Proceedings of Machine Learning Research vol 195:1–66, 2023 36th Annual Conference on Learning Theory

Statistical-Computational Tradeoffs in Mixed Sparse Linear Regression

Gabriel Arpino GA442@CAM.AC.UK
University of Cambridge

Ramji Venkataramanan RV285@CAM.AC.UK
University of Cambridge

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We consider the problem of mixed sparse linear regression with two components, where two k-sparse signals $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^p$ are to be recovered from n unlabelled noisy linear measurements. The sparsity is allowed to be sublinear in the dimension ($k = o(p)$), and the additive noise is assumed to be independent Gaussian with variance σ^2. Prior work has shown that the problem suffers from a $\frac{k}{\mathtt{SNR}^2}$-to-$\frac{k^2}{\mathtt{SNR}^2}$ statistical-to-computational gap, resembling other computationally challenging high-dimensional inference problems such as Sparse PCA and Robust Sparse Mean Estimation (Brennan and Bresler, 2020b); here $\mathtt{SNR} := \|\boldsymbol{\beta}_1\|_2/\sigma^2 = \|\boldsymbol{\beta}_2\|_2/\sigma^2$ is the signal-to-noise ratio. We establish the existence of a more extensive $\frac{k}{\mathtt{SNR}^2}$-to-$\frac{k^2(\mathtt{SNR}+1)^2}{\mathtt{SNR}^2}$ computational barrier for this problem through the method of low-degree polynomials, but show that the problem is computationally hard *only* in a very narrow symmetric parameter regime. We identify a smooth information-computation tradeoff between the sample complexity n and running time $\exp(\tilde{\Theta}(k^2(\mathtt{SNR}+1)^2/(n\mathtt{SNR}^2))$ for any randomized algorithm in this hard regime. Via a simple reduction, this provides novel rigorous evidence for the existence of a computational barrier to solving exact support recovery in sparse phase retrieval with sample complexity $n = \tilde{o}(k^2)$. Our second contribution is to analyze a simple thresholding algorithm which, outside of the narrow regime where the problem is hard, solves the associated mixed regression detection problem in $O(np)$ time and matches the sample complexity required for (non-mixed) sparse linear regression of $\frac{k(\mathtt{SNR}+1)}{\mathtt{SNR}} \log p$; this allows the recovery problem to be subsequently solved by state-of-the-art techniques from the dense case. As a special case of our results, we show that this simple algorithm is order-optimal among a large family of algorithms in solving exact signed support recovery in sparse linear regression. To the best of our knowledge, this is the first thorough study of the interplay between mixture symmetry, signal sparsity, and their joint impact on the computational hardness of mixed sparse linear regression.

Keywords: Low-Degree Polynomials, Computational Lower Bounds, Sparse Linear Regression, Mixture Models, High-Dimensional Statistics

1. Introduction

This work considers the problem of two-component mixed sparse linear regression ($\mathtt{MSLR}$), where the goal is to estimate two k-sparse signals $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^p$ from n *unlabelled* noisy linear measurements. The model is defined as follows.

Definition 1 ($\mathtt{MSLR}$) *For $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{w} \in \mathbb{R}^n$, and $\boldsymbol{z} \in \mathbb{R}^n$, consider the model:*

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_1 \odot \boldsymbol{z} + \boldsymbol{X}\boldsymbol{\beta}_2 \odot (1-\boldsymbol{z}) + \boldsymbol{w},$$

where $\odot$ denotes element-wise product between vectors, $X_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $w_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$, $z_i \overset{i.i.d.}{\sim}$ Bernoulli(ϕ), and $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^p$ each k-sparse. Given $(\boldsymbol{X}, \boldsymbol{y})$ the objective is to estimate $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$.

This model was introduced by Quandt and Ramsey (1978) and has since been widely studied in the machine learning and statistics communities; see, e.g., Städler et al. (2010); Chen et al. (2014); Yi et al. (2014); Fan et al. (2018); Javanmard et al. (2022) and the references therein.

If the latent variables $(z_i)_{i\in[n]}$ are observed, the problem reduces to solving two separate linear regressions. However, in many applications, the latent variables may be unknown as the data may come from different unlabelled sub-populations. The $\texttt{MSLR}$ model captures this effect and has been applied to a variety of settings including market segmentation (Wedel and Kamakura, 2000), music perception (Viele and Tong, 2002), health care (Deb and Holmes, 2000; Luo et al., 2022; Im et al., 2022), and various others (Li et al., 2022; Kazor and Hering, 2019). Variants of mixed regression models called hierarchical mixtures-of-experts have long been studied in the machine learning community (Jordan and Jacobs, 1994), where they have been used for ensemble learning, and in Gated Recurrent Units and Attention Networks (Makkuva et al., 2019).

The maximum-likelihood estimator is a natural choice for estimating the signals $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$. However, the resulting optimization problem is non-convex and NP-hard (Yi et al., 2014). The problem is therefore challenging both statistically and computationally, and a variety of efficient estimators have been proposed. These include spectral methods (Chaganty and Liang, 2013; Yi et al., 2014; Zhang et al., 2022), expectation-maximization (EM) (Khalili and Chen, 2007; Faria and Soromenho, 2010; Städler et al., 2010), alternating minimization (Yi et al., 2014; Shen and Sanghavi, 2019; Ghosh and Kannan, 2020), convex relaxation (Chen et al., 2014), moment descent methods (Li and Liang, 2018; Chen et al., 2020), and the use of tractable non-convex objectives (Zhong et al., 2016; Barik and Honorio, 2022).

Despite recent works addressing the statistical and computational feasibility of mixed linear regression (including but not limited to Azizyan et al. (2013); Pal et al. (2022, 2021)), little is understood about the problem in the high-dimensional sparse regime where both the sample size n and the sparsity k can be sublinear in the dimension p, over the range of all $\texttt{SNR}$ scalings, where $\texttt{SNR} := \|\boldsymbol{\beta}_1\|_2^2/\sigma^2 = \|\boldsymbol{\beta}_2\|_2^2/\sigma^2$ is the signal-to-noise ratio. This regime is motivated by a variety of recent statistical applications, ranging from biology to communications (we refer to the monographs Hastie et al. (2015); Giraud (2021) which contain multiple references). The assumptions in Definition 1 of i.i.d. Gaussian data rows $\boldsymbol{x}_i$ and additive Gaussian noise w_i have been often considered broadly in the high-dimensional statistics literature as an idealized assumption (e.g., Wainwright (2009a,b); Arias-Castro et al. (2011); Janson et al. (2017)).

One aspect of this formulation that is starting to become clear is that in a symmetric parameter regime, the $\texttt{MSLR}$ problem is hard, i.e., it cannot be solved by polynomial-time algorithms at the information-theoretically optimal sample complexity $n_{\texttt{IT}} = \tilde{\Theta}(k/\texttt{SNR}^2)$ (Fan et al., 2018). Exhaustive search typically yields statistically near-optimal estimators for the signal support sets, but the running time is exponential in k. The recent works of Brennan and Bresler (2020b); Fan et al. (2018) provided different ways of quantifying this phenomenon, evidencing a fundamental algorithmic barrier for algorithms performing at all sample complexities $n = \tilde{o}(k^2/\texttt{SNR}^2)$ and sparsities $k = o(\sqrt{p})$ in a very narrow and symmetric parameter regime which we call *Symmetric Balanced Mixture of Sparse Linear Regressions* ($\texttt{SB-MSLR}$), defined as

$$\texttt{SB-MSLR}: \ \phi = 1/2 \text{ and } \boldsymbol{\beta}_1 = -\boldsymbol{\beta}_2. \tag{1}$$

(Here we recall that ϕ is the mixture parameter in Definition 1, so $\phi = \frac{1}{2}$ implies that each y_i is equally likely to come from $\boldsymbol{\beta}_1$ or $\boldsymbol{\beta}_2$.) This phenomenon has been termed a $\frac{k}{\texttt{SNR}^2}$-to-$\frac{k^2}{\texttt{SNR}^2}$ *statistical-to-computational gap*, where the problem is solvable with order $k/\texttt{SNR}^2$ samples, but efficient algorithms require at least order $k^2/\texttt{SNR}^2$ samples. (Throughout this paper, by efficient algorithms we mean those with running time $O(p^\eta)$ for some constant $\eta > 0$.) This computational threshold is similar in order to those derived for a multitude of statistical estimation problems, from variants of Planted Clique, e.g., sparse PCA and robust mean estimation (Brennan and Bresler, 2020b). Notably, SB-MSLR is close to a prominent formulation of sparse phase retrieval where $\boldsymbol{y} = |\boldsymbol{X}\boldsymbol{\beta}| + \boldsymbol{w}$ (Brennan and Bresler, 2020b; Fan et al., 2018), which has been widely studied and is believed to possess a k-to-k^2 statistical-computational gap (Liu et al., 2021; Wu and Rebeschini, 2021).

The special case of sparse linear regression (SLR), where there is only one signal (i.e., $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ in Definition 1) has been extensively studied in the last few decades (Candes and Tao, 2005; Donoho, 2006; Wainwright, 2009b). For SLR, the statistical-computational gap is much smaller, but still exists. Indeed, in the regime where $k = o(p)$, the information-theoretically optimal sample complexity for SLR is of order $\frac{k\log(p/k)}{\log(1+\texttt{SNR})}$ (Wang et al., 2010; Reeves et al., 2019); in contrast, recent works such as Bandeira et al. (2022); Gamarnik and Zadik (2022) have established lower bounds in the regime $\texttt{SNR} \to \infty$ via the study of the Overlap Gap Property and Low Degree polynomials, and shown that a sample complexity of order at least $k \log p$ is required for efficiently solving SLR. Moreover, upper bounds of the same order can be obtained using a number of algorithms (Wainwright, 2009b; Bandeira et al., 2022; Gamarnik and Zadik, 2022).

In this paper, for both MSLR and the special case of SLR, we present new algorithmic lower bounds as well as upper bounds obtained by analyzing a simple thresholding algorithm. The thresholding algorithm, which we call CORR, was used by Bandeira et al. (2022) to obtain upper bounds for approximate support recovery (up to $o(k)$ errors) in SLR, in the setting of binary signal and $\texttt{SNR} \to \infty$ with growing k. In all our results, we make the dependence on SNR explicit, so that they hold for all SNR regimes, including $\texttt{SNR} = \Theta(1)$ and for $\texttt{SNR} = o(1)$. Before summarizing our results, we define the class of prior distributions we consider for the signals $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$.

Signal Priors We consider joint priors for $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ that are marginally uniform over k-sparse vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^p$ with equal norm $\|\boldsymbol{\beta}\|_2$. The case where the two signals have equal norm is more challenging as each entry of the observation $\boldsymbol{y}$ will have the same variance regardless of which signal it corresponds to. We denote such a prior by $\mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$, where the non-zero entries of each vector take values in $\mathcal{D} \subseteq \mathbb{R}$. We assume that $\boldsymbol{\beta}_{\texttt{min}} := \min\{|\beta| \mid \beta \in \mathcal{D}\} > 0$.

Notation We use boldface font for vectors and matrices and plain font to denote scalars (e.g. $\boldsymbol{a}$ and a, respectively). For $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{x}_i$ denotes the i-th row of this matrix and $\boldsymbol{X}_j$ the j-th column of this matrix. Throughout the work, we adopt the standard asymptotic notation $O(\cdot), \Omega(\cdot), o(\cdot), \omega(\cdot)$, and $\Theta(\cdot)$. We let $\tilde{O}(\cdot)$ and analogous variants denote these relations up to polylog factors. By $\lesssim, \gtrsim, \simeq$ we denote inequalities and equality up to constants, respectively. We let $[n] := \{1, 2 \cdots n\}$. For the MSLR setting in Definition 1 and the parameter regime SB-MSLR in (1), we let MSLR $\setminus$ SB-MSLR refer to the MSLR problem with associated parameters lying outside the SB-MSLR parameter regime.

1.1. Our Contributions

In what follows, our computational lower bound results hold in full generality for signals with bounded amplitude in the scaling regime $p \to \infty$, $n \to \infty$ and $k = o(\sqrt{p})$. Our algorithmic achievability results hold for general signals with high probability in the sublinear sparsity regime $p \to \infty$, $n \to \infty$, $k = o(p)$, and $n = \omega(k)$.

