JMLR: Workshop and Conference Proceedings vol 35:1–45, 2014

A Convex Formulation for Mixed Regression with Two Components: Minimax Optimal Rates

Yudong Chen YUDONG.CHEN@BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences, University of California, Berkeley.

Xinyang Yi YIXY@UTEXAS.EDU
Constantine Caramanis CONSTANTINE@UTEXAS.EDU
Department of Electrical and Computer Engineering, The University of Texas at Austin.

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, and provide upper bounds on the recovery errors for both arbitrary noise and stochastic noise settings. We also give *matching minimax lower bounds* (up to log factors), showing that under certain assumptions, our algorithm is information-theoretically optimal. Our results represent the first tractable algorithm guaranteeing successful recovery with tight bounds on recovery errors and sample complexity.

1. Introduction

This paper considers the problem of *mixed linear regression*, where the output variable we see comes from one of two unknown regressors. Thus we see data $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, where

$$y_i = z_i \cdot \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1^* \rangle + (1 - z_i) \cdot \langle \boldsymbol{x}_i, \boldsymbol{\beta}_2^* \rangle + e_i, \quad i = 1, \dots, n,$$

where $z_i \in \{0, 1\}$ can be thought of as a hidden label, and e_i is the noise. Given the label for each sample, the problem decomposes into two standard regression problems, and can be easily solved. Without it, 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 ($n = O(p)$) sample complexity. Exponential-effort brute force search 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 $O(p^6)$ sample complexity (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. However, its behavior is poorly understood, and in particular, no theoretical guarantees on global convergence are known.

Contributions. In this paper, we tackle both statistical and algorithmic objectives at once. The algorithms we give are efficient, specified by solutions of convex optimization problems; in the

1. It should be possible to improve the tensor rates to $O(p^4)$ for the case of Gaussian design

noiseless, arbitrary noise and stochastic noise regimes, they provide the best known sample complexity results; in the balanced case where nearly half the samples come from each of $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$, we provide matching minimax lower bounds, showing our results are optimal.

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 certain technical conditions, as long as the number of observations for each regressor satisfy $n_1, n_2 \gtrsim p$, our algorithm produces an estimator $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ which satisfies

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

  Note that this immediately implies exact recovery in the noiseless case with $O(p)$ samples.

- In the stochastic noise setting with sub-Gaussian noise and balanced labels, we show under the necessary assumption $n_1, n_2 \gtrsim p$ and a Gaussian design matrix, our estimate satisfies the following (ignoring polylog factors):

$$\|\hat{\boldsymbol{\beta}}_b - \boldsymbol{\beta}_b^*\|_2 \lesssim \begin{cases} \sigma\sqrt{\frac{p}{n}}, & \text{if } \gamma \geq \sigma, \\ \frac{\sigma^2}{\gamma}\sqrt{\frac{p}{n}}, & \text{if } \sigma\left(\frac{p}{n}\right)^{\frac{1}{4}} \leq \gamma \leq \sigma, \\ \sigma\left(\frac{p}{n}\right)^{\frac{1}{4}}, & \text{if } \gamma \leq \sigma\left(\frac{p}{n}\right)^{\frac{1}{4}} \end{cases}$$

  where $b = 1, 2$ and γ is any lower bound of $\|\boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^*\|_2$ and σ^2 is the variance of the noise e_i.

- 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, thus showing that the results obtained by our convex optimization solution 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 and exhibit several phases, thus showing a qualitatively different behavior than in standard regression and many other parametric problems (for which the scaling is $\sqrt{1/n}$).

2. Related Work and Contributions

Mixture models and latent variable modeling are very broadly used in a wide array of contexts far beyond regression. Subspace clustering (Elhamifar and Vidal, 2009; Soltanolkotabi et al., 2013; Wang and Xu, 2013), Gaussian mixture models (Hsu and Kakade, 2012; Azizyan et al., 2013) and k-means clustering are popular examples of unsupervised learning for mixture models. The most popular and broadly implemented approach to mixture problems, including mixed regression, is the so-called Expectation-Maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Peel, 2004). In fact, EM has been used for mixed regression for various application domains (Viele and Tong, 2002; Grün and Leisch, 2007). Despite its wide use, still little is known about its performance beyond local convergence (Wu, 1983).

One exception is the recent work in Yi et al. (2013), which considers mixed regression in the noiseless setting, where they propose an alternating minimization approach initialized by a grid

search and show that it recovers the regressors in the noiseless case with a sample complexity of $O(p \log^2 p)$. However, they do not provide guarantees in the noisy setting, and extension to this setting appears challenging. Another notable exception is the work in Stadler et al. (2010). There, EM is adapted to the high-dimensional sparse regression setting, where the regressors are known to be sparse. The authors use EM to solve a penalized (for sparsity) likelihood function. A 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.

Mixture models have been recently explored using the recently developed technology of tensors in Anandkumar et al. (2012); Hsu and Kakade (2012). In Chaganty and Liang (2013), the authors consider a tensor-based approach, regressing $\mathbf{x}^{\otimes 3}$ against y_i^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 models, as we are. Yet, the tensor approach requires $O(p^6)$ samples, which is several orders of magnitude more than the $O(p \cdot \text{polylog}(p))$ that our work requires. As noted in their work, the higher sampling requirement for using third order tensors seems intrinsic.

In this work we consider the setting with two mixture components. Many interesting applications have binary latent factors: gene mutation present/not, gender, healthy/sick individual, children/adult, etc.; see also the examples in Viele and Tong (2002). Theoretically, the minimax rate was previously unknown even in the two-component case. Extension to more than two components is of great interest.

Finally, we note that our focus is on estimating the regressors $(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$ rather than identifying the hidden labels $\{z_i\}$ or predicting the response y_i for future data points. The relationship between covariates and response is often equally (some times more) important as prediction. For example, the regressors may correspond to unknown signals or molecular structures, and the response-covariate pairs are linear measurements; here the regressors are themselves the object of interest. For many mixture problems, including clustering, identifying the labels accurately for all data points may be (statistically) impossible. Obtaining the regressors allows for an estimate of this label (see Sun et al., 2013, for a related setting).

3. Main Results

In this section we present this paper's main results. In addition, we present the precise setup and assumptions, and introduce the basic notation we use.

3.1. Problem Set Up

Suppose there are two unknown vectors $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ in $\mathbb{R}^p$. We observe n *noisy* linear measurements $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ which satisfy the following: for $b \in \{1, 2\}$ and $i \in \mathcal{I}_b \subseteq [n]$,

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

where $\mathcal{I}_1$ with $n_1 = |\mathcal{I}_1|$ and $\mathcal{I}_2$ with $n_2 = |\mathcal{I}_2|$ denote the subsets of the measurements corresponding to $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$, respectively. Given $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, the goal is to recover $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$. In particular, for the true regressor pair $\boldsymbol{\theta}^* = (\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$ and an estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ of it, we are interested in bounding the recovery error

$$\rho(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) := \min \left\{ \left\| \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^* \right\|_2 + \left\| \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^* \right\|_2, \left\| \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_2^* \right\|_2 + \left\| \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_1^* \right\|_2 \right\},$$

Algorithm 1 Estimate $\boldsymbol{\beta}^*$'s

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 $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{g}} \pm \sqrt{\hat{\lambda}}\hat{\boldsymbol{v}}$. Output: $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$

i.e., the total error in both regressors up to permutation. Unlike the noiseless setting, in the presence of noise, the correct labels are in general irrecoverable.

The key high-level insight that leads to our optimization formulations, is to work in the lifted space of $p \times p$ matrices, *yet without lifting to* 3*-tensors*. Using basic matrix concentration results not available for tensors, this ultimately allows us to provide optimal statistical rates. In this work, we seek to recover the following:

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

Clearly $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ can be recovered from $\boldsymbol{K}^*$ and $\boldsymbol{g}^*$. Indeed, note that

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

Let λ^* and $\boldsymbol{v}^*$ be the first eigenvalue-eigenvector pair of $\boldsymbol{J}^*$. We have $\sqrt{\lambda^*}\boldsymbol{v}^* := \pm\frac{1}{2}\left(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\right)$; together with $\boldsymbol{g}^*$ we can recover $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$. Given approximate versions $\hat{\boldsymbol{K}}$ and $\hat{\boldsymbol{g}}$ of $\boldsymbol{K}^*$ and $\boldsymbol{g}^*$, we obtain estimates $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ using a similar approach, which we give in Algorithm 1. We show below that in fact this recovery procedure is stable, so that if $\hat{\boldsymbol{K}}$ and $\hat{\boldsymbol{g}}$ are close to $\boldsymbol{K}^*$ and $\boldsymbol{g}^*$, Algorithm 1 outputs $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ that are close to $(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$.

We now give the two formulations for arbitrary and stochastic noise, and we state the main results of the paper. For the arbitrary noise case, while one can use the same quadratic objective as we do in arbitrary case, it turns out that the analysis is more complicated than considering a similar objective – an ℓ_1 objective. In the noiseless setting, our results immediately imply exact recovery with an optimal number of samples, and in fact remove the additional log factors in the sample complexity requirements in Yi et al. (2013). In both the arbitrary/adversarial noise setting and the stochastic noise setting, our results are information-theoretically optimal, as they match (up to at most a polylog factor) the minimax lower bounds we derive in Section 3.4.

