holds for some numerical constants $c, \underline{\delta}, \bar{\delta}$. For $b = 1, 2$, if $\mu > 1$ and $n_b \ge cpr$, then with probability $1 - \exp(-n_b)$, we have the following RIP-2:*

$$\underline{\delta}\left(\|\boldsymbol{Z}\|_F + \sigma\|\boldsymbol{z}\|_2\right) \le \|\mathcal{B}_b\boldsymbol{Z} - \boldsymbol{D}_b\boldsymbol{z}\|_1 \le \bar{\delta}\left(\|\boldsymbol{Z}\|_F + \sigma\|\boldsymbol{z}\|_2\right), \quad \forall\boldsymbol{z}\in\mathbb{R}^p, \forall\boldsymbol{Z}\in\mathbb{R}^{p\times p} \text{ with } \mathrm{rank}(\boldsymbol{Z}) \le r.$$

Using this we can now bound the last inequality in (31) above. First, note that for each $b = 1, 2$,

$$\begin{aligned}
\sum_{i=2}^M \left\|\mathcal{B}_b(\hat{\boldsymbol{H}}_i)\right\|_1 &\overset{(a)}{\le} \bar{\delta}\sum_{i=2}^M \left\|\hat{\boldsymbol{H}}_i\right\|_F \le \bar{\delta}\sum_{i=2}^M \frac{1}{\sqrt{K}}\left\|\hat{\boldsymbol{H}}_{i-1}\right\|_* \\
&\le \frac{\bar{\delta}}{\sqrt{K}}\left\|\hat{\boldsymbol{H}}_{T^\perp}\right\|_*, \qquad (32)
\end{aligned}$$

where (a) follows from the upper bound in Lemma 5 with σ set to 0. Then, applying the lower-bound in Lemma 5 to the first term in the parentheses in (31), and (32) to the second

term, we obtain

$$\begin{aligned}
&\sum_b \left\| n_b \mathcal{A}_b \left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right) + 2\boldsymbol{e}_b \circ \left(\boldsymbol{X}_b \hat{\boldsymbol{h}} \right) \right\|_2^2 \\
\geq & n \left(\sum_b \underline{\delta} \left\| \hat{\boldsymbol{H}}_T - 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right\|_F + 2\underline{\delta}\sigma \|\hat{\boldsymbol{h}}\|_2 - \bar{\delta} \sqrt{\frac{1}{K}} \left\| \hat{\boldsymbol{H}}_{T^\perp} \right\|_* \right)^2 \\
\gtrsim & n \left(\sum_b \underline{\delta}^2 \left\| \hat{\boldsymbol{H}}_T - 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right\|_F^2 + \underline{\delta}^2 \sigma^2 \|\hat{\boldsymbol{h}}\|_2^2 - \bar{\delta}^2 \frac{1}{K} \left\| \hat{\boldsymbol{H}}_{T^\perp} \right\|_*^2 \right).
\end{aligned}$$

Choosing K to be sufficiently large, and applying Lemma 2, we obtain

$$\begin{aligned}
&\sum_b \left\| n_b \mathcal{A}_b \left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right) + 2\boldsymbol{e}_b \circ \left(\boldsymbol{X}_b \hat{\boldsymbol{h}} \right) \right\|_2^2 \\
\gtrsim & n \left(\left\| \hat{\boldsymbol{H}}_T \right\|_F^2 + \gamma^2 \|\hat{\boldsymbol{h}}\|_2^2 + \sigma^2 \|\hat{\boldsymbol{h}}\|_2^2 - \frac{1}{100} \left\| \hat{\boldsymbol{H}}_{T^\perp} \right\|_*^2 \right). \quad (33)
\end{aligned}$$

Using (30), we further get

$$\begin{aligned}
&\sum_b \left\| n_b \mathcal{A}_b \left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right) + 2\boldsymbol{e}_b \circ \left(\boldsymbol{X}_b \hat{\boldsymbol{h}} \right) \right\|_2^2 \\
\gtrsim & n \Bigg[\left\| \hat{\boldsymbol{H}}_T \right\|_F^2 + \gamma^2 \|\hat{\boldsymbol{h}}\|_2^2 + \sigma^2 \|\hat{\boldsymbol{h}}\|_2^2 - \frac{1}{8} \left\| \hat{\boldsymbol{H}}_T \right\|_*^2 \\
& - \frac{1}{25} \left(\gamma^2 + \sigma^2 \right) \|\hat{\boldsymbol{h}}\|_2^2 \Bigg] \\
\gtrsim & \frac{1}{8} n \left(\left\| \hat{\boldsymbol{H}}_T \right\|_F + (\gamma + \sigma) \|\hat{\boldsymbol{h}}\|_2 \right)^2. \quad (34)
\end{aligned}$$

This completes Step (2), and we are ready to combine the results to obtain error bounds, as promised in Step (3) and by the theorem.

3) Step (3): Producing Error bounds: Combining (29) and (34), we get

$$n \left(\left\| \hat{\boldsymbol{H}}_T \right\|_F + (\gamma + \sigma) \|\hat{\boldsymbol{h}}\|_2 \right)^2 \lesssim \lambda \|\boldsymbol{H}_T\|_F + \lambda(\gamma + \sigma) \|\hat{\boldsymbol{h}}\|_2,$$

which implies $\left\| \hat{\boldsymbol{H}}_T \right\|_F + (\gamma + \sigma) \|\hat{\boldsymbol{h}}\|_2 \lesssim \frac{\lambda}{n}$. It follows that $\|\hat{\boldsymbol{h}}\|_2 \lesssim \frac{1}{n(\gamma+\sigma)} \lambda$ and

$$\begin{aligned}
\left\| \hat{\boldsymbol{H}} \right\|_F &\leq \left\| \hat{\boldsymbol{H}}_T \right\|_* + \left\| \hat{\boldsymbol{H}}_{T^\perp} \right\|_* \\
&\overset{(a)}{\leq} \frac{7}{2} \left\| \hat{\boldsymbol{H}}_T \right\|_* + (\gamma + \sigma) \|\hat{\boldsymbol{h}}\|_2 \\
&\overset{(b)}{\leq} \frac{7}{2} \cdot \sqrt{4} \left\| \hat{\boldsymbol{H}}_T \right\|_F + (\gamma + \sigma) \|\hat{\boldsymbol{h}}\|_2 \\
&\lesssim \frac{1}{n} \lambda,
\end{aligned}$$

where we use (30) in (a) and $\operatorname{rank}\left(\hat{\boldsymbol{H}}_T \right) \leq 4$ in (b). This completes Step (3) and the proof of the theorem.

D. Proof of Theorem 5

In this section we prove the error bounds in Theorem 5 for our nonconvex approach.

1) Statistical Error: The proof of part (a) of Theorem 5 follows along the same lines as the proof of Theorem 3. In addition, we need to derive RIP and error bounds of the negative term. For the optimal solution $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}}) = (\boldsymbol{K}^* + \hat{\boldsymbol{H}}, \boldsymbol{g}^* + \hat{\boldsymbol{h}})$, from the constraint we have that

$$\|\boldsymbol{K}^* + \hat{\boldsymbol{H}}\|_* \leq \|\boldsymbol{K}^*\|_*.$$

Using the decomposition $\hat{\boldsymbol{H}} = \hat{\boldsymbol{H}}_T + \hat{\boldsymbol{H}}_{T^\perp}$, we have $\|\boldsymbol{K}^* + \hat{\boldsymbol{H}}\|_* = \|\boldsymbol{K}^* + \hat{\boldsymbol{H}}_T\|_* + \|\hat{\boldsymbol{H}}_{T^\perp}\|_* \geq \|\boldsymbol{K}^*\|_* - \|\hat{\boldsymbol{H}}_T\|_* + \|\hat{\boldsymbol{H}}_{T^\perp}\|_*$. We thus have

$$\|\hat{\boldsymbol{H}}_{T^\perp}\|_* \leq \|\hat{\boldsymbol{H}}_T\|_*. \quad (35)$$

Following similar calculations as those in Section IV-C1, one can obtain the consequence of optimality as

$$\begin{aligned}
&\sum_b \left\| n_b \mathcal{A}_b \left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right) + 2\boldsymbol{e}_b \circ (\boldsymbol{X}_b \hat{\boldsymbol{h}}) \right\|_2^2 \\
&\underbrace{- \sum_b 4\sigma^2 \mathcal{A}_b(\hat{\boldsymbol{h}}\hat{\boldsymbol{h}}^\top)}_{S_5} \\
\leq & \sum_b 2 \left\langle \hat{\boldsymbol{H}} - 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top, n_b \mathcal{A}_b^* \boldsymbol{w}_b \right\rangle - \sum_b \left\langle \hat{\boldsymbol{h}}, 4\boldsymbol{X}_b^\top \operatorname{diag}(\boldsymbol{e}_b) \boldsymbol{w}_b \right\rangle \\
&\underbrace{- \sum_b 4\sigma^2 \mathcal{A}_b(\boldsymbol{\delta}_b^* \hat{\boldsymbol{h}}^\top)}_{S_6} \underbrace{- \sum_b 8\sigma^2 \langle \boldsymbol{e}_b, \boldsymbol{X}_b \hat{\boldsymbol{h}} \rangle}_{S_7}. \quad (36)
\end{aligned}$$

Compared to (27), the above equality does not have the nuclear norm term since we remove the regularization term. The additional terms S_5, S_6 and S_7 come from the negative term $-\sum_{i=1}^n 4\sigma^2 (y_i - \langle \boldsymbol{x}_i, \boldsymbol{g} \rangle)^2$.

By standard concentration result, when $n \gtrsim p/\epsilon^2$, with probability at least $1 - \exp(-p)$, $\|\frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top - \boldsymbol{I}\| \leq \epsilon$. Choosing $\epsilon = 0.1$, we have w.h.p.

$$S_5 = 4\sigma^2 \sum_{i=1}^n \langle \boldsymbol{x}_i \boldsymbol{x}_i^\top - \boldsymbol{I}, \hat{\boldsymbol{h}}\hat{\boldsymbol{h}}^\top \rangle + 4n\sigma^2 \|\hat{\boldsymbol{h}}\|_2^2 \leq 4.4 n\sigma^2 \|\hat{\boldsymbol{h}}\|_2^2.$$

Using the above result and (34), we obtain

$$\begin{aligned}
&\sum_b \left\| n_b \mathcal{A}_b \left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right) + 2\boldsymbol{e}_b \circ (\boldsymbol{X}_b \hat{\boldsymbol{h}}) \right\|_2^2 \\
&- \sum_b 4\sigma^2 \mathcal{A}_b(\hat{\boldsymbol{h}}\hat{\boldsymbol{h}}^\top) \\
\gtrsim & n \left(\left\| \hat{\boldsymbol{H}}_T \right\|_F + (\gamma + \sigma) \|\hat{\boldsymbol{h}}\|_2 \right)^2 - n\sigma^2 \|\hat{\boldsymbol{h}}\|_2^2 \\
\gtrsim & n \left(\left\| \hat{\boldsymbol{H}}_T \right\|_F + \gamma \|\hat{\boldsymbol{h}}\|_2 \right)^2, \quad (37)
\end{aligned}$$

where the second inequality follows from the assumption $\gamma/\sigma \geq c$ for sufficiently large constant c.

Now we turn to the right hand side of (36). By the Cauchy-Schwarz inequality, it is upper bounded by

$$\left(\|\hat{\boldsymbol{H}}_T\|_* + \|\hat{\boldsymbol{H}}_{T^\perp}\|_* \right) \cdot P + \|\hat{\boldsymbol{h}}\|_2 \cdot Q',$$

where P is given in (28), and we defined

$$Q' := 4\gamma P + W$$

and

$$W := \sqrt{p}\Big\| \sum_b \boldsymbol{X}_b^\top \mathrm{diag}(\boldsymbol{e}_b)\boldsymbol{w}_b + \sum_b \sigma^2 \boldsymbol{X}_b^\top \boldsymbol{X}_b \boldsymbol{\delta}_b^* + \sum_b 2\sigma^2 \boldsymbol{X}_b^\top \boldsymbol{e}_b \Big\|_\infty. \tag{38}$$

Using Lemma 4, we obtain $P \lesssim \sigma\gamma\sqrt{np}\log^3 n$. The following lemma gives an upper bound of the term W; see Appendix F-C for the proof.

Lemma 6. *Under the assumptions of the Theorem 5, we have that w.h.p.,*

$$W \lesssim \sigma^2\gamma\sqrt{np}\log^2 n.$$

We therefore conclude that the right hand side of (36) is at most on the order of

$$\sigma\gamma\sqrt{np}\log^3 n\left(\|\hat{\boldsymbol{H}}_T\|_* + \|\hat{\boldsymbol{H}}_{T^\perp}\|_*\right) + \sigma\gamma^2\sqrt{np}\log^3 n\|\hat{\boldsymbol{h}}\|_2. \tag{39}$$

Putting (37) and (39) together, we finish our proof by showing

$$\begin{aligned} & n\left(\left\|\hat{\boldsymbol{H}}_T\right\|_F + \gamma\|\hat{\boldsymbol{h}}\|_2\right)^2 \\ \lesssim & \sigma\gamma\sqrt{np}\log^3 n\left(\|\hat{\boldsymbol{H}}_T\|_* + \|\hat{\boldsymbol{H}}_{T^\perp}\|_*\right) + \sigma\gamma^2\sqrt{np}\log^3 n\|\hat{\boldsymbol{h}}\|_2 \\ \lesssim & \sigma\gamma\sqrt{np}\log^3 n\|\hat{\boldsymbol{H}}_T\|_F + \sigma\gamma^2\sqrt{np}\log^3 n\|\hat{\boldsymbol{h}}\|_2, \end{aligned}$$

where the second inequality follows from the conic constraint (35) and $\|\hat{\boldsymbol{H}}_T\|_* \le 2\|\hat{\boldsymbol{H}}_T\|_F$.

2) Optimization Error: Recall that $\boldsymbol{H}^{(t)} := \boldsymbol{K}^{(t)} - \hat{\boldsymbol{K}}$ and $\boldsymbol{h}^{(t)} := \boldsymbol{g}^{(t)} - \hat{\boldsymbol{g}}$, which represent the optimization error, and $\hat{\boldsymbol{H}} := \hat{\boldsymbol{K}} - \boldsymbol{K}^*$ and $\hat{\boldsymbol{h}} := \hat{\boldsymbol{g}} - \boldsymbol{g}^*$, which represent the statistical error. We need several auxiliary lemmas for the proof. First, we establish a result that shows that $\boldsymbol{H}^{(t)}$ belongs to a cone-like set. See Appendix F-D for the proof.

Lemma 7. *From program (8), due to the nuclear norm constraint, we have that for any $t = 0, 1, \ldots$,*

$$\left\|\boldsymbol{H}_{T^\perp}^{(t)}\right\|_* \le 3\left\|\boldsymbol{H}_T^{(t)}\right\|_* + 4\left\|\hat{\boldsymbol{H}}_T\right\|_*.$$

Next, we characterize the curvature of $\mathcal{L}_n$. To ease notation, for any $\boldsymbol{K}, \boldsymbol{K}' \in \mathbb{R}^{p\times p}$ and $\boldsymbol{g}, \boldsymbol{g}' \in \mathbb{R}^p$, we define

$$\begin{aligned} & Q_n(\boldsymbol{K}', \boldsymbol{g}'; \boldsymbol{K}, \boldsymbol{g}) \\ & := \mathcal{L}_n(\boldsymbol{K}', \boldsymbol{g}') - \mathcal{L}_n(\boldsymbol{K}, \boldsymbol{g}) - \langle \nabla_{\boldsymbol{K}}\mathcal{L}_n(\boldsymbol{K}, \boldsymbol{g}), \boldsymbol{K}' - \boldsymbol{K}\rangle \\ & \quad - \langle \nabla_{\boldsymbol{g}}\mathcal{L}_n(\boldsymbol{K}, \boldsymbol{g}), \boldsymbol{g}' - \boldsymbol{g}\rangle \\ & = \sum_b \Big\| n_b\mathcal{A}_b\left(-(\boldsymbol{K}' - \boldsymbol{K}) + 2\boldsymbol{\beta}_b^*(\boldsymbol{g}' - \boldsymbol{g})^\top\right) \\ & \quad + 2\boldsymbol{e}_b \circ (\boldsymbol{X}_b(\boldsymbol{g}' - \boldsymbol{g}))\Big\|_2^2 - \sum_{i=1}^n 4\sigma^2\langle \boldsymbol{x}_i, \boldsymbol{g}' - \boldsymbol{g}\rangle^2. \end{aligned} \tag{40}$$

The next lemma shows the smoothness of $\mathcal{L}_n$. Its proof is given in Appendix C.

Lemma 8. *Given $Q_n(\boldsymbol{K}', \boldsymbol{g}'; \boldsymbol{K}, \boldsymbol{g})$ that is defined in (40), we have that there exist constants $\{c_i\}_{i=1}^2$ such that with probability at least $1 - c_2/n$, the inequality*

$$\begin{aligned} & Q_n(\boldsymbol{K}', \boldsymbol{g}'; \boldsymbol{K}, \boldsymbol{g}) \\ & \le c_1 n\sqrt{p}^3\log n\|\boldsymbol{K}' - \boldsymbol{K}\|_F^2 + c_1 n\log n\gamma^2\|\boldsymbol{g}' - \boldsymbol{g}\|_2^2 \end{aligned} \tag{41}$$

holds for any $\boldsymbol{K}, \boldsymbol{K}' \in \mathbb{R}^{p\times p}$ and $\boldsymbol{g}, \boldsymbol{g}' \in \mathbb{R}^p$.

Next we turn to the lower bound of Q. We first let $\boldsymbol{K}' = \boldsymbol{K}^{(t)}$, $\boldsymbol{g}' = \boldsymbol{g}^{(t)}$, $\boldsymbol{K} = \hat{\boldsymbol{K}}$, $\boldsymbol{g} = \hat{\boldsymbol{g}}$. Then we have

$$\begin{aligned} & Q_n(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}}; \boldsymbol{K}^{(t)}, \boldsymbol{g}^{(t)}) \\ & = \sum_b \left\| n_b\mathcal{A}_b\left(-\boldsymbol{H}^{(t)} + 2\boldsymbol{\beta}_b^*\boldsymbol{h}^{(t)\top}\right) + 2\boldsymbol{e}_b \circ \left(\boldsymbol{X}_b\boldsymbol{h}^{(t)}\right)\right\|_2^2 \\ & \quad - \sum_{i=1}^n 4\sigma^2\langle \boldsymbol{x}_i, \boldsymbol{h}^{(t)}\rangle^2 \\ & \gtrsim n\left(\left\|\boldsymbol{H}_T^{(t)}\right\|_F^2 + \gamma^2\|\boldsymbol{h}^{(t)}\|_2^2 + \sigma^2\|\boldsymbol{h}^{(t)}\|_2^2 - \frac{1}{100}\left\|\boldsymbol{H}_{T^\perp}^{(t)}\right\|_*^2\right) \\ & \quad - \sum_b 4\sigma^2\mathcal{A}_b(\hat{\boldsymbol{h}}\hat{\boldsymbol{h}}^\top) \\ & \gtrsim n\left(\left\|\boldsymbol{H}_T^{(t)}\right\|_F^2 + \gamma^2\|\boldsymbol{h}^{(t)}\|_2^2 + \sigma^2\|\boldsymbol{h}^{(t)}\|_2^2 - \frac{1}{100}\left\|\boldsymbol{H}_{T^\perp}^{(t)}\right\|_*^2\right) \\ & \quad - n\sigma^2\|\boldsymbol{h}^{(t)}\|_2^2, \end{aligned}$$

where the first inequality follows from (33) by replacing the $(\hat{\boldsymbol{H}}, \hat{\boldsymbol{h}})$ there by $(\boldsymbol{H}^{(t)}, \boldsymbol{h}^{(t)})$; the second inequality is from the concentration property of operator $\mathcal{A}_b$ as in the proof of part (a). Using Lemma 7 to bound $\|\boldsymbol{H}_{T^\perp}^{(t)}\|_*^2$, we further have

$$\begin{aligned} & Q_n(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}}; \boldsymbol{K}^{(t)}, \boldsymbol{g}^{(t)}) \\ & \gtrsim n\left(0.38\left\|\boldsymbol{H}_T^{(t)}\right\|_F^2 + \gamma^2\|\boldsymbol{h}^{(t)}\|_2^2 + \sigma^2\|\boldsymbol{h}^{(t)}\|_2^2 - 0.32\left\|\hat{\boldsymbol{H}}_T\right\|_*^2\right) \\ & \quad - n\sigma^2\|\boldsymbol{h}^{(t)}\|_2^2 \\ & \gtrsim n\left\|\boldsymbol{H}^{(t)}\right\|_F^2 + n\gamma^2\|\boldsymbol{h}^{(t)}\|_2^2 - n\left\|\hat{\boldsymbol{H}}\right\|_F^2. \end{aligned} \tag{42}$$

We need the following result to connect the established curvature property to the optimization error. Its proof is given in Appendix F-E.