Computational Lower Bounds for $\texttt{MSLR}$ We provide novel rigorous evidence through the study of low-degree polynomials (Kunisky et al., 2022; Schramm and Wein, 2022; Hopkins, 2018) that there exists a fundamental algorithmic barrier to solving a detection (hypothesis testing) variant of $\texttt{SB-MSLR}$ at all sample complexities $n = o(\frac{k^2(\texttt{SNR}+1)^2}{\texttt{SNR}^2} \cdot \frac{1}{\log p})$ and sparsities $k = o(\sqrt{p})$. Moreover, we show that this computational barrier implies a smooth tradeoff between sample and time complexities, preventing algorithms with running time less than $\exp(\tilde{\Theta}(\frac{k^2}{n} \cdot (\texttt{SNR}+1)^2/\texttt{SNR}^2))$ from succeeding. These results extend those of Brennan and Bresler (2020b); Fan et al. (2018) by showing that $\texttt{SB-MSLR}$ has a significant statistical-to-computational gap in all $\texttt{SNR}$ regimes (including the noiseless and $\texttt{SNR} = \omega(1)$ regimes), and by identifying a smooth tradeoff between sample size and running time in the hard regime.

We then provide polynomial-time reductions between the detection and recovery variants of $\texttt{SB-MSLR}$, for signals taking nonzero values in $\{1, -1\}$, translating our hardness results to evidence that exact support recovery is just as hard for growing $\texttt{SNR}$ values. We also show that any $\texttt{MSLR}$ regime containing $\texttt{SB-MSLR}$ as a subproblem must be hard, by reducing the $\texttt{SB-MSLR}$ exact recovery problem to exact recovery in the more general *Partially Symmetric Balanced* $\texttt{MSLR}$ regime, or $\texttt{PSB-MSLR}$, where

$$\texttt{PSB-MSLR} : \phi = \frac{1}{2}, \text{ and } \boldsymbol{\beta}_{1,j} = -\boldsymbol{\beta}_{2,j} \text{ for } j \in J \subseteq \operatorname{supp}(\boldsymbol{\beta}_1) \cap \operatorname{supp}(\boldsymbol{\beta}_2), \text{ with } |J| = \Theta(k). \quad (2)$$

Our computational lower bounds for the noiseless version of $\texttt{SB-MSLR}$ yield equivalent lower bounds for exact support recovery in sparse phase retrieval, where $\boldsymbol{y} = |\boldsymbol{X}\boldsymbol{\beta}| + \boldsymbol{w}$. This provides novel rigorous evidence of a computational barrier and a smooth information-computation tradeoff for solving exact support recovery in sparse phase retrieval with $n = \tilde{o}(k^2)$ samples, addressing a prominent open question on the hardness of this problem (Liu et al., 2021; Brennan and Bresler, 2020b; Wu and Rebeschini, 2021).

Algorithms for $\texttt{MSLR}$ Perhaps surprisingly, however, we prove that the above algorithmic barrier vanishes outside of $\texttt{SB-MSLR}$. We show that a simple thresholding algorithm called $\texttt{CORR}$ solves the detection variant of $\texttt{MSLR}$ outside of $\texttt{SB-MSLR}$ with $O(np)$ running time and sample complexity n of order $\frac{k(\texttt{SNR}+1)}{\texttt{SNR}} \log p$, matching that required for efficiently solving sparse linear regression. We note that $\texttt{SB-MSLR}$ is a very narrow parameter regime. Indeed, for signal priors (on the non-zero values) that are absolutely continuous with respect to the Lebesgue measure, the constraint (1) almost surely does not hold, and therefore, $\texttt{CORR}$ succeeds on a set of measure one.

In terms of the original recovery problem, $\texttt{CORR}$ is proven to exactly recover the joint support of both signals outside of a regime slightly broader than $\texttt{PSB-MSLR}$ (see Theorem 8 for a precise statement). Recovery of the joint support then reduces the problem to the dense or proportionally-sparse case ($k/p = n/p = \Theta(1)$) where existing algorithms can infer $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ exactly. This extends the recent work of Mazumdar and Pal (2022) which provides an exact joint support recovery algorithm for the case of binary signals (drawn from $\{0, 1\}^p$) with sample complexity of order $\frac{k(\texttt{SNR}+1)}{\texttt{SNR}} \log^3 p$. We highlight that the assumption of binary signals with all the non-zero entries equal to 1 is restrictive as it does not encompass the important regimes $\texttt{SB-MSLR}$, $\texttt{PSB-MSLR}$ where the problem is hard. We can summarize the algorithmically hard parameter regimes in set notation as:

$$\underset{\text{“Low-degree hard detection”}}{\texttt{SB-MSLR}} \subset \underset{\substack{\text{“Exact support recovery}\\ \text{is hard by reduction”}}}{\texttt{PSB-MSLR}} \subset \texttt{MSLR}.$$

Lower Bounds and Algorithms for SLR Our results also provide clarity into the computational barriers that arise in the special case of sparse linear regression (SLR), where $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. As mentioned in the introduction, previous authors have established that a sample complexity of at least order $k \log p$ is required for efficient algorithms (Bandeira et al., 2022; Gamarnik and Zadik, 2022), with matching algorithmic upper bound results available for the case $\mathtt{SNR} \to \infty$ (Wainwright, 2009b; Bandeira et al., 2022; Gamarnik and Zadik, 2022). We extend these findings and provide rigorous low-degree evidence that polynomial-time algorithms require sample complexity of order at least $n_{\mathtt{alg}}^{\mathtt{SLR}} := \frac{k(\mathtt{SNR}+1)}{\mathtt{SNR}} \log p$ for the detection variant of SLR (in the regime where $\|\boldsymbol{\beta}\|_2^2$ is of order k). Our proof technique consists of a vanilla low-degree calculation for SLR; this is different from the approach of Bandeira et al. (2022), who established a connection between the low-degree method and the Franz-Parisi criterion to obtain computational lower bounds for SLR. Our direct proof technique allows us to explicitly quantify the role of SNR in the problem.

Furthermore, we prove that CORR solves both detection and signed support recovery in SLR with $9n_{\mathtt{alg}}^{\mathtt{SLR}}$ samples, for all SNR scalings and general sparse signal priors. Moreover, it runs in $O(np)$ time which can be significantly more efficient than alternative solutions such as the Lasso depending on the convergence criterion used (Wainwright, 2009b). This in turn certifies the order optimality of CORR for exact signed support recovery in SLR with respect to the class of algorithms that are analytic polynomials of the input of degree at most $O(\log p)$ (including spectral methods running in $O(\log p)$ iterations). We note that the statistical-computational gap in SLR between $\frac{k \log(p/k)}{\log(1+\mathtt{SNR})}$ and $n_{\mathtt{alg}}^{\mathtt{SLR}}$ is only up to multiplicative constants unless $\mathtt{SNR} = \omega(1)$.

Our contributions are summarized along with existing results in Table 1 below, for signals taking values in $\{-1, 0, 1\}$. In Table 1, $n_{\mathtt{IT}}$ denotes the information-theoretic threshold for detection (and by reduction, recovery) and $n_{\mathtt{alg}}$ denotes the sample threshold for efficient algorithms. Importantly, we show that MSLR behaves like SLR outside of the narrow SB-MSLR regime, and reconcile existing results in the literature proving achievable sample complexity of order k in the binary case (Mazumdar and Pal, 2022) but of order k^2 in the general case (Städler et al., 2010). These results lead us to believe that the $\frac{k}{\mathtt{SNR}^2}$-to-$\frac{k^2(\mathtt{SNR}+1)^2}{\mathtt{SNR}^2}$ gap arises from brittle symmetries in the signals, and that SB-MSLR and SLR are computationally very different problems, the former only inefficiently solvable in high-dimensional settings.

1.2. Connections to Previous Work

Among the first works rigorously evidencing statistical-to-computational gaps was that of Barak et al. (2016) who proved a tight computational lower bound for the Planted Clique (PC) problem using the sum-of-squares (SOS) hierarchy. Based on the SOS method, Hopkins (2018) then formulated a conjecture (a version of Conjecture 2 described in the next subsection) on the optimality of low-degree polynomials for hypothesis testing. This approach has yielded evidence for computational barriers in high-dimensional inference problems such as sparse PCA (Hopkins and Steurer, 2017; Bandeira et al., 2020). Other approaches to evidencing computational barriers include the failure of classes of algorithms such as statistical query (Diakonikolas et al., 2019), local (Linial, 1992; Gamarnik and Sudan, 2017) and message passing algorithms (Zdeborova and Krzakala, 2016; Krzakala et al., 2007), and the reduction from variants of canonical "hard" problems such as Planted Clique (Berthet and Rigollet, 2013; Brennan and Bresler, 2020b).

Notably, the problem of high-dimensional MSLR has attracted attention as the special case of SB-MSLR has been shown to exhibit a k-to-k^2 statistical-to-computational gap, which we more pre-

	Information-theoretic lower bound $n_{\texttt{IT}}$	Algorithmic lower bound $n_{\texttt{alg}}$	Algorithms
MSLR (Previous)	$\tilde{\Theta}(k/\texttt{SNR}^2)$ (Fan et al., 2018)	$\tilde{\Theta}(k^2/\texttt{SNR}^2)$ (Fan et al., 2018; Brennan and Bresler, 2020b)	ℓ_1-penalization $\left(n = \Omega(k^2), \texttt{SNR} \to \infty\right)$; Polynomial Identities $\left(\text{for 0-1 valued signals, } n = \Omega(\frac{k(\texttt{SNR}+1)}{\texttt{SNR}} \log^3 p)\right)$ (Städler et al., 2010; Mazumdar and Pal, 2022)
MSLR (This Work)			
SB-MSLR,PSB-MSLR		$\Theta\left(\frac{k^2(\texttt{SNR}+1)^2}{\texttt{SNR}^2}\frac{1}{\log p}\right)$	
MSLR \ SB-MSLR			CORR $\left(n = \Omega(\frac{k(\texttt{SNR}+1)}{\texttt{SNR}} \log p)\right)$
SLR (Previous)	$\Theta\left(\frac{2k\log(p/k)}{\log_2(1+\texttt{SNR})}\right)$ (Wang et al., 2010; Gamarnik and Zadik, 2022; Reeves et al., 2019)	$\Theta(k \log p)$ (Wainwright, 2009b; Gamarnik and Zadik, 2022; Bandeira et al., 2022; Arpino, 2021)	Lasso, CORR , Search, OMP ($n = \Omega(k \log p)$, $\texttt{SNR} \to \infty$) (Wainwright, 2009b; Bandeira et al., 2022; Gamarnik and Zadik, 2022; Wainwright, 2009a; Cai and Wang, 2011)
SLR (This Work)		$\Theta\left(\frac{k(\texttt{SNR}+1)}{\texttt{SNR}} \log p\right)$	CORR ($n \geq \frac{8k(\texttt{SNR}+1)}{\beta^2_{\texttt{min}}\texttt{SNR}} \log 2p$)

Table 1: Summary of contributions for signals taking values in $\{-1, 0, 1\}$.

cisely define as a $\frac{k}{\texttt{SNR}^2}$-to-$\frac{k^2}{\texttt{SNR}^2}$ gap. This was identified through the study of average-case reductions from Planted Clique (Brennan and Bresler, 2020b) and the statistical query model (Fan et al., 2018). After noticing that no polynomial-time algorithms for SB-MSLR were known to succeed below sample complexity $\tilde{\Theta}(k^2/\texttt{SNR}^2)$, Fan et al. (2018) derived lower bounds on the information-theoretic and computational limits of an associated detection problem. Specifically, they proved that the information-theoretic minimal sample complexity is $n = \tilde{\Theta}(k/\texttt{SNR}^2)$, while statistical query algorithms (and conjecturally polynomial-time algorithms) are proven to fail for all sample complexities below the larger threshold of $n = \tilde{o}\left(k^2/\texttt{SNR}^2\right)$. This matches in order the failure threshold of many existing algorithms in the literature, although it has not been rigorously shown that the computational lower bound is tight.

Similarly, Brennan and Bresler (2020b) proved that the associated detection problem we consider in this work (SB-MSLR − D) reduces to a variant of the PC detection problem termed "Secret Leakage PC" in a regime contained within sample complexity $n = o\left(k^2/\texttt{SNR}^2\right)$. The detection version of Planted Clique can be formulated as that of identifying whether a clique of size k has been artifically "planted" in an Erdös-Rényi graph of size n. The problem can be solved by exhaustive search for $k = \Omega(\log n)$. The Planted Clique conjecture is that there is no polynomial time algorithm solving PC if $k = o(\sqrt{n})$. There are a variety of sources of evidence for the PC conjecture, see Feldman et al. (2013); Barak et al. (2016); Brennan and Bresler (2020b) and the references therein.