Notation. We use lower case bold letters to denote vectors, and capital bold-face letters for matrices. For a vector $\boldsymbol{\theta}$, θ_i and $\theta(i)$ both denote its i-th coordinate. We use standard notation for matrix and vector norms, e.g., $\|\cdot\|_*$ to denote the nuclear norm (as known as the trace norm, which is the sum of the singular values of a matrix), $\|\cdot\|_F$ the Frobenius norm, and $\|\cdot\|$ the operator norm. We define a quantity we use repeatedly. Let

$$\alpha := \frac{\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|_2^2}{\|\boldsymbol{\beta}_1^*\|_2^2 + \|\boldsymbol{\beta}_2^*\|^2}. \tag{3}$$

Note that $\alpha > 0$ when $\boldsymbol{\beta}_1^* \neq \boldsymbol{\beta}_2^*$, and is always bounded by 2. We say a number c is a *numerical constant* if c is independent of the dimension p, the number of measurements n and the quantity α. For ease of parsing, we typically use c to denote a *large* constant, and $\frac{1}{c}$ for a *small* constant.

3.2. Arbitrary Noise

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

We consider the following convex program:

$$\min_{\boldsymbol{K},\boldsymbol{g}} \quad \|\boldsymbol{K}\|_* \tag{4}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \left| -\left\langle \boldsymbol{x}_i\boldsymbol{x}_i^\top, \boldsymbol{K} \right\rangle + 2y_i \left\langle \boldsymbol{x}_i, \boldsymbol{g} \right\rangle - y_i^2 \right| \le \eta. \tag{5}$$

The intuition is that in the noiseless case with $\boldsymbol{e} = \boldsymbol{0}$, if we substitute the desired solution $(\boldsymbol{K}^*, \boldsymbol{g}^*)$ given by (2) into the above program, the LHS of (5) becomes zero; moreover, the rank of $\boldsymbol{K}^*$ is 2, and minimizing the nuclear norm term in (4) 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 arbitrary noise. Theorem 1 provides guarantees on how close the optimal solution $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ is to $(\boldsymbol{K}^*, \boldsymbol{g}^*)$; then the companion result, Theorem 2, provides quality bounds on $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$, produced by using Algorithm 1 on the output $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$.

Theorem 1 (Arbitrary Noise) *There exist numerical positive constants $c_1, \ldots, c_6$ such that the following holds. Assume $\frac{n_1}{n_2}, \frac{n_2}{n_1} = \Theta(1)$. Suppose, moreover, that (1) $\mu > 1$ and $\alpha > 0$; (2) $\min\{n_1, n_2\} \ge c_3 \frac{1}{\alpha} p$; (3) the parameter η satisfies*

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

and (4) the noise satisfies

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

Then, with probability at least $1 - c_1 \exp(-c_2 n)$, any optimal solution $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ to the program (4)–(5) satisfies

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

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

We then use Algorithm 1 to estimate $(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$, which is stable as shown by the theorem below.

2. Recall that, as shown in Yi et al. (2013), the general deterministic covariate mixed regression problem is NP-hard even in the noiseless setting.

Theorem 2 (Estimating β^*, arbitrary noise) *Suppose conditions 1–4 in Theorem 1 hold, and* $\eta \asymp \sqrt{n}\,\|e\|_2\,\|\beta_2^* - \beta_1^*\|_2$. *Then with probability at least* $1 - c_1\exp(-c_2 n)$, *the output* $\hat{\theta} = (\hat{\beta}_1, \hat{\beta}_2)$ *of Algorithm 1 satisfies*

$$\rho(\hat{\theta}, \theta^*) \leq \frac{1}{c_3\sqrt{\alpha}}\frac{\|e\|_2}{\sqrt{n}}, \quad b = 1, 2.$$

Theorem 2 immediately implies exact recovery in the noiseless case.

Corollary 3 (Exact Recovery) *Suppose* $e = 0$, *the conditions 1 and 2 in Theorem 1 hold, and* $\eta = 0$. *Then with probability at least* $1 - c_1\exp(-c_2 n)$, *Algorithm 1 returns the true* $\{\beta_1^*, \beta_2^*\}$.

Discussion of Assumptions:

(1) In Theorem 1, the condition $\mu > 1$ is satisfied, for instance, if $\{x_i\}$ is Gaussian (with $\mu = 3$). Moreover, this condition is in general necessary. To see this, suppose each $x_i(l)$ is a Rademacher ± 1 variable, which has $\mu = 1$, and $\beta_1^*, \beta_2^* \in \mathbb{R}^2$. The response variable y_i must have the form

$$y_i = \pm(\beta_b^*)_1 \pm (\beta_b^*)_2.$$

Consider two possibilities: $\beta_1^* = -\beta_2^* = (1, 0)^\top$ or $\beta_1^* = -\beta_2^* = (0, 1)^\top$. In both cases, (x_i, y_i) may take any one of the values in $\{\pm 1\}^2 \times \{\pm 1\}$ with equal probabilities. Thus, it is impossible to distinguish between these two possibilities.

(2) The condition $\alpha > 0$ holds if β_1^* and β_2^* are not equal. Suppose α is lower-bounded by a constant. The main assumption on the noise, namely, $\|e\|_2 \lesssim \sqrt{n}\left(\|\beta_1^*\|_2 + \|\beta_2^*\|_2\right)$ (the condition 4 in Theorem 1) cannot be substantially relaxed if we want a bound on $\|\hat{g} - g^*\|_2$. Indeed, if $|e_i| \gtrsim \|\beta_b^*\|_2$ for all i, then an adversary may choose e_i such that

$$y_i = x_i^\top \beta_b^* + e_i = 0, \quad \forall i,$$

in which case the convex program (4)–(5) becomes independent of g. That said, the case with condition 4 violated can be handled trivially. Suppose $\|e\|_2 \geq c_4\sqrt{\alpha n}\left(\|\beta_1^*\|_2 + \|\beta_2^*\|_2\right)$ for any constant c_4. A standard argument for ordinal linear regression shows that the blind estimator $\hat{\beta} := \min_\beta \sum_{i\in\mathcal{I}_1\cup\mathcal{I}_2}\left|x_i^\top\beta - y_i\right|$ satisfies w.h.p.

$$\max\left\{\left\|\hat{\beta} - \beta_1^*\right\|_2, \left\|\hat{\beta} - \beta_2^*\right\|_2\right\} \lesssim \frac{\|e\|_2}{\sqrt{n}},$$

and this bound is optimal (see the minimax lower bound in Section 3.4). Therefore, the condition 4 in Theorem 1 is not really restrictive, i.e., the case when it holds is precisely the interesting setting.

(3) Finally, note that if $n_1/n_2 = o(1)$ or $n_2/n_1 = o(1)$, then a single β^* explains 100% (asymptotically) of the observed data. Moreover, the standard least squares solution recovers this β^* at the same rates as in standard (not mixed) regression.

Optimality of sample complexity. The sample complexity requirements of Theorem 2 and Corollary 3 are optimal. The results require the number of samples n_1, n_2 to be $\Omega(p)$. Since we are estimating two p dimensional vectors without any further structure, this result cannot be improved.

3.3. Stochastic Noise and Consistency

We now consider the stochastic noise setting. We show that for Gaussian covariate in the balanced setting, we have asymptotic consistency and the rates we obtain match information-theoretic bounds we give in Section 3.4, and hence are minimax optimal. Specifically, our setup is as follows. We assume the covariates $\{\boldsymbol{x}_i\}$ have i.i.d. Gaussian entries with zero mean and unit variance . For the noise, we assume $\{e_i\}$ are i.i.d., zero-mean sub-Gaussian with $\mathbb{E}\left[e_i^2\right] = \sigma^2$ and their sub-Gaussian norm $\|e_i\|_{\psi_2} \leq c\sigma$ for some absolute constant c, and are independent of $\{\boldsymbol{x}_i\}$.

Much like in standard regression, the independence assumption on $\{e_i\}$ makes the least-squares objective analytically convenient. In particular, we consider a Lagrangian formulation, regularizing the squared loss objective with the nuclear norm of $\boldsymbol{K}$. Thus, we solve the following:

$$\min_{\boldsymbol{K},\boldsymbol{g}} \sum_{i=1}^{n} \left(- \left\langle \boldsymbol{x}_i \boldsymbol{x}_i^\top, \boldsymbol{K} \right\rangle + 2y_i \left\langle \boldsymbol{x}_i, \boldsymbol{g} \right\rangle - y_i^2 + \sigma^2 \right)^2 + \lambda \left\| \boldsymbol{K} \right\|_* . \tag{6}$$

We assume the noise variance σ^2 is known and can be estimated.[3] As with the arbitrary noise case, our first theorem guarantees $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ is close to $(\boldsymbol{K}^*, \boldsymbol{g}^*)$, and then a companion theorem gives error bounds on estimating $\boldsymbol{\beta}_b^*$.