Lemma 9. *Suppose function Q_n satisfies the following two conditions with functions $\overline{Q}_n$ and $\underline{Q}_n$ that map $\mathbb{R}^{p\times p} \times \mathbb{R}^p$ to $\mathbb{R}$:*

- *For $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ and every $(\boldsymbol{K}^{(t)}, \boldsymbol{g}^{(t)})$, $Q_n(\boldsymbol{K}^{(t)}, \boldsymbol{g}^{(t)}; \hat{\boldsymbol{K}}, \hat{\boldsymbol{g}}) \ge \underline{Q}_n(\boldsymbol{H}^{(t)}, \boldsymbol{h}^{(t)})$;*
- *For any $\boldsymbol{K}, \boldsymbol{K}' \in \mathbb{R}^{p\times p}$ and $\boldsymbol{g}, \boldsymbol{g}' \in \mathbb{R}^p$, $Q_n(\boldsymbol{K}', \boldsymbol{g}'; \boldsymbol{K}, \boldsymbol{g}) \le \overline{Q}_n(\boldsymbol{K}' - \boldsymbol{K}, \boldsymbol{g}' - \boldsymbol{g})$.*

Then we have

$$\begin{aligned} & \|\boldsymbol{H}^{(t+1)}\|_F^2 + \gamma^2\|\boldsymbol{h}^{(t+1)}\|_2^2 \\ & \le \|\boldsymbol{H}^{(t)}\|_F^2 + \gamma^2\|\boldsymbol{h}^{(t)}\|_2^2 - \frac{2}{\eta}\underline{Q}_n\left(\boldsymbol{H}^{(t)}, \boldsymbol{h}^{(t)}\right) - \|\boldsymbol{\Delta}^{(t)}\|_F^2 \\ & \quad - \gamma^2\|\boldsymbol{\delta}^{(t)}\|_2^2 + \frac{2}{\eta}\overline{Q}_n\left(\boldsymbol{\Delta}^{(t)}, \boldsymbol{\delta}^{(t)}\right), \end{aligned}$$

where we let $\boldsymbol{\Delta}^{(t)} := \boldsymbol{K}^{(t+1)} - \boldsymbol{K}^{(t)}$ and $\boldsymbol{\delta}^{(t)} := \boldsymbol{g}^{(t+1)} - \boldsymbol{g}^{(t)}$.

Now we are ready to prove Theorem 5 part (b). Plugging the established bounds of the function Q_n, (41) and (42), into Lemma 9, we obtain that for some constants c_1, c_2,

$$\begin{aligned} & \|\boldsymbol{H}^{(t+1)}\|_F^2 + \gamma^2\|\boldsymbol{h}^{(t+1)}\|_2^2 \\ & \le \left(1 - \frac{c_1 n}{\eta}\right)\left(\|\boldsymbol{H}^{(t)}\|_F^2 + \gamma^2\|\boldsymbol{h}^{(t)}\|_2^2\right) + \frac{c_1 n}{\eta}\|\hat{\boldsymbol{H}}\|_F^2 \\ & + \left(\frac{c_2 n\log n}{\eta} - 1\right)\|\boldsymbol{\delta}^{(t)}\|_2^2 + \frac{c_2 n\sqrt{p}^3\log n}{\eta}\|\boldsymbol{\Delta}^{(t)}\|_F^2 - \|\boldsymbol{\Delta}^{(t)}\|_F^2. \end{aligned}$$

Therefore by setting $\eta > c_3 n\sqrt{p}^3 \log n$ for sufficiently large constant c_3, we have

$$\begin{aligned}&\|\boldsymbol{H}^{(t+1)}\|_F^2 + \gamma^2\|\boldsymbol{h}^{(t+1)}\|_2^2\\ &\le \left(1-\frac{c_1 n}{\eta}\right)\left(\|\boldsymbol{H}^{(t)}\|_F^2 + \gamma^2\|\boldsymbol{h}^{(t)}\|_2^2\right) + \frac{c_1 n}{\eta}\|\hat{\boldsymbol{H}}\|_F^2.\end{aligned}$$

Applying this bound recursively to $t = 0, 1, 2, \ldots$ proves our result.

E. Proofs of Theorems 2, 4 and 6

In this section, we show that an error bound on the input $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ of Algorithm 1 implies an error bound on its output $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$. Recall the quantities $\hat{\boldsymbol{J}}$, $\boldsymbol{J}^*$, $\hat{\lambda}$, λ^*,$\hat{\boldsymbol{v}}$ and $\boldsymbol{v}^*$ defined in Section III-A and in Algorithm 1.

The main component of the proof is a general perturbation bound. We prove these in the first section below, and then use them to prove Theorems 2, 4 and 6 in the three subsequent sections.

1) Perturbation Bound: We require the following perturbation bound.

Lemma 10. *If* $\left\|\hat{\boldsymbol{J}} - \boldsymbol{J}^*\right\|_F \le \delta$, *then*

$$\left\|\sqrt{\hat{\lambda}}\hat{\boldsymbol{v}} - \sqrt{\lambda^*}\boldsymbol{v}^*\right\|_2 \le 10\min\left\{\frac{\delta}{\sqrt{\|\boldsymbol{J}^*\|}}, \sqrt{\delta}\right\}.$$

Proof. By Weyl's inequality, we have

$$\left|\hat{\lambda} - \lambda^*\right| \le \left\|\hat{\boldsymbol{J}} - \boldsymbol{J}^*\right\| \le \delta.$$

This implies

$$\left|\sqrt{\hat{\lambda}} - \sqrt{\lambda^*}\right| = \left|\frac{\hat{\lambda} - \lambda^*}{\sqrt{\hat{\lambda}} + \sqrt{\lambda^*}}\right| \le 2\min\left\{\frac{\delta}{\sqrt{\lambda^*}}, \sqrt{\delta}\right\}. \quad (43)$$

Using Weyl's inequality and the Davis-Kahan sine theorem, we obtain

$$|\sin\angle(\hat{\boldsymbol{v}}, \boldsymbol{v}^*)| \le \min\left\{\frac{2\|\hat{\boldsymbol{K}} - \boldsymbol{K}^*\|}{\|\boldsymbol{K}^*\|}, 1\right\} \le \min\left\{\frac{2\delta}{\lambda^*}, 1\right\}. \quad (44)$$

On the other hand, we have

$$\begin{aligned}&\left\|\hat{\boldsymbol{v}}\sqrt{\hat{\lambda}} - \boldsymbol{v}^*\sqrt{\lambda^*}\right\|_2\\ &\le \left\|\hat{\boldsymbol{v}}\sqrt{\hat{\lambda}} - \boldsymbol{v}^*\sqrt{\hat{\lambda}}\right\|_2 + \left\|\boldsymbol{v}^*\sqrt{\hat{\lambda}} - \boldsymbol{v}^*\sqrt{\lambda^*}\right\|_2\\ &= \sqrt{\hat{\lambda}}\left\|\hat{\boldsymbol{v}} - \boldsymbol{v}^*\right\|_2 + \|\boldsymbol{v}^*\|_2\left|\sqrt{\hat{\lambda}} - \sqrt{\lambda^*}\right|\\ &= \left(\sqrt{\lambda^*} + \sqrt{\hat{\lambda}} - \sqrt{\lambda^*}\right)\left\|\hat{\boldsymbol{v}} - \boldsymbol{v}^*\right\|_2 + \|\boldsymbol{v}^*\|_2\left|\sqrt{\hat{\lambda}} - \sqrt{\lambda^*}\right|\\ &\le \sqrt{\lambda^*}\left\|\hat{\boldsymbol{v}} - \boldsymbol{v}^*\right\|_2 + 3\left|\sqrt{\hat{\lambda}} - \sqrt{\lambda^*}\right|,\end{aligned}$$

where in the last inequality we use the fact that $\|\boldsymbol{v}^*\| = \|\hat{\boldsymbol{v}}\| = 1$. Elementary calculation shows that

$$\left\|\hat{\boldsymbol{v}} - \boldsymbol{v}^*\right\|_2 = 2\left|\sin\frac{1}{2}\angle(\hat{\boldsymbol{v}}, \boldsymbol{v}^*)\right| \le \sqrt{2}\left|\sin\angle(\hat{\boldsymbol{v}}, \boldsymbol{v}^*)\right|.$$

It follows that

$$\begin{aligned}&\left\|\hat{\boldsymbol{v}}\sqrt{\hat{\lambda}} - \boldsymbol{v}^*\sqrt{\lambda^*}\right\|_2\\ &\le \sqrt{2}\sqrt{\lambda^*}\left|\sin\angle(\hat{\boldsymbol{v}}, \boldsymbol{v}^*)\right| + 3\left|\sqrt{\hat{\lambda}} - \sqrt{\lambda^*}\right|\\ &\le \sqrt{2}\min\left\{\frac{2\delta}{\sqrt{\lambda^*}}, \sqrt{\lambda^*}\right\} + 6\min\left\{\frac{\delta}{\sqrt{\lambda^*}}, \sqrt{\delta}\right\}\\ &\le 10\min\left\{\frac{\delta}{\sqrt{\lambda^*}}, \sqrt{\delta}\right\},\end{aligned}$$

where we use (43) and (44) in the second inequality. This concludes the proof.

We can now use this perturbation result to provide guarantees on recovering $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ given noisy versions of $\boldsymbol{g}^*$ and $\boldsymbol{K}^*$. To this end, suppose we are given $\hat{\boldsymbol{K}}$ and $\hat{\boldsymbol{g}}$ which satisfy

$$\left\|\hat{\boldsymbol{K}} - \boldsymbol{K}^*\right\|_F \le \delta_K, \qquad \|\hat{\boldsymbol{g}} - \boldsymbol{g}^*\|_2 \le \delta_g.$$

Then by the triangle inequality we have

$$\left\|\hat{\boldsymbol{J}} - \boldsymbol{J}^*\right\|_F \le \delta_K + 2\delta_g\|\boldsymbol{g}^*\|_2 + \delta_g^2.$$

Therefore, up to relabeling b, we have

$$\begin{aligned}&\left\|\hat{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_b^*\right\|_2\\ &\le \|\hat{\boldsymbol{g}} - \boldsymbol{g}^*\|_2 + \left\|\sqrt{\hat{\lambda}}\hat{\boldsymbol{v}} - \sqrt{\lambda^*}\boldsymbol{v}^*\right\|_2\\ &\lesssim \delta_g + \min\left\{\frac{\delta_K + 2\delta_g\|\boldsymbol{g}^*\|_2 + \delta_g^2}{\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|_2}, \sqrt{\delta_K + 2\delta_g\|\boldsymbol{g}^*\|_2 + \delta_g^2}\right\},\end{aligned} \quad (45)$$

where the second inequality follows from Lemma 10 and $\lambda^* = \frac{1}{4}\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|_2^2$.

2) Proof of Theorem 2 (Arbitrary Noise): In the case of arbitrary noise, as set up above, Theorem 1 guarantees the following:

$$\begin{aligned}\delta_K &\asymp \frac{\sqrt{n}\|\boldsymbol{e}\|_2\|\boldsymbol{\beta}_2^* - \boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{e}\|_2^2}{\sqrt{\alpha}n} \lesssim \frac{1}{\sqrt{\alpha}}\frac{\|\boldsymbol{e}\|_2}{\sqrt{n}}\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|,\\ \delta_g &\asymp \frac{\sqrt{n}\|\boldsymbol{e}\|_2\|\boldsymbol{\beta}_2^* - \boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{e}\|_2^2}{\sqrt{\alpha}n\left(\|\boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^*\|_2\right)} \lesssim \frac{\|\boldsymbol{e}\|_2}{\sqrt{n}}.\end{aligned}$$

where we use the assumption $\|\boldsymbol{e}\|_2 \le \frac{\sqrt{\alpha}}{c_4}\sqrt{n}\left(\|\boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^*\|_2\right) \asymp \frac{1}{c_4}\sqrt{n}\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|_2$. Using (45), we get that up to relabeling b,

$$\begin{aligned}&\left\|\hat{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_b^*\right\|_2\\ &\lesssim \frac{\|\boldsymbol{e}\|_2}{\sqrt{n}} + \min\left\{\frac{1}{\sqrt{\alpha}}\frac{\|\boldsymbol{e}\|_2}{\sqrt{n}} + \frac{\|\boldsymbol{e}\|_2^2}{n\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|_2},\right.\\ &\qquad\left.\sqrt{\frac{1}{\sqrt{\alpha}}\frac{\|\boldsymbol{e}\|_2}{\sqrt{n}}\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|_2 + \frac{\|\boldsymbol{e}\|_2^2}{n}}\right\}\\ &\lesssim \frac{1}{\sqrt{\alpha}}\frac{\|\boldsymbol{e}\|_2}{\sqrt{n}} + \min\left\{\frac{\|\boldsymbol{e}\|_2^2}{n\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|_2}, \sqrt{\frac{1}{\sqrt{\alpha}}\frac{\|\boldsymbol{e}\|_2}{\sqrt{n}}\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|_2}\right\}\\ &\le \frac{1}{\sqrt{\alpha}}\frac{\|\boldsymbol{e}\|_2}{\sqrt{n}}.\end{aligned}$$

3) Proof of Theorem 4 (Stochastic Noise): Next consider the setting with stochastic noise. Under the assumptions of Theorem 4, Theorem 3 guarantees the following bounds on the errors in recovering $\boldsymbol{K}^*$ and $\boldsymbol{g}^*$:

$$\delta_K \asymp \sigma \left(\|\boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^*\|_2 + \sigma\right) \sqrt{\frac{p}{n}} \log^4 n,$$
$$\delta_g \asymp \sigma \sqrt{\frac{p}{n}} \log^4 n.$$

If we let $\gamma = \|\beta_1^*\|_2 + \|\beta_2^*\|_2$, then this means

$$\begin{aligned}
&\delta_K + 2\delta_g \|\boldsymbol{g}^*\|_2 + \delta_g^2 \\
\asymp &\sigma\gamma \sqrt{\frac{p}{n}} \log^4 n + \sigma^2 \sqrt{\frac{p}{n}} \log^4 n + \sigma^2 \frac{p}{n} \log^8 n \\
\lesssim &\sigma\gamma \sqrt{\frac{p}{n}} \log^4 n + \sigma^2 \sqrt{\frac{p}{n}} \log^4 n,
\end{aligned}$$

where the last inequality follows from the assumption that $n \geq cp \log^8 n$ for some $c > 1$. Combining these with (45), we obtain that up to relabeling of b,

$$\begin{aligned}
&\left\|\hat{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_b^*\right\|_2 \\
&\lesssim \sigma \sqrt{\frac{p}{n}} \log^4 n \\
&\quad + \min\left\{ \frac{\sigma\gamma\sqrt{\frac{p}{n}} + \sigma^2\sqrt{\frac{p}{n}}}{\sqrt{\alpha}\gamma}, \sqrt{\sigma\gamma\sqrt{\frac{p}{n}} + \sigma^2\sqrt{\frac{p}{n}}} \right\} \log^4 n \\
&\lesssim \sigma \sqrt{\frac{p}{n}} \log^4 n + \min\left\{ \frac{\sigma^2\sqrt{\frac{p}{n}}}{\gamma}, \sqrt{\sigma\gamma\sqrt{\frac{p}{n}} + \sigma^2\sqrt{\frac{p}{n}}} \right\} \log^4 n,
\end{aligned}$$

where the last inequality follows from α being lower-bounded by a constant. Observe that the minimization in the last RHS is no larger than $\sigma\sqrt{\frac{p}{n}}$ if $\gamma \geq \sigma$, and equals $\min\left\{\frac{\sigma^2\sqrt{\frac{p}{n}}}{\gamma}, \sigma\left(\frac{p}{n}\right)^{1/4}\right\}$ if $\gamma < \sigma$. It follows that

$$\begin{aligned}
&\left\|\hat{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_b^*\right\|_2 \\
&\lesssim \sigma \sqrt{\frac{p}{n}} \log^4 n + \min\left\{ \frac{\sigma^2\sqrt{\frac{p}{n}}}{\gamma}, \sigma\left(\frac{p}{n}\right)^{1/4} \right\} \log^4 n.
\end{aligned}$$

4) Proof of Theorem 6 (Nonconvex Formulation): Under our assumption, we have $\|\boldsymbol{K}^{(T)} - \hat{\boldsymbol{K}}\|_F \lesssim \|\hat{\boldsymbol{H}}\|_F, \|\boldsymbol{g}^{(T)} - \hat{\boldsymbol{g}}\|_2 \lesssim \|\hat{\boldsymbol{h}}\|_2$. From the triangle inequality, we have

$$\begin{aligned}
\|\boldsymbol{K}^{(T)} - \boldsymbol{K}^*\|_F &\leq \|\boldsymbol{K}^{(T)} - \hat{\boldsymbol{K}}\|_F + \|\hat{\boldsymbol{H}}\|_F \\
&\lesssim \sigma \left(\|\boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^*\|_2\right) \sqrt{\frac{p}{n}} \log^3 n,
\end{aligned}$$

and

$$\|\boldsymbol{g}^{(T)} - \boldsymbol{g}^*\|_2 \leq \|\boldsymbol{g}^{(T)} - \hat{\boldsymbol{g}}\|_2 + \|\hat{\boldsymbol{h}}\|_2 \lesssim \sigma \sqrt{\frac{p}{n}} \log^3 n.$$

From here the proof follows along the lines of Section IV-E3. We omit the details.

F. Proof Outlines of Theorems 7 and 8

Next, we provide the main steps of proving the minimax lower bounds in Theorems 7 and 8, and postpone the full proofs to Appendix D and E. The high-level ideas in the proofs of Theorems 7 and 8 are similar: we use a standard argument [5, 38, 35] to convert the estimation problem into a hypothesis testing problem, and then use information-theoretic inequalities to lower bound the error probability in hypothesis testing. In particular, recall the definition of the set $\Theta(\underline{\gamma})$ of regression vector pairs in (10); we construct a δ-packing $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ of $\Theta(\underline{\gamma})$ in the metric ρ, and use the following inequality:

$$\inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}^* \in \Theta(\underline{\gamma})} \mathbb{E}\left[\rho(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)\right] \geq \delta \inf_{\tilde{\boldsymbol{\theta}}} \mathbb{P}\left(\tilde{\boldsymbol{\theta}} \neq \boldsymbol{\theta}^*\right), \tag{46}$$

where on the RHS $\boldsymbol{\theta}^*$ is assumed to be sampled uniformly at random from Θ. To lower-bound the minimax expected error by $\frac{1}{2}\delta$, it suffices to show that the probability on the last RHS is at least $\frac{1}{2}$. By Fano's inequality [17], we have

$$\mathbb{P}\left(\tilde{\boldsymbol{\theta}} \neq \boldsymbol{\theta}^*\right) \geq 1 - \frac{I\left(\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\theta}^*\right) + \log 2}{\log M}. \tag{47}$$

It remains to construct a packing set Θ with the appropriate separation δ and cardinality M, and to upper-bound the mutual information $I\left(\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\theta}^*\right)$. We show how to do this for Part 2 of Theorem 8, for which the desired separation is $\delta = 2c_0 \frac{\sigma^2}{\kappa}\sqrt{\frac{p}{n}}$, where $\kappa = \frac{\gamma}{2}$. Let $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M\}$ be a $\frac{p-1}{16}$-packing of $\{0,1\}^{p-1}$ in Hamming distance with $\log M \geq (p-1)/16$, which exists by the Varshamov-Gilbert bound [29]. We construct Θ by setting $\boldsymbol{\theta}_i := (\boldsymbol{\beta}_i, -\boldsymbol{\beta}_i)$ for $i = 1, \ldots, M$ with

$$\boldsymbol{\beta}_i = \kappa_0 \boldsymbol{\epsilon}_p + \sum_{j=1}^{p-1} \left(2\boldsymbol{\xi}_i(j) - 1\right) \tau \boldsymbol{\epsilon}_j,$$

where $\tau = \frac{4\delta}{\sqrt{p-1}}$, $\kappa_0^2 = \kappa^2 - (p-1)\tau^2$, and $\boldsymbol{\epsilon}_j$ is the j^{th} standard basis in $\mathbb{R}^p$. We verify that this Θ indeed defines a δ-packing of $\Theta(\underline{\gamma})$, and moreover satisfies $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}\|^2 \leq 16\delta^2$ for all $i \neq i'$. To bound the mutual information, we observe that by independence between $\boldsymbol{X}$ and $\boldsymbol{\theta}^*$, we have

$$\begin{aligned}
I\left(\boldsymbol{\theta}^*; \boldsymbol{X}, \boldsymbol{y}\right) \leq &\frac{1}{M^2} \sum_{1 \leq i, i' \leq M} D\left(\mathbb{P}_i \| \mathbb{P}_{i'}\right) \\
= &\frac{1}{M} \sum_{1 \leq i, i' \leq M} \sum_{j=1}^{n} \mathbb{E}_{\boldsymbol{X}}\left[D\left(\mathbb{P}_{i,\boldsymbol{X}}^{(j)} \| \mathbb{P}_{i',\boldsymbol{X}}^{(j)}\right)\right],
\end{aligned}$$

where $\mathbb{P}_{i,\boldsymbol{X}}^{(j)}$ denotes the distribution of y_j conditioned on $\boldsymbol{X}$ and $\boldsymbol{\theta}^* = \boldsymbol{\theta}_i$. The remaining and crucial step is to obtain sharp upper bounds on the above KL-divergence between two mixtures of one-dimensional Gaussian distributions. This requires some technical calculations, from which we obtain

$$\mathbb{E}_{\boldsymbol{X}} D\left(\mathbb{P}_{i,\boldsymbol{X}}^{(j)} \| \mathbb{P}_{i',\boldsymbol{X}}^{(j)}\right) \leq \frac{c' \left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}\right\|^2 \kappa^2}{\sigma^4}.$$

We conclude that $I(\boldsymbol{\theta}^*; \boldsymbol{X}, \boldsymbol{y}) \leq \frac{1}{4}\log M$. Combining with (46) and (47) proves Part 2 of Theorem 8. Theorem 7 and the remaining part of Theorem 8 are proved in a similar manner.