The results above provide evidence for a $\frac{k}{\texttt{SNR}^2}$-to-$\frac{k^2}{\texttt{SNR}^2}$ statistical-to-computational gap between the information-theoretic and the computational limits of SB-MSLR. More broadly, the work in Brennan and Bresler (2020a) makes a step towards understanding the pervasiveness of k-to-k^2 gaps in high-dimensional statistics by showing that efficient algorithms for learning mixtures with k-

sparse means require at least $\tilde{\Omega}(k^2)$ sample complexity. In Theorem 4, we sharpen the existing computational lower bounds for SB-MSLR, evidencing a more extensive $\frac{k}{\mathtt{SNR}^2}$-to-$\frac{k^2(\mathtt{SNR}+1)^2}{\mathtt{SNR}^2}$ gap, which unlike earlier lower bounds, indicates a significant computational barrier even in the noiseless regime ($\mathtt{SNR} = \infty$).

1.3. The Low-Degree Method

The low-degree method is a framework for obtaining lower bounds on the complexity of hypothesis testing problems, that emerged from the study of the sum-of-squares hierarchy (Barak et al., 2016; Hopkins et al., 2017; Hopkins and Steurer, 2017; Hopkins, 2018). The low-degree method boils down to rigorously ruling out the possibility of low-degree polynomial functions of the input for solving a given hypothesis testing problem. Consider the setting of simple binary hypothesis testing, where one seeks to to distinguish between two distributions $\mathbb{P}_N$ and $\mathbb{Q}_N$ over $\mathbb{R}^N$, where N is the (potentially growing) problem size. Given a sample $\boldsymbol{x}$ drawn from either $\mathbb{P}_N$ or $\mathbb{Q}_N$, the goal is to identify whether $\boldsymbol{x}$ originated from the former or the latter through a hypothesis test. In our setting of MSLR, we can view $N = np + n$ as the total dimension of our data $(\boldsymbol{X}, \boldsymbol{y})$, and notice that $\log N = O(\log p)$. We consider two notions of success in testing:

- **Strong Detection/Distinguishing**: the test succeeds with probability $1 - o(1)$ as $p \to \infty$.
- **Weak Detection/Distinguishing**: the test succeeds with probability $\frac{1}{2} + \epsilon$ for some constant $\epsilon > 0$.

A *degree-D polynomial algorithm* denotes a sequence of (possibly random) multivariate polynomials $g_N : \mathbb{R}^N \to \mathbb{R}$ of degree D, and $f_{\leq D}$ we denotes the orthogonal projection of a function f onto the space of degree-D polynomials. Over the last decade, it has been established that for a large array of high-dimensional testing problems (including sparse PCA, planted clique, community detection, and many others), the class of degree-$O(\log p)$ polynomial algorithms is strictly as powerful as the best known polynomial-time algorithms (Bandeira et al., 2020; Ding et al., 2023; Hopkins, 2018; Hopkins and Steurer, 2017; Hopkins et al., 2017; Kunisky et al., 2022). This is formalized in the following conjecture.

Conjecture 2 (The Low Degree Conjecture Coja-Oghlan et al. (2022); Hopkins (2018)) *Define the chi-square divergence between $\mathbb{P}_N$ and $\mathbb{Q}_N$ as $\chi^2(\mathbb{P}_N \| \mathbb{Q}_N) := \mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}_N} \frac{d\mathbb{P}_N(\boldsymbol{x})}{d\mathbb{Q}_N(\boldsymbol{x})}^2 - 1$, and let $\chi^2_{\leq D}(\mathbb{P}_N \| \mathbb{Q}_N)$ be its projection onto the space of degree-D polynomials.*

- *If $\chi^2_{\leq D}(\mathbb{P}_N \| \mathbb{Q}_N) = O(1)$ for some $D = \omega(\log N)$, strong detection has no polynomial-time algorithm and furthermore requires runtime $\exp(\tilde{\Omega}(D))$.*
- *If $\chi^2_{\leq D}(\mathbb{P}_N \| \mathbb{Q}_N) = o(1)$ for some $D = \omega(\log N)$, weak detection has no polynomial-time algorithm and furthermore requires runtime $\exp(\tilde{\Omega}(D))$.*

A variety of state-of-the-art algorithms can be approximated by low-degree polynomials and therefore rigorously ruled out by low-degree lower bounds of the above form, including the important class of spectral methods (see Theorem 4.4 of Kunisky et al. (2022)), and all statistical query algorithms (Brennan et al., 2021). Recent works have also proven the equivalence between low-degree polynomial algorithms and well-established algorithmic solutions derived from statistical physics in certain classes of problems (Bandeira et al., 2022; Montanari and Wein, 2022). Although degree $O(\log p)$ polynomials are not proven to encompass all polynomial-time algorithms, the success of such a polynomial in hypothesis testing tends to indicate the success of general polynomial-time algorithms. In this light, we aim to provide concrete evidence for computational hardness in MSLR and SLR by proving a low-degree lower bound of the form $\chi^2_{\leq D}(\mathbb{P}_N \| \mathbb{Q}_N) = O(1)$ for an associated detection problem, which can then be reduced to recovery. For more background on the low-degree method, see Appendix A.

2. Main Results

2.1. Lower bounds for MSLR

We begin by defining a detection variant of MSLR, where given $(\boldsymbol{X}, \boldsymbol{y})$ the goal is to distinguish between two hypotheses: one in which the data correspond to the MSLR model, and another in which $\boldsymbol{X}$ and $\boldsymbol{y}$ are independent.

Definition 3 (Detection Variant $\mathtt{MSLR}-\mathtt{D}$**)** *For* $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\sigma > 0$, *and* $\boldsymbol{w} \in \mathbb{R}^n$, *consider the following hypothesis testing problem:*

$$\mathbb{P}(\boldsymbol{X}) \otimes \mathbb{P}(\boldsymbol{y}) : \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} \\ \sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1} \cdot \boldsymbol{w} \end{bmatrix}$$

$$\mathbb{P}(\boldsymbol{X}, \boldsymbol{y}) : \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} \\ \frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_1 \odot \boldsymbol{z} + \frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_2 \odot (1-\boldsymbol{z}) + \boldsymbol{w} \end{bmatrix}$$

where $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \sim \mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$, *and* $X_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $w_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $z_i \overset{i.i.d.}{\sim}$ *Bernoulli*(ϕ). *The task is to construct a function* f *which strongly distinguishes* $\mathbb{P}(\boldsymbol{X}) \otimes \mathbb{P}(\boldsymbol{y})$ *from* $\mathbb{P}(\boldsymbol{X}, \boldsymbol{y})$.

Notice that the marginal distributions of $\mathbb{P}(\boldsymbol{X}) \otimes \mathbb{P}(\boldsymbol{y})$ and $\mathbb{P}(\boldsymbol{X}, \boldsymbol{y})$ are equal, so as to rule out solutions that simply threshold the moments of $\boldsymbol{y}$ and ignore $\boldsymbol{X}$. The corresponding detection variant of SB-MSLR, denoted by $\mathtt{SB\text{-}MSLR}-\mathtt{D}$, is defined similarly to $\mathtt{MSLR}-\mathtt{D}$ in the parameter regime of SB-MSLR given in (1). From this formulation we obtain the following hardness result for $\mathtt{SB\text{-}MSLR}-\mathtt{D}$. The proof is given in Appendix B.3.

Theorem 4 (Low-degree lower bound for $\mathtt{SB\text{-}MSLR}-\mathtt{D}$**)** *Consider the setting of* $\mathtt{SB\text{-}MSLR}-\mathtt{D}$ *with* $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \sim \mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$, *and bounded amplitude signals* ($\boldsymbol{\beta}_{min} = \Theta(\|\boldsymbol{\beta}\|_\infty)$). *For sample sizes* n *where* $n = \omega(\max\{k, \log p\})$ *and* $n = o\left(\frac{k^2(SNR+1)^2}{SNR^2} \cdot \frac{1}{\log p}\right)$, *Conjecture 2 implies that any randomized algorithm requires running time* $\exp\left(\tilde{\Omega}\left(\min\left\{\frac{k^2(SNR+1)^2}{nSNR^2}, n\right\}\right)\right)$ *to solve* $\mathtt{SB\text{-}MSLR}-\mathtt{D}$ *in the regime* $k = o(\sqrt{p})$.

Theorem 4 is our main low-degree hardness result. There are three regimes of interest, which we describe in terms of $n_{\mathtt{alg}}^{\mathtt{SB\text{-}MSLR}} := \frac{k^2(\mathrm{SNR}+1)^2}{\mathrm{SNR}^2}$. First, if $n = \Omega\left(n_{\mathtt{alg}}^{\mathtt{SB\text{-}MSLR}}/\log p\right)$, the lower bound on the running time in Theorem 4 equals $e^{\tilde{O}(\log p)}$, and hence does not rule out polynomial-time solutions. Otherwise, Theorem 4 (via Conjecture 2) implies a smooth tradeoff between sample size n and super-polynomial (but sub-exponential) running time $\exp\left(\tilde{\Omega}\left(n_{\mathtt{alg}}^{\mathtt{SB\text{-}MSLR}}/n\right)\right)$, for $n = \omega((n_{\mathtt{alg}}^{\mathtt{SB\text{-}MSLR}})^{\frac{1}{2}})$; this is reminiscent of a similar tradeoff in Sparse PCA (Ding et al., 2023). In the third case, where $n = o((n_{\texttt{alg}}^{\texttt{SB-MSLR}})^{\frac{1}{2}})$, Theorem 4 implies that $e^{\tilde{\Omega}(n)}$ running time is required. Thus there are three distinct computational regimes depending on the sample complexity n: the first permitting polynomial-time solutions, the second enforcing a smooth inversely related information-computation tradeoff, and the last implying an exponential increase in running time as the sample size increases. This extends the results of (Brennan and Bresler, 2020b; Fan et al., 2018) which indicated that the $n = \tilde{o}(k^2/\texttt{SNR}^2)$ sample regime presents statistical-query and planted-clique related algorithmic barriers for $\texttt{SB-MSLR} - \texttt{D}$ with signals in $\{-1, 0, 1\}^p$; note that a lower bound of order $k^2/\texttt{SNR}^2$ is vacuous in the noiseless setting, as well as in the natural setting where $\texttt{SNR} = \frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} = \Theta(k)$.

The work in Fan et al. (2018) proved that the information-theoretic minimal sample complexity of $\texttt{SB-MSLR} - \texttt{D}$ is $n = \tilde{\Theta}(k/\texttt{SNR}^2)$, which is vacuous for $\texttt{SNR} = \omega(\sqrt{k})$. The information-theoretic minimal sample complexity of the related sparse phase retrieval ($\texttt{SPR}$) detection problem, however, is known to be of order $k \log p$ for a broad class of signal-to-noise ratios (see, for example, Theorem 3.2 in Cai et al. (2016) and Section 6.1 in Lecué and Mendelson (2015)). By straightforward reductions from $\texttt{SLR}$ to $\texttt{SB-MSLR}$ to $\texttt{SPR}$, one can show that the information-theoretic sample complexity of detection in $\texttt{SB-MSLR}$ lies between $\frac{k \log(p/k)}{\log(1+\texttt{SNR})}$ and $k \log p$. In this light, Theorem 4 certifies a statistical-computational gap in $\texttt{SB-MSLR} - \texttt{D}$ of order at least k for broad $\texttt{SNR}$ regimes.

We highlight that Theorem 4 rigorously rules out the success of analytic polynomials of the input of degree at most $O(\log p)$, including spectral methods. The $k = o(\sqrt{p})$ assumption is often standard for detection lower bounds where the signal is k-sparse (see (Brennan and Bresler, 2020b; Fan et al., 2018; Ding et al., 2023) and references therein), and can at times be lifted by conditioning away a certain bad event (Bandeira et al., 2022).

Remark 5 *We have included the bounded amplitude assumption in Theorem 4 for interpretability. The dependence on $\|\boldsymbol{\beta}\|_\infty$ can be made explicit by replacing k^2 in Theorem 4 with $\|\boldsymbol{\beta}\|_2^4/\|\boldsymbol{\beta}\|_\infty^4$. We believe the dependence on $\|\boldsymbol{\beta}\|_\infty$ is an artifact of the proof technique; see Appendix B.3.*

Through Theorem 28 in Appendix C, we provide a polynomial-time reduction from $\texttt{SB-MSLR} - \texttt{D}$ to exact support recovery in $\texttt{PSB-MSLR}$, for signals in $\{-1, 0, 1\}^p$ and $\texttt{SNR} = \omega(1)$, transferring hardness from Theorem 4 to this case. In Appendix C.3, we provide a polynomial-time reduction from $\texttt{SB-MSLR} - \texttt{D}$ to both exact support recovery and detection in sparse phase retrieval ($\texttt{SPR}$) for signals with non-zero entries in $\{-1, 0, 1\}^p$, translating the hardness results of Theorem 4 to $\texttt{SPR}$. This provides novel rigorous evidence for the conjecture that $\texttt{SPR}$ is computationally infeasible for sample sizes $n = \tilde{o}(k^2)$ (Wu and Rebeschini, 2021; Li et al., 2022; Brennan and Bresler, 2020b).