Theorem 4 *For any constant $0 < c_3 < 2$, there exist numerical positive constant c_1, c_2, c_4, c_5, c_6, which might depend on c_3, such that the following hold. Assume $\frac{n_1}{n_2}, \frac{n_2}{n_1} = \Theta(1)$. Suppose: (1) $\alpha \geq c_3$; (2) $\min\{n_1, n_2\} \geq c_4 p$; (3) $\{\boldsymbol{x}_i\}$ are Gaussian; and (4) λ satisfies $\lambda \geq c_5 \sigma \left(\|\boldsymbol{\beta}_1^*\|_2 + \|\boldsymbol{\beta}_2^*\|_2 + \sigma \right) \left(\sqrt{np} + |n_1 - n_2 \right.$ With probability at least $1 - c_1 n^{-c_2}$, any optimal solution $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ to the regularized least squares program (6) satisfies*

$$\left\| \hat{\boldsymbol{K}} - \boldsymbol{K}^* \right\|_F \leq c_6 \frac{1}{n} \lambda,$$
$$\|\hat{\boldsymbol{g}} - \boldsymbol{g}^*\|_2 \leq c_6 \frac{1}{n \left(\|\boldsymbol{\beta}_1^*\| + \|\boldsymbol{\beta}_2^*\| + \sigma \right)} \lambda.$$

The bounds in the above theorem depend on $|n_1 - n_2|$. This appears as a result of the objective function in the formulation (6) and not an artifact of our analysis.[4] Nevertheless, in the balanced setting with $|n_1 - n_2|$ small, we have consistency with optimal convergence rate. In this case, running Algorithm 1 on the optimal solution $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ of the program (6) to estimate the $\boldsymbol{\beta}^*$'s, we have the following guarantees.

Theorem 5 (Estimating $\boldsymbol{\beta}^*$, stochastic noise) *Suppose $|n_1 - n_2| = O(\sqrt{n \log n})$, the conditions 1–3 in Theorem 4 hold, $\lambda \asymp \sigma \left(\|\boldsymbol{\beta}_1^*\| + \|\boldsymbol{\beta}_2^*\| + \sigma \right) \sqrt{np} \log^3 n$, and $n \geq c_3 p \log^8 n$. 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*

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

3. We note that similar assumptions are made in Chaganty and Liang (2013). It might be possible to avoid the dependence on σ by using a symmetrized error term (see, e.g. Cai and Zhang, 2013).
4. Intuitively, if the majority of the observations are generated by one of the $\boldsymbol{\beta}_b^*$, then the objective produces a solution that biases toward this $\boldsymbol{\beta}_b^*$ since this solution fits more observations. It might be possible to compensate for such bias by optimizing a different objective.

Notice 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 shall see that these three terms match well with the information-theoretic lower bounds given in Section 3.4, and represent three phases of the error rate.

Discussion of Assumptions. The theoretical results in this sub-section assume Gaussian covariate distribution in addition to sub-Gaussianity of the noise. This 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 covariate. We believe this additional cost is an artifact of our analysis.

3.4. Minimax Lower Bounds

In this subsection, we derive minimax lower bounds on the estimation errors for both the arbitrary and stochastic noise settings. Recall that $\boldsymbol{\theta}^* := (\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*) \in \mathbb{R}^p \times \mathbb{R}^p$ is the true regressor pairs, and we use $\hat{\boldsymbol{\theta}} \equiv \hat{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{y}) = \left(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2\right)$ to denote any estimator, which is a measurable function of the observed data $(\boldsymbol{X}, \boldsymbol{y})$. For any $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $\boldsymbol{\theta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ in $\mathbb{R}^p \times \mathbb{R}^p$, we have defined the error (semi)-metric

$$\rho\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right) := \min\left\{\left\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1'\right\|_2 + \left\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_2'\right\|_2, \left\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2'\right\|_2 + \left\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1'\right\|_2\right\}.$$

Remark 6 *We show in the appendix that $\rho(\cdot, \cdot)$ satisfies the triangle inequality.*

We consider the following class of parameters:

$$\Theta(\underline{\gamma}) := \left\{\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \in \mathbb{R}^p \times \mathbb{R}^p : 2\left\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\right\| \geq \left\|\boldsymbol{\beta}_1\right\| + \left\|\boldsymbol{\beta}_2\right\| \geq \underline{\gamma}\right\}, \tag{7}$$

i.e., pairs of regressors whose norms and separation are lower bounded.

We first consider the arbitrary noise setting, where the noise $\boldsymbol{e}$ is assumed to lie in the ℓ_2-ball $\mathbb{B}(\epsilon) := \{\boldsymbol{\alpha} \in \mathbb{R}^n : \|\boldsymbol{\alpha}\|_2 \leq \epsilon\}$ and otherwise arbitrary. We have the following theorem.

Theorem 7 (Lower bound, arbitrary noise) *There exist universal constants $c_0, c_1 > 0$ such that the following is true. If $n \geq c_1 p$, then for any $\underline{\gamma} > 0$ and any hidden labels $\boldsymbol{z} \in \{0,1\}^n$, we have*

$$\inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}^* \in \Theta(\underline{\gamma})} \sup_{\boldsymbol{e} \in \mathbb{B}(\epsilon)} \rho(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \geq c_0 \frac{\epsilon}{\sqrt{n}} \tag{8}$$

with probability at least $1 - n^{-10}$, where the probability is w.r.t. 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 cannot be improved. Therefore, 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 holds when $n \gtrsim p$.

For the stochastic noise setting, we further assume the two components have equal mixing weights. Recall that $z_i \in \{0,1\}$ is the i-th hidden label, i.e., $z_i = 1$ if and only if $i \in \mathcal{I}_1$ for $i = 1, \ldots, n$. We have the following theorem.

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

1. *For any $\underline{\gamma} > \sigma$, we have*

$$\inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}^* \in \Theta(\underline{\gamma})} \mathbb{E}_{\boldsymbol{X},\boldsymbol{z},\boldsymbol{e}} \left[\rho(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\right] \geq c_0 \sigma \sqrt{\frac{p}{n}}. \tag{9}$$

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

$$\inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}^* \in \Theta(\underline{\gamma})} \mathbb{E}_{\boldsymbol{X},\boldsymbol{z},\boldsymbol{e}} \left[\rho(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\right] \geq c_0 \frac{\sigma^2}{\underline{\gamma}} \sqrt{\frac{p}{n}}. \tag{10}$$

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

$$\inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}^* \in \Theta(\underline{\gamma})} \mathbb{E}_{\boldsymbol{X},\boldsymbol{z},\boldsymbol{e}} \left[\rho(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\right] \geq c_0 \sigma \left(\frac{p}{n}\right)^{1/4}. \tag{11}$$

Here $\mathbb{E}_{\boldsymbol{X},\boldsymbol{z},\boldsymbol{e}}[\cdot]$ denotes the expectation w.r.t. the covariate $\boldsymbol{X}$, the hidden labels $\boldsymbol{z}$ and the noise $\boldsymbol{e}$.

We see that the three lower bounds in the above theorem match the three terms in the upper bound given in Theorem 5 respectively up to a polylog factor, proving the minimax optimality of the error bounds of our convex formulation. Therefore, Theorems 5 and 8 together establish the following minimax error rate (up to a polylog factor) in the stochastic noise setting:

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

where $\underline{\gamma}$ is any lower bound on $\|\boldsymbol{\beta}_1^*\| + \|\boldsymbol{\beta}_2^*\|$. Notice how the scaling of the minimax error rate exhibits three phases depending on the Signal-to-Noise Ratio (SNR) $\underline{\gamma}/\sigma$. (1) 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. (2) 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 $1/n^{\frac{1}{4}}$ and is associated with the demixing of the two components $\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*$. (3) 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 a related phenomenon, see Azizyan et al. (2013); Chen (1995).

4. Proof Outline

In this section, we provide the outline and the key ideas in the proofs of Theorems 1, 4, 7 and 8. The complete proofs, along with the perturbation results of Theorems 2, 5, are deferred to the appendix.

The main hurdle is proving strict curvature near the desired solution $(\boldsymbol{K}^*, \boldsymbol{g}^*)$ in the allowable directions. This is done by demonstrating that a linear operator related to the ℓ_1/ℓ_2 errors satisfies a restricted-isometry-like condition, and that this in turn implies a strict convexity condition along the cone centered at $(\boldsymbol{K}^*, \boldsymbol{g}^*)$ of all directions defined by potential optima.

4.1. 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 $\{\boldsymbol{x}_{b,i}^\top, i = 1, \ldots, n_b\}$. For $b = 1, 2$ and $j = 1, \ldots, \lfloor n_b/2 \rfloor$, define the matrix $\boldsymbol{B}_{b,j} := \boldsymbol{x}_{b,2j}\boldsymbol{x}_{b,2j}^\top - \boldsymbol{x}_{b,2j-1}\boldsymbol{x}_{b,2j-1}^\top$. Also let $\boldsymbol{e}_b := [e_{b,1} \ \cdots \ e_{b,n_b}]^\top \in \mathbb{R}^{n_b}$.

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 $j = 1, \ldots, \lfloor n_b \rfloor$,

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

For each $b = 1, 2$, we also 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

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

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

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

Denote the optimal solution to the optimization problem of interest (either (4) or (6)) 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 $\hat{\boldsymbol{H}}_T^\perp := \mathcal{P}_{T^\perp}\hat{\boldsymbol{H}}$.

4.2. Upper Bounds for Arbitrary Noise: Proof Outline

The proof follows from three main steps.