V. Conclusion

This paper provides a computationally and statistically efficient algorithm for mixed regression with two components. To the best of our knowledge, the is the first efficient algorithm that can provide $O(p)$ sample complexity guarantees. Under certain conditions, we prove matching lower bounds, thus demonstrating our algorithm achieves the minimax optimal rates. There are several interesting open questions that remain. Most immediate is the issue of understanding the degree to which the assumptions currently required for minimax optimality can be removed or relaxed. The extension to more than two components is important, though how to do this within the current framework is not obvious.

At its core, the approach here is a method of moments, as the convex optimization formulation produces an estimate of the cross moments, $(\boldsymbol{\beta}_1^* \boldsymbol{\beta}_2^{*\top} + \boldsymbol{\beta}_2^* \boldsymbol{\beta}_1^{*\top})$. An interesting aspect of these results is the significant improvement in sample complexity guarantees this tailored approach brings, compared to a more generic implementation of the tensor machinery which requires use of third order moments. Given the statistical and also computational challenges related to third order tensors, understanding the connections more carefully seems to be an important future direction.

Appendix A
Proof of Lemma 4

We now move to the proof of Lemma 4, which bounds the noise terms P and Q. Note that

$$\begin{aligned} P &= 2\sum_b \|n_b \mathcal{A}_b^* \boldsymbol{w}_b\| \\ &\le \underbrace{2\sum_b \|n_b \mathcal{A}_b^* \boldsymbol{w}_{1,b}\|}_{S_1} + \underbrace{2\sum_b \|n_b \mathcal{A}_b^* \boldsymbol{w}_{2,b}\|}_{S_2}, \end{aligned}$$

and

$$\begin{aligned} Q &= \sum_b 4\|\boldsymbol{\beta}_b^*\|_2 \|n_b \mathcal{A}_b^* \boldsymbol{w}_b\| + \sqrt{p}\left\|\sum_b 4\boldsymbol{X}_b^\top \mathrm{diag}(\boldsymbol{e}_b)\boldsymbol{w}_b\right\|_\infty \\ &\le 4\gamma P + \underbrace{\sqrt{p}\left\|\sum_b 4\boldsymbol{X}_b^\top \mathrm{diag}(\boldsymbol{e}_b)\boldsymbol{w}_{1,b}\right\|_\infty}_{S_3} \\ &\quad + \underbrace{\sqrt{p}\left\|\sum_b 4\boldsymbol{X}_b^\top \mathrm{diag}(\boldsymbol{e}_b)\boldsymbol{w}_{2,b}\right\|_\infty}_{S_4}. \end{aligned}$$

Therefore, the lemma is implied if we can show

$$S_1 + S_2 \le \frac{\lambda}{2}, \quad S_3 + S_4 \le \sigma\lambda, \qquad \text{w.h.p.}$$

But $\lambda \gtrsim \sigma(\gamma+\sigma)\left(\sqrt{np} + |n_1 - n_2|\sqrt{p}\right)\log^3 n$ by assumption of Theorem 3. Therefore, the lemma follows if each of the following bounds holds w.h.p.

$$\begin{aligned} S_1 &\lesssim \sigma\gamma\sqrt{np}\log^3 n, \\ S_2 &\lesssim \sigma^2\sqrt{np}\log^3 n, \\ S_3 &\lesssim \sigma^2\gamma\left(\sqrt{np} + |n_1 - n_2|\sqrt{p}\right)\log^2 n, \\ S_4 &\lesssim \sigma^3\sqrt{np}\log^2 n. \end{aligned}$$

We now prove these bounds.

a) Term S_1:: Note that $\gamma \ge \|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|_2$, so the desired bound on S_1 follows from the lemma below, which is proved in Section F-F.

Lemma 11. *Suppose $\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*$ is supported on the first coordinate. Then w.h.p.*

$$S_1 \lesssim \|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|_2 \, \sigma\sqrt{np}\log^3 n.$$

b) Term S_2:: By definition, we have

$$S_2 = 2\sum_b \left\|\sum_{i=1}^{n_b}\left(\boldsymbol{e}_{b,i}^2 - \sigma^2\right)\boldsymbol{x}_{b,i}\boldsymbol{x}_{b,i}^\top\right\|.$$

Here each $\boldsymbol{e}_{b,i}^2 - \sigma^2$ is zero-mean, $\lesssim \sigma^2 \log n$ almost surely, and has variance $\lesssim \sigma^4$. The quantity inside the spectral norm is the sum of independent zero-mean bounded matrices. An application of the Matrix Bernstein inequality [28] gives

$$\left\|\sum_{i=1}^{n_b}\left(\boldsymbol{e}_{b,i}^2 - \sigma^2\right)\boldsymbol{x}_{b,i}\boldsymbol{x}_{b,i}^\top\right\| \lesssim \sigma^2\sqrt{n_b p}\log^3 n_b,$$

for each $b = 1, 2$. The desired bound follows.

c) Term S_3:: We have

$$\begin{aligned} S_3/4 &= \sqrt{p}\left\|\sum_b \boldsymbol{X}_b^\top \mathrm{diag}\left(\boldsymbol{e}_b\right)\left(-\boldsymbol{e}_b \circ \left(\boldsymbol{X}_b\boldsymbol{\delta}_b^*\right)\right)\right\|_\infty \\ &= \sqrt{p}\left\|\sum_b \boldsymbol{X}_b^\top \mathrm{diag}\left(\boldsymbol{e}_b^2\right)\boldsymbol{X}_b\boldsymbol{\delta}_b^*\right\|_\infty \\ &= \sqrt{p}\max_{l\in[p]}\left|\sum_b \left(\boldsymbol{e}_b^2 \circ \boldsymbol{X}_{b,l}\right)^\top \boldsymbol{X}_b\boldsymbol{\delta}_b^*\right|, \end{aligned}$$

where $\boldsymbol{X}_{b,l}$ is the l^{th} column of $\boldsymbol{X}_b$. WLOG, we assume $n_1 \ge n_2$. Observe that for each $l \in [p]$,

$$\begin{aligned} &\sum_b \left(\boldsymbol{e}_b^2 \circ \boldsymbol{X}_{b,l}\right)^\top \boldsymbol{X}_b\boldsymbol{\delta}_b^* \\ &= \underbrace{\sum_{i=1}^{n_2}\left(e_{1,i}^2\boldsymbol{x}_{1,i}(l)\boldsymbol{x}_{1,i}^\top - e_{2,i}^2\boldsymbol{x}_{2,i}(l)\boldsymbol{x}_{2,i}^\top\right)\boldsymbol{\delta}_1^*}_{S_{3,1,l}} \\ &\quad + \underbrace{\sum_{i=n_2+1}^{n_1} e_{1,i}^2\boldsymbol{x}_{1,i}(l)\boldsymbol{x}_{1,i}^\top\boldsymbol{\delta}_1^*}_{S_{3,2,l}}. \end{aligned}$$

Let $\boldsymbol{\epsilon}_i$ be the i^{th} standard basis vector in $\mathbb{R}^n$. The term $S_{3,1,l}$ can be written as

$$\begin{aligned}S_{3,1,l} =& \sum_{i=1}^{n_2} \left(\boldsymbol{x}_{1,i}^\top \left(e_{1,i}^2 \boldsymbol{\epsilon}_l \boldsymbol{\delta}_1^{*\top}\right) \boldsymbol{x}_{1,i} - \boldsymbol{x}_{2,i}^\top \left(e_{2,i}^2 \boldsymbol{\epsilon}_l \boldsymbol{\delta}_1^{*\top}\right) \boldsymbol{x}_{2,i}\right)\\ =& \boldsymbol{\chi}^\top \boldsymbol{G} \boldsymbol{\chi},\end{aligned}$$

where we defined

$$\begin{aligned}\boldsymbol{\chi}^\top :=& \big[e_{1,1}\boldsymbol{x}_{1,1}^\top \;\; e_{1,2}\boldsymbol{x}_{1,2}^\top \;\; \cdots \;\; e_{1,n_2}\boldsymbol{x}_{1,n_2}^\top\\ & e_{2,1}\boldsymbol{x}_{2,1}^\top \;\; e_{2,2}\boldsymbol{x}_{2,2}^\top \;\; \cdots \;\; e_{2,n_2}\boldsymbol{x}_{2,n_2}^\top\big] \in \mathbb{R}^{2n_2p}\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{G} :=& \mathrm{diag}\big(\boldsymbol{\epsilon}_l\boldsymbol{\delta}_1^{*\top}, \boldsymbol{\epsilon}_l\boldsymbol{\delta}_1^{*\top}, \ldots, \boldsymbol{\epsilon}_l\boldsymbol{\delta}_1^{*\top},\\ & -\boldsymbol{\epsilon}_l\boldsymbol{\delta}_1^{*\top}, -\boldsymbol{\epsilon}_l\boldsymbol{\delta}_1^{*\top}, \ldots, -\boldsymbol{\epsilon}_l\boldsymbol{\delta}_1^{*\top}\big) \in \mathbb{R}^{2n_2p\times 2n_2p};\end{aligned}$$

in other words, $\boldsymbol{G}$ is the block-diagonal matrix with $\left\{\pm\boldsymbol{\epsilon}_l\boldsymbol{\delta}_1^{*\top}\right\}$ on its diagonal. Note that $\mathbb{E}S_{3,1,l} = 0$, and the entries of $\boldsymbol{\chi}$ are i.i.d. sub-Gaussian with parameter bounded by $\sigma\sqrt{\log n}$. Using the Hanson-Wright inequality (e.g., [25]), we obtain w.h.p.

$$\max_{l\in[p]} |S_{3,1,l}| \lesssim \|\boldsymbol{G}\|_F \sigma^2 \log^2 n \le \sigma^2\sqrt{2n}\gamma\log^2 n.$$

Since $\boldsymbol{\delta}_1^*$ is supported on the first coordinate, the term $S_{3,2,l}$ can be bounded w.h.p. by

$$\begin{aligned}\max_{l\in[p]} |S_{3,2,l}| =& \max_{l\in[p]} \left|\sum_{i=n_2+1}^{n_1} e_{1,i}^2 \boldsymbol{x}_{1,i}(l)\boldsymbol{x}_{1,i}(1)\boldsymbol{\delta}_1^*(1)\right|\\ \lesssim& (n_1-n_2)\,\sigma^2\gamma\log^2 n\end{aligned}$$

using Hoeffding's inequality. It follows that w.h.p.,

$$\begin{aligned}S_3 \lesssim& \sqrt{p}\max_{l\in[p]}\left(|S_{3,1,l}| + |S_{3,2,l}|\right)\\ \lesssim& \sigma^2\gamma\left(\sqrt{np} + |n_1-n_2|\sqrt{p}\right)\log^2 n.\end{aligned}$$

d) Term S_4:: We have w.h.p.

$$\begin{aligned}S_4/4 \le& \sqrt{p}\sum_b \left\|\boldsymbol{X}^\top\left(\boldsymbol{e}_b \circ \boldsymbol{w}_{2,b}\right)\right\|_\infty\\ \overset{(a)}{\lesssim}& \sqrt{p\log n}\sum_b \|\boldsymbol{e}_b\circ\boldsymbol{w}_{2,b}\|_2\\ =& \sqrt{p\log n}\sum_b \left\|\boldsymbol{e}_b^3 - \sigma^2\boldsymbol{e}_b\right\|_2\\ \overset{(b)}{\lesssim}& \sigma^3\sqrt{np}\log^2 n,\end{aligned}$$

where in (a) we use the independence between $\boldsymbol{X}$ and $\boldsymbol{e}_b\circ\boldsymbol{w}_{2,b}$ and the standard sub-Gaussian concentration inequality (e.g., [30]), and (b) follows from the boundedness of $\boldsymbol{e}$.

Appendix B
Proof of Lemma 5

The proofs for $b = 1$ and 2 are identical, so we omit the subscript b. WLOG we may assume $\sigma = 1$. Our proof generalizes the proof of an RIP-type result in [16]

Fix $\boldsymbol{Z}$ and $\boldsymbol{z}$. Let $\xi_j := \langle\boldsymbol{B}_j, \boldsymbol{Z}\rangle$ and $\nu := \|\boldsymbol{Z}\|_F$. We already know that ξ_j is a sub-exponential random variable with $\|\xi_j\|_{\psi_1} \le c_1\nu$ and $\|\xi_j - \mathbb{E}[\xi_j]\|_{\psi_1} \le 2c_1\nu$.

Let $\gamma_j = \langle\boldsymbol{d}_j, \boldsymbol{z}\rangle$ and $\omega := \|\boldsymbol{z}\|_2$. It is easy to check that γ_j is sub-Gaussian with $\|\gamma_j\|_{\psi_2} \le c_1\mu$. It follows that $\|\xi_j - \gamma_j\|_{\psi_1} \le c_1(\nu+\omega)$.

Note that

$$\|\mathcal{B}\boldsymbol{Z} - \boldsymbol{D}\boldsymbol{z}\|_1 = \sum_{j=1}^{n/2}\frac{2}{n}|\xi_j - \gamma_j|.$$

Therefore, applying the Bernstein-type inequality for the sum of sub-exponential variables [30], we obtain

$$\begin{aligned}&\mathbb{P}\left[\left|\|\mathcal{B}\boldsymbol{Z} - \boldsymbol{D}\boldsymbol{z}\|_1 - \mathbb{E}|\xi_j - \gamma_j|\right| \ge t\right]\\ \le& 2\exp\left[-c\min\left\{\frac{t^2}{c_2(\nu+\mu)^2/n}, \frac{t}{c_2(\nu+\mu)/n}\right\}\right].\end{aligned}$$

Setting $t = (\nu+\sigma\omega)/c_3$ for any $c_3 > 1$, we get

$$\mathbb{P}\left[\left|\|\mathcal{B}\boldsymbol{Z} - \boldsymbol{D}\boldsymbol{z}\|_1 - \mathbb{E}|\xi_j - \gamma_j|\right| \ge \frac{\nu+\omega}{c_3}\right] \le 2\exp\left[-c_4 n\right]. \tag{48}$$

But sub-exponentiality implies

$$\mathbb{E}[|\xi_j - \gamma_j|] \le \|\xi_j - \gamma_j\|_{\psi_1} \le c_2(\nu+\mu),$$

and

$$\mathbb{P}\left[\|\mathcal{B}\boldsymbol{Z} - \boldsymbol{D}\boldsymbol{z}\|_1 \ge \left(c_2 + \frac{1}{c_3}\right)(\nu+\omega)\right] \le 2\exp\left[-c_4 n\right].$$

Now, note that

$$\mathbb{E}[|\xi_j - \gamma_j|] \ge \sqrt{\frac{\left(\mathbb{E}\left[(\xi_j-\gamma_j)^2\right]\right)^3}{\mathbb{E}\left[(\xi_j-\gamma_j)^4\right]}}.$$

We bound the numerator and denominator. By sub-exponentiality, we have $\mathbb{E}\left[(\xi_j-\gamma_j)^4\right] \le c_5(\nu+\omega)^4$. On the other hand, note that

$$\begin{aligned}&\mathbb{E}(\xi_j-\gamma_j)^2\\ =& \mathbb{E}\left(\langle\boldsymbol{B}_j,\boldsymbol{Z}\rangle - \langle\boldsymbol{d}_j,\boldsymbol{z}\rangle\right)^2\\ =& \mathbb{E}\langle\boldsymbol{B}_j,\boldsymbol{Z}\rangle^2 + \mathbb{E}\langle\boldsymbol{d}_j,\boldsymbol{z}\rangle^2 - 2\mathbb{E}\left[\langle\boldsymbol{B}_j,\boldsymbol{Z}\rangle\langle\boldsymbol{d}_j,\boldsymbol{z}\rangle\right]\\ =& \mathbb{E}\langle\boldsymbol{B}_j,\boldsymbol{Z}\rangle^2 + \mathbb{E}\left\langle\boldsymbol{d}_j\boldsymbol{d}_j^\top, \boldsymbol{z}\boldsymbol{z}^\top\right\rangle\\ & - 2\mathbb{E}\left[\langle\boldsymbol{B}_j,\boldsymbol{Z}\rangle\langle e_{2j}\boldsymbol{x}_{2j} - e_{2j-1}\boldsymbol{x}_{2j-1}, \boldsymbol{z}\rangle\right]\\ =& \mathbb{E}\langle\boldsymbol{B}_j,\boldsymbol{Z}\rangle^2 + \mathbb{E}\left\langle\boldsymbol{d}_j\boldsymbol{d}_j^\top, \boldsymbol{z}\boldsymbol{z}^\top\right\rangle - 2\mathbb{E}[e_{2j}]\,\mathbb{E}\left[\langle\boldsymbol{B}_j,\boldsymbol{Z}\rangle\langle\boldsymbol{x}_{2j},\boldsymbol{z}\rangle\right]\\ & - 2\mathbb{E}[e_{2j-1}]\,\mathbb{E}\left[\langle\boldsymbol{B}_{j-1},\boldsymbol{Z}\rangle\langle\boldsymbol{x}_{2j-1},\boldsymbol{z}\rangle\right]\\ =& \mathbb{E}\langle\boldsymbol{B}_j,\boldsymbol{Z}\rangle^2 + \mathbb{E}\left\langle\boldsymbol{d}_j\boldsymbol{d}_j^\top, \boldsymbol{z}\boldsymbol{z}^\top\right\rangle,\end{aligned}$$

where in the last equality we use the fact that $\{e_i\}$ are independent of $\{\boldsymbol{x}_i\}$ and $\mathbb{E}[e_i] = 0$ for all i. We already know

$$\begin{aligned}\mathbb{E}\langle\boldsymbol{B}_j,\boldsymbol{Z}\rangle^2 =& \langle\mathbb{E}\left[\langle\boldsymbol{B}_j,\boldsymbol{Z}\rangle\boldsymbol{B}_j\right],\boldsymbol{Z}\rangle\\ =& 4\|\boldsymbol{Z}\|_F^2 + 2(\mu-3)\|\mathrm{diag}(\boldsymbol{Z})\|_F^2\\ \ge& 2(\mu-1)\|\boldsymbol{Z}\|_F^2.\end{aligned}$$

Some calculation shows that

$$\begin{aligned}\mathbb{E}\left\langle\boldsymbol{d}_j\boldsymbol{d}_j^\top, \boldsymbol{z}\boldsymbol{z}^\top\right\rangle =& \left\langle\mathbb{E}\left[e_{2j}^2\boldsymbol{x}_{2j}\boldsymbol{x}_{2j}^\top + e_{2j}^2\boldsymbol{x}_{2j}\boldsymbol{x}_{2j}^\top\right], \boldsymbol{z}\boldsymbol{z}^\top\right\rangle\\ =& 2\left\langle\boldsymbol{I}, \boldsymbol{z}\boldsymbol{z}^\top\right\rangle = 2\|\boldsymbol{z}\|^2.\end{aligned}$$

It follows that

$$\mathbb{E}(\xi_j-\gamma_j)^2 \ge 2(\mu-1)\|\boldsymbol{Z}\|_F^2 + 2\|\boldsymbol{z}\|^2 \ge c_6\left(\nu^2+\omega^2\right),$$

where the inequality holds when $\mu > 1$. We therefore obtain

$$\mathbb{E}\left[|\xi_j - \gamma_j|\right] \geq c_7 \frac{\sqrt{(\nu^2+\omega^2)^3}}{(\nu+\omega)^2} \geq c_8(\nu+\omega).$$

Substituting back to (48), we get

$$\mathbb{P}\left[\|\mathcal{B}\boldsymbol{Z} - \boldsymbol{D}\boldsymbol{z}\|_1 \leq \left(c_8 - \frac{1}{c_3}\right)(\nu+\omega)\right] \leq 2\exp\left[-c_4 n\right].$$

To complete the proof of the lemma, we use an ϵ-net argument. Let $\mathcal{S}_r$ be the set

$$\left\{(\boldsymbol{Z},\boldsymbol{z}) \in \mathbb{R}^{p\times p}\times\mathbb{R}^p : \mathrm{rank}(\boldsymbol{Z}) \leq r, \|\boldsymbol{Z}\|_F^2 + \|\boldsymbol{z}\|_2^2 = 1\right\}.$$

We need the following lemma, which is proved in Section F-G.