2.2. Algorithms for MSLR

We denote the support sets of $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ by $\mathcal{S}_1, \mathcal{S}_2$, respectively. Note that $|\mathcal{S}_1| = |\mathcal{S}_2| = k$. Let us define the following quantities:

$$\langle \boldsymbol{\beta} \rangle_{\texttt{min}}^2 := \min_{j \in \mathcal{S}_1 \cup \mathcal{S}_2} (\phi \boldsymbol{\beta}_{1,j} + (1-\phi)\boldsymbol{\beta}_{2,j})^2,$$

$$\langle \boldsymbol{\beta} \rangle_{>0}^2 := \min_{\substack{j \in \mathcal{S}_1 \cup \mathcal{S}_2 \\ (\phi \boldsymbol{\beta}_{1,j} + (1-\phi)\boldsymbol{\beta}_{2,j}) > 0}} (\phi \boldsymbol{\beta}_{1,j} + (1-\phi)\boldsymbol{\beta}_{2,j})^2.$$

Note that $\langle \boldsymbol{\beta} \rangle_{>0}^2 > 0$ for $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \sim \mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$ outside of the $\texttt{SB-MSLR}$ regime. Also recall that $\boldsymbol{\beta}_{\texttt{min}} = \min\{|\beta| \mid \beta \in \mathcal{D}\} > 0$.

Definition 6 (CORR) *Let* CORR *be the algorithm that outputs an estimate of the joint support set* $\mathcal{S}_1 \cup \mathcal{S}_2$ *of* $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ *according to* $\widehat{\mathcal{S}_1 \cup \mathcal{S}_2} = \left\{ j \in [p] : \left| \frac{\langle \boldsymbol{X}_j, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| \geq \tau \right\}$*, where* $\tau = \sqrt{2(1+\frac{\epsilon}{2})\log 2p}$ *for some* $\epsilon \in (0,1)$.

Theorem 7 (Success of CORR **on** MSLR − D **outside** SB-MSLR**)** *Consider the general setting of* $\texttt{MSLR} - \texttt{D} \setminus$ SB-MSLR *with* $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \sim \mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$*. Let* $\epsilon \in (0,1)$ *be the parameter used in* CORR *. Then provided*

$$n \geq \frac{32(1+\epsilon)}{\min\{\phi^2 \boldsymbol{\beta}_{min}^2,\ (1-\phi)^2 \boldsymbol{\beta}_{min}^2,\ \langle \boldsymbol{\beta} \rangle_{>0}^2\}} \frac{\|\boldsymbol{\beta}\|_2^2\,(SNR+1)}{SNR} \log 2p,$$

the CORR *algorithm solves strong detection in* $\texttt{MSLR} - \texttt{D} \setminus$ SB-MSLR*.*

The proof of Theorem 7 is given in Appendix D.2. In the natural setting where $\|\boldsymbol{\beta}\|_2^2$ is of order k, the theorem implies that CORR solves MSLR − D outside of the SB-MSLR regime with *square-root* the number of samples implied by the low-degree lower bound in Theorem 4, up to log factors. Indeed, the sample complexity in Theorem 7 matches the optimal sample complexity for the simpler SLR − D problem; see Theorem 10 below. This theorem effectively quantifies the extent to which one can solve MSLR with the sample complexity of SLR. The proof of Theorem 7 also holds in the more general case where $\langle \boldsymbol{\beta} \rangle_{>0}^2 > 0$ and the signal norms $\|\boldsymbol{\beta}_1\|_2, \|\boldsymbol{\beta}_2\|_2$ are not constrained to be equal. For signal priors on the nonzero entries that are absolutely continuous with respect to the Lebesgue measure, the event $\{\langle \boldsymbol{\beta} \rangle_{>0}^2 > 0\}$ has measure one, as $\phi \boldsymbol{\beta}_1 + (1-\phi)\boldsymbol{\beta}_2 \neq 0$ is almost surely satisfied.

Theorem 8 (Sucess of CORR **for recovery in** MSLR **for** $\langle \boldsymbol{\beta} \rangle_{\texttt{min}}^2 > 0$**)** *Consider the general setting of* MSLR *with either* $\sigma = 0, \phi \neq 1/2$ *(noiseless), or* $\phi = 1/2, SNR = \Omega(k)$ *(balanced). Let* $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \sim \mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$*. Let* $\epsilon \in (0,1)$ *be the parameter used in* CORR *, and*

$$n \geq \frac{32(1+\epsilon)}{\min\{\phi^2 \boldsymbol{\beta}_{min}^2, (1-\phi)^2 \boldsymbol{\beta}_{min}^2, \langle \boldsymbol{\beta} \rangle_{min}^2\}} \frac{\|\boldsymbol{\beta}\|_2^2(SNR+1)}{SNR} \log 2p.$$

Then there exists an algorithm which, in combination with CORR *, exactly recovers* $\boldsymbol{\beta}_1$ *and* $\boldsymbol{\beta}_2$ *(up to relabeling) with probability at least* $1 - c_1(\frac{k}{p} + ke^{-c_2 n} + \frac{k}{n} + \frac{1}{p^{c_2}})$ *for constants* $c_1, c_2 > 0$.

The proof of Theorem 8, given in Appendix D.3, first uses CORR for support recovery, followed by existing recovery algorithms for the noiseless and balanced cases of dense ($k/p = \Theta(1)$) mixed linear regression (Yi et al., 2014; Chen et al., 2014). Under the condition $\langle \boldsymbol{\beta} \rangle_{\texttt{min}}^2 > 0$, which is slightly more restrictive than SB-MSLR, Theorem 8 yields a sample complexity of the same order as that for SLR. We note that for signal priors on the nonzero entries that are absolutely continuous with respect to the Lebesgue measure, the event $\{\langle \boldsymbol{\beta} \rangle_{\texttt{min}}^2 > 0\}$ has measure one. We highlight that the noiseless case can be formulated as a mixed variant of compressed sensing with independent Gaussian design (Yu and Sapiro, 2011).

Remark 9 *The restriction to the noiseless and balanced cases in Theorem 8 is due to the guarantees provided by existing algorithms in the dense case, for which experiments indicate success far beyond these regimes (Yi et al., 2014; Chen et al., 2014).*

2.3. Lower bounds for Sparse Linear Regression (SLR)

We define the detection variant of SLR, called $\mathtt{SLR}-\mathtt{D}$, as per Definition 3 with the constraint $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. The following lower bound for $\mathtt{SLR}-\mathtt{D}$ is proved in Appendix B.2.

Theorem 10 (Low-degree lower bound for $\mathtt{SLR}-\mathtt{D}$) *Consider the setting of $\mathtt{SLR}-\mathtt{D}$ (Definition 3 under $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$) with $\boldsymbol{\beta} \sim \mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$. For $n = \omega(\log p)$ and $n \leq (1-\epsilon)(1-2\theta)\frac{\|\boldsymbol{\beta}\|_2^2}{\|\boldsymbol{\beta}\|_\infty^2}\frac{(SNR+1)}{SNR}\log p$ for any $\epsilon \in (0,1)$, Conjecture 2 implies that any randomized algorithm requires running time $e^{\tilde{\Omega}(n)}$ to solve $\mathtt{SLR}-\mathtt{D}$ in the regime $k = O(p^\theta) \leq \sqrt{p}$ with $\theta \in (0, 1/2]$.*

In the natural setting where $\|\boldsymbol{\beta}\|_2^2$ is of order k and the entries have bounded amplitude, the low-degree lower bound on n is of order $\frac{k(\mathtt{SNR}+1)}{\mathtt{SNR}}\log p$. This matches the order of existing lower bounds for SLR in (Bandeira et al., 2022; Gamarnik and Zadik, 2022), but has the advantage of being valid for all SNR regimes and generic priors on the sparse signal $\boldsymbol{\beta}$. We believe that the dependence of the bound on $\|\boldsymbol{\beta}\|_\infty$ is an artifact of the proof technique; see Appendix B.2.

A reduction from $\mathtt{SLR}-\mathtt{D}$ to SLR follows similarly to the reduction from $\mathtt{SB\text{-}MSLR}-\mathtt{D}$ to SB-MSLR, which is given in Appendix C.

2.4. Algorithms for SLR

Theorem 11 *Consider the setting of SLR with $\boldsymbol{\beta} \sim \mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$. Let $\epsilon \in (0,1)$ be the parameter used in CORR . Then for $n \geq \frac{8(1+\epsilon)}{\boldsymbol{\beta}_{\min}^2}\|\boldsymbol{\beta}\|_2^2\frac{(SNR+1)}{SNR}\log 2p$, we have that CORR solves strong detection in* $\mathtt{SLR}-\mathtt{D}$.

We next consider a slight variant of CORR that recovers the *signed* support of $\boldsymbol{\beta}$. It produces $\hat{\boldsymbol{\beta}}$ with entries given by

$$\hat{\boldsymbol{\beta}}_j = \mathbb{1}\left\{\left|\frac{\langle \boldsymbol{X}_j, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \geq \sqrt{2(1+\epsilon/2)\log 2p}\right\} \operatorname{sign}\left(\frac{\langle \boldsymbol{X}_j, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right), \quad \text{for } j \in [p], \tag{3}$$

where $\operatorname{sign}(x)$ equals 1 for $x > 0$, equals -1 for $x < 0$, and 0 for $x = 0$.

Theorem 12 *Consider the setting of SLR with $\boldsymbol{\beta} \sim \mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$. Let $\epsilon \in (0,1)$ be the parameter used in the above variant of CORR . Then for $n \geq \frac{8(1+\epsilon)}{\boldsymbol{\beta}_{\min}^2}\|\boldsymbol{\beta}\|_2^2\frac{(SNR+1)}{SNR}\log 2p$, the vector $\hat{\boldsymbol{\beta}}$ in (3) equals the signed support of $\boldsymbol{\beta}$ with probability at least $1 - (\frac{k}{p} + 2ke^{-c_2 n} + \frac{1}{p^{c_2}})$ for some constant $c_2 > 0$.*

The proofs of Theorem 11 and Theorem 12 are given in Appendix D.4. The sample complexity required for the success of CORR matches the low-degree lower bound in Theorem 10 up to constants, which rigorously certifies the order optimality of CORR among low-degree polynomial algorithms, including spectral methods running in $O(\log p)$ iterations, in all SNR regimes. These achievable sample complexities also match those of previous work (Wainwright, 2009b; Bandeira et al., 2022; Gamarnik and Zadik, 2022; Cai and Wang, 2011; Donoho and Tanner, 2010), with the important extension that they hold for all SNR scalings and general sparse signal priors.

3. Proof Ideas

Low-Degree Lower Bounds Theorem 4 amounts to proving that $\chi^2_{\leq D}(\mathbb{P}(\boldsymbol{X},\boldsymbol{y})\|\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y})) = O(1)$ in $\texttt{MSLR}-\texttt{D}$ (Definition 3) with $\boldsymbol{\beta}_1 = -\boldsymbol{\beta}_2$ and $\phi = 1/2$, for n in the regime specified in the theorem. We rewrite the expression for $\chi^2_{\leq D}(\mathbb{P}(\boldsymbol{X},\boldsymbol{y})\|\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y}))$ in Conjecture 2 in terms of multivariate Hermite polynomials in the data $(\boldsymbol{X},\boldsymbol{y})$ of degree up to D. For $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_{np+n}]$, with $\alpha_i \in \mathbb{N}$, the normalized Hermite polynomial of order $\boldsymbol{\alpha}$ is denoted by $\frac{\tilde{H}_{\boldsymbol{\alpha}}(\boldsymbol{X},\boldsymbol{y})}{\sqrt{\boldsymbol{\alpha}!}}$. The precise definition of the polynomial is given in Appendix B, but the key fact we will use is that $\left\{\frac{\tilde{H}_{\boldsymbol{\alpha}}}{\sqrt{\boldsymbol{\alpha}!}}\right\}$ form an orthonormal system with respect to the null distribution in Definition 3 (see Proposition 18 in Appendix B). Using this, we have

$$\begin{aligned}
\chi^2_{\leq D}(\mathbb{P}(\boldsymbol{X},\boldsymbol{y})\|\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y})) + 1 &= \mathbb{E}_{(\boldsymbol{X},\boldsymbol{y})\sim\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y})}\left(\frac{d\mathbb{P}(\boldsymbol{X},\boldsymbol{y})}{d(\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y}))}\right)^2_{\leq D} \\
&= \sum_{0\leq|\boldsymbol{\alpha}|\leq D}\frac{1}{\boldsymbol{\alpha}!}\mathop{\mathbb{E}}_{\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y})}\left[\frac{d\mathbb{P}(\boldsymbol{X},\boldsymbol{y})}{d(\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y}))}\tilde{H}_{\boldsymbol{\alpha}}(\boldsymbol{X},\boldsymbol{y})\right]^2 \\
&= \sum_{0\leq|\boldsymbol{\alpha}|\leq D}\frac{1}{\boldsymbol{\alpha}!}\mathop{\mathbb{E}}_{\mathbb{P}(\boldsymbol{X},\boldsymbol{y})}\left[\tilde{H}_{\boldsymbol{\alpha}}(\boldsymbol{X},\boldsymbol{y})\right]^2, \qquad (4)
\end{aligned}$$

where $\boldsymbol{\alpha}! = \prod_i \alpha_i!$. The key element of the proof involves subsequently upper bounding (4) through Hermite polynomial identities and multinomial-theorem manipulations, yielding a weighted sum over D moments of the overlap $\langle\boldsymbol{\beta}_1^{(1)}, \boldsymbol{\beta}_1^{(2)}\rangle$, where $\boldsymbol{\beta}_1^{(1)}, \boldsymbol{\beta}_1^{(2)}$ are two i.i.d copies of the signal $\boldsymbol{\beta}_1$ (see Lemma 24). Each of these D moments can be bounded for $k \leq \sqrt{p}$, allowing the entire sum over $D \simeq \min\left\{\frac{k^2(\text{SNR}+1)^2}{n\text{SNR}^2}, n\right\}$ terms to converge, and yielding the result. The case of $\texttt{SLR}$ in Theorem 10 is similar, but with the simplification $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, we can afford to set $D \simeq n$ and still have this sum converge, yielding the key difference in lower bounds between $\texttt{SLR}$ and $\texttt{SB-MSLR}$.