(1) First, the ℓ_1 error term that in this formulation appears in the LHS of the constraint (5) in the optimization, is naturally related to the operators $\mathcal{A}_b$. Using the definitions above, for any feasible $(\boldsymbol{K}, \boldsymbol{g}) = (\boldsymbol{K}^* + \boldsymbol{H}, \boldsymbol{g}^* + \boldsymbol{h})$, the constraint (5) in the optimization program can be rewritten as

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

This inequality holds in particular for $\boldsymbol{H} = \boldsymbol{0}$ and $\boldsymbol{h} = \boldsymbol{0}$ under the conditions of the theorem, as well as for $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{h}}$ associated with the optimal solution since it is feasible. Now, using directly the definitions for $\mathcal{A}_b$ and $\mathcal{B}_b$, and a simple triangle inequality, we obtain that

$$\lfloor n_b/2 \rfloor \left\| \mathcal{B}_b\left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}^*_b \hat{\boldsymbol{h}}^\top\right) \right\|_1 \leq n_b \left\| \mathcal{A}_b\left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}^*_b \hat{\boldsymbol{h}}^\top\right) \right\|_1 .$$

From the last two display equations, and using now the assumptions on η and on $\boldsymbol{e}$, we obtain an upper bound for $\mathcal{B}$ using the error bound η:

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

(2) Next, we obtain a lower-bound on the last LHS by showing the operator $\mathcal{B}$ is an approximate isometry on low-rank matrices. Note that we want to bound the $\|\cdot\|_2$ norm of $\hat{\boldsymbol{h}}$ and the Frobenius norm of $\hat{\boldsymbol{H}}$, though we currently have an ℓ_1-norm bound on $\mathcal{B}$ in terms of η, above. Thus, the RIP-like condition we require needs to relate these two norms. We show that with high probability, for low-rank matrices,

$$\underline{\delta} \left\| \boldsymbol{Z} \right\|_F \leq \left\| \mathcal{B}_b \boldsymbol{Z} \right\|_1 \leq \bar{\delta} \left\| \boldsymbol{Z} \right\|_F, \quad \forall \boldsymbol{Z} \in \mathbb{R}^{p \times p} \text{ with } \operatorname{rank}(\boldsymbol{Z}) \leq \rho.$$

Proving this RIP-like result is done using concentration and an ϵ-net argument, and requires the assumption $\mu > 1$. We then use this and the optimality of $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$ to obtain the desired lower-bounds

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

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

(3) The remainder of the proof involves combining the upper and lower bounds obtain in the last two steps. After some algebraic manipulations, and use of conditions in the assumptions of the theorem, we obtain the desired recovery error bounds

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

4.3. Upper Bounds for Stochastic Noise: Proof Outline

The main conceptual flow of the proof for the stochastic setting is quite similar to the deterministic noise case, though some significant additional steps are required, in particular, the proof of a second RIP-like result.

(1) 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:

$$\sum_b \left\| n_b \mathcal{A}_b \left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right) + 2\boldsymbol{e}_b \circ (\boldsymbol{X}_b \hat{\boldsymbol{h}}) \right\|_2^2 \leq \lambda \left(\frac{3}{2} \left\| \hat{\boldsymbol{H}}_T \right\|_* - \frac{1}{2} \left\| \hat{\boldsymbol{H}}_T^\perp \right\|_* \right) + \lambda (\gamma + \sigma) \left\| \hat{\boldsymbol{h}} \right\|_2,$$

where we have defined the parameter $\gamma := \left\| \boldsymbol{\beta}_1^* \right\|_2 + \left\| \boldsymbol{\beta}_2^* \right\|_2$. The proof of this inequality involves carefully bounding several noise-related terms using concentration. A consequence of this inequality is that $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{h}}$ cannot be arbitrary, and must live in a certain cone.

(2) The RIP-like condition for $\mathcal{B}_b$ in the stochastic case is more demanding. We prove a second RIP-like condition for $\|\mathcal{B}_b\mathbf{Z} - \boldsymbol{D}_b\mathbf{z}\|_1$, using the Frobenius norm of $\boldsymbol{Z}$ and the ℓ_2-norm of $\mathbf{Z}$:

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

We then bound $\mathcal{A}$ by terms involving $\mathcal{B}$, and then invoke the above RIP condition and the cone constraint to obtain the following lower bound:

$$\sum_b \left\| n_b\mathcal{A}_b\left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^*\hat{\boldsymbol{h}}^\top\right) + 2\boldsymbol{e}_b \circ \left(\boldsymbol{X}_b\hat{\boldsymbol{h}}\right)\right\|_2^2 \gtrsim \frac{1}{8}n\left(\left\|\hat{\boldsymbol{H}}_T\right\|_F + (\gamma+\sigma)\|\hat{\boldsymbol{h}}\|_2\right)^2.$$

(3) We now put together the upper and lower bounds in Step (1) and Step (2). This gives

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

from which it eventually follows that

$$\|\hat{\boldsymbol{h}}\|_2 \lesssim \frac{1}{n(\gamma+\sigma)}\lambda, \qquad \left\|\hat{\boldsymbol{H}}\right\|_F \lesssim \frac{1}{n}\lambda.$$

4.4. Lower Bounds: Proof Outline

The high-level ideas in the proofs of Theorems 7 and 8 are similar: we use a standard argument (Yu, 1997; Yang and Barron, 1999; Birgé, 1983) 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 regressor pairs in (7); 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:

$$\inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}^* \in \Theta(\underline{\gamma})} \mathbb{E}\left[\rho(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)\right] \ge \delta \inf_{\tilde{\boldsymbol{\theta}}} \mathbb{P}\left(\tilde{\boldsymbol{\theta}} \ne \boldsymbol{\theta}^*\right), \tag{12}$$

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 (Cover and Thomas, 2012), we have

$$\mathbb{P}\left(\tilde{\boldsymbol{\theta}} \ne \boldsymbol{\theta}^*\right) \ge 1 - \frac{I\left(\boldsymbol{y}, \boldsymbol{X}; \boldsymbol{\theta}^*\right) + \log 2}{\log M}. \tag{13}$$

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 $\log M \ge (p-1)/16$, which exists by the Varshamov-Gilbert bound (Tsybakov, 2009). 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

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

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 (12) and (13) proves Part 2 of Theorem 8. Theorem 7 and Parts 1, 3 of Theorem 8 are proved in a similar manner.

5. 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.

Acknowledgments

We thank Yuxin Chen for illuminating conversations on the topic. We acknowledge support from NSF Grants EECS-1056028, CNS-1302435, CCF-1116955, and the USDOT UTC – D-STOP Center at UT Austin.

References

Anima Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *CoRR*, abs/1210.7559, 2012.

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

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

T Tony Cai and Anru Zhang. ROP: Matrix recovery via rank-one projections. *arXiv preprint arXiv:1310.5791*, 2013.

Emmanuel 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.

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

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

Yuxin Chen, Yuejie Chi, and Andrea Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *arXiv preprint arXiv:1310.0807*, 2013.

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

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.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2790–2797. IEEE, 2009.

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.

Daniel Hsu and Sham M. Kakade. Learning gaussian mixture models: Moment methods and spectral decompositions. *CoRR*, abs/1206.5766, 2012.

Geoffrey McLachlan and David Peel. *Finite Mixture Models*. Wiley series in probability and statistics: Applied probability and statistics. Wiley, 2004. ISBN 9780471654063. URL http://books.google.com/books?id=7M5vK8OpXZ4C.

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.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *arXiv preprint arXiv:1306.2872*, 2013.

Mahdi Soltanolkotabi, Ehsan Elhamifar, and Emmanuel Candes. Robust subspace clustering. *arXiv preprint arXiv:1301.2603*, 2013.

Nicolas Stadler, Peter Buhlmann, and Sara Geer. L1-penalization for mixture regression models. *TEST*, 19(2):209–256, 2010. ISSN 1133-0686.

Yuekai Sun, Stratis Ioannidis, and Andrea Montanari. Learning mixtures of linear classifiers. *arXiv preprint arXiv:1311.2547*, 2013.

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

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

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

Kert Viele and Barbara Tong. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12(4), 2002. ISSN 0960-3174. URL http://dx.doi.org/10.1023/A%3A1020779827503.

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

CF Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. *Arxiv preprint arxiv:1310.3745*, 2013.

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

Supplemental Results

Appendix A. Proofs of Theorems 2 and 5

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 3.1 and in Algorithm 1.

A key component of the proof involves some perturbation bounds. We prove these in the first section below, and then use them to prove Theorems 2 and 5 in the two subsequent sections.

A.1. Perturbation Bounds

We require the following perturbation bounds.

Lemma 9 *If* $\left\|\hat{\boldsymbol{J}} - \boldsymbol{J}^*\right\|_F \leq \delta$, *then*

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

Proof By Weyl's inequality, we have

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

This implies

$$\left|\sqrt{\hat{\lambda}} - \sqrt{\lambda^*}\right| = \left|\frac{\hat{\lambda} - \lambda^*}{\sqrt{\hat{\lambda}} + \sqrt{\lambda^*}}\right| \leq 2\min\left\{\frac{\delta}{\sqrt{\lambda^*}}, \sqrt{\delta}\right\}. \tag{14}$$

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

$$|\sin\angle(\hat{\boldsymbol{v}}, \boldsymbol{v}^*)| \leq \min\left\{\frac{2\|\hat{\boldsymbol{K}} - \boldsymbol{K}^*\|}{\|\boldsymbol{K}^*\|}, 1\right\} \leq \min\left\{\frac{2\delta}{\lambda^*}, 1\right\}. \tag{15}$$

On the other hand, we have

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

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

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

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 (14) and (15) in the second inequality. 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 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

$$\begin{aligned}\left\|\hat{\boldsymbol{\beta}}_b-\boldsymbol{\beta}_b^*\right\|_2 &\le \|\hat{\boldsymbol{g}}-\boldsymbol{g}^*\|_2+\left\|\sqrt{\hat{\lambda}}\hat{\boldsymbol{v}}-\sqrt{\lambda^*}\boldsymbol{v}^*\right\|_2 \\ &\lesssim \delta_g+\min\left\{\frac{\delta_K+2\delta_g\|\boldsymbol{g}^*\|_2+\delta_g^2}{\|\boldsymbol{\beta}_1^*-\boldsymbol{\beta}_2^*\|_2},\sqrt{\delta_K+2\delta_g\|\boldsymbol{g}^*\|_2+\delta_g^2}\right\},\end{aligned}\tag{16}$$

where the second inequality follows from Lemma 9 and $\lambda^*=\frac{1}{4}\|\boldsymbol{\beta}_1^*-\boldsymbol{\beta}_2^*\|_2^2$.