Lemma 12. *For each $\epsilon > 0$ and $r \geq 1$, there exists a set $\mathcal{N}_r(\epsilon)$ with $|\mathcal{N}_r(\epsilon)| \leq \left(\frac{40}{\epsilon}\right)^{10pr}$ which is an ϵ-covering of $\mathcal{S}_r$, meaning that for all $(\boldsymbol{Z},\boldsymbol{z}) \in \mathcal{S}_r$, there exists $\left(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{z}}\right) \in \mathcal{N}_r(\epsilon)$ such that*

$$\sqrt{\left\|\tilde{\boldsymbol{Z}} - \boldsymbol{Z}\right\|_F^2 + \|\tilde{\boldsymbol{z}} - \boldsymbol{z}\|_2^2} \leq \epsilon.$$

Note that $\frac{1}{\sqrt{2}}\left(\|\boldsymbol{Z}\|_F + \|\boldsymbol{z}\|_2\right) \leq \sqrt{\|\boldsymbol{Z}\|_F^2 + \|\boldsymbol{z}\|_2^2} \leq \|\boldsymbol{Z}\|_F + \|\boldsymbol{z}\|_2$ for all $\boldsymbol{Z}$ and $\boldsymbol{z}$. Therefore, up to a change of constant, it suffices to prove Lemma 5 for all $(\boldsymbol{Z},\boldsymbol{z})$ in $\mathcal{S}_r$. By the union bound and Lemma 12, we have

$$\begin{aligned}&\mathbb{P}\left(\max_{(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{z}})\in\mathcal{N}_r(\epsilon)}\left\|\mathcal{B}\tilde{\boldsymbol{Z}} - \boldsymbol{D}\tilde{\boldsymbol{z}}\right\|_1 \leq 2\left(c_2 + \frac{1}{c_3}\right)\right)\\ \geq& 1 - |\mathcal{N}_r(\epsilon)|\cdot\exp\left(-c_4 n\right) \geq 1 - \exp(-c_4 n/2),\end{aligned}$$

when $n \geq (2/c_4)\cdot 10pr\log(40/\epsilon)$. On this event, we have

$$\begin{aligned}\bar{M} :=& \sup_{(\boldsymbol{Z},\boldsymbol{z})\in\mathcal{S}_r}\|\mathcal{B}\boldsymbol{Z} - \boldsymbol{D}\boldsymbol{z}\|_1\\ \leq& \max_{(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{z}})\in\mathcal{N}_r(\epsilon)}\left\|\mathcal{B}\tilde{\boldsymbol{Z}} - \boldsymbol{D}\tilde{\boldsymbol{z}}\right\|_1\\ &+ \sup_{(\boldsymbol{Z},\boldsymbol{z})\in\mathcal{S}_r}\left\|\mathcal{B}(\boldsymbol{Z}-\tilde{\boldsymbol{Z}}) - \boldsymbol{D}(\boldsymbol{z}-\tilde{\boldsymbol{z}})\right\|_1\\ \leq& 2\left(c_2 + \frac{1}{c_3}\right) + \sup_{\boldsymbol{Z}\in S_r}\sqrt{\left\|\boldsymbol{Z}-\tilde{\boldsymbol{Z}}\right\|_F^2 + \|\boldsymbol{z}-\tilde{\boldsymbol{z}}\|_2^2}\\ &\times \sup_{(\boldsymbol{Z}',\boldsymbol{z}')\in\mathcal{S}_{2r}}\|\mathcal{B}\boldsymbol{Z}' - \boldsymbol{D}\boldsymbol{z}'\|_1\\ \leq& 2\left(c_2 + \frac{1}{c_3}\right) + \epsilon\sup_{(\boldsymbol{Z}',\boldsymbol{z}')\in\mathcal{S}_{2r}}\|\mathcal{B}\boldsymbol{Z}' - \boldsymbol{D}\boldsymbol{z}'\|_1.\end{aligned}$$

Note that for $(\boldsymbol{Z}',\boldsymbol{z}') \in \mathcal{S}_{2r}$, we can write $\boldsymbol{Z}' = \boldsymbol{Z}_1' + \boldsymbol{Z}_2'$ such that $\boldsymbol{Z}_1', \boldsymbol{Z}_2'$ both have rank r and $1 = \|\boldsymbol{Z}'\|_F \geq \max\left\{\|\boldsymbol{Z}_1'\|_F, \|\boldsymbol{Z}_2'\|_F\right\}$. So

$$\begin{aligned}&\sup_{(\boldsymbol{Z}',\boldsymbol{z}')\in\mathcal{S}_{2r}}\|\mathcal{B}\boldsymbol{Z}' - \boldsymbol{D}\boldsymbol{z}'\|_1\\ \leq& \sup_{\boldsymbol{Z}'\in\mathcal{S}_{2r}}\|\mathcal{B}\boldsymbol{Z}_1' - \boldsymbol{D}\boldsymbol{z}'\|_1 + \sup_{\boldsymbol{Z}'\in\mathcal{S}_{2r}}\|\mathcal{B}\boldsymbol{Z}_2'\|_1 \qquad (49)\\ \leq& 2\bar{M}.\end{aligned}$$

Combining the last two displayed equations and choosing $\epsilon = \frac{1}{4}$, we obtain

$$\bar{M} \leq \bar{\delta} := \frac{2}{1-2\epsilon}\left(c_2 + \frac{1}{c_3}\right),$$

with probability at least $1 - \exp(-c_9 n)$. Note that $\bar{\delta}$ is a constant independent of p and r (but it might depend on $\mu := \mathbb{E}\left[(\boldsymbol{x}_i)_l^4\right]$).

For a possibly different ϵ', we have

$$\begin{aligned}&\inf_{(\boldsymbol{Z},\boldsymbol{z})\in\mathcal{S}_r}\|\mathcal{B}\boldsymbol{Z} - \boldsymbol{D}\boldsymbol{z}\|_1\\ \geq& \min_{(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{z}})\in\mathcal{N}_r(\epsilon)}\left\|\mathcal{B}\tilde{\boldsymbol{Z}} - \tilde{\boldsymbol{z}}\right\|_1\\ &- \sup_{(\boldsymbol{Z},\boldsymbol{z})\in\mathcal{S}_r}\left\|\mathcal{B}(\boldsymbol{Z}-\tilde{\boldsymbol{Z}}) - \boldsymbol{D}(\boldsymbol{z}-\tilde{\boldsymbol{z}})\right\|_1.\end{aligned}$$

By the union bound, we have

$$\begin{aligned}&\mathbb{P}\left(\min_{(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{z}})\in\mathcal{N}_r(\epsilon)}\left\|\mathcal{B}\tilde{\boldsymbol{Z}} - \tilde{\boldsymbol{z}}\right\|_1 \geq \left(c_7 - \frac{1}{c_3}\right)\right)\\ \geq& 1 - \exp\left(-c_4 n + 10pr\log(40/\epsilon')\right)\\ \geq& 1 - \exp(-c_4 n/2),\end{aligned}$$

provided $n \geq (2/c_4)\cdot 10pr\log(40/\epsilon')$. On this event, we have

$$\begin{aligned}\inf_{(\boldsymbol{Z},\boldsymbol{z})\in\mathcal{S}_r}\|\mathcal{B}\boldsymbol{Z} - \boldsymbol{D}\boldsymbol{z}\|_1 &\overset{(a)}{\geq} \left(c_7 - \frac{1}{c_3}\right) - 2\epsilon'\bar{M}\\ &\overset{(b)}{\geq} \left(c_7 - \frac{1}{c_3}\right) - 2\epsilon'\bar{\delta},\end{aligned}$$

where (a) follows from (49) and (b) follows from the the upper-bound on $\bar{M}$ we just established. We complete the proof by choosing ϵ' to be a sufficiently small constant such that $\underline{\delta} := \left(c_7 - \frac{1}{c_3}\right) - 2\epsilon'\bar{\delta} > 0$.

Appendix C
Proof of Lemma 8

Let $\boldsymbol{H} = \boldsymbol{K}' - \boldsymbol{g}'$ and $\boldsymbol{h} = \boldsymbol{g}' - \boldsymbol{g}$. We observe that

$$\begin{aligned}&Q_n(\boldsymbol{K}',\boldsymbol{g}';\boldsymbol{K},\boldsymbol{g})\\ =& \sum_{i=1}^n\left(-\langle\boldsymbol{x}_i\boldsymbol{x}_i,\boldsymbol{H}\rangle + 2y_i\langle\boldsymbol{x}_i,\boldsymbol{h}\rangle\right)^2 - \sum_{i=1}^n 4\sigma^2\langle\boldsymbol{x}_i,\boldsymbol{h}\rangle^2.\\ \leq& \sum_{i=1}^n 2\langle\boldsymbol{x}_i\boldsymbol{x}_i,\boldsymbol{H}\rangle^2 + 8\max_{i\in[n]}\{y_i^2\}\left(\sum_{i=1}^n\langle\boldsymbol{x}_i.\boldsymbol{h}\rangle^2\right)\end{aligned}$$

By standard concentration results, we have $\Pr\left(\sum_{i=1}^n\langle\boldsymbol{x}_i,\boldsymbol{h}\rangle^2 \geq 1.1\|\boldsymbol{h}\|_2^2\right) \leq c_1\exp(-c_2 p)$ for any $\boldsymbol{h}\in\mathbb{R}^p$. We thus obtain that w.h.p.

$$\begin{aligned}&Q_n(\boldsymbol{K}',\boldsymbol{g}';\boldsymbol{K},\boldsymbol{g})\\ \leq& \sum_{i=1}^n 2\langle\boldsymbol{x}_i\boldsymbol{x}_i,\boldsymbol{H}\rangle^2 + 9n\cdot\max_{i\in[n]}\{y_i^2\}\cdot\|\boldsymbol{h}\|_2^2.\end{aligned}$$

We need the following lemma, which is proved in Section F-H, to bound the first term on the right hand side of the above inequality.

Lemma 13. *Suppose $\{\boldsymbol{x}_i\}_{i=1}^n$ are i.i.d. p-dimensional centered sub-Gaussian random vectors with norm $\|\boldsymbol{x}_i\|_{\psi_2} \leq c$ for some constant c and identity covariance matrix $\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i^\top) = \boldsymbol{I}$. There exist constants c_i such that for any matrix $\boldsymbol{H}\in\mathbb{R}^{p\times p}$, we have*

$$\frac{1}{n}\sum_{i=1}^n\left\langle\boldsymbol{x}_i\boldsymbol{x}_i^\top,\boldsymbol{H}\right\rangle^2 \leq c_1 p\log n\cdot\|\boldsymbol{H}\|_*\cdot\|\boldsymbol{H}\|_F,$$

with probability at least $1-c_2n^{-1}$ *under condition* $n \geq c_3 p$.

Note that $\|\boldsymbol{H}\|_* \leq \sqrt{p}\|\boldsymbol{H}\|_F$ for any $\boldsymbol{H}$. We have that $\sum_{i\in[n]}\langle \boldsymbol{x}_i\boldsymbol{x}_i^\top, \boldsymbol{H}\rangle^2 \lesssim n\sqrt{p}^3 \log n \|\boldsymbol{H}\|_F^2$. Also note that

$$\Pr\left(\max_{i\in[n]}|y_i| > t\right) \leq n\exp\left(1-t^2/(\gamma^2+\sigma^2)\right).$$

Setting $t \asymp (\gamma+\sigma)\sqrt{\log n}$ in the above inequality leads to our result.

APPENDIX D
PROOF OF THEOREM 7

We need some additional notation. Let $\boldsymbol{z} := (z_1, z_2, \ldots, z_n)^\top \in \{0,1\}^n$ be the vector of hidden labels with $z_i = 1$ if and only if $i \in \mathcal{I}_1$. We use $\boldsymbol{y}(\boldsymbol{\theta}^*, \boldsymbol{X}, \boldsymbol{e}, \boldsymbol{z})$ to denote the value of the response vector $\boldsymbol{y}$ given $\boldsymbol{\theta}^*$, $\boldsymbol{X}$, $\boldsymbol{e}$ and $\boldsymbol{z}$, that is,

$$\boldsymbol{y}\left(\boldsymbol{\theta}^*, \boldsymbol{X}, \boldsymbol{e}, \boldsymbol{z}\right) = \boldsymbol{z} \circ (\boldsymbol{X}\boldsymbol{\beta}_1^*) + (\boldsymbol{1}-\boldsymbol{z}) \circ (\boldsymbol{X}\boldsymbol{\beta}_2^*) + \boldsymbol{e},$$

where $\boldsymbol{1}$ is the all-ones vector in $\mathbb{R}^n$ and $\circ$ denotes element-wise product.

By standard results, we know that with probability at least $1-n^{-10}$,

$$\|\boldsymbol{X}\boldsymbol{\alpha}\|_2 \leq 2\sqrt{n}\,\|\boldsymbol{\alpha}\|_2, \forall \boldsymbol{\alpha} \in \mathbb{R}^p. \tag{50}$$

Hence it suffices to prove (11) in the theorem statement assuming (50) holds.

Let $\boldsymbol{v}$ be an arbitrary unit vector in $\mathbb{R}^p$. We define $\delta := c_0\frac{\epsilon}{\sqrt{n}}$, $\boldsymbol{\theta}_1 := \left(\frac{1}{2}\underline{\gamma}\boldsymbol{v}, -\frac{1}{2}\underline{\gamma}\boldsymbol{v}\right)$ and $\boldsymbol{\theta}_2 = \left(\frac{1}{2}\underline{\gamma}\boldsymbol{v} + \delta\boldsymbol{v}, -\frac{1}{2}\underline{\gamma}\boldsymbol{v} - \delta\boldsymbol{v}\right)$. Note that $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta(\underline{\gamma})$ as long as c_0 is sufficiently small, and $\rho\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right) = 2\delta$. We further define $\boldsymbol{e}_1 := \boldsymbol{0}$ and $\boldsymbol{e}_2 := -\delta\left(2\boldsymbol{z}-\boldsymbol{1}\right) \circ (\boldsymbol{X}\boldsymbol{v})$. Note that $\|\boldsymbol{e}_2\| \leq 2\sqrt{n}\delta \leq \epsilon$ by (50), so $\boldsymbol{e}_1, \boldsymbol{e}_2 \in \mathbb{B}(\epsilon)$. If we set $\boldsymbol{y}_i = \boldsymbol{y}\left(\boldsymbol{\theta}_i, \boldsymbol{X}, \boldsymbol{e}_i, \boldsymbol{z}\right)$ for $i = 1, 2$, then we have

$$\begin{aligned}
\boldsymbol{y}_2 &= \boldsymbol{z} \circ \left(\boldsymbol{X}(\frac{1}{2}\underline{\gamma}\boldsymbol{v} + \delta\boldsymbol{v})\right) + (\boldsymbol{1}-\boldsymbol{z}) \circ \left(\boldsymbol{X}(-\frac{1}{2}\underline{\gamma}\boldsymbol{v} - \delta\boldsymbol{v})\right) + \boldsymbol{e}_2 \\
&= (2\boldsymbol{z}-\boldsymbol{1}) \circ \left(\boldsymbol{X}(\frac{1}{2}\underline{\gamma}\boldsymbol{v} + \delta\boldsymbol{v})\right) - \delta\left(2\boldsymbol{z}-\boldsymbol{1}\right) \circ (\boldsymbol{X}\boldsymbol{v}) \\
&= (2\boldsymbol{z}-\boldsymbol{1}) \circ \left(\boldsymbol{X}(\frac{1}{2}\underline{\gamma}\boldsymbol{v})\right) + \boldsymbol{e}_1 \\
&= \boldsymbol{y}_1,
\end{aligned}$$

which holds for any $\boldsymbol{X}$ and $\boldsymbol{z}$. Therefore, for any $\hat{\boldsymbol{\theta}}$, we have

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}^*\in\Theta(\underline{\gamma})} \sup_{\boldsymbol{e}\in\mathbb{B}(\epsilon)} \rho\left(\hat{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{y}), \boldsymbol{\theta}^*\right) \\
&\geq \frac{1}{2}\rho\left(\hat{\boldsymbol{\theta}}\left(\boldsymbol{X}, \boldsymbol{y}_1\right), \boldsymbol{\theta}_1\right) + \frac{1}{2}\rho\left(\hat{\boldsymbol{\theta}}\left(\boldsymbol{X}, \boldsymbol{y}_2\right), \boldsymbol{\theta}_2\right) \\
&= \frac{1}{2}\rho\left(\hat{\boldsymbol{\theta}}\left(\boldsymbol{X}, \boldsymbol{y}_1\right), \boldsymbol{\theta}_1\right) + \frac{1}{2}\rho\left(\hat{\boldsymbol{\theta}}\left(\boldsymbol{X}, \boldsymbol{y}_1\right), \boldsymbol{\theta}_2\right) \\
&\geq \frac{1}{2}\rho\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right) \\
&= \delta,
\end{aligned}$$

where the second inequality holds because ρ is a metric and satisfies the triangle inequality. Taking the infimum over $\hat{\boldsymbol{\theta}}$ proves the theorem.

APPENDIX E
PROOF OF THEOREM 8

Throughout the proof we set $\kappa := \frac{1}{2}\underline{\gamma}$.

A. Part 1 of the Theorem

We prove the first part of the theorem by establishing a lower-bound for standard linear regression. Set $\delta_1 := c_0\sigma\sqrt{\frac{p-1}{n}}$, and define the (semi)-metric $\rho_1(\cdot,\cdot)$ by $\rho_1(\boldsymbol{\beta}, \boldsymbol{\beta}') = \min\left\{\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|, \|\boldsymbol{\beta}+\boldsymbol{\beta}'\|\right\}$. We begin by constructing a δ_1-packing set $\Phi_1 := \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M\}$ of $\mathbb{G}^p(\kappa) := \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\| \geq \kappa\}$ in the metric ρ_1. We need a packing set of the hypercube $\{0,1\}^{p-1}$ in the Hamming distance.

Lemma 14. *For* $p \geq 16$, *there exists* $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M\} \subset \{0,1\}^{p-1}$ *such that* $M \geq 2^{(p-1)/16}$ *and*

$$\min\left\{\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|_0, \|\boldsymbol{\xi}_i + \boldsymbol{\xi}_j\|_0\right\} \geq \frac{p-1}{16}, \quad \forall 1 \leq i < j \leq M.$$

See Section F-I for the proof. Let $\tau := 2c_0\sigma\sqrt{\frac{1}{n}}$ for some absolute constant $c_0 > 0$ that is sufficiently small, and $\kappa_0^2 := \kappa^2 - (p-1)\tau^2$. Note that $\kappa_0 \geq 0$ since $\underline{\gamma} \geq \sigma$ by assumption. For $i = 1, \ldots, M$, we set

$$\boldsymbol{\beta}_i = \kappa_0\boldsymbol{\epsilon}_p + \sum_{j=1}^{p-1}\left(2\boldsymbol{\xi}_i(j) - 1\right)\tau\boldsymbol{\epsilon}_j,$$

where $\boldsymbol{\epsilon}_j$ is the j^{th} standard basis in $\mathbb{R}^p$ and $\boldsymbol{\xi}_i(j)$ is the j^{th} coordinate of $\boldsymbol{\xi}_i$. Note that $\|\boldsymbol{\beta}_i\|_2 = \kappa, \forall i \in [M]$, so $\Phi_1 = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M\} \subset \mathbb{G}^p(\kappa)$. We also have that for all $1 \leq i < j \leq M$,

$$\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2^2 \leq (p-1)\tau^2 = 4c_0^2\frac{\sigma^2(p-1)}{n}. \tag{51}$$

Moreover, we have

$$\begin{aligned}
\rho^2\left(\boldsymbol{\beta}_i, \boldsymbol{\beta}_j\right) &= \min\left\{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2^2, \|\boldsymbol{\beta}_i + \boldsymbol{\beta}_j\|_2^2\right\} \\
&\geq 4\tau^2 \min\left\{\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|_0, \|\boldsymbol{\xi}_i + \boldsymbol{\xi}_j\|_0\right\} \\
&\geq 4 \cdot 4c_0^2\frac{\sigma^2}{n} \cdot \frac{p-1}{16} = \delta_1^2,
\end{aligned} \tag{52}$$

so $\Phi_1 = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M\}$ is a δ_1-packing of $\mathbb{G}^p(\kappa)$ in the metric ρ_1.

Suppose $\boldsymbol{\beta}^*$ is sampled uniformly at random from the set Φ_1. For $i = 1, \ldots, M$, let $\mathbb{P}_{i,\boldsymbol{X}}$ denote the distribution of $\boldsymbol{y}$ conditioned on $\boldsymbol{\beta}^* = \boldsymbol{\beta}_i$ and $\boldsymbol{X}$, and $\mathbb{P}_i$ denote the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{y}$ conditioned on $\boldsymbol{\beta}^* = \boldsymbol{\beta}_i$. Because $\boldsymbol{X}$ is independent of $\boldsymbol{z}$,$\boldsymbol{e}$ and $\boldsymbol{\beta}^*$, we have

$$\begin{aligned}
D\left(\mathbb{P}_i \| \mathbb{P}_{i'}\right) &= \mathbb{E}_{\mathbb{P}_i(\boldsymbol{X},\boldsymbol{y})} \log\frac{p_i(\boldsymbol{X}, \boldsymbol{y})}{p_{i'}(\boldsymbol{X}, \boldsymbol{y})} \\
&= \mathbb{E}_{\mathbb{P}_i(\boldsymbol{X},\boldsymbol{y})} \log\frac{p_i(\boldsymbol{y}|\boldsymbol{X})}{p_{i'}(\boldsymbol{y}|\boldsymbol{X})} \\
&= \mathbb{E}_{\mathbb{P}(\boldsymbol{X})}\left[\mathbb{E}_{\mathbb{P}_i(\boldsymbol{y}|\boldsymbol{X})}\left[\log\frac{p_i(\boldsymbol{y}|\boldsymbol{X})}{p_{i'}(\boldsymbol{y}|\boldsymbol{X})}\right]\right] \\
&= \mathbb{E}_{\boldsymbol{X}}\left[D\left(\mathbb{P}_{i,\boldsymbol{X}} \| \mathbb{P}_{i',\boldsymbol{X}}\right)\right].
\end{aligned}$$

Using the above equality and the convexity of the mutual information, we get that

$$\begin{aligned} I\left(\boldsymbol{\beta}^*; \boldsymbol{X}, \boldsymbol{y}\right) \leq & \frac{1}{M^2} \sum_{1 \leq i, i' \leq M} D\left(\mathbb{P}_i \| \mathbb{P}_{i'}\right) \\ = & \frac{1}{M^2} \sum_{1 \leq i, i' \leq M} \mathbb{E}_{\boldsymbol{X}}\left[D\left(\mathbb{P}_{i, \boldsymbol{X}} \| \mathbb{P}_{i', \boldsymbol{X}}\right)\right] \\ = & \frac{1}{M^2} \sum_{1 \leq i, i' \leq M} \mathbb{E}_{\boldsymbol{X}} \frac{\left\|\boldsymbol{X} \boldsymbol{\beta}_i - \boldsymbol{X} \boldsymbol{\beta}_{i'}\right\|^2}{2 \sigma^2} \\ = & \frac{1}{M^2} \sum_{1 \leq i, i' \leq M} \frac{n\left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}\right\|^2}{2 \sigma^2}. \end{aligned}$$

It follows from (51) that

$$I\left(\boldsymbol{\beta}^*; \boldsymbol{X}, \boldsymbol{y}\right) \leq 8 c_0^2 p \leq \frac{1}{2}\left(\log_2 M\right) / \left(\log_2 e\right) = \frac{1}{4} \log M,$$

provided c_0 is sufficiently small. Following a standard argument [5, 35, 38] to transform the estimation problem into a hypothesis testing problem (cf. Eq. (46) and (47)), we obtain

$$\begin{aligned} & \inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}^* \in \mathbb{G}^p(\kappa)} \mathbb{E}_{\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{e}}\left[\rho_1\left(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*\right)\right] \\ & \geq \delta_1\left(1 - \frac{I\left(\boldsymbol{\beta}^*; \boldsymbol{X}, \boldsymbol{y}\right) + \log 2}{\log M}\right) \\ & \geq \frac{1}{2} \delta_1 = \frac{1}{2} c_0 \sigma \sqrt{\frac{p}{n}}. \end{aligned}$$

This establishes a minimax lower bound for standard linear regression. Now observe that given any standard linear regression problem with regression vector $\boldsymbol{\beta}^* \in \mathbb{G}^p(\kappa)$, we can reduce it to a mixed regression problem with $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^*, -\boldsymbol{\beta}^*) \in \Theta(\underline{\gamma})$ by multiplying each y_i by a Rademacher ± 1 variable. Part 1 of the theorem hence follows.