Reductions from detection to recovery We follow the procedure for average-case reductions outlined by Brennan and Bresler (2020b). We transfer computational hardness from $\texttt{SB-MSLR}-\texttt{D}$ to recovery in $\texttt{SB-MSLR}$ by forming an average-case reduction for k-sparse signals in $\{-1,0,1\}^p$. Denote the parameter regime of Theorem 4 as the "critical" parameter regime. Given any sequence of parameters $\mathcal{P}$ in the critical regime, we construct another sequence of parameters $\mathcal{P}'$ in the critical regime with the following property: if there exists a randomized polynomial-time algorithm $\mathcal{A}'$ solving exact recovery in $\texttt{PSB-MSLR}$ with parameter scaling $\mathcal{P}'$, then we can construct a randomized polynomial-time algorithm solving $\texttt{SB-MSLR}-\texttt{D}$ with parameter scaling $\mathcal{P}$. This would in turn contradict Theorem 4, implying computational hardness of exact recovery in $\texttt{PSB-MSLR}$ in the critical regime. We first provide an average case reduction from $\texttt{SB-MSLR}-\texttt{D}$ to exact recovery in $\texttt{SB-MSLR}$ in Lemma 26, and then reduce exact recovery in $\texttt{SB-MSLR}$ to exact recovery in $\texttt{PSB-MSLR}$ in Theorem 28.

The $\texttt{CORR}$ algorithm For $\texttt{MSLR}$, the proofs of Theorems 7 and 8 crucially rely on Theorem 37, which shows that $\texttt{CORR}$ recovers the joint support of the signals $(\mathcal{S}_1\cup\mathcal{S}_2)$ if n satisfies the condition in Theorem 8. To prove Theorem 37, we analyze the quantity $u_j := \frac{\langle\boldsymbol{X}_j,\boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}$ in three cases. When $j \in (\mathcal{S}_1\cup\mathcal{S}_2)^{\complement}$, we have that $u_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ for $j\in[p]$ by the independence of $\boldsymbol{X}_j$ and $\boldsymbol{y}$. The typical value of $\max_{j\in[p]} u_j$ in this case is $\sqrt{2\log p}$, and we can bound the probability of a false positive by standard concentration bounds, detailed in Lemma 45. When $j \in \mathcal{S}_1 \cap \mathcal{S}_2$, we show that conditioned on $\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2$, u_j is normally distributed with mean

$$\mathbb{E}\left[u_j \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right] = \frac{\|\boldsymbol{y}_{\{\boldsymbol{z}=1\}}\|_2^2 \boldsymbol{\beta}_{1,j} + \|\boldsymbol{y}_{\{\boldsymbol{z}=0\}}\|_2^2 \boldsymbol{\beta}_{2,j}}{\|\boldsymbol{y}\|_2(\|\boldsymbol{\beta}\|_2^2 + \sigma^2)}, \tag{5}$$

and variance less than 1 (Lemma 40). Here, $\|\boldsymbol{y}_{\{\boldsymbol{z}=1\}}\|_2$ denotes the norm of the vector with entries $(y_i 1_{\{z_i=1\}})_{i\in[n]}$. For large n, p and $j \in [p]$, the typical value of the conditional mean above is $\sqrt{\frac{n}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}}\,(\phi\boldsymbol{\beta}_{1,j} + (1-\phi)\boldsymbol{\beta}_{2,j})$, which is greater than $\sqrt{2(1+\epsilon/2)\log 2p}$ for

$$n \geq \frac{2(1+\epsilon)}{\langle\boldsymbol{\beta}\rangle_{\mathtt{min}}^2}(\|\boldsymbol{\beta}\|_2^2 + \sigma^2)\log 2p \simeq (1+\epsilon)\frac{k(\mathtt{SNR}+1)}{\langle\boldsymbol{\beta}\rangle_{\mathtt{min}}^2\mathtt{SNR}}\log 2p.$$

The remaining case $j \in \mathcal{S}_1 \Delta \mathcal{S}_2$ is similar, with the conditional mean obtained by setting $\boldsymbol{\beta}_{2,j} = 0$ in (5). The results for SLR in Theorems 11, 12 follow a similar reasoning, with $\mathbb{E}\left[u_j \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right] = \frac{\boldsymbol{\beta}_j\|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}$.

4. Discussion

In this work we rigorously characterize the computational hardness of Mixed Sparse Linear Regression (MSLR) through the method of low-degree polynomials. We evidence that in the highly symmetric SB-MSLR regime, randomized polynomial-time algorithms cannot solve an associated detection problem with sample complexity $n = \tilde{o}\left(\frac{k^2(\mathtt{SNR}+1)^2}{\mathtt{SNR}^2}\right)$, revealing a statistical-computational gap of order at least k. Outside of the SB-MSLR regime, however, a simple polynomial-time algorithm CORR succeeds in solving detection with minimal sample complexity.

We note that our low-degree statistical-computational gap for SB-MSLR persists even in the noiseless ($\mathtt{SNR} = \infty$) regime. Recent discoveries have highlighted that evidence for statistical-to-computational gaps do not always hold in the noiseless setting. Examples include "brittle" algorithms such as Gaussian elimination "breaking" the statistical-to-computational gap in learning parities (Zadik et al., 2022). It was also recently found in Zadik et al. (2022) that the LLL family of algorithms, originating from cryptography, can break the statistical-to-computational gaps predicted in certain noiseless clustering problems. Further, in a recent talk by Zadik (2021), a proof sketch was presented for a lattice-based algorithm that can recover $\boldsymbol{\beta} \in \mathbb{R}^p$ in *dense* noiseless phase retrieval with $p+1$ measurements — this breaks a conjectured statistical-to-computational gap for dense phase retrieval, but the algorithm does not capture the sparse problem structure present in MSLR. To the best of our knowledge, noiseless inference in SB-MSLR and sparse phase retrieval still cannot be achieved with fewer than order $k^2 \log p$ samples for the case of k-sparse $\boldsymbol{\beta} \in \{-1, 0, 1\}^p$. Such an achievement, if possible, would constitute an interesting and novel contribution.

Acknowledgments

GA was supported by a Cambridge Trust scholarship and funding from Invenia Labs. GA thanks Alex Wein, Ilias Zadik, Afonso Bandeira, Nicolò Grometto, Daniil Dmitriev for helpful conversations regarding statistical-computational gaps in sparse linear regression and phase retrieval.

References

Ery Arias-Castro, Emmanuel J. Candès, and Yaniv Plan. Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556, 2011.

Gabriel Arpino. Computational Hardness of Sparse High-Dimensional Linear Regression, 2021. ETH-Zürich Master's Thesis.

Martin Azizyan, Aarti Singh, and Larry Wasserman. Minimax Theory for High-dimensional Gaussian Mixtures with Sparse Mean Separation. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Afonso S. Bandeira, Dmitriy Kunisky, and Alexander S. Wein. Computational Hardness of Certifying Bounds on Constrained PCA Problems. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*, volume 151, pages 78:1–78:29, 2020.

Afonso S. Bandeira, Ahmed El Alaoui, Samuel B. Hopkins, Tselil Schramm, Alexander S. Wein, and Ilias Zadik. The Franz-Parisi Criterion and Computational Trade-offs in High Dimensional Statistics. In *Advances in Neural Information Processing Systems*, 2022.

Boaz Barak, Samuel B. Hopkins, Jonathan Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin. A Nearly Tight Sum-of-Squares Lower Bound for the Planted Clique Problem. In *Proceedings - 57th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2016*, pages 428–437, December 2016.

Adarsh Barik and Jean Honorio. Sparse mixed linear regression with guarantees: Taming an intractable problem with invex relaxation. In *International Conference on Machine Learning*, volume 162, pages 1627–1646, 2022.

Quentin Berthet and Philippe Rigollet. Complexity Theoretic Lower Bounds for Sparse Principal Component Detection. In *Proceedings of the 26th Annual Conference on Learning Theory*, pages 1046–1066, 2013.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Matthew Brennan and Guy Bresler. Average-Case Lower Bounds for Learning Sparse Mixtures, Robust Estimation and Semirandom Adversaries. Technical report, 2020a. arXiv:1908.06130.

Matthew Brennan and Guy Bresler. Reducibility and Statistical-Computational Gaps from Secret Leakage. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 648–847, 2020b.

Matthew S. Brennan, Guy Bresler, Sam Hopkins, Jerry Li, and Tselil Schramm. Statistical Query Algorithms and Low Degree Tests Are Almost Equivalent. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pages 774–774, 2021.

T. Tony Cai and Lie Wang. Orthogonal Matching Pursuit for Sparse Signal Recovery With Noise. *IEEE Transactions on Information Theory*, 57(7):4680–4688, 2011.

T. Tony Cai, Xiaodong Li, and Zongming Ma. Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *The Annals of Statistics*, October 2016. Publisher: Institute of Mathematical Statistics.

E.J. Candes and T. Tao. Decoding by Linear Programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase Retrieval via Wirtinger Flow: Theory and Algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

Arun Tejasvi Chaganty and Percy Liang. Spectral Experts for Estimating Mixtures of Linear Regressions. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1040–1048, May 2013.

Sitan Chen, Jerry Li, and Zhao Song. Learning mixtures of linear regressions in subexponential time via Fourier moments. *Proceedings of the 52nd Annual ACM Symposium on Theory of Computing*, pages 587–600, 2020.

Yudong Chen, Xinyang Yi, and Constantine Caramanis. A Convex Formulation for Mixed Regression with Two Components: Minimax Optimal Rates. In *Proceedings of The 27th Conference on Learning Theory*, pages 560–604, 2014.

Amin Coja-Oghlan, Oliver Gebhard, Max Hahn-Klimroth, Alexander S. Wein, and Ilias Zadik. Statistical and Computational Phase Transitions in Group Testing. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 4764–4781, June 2022.

Partha Deb and Ann M. Holmes. Estimates of use and costs of behavioural health care: a comparison of standard and finite mixture models. *Health Economics*, 9(6):475–489, September 2000.

Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754, USA, January 2019.

Yunzi Ding, Dmitriy Kunisky, Alexander S. Wein, and Afonso S. Bandeira. Subexponential-Time Algorithms for Sparse PCA. *Foundations of Computational Mathematics*, 2023.

David L. Donoho and Jared Tanner. Counting the Faces of Randomly-Projected Hypercubes and Orthants, with Applications. *Discrete & Computational Geometry*, 43(3):522–541, April 2010.

D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, April 2006.

Jianqing Fan, Han Liu, Zhaoran Wang, and Zhuoran Yang. Curse of Heterogeneity: Computational Barriers in Sparse Mixture Models and Phase Retrieval. Technical report, arXiv, August 2018. arXiv:1808.06996.

Susana Faria and Gilda Soromenho. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, 2010.

Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proceedings of the forty-fifth annual ACM symposium on Theory of Computing*, pages 655–664, June 2013.