We shall apply this result to the optimal solution $(\hat{\boldsymbol{K}},\hat{\boldsymbol{g}})$ obtained in the arbitrary noise setting, and in the stochastic noise setting, and thus prove Theorems 2 and 5.

A.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 (16), we get that up to relabeling b,

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

A.3. Proof of Theorem 5 (Stochastic Noise)

Next consider the setting with stochastic noise. Under the assumption of Theorem 5, Theorem 4 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 last inequality follows from the assumption that $n \geq p \log^8 n$ for some $c > 1$. Combining these with (16), we obtain that up to relabeling of b,

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

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

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

Appendix B. Proof of Theorem 1

We now fill in the details for the proof outline given in Section 4.2, and complete the proof of Theorem 1 for the arbitrary noise setting. Some of the more technical or tedious proofs are relegated to the appendix. As in the proof outline, we assume the optimal solution to the optimization is $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}}) = (\boldsymbol{K}^* + \hat{\boldsymbol{H}}, \boldsymbol{g}^* + \hat{\boldsymbol{h}})$, and 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

$$\left\|\hat{\boldsymbol{K}}\right\|_* - \|\boldsymbol{K}^*\|_* \geq \left\|\boldsymbol{K}^* + \hat{\boldsymbol{H}}_T^\perp\right\|_* - \left\|\hat{\boldsymbol{H}}_T\right\|_* - \|\boldsymbol{K}^*\|_* = \left\|\hat{\boldsymbol{H}}_T^\perp\right\|_* - \left\|\hat{\boldsymbol{H}}_T\right\|_*. \tag{17}$$

B.1. Step (1): Consequence of Feasibility

This step uses feasibility of the solution, to get a bound on $\mathcal{B}$ in terms of the error parameter η.

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

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

Therefore, the constraint (5) is equivalent to

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

Using the notation from Section 4.1, this can be rewritten as

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

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 Vershynin (2010), 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

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

where we use the assumptions on $\boldsymbol{e}$ and η in (a) and (b), respectively. This 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 display equations and (19), we further have

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

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

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

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

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

This concludes Step (1) of the proof.

B.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 10 *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 the following:*

$$\underline{\delta} \left\| \boldsymbol{Z} \right\|_F \leq \left\| \mathcal{B}_b \boldsymbol{Z} \right\|_1 \leq \bar{\delta} \left\| \boldsymbol{Z} \right\|_F, \quad \forall \boldsymbol{Z} \in \mathbb{R}^{p \times p} \text{ with } \mathrm{rank}(\boldsymbol{Z}) \leq \rho.$$

We defer the proof of this lemma to the appendix, where in fact we show it is a special case of a similar result we use in Section C.

We now turn to the implications of this lemma, 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 $\left\| \hat{\boldsymbol{K}} \right\|_* \leq \left\| \boldsymbol{K}^* \right\|_*$ 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 $\mathrm{rank}(\hat{\boldsymbol{H}}_i) \leq K$ and the smallest singular value of $\hat{\boldsymbol{H}}_i$ is larger than the largest singular value of $\hat{\boldsymbol{H}}_{i+1}$ (cf. Recht et al. (2010)). By Lemma 10, we get that for each $b = 1, 2$,

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

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 10 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

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

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

Lemma 11 *We have*

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

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 $\text{rank}(\hat{\boldsymbol{H}}_T) \leq 4$. This concludes the proof of Step (2).

B.3. Step (3): Producing Error Bounds

We now combine the result of the three 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

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

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

Appendix C. Proof of Theorem 4

We follow the three steps from the proof outline in Section 4.3, to give the proof of Theorem 4 for the stochastic noise setting. We continue to use the notation given in Section 4.1. For each $b = 1, 2$, we define the vector $\boldsymbol{d}_{b,j} = e_{b,2j}\boldsymbol{x}_{b,2j} - e_{b,2j-1}\boldsymbol{x}_{b,2j-1}$ for $j = 1, \ldots, \lfloor n_b/2 \rfloor$, as well as the vectors $\boldsymbol{c}_{b,i} := y_{b,i}\boldsymbol{x}_{b,i}$ for $i \in [n_b]$. We let $\boldsymbol{D}_b := (\lfloor n_b/2 \rfloor)^{-1}\left[\boldsymbol{d}_{b,1}, \ldots, \boldsymbol{d}_{b,\lfloor n_b/2 \rfloor}\right]^\top \in \mathbb{R}^{\lfloor n_b/2 \rfloor \times p}$. We also define the shorthand

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

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 12 *There exists an absolute constant $c > 0$ such that, if the conclusion of Theorem 4 holds w.h.p. with the additional assumption that*

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

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

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

C.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

$$\begin{aligned}
&\sum_{i=1}^{n} \left(- \left\langle \boldsymbol{x}_i \boldsymbol{x}_i^\top, \hat{\boldsymbol{K}} \right\rangle + 2 y_i \left\langle \boldsymbol{x}_i, \hat{\boldsymbol{g}} \right\rangle - y_i^2 + \sigma^2 \right)^2 + \lambda \left\| \hat{\boldsymbol{K}} \right\|_* \\
\leq &\sum_{i=1}^{n} \left(- \left\langle \boldsymbol{x}_i \boldsymbol{x}_i^\top, \boldsymbol{K}^* \right\rangle + 2 y_i \left\langle \boldsymbol{x}_i, \boldsymbol{g}^* \right\rangle - y_i^2 + \sigma^2 \right)^2 + \lambda \left\| \boldsymbol{K}^* \right\|_* .
\end{aligned}$$

Using the expression (18), we have

$$\begin{aligned}
&\sum_{i=1}^{n} \left(- \left\langle \boldsymbol{x}_i \boldsymbol{x}_i^\top, \hat{\boldsymbol{H}} \right\rangle + 2(\boldsymbol{x}_i^\top \boldsymbol{\beta}_b^* + e_i) \left\langle \boldsymbol{x}_i, \hat{\boldsymbol{h}} \right\rangle - e_i \boldsymbol{x}_i^\top \boldsymbol{\delta}_b^* - (e_i^2 - \sigma^2) \right)^2 + \lambda \left\| \hat{\boldsymbol{K}} \right\|_* \\
\leq &\sum_{i=1}^{n} \left(- e_i \boldsymbol{x}_i^\top \boldsymbol{\delta}_b^* - (e_i^2 - \sigma^2) \right)^2 + \lambda \left\| \boldsymbol{K}^* \right\|_* .
\end{aligned}$$

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

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

Expanding the squares and rearranging terms, we obtain

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

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

$$\begin{aligned}
P &:= 2 \sum_b \| n_b \mathcal{A}_b^* \boldsymbol{w}_b \| , \\
Q &:= \sum_b \| \boldsymbol{\beta}_b^* \|_2 \| n_b \mathcal{A}_b^* \boldsymbol{w}_b \| + \sqrt{p} \left\| \sum_b 2 \boldsymbol{X}_b^\top \mathrm{diag}(\boldsymbol{e}_b) \boldsymbol{w}_b \right\|_\infty ,
\end{aligned}$$

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 Section C.4.

Lemma 13 *Under the assumption of the theorem, we have* $\lambda \geq 2P$ *and* $\lambda \geq \frac{1}{\sigma+\gamma}Q$ *with high probability.*

Applying the lemma, we get

$$\sum_b \left\| n_b \mathcal{A}_b \left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right) + 2\boldsymbol{e}_b \circ (\boldsymbol{X}_b \hat{\boldsymbol{h}}) \right\|_2^2 \leq \lambda \left(\frac{3}{2} \left\| \hat{\boldsymbol{H}}_T \right\|_* - \frac{1}{2} \left\| \hat{\boldsymbol{H}}_T^\perp \right\|_* \right) + \lambda (\gamma + \sigma) \|\hat{\boldsymbol{h}}\|_2. \tag{27}$$

Since the right hand side of (27) is non-negative, we obtain the following cone constraint for the optimal solution:

$$\left\| \hat{\boldsymbol{H}}_T^\perp \right\|_* \leq \frac{5}{2} \left\| \hat{\boldsymbol{H}}_T \right\|_* + (\gamma + \sigma) \|\hat{\boldsymbol{h}}\|_2. \tag{28}$$