B. Part 2 of the Theorem

Let $\delta_2 := 2 c_0 \frac{\sigma^2}{\kappa} \sqrt{\frac{p-1}{n}}$. We first construct a δ_2-packing set $\Theta_2 := \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ of $\Theta(\underline{\gamma})$ in the metric $\rho(\cdot, \cdot)$. Set $\tau := 2 c_0 \frac{\sigma^2}{\kappa} \sqrt{\frac{1}{n}}$ and $\kappa_0^2 := \kappa^2 - (p-1)\tau^2$. Note that $\kappa_0 \geq 0$ under the assumption $\kappa \geq c_1 \sigma \left(\frac{p}{n}\right)^{1/4}$ provided that c_0 is small enough. For $i = 1, \ldots, M$, we set $\boldsymbol{\theta}_i := (\boldsymbol{\beta}_i, -\boldsymbol{\beta}_i)$ with

$$\boldsymbol{\beta}_i = \kappa_0 \boldsymbol{\epsilon}_p + \sum_{j=1}^{p-1} \left(2 \boldsymbol{\xi}_i(j) - 1\right) \tau \boldsymbol{\epsilon}_j,$$

where $\{\boldsymbol{\xi}_i\}$ are the vectors in Lemma 14. Note that $\|\boldsymbol{\beta}_i\| = \kappa$ for all i, so $\Theta_2 = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M\} \subset \Theta(\underline{\gamma})$. We also have that for all $1 \leq i < i' \leq M$,

$$\left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}\right\|^2 \leq p \tau^2 = 4 c_0^2 \frac{\sigma^4 p}{\kappa^2 n}. \tag{53}$$

Moreover, we have

$$\begin{aligned} \rho^2\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i'}\right) & = 4 \min\left\{\left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}\right\|^2, \left\|\boldsymbol{\beta}_i + \boldsymbol{\beta}_{i'}\right\|^2\right\} \\ & \geq 16 \tau^2 \min\left\{\left\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_{i'}\right\|_0, \left\|\boldsymbol{\xi}_i + \boldsymbol{\xi}_{i'}\right\|_0\right\} \\ & \geq 16 \cdot 4 c_0^2 \frac{\sigma^4}{\kappa^2 n} \cdot \frac{p-1}{16} = \delta_2^2, \end{aligned} \tag{54}$$

so $\Theta_2 = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ forms a δ_2-packing of $\Theta(\underline{\gamma})$ in the metric ρ.

Suppose $\boldsymbol{\theta}^*$ is sampled uniformly at random from the set Θ_2. For $i = 1, \ldots, M$, let $\mathbb{P}_{i, \boldsymbol{X}}^{(j)}$ denote the distribution of $\boldsymbol{y}_j$ conditioned on $\boldsymbol{\theta}^* = \boldsymbol{\theta}_i$ and $\boldsymbol{X}$, $\mathbb{P}_{i, \boldsymbol{X}}$ denote the distribution of $\boldsymbol{y}$ conditioned on $\boldsymbol{\theta}^* = \boldsymbol{\theta}_i$ and $\boldsymbol{X}$, and $\mathbb{P}_i$ denote the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{y}$ conditioned on $\boldsymbol{\theta}^* = \boldsymbol{\theta}_i$. We need the following bound on the KL divergence between two mixtures of univariate Gaussians. For any $a > 0$, we use $\mathbb{Q}_a$ to denote the distribution of the equal-weighted mixture of two Gaussian distributions $\mathcal{N}(a, \sigma^2)$ and $\mathcal{N}(-a, \sigma^2)$.

Lemma 15. *The following bound holds for any $u, v \geq 0$:*

$$\begin{aligned} & D\left(\mathbb{Q}_u \| \mathbb{Q}_v\right) \\ & \leq \frac{u^2 - v^2}{2 \sigma^4} u^2 + \frac{v^3 \max\{0, v-u\}}{2 \sigma^8}\left(u^4 + 6 u^2 \sigma^2 + 3 \sigma^4\right). \end{aligned}$$

See Section F-J for the proof. Note that $\mathbb{P}_{i, \boldsymbol{X}}^{(j)} = \mathbb{Q}_{\left|\boldsymbol{x}_j^\top \boldsymbol{\beta}_i\right|}$. Using $\mathbb{P}_{i, \boldsymbol{X}} = \otimes_{j=1}^n \mathbb{P}_{i, \boldsymbol{X}}^{(j)}$ and the above lemma, we have

$$\begin{aligned} & \mathbb{E}_{\boldsymbol{X}} D\left(\mathbb{P}_{i, \boldsymbol{X}} \| \mathbb{P}_{i', \boldsymbol{X}}\right) \\ = & \sum_{j=1}^n \mathbb{E}_{\boldsymbol{X}} D\left(\mathbb{P}_{i, \boldsymbol{X}}^{(j)} \| \mathbb{P}_{i', \boldsymbol{X}}^{(j)}\right) \\ \leq & n \mathbb{E} \frac{\left|\boldsymbol{x}_1^\top \boldsymbol{\beta}_i\right|^2 - \left|\boldsymbol{x}_1^\top \boldsymbol{\beta}_{i'}\right|^2}{2 \sigma^4}\left|\boldsymbol{x}_j^\top \boldsymbol{\beta}_i\right|^2 \\ & + n \mathbb{E}_{\boldsymbol{X}} \frac{\left|\boldsymbol{x}_1^\top \boldsymbol{\beta}_{i'}\right|^3 \max\left\{0, \left|\boldsymbol{x}_1^\top \boldsymbol{\beta}_{i'}\right| - \left|\boldsymbol{x}_1^\top \boldsymbol{\beta}_i\right|\right\}}{2 \sigma^8} \\ & \times\left(\left|\boldsymbol{x}_1^\top \boldsymbol{\beta}_i\right|^4 + 6\left|\boldsymbol{x}_1^\top \boldsymbol{\beta}_i\right|^2 \sigma^2 + 3 \sigma^4\right). \end{aligned}$$

To bound the expectations in the last RHS, we need a simple technical lemma proved in Section F-K.

Lemma 16. *Suppose $\boldsymbol{x} \in \mathbb{R}^p$ has i.i.d. standard Gaussian components, and $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^p$ are any fixed vectors with $\|\boldsymbol{\alpha}\|_2 = \|\boldsymbol{\beta}\|_2$. There exists an absolute constant $\bar{c}$ such that for any non-negative integers k, l with $k + l \leq 8$,*

$$\mathbb{E}\left|\boldsymbol{x}^\top \boldsymbol{\alpha}\right|^k\left|\boldsymbol{x}^\top \boldsymbol{\beta}\right|^l \leq \bar{c}\|\boldsymbol{\alpha}\|^k\|\boldsymbol{\beta}\|^l.$$

Moreover, we have

$$\mathbb{E}_{\boldsymbol{X}}\left[\left(\left|\boldsymbol{x}^\top \boldsymbol{\alpha}\right|^2 - \left|\boldsymbol{x}^\top \boldsymbol{\beta}\right|^2\right)\left|\boldsymbol{x}^\top \boldsymbol{\alpha}\right|^2\right] \leq 2\|\boldsymbol{\alpha}\|\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|^2.$$

$$\mathbb{E}\left(\left|\boldsymbol{x}^\top \boldsymbol{\alpha}\right|^2 - \left|\boldsymbol{x}^\top \boldsymbol{\beta}\right|^2\right)^2 \leq \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|^4.$$

Using the above lemma and the fact that $\|\boldsymbol{\beta}_i\|_2 = \|\boldsymbol{\beta}_{i'}\|_2 = \kappa$ for all $1 \leq i < i' \leq M$, we have

$$\mathbb{E}_{\boldsymbol{X}} \frac{\left|\boldsymbol{x}_1^\top \boldsymbol{\beta}_i\right|^2 - \left|\boldsymbol{x}_1^\top \boldsymbol{\beta}_{i'}\right|^2}{2 \sigma^4}\left|\boldsymbol{x}_1^\top \boldsymbol{\beta}_i\right|^2 \leq \frac{1}{2 \sigma^4} \kappa^2\left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}\right\|^2,$$

and for some universal constant $c'>0$,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}\frac{\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_{i'}\right|^3\max\left\{0,\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_{i'}\right|-\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_i\right|\right\}}{2\sigma^8}\\
&\quad\times\left(\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_i\right|^4+6\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_i\right|^2\sigma^2+3\sigma^4\right)\\
\le&\frac{1}{2\sigma^8}\mathbb{E}_{\boldsymbol{X}}\max\left\{0,\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_{i'}\right|^2-\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_i\right|^2\right\}\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_{i'}\right|^2\\
&\quad\times\left(\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_i\right|^4+6\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_i\right|^2\sigma^2+3\sigma^4\right)\\
\overset{(a)}{\le}&\frac{1}{2\sigma^4}\sqrt{\mathbb{E}_{\boldsymbol{X}}\left(\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_{i'}\right|^2-\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_i\right|^2\right)^2\cdot\frac{1}{\sigma^8}\mathbb{E}_{\boldsymbol{X}}\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_{i'}\right|^4}\\
&\quad\times\sqrt{\left(\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_i\right|^4+6\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_i\right|^2\sigma^2+3\sigma^4\right)^2}\\
\overset{(b)}{\le}&\frac{1}{2\sigma^4}\sqrt{\left\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_i\right\|^4\cdot c'^2\left\|\boldsymbol{\beta}_{i'}\right\|^4}=\frac{c'}{2\sigma^4}\left\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_i\right\|^2\kappa^2,
\end{aligned}$$

where (a) follows from the Cauchy-Schwarz inequality, and (b) follows from the first and third inequalities in Lemma 16 as well as $\|\boldsymbol{\beta}_i\|=\|\boldsymbol{\beta}_{i'}\|=\kappa\le\sigma$. It follows that

$$\mathbb{E}_{\boldsymbol{X}}D\left(\mathbb{P}_{i,\boldsymbol{X}}\|\mathbb{P}_{i',\boldsymbol{X}}\right)\le n\cdot\frac{c'\left\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i'}\right\|^2\kappa^2}{\sigma^4}\le c''p,$$

where the last inequality follows from (53) and c'' can be made sufficiently small by choosing c_0 small enough. We therefore obtain

$$\begin{aligned}
I\left(\boldsymbol{\theta}^*;\boldsymbol{X},\boldsymbol{y}\right)\le&\frac{1}{M^2}\sum_{1\le i,i'\le M}D\left(\mathbb{P}_i\|\mathbb{P}_{i'}\right)\\
=&\frac{1}{M}\sum_{1\le i,i'\le M}\mathbb{E}_{\boldsymbol{X}}\left[D\left(\mathbb{P}_{i,\boldsymbol{X}}\|\mathbb{P}_{i',\boldsymbol{X}}\right)\right]\le c''p\le\frac{1}{4}\log M
\end{aligned}$$

using $M\ge 2^{(p-1)/16}$. Following a standard argument [38, 35, 5] to transform the estimation problem into a hypothesis testing problem (cf. Eq. (46) and (47)), we obtain

$$\begin{aligned}
&\inf_{\hat{\boldsymbol{\theta}}}\sup_{\boldsymbol{\theta}^*\in\Theta(\underline{\gamma})}\mathbb{E}_{\boldsymbol{X},\boldsymbol{z},\boldsymbol{e}}\left[\rho\left(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}^*\right)\right]\\
&\ge\delta_2\left(1-\frac{I\left(\boldsymbol{\theta}^*;\boldsymbol{X},\boldsymbol{y}\right)+\log 2}{\log M}\right)\\
&\ge\frac{1}{2}\delta_2=c_0\frac{\sigma^2}{\kappa}\sqrt{\frac{p}{n}}.
\end{aligned}$$

C. Part 3 of the Theorem

The proof follows similar lines as Part 2. Let $\delta_3:=2c_0\sigma\left(\frac{p}{n}\right)^{1/4}$. Again we first construct a δ_3-packing set $\Theta_3:=(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M)$ of $\Theta(\underline{\gamma})$ in the metric $\rho(\cdot,\cdot)$. Set $\tau:=\frac{2c_0\sigma}{\sqrt{p-1}}\left(\frac{p}{n}\right)^{1/4}$. For $i=1,\ldots,M$, we set $\boldsymbol{\theta}_i=(\boldsymbol{\beta}_i,-\boldsymbol{\beta}_i)$ with

$$\boldsymbol{\beta}_i=\sum_{j=1}^{p-1}\left(2\boldsymbol{\xi}_i(j)-1\right)\tau\boldsymbol{\epsilon}_j,$$

where $\{\boldsymbol{\xi}_i\}$ are the vectors from Lemma 14. Note that $\|\boldsymbol{\beta}_i\|_2=\sqrt{p-1}\tau=2c_0\sigma\left(\frac{p}{n}\right)^{1/4}\ge c_1\sigma\left(\frac{p}{n}\right)^{1/4}\ge\kappa$ provided c_1 is sufficiently small, so $\Theta_3=\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M\}\subset\Theta(\underline{\gamma})$. We also have for all $1\le i<i'\le M$,

$$\begin{aligned}
\rho^2\left(\boldsymbol{\beta}_i,\boldsymbol{\beta}_{i'}\right)=&4\min\left\{\left\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i'}\right\|_2^2\left\|\boldsymbol{\beta}_i+\boldsymbol{\beta}_{i'}\right\|_2^2\right\}\\
\ge&16\tau^2\min\left\{\left\|\boldsymbol{\xi}_i-\boldsymbol{\xi}_{i'}\right\|_0,\left\|\boldsymbol{\xi}_i+\boldsymbol{\xi}_{i'}\right\|_0\right\}\\
=&16\cdot\frac{4c_0^2\sigma^2}{p-1}\sqrt{\frac{p}{n}}\cdot\frac{p-1}{16}\ge\delta_3^2,
\end{aligned}\tag{55}$$

so $\Theta_3=\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M\}$ is a δ_3-packing of $\Theta(\underline{\gamma})$ in the metric ρ.

Suppose $\boldsymbol{\theta}^*$ is sampled uniformly at random from the set Θ_2. Define $\mathbb{P}_{i,\boldsymbol{X}},\mathbb{P}_{i,\boldsymbol{X}}^{(j)}$ and $\mathbb{P}_i$ as in the proof of Part 2 of the theorem. We have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}D\left(\mathbb{P}_{i,\boldsymbol{X}}|\mathbb{P}_{i',\boldsymbol{X}}\right)\\
=&\sum_{j=1}^n\mathbb{E}_{\boldsymbol{X}}D\left(\mathbb{P}_{i,\boldsymbol{X}}^{(j)}\|\mathbb{P}_{i',\boldsymbol{X}}^{(j)}\right)\\
\overset{(a)}{\le}&n\mathbb{E}_{\boldsymbol{X}}\frac{\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_i\right|^2-\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_{i'}\right|^2}{2\sigma^4}\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_i\right|^2\\
&+n\mathbb{E}_{\boldsymbol{X}}\frac{\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_{i'}\right|^3\max\left\{0,\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_{i'}\right|-\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_i\right|\right\}}{2\sigma^8}\\
&\times\left(\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_i\right|^4+6\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_i\right|^2\sigma^2+3\sigma^4\right)\\
\le&\frac{n}{2\sigma^4}\mathbb{E}_{\boldsymbol{X}}\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_i\right|^4+\frac{n}{2\sigma^8}\mathbb{E}_{\boldsymbol{X}}\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_{i'}\right|^4\\
&\times\left(\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_i\right|^4+6\left|\boldsymbol{x}_1^\top\boldsymbol{\beta}_i\right|^2\sigma^2+3\sigma^4\right)\\
\overset{(b)}{\le}&\frac{n}{2\sigma^4}\bar{c}\left\|\boldsymbol{\beta}_i\right\|^4+\frac{n}{2\sigma^8}\bar{c}\left\|\boldsymbol{\beta}_{i'}\right\|^4\left(\left\|\boldsymbol{\beta}_i\right\|^4+6\sigma^2\left\|\boldsymbol{\beta}_i\right\|^2+9\sigma^4\right)\\
\overset{(c)}{\le}&c'p,
\end{aligned}$$

where (a) follows from Lemma 15, (b) follows from Lemma 16, (c) follows from $\|\boldsymbol{\beta}_i\|=2c_0\sigma\left(\frac{p}{n}\right)^{1/4}\le\sigma,\forall i$, and c' is a sufficiently small absolute constant. It follows that

$$I\left(\boldsymbol{\theta}^*;\boldsymbol{X},\boldsymbol{y}\right)\le\frac{1}{M}\sum_{1\le i,i'\le M}\mathbb{E}_{\boldsymbol{X}}D\left(\mathbb{P}_i\|\mathbb{P}_{i'}\right)\le c'p\le\frac{1}{4}\log M,$$

since $M\ge 2^{(p-1)/8}$. Again transforming the estimation problem into a hypothesis testing problem, we obtain

$$\begin{aligned}
&\inf_{\hat{\boldsymbol{\theta}}}\sup_{\boldsymbol{\theta}^*\in\Theta(\underline{\gamma})}\mathbb{E}_{\boldsymbol{X},\boldsymbol{z},\boldsymbol{e}}\left[\rho\left(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}^*\right)\right]\\
&\ge\delta_3\left(1-\frac{I\left(\boldsymbol{\theta}^*;\boldsymbol{X},\boldsymbol{y}\right)+\log 2}{\log M}\right)\\
&\ge\frac{1}{2}\delta_3=c_0\sigma\left(\frac{p}{n}\right)^{1/4}.
\end{aligned}$$

APPENDIX F
PROOFS OF AUXILIARY RESULTS

A. Proof of Lemma 2

Simple algebra shows that

$$\begin{aligned}
&\sum_b\left\|\hat{\boldsymbol{H}}_T-2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top\right\|_F^2\\
=&2\left\|\hat{\boldsymbol{H}}_T-(\boldsymbol{\beta}_1^*+\boldsymbol{\beta}_2^*)\hat{\boldsymbol{h}}^\top\right\|_F^2+2\left\|\boldsymbol{\beta}_1^*-\boldsymbol{\beta}_2^*\right\|_2^2\left\|\hat{\boldsymbol{h}}\right\|_2^2\\
\ge&2\left\|\boldsymbol{\beta}_1^*-\boldsymbol{\beta}_2^*\right\|_2^2\left\|\hat{\boldsymbol{h}}\right\|_2^2\ge\alpha\left(\left\|\boldsymbol{\beta}_1^*\right\|_2+\left\|\boldsymbol{\beta}_2^*\right\|\right)^2\left\|\hat{\boldsymbol{h}}\right\|_2^2,
\end{aligned}$$

and

$$\begin{aligned}
&\sum_b \left\|\hat{\boldsymbol{H}}_T - 2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}\right\|_F^2 \\
=&4\left(\|\boldsymbol{\beta}_1^*\|_2^2 + \|\boldsymbol{\beta}_2^*\|^2\right)\left\|\hat{\boldsymbol{h}} - \frac{\hat{\boldsymbol{H}}_T(\boldsymbol{\beta}_1^* + \boldsymbol{\beta}_2^*)}{2\|\boldsymbol{\beta}_1^*\|_2^2 + 2\|\boldsymbol{\beta}_2^*\|^2}\right\|_2^2 \\
&+ \frac{2\left(\|\boldsymbol{\beta}_1^*\|_2^2 + \|\boldsymbol{\beta}_2^*\|^2\right)\left\|\hat{\boldsymbol{H}}_T\right\|_F^2 - \left\|\hat{\boldsymbol{H}}_T(\boldsymbol{\beta}_1^* + \boldsymbol{\beta}_2^*)\right\|_2^2}{\|\boldsymbol{\beta}_1^*\|_2^2 + \|\boldsymbol{\beta}_2^*\|^2} \\
\overset{(a)}{\geq}& \frac{2\left(\|\boldsymbol{\beta}_1^*\|_2^2 + \|\boldsymbol{\beta}_2^*\|^2\right)\left\|\hat{\boldsymbol{H}}_T\right\|_F^2 - \left\|\hat{\boldsymbol{H}}_T\right\|_F^2\|\boldsymbol{\beta}_1^* + \boldsymbol{\beta}_2^*\|_2^2}{\|\boldsymbol{\beta}_1^*\|_2^2 + \|\boldsymbol{\beta}_2^*\|^2} \\
=&\alpha\left\|\hat{\boldsymbol{H}}_T\right\|_F^2,
\end{aligned}$$

where the inequality (a) follows from $\left\|\hat{\boldsymbol{H}}_T\right\| \leq \left\|\hat{\boldsymbol{H}}_T\right\|_F$. Combining the last two displayed equations with the simple inequality

$$\sum_b \left\|\hat{\boldsymbol{H}}_T - 2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}\right\|_F \geq \sqrt{\sum_b \left\|\hat{\boldsymbol{H}}_T - 2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}\right\|_F^2},$$

we obtain

$$\begin{aligned}
\sum_b \left\|\hat{\boldsymbol{H}}_T - 2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}\right\|_F &\geq \sqrt{\alpha}\left(\|\boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^*\|_2\right)\|\hat{\boldsymbol{h}}\|_2, \\
\sum_b \left\|\hat{\boldsymbol{H}}_T - 2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}\right\|_F &\geq \sqrt{\alpha}\left\|\hat{\boldsymbol{H}}_T\right\|_F.
\end{aligned}$$