David Gamarnik and Madhu Sudan. Limits of Local Algorithms Over Sparse Random Graphs. *The Annals of Probability*, 45(4):2353–2376, 2017.

David Gamarnik and Ilias Zadik. High dimensional regression with binary coefficients. estimating squared error and a phase transtition. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 948–953, June 2017.

David Gamarnik and Ilias Zadik. Sparse high-dimensional linear regression. Estimating squared error and a phase transition. *The Annals of Statistics*, 50(2):880–903, April 2022.

Avishek Ghosh and Ramchandran Kannan. Alternating Minimization Converges Super-Linearly for Mixed Linear Regression. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020.

Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman & Hall/CRC monographs on statistics and applied probability. CRC Press, second edition edition, 2021.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Number 143. CRC Press, Taylor & Francis Group, 2015.

Samuel B. Hopkins and David Steurer. Efficient Bayesian Estimation from Few Samples: Community Detection and Related Problems. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 379–390, October 2017.

Samuel B. Hopkins, Pravesh K. Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The Power of Sum-of-Squares for Detecting Hidden Structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 720–731, October 2017.

Samuel Brink Klevit Hopkins. Statistical Inference and the Sum of Squares Method. 2018. Cornell University Library.

Yunju Im, Yuan Huang, Jian Huang, and Shuangge Ma. Bayesian hierarchical finite mixture of regression for histopathological imaging-based cancer data analysis. *Statistics in Medicine*, 41 (6):1009–1022, 2022.

Amnon Jakimovski, Ambikeshwar Sharma, and József Szabados. Hermite and Hermite-Birkhoff Interpolation and Walsh Equiconvergence. In *Walsh Equiconvergence of Complex Interpolating Polynomials*, pages 25–54. Springer Monographs in Mathematics, 2006.

Lucas Janson, Rina Foygel Barber, and Emmanuel Candès. EigenPrism: inference for high dimensional signal-to-noise ratios. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1037–1065, 2017.

Adel Javanmard, Simeng Shao, and Jacob Bien. Prediction Sets for High-Dimensional Mixture of Experts Models, October 2022. arXiv:2210.16710.

Michael I. Jordan and Robert A. Jacobs. Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation*, 6(2):181–214, March 1994.

Karen Kazor and Amanda S. Hering. Mixture of Regression Models for Large Spatial Datasets. *Technometrics*, 61(4):507–523, 2019.

Abbas Khalili and Jiahua Chen. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102(479):1025–1038, 2007.

Florent Krzakala, Andrea Montanari, Federico Ricci-Tersenghi, Guilhem Semerjian, and Lenka Zdeborova. Gibbs states and the set of solutions of random constraint satisfaction problems. *Proceedings of the National Academy of Sciences*, 104:10318–10323, June 2007.

Dmitriy Kunisky, Alexander S. Wein, and Afonso S. Bandeira. Notes on Computational Hardness of Hypothesis Testing: Predictions Using the Low-Degree Likelihood Ratio. In *Mathematical Analysis, its Applications and Computation*, Springer Proceedings in Mathematics & Statistics, pages 1–50, 2022.

Guillaume Lecué and Shahar Mendelson. Minimax rate of convergence and the performance of empirical risk minimization in phase recovery. *Electronic Journal of Probability*, 2015.

Yan Li, Chun Yu, Yize Zhao, Weixin Yao, Robert H. Aseltine, and Kun Chen. Pursuing sources of heterogeneity in modeling clustered population. *Biometrics*, 78(2):716–729, 2022.

Yuanzhi Li and Yingyu Liang. Learning mixtures of linear regressions with nearly optimal complexity. *Conference On Learning Theory*, pages 1125–1144, 2018.

Nathan Linial. Locality in Distributed Graph Algorithms. *SIAM Journal on Computing*, 21(1): 193–201, February 1992.

Zhaoqiang Liu, Subhroshekhar Ghosh, and Jonathan Scarlett. Towards Sample-Optimal Compressive Phase Retrieval with Sparse and Generative Priors. In *Advances in Neural Information Processing Systems*, 2021.

Ziye Luo, Xinyue Yao, Yifan Sun, and Xinyan Fan. Regression-based heterogeneity analysis to identify overlapping subgroup structure in high-dimensional data. *Biometrical Journal*, 64(6): 1109–1141, 2022.

Ashok Makkuva, Pramod Viswanath, Sreeram Kannan, and Sewoong Oh. Breaking the gridlock in Mixture-of-Experts: Consistent and Efficient Algorithms. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4304–4313, May 2019.

Arya Mazumdar and Soumyabrata Pal. On Learning Mixture Models with Sparse Parameters. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 9182–9213, May 2022.

Andrea Montanari and Alexander S. Wein. Equivalence of Approximate Message Passing and Low-Degree Polynomials in Rank-One Matrix Estimation, December 2022. arXiv:2212.06996.

Andrea Montanari, Daniel Reichman, and Ofer Zeitouni. On the Limitation of Spectral Methods: From the Gaussian Hidden Clique Problem to Rank-One Perturbations of Gaussian Tensors. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Keith B. Oldham, Jan C. Myland, Jerome Spanier, and Jan Myland. *An Atlas of functions: with equator, the atlas function calculator*. Mathematics and Statistics. Springer US Springer e-books, New York, NY, 2009. ISBN 978-0-387-48807-3.

Soumyabrata Pal, Arya Mazumdar, and Venkata Gandikota. Support recovery of sparse signals from a mixture of linear measurements. In *Advances in Neural Information Processing Systems*, 2021.

Soumyabrata Pal, Arya Mazumdar, Rajat Sen, and Avishek Ghosh. On Learning Mixture of Linear Regressions in the Non-Realizable Setting. In *Proceedings of the 39th International Conference on Machine Learning*, pages 17202–17220, 2022.

Richard E. Quandt and James B. Ramsey. Estimating Mixtures of Normal Distributions and Switching Regressions. *Journal of the American Statistical Association*, 73(364):730–738, December 1978.

Galen Reeves, Jiaming Xu, and Ilias Zadik. The All-or-Nothing Phenomenon in Sparse Linear Regression. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 2652–2663, June 2019.

Tselil Schramm and Alexander S. Wein. Computational Barriers to Estimation from Low-Degree Polynomials. *The Annals of Statistics*, 50(3), June 2022.

Yanyao Shen and Sujay Sanghavi. Iterative least trimmed squares for mixed linear regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Nicolas Städler, Peter Bühlmann, and Sara van de Geer. l1-penalization for mixture regression models. *TEST*, 19(2):209–256, August 2010.

Ramon van Handel. Probability in High Dimension. Technical report, Defense Technical Information Center, Fort Belvoir, VA, June 2014.

Kert Viele and Barbara Tong. Modeling with Mixtures of Linear Regressions. *Statistics and Computing*, 12:315–330, October 2002.

Martin J. Wainwright. Information-Theoretic Limits on Sparsity Recovery in the High-Dimensional and Noisy Setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741, December 2009a.

Martin J. Wainwright. Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using ℓ_1-Constrained Quadratic Programming (Lasso). *IEEE Transactions on Information Theory*, 55 (5):2183–2202, May 2009b.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 1 edition, February 2019.

Wei Wang, Martin J. Wainwright, and Kannan Ramchandran. Information-Theoretic Limits on Sparse Signal Recovery: Dense versus Sparse Measurement Matrices. *IEEE Transactions on Information Theory*, 56(6):2967–2979, June 2010.

Michel Wedel and Wagner A Kamakura. *Market Segmentation: Conceptual and Methodological Foundations*. Springer Conceptual and Methodological Foundations, 2000.

Fan Wu and Patrick Rebeschini. Hadamard Wirtinger Flow for Sparse Phase Retrieval. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 982–990, 2021.

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pages 613–621, 2014.

Guoshen Yu and Guillermo Sapiro. Statistical Compressed Sensing of Gaussian Mixture Models. *IEEE Transactions on Signal Processing*, 59(12):5842–5858, December 2011. ISSN 1941-0476. doi: 10.1109/TSP.2011.2168521. Conference Name: IEEE Transactions on Signal Processing.

Ilias Zadik. The surprising power of the Lenstra-Lenstra-Lovasz algorithm for noiseless inference. Simons Institute IFML/CCSI Symposium, 2021. URL https://youtu.be/QGUEaXII4Kw.

Ilias Zadik, Min Jae Song, Alexander S. Wein, and Joan Bruna. Lattice-Based Methods Surpass Sum-of-Squares in Clustering. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 1247–1248, June 2022.

Lenka Zdeborova and Florent Krzakala. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65(5):453–552, September 2016.

Yihan Zhang, Marco Mondelli, and Ramji Venkataramanan. Precise Asymptotics for Spectral Methods in Mixed Generalized Linear Models, December 2022. arXiv:2211.11368.

Kai Zhong, Prateek Jain, and Inderjit S. Dhillon. Mixed linear regression with multiple components. In *Advances in Neural Information Processing Systems*, pages 2190–2198, 2016.

Contents

Appendix A. Additional background on the Low Degree Method

In this section, we give additional background on the low-degree method, the chi-squared divergence and its orthogonal projection onto the space of low-degree polynomials. Consider the setting in Section 1.3, where the task is to distinguish between two probability distributions $\mathbb{P}_N$ and $\mathbb{Q}_N$ over $\mathbb{R}^N$ where N is the (potentially growing) problem size. Given a sample $\boldsymbol{x}$ drawn from $\mathbb{P}_N$ or $\mathbb{Q}_N$, one seeks to identify whether $\boldsymbol{x}$ originated from the former or the latter through a hypothesis test. Recall the notions of strong and weak detection from Section 1.3.

One powerful method of identifying whether strong or weak detection is possible is through the study of the *chi-squared divergence* $\chi^2(\mathbb{P}_N \| \mathbb{Q}_N)$. Indeed, assume that $\mathbb{P}_N$ is absolutely continuous with respect to $\mathbb{Q}_N$, and let $\mathsf{L} = \frac{d\mathbb{P}_N}{d\mathbb{Q}_N}$ be the likelihood ratio. We have:

$$\begin{aligned}\chi^2(\mathbb{P}_N \| \mathbb{Q}_N) &:= \mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}_N} \mathsf{L}(\boldsymbol{x})^2 - 1 \\ &= \sup_{f:\mathbb{R}^p\to\mathbb{R}} \frac{(\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_N} f(\boldsymbol{x}))^2}{\mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}_N} f(\boldsymbol{x})^2} - 1 \\ &= \sup_{\substack{f:\mathbb{R}^p\to\mathbb{R} \\ \mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}_N} f(\boldsymbol{x})=0}} \frac{(\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_N} f(\boldsymbol{x}))^2}{\mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}_N} f(\boldsymbol{x})^2},\end{aligned}$$

where the equivalences follow from standard arguments (see Kunisky et al. (2022)). Interpreting the above result, the chi-squared divergence represents optimality in the L^2 sense. It relates to the squared maximum expectation any function can have under $\mathbb{P}_N$, while still being bounded in the space $L^2(\mathbb{Q}_N)$. In fact, the chi-square divergence between two distributions can rigorously characterize their behaviour under testing:

Lemma 13 (Adapted from Lemma 2 of Montanari et al. (2015) and Lemma 7.1 of Coja-Oghlan et al. (2022))

- *If $\chi^2(\mathbb{P}_N \| \mathbb{Q}_N) = O(1)$ as $N \to \infty$, then strong detection is impossible.*
- *If $\chi^2(\mathbb{P}_N \| \mathbb{Q}_N) = o(1)$ as $N \to \infty$, then weak detection is impossible.*

This result is powerful, as it identifies the chi-square divergence as a sufficient quantity for finding identifying results in testing. Note however, that this quantity reveals nothing with regards to *computation*.