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

C.2. Step (2): RIP and Lower Bounds

We can get a lower bound to the expression in the LHS of (27) 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 $\text{rank}(\hat{\boldsymbol{H}}_i) \leq K$ and the smallest singular value of $\hat{\boldsymbol{H}}_i$ is larger than the largest singular value of $\hat{\boldsymbol{H}}_{i+1}$. Then we have the following chain of inequalities:

$$\begin{aligned}
&\sum_b \left\| n_b \mathcal{A}_b \left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right) + 2\boldsymbol{e}_b \circ \left(\boldsymbol{X}_b \hat{\boldsymbol{h}} \right) \right\|_2^2 \\
&\overset{(a)}{\geq} \sum_b \left\| n_b \mathcal{B}_b \left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right) + 2n_b \boldsymbol{D}_b \hat{\boldsymbol{h}} \right\|_2^2 \\
&\overset{(b)}{\geq} \sum_b n_b \left\| \mathcal{B}_b \left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right) + 2\boldsymbol{D}_b \hat{\boldsymbol{h}} \right\|_1^2 \\
&\overset{(c)}{\gtrsim} n \left(\sum_b \left\| \mathcal{B}_b \left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right) + 2\boldsymbol{D}_b \hat{\boldsymbol{h}} \right\|_1 \right)^2 \\
&\overset{(d)}{\geq} n \left(\sum_b \left\| \mathcal{B}_b \left(-\hat{\boldsymbol{H}}_T + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top + \hat{\boldsymbol{H}}_1 \right) + 2\boldsymbol{D}_b \hat{\boldsymbol{h}} \right\|_1 - \sum_b \sum_{i=2}^M \left\| \mathcal{B}_b (\hat{\boldsymbol{H}}_i) \right\|_1 \right)^2.
\end{aligned} \tag{29}$$

Here (a) follows from the definitions of $\mathcal{A}_b$ and $\mathcal{B}_b$ and the triangle inequality, (b) follows from $\|\boldsymbol{u}\|_2 \geq \frac{1}{n_b} \|\boldsymbol{u}\|_1$ 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 (29) in terms of $\|\hat{\boldsymbol{h}}\|_2$ and $\|\hat{\boldsymbol{H}}\|_F$, we need an extension of the previous RIP-like result from Lemma 10, in order to deal with the first term in (29). The following lemma is proved in the appendix.

Lemma 14 *The following holds for some numerical constants* $c, \underline{\delta}, \bar{\delta}$*. For* $b = 1, 2$*, if* $\mu > 1$ *and* $n_b \geq cpr$*, then with probability* $1 - \exp(-n_b)$*, we have the following RIP-2:*

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

Using this we can now bound the last inequality in (29) above. First, note that for each $b = 1, 2$,

$$\sum_{i=2}^{M} \left\| \mathcal{B}_b(\hat{\boldsymbol{H}}_i) \right\|_1 \overset{(a)}{\leq} \bar{\delta} \sum_{i=2}^{M} \left\| \hat{\boldsymbol{H}}_i \right\|_F \leq \bar{\delta} \sum_{i=2}^{M} \frac{1}{\sqrt{K}} \left\| \hat{\boldsymbol{H}}_{i-1} \right\|_* \leq \frac{\bar{\delta}}{\sqrt{K}} \left\| \hat{\boldsymbol{H}}_{T^\perp} \right\|_*, \tag{30}$$

where (a) follows from the upper bound in Lemma 14 with σ set to 0. Then, applying the lower-bound in Lemma 14 to the first term in the parentheses in (29), and (30) 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 11, we obtain

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

Using (28), we further get

$$\begin{aligned}
&\sum_b \left\| n_b \mathcal{A}_b \left(-\hat{\boldsymbol{H}} + 2\boldsymbol{\beta}_b^* \hat{\boldsymbol{h}}^\top \right) + 2\boldsymbol{e}_b \circ \left(\boldsymbol{X}_b \hat{\boldsymbol{h}} \right) \right\|_2^2 \\
\gtrsim & n \left[\left\| \hat{\boldsymbol{H}}_T \right\|_F^2 + \gamma^2 \|\hat{\boldsymbol{h}}\|_2^2 + \sigma^2 \|\hat{\boldsymbol{h}}\|_2^2 - \frac{1}{8} \left\| \hat{\boldsymbol{H}}_T \right\|_*^2 - \frac{1}{25} \left(\gamma^2 + \sigma^2 \right) \|\hat{\boldsymbol{h}}\|_2^2 \right] \\
\gtrsim & \frac{1}{8} n \left(\left\| \hat{\boldsymbol{H}}_T \right\|_F + (\gamma + \sigma) \|\hat{\boldsymbol{h}}\|_2 \right)^2.
\end{aligned} \tag{31}$$

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.

C.3. Step (3): Producing Error bounds

Combining (27) and (31), 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 (28) in (a) and rank $\left(\hat{\boldsymbol{H}}_T\right) \leq 4$ in (b). This completes Step (3) and the proof of the theorem.

C.4. Proof of Lemma 13

We now move to the proof of Lemma 13, which bounds the noise terms P and Q. Note that

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

and

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

So the lemma is implied if we can show

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

But $\lambda \gtrsim \sigma(\gamma+\sigma)\left(\sqrt{np} + |n_1 - n_2|\sqrt{p}\right)\log^3 n$ by assumption of Theorem 4. 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.

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

Lemma 15 *Suppose $\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*$ is supported on the first coordinate. Then w.h.p.*

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

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 Tropp (2012) gives

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

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

Term S_3: We have

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

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

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

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

$$\begin{aligned} \boldsymbol{\chi}^\top :=& \left[\begin{array}{cccccccc} e_{1,1}\boldsymbol{x}_{1,1}^\top & e_{1,2}\boldsymbol{x}_{1,2}^\top & \cdots & e_{1,n_2}\boldsymbol{x}_{1,n_2}^\top & e_{2,1}\boldsymbol{x}_{2,1}^\top & e_{2,2}\boldsymbol{x}_{2,2}^\top & \cdots & e_{2,n_2}\boldsymbol{x}_{2,n_2}^\top \end{array} \right] \in \mathbb{R}^{2n_2 p} \\ \boldsymbol{G} :=& \operatorname{diag}\left(\boldsymbol{\epsilon}_l \boldsymbol{\delta}_1^{*\top}, \boldsymbol{\epsilon}_l \boldsymbol{\delta}_1^{*\top}, \ldots, \boldsymbol{\epsilon}_l \boldsymbol{\delta}_1^{*\top}, -\boldsymbol{\epsilon}_l \boldsymbol{\delta}_1^{*\top}, -\boldsymbol{\epsilon}_l \boldsymbol{\delta}_1^{*\top}, \ldots, -\boldsymbol{\epsilon}_l \boldsymbol{\delta}_1^{*\top} \right) \in \mathbb{R}^{2n_2 p \times 2n_2 p}; \end{aligned}$$

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., Rudelson and Vershynin (2013)), we obtain w.h.p.

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

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

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

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

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

Term S_4: We have w.h.p.

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

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., Vershynin (2010)), and (b) follows from the boundedness of $\boldsymbol{e}$.

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 $\boldsymbol{z}$, i.e.,

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

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

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

$$\|\boldsymbol{X}\boldsymbol{\alpha}\|_2 \leq 2\sqrt{n}\,\|\boldsymbol{\alpha}\|_2\,, \forall \boldsymbol{\alpha} \in \mathbb{R}^p. \tag{32}$$

Hence it suffices to proves (8) in the theorem statment assuming (32) holds.

Let $\boldsymbol{v}$ be an arbibrary 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 $\boldsymbol{e}_1 := \mathbf{0}$ and $\boldsymbol{e}_2 := -\delta\left(2\boldsymbol{z} - \mathbf{1}\right) \circ (\boldsymbol{X}\boldsymbol{v})$. Note that $\|\boldsymbol{e}_2\| \leq 2\sqrt{n}\delta \leq \epsilon$ by (32), 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

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

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

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

E.1. 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 $\rho_1(\boldsymbol{\beta}, \boldsymbol{\beta}') = \min\{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|, \|\boldsymbol{\beta} + \boldsymbol{\beta}'\|\}$. 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 16 *For $p \geq 16$, there exists $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M\} \subset \{0,1\}^{p-1}$ such that*

$$\begin{aligned}M &\geq 2^{(p-1)/16}, \\ \min\left\{\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|_0, \|\boldsymbol{\xi}_i + \boldsymbol{\xi}_j\|_0\right\} &\geq \frac{p-1}{16}, \forall 1 \leq i < j \leq M.\end{aligned}$$

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

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

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 $1 \leq i < j \leq M$,

$$\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2^2 \leq (p-1)\tau^2 = 4c_0^2\frac{\sigma^2(p-1)}{n}. \tag{33}$$

Moreover, we have

$$\begin{aligned}\rho^2(\boldsymbol{\beta}_i, \boldsymbol{\beta}_j) &= \min\left\{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2^2, \|\boldsymbol{\beta}_i + \boldsymbol{\beta}_j\|_2^2\right\} \\ &\geq 4\tau^2 \min\left\{\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|_0, \|\boldsymbol{\xi}_i + \boldsymbol{\xi}_j\|_0\right\} \geq 4 \cdot 4c_0^2\frac{\sigma^2}{n} \cdot \frac{p-1}{16} = \delta_1^2.\end{aligned} \tag{34}$$

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 $\boldsymbol{X}$ are 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 (33) that

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

provided c_0 is sufficiently small. Following a standard argument (Yu, 1997; Yang and Barron, 1999; Birgé, 1983) to transform the estimation problem into a hypothesis testing problem (cf. Eq. (12) and (13)), 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 regressor $\boldsymbol{\beta}^* \in \mathbb{G}^p(\kappa)$, 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.