B. Proof of Lemma 3

Without loss of generality, we may assume $\sigma = 1$. Set $L := \sqrt{c\log n}$ for some c sufficiently large. For each $i \in [n]$, we define the event $\mathcal{E}_i = \{|e_i| \leq L\}$ and the truncated random variables

$$\bar{e}_i = e_i\mathbf{1}\left(\mathcal{E}_i\right),$$

where $\mathbf{1}(\cdot)$ is the indicator function and c is some sufficiently large numeric constant. Let $m_i := \mathbb{E}\left[e_i\mathbf{1}\left(\mathcal{E}_i^c\right)\right]$ and $s_i := \sqrt{\mathbb{E}\left[e_i^2\mathbf{1}\left(\mathcal{E}_i^c\right)\right]}$. WLOG we assume $m_i \geq 0$. Note that the following equation holds almost surely:

$$e_i^2\mathbf{1}\left(\mathcal{E}_i^c\right) = |e_i|\cdot|e_i|\,\mathbf{1}\left(\mathcal{E}_i^c\right) \geq L\cdot|e_i|\,\mathbf{1}\left(\mathcal{E}_i^c\right) \geq L\cdot e_i\mathbf{1}\left(\mathcal{E}_i^c\right).$$

Taking the expectation of both sides gives $s_i^2 \geq Lm_i$. We further define

$$\tilde{e}_i := \bar{e}_i + L\epsilon_i^+ - L\epsilon_i^-,$$

where ϵ_i^+ and ϵ_i^- are independent random variables distributed as $\mathrm{Ber}(\nu_i^+)$ and $\mathrm{Ber}(\nu_i^-)$, respectively, with

$$\nu_i^+ := \frac{1}{2}\left(\frac{m_i}{L} + \frac{s_i^2}{L^2}\right), \quad \nu_i^- := \frac{1}{2}\left(-\frac{m_i}{L} + \frac{s_i^2}{L^2}\right).$$

Note that $m_i \geq 0$ and $s_i^2 \geq Lm_i$ implies that $\nu_i^+, \nu_i^- \geq 0$. We show below that $\nu_i^+, \nu_i^- \leq 1$ so the random variables ϵ_i^+ and ϵ_i^- are well-defined.

With this setup, we now characterize the distribution of $\tilde{e}_i$. Note that

$$\begin{aligned}
\mathbb{E}\left[L\epsilon_i^+ - L\epsilon_i^-\right] &= m_i, \\
\mathbb{E}\left[(L\epsilon_i^+)^2 + (L\epsilon_i^-)^2\right] &= s_i^2,
\end{aligned}$$

which means

$$\begin{aligned}
\mathbb{E}\left[\tilde{e}_i\right] &= \mathbb{E}\left[\bar{e}_i\right] + \mathbb{E}\left[e_i\mathbf{1}\left(\mathcal{E}_i^c\right)\right] = \mathbb{E}\left[e_i\right] = 0. \\
Var\left[\tilde{e}_i^2\right] &= \mathbb{E}\left[\bar{e}_i^2\right] + \mathbb{E}\left[e_i^2\mathbf{1}\left(\mathcal{E}_i^c\right)\right] = \mathbb{E}\left[e_i^2\right] = 1.
\end{aligned}$$

Moreover, $\tilde{e}_i$ is bounded by $3L$ almost surely, which means it is sub-Gaussian with sub-Gaussian norm at most $3L$. Also note that

$$\begin{aligned}
m_i &\leq \mathbb{E}\left[|e_i\mathbf{1}\left(\mathcal{E}_i^c\right)|\right] \\
&= \int_0^\infty \mathbb{P}\left(|e_i\mathbf{1}\left(\mathcal{E}_i^c\right)| \geq t\right)dt \\
&= L\cdot\mathbb{P}(|e_i| \geq L) + \int_L^\infty \mathbb{P}\left(|e_i| \geq t\right)dt \\
&\leq \sqrt{c\log n}\frac{1}{n^{c_1}} + \int_L^\infty e^{1-t^2}dt \leq \frac{4}{n^{c_2}}
\end{aligned}$$

for some large constant c_1 and c_2 by sub-Gaussianity of e_i. A similar calculation gives

$$s_i^2 = \mathbb{E}\left[e_i^2\mathbf{1}\left(\mathcal{E}^c\right)\right] \lesssim \frac{1}{n^{c_2}}.$$

This implies $\nu_i^+, \nu_i^- \lesssim \frac{1}{n^{c_2}}$, or equivalently $L\epsilon_i^+ - L\epsilon_i^- = 0$ w.h.p. We also have $\bar{e}_i = e_i$ w.h.p. by sub-Gaussianity of e_i. It follows that $\tilde{e}_i = \bar{e}_i + L\epsilon_i^+ - L\epsilon_i^- = e_i$ w.h.p. Moreover, $\tilde{e}_i$ and e_i have the same mean and variance.

We define the variables $\{(\tilde{\boldsymbol{x}}_i)_l, i \in [n], l \in [p]\}$ in a similar manner. Each $(\tilde{\boldsymbol{x}}_i)_l$ is sub-Gaussian, bounded by L a.s., has mean 0 and variance 1, and equals $(\boldsymbol{x}_i)_l$ w.h.p.

Now suppose the conclusion of Theorem 3 holds w.h.p. for the program (7) with $\{(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)\}$ as the input, where $\tilde{y}_i = \tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}_b^* + \tilde{e}_i$ for all $i \in \mathcal{I}_b$ and $b = 1, 2$. We know that $\boldsymbol{e} = \tilde{\boldsymbol{e}}$ and $\boldsymbol{x}_i = \tilde{\boldsymbol{x}}_i, \forall i$ with high probability. On this event, the program above is identical to the original program with $\{(\boldsymbol{x}_i, y_i)\}$ as the input. Therefore, the conclusion of the theorem also holds w.h.p. for the original program.

C. Proof of Lemma 6

We need to bound

$$\begin{aligned}
W =&\sqrt{p}\left\|\sum_b \boldsymbol{X}_b^\top\mathrm{diag}(\boldsymbol{e}_b)\boldsymbol{w}_b + \sum_b \sigma^2\boldsymbol{X}_b^\top\boldsymbol{X}_b\boldsymbol{\delta}_b^*\right. \\
&\left.+ \sum_b 2\sigma^2\boldsymbol{X}_b^\top\boldsymbol{e}_b\right\|_\infty \\
\leq&\sqrt{p}\sum_b\left\|\sum_{i\in\mathcal{I}_b}(\sigma^2 - e_i^2)\langle\boldsymbol{x}_i, \boldsymbol{\delta}_b^*\rangle\boldsymbol{x}_i + \sum_{i\in\mathcal{I}_b}(e_i^2 + \sigma^2)e_i\boldsymbol{x}_i\right\|_\infty \\
\leq&\underbrace{\sqrt{p}\sum_b\left\|\sum_{i\in\mathcal{I}_b}(\sigma^2 - e_i^2)\langle\boldsymbol{x}_i, \boldsymbol{\delta}_b^*\rangle\boldsymbol{x}_i\right\|_\infty}_{S_9} \\
&+ \underbrace{\sqrt{p}\sum_b\left\|\sum_{i\in\mathcal{I}_b}(e_i^2 + \sigma^2)e_i\boldsymbol{x}_i\right\|_\infty}_{S_{10}}.
\end{aligned}$$

Lemma 3 makes it possible to assume the boundedness of e_i with the loss of only small probability, we thus have

that $(e_i^2+\sigma^2)e_i\boldsymbol{x}_i$ is sub-Gaussian random vector with Orlicz norm $O(\sigma^3\sqrt{\log n}^3)$. Therefore the standard sub-Gaussian concentration result leads to that the following inequality holds with high probability

$$S_{10} \lesssim \sigma^3\sqrt{n_1}\log^2 n_1 + \sigma^3\sqrt{n_2}\log^2 n_2 \lesssim \sigma^3\sqrt{n}\log^2 n,$$

where the last inequality follows from $\min\{n_1/n_2, n_2/n_1\} = \Omega(1)$.

Since $\langle \boldsymbol{x}_i, \boldsymbol{\delta}_b^*\rangle$ is sub-Gaussian random variable with Orlicz norm $O(\gamma)$ and $(\sigma^2 - e_i^2)\boldsymbol{x}_i$ is sub-Gaussian random vector with Orlicz norm $O(\sigma^2\log n)$, their product $(\sigma^2 - e_i^2)\langle \boldsymbol{x}_i, \boldsymbol{\delta}_b^*\rangle\boldsymbol{x}_i$ is centered sub-exponential random vector with Orlicz norm $O(\gamma\sigma^2\log n)$. By standard concentration inequality of sub-exponential random variables (see, e.g., Corollary 5.17 in [30]), we obtain that for every $\epsilon > 0$ and $K = \gamma\sigma^2\log n$,

$$\begin{aligned}&\Pr\left(\left\|\sum_{i\in\mathcal{I}_b}(\sigma^2-e_i^2)\langle \boldsymbol{x}_i, \boldsymbol{\delta}_b^*\rangle\boldsymbol{x}_i + \sum_{i\in\mathcal{I}_b}(e_i^2+\sigma^2)e_i\boldsymbol{x}_i\right\|_\infty > \epsilon n_b\right)\\ &\le 2p\exp\left(-c\min\left\{\frac{\epsilon^2}{K^2}, \frac{\epsilon}{K}\right\}n_b\right).\end{aligned}$$

For each $b = 1, 2$, by setting $\epsilon = c_1 K\sqrt{\log p/n_b}$, we have that when $n_b \gtrsim \log p$, with probability at least $1 - 1/p$,

$$S_9 \lesssim \gamma\sigma^2(\sqrt{n_1}+\sqrt{n_2})\sqrt{\log p}\log n \lesssim \gamma\sigma^2\sqrt{n}\log^2 n.$$

Putting all ingredients together, we have that $W \lesssim \gamma\sigma^2\sqrt{pn}\log^2 n$ with high probability.

D. Proof of Lemma 7

Using the inequalities $\left\|\boldsymbol{K}^{(t)}\right\|_* \le \|\boldsymbol{K}^*\|_*$ and $\left\|\hat{\boldsymbol{K}}\right\|_* \le \|\boldsymbol{K}^*\|_*$, we have

$$\begin{aligned}&\left\|\left(\boldsymbol{K}^{(t)} - \boldsymbol{K}^*\right)_{T^\perp}\right\|_* \le \left\|\left(\boldsymbol{K}^{(t)} - \boldsymbol{K}^*\right)_{T}\right\|_*, \quad \text{and}\\ &\left\|\hat{\boldsymbol{H}}_{T^\perp}\right\|_* \le \left\|\hat{\boldsymbol{H}}_T\right\|_*.\end{aligned} \tag{56}$$

For the term $\boldsymbol{H}^{(t)}$ we have

$$\begin{aligned}\left\|\boldsymbol{H}^{(t)}\right\|_* &\le \left\|\boldsymbol{K}^{(t)} - \boldsymbol{K}^*\right\|_* + \left\|\hat{\boldsymbol{H}}\right\|_*\\ &\overset{(a)}{=} \left\|\left(\boldsymbol{K}^{(t)} - \boldsymbol{K}^*\right)_{T^\perp}\right\|_* + \left\|\left(\boldsymbol{K}^{(t)} - \boldsymbol{K}^*\right)_{T}\right\|_*\\ &\quad + \left\|\hat{\boldsymbol{H}}_{T^\perp}\right\|_* + \left\|\hat{\boldsymbol{H}}_T\right\|_*\\ &\overset{(b)}{\le} 2\left\|\left(\boldsymbol{K}^{(t)} - \boldsymbol{K}^*\right)_{T}\right\|_* + 2\left\|\hat{\boldsymbol{H}}_T\right\|_*\\ &\overset{(c)}{\le} 2\left\|\boldsymbol{H}_T^{(t)}\right\|_* + 4\left\|\hat{\boldsymbol{H}}_T\right\|_*,\end{aligned}$$

where the inequality (b) follows from (56); (a) and (c) are from the triangle inequality.

Also note that $\left\|\boldsymbol{H}^{(t)}\right\|_* \ge \left\|\boldsymbol{H}_{T^\perp}^{(t)}\right\|_* - \left\|\boldsymbol{H}_T^{(t)}\right\|_*$. Putting the lower and upper bounds of $\|\boldsymbol{H}^{(t)}\|_*$ together, we complete our proof.

E. Proof of Lemma 9

Let

$$\begin{aligned}&\Psi_t(\boldsymbol{K}, \boldsymbol{g})\\ &:= \left\langle \nabla_{\boldsymbol{K}}\mathcal{L}_n^{(t)}, \boldsymbol{K}\right\rangle + \left\langle \nabla_{\boldsymbol{g}}\mathcal{L}_n^{(t)}, \boldsymbol{g}\right\rangle + \frac{\eta}{2}\left\|\boldsymbol{K} - \boldsymbol{K}^{(t)}\right\|_F^2\\ &\quad + \frac{\eta\gamma^2}{2}\left\|\boldsymbol{g} - \boldsymbol{g}^{(t)}\right\|_2^2.\end{aligned}$$

From the optimality of $\left(\boldsymbol{K}^{(t+1)}, \boldsymbol{g}^{(t)}\right)$, we have

$$\begin{aligned}&\left\langle \nabla_{\boldsymbol{K}}\Psi_t(\boldsymbol{K}^{(t+1)}, \boldsymbol{g}^{(t+1)}), \hat{\boldsymbol{K}} - \boldsymbol{K}^{(t+1)}\right\rangle\\ &+ \left\langle \nabla_{\boldsymbol{g}}\Psi_t(\boldsymbol{K}^{(t+1)}, \boldsymbol{g}^{(t+1)}), \hat{\boldsymbol{g}} - \boldsymbol{g}^{(t+1)}\right\rangle \ge 0.\end{aligned}$$

We thus have

$$\begin{aligned}&\left\langle \nabla_{\boldsymbol{K}}\mathcal{L}_n^{(t)}, \hat{\boldsymbol{K}} - \boldsymbol{K}^{(t+1)}\right\rangle + \left\langle \nabla_{\boldsymbol{g}}\mathcal{L}_n^{(t)}, \hat{\boldsymbol{g}} - \boldsymbol{g}^{(t+1)}\right\rangle\\ &\ge \eta\left\langle \boldsymbol{K}^{(t)} - \boldsymbol{K}^{(t+1)}, \hat{\boldsymbol{K}} - \boldsymbol{K}^{(t+1)}\right\rangle\\ &\quad + \eta\gamma^2\left\langle \boldsymbol{g}^{(t)} - \boldsymbol{g}^{(t+1)}, \hat{\boldsymbol{g}} - \boldsymbol{g}^{(t+1)}\right\rangle.\end{aligned} \tag{57}$$

Using the first condition (lower bound), we have

$$\begin{aligned}&\mathcal{L}_n(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})\\ &\ge \mathcal{L}_n^{(t)} + \left\langle \nabla_{\boldsymbol{K}}\mathcal{L}_n^{(t)}, \hat{\boldsymbol{K}} - \boldsymbol{K}^{(t)}\right\rangle + \left\langle \nabla_{\boldsymbol{g}}\mathcal{L}_n^{(t)}, \hat{\boldsymbol{g}} - \boldsymbol{g}^{(t)}\right\rangle\\ &\quad + \underline{Q}_n(\boldsymbol{H}^{(t)}, \boldsymbol{h}^{(t)})\\ &\overset{(a)}{\ge} \mathcal{L}_n^{(t)} + \left\langle \nabla_{\boldsymbol{K}}\mathcal{L}_n^{(t)}, \boldsymbol{K}^{(t+1)} - \boldsymbol{K}^{(t)}\right\rangle\\ &\quad + \left\langle \nabla_{\boldsymbol{g}}\mathcal{L}_n^{(t)}, \boldsymbol{g}^{(t+1)} - \boldsymbol{g}^{(t)}\right\rangle + \underline{Q}_n(\boldsymbol{H}^{(t)}, \boldsymbol{h}^{(t)})\\ &\quad + \eta\left\langle \boldsymbol{K}^{(t)} - \boldsymbol{K}^{(t+1)}, \hat{\boldsymbol{K}} - \boldsymbol{K}^{(t+1)}\right\rangle\\ &\quad + \eta\gamma^2\left\langle \boldsymbol{g}^{(t)} - \boldsymbol{g}^{(t+1)}, \hat{\boldsymbol{g}} - \boldsymbol{g}^{(t+1)}\right\rangle.\end{aligned}$$

Now applying the second condition (upper bound), we have

$$\begin{aligned}&\mathcal{L}_n(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})\\ &\ge \mathcal{L}_n^{(t+1)} - \overline{Q}_n(\boldsymbol{\Delta}^{(t)}, \boldsymbol{\delta}^{(t)}) + \underline{Q}_n(\boldsymbol{H}^{(t)}, \boldsymbol{h}^{(t)})\\ &\quad + \eta\left\langle \boldsymbol{K}^{(t)} - \boldsymbol{K}^{(t+1)}, \hat{\boldsymbol{K}} - \boldsymbol{K}^{(t+1)}\right\rangle\\ &\quad + \eta\gamma^2\left\langle \boldsymbol{g}^{(t)} - \boldsymbol{g}^{(t+1)}, \hat{\boldsymbol{g}} - \boldsymbol{g}^{(t+1)}\right\rangle.\end{aligned}$$

Applying $\mathcal{L}_n^{(t+1)} \ge \mathcal{L}_n(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ and rearranging terms yield

$$\begin{aligned}&\eta\left\langle \boldsymbol{K}^{(t)} - \boldsymbol{K}^{(t+1)}, \hat{\boldsymbol{K}} - \boldsymbol{K}^{(t)}\right\rangle + \eta\gamma^2\left\langle \boldsymbol{g}^{(t)} - \boldsymbol{g}^{(t+1)}, \hat{\boldsymbol{g}} - \boldsymbol{g}^{(t)}\right\rangle\\ &\le \overline{Q}_n(\boldsymbol{\Delta}^{(t)}, \boldsymbol{\delta}^{(t)}) - \underline{Q}_n(\boldsymbol{H}^{(t)}, \boldsymbol{h}^{(t)}) - \eta\|\boldsymbol{\Delta}^{(t)}\|_F^2\\ &\quad - \eta\gamma^2\|\boldsymbol{\delta}^{(t)}\|_2^2.\end{aligned} \tag{58}$$

Finally, by expanding $\|\boldsymbol{H}^{(t+1)}\|_F^2$ and $\|\boldsymbol{h}^{(t+1)}\|_2^2$ through

$$\begin{aligned}&\|\boldsymbol{H}^{(t+1)}\|_F^2 + \gamma^2\|\boldsymbol{h}^{(t+1)}\|_2^2\\ &\le \|\boldsymbol{H}^{(t)}\|_F^2 + \gamma^2\|\boldsymbol{h}^{(t)}\|_2^2 + 2\left\langle \boldsymbol{K}^{(t)} - \boldsymbol{K}^{(t+1)}, \hat{\boldsymbol{K}} - \boldsymbol{K}^{(t)}\right\rangle\\ &\quad + 2\gamma^2\left\langle \boldsymbol{g}^{(t)} - \boldsymbol{g}^{(t+1)}, \hat{\boldsymbol{g}} - \boldsymbol{g}^{(t)}\right\rangle + \|\boldsymbol{\Delta}^{(t)}\|_F^2 + \gamma^2\|\boldsymbol{\delta}^{(t)}\|_2^2\end{aligned}$$

and applying (58), we complete the proof.