The computational analogue of the chi-square divergence is the *degree-D chi-square divergence* $\chi^2_{\leq D}(\mathbb{P}_N \| \mathbb{Q}_N)$. This quantity measures whether $\mathbb{P}_N$ and $\mathbb{Q}_N$ can be distinguished by a degree-D polynomial of the input $\boldsymbol{x}$. Consider the Hilbert Space $L^2(\mathbb{Q}_N)$, where for functions $f, g : \mathbb{R}^p \to \mathbb{R}$ we have the inner product $\langle f, g\rangle := \mathbb{E}_{\boldsymbol{x}\sim Q_N}[f(\boldsymbol{x})g(\boldsymbol{x})]$ and the corresponding norm $\|f\|_{\mathbb{Q}_N} = \sqrt{\langle f, f\rangle_{\mathbb{Q}_N}}$. Additionally, denote $\mathbb{R}[\boldsymbol{x}]_{\leq D}$ as the space of multivariate polynomials from $\mathbb{R}^p$ to $\mathbb{R}$ of degree at most D, and let $f^{\leq D}$ denote the orthogonal projection of f onto $R[\boldsymbol{x}]_{\leq D}$ in $L^2(\mathbb{Q}_N)$. We can then define $\chi^2_{\leq D}(\mathbb{P}_N \| \mathbb{Q}_N)$ as follows:

$$\begin{aligned}\chi^2_{\leq D}(\mathbb{P}_N \| \mathbb{Q}_N) &:= \mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}_N} \mathsf{L}^{\leq D}(\boldsymbol{x})^2 - 1 \\ &= \|\mathsf{L}^{\leq D}\|^2_{Q_N} - 1 \\ &= \sup_{f\in\mathbb{R}[\boldsymbol{x}]_{\leq D}} \frac{(\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_N} f(\boldsymbol{x}))^2}{\mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}_N} f(\boldsymbol{x})^2} - 1 \\ &= \sup_{\substack{f\in\mathbb{R}[\boldsymbol{x}]_{\leq D} \\ \mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}_N} f(\boldsymbol{x})=0}} \frac{(\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_N} f(\boldsymbol{x}))^2}{\mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}_N} f(\boldsymbol{x})^2}.\end{aligned} \tag{6}$$

The proof of this result can be found in Hopkins (2018); Kunisky et al. (2022). The low-degree chi-square divergence can therefore interpreted analogously to chi-square divergence: it quantifies the maximum expectation any low-degree function can have under $\mathbb{P}_N$ while still being in the degree-D polynomial subspace of $L^2(\mathbb{Q}_N)$. We then have the analogue of Lemma 13 for low-degree polynomial functions of the input and, conjecturally, general polynomial-time algorithms, given by Conjecture 2.

In this work, we consider testing between distributions that do not simply consist of *signal plus noise*, but instead of linearly transformed signals plus noise. Along with the recent work in Bandeira et al. (2022); Arpino (2021), this is, to the best of our knowledge, among the first applications of the low-degree method to such problems, which were previously believed to be out of reach from current methods (Schramm and Wein, 2022).

Appendix B. Proofs of Low-Degree Lower Bounds

Preliminaries and Notation. All results concerning the low-degree hardness of the associated problems are asymptotic in p, as we take $p \to \infty$ first. We use the conventions from Schramm and Wein (2022). Let $\mathbb{N} = \{0, 1, 2, \cdots\}$ and $[n] = \{1, 2 \cdots n\}$. We define $0^0 := 1$. We denote by boldface a multiset or vector, so for $\boldsymbol{\alpha} \in \mathbb{N}^n$ we mean $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_n]$ for $\alpha_i \in \mathbb{N}, \forall i \in [n]$. For $\boldsymbol{\alpha} \in \mathbb{N}^n$, define $|\boldsymbol{\alpha}| = \sum_{i=1}^n \alpha_i$, $\boldsymbol{\alpha}! = \prod_{i=1}^n \alpha_i!$ and (for $\boldsymbol{X} \in \mathbb{R}^n$) $\boldsymbol{X}^{\boldsymbol{\alpha}} = \prod_{i=1}^n X_i^{\alpha_i}$. Let $\texttt{abs}(\boldsymbol{\alpha})$ denote the entry-wise absolute value operation on the vector $\boldsymbol{\alpha}$. We use $\boldsymbol{\alpha} \geq \boldsymbol{\beta}$ to mean $\alpha_i \geq \boldsymbol{\beta}_i$ for all i. The operations $\boldsymbol{\alpha} + \boldsymbol{\beta}$ and $\boldsymbol{\alpha} - \boldsymbol{\beta}$ are performed entrywise. For $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}^n$ with $\boldsymbol{\alpha} \geq \boldsymbol{\beta}$, define $\binom{\boldsymbol{\alpha}}{\boldsymbol{\beta}} = \prod_{i=1}^n \binom{\alpha_i}{\boldsymbol{\beta}_i}$. We use subindices to denote subsets of a vector or multiset, so for $\boldsymbol{\alpha} \in \mathbb{N}^{n \times (p+1)}$, we let $\boldsymbol{\alpha}_{p+1} := [\alpha_{1,p+1}, \ldots, \alpha_{n,p+1}]$ denote the $p+1$th column of the matrix $\boldsymbol{\alpha}$. We let $\boldsymbol{\alpha}_{\cdot,:p}$ denote the entire $n \times p$ submatrix obtained by selecting only up to the pth column, and $\boldsymbol{\alpha}_{i,:p}$ the vector consisting of elements from the ith row up to the pth column. We denote by $[\boldsymbol{A} \;\; \boldsymbol{y}]$ the matrix formed through the horizontal concatenation of $\boldsymbol{y} \in \mathbb{R}^n$ onto $\boldsymbol{A} \in \mathbb{R}^{n \times p}$, forming an $n \times (p+1)$ real matrix. Unless otherwise indicated, we let $\|\cdot\| := \|\cdot\|_{\mathbb{Q}_p}$ and $\langle \cdot, \cdot \rangle := \langle \cdot, \cdot \rangle_{\mathbb{Q}_p}$. We use $\mathbb{1}$ to denote the indicator function.

The univariate Hermite polynomials $H_k(x)$ for $k \geq 0$ are defined by the recursion $H_0(x) = 1$, and $H_{k+1}(x) = xH_k(x) - H_k'(x)$. For $\boldsymbol{\alpha} \in \mathbb{N}^N$, let $H_{\boldsymbol{\alpha}}$ denote the *multivariate* Hermite polynomial of order $\boldsymbol{\alpha}$, defined as $H_{\boldsymbol{\alpha}}(\boldsymbol{u}) = \prod_{i=1}^N H_{\alpha_i}(u_i)$, for $\boldsymbol{u} \in \mathbb{R}^N$. For $N \in \mathbb{N}$, the normalized N-variate Hermite polynomials $\frac{1}{\sqrt{\boldsymbol{\alpha}!}} H_{\boldsymbol{\alpha}}$ form a complete orthonormal system of (multivariate) polynomials for $L^2(\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N))$ (see Kunisky et al. (2022)).

In what follows, we give further basic facts regarding Hermite polynomials (see Kunisky et al. (2022) for more detailed descriptions), along with two auxiliary combinatorial lemmas that will be of use for the main proofs.

Proposition 14 (Gaussian Integration by Parts, Prop. 2.10 in Kunisky et al. (2022)) *If $f : \mathbb{R} \to \mathbb{R}$ is k-times continuously differentiable and $f(y)$ and its first k derivatives are bounded by $\mathcal{O}(\exp(|y|^\alpha))$ for some $\alpha \in (0, 2)$, then*

$$\mathbb{E}_{y \sim \mathcal{N}(0,1)} \left[H_k(y) f(y) \right] = \mathbb{E}_{y \sim \mathcal{N}(0,1)} \left[\frac{d^k f}{dy^k}(y) \right].$$

Proposition 15 (Hermite derivative (Jakimovski et al., 2006)) *For $n \in \mathbb{N}$, $m \in \mathbb{N}$:*

$$H_n^{(m)}(x) = \frac{n!}{(n-m)!} H_{n-m}(x).$$

Proposition 16 (Hermite sum formula, Prop 3.1 in Schramm and Wein (2022)) *For any $k \in \mathbb{N}$ and $z, \mu \in \mathbb{R}$,*

$$H_k(z+\mu) = \sum_{l=0}^{k} \binom{k}{l} \mu^{k-l} H_l(z).$$

Proposition 17 (Hermite multiplication formula Oldham et al. (2009)) *For $\gamma \in \mathbb{R}$,*

$$H_n(\gamma x) = \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \gamma^{n-2i} \left(\gamma^2 - 1\right)^i \binom{n}{2i} \frac{(2i)!}{i!} 2^{-i} H_{n-2i}(x).$$

Proposition 18 *Consider the null distribution $\mathbb{P}(\boldsymbol{X}) \otimes \mathbb{P}(\boldsymbol{y})$ whose law given by*

$$\mathcal{N}(0,1)^{\otimes n \times p} \otimes \mathcal{N}\left(0, \frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1\right)^{\otimes (p+1)}.$$

Let $\boldsymbol{u} = [\boldsymbol{X} \ \boldsymbol{y}] \in \mathbb{R}^{N \times (p+1)}$. Then, an orthonormal system with respect to this null distribution, indexed by $\boldsymbol{\alpha} \in \mathbb{N}^{n \times (p+1)}$, is given by

$$\frac{1}{\sqrt{\boldsymbol{\alpha}!}} \tilde{H}_{\boldsymbol{\alpha}}(\boldsymbol{u}) := \frac{1}{\sqrt{\boldsymbol{\alpha}!}} \prod_{i=1}^{n} \prod_{j=1}^{p} H_{\alpha_{i,j}}(u_{i,j}) H_{\alpha_{i,p+1}}\left(\frac{u_{i,p+1}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}+1}}\right).$$

Proof Let $\boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)} \in \mathbb{N}^{n \times (p+1)}$. Then,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y})} \frac{1}{\sqrt{\boldsymbol{\alpha}^{(1)}!}} \tilde{H}_{\boldsymbol{\alpha}^{(1)}}(\boldsymbol{u}) \frac{1}{\sqrt{\boldsymbol{\alpha}^{(2)}!}} \tilde{H}_{\boldsymbol{\alpha}^{(2)}}(\boldsymbol{u}) \\
&= \frac{1}{\sqrt{\boldsymbol{\alpha}^{(1)}!\boldsymbol{\alpha}^{(2)}!}} \mathbb{E}_{\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y})} \prod_{i=1}^{n} \prod_{j=1}^{p} H_{\alpha^{(1)}_{i,j}}(u_{i,j}) H_{\alpha^{(1)}_{i,p+1}}\left(\frac{u_{i,p+1}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}+1}}\right) H_{\alpha^{(2)}_{i,j}}(u_{i,j}) H_{\alpha^{(2)}_{i,p+1}}\left(\frac{u_{i,p+1}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}+1}}\right) \\
&= \frac{1}{\sqrt{\boldsymbol{\alpha}^{(1)}!\boldsymbol{\alpha}^{(2)}!}} \mathbb{E} \prod_{i=1}^{n} \prod_{j=1}^{p} H_{\alpha^{(1)}_{i,j}}(u_{i,j}) H_{\alpha^{(2)}_{i,j}}(u_{i,j}) \, \mathbb{E} \prod_{i=1}^{n} H_{\alpha^{(1)}_{i,p+1}}\left(\frac{u_{i,p+1}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}+1}}\right) H_{\alpha^{(2)}_{i,p+1}}\left(\frac{u_{i,p+1}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}+1}}\right) \\
&= \frac{1}{\sqrt{\boldsymbol{\alpha}^{(1)}!\boldsymbol{\alpha}^{(2)}!}} \mathbb{E} \prod_{i=1}^{n} \prod_{j=1}^{p} H_{\alpha^{(1)}_{i,j}}(u_{i,j}) H_{\alpha^{(2)}_{i,j}}(u_{i,j}) \, \mathbb{E} \prod_{i=1}^{n} H_{\alpha^{(1)}_{i,p+1}}(w_i) H_{\alpha^{(2)}_{i,p+1}}(w_i) \\
&= \frac{1}{\sqrt{\boldsymbol{\alpha}^{(1)}!\boldsymbol{\alpha}^{(2)}!}} \sqrt{\boldsymbol{\alpha}^{(1)}!\boldsymbol{\alpha}^{(2)}!} \prod_{i=1}^{n} \prod_{j=1}^{p} \mathbb{1}_{\{\alpha^{(1)}_{i,j} = \alpha^{(2)}_{i,j}\}} \prod_{i=1}^{n} \mathbb{1}_{\{\alpha^{(1)}_{i,p+1} = \alpha^{(2)}_{i,p+1}\}} \\
&= \mathbb{1}_{\boldsymbol{\alpha}^{(1)} = \boldsymbol{\alpha}^{(2)}},
\end{aligned}$$

where $w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ are independent of all other variables for $i \in [n]$. ■

Lemma 19 *For $\boldsymbol{\beta} \in \mathbb{N}$ even:*

$$\sum_{\xi=0}^{\frac{\boldsymbol{\beta}}{2}} \binom{\boldsymbol{\beta}}{2\xi} \frac{(2\xi)!}{\xi!} \left(-\frac{1}{2}\right)^{\xi} (\boldsymbol{\beta} - 2\xi - 1)!! = \mathbb{1}_{\boldsymbol{\beta}=0}.$$