E.2. Part 2 of the Theorem

Let $\delta_2 := 2c_0 \frac{\sigma^2}{\kappa} \sqrt{\frac{p-1}{n}}$. 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 $\tau := 2c_0 \frac{\sigma^2}{\kappa} \sqrt{\frac{1}{n}}$ 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 16. 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 $1 \leq i < i' \leq M$,

$$\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}\|^2 \leq p\tau^2 = 4c_0^2 \frac{\sigma^4 p}{\kappa^2 n}. \tag{35}$$

Moreover, we have

$$\begin{aligned}
\rho^2\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i'}\right) &= 4 \min\left\{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}\|^2, \|\boldsymbol{\beta}_i + \boldsymbol{\beta}_{i'}\|^2\right\} \\
&\geq 16\tau^2 \min\left\{\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_{i'}\|_0, \|\boldsymbol{\xi}_i + \boldsymbol{\xi}_{i'}\|_0\right\} \geq 16 \cdot 4c_0^2 \frac{\sigma^4}{\kappa^2 n} \cdot \frac{p-1}{16} = \delta_2^2,
\end{aligned} \tag{36}$$

so $\Theta_2 = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ forms a δ_2-packing of the $\Theta(\underline{\gamma})$ 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 17 *The following bounds holds for any $u, v \geq 0$:*

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

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

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

To bound the expectations in the last RHS, we need a simple technical lemma.

Lemma 18 *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

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

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

Using the above lemma and the fact that $\left\|\boldsymbol{\beta}_i\right\|_2=\left\|\boldsymbol{\beta}_{i'}\right\|_2=\kappa$ 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 $c'>0$,

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

where (a) follows from Cauchy-Schwarz inequality, and (b) follows from the first and third inequalities in Lemma 18 as well as $\left\|\boldsymbol{\beta}_i\right\|=\left\|\boldsymbol{\beta}_{i'}\right\|=\kappa\leq\sigma$. It follows that

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

where the last inequality follows from (35) and c'' can be made sufficiently small by choosing c_0 small enough. We therefore obtain

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

using $M\geq 2^{(p-1)/16}$. Following a standard argument (Yu, 1997; Yang and Barron, 1999; Birgé, 1983) to transform the estimation problem into a hypothesis testing problem (cf. Eq. (12) and (13)), we obtain

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

E.3. 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

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

where $\{\boldsymbol{\xi}_i\}$ are the vectors from Lemma 16. Note that $\|\boldsymbol{\beta}_i\|_2 = \sqrt{p-1}\tau = 2c_0\sigma\left(\frac{p}{n}\right)^{1/4} \geq c_1\sigma\left(\frac{p}{n}\right)^{1/4} \geq \kappa$ 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 $1 \leq i < i' \leq M$,

$$\begin{aligned}\rho^2(\boldsymbol{\beta}_i, \boldsymbol{\beta}_{i'}) &= 4\min\left\{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}\|_2^2, \|\boldsymbol{\beta}_i + \boldsymbol{\beta}_{i'}\|_2^2\right\} \\ &\geq 16\tau^2 \min\left\{\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_{i'}\|_0, \|\boldsymbol{\xi}_i + \boldsymbol{\xi}_{i'}\|_0\right\} = 16 \cdot \frac{4c_0^2\sigma^2}{p-1}\sqrt{\frac{p}{n}} \cdot \frac{p-1}{16} \geq \delta_3^2, \end{aligned} \tag{37}$$

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

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

where (a) follows from Lemma 17, (b) follows from Lemma 18, (c) follows from $\|\boldsymbol{\beta}_i\| = 2c_0\sigma\left(\frac{p}{n}\right)^{1/4} \leq \sigma, \forall i$, and c' is a sufficiently small absolute constant. It follows that

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

since $M \geq 2^{(p-1)/8}$. Following a standard argument (Yu, 1997; Yang and Barron, 1999; Birgé, 1983) to transform the estimation problem into a hypothesis testing problem (cf. Eq. (12) and (13)),

we obtain

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

Appendix F. Proofs of Technical Lemmas

F.1. Proof of Lemma 11

Simple algebra shows that

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

and

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

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

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

F.2. Proof of Lemmas 10 and 14

Setting $\sigma = 0$ in Lemma 14 recovers Lemma 10. So we only need to prove Lemma 14. 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 Chen et al. (2013)

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 $\|\xi_j - \mathbb{E}[\xi_j]\|_{\psi_1} \le 2c_1\nu$.

On the other hand, 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 Vershynin (2010), we obtain

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

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

$$\mathbb{P}\left[\left| \|\mathcal{B}\boldsymbol{Z} - \boldsymbol{D}\boldsymbol{z}\|_1 - \mathbb{E}|\xi_j - \gamma_j| \right| \ge \frac{\nu+\omega}{c_3} \right] \le 2\exp[-c_4 n]. \tag{38}$$

But sub-exponentiality implies

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

Hence

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

On the other hand, note that

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

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

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

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

$$\mathbb{E}\langle \boldsymbol{B}_j, \boldsymbol{Z}\rangle^2 = \langle \mathbb{E}[\langle \boldsymbol{B}_j, \boldsymbol{Z}\rangle \boldsymbol{B}_j], \boldsymbol{Z}\rangle = 4\|\boldsymbol{Z}\|_F^2 + 2(\mu-3)\|\mathrm{diag}(\boldsymbol{Z})\|_F^2 \geq 2(\mu-1)\|\boldsymbol{Z}\|_F^2.$$

Some calculation shows that

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

It follows that

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

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

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

Substituting back to (38), we get

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

To complete the proof of the lemma, we use an ϵ-net argument. Define the set

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

We need the following lemma, which is proved in Appendix F.2.1.

Lemma 19 *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 14 for all $(\boldsymbol{Z}, \boldsymbol{z})$ in $\mathcal{S}_r$. By the union bound and Lemma 19, we have

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

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

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

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

$$\sup_{(\boldsymbol{Z}', \boldsymbol{z}') \in \mathcal{S}_{2r}} \left\|\mathcal{B}\boldsymbol{Z}' - \boldsymbol{D}\boldsymbol{z}'\right\|_1 \leq \sup_{\boldsymbol{Z}' \in \mathcal{S}_{2r}} \left\|\mathcal{B}\boldsymbol{Z}_1' - \boldsymbol{D}\boldsymbol{z}'\right\|_1 + \sup_{\boldsymbol{Z}' \in \mathcal{S}_{2r}} \left\|\mathcal{B}\boldsymbol{Z}_2'\right\|_1 \leq 2\bar{M}. \tag{39}$$

Combining the last two display 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 might depend on $\mu := \mathbb{E}\left[(\boldsymbol{x}_i)_l^4\right]$).

For a possibly different ϵ', we have

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

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

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

where (a) follows from (39) 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 $\underline{\delta} := \left(c_7 - \frac{1}{c_3}\right) - 2\epsilon'\bar{\delta} > 0$.

F.2.1. PROOF OF LEMMA 19

Proof Define the sphere

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

Let $\mathcal{M}_r(\epsilon/2, 1)$ be the smallest $\epsilon/2$-net of $\mathcal{T}_r'(1)$. We know $|\mathcal{M}_r(\epsilon/2, 1)| \leq \left(\frac{20}{\epsilon}\right)^{6pr}$ by Candès and Plan (2011). For any $0 \leq b \leq 1$, 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 $|\mathcal{M}_r(\epsilon/2, b)| = |\mathcal{M}_r(\epsilon/2, 1)| \leq \left(\frac{20}{\epsilon}\right)^{6pr}$. Let $k := \lfloor 2/\epsilon \rfloor \leq 2/\epsilon$. 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 $\bar{\mathcal{T}}_r := \{\boldsymbol{Z} \in \mathbb{R}^{p \times p} : \text{rank}(\boldsymbol{Z}) \leq r, \|\boldsymbol{Z}\|_F \leq 1\}$, with the additional property that every $\boldsymbol{Z}$'s nearest neighbor $\tilde{\boldsymbol{Z}}$ in $\bar{\mathcal{M}}_r(\epsilon)$ satisfies $\left\|\tilde{\boldsymbol{Z}}\right\|_F \leq \|\boldsymbol{Z}\|_F$. 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

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

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

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

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 $\|\tilde{\boldsymbol{z}}\|_2 \le \|\boldsymbol{z}\|_2$.

Define the ball $\bar{\mathcal{S}}_r := \left\{(\boldsymbol{Z}, \boldsymbol{z}) \in \mathbb{R}^{p\times p} \times \mathbb{R}^p : \text{rank}(\boldsymbol{Z}) \le r, \|\boldsymbol{Z}\|_F^2 + \|\boldsymbol{z}\|_2^2 \le 1\right\}$. 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,

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

F.3. Proof of Lemma 12

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 $\mathcal{E}_i = \{|e_i| \le L\}$ 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 $m_i \ge 0$. Note that the following equation holds almost surely:

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

Taking the expectation of both sides gives $s_i^2 \ge L m_i$. 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 Ber(ν_i^+) and Ber(ν_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 4 holds w.h.p. for the program (6) 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.

F.4. Proof of Lemma 15

Proof We need to bound

$$S_{1,1} = 2\sum_b \left\| \sum_{i=1}^{n_b} e_{b,i} \boldsymbol{x}_{b,i} \boldsymbol{x}_{b,i}^\top \cdot \boldsymbol{x}_{b,i}^\top \left(\boldsymbol{\beta}_b^* - \boldsymbol{\beta}_{-b}^*\right) \right\|,$$

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.

$$\|\boldsymbol{E}\| := \left\| \sum_{i=1}^{n} e_i \boldsymbol{x}_i \boldsymbol{x}_i^\top \cdot \boldsymbol{x}_i^\top \boldsymbol{\delta}_1^* \right\| \lesssim \sigma \left\|\boldsymbol{\delta}_1^*\right\|_2 \sqrt{np} \log^3 n. \tag{40}$$

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

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

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 Tropp (2012); 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

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

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, and

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

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

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 (40) holds w.h.p., which completes the proves of the lemma.