F. Proof of Lemma 11

We need to bound

$$S_1 = 2\sum_b \left\| \sum_{i\in\mathcal{I}_b} e_{b,i}\boldsymbol{x}_{b,i}\boldsymbol{x}_{b,i}^\top \cdot \boldsymbol{x}_{b,i}^\top \left(\boldsymbol{\beta}_b^* - \boldsymbol{\beta}_{-b}^*\right)\right\|,$$

where $\boldsymbol{\beta}_b^* - \boldsymbol{\beta}_{-b}^*$ is supported on the first coordinate. Because $n_1 \asymp n_2 \asymp n$ and $\{(e_{b,i}, \boldsymbol{x}_{b,i})\}$ are identically distributed, it suffices to prove w.h.p.

$$\|\boldsymbol{E}\| := \left\| \sum_{i=1}^n e_i\boldsymbol{x}_i\boldsymbol{x}_i^\top \cdot \boldsymbol{x}_i^\top\boldsymbol{\delta}_1^* \right\| \lesssim \sigma \left\|\boldsymbol{\delta}_1^*\right\|_2 \sqrt{np}\log^3 n. \quad (59)$$

Let $\bar{\boldsymbol{x}}_i \in \mathbb{R}^1$ and $\underline{\boldsymbol{x}}_i \in \mathbb{R}^{p-1}$ be the subvectors of $\boldsymbol{x}_i$ corresponding to the first and the last $p-1$ coordinates, respectively. We define $\bar{\boldsymbol{\delta}}_1^*$ similarly; note that $\left\|\bar{\boldsymbol{\delta}}_1^*\right\| = \left\|\boldsymbol{\delta}_1^*\right\|$.

Note that $\boldsymbol{E} := \sum_i e_i\boldsymbol{x}_i\boldsymbol{x}_i^\top \cdot \bar{\boldsymbol{x}}_i^\top\bar{\boldsymbol{\delta}}_1^*$ due to the support of $\boldsymbol{\delta}_1^*$. We partition $\boldsymbol{E} \in \mathbb{R}^{p\times p}$ as

$$\boldsymbol{E} = \begin{bmatrix} \boldsymbol{E}_1 & \boldsymbol{E}_{12} \\ \boldsymbol{E}_{12}^\top & \boldsymbol{E}_2 \end{bmatrix},$$

where $\boldsymbol{E}_1 \in \mathbb{R}^{1\times 1}$, $\boldsymbol{E}_2 \in \mathbb{R}^{(p-1)\times(p-1)}$ and $\boldsymbol{E}_{12} \in \mathbb{R}^{1\times p}$. We have

$$\|\boldsymbol{E}\| \le \|\boldsymbol{E}_1\| + \|\boldsymbol{E}_2\| + 2\|\boldsymbol{E}_{12}\|.$$

We bound each term separately.

Consider $\boldsymbol{E}_1 = \sum_i e_i\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^\top \cdot \bar{\boldsymbol{x}}_i^\top\bar{\boldsymbol{\delta}}_1^*$. We condition on $\{\bar{\boldsymbol{x}}_i\}$. Note that $\|\bar{\boldsymbol{x}}_i\|_2 \lesssim \sqrt{\log n}$ and $\left|\bar{\boldsymbol{x}}_i^\top\bar{\boldsymbol{\delta}}_1^*\right| \lesssim \|\boldsymbol{\delta}_1^*\|\sqrt{\log n}$ a.s. by boundedness of $\boldsymbol{x}_i$. Since $\{e_i\}$ are independent of $\{\bar{\boldsymbol{x}}_i\}$, we have

$$\mathbb{P}\left[\|\boldsymbol{E}_1\| \lesssim \sigma\|\boldsymbol{\delta}_1^*\|\sqrt{n}\log^2 n \middle| \{\bar{\boldsymbol{x}}_i\}\right] \ge 1 - n^{-10},$$

w.h.p. using Hoeffding's inequality. Integrating over $\{\bar{\boldsymbol{x}}_i\}$ proves $\|\boldsymbol{E}_1\| \lesssim \sigma\|\boldsymbol{\delta}_1^*\|\sqrt{n}\log^2 n$, w.h.p.

Consider $\boldsymbol{E}_2 = \sum_i e_i\underline{\boldsymbol{x}}_i\underline{\boldsymbol{x}}_i^\top \cdot \bar{\boldsymbol{x}}_i^\top\bar{\boldsymbol{\delta}}_1^*$. We condition on the event $\mathcal{F} := \left\{\forall i : \left|\bar{\boldsymbol{x}}_i^\top\bar{\boldsymbol{\delta}}_1^*\right| \lesssim \|\boldsymbol{\delta}_1^*\|\sqrt{\log n}\right\}$, which occurs with high probability and is independent of e_i and $\underline{\boldsymbol{x}}_i$. We shall apply the matrix Bernstein inequality [28]; to this end, we compute:

$$\left\|e_i\underline{\boldsymbol{x}}_i\underline{\boldsymbol{x}}_i^\top \cdot \bar{\boldsymbol{x}}_i^\top\bar{\boldsymbol{\delta}}_1^*\right\| \lesssim \sigma p\|\boldsymbol{\delta}_1^*\|\log^2 n, \quad \text{a.s.}$$

by boundedness, and

$$\begin{aligned}&\left\|\sum_i \mathbb{E}e_i^2\left(\underline{\boldsymbol{x}}_i\underline{\boldsymbol{x}}_i^\top\right)^2 \cdot \left(\bar{\boldsymbol{x}}_i^\top\bar{\boldsymbol{\delta}}_1^*\right)^2\right\| \\ \le& n\sigma^2 \max_i \left|\bar{\boldsymbol{x}}_i^\top\bar{\boldsymbol{\delta}}_1^*\right|^2 \left\|\mathbb{E}\left(\underline{\boldsymbol{x}}_i\underline{\boldsymbol{x}}_i^\top\right)^2\right\| \\ \le& np\sigma^2\left\|\boldsymbol{\delta}_1^*\right\|^2\log n.\end{aligned}$$

Applying the Matrix Bernstein inequality then gives

$$\|\boldsymbol{E}_2\| \lesssim \sigma\|\boldsymbol{\delta}_1^*\|(p+\sqrt{np})\log^2 n \le \sigma\|\boldsymbol{\delta}_1^*\|\sqrt{np}\log^3 n,$$

w.h.p., where we use $n \gtrsim p$ in the last inequality.

Consider $\boldsymbol{E}_{12} = \sum_i e_i\bar{\boldsymbol{x}}_i\underline{\boldsymbol{x}}_i^\top \cdot \bar{\boldsymbol{x}}_i^\top\bar{\boldsymbol{\delta}}_1^*$. We again condition on the event $\mathcal{F}$ and use the matrix Bernstein inequality. Observe that

$$\left\|e_i\bar{\boldsymbol{x}}_i\underline{\boldsymbol{x}}_i^\top \cdot \bar{\boldsymbol{x}}_i^\top\bar{\boldsymbol{\delta}}_1^*\right\| \lesssim \sigma\sqrt{p}\|\boldsymbol{\delta}_1^*\|\log^2 n, \quad \text{a.s.}$$

by boundedness. Moreover, we have

$$\begin{aligned}&\left\|\sum_i \mathbb{E}e_i^2\left(\bar{\boldsymbol{x}}_i^\top\bar{\boldsymbol{\delta}}_b\right)^2\left(\underline{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^\top\right)\left(\bar{\boldsymbol{x}}_i\underline{\boldsymbol{x}}_i^\top\right)\right\| \\ \le& n\sigma^2\max_i\left|\bar{\boldsymbol{x}}_i^\top\bar{\boldsymbol{\delta}}_1^*\right|^2\|\bar{\boldsymbol{x}}_i\|^2\left\|\mathbb{E}\underline{\boldsymbol{x}}_i\underline{\boldsymbol{x}}_i^\top\right\| \\ \lesssim& n\sigma^2\left\|\boldsymbol{\delta}_1^*\right\|^2\log^2 n\end{aligned}$$

and

$$\begin{aligned}&\left\|\sum_i \mathbb{E}e_i^2\left(\bar{\boldsymbol{x}}_i^\top\bar{\boldsymbol{\delta}}_b\right)^2\left(\bar{\boldsymbol{x}}_i\underline{\boldsymbol{x}}_i^\top\right)\left(\underline{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^\top\right)\right\| \\ \le& n\sigma^2\max_i\left|\bar{\boldsymbol{x}}_i^\top\bar{\boldsymbol{\delta}}_1^*\right|^2\left\|\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^\top\right\|\mathbb{E}\left[\underline{\boldsymbol{x}}_i^\top\underline{\boldsymbol{x}}_i\right] \\ \lesssim& np\sigma^2\left\|\boldsymbol{\delta}_1^*\right\|^2\log^2 n.\end{aligned}$$

Applying the Matrix Bernstein inequality then gives

$$\|\boldsymbol{E}_{12}\| \lesssim \sigma\|\boldsymbol{\delta}_1^*\|\sqrt{np}\log^3 n.$$

Combining these bounds on $\|\boldsymbol{E}_i\|$, $i = 1,2,3$, we conclude that (59) holds w.h.p., which completes the proves of the lemma.

G. Proof of Lemma 12

Define the sphere

$$\mathcal{T}_r(b) := \left\{\boldsymbol{Z} \in \mathbb{R}^{p\times p} : \operatorname{rank}(\boldsymbol{Z}) \le r, \|\boldsymbol{Z}\|_F = b\right\}.$$

Let $\mathcal{M}_r(\epsilon/2, 1)$ be the smallest $\epsilon/2$-net of $\mathcal{T}_r'(1)$. We know $|\mathcal{M}_r(\epsilon/2,1)| \le \left(\frac{20}{\epsilon}\right)^{6pr}$ by [8]. For any $0 \le b \le 1$, we know $\mathcal{M}_r(\epsilon/2, b) := \{b\boldsymbol{Z} : \boldsymbol{Z} \in \mathcal{M}(\epsilon/2,1)\}$ is an $\epsilon/2$-net of $\mathcal{T}_r'(b)$, with $|\mathcal{M}_r(\epsilon/2,b)| = |\mathcal{M}_r(\epsilon/2,1)| \le \left(\frac{20}{\epsilon}\right)^{6pr}$. Let $k := \lfloor 2/\epsilon \rfloor \le 2/\epsilon$. Consider the set $\bar{\mathcal{M}}_r(\epsilon) = \{0\} \cup \bigcup_{i=1}^k \mathcal{M}_r(\epsilon/2, i\epsilon/2)$. We claim that $\bar{\mathcal{M}}_r(\epsilon)$ is an ϵ-net of the ball $\bar{\mathcal{T}}_r := \{\boldsymbol{Z} \in \mathbb{R}^{p\times p} : \operatorname{rank}(\boldsymbol{Z}) \le r, \|\boldsymbol{Z}\|_F \le 1\}$, with the additional property that every $\boldsymbol{Z}$'s nearest neighbor $\tilde{\boldsymbol{Z}}$ in $\bar{\mathcal{M}}_r(\epsilon)$ satisfies $\left\|\tilde{\boldsymbol{Z}}\right\|_F \le \|\boldsymbol{Z}\|_F$. To see this, note that for any $\boldsymbol{Z} \in \bar{\mathcal{T}}(r)$, there must be some $0 \le i \le k$ such that $i\epsilon/2 \le \|\boldsymbol{Z}\|_F \le (i+1)\epsilon/2$. Define $\boldsymbol{Z}' := i\epsilon\boldsymbol{Z}/(2\|\boldsymbol{Z}\|_F)$, which is in $\mathcal{T}_r(i\epsilon/2)$. We choose $\tilde{\boldsymbol{Z}}$ to be the point in $\mathcal{M}_r(\epsilon/2, i\epsilon/2)$ that is closest to $\boldsymbol{Z}'$. We have

$$\begin{aligned}\left\|\tilde{\boldsymbol{Z}} - \boldsymbol{Z}\right\|_F \le& \left\|\tilde{\boldsymbol{Z}} - \boldsymbol{Z}'\right\|_F + \|\boldsymbol{Z}' - \boldsymbol{Z}\|_F \\ \le& \epsilon/2 + (\|\boldsymbol{Z}\|_F - i\epsilon/2) \le \epsilon,\end{aligned}$$

and $\left\|\tilde{\boldsymbol{Z}}\right\|_F = i\epsilon/2 \le \|\boldsymbol{Z}\|_F$. The cardinality of $\bar{\mathcal{M}}_r(\epsilon)$ satisfies

$$\begin{aligned}\left|\bar{\mathcal{M}}_r(\epsilon)\right| \le& 1 + \sum_{i=1}^k |\mathcal{M}_r(\epsilon/2, k\epsilon/2)| \\ \le& 1 + \frac{1}{\epsilon}\left(\frac{20}{\epsilon}\right)^{6pr} \le \left(\frac{20}{\epsilon}\right)^{7pr}.\end{aligned}$$

We know that the smallest $\epsilon/2$-net $\mathcal{M}'(\epsilon/2,1)$ of the sphere $\mathcal{T}'(1) := \{\boldsymbol{z} \in \mathbb{R}^p : \|\boldsymbol{z}\| = 1\}$ satisfies $|\mathcal{M}'(\epsilon/2,1)| \le \left(\frac{20}{\epsilon}\right)^p$. It follows from an argument similar to above that there is an ϵ-covering $\bar{\mathcal{M}}'(\epsilon)$ of the ball $\bar{\mathcal{T}}' := \{\boldsymbol{z} \in \mathbb{R}^p : \|\boldsymbol{z}\| \le 1\}$ with cardinality $\left|\bar{\mathcal{M}}'(\epsilon)\right| \le \left(\frac{20}{\epsilon}\right)^{2p}$ and the property that every $\boldsymbol{z}$'s nearest neighbor $\tilde{\boldsymbol{z}}$ in $\bar{\mathcal{M}}'(\epsilon)$ satisfies $\|\tilde{\boldsymbol{z}}\|_2 \le \|\boldsymbol{z}\|_2$.

Let $\bar{\mathcal{S}}_r$ be the set

$$\left\{(\boldsymbol{Z},\boldsymbol{z})\in\mathbb{R}^{p\times p}\times\mathbb{R}^p:\mathrm{rank}(\boldsymbol{Z})\le r,\|\boldsymbol{Z}\|_F^2+\|\boldsymbol{z}\|_2^2\le 1\right\}.$$

We claim that $\bar{\mathcal{N}}_r(\sqrt{2}\epsilon):=\left(\bar{\mathcal{M}}_r(\epsilon)\times\bar{\mathcal{M}}'(\epsilon)\right)\cap\bar{\mathcal{S}}_r$ is an $\sqrt{2}\epsilon$-net of $\bar{\mathcal{S}}_r$. To see this, for any $(\boldsymbol{Z},\boldsymbol{z})\in\bar{\mathcal{S}}_r\subset\bar{\mathcal{T}}(r)\times\bar{\mathcal{T}}'$, we let $\tilde{\boldsymbol{Z}}$ ($\tilde{\boldsymbol{z}}$, resp.) be the point in $\bar{\mathcal{M}}_r(\epsilon)$ ($\bar{\mathcal{M}}'(\epsilon)$, resp.) closest to $\boldsymbol{Z}$ ($\boldsymbol{z}$, resp.) We have

$$\sqrt{\left\|\tilde{\boldsymbol{Z}}-\boldsymbol{Z}\right\|_F^2+\|\tilde{\boldsymbol{z}}-\boldsymbol{z}\|_2^2}\le\sqrt{\epsilon^2+\epsilon^2}=\sqrt{2}\epsilon,$$

and $\left\|\tilde{\boldsymbol{Z}}\right\|_F^2+\|\tilde{\boldsymbol{z}}\|_2^2\le\|\boldsymbol{Z}\|_F^2+\|\boldsymbol{z}\|_2^2\le 1$.

Let $\mathcal{N}_r(\sqrt{2}\epsilon)$ be the projection of the set $\bar{\mathcal{N}}_r(\sqrt{2}\epsilon)$ onto the sphere $\mathcal{S}_r$. Since projection does not increase distance, we are guaranteed that $\mathcal{N}_r(\sqrt{2}\epsilon)$ is an $\sqrt{2}\epsilon$-net of $\mathcal{S}_r$. Moreover,

$$\begin{aligned}&\left|\mathcal{N}_r(\sqrt{2}\epsilon)\right|\\&\le\left|\bar{\mathcal{N}}_r(\sqrt{2}\epsilon)\right|\le\left|\bar{\mathcal{M}}_r(\epsilon)\right|\times\left|\bar{\mathcal{M}}'(\epsilon)\right|\le\left(\frac{20}{\epsilon}\right)^{10pr}.\end{aligned}$$

H. Proof of Lemma 13

For simplicity, let's assume $\boldsymbol{H}$ is symmetric. Later on one can check our proof works for any general matrix $\boldsymbol{H}$ as well. Suppose $\boldsymbol{H}=\sum_{j\in[p]}\sigma_j\boldsymbol{u}_j\boldsymbol{u}_j^\top$, where σ_j is j^{th} eigenvalue and $\boldsymbol{u}_j$ is the corresponding eigenvector with unit Euclidean norm. We have

$$\begin{aligned}&\frac{1}{n}\sum_{i\in[n]}\left\langle\boldsymbol{x}_i\boldsymbol{x}_i^\top,\boldsymbol{H}\right\rangle^2\\&=\frac{1}{n}\sum_{i\in[n]}\left(\sum_{j\in[p]}\sigma_j\left\langle\boldsymbol{x}_i,\boldsymbol{u}_j\right\rangle^2\right)^2\\&=\frac{1}{n}\sum_{i\in[n]}\left(\sum_{j,k\in[p]}\sigma_j\sigma_k\left\langle\boldsymbol{x}_i,\boldsymbol{u}_j\right\rangle^2\left\langle\boldsymbol{x}_i,\boldsymbol{u}_k\right\rangle^2\right)\\&\le\frac{1}{n}\sum_{i\in[n]}\left(\sum_{j\in[p]}|\sigma_j|\langle\boldsymbol{x}_i,\boldsymbol{u}_j\rangle^2\cdot\left\langle\boldsymbol{x}_i\boldsymbol{x}_i^\top,\sum_{k\in[p]}|\sigma_k|\boldsymbol{u}_k\boldsymbol{u}_k^\top\right\rangle\right)\\&\le\frac{1}{n}\max_{i\in[n]}\left\{\|\boldsymbol{x}_i\|_2^2\right\}\cdot\|\boldsymbol{H}\|_F\cdot\sum_{i\in[n]}\left(\sum_{j\in[p]}|\sigma_j|\langle\boldsymbol{x}_i,\boldsymbol{u}_j\rangle^2\right).\end{aligned}\tag{60}$$

By using standard concentration result, for some constants c,c',c'', $\|n^{-1}\sum_{i\in[n]}\boldsymbol{x}_i\boldsymbol{x}_i-\mathbf{I}\|\le 0.1$ with probability at least $1-c\exp(-c'p)$ under assumption $n\ge c''p$. We thus have that w.h.p. $\frac{1}{n}\sum_{i\in[n]}\langle\boldsymbol{x}_i,\boldsymbol{u}_j\rangle^2\le 1.1$ for all $j\in[p]$. Using this result, we continue (60) with

$$\begin{aligned}&\frac{1}{n}\sum_{i\in[n]}\left\langle\boldsymbol{x}_i\boldsymbol{x}_i^\top,\boldsymbol{H}\right\rangle^2\\&\le 1.1\max_{i\in[n]}\left\{\|\boldsymbol{x}_i\|_2^2\right\}\cdot\|\boldsymbol{H}\|_F\left(\sum_{j\in[p]}|\sigma_j|\right)\\&=1.1\max_{i\in[n]}\left\{\|\boldsymbol{x}_i\|_2^2\right\}\cdot\|\boldsymbol{H}\|_F\cdot\|\boldsymbol{H}\|_*.\end{aligned}$$

For each $j\in[p]$, let x_{ij} denote the jth coordinate of $\boldsymbol{x}_i$. By union bound, for some constant c_1,c_2,

$$\Pr\left(\max_{i\in[n],j\in[p]}|x_{ij}|>t\right)\le np\cdot\exp(1-c_2t^2).$$

Choosing $t\asymp\sqrt{\log(np)}$, we thus have w.h.p.

$$\max_{i\in[n]}\left\{\|\boldsymbol{x}_i\|_2^2\right\}\lesssim p\cdot\log(np),$$

which completes our proof.

I. Proof of Lemma 14

We need a standard result on packing the unit hypercube.

Lemma 17 (Varshamov-Gilbert Bound, [29])**.** *For $p\ge 15$, there exists a set $\Omega_0=\{\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_{M_0}\}\subset\{0,1\}^{p-1}$ such that $M\ge 2^{(p-1)/8}$ and $\|\boldsymbol{\xi}_i-\boldsymbol{\xi}_j\|_0\ge\frac{p-1}{8}$, $\forall 1\le i<j\le M_0$.*

We claim that for $i\in[M_0]$, there is at most one $\bar{i}\in[M_0]$ with $\bar{i}\ne i$ such that

$$\|\boldsymbol{\xi}_i-(-\boldsymbol{\xi}_{\bar{i}})\|_0<\frac{p-1}{16};\tag{61}$$

otherwise if there are two distinct i_1, i_2 that satisfy the above inequality, then they also satisfy

$$\|\boldsymbol{\xi}_{i_1}-\boldsymbol{\xi}_{i_2}\|_0\le\|\boldsymbol{\xi}_{i_1}-(-\boldsymbol{\xi}_i)\|_0+\|\boldsymbol{\xi}_{i_2}-(-\boldsymbol{\xi}_i)\|_0<\frac{p-1}{8},$$

which contradicts Lemma 17. Consequently, for each $i\in[M_0]$, we use $\bar{i}$ to denote the unique index in $[M_0]$ that satisfies (61) if such an index exists.

We construct a new set $\Omega\subseteq\Omega_0$ by deleting elements from Ω_0: Sequentially for $i=1,2,\dots,M$, we delete $\boldsymbol{\xi}_{\bar{i}}$ from Ω_0 if $\bar{i}$ exists and both $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_{\bar{i}}$ have not been deleted. Note that at most half of the elements in Ω are deleted in this procedure. The resulting $\Omega=\{\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\dots,\boldsymbol{\xi}_M\}$ thus satisfies $M\ge 2^{(p-1)/16}$ and

$$\min\left\{\|\boldsymbol{\xi}_i-\boldsymbol{\xi}_j\|_0,\|\boldsymbol{\xi}_i+\boldsymbol{\xi}_j\|_0\right\}\ge\frac{p-1}{16},\forall 1\le i<j\le M.$$

J. Proof of Lemma 15

By rescaling, it suffices to prove the lemma for $\sigma=1$. Let $\psi(x):=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$ be the density function of the standard Normal distribution. The density function of $\mathbb{Q}_u$ is

$$f_u(x)=\frac{1}{2}\psi(x-u)+\frac{1}{2}\psi(x+u),$$

and the density of $\mathbb{Q}_v$ is given similarly. We compute

$$\begin{aligned}
&D\left(\mathbb{Q}_u\|\mathbb{Q}_v\right)\\
&=\int_{-\infty}^{\infty} f_u(x)\log\frac{f_u(x)}{f_v(x)}dx\\
&=\frac{1}{2}\int_{-\infty}^{\infty}\left[\psi\left(x-u\right)+\psi\left(x-u\right)\right]\\
&\qquad\times\log\left[\frac{\exp\left(-\frac{(x-u)^2}{2}\right)+\exp\left(-\frac{(x+u)^2}{2}\right)}{\exp\left(-\frac{(x-v)^2}{2}\right)+\exp\left(-\frac{(x+v)^2}{2}\right)}\right]dx\\
&=\frac{1}{2}\int_{-\infty}^{\infty}\left[\psi\left(x-u\right)+\psi\left(x-u\right)\right]\\
&\qquad\times\log\left[\frac{\exp\left(xu-\frac{u^2}{2}\right)+\exp\left(-xu-\frac{u^2}{2}\right)}{\exp\left(xv-\frac{v^2}{2}\right)+\exp\left(-xv-\frac{v^2}{2}\right)}\right]dx\\
&=\frac{1}{2}\int_{-\infty}^{\infty}\left[\psi\left(x-u\right)+\psi\left(x-u\right)\right]\\
&\qquad\times\log\left[\exp\left(-\frac{u^2-v^2}{2}\right)\frac{\exp\left(xu\right)+\exp\left(-xu\right)}{\exp\left(xv\right)+\exp\left(-xv\right)}\right]dx\\
&=\frac{1}{2}\int_{-\infty}^{\infty}\left[\psi\left(x-u\right)+\psi\left(x-u\right)\right]\\
&\qquad\times\left[-\frac{u^2-v^2}{2}+\log\frac{\cosh\left(xu\right)}{\cosh\left(xv\right)}\right]dx\\
&=-\frac{u^2-v^2}{2}\\
&\qquad+\frac{1}{2}\int_{-\infty}^{\infty}\left[\psi\left(x-u\right)+\psi\left(x-u\right)\right]\log\frac{\cosh\left(xu\right)}{\cosh\left(xv\right)}dx.
\end{aligned}\tag{62}$$

By Taylor's Theorem, the expansion of $\log\cosh(y)$ at the point a satisfies

$$\begin{aligned}
&\log\cosh(y)\\
&=\log\cosh(a)+(y-a)\tanh(a)+\frac{1}{2}(y-a)^2\operatorname{sech}^2(u)\\
&\quad-\frac{1}{3}(y-a)^3\tanh(\xi)\operatorname{sech}^2(\xi)
\end{aligned}$$

for some number ξ between a and y. Let $w:=\frac{u+v}{2}$. We expand $\log\cosh(xu)$ and $\log\cosh(xv)$ separately using the above equation, which gives that for some ξ_1 between u and w, and some ξ_2 between v and w,

$$\begin{aligned}
&\log\cosh\left(xu\right)-\log\cosh\left(xv\right)\\
&=x(u-v)\tanh\left(xw\right)+\frac{x^2[(u-w)^2-(v-w)^2]}{2}\operatorname{sech}^2\left(xw\right)\\
&\quad-\frac{x^3\left(u-w\right)^3}{3}\tanh(x\xi_1)\operatorname{sech}^2(x\xi_1)\\
&\quad+\frac{x^3\left(v-w\right)^3}{3}\tanh(x\xi_2)\operatorname{sech}^2(x\xi_2)\\
&=x(u-v)\tanh\left(\frac{x(u+v)}{2}\right)+\frac{-x^3}{3}\left(\frac{u-v}{2}\right)^3\\
&\quad\times\left[\tanh(x\xi_1)\operatorname{sech}^2(x\xi_1)+\tanh(x\xi_2)\operatorname{sech}^2(x\xi_2)\right],
\end{aligned}\tag{63}$$

where the last equality follows from $u-w=w-v=\frac{u-v}{2}$. We bound the RHS of (63) by distinguishing two cases.

Case 1: $u\geq v\geq 0$.*:* Because $\tanh(x\xi_1)$ and $\tanh(x\xi_2)$ have the same sign as x^3, the second term in (63) is negative. Moreover, we have $x\tanh\left(\frac{x(u+v)}{2}\right)\leq x\cdot\frac{x(u+v)}{2}$ since $\frac{u+v}{2}\geq 0$. It follows that

$$\log\cosh\left(xu\right)-\log\cosh\left(xv\right)\leq\frac{x^2(u-v)(u+v)}{2},$$

Substituting back to (62), we obtain

$$\begin{aligned}
&D\left(\mathbb{Q}_u\|\mathbb{Q}_v\right)\\
&\leq-\frac{u^2-v^2}{2}+\frac{1}{2}\int_{-\infty}^{\infty}\left[\psi(x-u)+\psi(x+u)\right]\cdot\frac{x^2(u^2-v^2)}{2}dx\\
&=-\frac{u^2-v^2}{2}+\frac{u^2-v^2}{2}(u^2+1)=\frac{u^2-v^2}{2}u^2.
\end{aligned}$$

Case 2: $v\geq u\geq 0$.*:* Let $h(y):=\tanh(y)-y+\frac{y^3}{3}$. Taking the first order Taylor's expansion at the origin, we know that for any $y\geq 0$ and some $0\leq\xi\leq y$, $h(y)=-2\left(\tanh(\xi)\operatorname{sech}^2(\xi)-\xi\right)y^2\geq 0$ since $\tanh(\xi)\operatorname{sech}^2(\xi)\leq\xi\cdot 1^2$ for all $\xi\geq 0$. This means $\tanh(y)\geq y-\frac{y^3}{3},\forall y\geq 0$. Since $u-v\leq 0$ and $\tanh(\cdot)$ is an odd function, we have

$$\begin{aligned}
&x(u-v)\tanh\left(x(u+v)\right)\\
&\leq x(u-v)\left[x(u+v)-\frac{1}{3}\left(xx(u+v)\right)^3\right].
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
&x\left[\tanh(x\xi_1)\operatorname{sech}^2(x\xi_1)+\tanh(x\xi_2)\operatorname{sech}^2(x\xi_2)\right]\\
&\overset{(a)}{\leq}x(x\xi_1+x\xi_2)\overset{(b)}{\leq}x\cdot 2vx,
\end{aligned}$$

where (a) follows from $\operatorname{sech}^2(y)\leq 1$ and $0\leq y\tanh(y)\leq y^2$ for all y, and (b) follows from $\xi_1,\xi_2\leq v$ since $v\geq w\geq u\geq 0$. Combining the last two displayed equations with (63), we obtain

$$\begin{aligned}
&\log\cosh\left(xu\right)-\log\cosh\left(xv\right)\\
&\leq x(u-v)\left[\frac{x(u+v)}{2}-\frac{1}{3}\left(\frac{x(u+v)}{2}\right)^3\right]+\frac{x^3}{3}\left(\frac{v-u}{2}\right)^3(2vx).
\end{aligned}$$

When $u\leq v$, we get

$$\begin{aligned}
&D\left(\mathbb{Q}_u\|\mathbb{Q}_v\right)\\
&\leq-\frac{u^2-v^2}{2}+\frac{1}{2}\int_{-\infty}^{\infty}\left[\psi(x-u)+\psi(x+v)\right]\\
&\quad\times\left[\frac{u^2-v^2}{2}x^2+\frac{v-u}{3}\left(\frac{u+v}{2}\right)^3x^4+\frac{2v}{3}\left(\frac{v-u}{2}\right)^3x^4\right]dx\\
&=-\frac{u^2-v^2}{2}+\frac{u^2-v^2}{2}(u^2+1)+\left[\frac{(v-u)(u+v)^3}{48}+\frac{v\left(v-u\right)^3}{24}\right]\\
&\quad\times\int_{-\infty}^{\infty}\left[\psi(x-u)+\psi(x+u)\right]x^4dx\\
&=\frac{u^2-v^2}{2}u^2+\left[\frac{(v-u)(u+v)^3}{24}+\frac{2v\left(v-u\right)^3}{24}\right]\left(u^4+6u^2+3\right)\\
&\leq\frac{u^2-v^2}{2}u^2+(v-u)\left[\frac{(2v)^3}{24}+\frac{2v\left(v\right)^2}{24}\right]\left(u^4+6u^2+3\right)\\
&\leq\frac{u^2-v^2}{2}u^2+(v-u)\frac{v^3}{2}\left(u^4+6u^2+3\right).
\end{aligned}$$

Combining the two cases, we conclude that

$$D\left(\mathbb{Q}_u \| \mathbb{Q}_v\right) \leq \frac{u^2-v^2}{2} u^2 + \frac{v^3 \max\left\{0, v-u\right\}}{2} (u^4+6u^2+3).$$

K. Proof of Lemma 16

We recall that for any standard Gaussian variable $z \sim \mathcal{N}(0,1)$, there exists a universal constant $\bar{c}$ such that $\mathbb{E}\left[|z|^k\right] \leq \bar{c}$ for all $k \leq 16$. Now observe that $\mu := \boldsymbol{x}^\top \boldsymbol{\alpha} \sim \mathcal{N}(0, \|\boldsymbol{\alpha}\|^2)$ and $\nu := \boldsymbol{x}^\top \boldsymbol{\beta} \sim \mathcal{N}(0, \|\boldsymbol{\beta}\|^2)$. Because $\boldsymbol{x}^\top \boldsymbol{\alpha} / \|\boldsymbol{\alpha}\| \sim \mathcal{N}(0,1)$ and $\boldsymbol{x}^\top \boldsymbol{\beta} / \|\boldsymbol{\beta}\| \sim \mathcal{N}(0,1)$, it follows from the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\mathbb{E}\left[\left|\boldsymbol{x}^\top \boldsymbol{\alpha}\right|^k \left|\boldsymbol{x}^\top \boldsymbol{\beta}\right|^l\right] \\
\leq& \|\boldsymbol{\alpha}\|^k \|\boldsymbol{\beta}\|^l \sqrt{\mathbb{E}\left|\frac{\boldsymbol{x}^\top \boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|}\right|^{2k} \mathbb{E}_{\boldsymbol{X}} \left|\frac{\boldsymbol{x}^\top \boldsymbol{\beta}}{\|\boldsymbol{\beta}\|}\right|^{2l}} \\
\leq& \bar{c} \|\boldsymbol{\alpha}\|^k \|\boldsymbol{\beta}\|^l .
\end{aligned}$$

This proves the first inequality in the lemma.

For the second inequality in the lemma, note that

$$\begin{aligned}
&\mathbb{E}\left[\left(\left|\boldsymbol{x}^\top \boldsymbol{\alpha}\right|^2 - \left|\boldsymbol{x}^\top \boldsymbol{\beta}\right|^2\right) \left|\boldsymbol{x}^\top \boldsymbol{\alpha}\right|^2\right] \\
&= \mathbb{E}\left|\boldsymbol{x}^\top \boldsymbol{\alpha}\right|^4 - \mathbb{E}\left|\boldsymbol{x}^\top \boldsymbol{\alpha}\right|^2 \left|\boldsymbol{x}^\top \boldsymbol{\beta}\right|^2 \\
&= 3\|\boldsymbol{\alpha}\|^4 - \mathbb{E}\left|\boldsymbol{x}^\top \boldsymbol{\alpha}\right|^2 \left|\boldsymbol{x}^\top \boldsymbol{\beta}\right|^2 .
\end{aligned}$$

But

$$\begin{aligned}
&\mathbb{E}\left|\boldsymbol{x}^\top \boldsymbol{\alpha}\right|^2 \left|\boldsymbol{x}^\top \boldsymbol{\beta}\right|^2 \\
&= \mathbb{E}\left(\alpha_1 x_1 + \cdots + \alpha_p x_p\right)^2 \left(x_1 \beta_1 + \cdots + x_p \beta_p\right)^2 \\
&= \mathbb{E}\sum_{i=1}^p x_i^4 \alpha_i^2 \beta_i^2 + \mathbb{E}\sum_{i \neq j} x_i^2 x_j^2 \alpha_i^2 \beta_j^2 + 2\mathbb{E}\sum_{i \neq j} x_i^2 x_j^2 \alpha_i \alpha_j \beta_i \beta_j \\
&= 3\sum_{i=1}^p \alpha_i^2 \beta_i^2 + \sum_{i \neq j} \alpha_i^2 \beta_j^2 + 2\sum_{i \neq j} \alpha_i \alpha_j \beta_i \beta_j \\
&= 2\sum_{i=1}^p \alpha_i^2 \beta_i^2 + \sum_{i,j} \alpha_i^2 \beta_j^2 + 2\sum_{i \neq j} \alpha_i \alpha_j \beta_i \beta_j \\
&= \|\boldsymbol{\alpha}\|^2 \|\boldsymbol{\beta}\|^2 + 2\sum_{i,j} \alpha_i \alpha_j \beta_i \beta_j \\
&= \|\boldsymbol{\alpha}\|^2 \|\boldsymbol{\beta}\|^2 + 2\left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle^2 . \qquad (64)
\end{aligned}$$

It follows that

$$\begin{aligned}
&\mathbb{E}\left[\left(\left|\boldsymbol{x}^\top \boldsymbol{\alpha}\right|^2 - \left|\boldsymbol{x}^\top \boldsymbol{\beta}\right|^2\right) \left|\boldsymbol{x}^\top \boldsymbol{\alpha}\right|^2\right] \\
=& 3\|\boldsymbol{\alpha}\|^4 - \|\boldsymbol{\alpha}\|^2 \|\boldsymbol{\beta}\|^2 - 2\left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle^2 \\
=& 2\|\boldsymbol{\alpha}\|^4 - 2\left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle^2 \\
\leq& 2\|\boldsymbol{\alpha}\|^4 + 2\left(\|\boldsymbol{\alpha}\|^2 - \left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle\right)^2 - 2\left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle^2 \\
=& 4\|\boldsymbol{\alpha}\|^4 - 4\|\boldsymbol{\alpha}\|^2 \left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle \\
=& 2\|\boldsymbol{\alpha}\|^2 \left(\|\boldsymbol{\alpha}\|^2 - 2\left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle + \|\boldsymbol{\beta}\|^2\right) \\
\leq& 2\|\boldsymbol{\alpha}\|^2 \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|^2 .
\end{aligned}$$

For the third inequality in the lemma, we use the equality (64) to obtain

$$\begin{aligned}
&\mathbb{E}\left(\left|\boldsymbol{x}^\top \boldsymbol{\alpha}\right|^2 - \left|\boldsymbol{x}^\top \boldsymbol{\beta}\right|^2\right)^2 \\
=& \mathbb{E}\left|\boldsymbol{x}^\top \boldsymbol{\alpha}\right|^4 - 2\mathbb{E}\left|\boldsymbol{x}^\top \boldsymbol{\alpha}\right|^2 \left|\boldsymbol{x}^\top \boldsymbol{\beta}\right|^2 + \mathbb{E}\left|\boldsymbol{x}^\top \boldsymbol{\beta}\right|^4 \\
=& 6\|\boldsymbol{\alpha}\|^4 - 2\|\boldsymbol{\alpha}\|^2 \|\boldsymbol{\beta}\|^2 - 4\left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle^2 = 4\|\boldsymbol{\alpha}\|^4 - 4\left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle^2 \\
\leq& 4\|\boldsymbol{\alpha}\|^4 - 4\left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle^2 + 2\left(\|\boldsymbol{\alpha}\|^2 - 2\left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle\right)^2 \\
=& 5\|\boldsymbol{\alpha}\|^4 + 4\left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle^2 - 8\|\boldsymbol{\alpha}\|^2 \left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle \\
\leq& 4\left[\|\boldsymbol{\alpha}\|^4 + \left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle^2 - 2\|\boldsymbol{\alpha}\|^2 \left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle\right] \\
=& \left(2\|\boldsymbol{\alpha}\|^2 - 2\left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle\right)^2 \\
=& \left(\|\boldsymbol{\alpha}\|^2 - 2\left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle + \|\boldsymbol{\beta}\|^2\right)^2 = \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|^4 .
\end{aligned}$$

REFERENCES

[1] Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.*, 40(5):2452–2482, 10 2012.

[2] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

[3] Martin Azizyan, Aarti Singh, and Larry Wasserman. Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. *arXiv preprint arXiv:1306.2035*, 2013.

[4] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *arXiv preprint arXiv:1408.2156*, 2014.

[5] Lucien Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. verw. Gebiete*, 65(2):181–237, 1983.

[6] T. Tony Cai and Anru Zhang. Rop: Matrix recovery via rank-one projections. *Ann. Statist.*, 43(1):102–138, 02 2015.

[7] Emmanuel J. Candès and Xiaodong Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 14(5):1017–1026, 2014.

[8] Emmanuel J. Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[9] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[10] Emmanuel J. Candès, Thomas Strohmer, and Vladislav Voroninski. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *arXiv preprint arXiv:1109.4499*, 2011.

[11] Emmanuel J. Candès, Thomas Strohmer, and Vladislav Voroninski. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming.

Communications on Pure and Applied Mathematics, 66(8):1241–1274, 2013.

[12] Arun Chaganty and Percy Liang. Spectral experts for estimating mixtures of linear regressions. In *International Conference on Machine Learning (ICML)*, 2013.

[13] Moses Charikar, Sudipto Guha, Éva Tardos, and David B. Shmoys. A constant-factor approximation algorithm for the k-median problem (extended abstract). In *Proceedings of the Thirty-first Annual ACM Symposium on Theory of Computing*, STOC '99, pages 1–10, New York, NY, USA, 1999. ACM.

[14] Jiahua Chen. Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, pages 221–233, 1995.

[15] Yuxin Chen and Emmanuel J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *arXiv preprint arXiv:1505.05114*, 2015.

[16] Yuxin Chen, Yuejie Chi, and Andrea Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *Information Theory, IEEE Transactions on*, 61(7):4034–4059, 2015.

[17] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley, 2012.

[18] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[19] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(11):2765–2781, 2013.

[20] Bettina Grün and Friedrich Leisch. Applications of finite mixtures of regression models. *URL: http://cran.r-project.org/web/packages/flexmix/vignettes/regression-examples.pdf*, 2007.

[21] Daniel Hsu and Sham M. Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20. ACM, 2013.

[22] Geoffrey McLachlan and David Peel. *Finite Mixture Models*. Wiley series in probability and statistics: Applied probability and statistics. Wiley, 2004.

[23] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. *Signal Processing, IEEE Transactions on*, 63(18):4814–4826, 2015.

[24] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review*, 52(471), 2010.

[25] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.*, 18:no. 82, 1–9, 2013.

[26] Mahdi Soltanolkotabi, Ehsan Elhamifar, and Emmanuel J. Candès. Robust subspace clustering. *Ann. Statist.*, 42(2):669–699, 04 2014.

[27] Nicolas Stadler, Peter Buhlmann, and Sara Geer. L1-penalization for mixture regression models. *TEST*, 19(2):209–256, 2010.

[28] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[29] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

[30] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Arxiv preprint arxiv:1011.3027*, 2010.

[31] Kert Viele and Barbara Tong. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12(4), 2002.

[32] Yu-Xiang Wang and Huan Xu. Noisy sparse subspace clustering. In *Proceedings of The 30th International Conference on Machine Learning*, pages 89–97, 2013.

[33] Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *arXiv preprint arXiv:1412.8729*, 2014.

[34] C. F. Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[35] Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

[36] Xinyang Yi and Constantine Caramanis. Regularized em algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 1567–1575, 2015.

[37] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *Proceedings of The 31st International Conference on Machine Learning*, pages 613–621, 2014, Arxiv preprint arxiv:1310.3745.

[38] Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

Yudong Chen is an Assistant Professor at the School of Operations Research and Information Engineering at Cornell University. From 2013 to 2015 he was a postdoctoral scholar at the Department of Electrical Engineering and Computer Sciences at University of California, Berkeley. He received his Ph.D. in electrical and computer engineering from the University of Texas at Austin in 2013. He obtained his BS and MS degrees from Tsinghua University, Beijing, China. His research interests include machine learning, high-dimensional and robust statistics, convex and non-convex optimization, and applications in networks and financial systems.

Xinyang Yi graduated with a Ph.D. in Electrical and Computer Engineering from The University of Texas at Austin. He obtained his B.E. degree from Tsinghua University in 2012. His research interests include machine learning and non-convex optimization.

Constantine Caramanis (M06) received his Ph.D. in electrical engineering and computer science from the Massachusetts Institute of Technology, and his A.B. in Mathematics from Harvard. Since 2006, he has been on the faculty in the Department of Electrical and Computer Engineering at The University of Texas at Austin. He received the NSF CAREER award in 2011. His current research interests include robust and large scale optimization and control, machine learning and high-dimensional statistics, with applications to large scale networks.