F.5. Proof of Remark 6

$\rho(\cdot, \cdot)$ satisfies the triangle inequality because

$$\begin{aligned}
&\rho\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right) + \rho\left(\boldsymbol{\theta}, \boldsymbol{\theta}''\right) \\
=& \min\left\{\left\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1'\right\|_2 + \left\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_2'\right\|_2, \left\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2'\right\|_2 + \left\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1'\right\|_2\right\} \\
&+ \min\left\{\left\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1''\right\|_2 + \left\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_2''\right\|_2, \left\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2''\right\|_2 + \left\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1''\right\|_2\right\} \\
=& \min\left\{\left\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1'\right\|_2 + \left\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_2'\right\|_2 + \min\left\{\left\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1''\right\|_2 + \left\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_2''\right\|_2, \left\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2''\right\|_2 + \left\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1''\right\|_2\right\},\right. \\
&\qquad \left.\left\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2'\right\|_2 + \left\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1'\right\|_2 + \min\left\{\left\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1''\right\|_2 + \left\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_2''\right\|_2, \left\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2''\right\|_2 + \left\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1''\right\|_2\right\}\right\} \\
\geq& \min\left\{\min\left\{\left\|\boldsymbol{\beta}_1' - \boldsymbol{\beta}_1''\right\|_2 + \left\|\boldsymbol{\beta}_2' - \boldsymbol{\beta}_2''\right\|_2, \left\|\boldsymbol{\beta}_1' - \boldsymbol{\beta}_2''\right\|_2 + \left\|\boldsymbol{\beta}_2' - \boldsymbol{\beta}_1''\right\|_2\right\},\right. \\
&\qquad \left.+ \min\left\{\left\|\boldsymbol{\beta}_2' - \boldsymbol{\beta}_1''\right\|_2 + \left\|\boldsymbol{\beta}_1' - \boldsymbol{\beta}_2''\right\|_2, \left\|\boldsymbol{\beta}_2' - \boldsymbol{\beta}_2''\right\|_2 + \left\|\boldsymbol{\beta}_1' - \boldsymbol{\beta}_1''\right\|_2\right\}\right\} \\
=& \min\left\{\left\|\boldsymbol{\beta}_1' - \boldsymbol{\beta}_1''\right\|_2 + \left\|\boldsymbol{\beta}_2' - \boldsymbol{\beta}_2''\right\|_2, \left\|\boldsymbol{\beta}_1' - \boldsymbol{\beta}_2''\right\|_2 + \left\|\boldsymbol{\beta}_2' - \boldsymbol{\beta}_1''\right\|_2\right\}.
\end{aligned}$$

F.6. Proof of Lemma 16

We need a standard result on packing the unit hypercube.

Lemma 20 (Varshamov-Gilbert Bound, Tsybakov (2009)) *For $p \geq 15$, there exists a set $\Omega_0 = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{M_0}\} \subset \{0, 1\}^{p-1}$ such that $M \geq 2^{(p-1)/8}$ and $\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|_0 \geq \frac{p-1}{8}$, $\forall 1 \leq i < j \leq M_0$.*

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

$$\|\boldsymbol{\xi}_i - (-\boldsymbol{\xi}_{\bar{i}})\|_0 < \frac{p-1}{16}; \tag{41}$$

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

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

which contradicts Lemma 20. Consequently, for each $i \in [M_0]$, we use $\bar{i}$ to denote the unique index in $[M_0]$ that satisfies (41) if such an index exists.

We construct a new set $\Omega \subseteq \Omega_0$ by deleting elements from Ω_0: Sequentially for $i = 1, 2, \ldots, M$, 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 $\Omega = \{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_M\}$ thus satisfies

$$\begin{aligned}
M &\geq 2^{(p-1)/16}, \\
\min\left\{\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|_0, \|\boldsymbol{\xi}_i + \boldsymbol{\xi}_j\|_0\right\} &\geq \frac{p-1}{16}, \forall 1 \leq i < j \leq M.
\end{aligned}$$

F.7. Proof of Lemma 17

Proof 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

$$\begin{aligned}
D\left(\mathbb{Q}_u \| \mathbb{Q}_v\right) &= \int_{-\infty}^{\infty} f_u(x) \log \frac{f_u(x)}{f_v(x)} dx \\
&= \frac{1}{2}\int_{-\infty}^{\infty} \left[\psi\left(x-u\right) + \psi\left(x-u\right)\right] \log \left[\frac{\exp\left(-\frac{(x-u)^2}{2}\right) + \exp\left(-\frac{(x+u)^2}{2}\right)}{\exp\left(-\frac{(x-v)^2}{2}\right) + \exp\left(-\frac{(x+v)^2}{2}\right)}\right] dx \\
&= \frac{1}{2}\int_{-\infty}^{\infty} \left[\psi\left(x-u\right) + \psi\left(x-u\right)\right] \log \left[\frac{\exp\left(xu - \frac{u^2}{2}\right) + \exp\left(-xu - \frac{u^2}{2}\right)}{\exp\left(xv - \frac{v^2}{2}\right) + \exp\left(-xv - \frac{v^2}{2}\right)}\right] dx \\
&= \frac{1}{2}\int_{-\infty}^{\infty} \left[\psi\left(x-u\right) + \psi\left(x-u\right)\right] \log \left[\exp\left(-\frac{u^2-v^2}{2}\right) \frac{\exp\left(xu\right) + \exp\left(-xu\right)}{\exp\left(xv\right) + \exp\left(-xv\right)}\right] dx \\
&= \frac{1}{2}\int_{-\infty}^{\infty} \left[\psi\left(x-u\right) + \psi\left(x-u\right)\right] \left[-\frac{u^2-v^2}{2} + \log \frac{\cosh\left(xu\right)}{\cosh\left(xv\right)}\right] dx \\
&= -\frac{u^2-v^2}{2} + \frac{1}{2}\int_{-\infty}^{\infty} \left[\psi\left(x-u\right) + \psi\left(x-u\right)\right] \log \frac{\cosh\left(xu\right)}{\cosh\left(xv\right)} dx \qquad (42)
\end{aligned}$$

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

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

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,

$$\begin{aligned}
&\log\cosh\left(xu\right) - \log\cosh\left(xv\right) \\
=&x(u-v)\tanh\left(xw\right) + \frac{x^2\left[\left(u-w\right)^2 - \left(v-w\right)^2\right]}{2}\operatorname{sech}^2\left(xw\right) \\
&- \frac{x^3\left(u-w\right)^3}{3}\tanh(x\xi_1)\operatorname{sech}^2(x\xi_1) + \frac{x^3\left(v-w\right)^3}{3}\tanh(x\xi_2)\operatorname{sech}^2(x\xi_2) \\
=&x(u-v)\tanh\left(\frac{x(u+v)}{2}\right) + \frac{-x^3}{3}\left(\frac{u-v}{2}\right)^3\left[\tanh(x\xi_1)\operatorname{sech}^2(x\xi_1) + \tanh(x\xi_2)\operatorname{sech}^2(x\xi_2)\right], \qquad (43)
\end{aligned}$$

where the last equality follows from $u - w = w - v = \frac{u-v}{2}$. We bound the RHS of (43) 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 (43) 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

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

Substituting back to (42), we obtain

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

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

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

On the other hand, we have

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

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 display equations with (43), we obtain

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

when $a \leq b$, we get

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

Combining the two cases, we conclude that

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

■

F.8. Proof of Lemma 18

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,

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

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

$$\begin{aligned}
\mathbb{E}\left|\boldsymbol{x}^\top \boldsymbol{\alpha}\right|^2 \left|\boldsymbol{x}^\top \boldsymbol{\beta}\right|^2 &= \mathbb{E}\left(\alpha_1 x_1 + \cdots + \alpha_p x_p\right)^2 \left(x_1 \beta_1 + \cdots + x_p \beta_p\right)^2 \\
&= \mathbb{E}\sum_{i=1}^p x_i^4 \alpha_i^2 \beta_i^2 + \mathbb{E}\sum_{i \neq j} x_i^2 x_j^2 \alpha_i^2 \beta_j^2 + 2\mathbb{E}\sum_{i \neq j} x_i^2 x_j^2 \alpha_i \alpha_j \beta_i \beta_j \\
&= 3\sum_{i=1}^p \alpha_i^2 \beta_i^2 + \sum_{i \neq j} \alpha_i^2 \beta_j^2 + 2\sum_{i \neq j} \alpha_i \alpha_j \beta_i \beta_j \\
&= 2\sum_{i=1}^p \alpha_i^2 \beta_i^2 + \sum_{i,j} \alpha_i^2 \beta_j^2 + 2\sum_{i \neq j} \alpha_i \alpha_j \beta_i \beta_j \\
&= \|\boldsymbol{\alpha}\|^2 \|\boldsymbol{\beta}\|^2 + 2\sum_{i,j} \alpha_i \alpha_j \beta_i \beta_j \\
&= \|\boldsymbol{\alpha}\|^2 \|\boldsymbol{\beta}\|^2 + 2\left\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \right\rangle^2 . \qquad (44)
\end{aligned}$$

It follows that

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

For the third inequality in the lemma, we use the equality (44) to obtain

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