Proof We have:

$$\begin{aligned}
\sum_{\xi=0}^{\frac{\boldsymbol{\beta}}{2}} \binom{\boldsymbol{\beta}}{2\xi} \frac{(2\xi)!}{\xi!} \left(-\frac{1}{2}\right)^{\xi} (\boldsymbol{\beta} - 2\xi - 1)!! &= \sum_{\xi=0}^{\frac{\boldsymbol{\beta}}{2}} \frac{\boldsymbol{\beta}!}{\xi!(\boldsymbol{\beta} - 2\xi)!!} \left(\frac{-1}{2}\right)^{\xi} \\
&= \sum_{\xi=0}^{\frac{\boldsymbol{\beta}}{2}} \frac{\boldsymbol{\beta}!}{\xi! \cdot (\frac{\boldsymbol{\beta}}{2} - \xi)! \cdot 2^{\frac{\boldsymbol{\beta}}{2} - \xi}} \left(\frac{-1}{2}\right)^{\xi} \\
&= \frac{\boldsymbol{\beta}!}{(\frac{\boldsymbol{\beta}}{2})! \cdot 2^{\frac{\boldsymbol{\beta}}{2}}} \sum_{\xi=0}^{\frac{\boldsymbol{\beta}}{2}} \binom{\frac{\boldsymbol{\beta}}{2}}{\xi} (-1)^{\xi} \\
&= \frac{\boldsymbol{\beta}!}{(\frac{\boldsymbol{\beta}}{2})! \cdot 2^{\frac{\boldsymbol{\beta}}{2}}} (1 + (-1))^{\frac{\boldsymbol{\beta}}{2}} \\
&= \mathbb{1}_{\boldsymbol{\beta}=0}.
\end{aligned}$$

■

Lemma 20 *For $p \geq 4$ and $k \leq \sqrt{p}$, it holds that $\frac{p^k}{4k!} \leq \binom{p}{k}$.*

Proof Note that $\binom{p}{k} \geq \frac{p^k}{4k!}$ if and only if:

$$\prod_{j=1}^{k-1} \left(1 - \frac{j}{p}\right) \geq \frac{1}{4}$$

Then applying the $k \leq \sqrt{p}$ assumption:

$$\prod_{j=1}^{k-1} \left(1 - \frac{j}{p}\right) \geq \prod_{i=1}^{\lfloor \sqrt{p} \rfloor} \left(1 - \frac{j}{p}\right) \geq \left(1 - \frac{1}{\sqrt{p}}\right)^{\sqrt{p}}$$

Now notice that for $\sqrt{p} \geq 2$, we have that $(1 - \frac{1}{\sqrt{p}})^{\sqrt{p}} \geq \frac{1}{4}$, leading to the desired result. ■

B.1. Low-degree analysis for MSLR: general mixtures

In subsection, we prove two technical lemmas. The first (Lemma 21) derives an expression for the projection of the likelihood ratio onto the multivariate Hermite polynomial $\tilde{H}_{\boldsymbol{\alpha}}$ defined in Proposition 18. The second lemma (Lemma 22) derives an explicit expression for the low-degree chi-squared divergence.

Lemma 21 *Let $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \sim \mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$. Let $\mathsf{L} = \frac{d\mathbb{P}(\boldsymbol{X},\boldsymbol{y})}{d\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y})}$ be the likelihood ratio, and $\tilde{H}_{\boldsymbol{\alpha}}$ the Hermite polynomial defined in Proposition 18. Then, for $\boldsymbol{\alpha} \in \mathbb{N}^{n\times(p+1)}$, we have*

$$\langle \mathsf{L}, \tilde{H}_{\boldsymbol{\alpha}} \rangle = \left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}+1}} \right)^{|\boldsymbol{\alpha}_{p+1}|} \|\boldsymbol{\beta}\|_2^{|\boldsymbol{\alpha}_{p+1}|-|\boldsymbol{\alpha}_{\cdot,:p}|} \boldsymbol{\alpha}_{p+1}! \cdot \prod_{i=1}^{n} \mathbb{1}_{\{\alpha_{i,p+1}-|\boldsymbol{\alpha}_{\cdot,:p}|=0\}} \mathop{\mathbb{E}}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} \left[\prod_{i=1}^{n}\prod_{j=1}^{p} (\boldsymbol{\beta}_{1,j} z_i + \boldsymbol{\beta}_{2,j}(1-z_i))^{\alpha_{i,j}} \right].$$

Proof We begin by expanding the inner product:

$$\begin{aligned}
&\langle \mathsf{L}, \tilde{H}_{\boldsymbol{\alpha}} \rangle \\
&= \mathop{\mathbb{E}}_{\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y})} \left[\frac{d\mathbb{P}(\boldsymbol{X},\boldsymbol{y})}{d\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y})} \tilde{H}_{\boldsymbol{\alpha}}(\boldsymbol{X},\boldsymbol{y}) \right] \\
&= \mathop{\mathbb{E}}_{\mathbb{P}(\boldsymbol{X},\boldsymbol{y})} \left[\prod_{i=1}^{n} \left(\prod_{j=1}^{p} H_{\alpha_{i,j}}(X_{i,j}) \right) H_{\alpha_{i,p+1}} \left(\frac{\left(\frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_1 \odot \boldsymbol{z} + \frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_2 \odot (1-\boldsymbol{z}) + \boldsymbol{w}\right)_i}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}+1}} \right) \right],
\end{aligned}$$

and applying Gaussian Integration by Parts (Proposition 14) we obtain

$$\begin{aligned}
&\langle \mathsf{L}, \tilde{H}_{\boldsymbol{\alpha}} \rangle \\
&= \mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{z},\boldsymbol{w}} \prod_{i=1}^{n} \left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}+1}} \right)^{|\boldsymbol{\alpha}_{i,:p}|} \frac{\alpha_{i,p+1}!}{(\alpha_{i,p+1}-|\boldsymbol{\alpha}_{i,:p}|)!} \\
&\qquad \cdot \prod_{j=1}^{p} (\boldsymbol{\beta}_{1,j} z_i + \boldsymbol{\beta}_{2,j}(1-z_i))^{\alpha_{i,j}} H_{\alpha_{i,p+1}-|\boldsymbol{\alpha}_{\boldsymbol{i},\boldsymbol{:p}}|} \left(\frac{\left(\frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_1 \odot \boldsymbol{z} + \frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_2 \odot (1-\boldsymbol{z}) + \boldsymbol{w}\right)_i}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}+1}} \right) \\
&= \left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}+1}} \right)^{|\boldsymbol{\alpha}_{\cdot,:p}|} \mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{z},\boldsymbol{w}} \prod_{i=1}^{n} \frac{\alpha_{i,p+1}!}{(\alpha_{i,p+1}-|\boldsymbol{\alpha}_{i,:p}|)!} \\
&\qquad \cdot \prod_{j=1}^{p} (\boldsymbol{\beta}_{1,j} z_i + \boldsymbol{\beta}_{2,j}(1-z_i))^{\alpha_{i,j}} H_{\alpha_{i,p+1}-|\boldsymbol{\alpha}_{\boldsymbol{i},\boldsymbol{:p}}|} \left(\frac{\left(\frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_1 \odot \boldsymbol{z} + \frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_2 \odot (1-\boldsymbol{z}) + \boldsymbol{w}\right)_i}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}+1}} \right).
\end{aligned}$$

We then apply the Hermite multiplication and addition formulas outlined in Propositions 16 and 17:

$$
\begin{aligned}
&\langle \mathsf{L}, \tilde{H}_{\boldsymbol{\alpha}} \rangle \\
&= \left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1}} \right)^{|\boldsymbol{\alpha}_{\cdot,:p}|} \mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{z},\boldsymbol{w}} \prod_{i=1}^{n} \frac{\alpha_{i,p+1}!}{(\alpha_{i,p+1} - |\alpha_{i,:p}|)!} \prod_{j=1}^{p} (\boldsymbol{\beta}_{1,j} z_i + \boldsymbol{\beta}_{2,j}(1-z_i))^{\alpha_{i,j}} \\
&\cdot \sum_{\xi=0}^{\lfloor \frac{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2} \rfloor} \left(\frac{1}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1}} \right)^{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi} \left(\frac{1}{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1} - 1 \right)^{\xi} \binom{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2\xi} \frac{(2\xi)!}{\xi!} 2^{-\xi} \\
&\cdot H_{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi} \left(\left(\frac{1}{\sigma} \boldsymbol{X}\boldsymbol{\beta}_1 \odot \boldsymbol{z} + \frac{1}{\sigma} \boldsymbol{X}\boldsymbol{\beta}_2 \odot (1-\boldsymbol{z}) + \boldsymbol{w} \right)_i \right) \\
&= \left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1}} \right)^{|\boldsymbol{\alpha}_{\cdot,:p}|} \mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{z},\boldsymbol{w}} \prod_{i=1}^{n} \frac{\alpha_{i,p+1}!}{(\alpha_{i,p+1} - |\alpha_{i,:p}|)!} \prod_{j=1}^{p} (\boldsymbol{\beta}_{1,j} z_i + \boldsymbol{\beta}_{2,j}(1-z_i))^{\alpha_{i,j}} \\
&\cdot \sum_{\xi=0}^{\lfloor \frac{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2} \rfloor} \left(\frac{1}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1}} \right)^{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi} \left(\frac{1}{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1} - 1 \right)^{\xi} \binom{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2\xi} \frac{(2\xi)!}{\xi!} 2^{-\xi} \\
&\cdot \sum_{\eta=0}^{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi} \binom{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi}{\eta} \left(\left(\frac{1}{\sigma} \boldsymbol{X}\boldsymbol{\beta}_1 \odot \boldsymbol{z} + \frac{1}{\sigma} \boldsymbol{X}\boldsymbol{\beta}_2 \odot (1-\boldsymbol{z}) \right)_i \right)^{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi - \eta} \underbrace{H_\eta(w_i)}_{\neq 0 \text{ only if } \eta = 0},
\end{aligned}
$$

which we simplify by noting that $H_\eta(w_i) \neq 0$ only if $\eta = 0$ to obtain:

$$
\begin{aligned}
&\langle \mathsf{L}, \tilde{H}_{\boldsymbol{\alpha}} \rangle \\
&= \left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1}} \right)^{|\boldsymbol{\alpha}_{\cdot,:p}|} \mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{z}} \prod_{i=1}^{n} \frac{\alpha_{i,p+1}!}{(\alpha_{i,p+1} - |\alpha_{i,:p}|)!} \prod_{j=1}^{p} (\boldsymbol{\beta}_{1,j} z_i + \boldsymbol{\beta}_{2,j}(1-z_i))^{\alpha_{i,j}} \\
&\cdot \sum_{\xi=0}^{\lfloor \frac{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2} \rfloor} \left(\frac{1}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1}} \right)^{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi} \left(\frac{1}{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1} - 1 \right)^{\xi} \binom{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2\xi} \frac{(2\xi)!}{\xi!} 2^{-\xi} \\
&\cdot \left(\left(\frac{1}{\sigma} \boldsymbol{X}\boldsymbol{\beta}_1 \odot \boldsymbol{z} + \frac{1}{\sigma} \boldsymbol{X}\boldsymbol{\beta}_2 \odot (1-\boldsymbol{z}) \right)_i \right)^{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi}.
\end{aligned}
$$

Now switching the sum with the product and grouping terms we obtain:

$$
\begin{aligned}
&\langle \mathsf{L}, \tilde{H}_{\boldsymbol{\alpha}} \rangle \\
&= \left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1}} \right)^{|\boldsymbol{\alpha}_{\cdot,:p}|} \prod_{i=1}^{n} \frac{\alpha_{i,p+1}!}{(\alpha_{i,p+1} - |\alpha_{i,:p}|)!} \mathop{\mathbb{E}}_{[\boldsymbol{X},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{z}]\sim\mathbb{P}(\boldsymbol{X},y)} \prod_{j=1}^{p} \left(\boldsymbol{\beta}_{1,j} z_i + \boldsymbol{\beta}_{2,j}(1-z_i)\right)^{\alpha_{i,j}} \\
&\cdot \sum_{\xi=0}^{\lfloor \frac{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2} \rfloor} \binom{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2\xi} \frac{(2\xi)!}{\xi!} \left(\frac{\left(\frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_1 \odot \boldsymbol{z} + \frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_2 \odot (1-\boldsymbol{z})\right)_i}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1}} \right)^{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi} \left(\frac{-\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}}{2\left(\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1\right)} \right)^{\xi} \\
&= \left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1}} \right)^{|\boldsymbol{\alpha}_{\cdot,:p}|} \frac{\boldsymbol{\alpha}_{p+1}!}{(\boldsymbol{\alpha}_{p+1} - |\boldsymbol{\alpha}_{\cdot,:p}|)!} \mathop{\mathbb{E}}_{[\boldsymbol{X},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{z}]\sim\mathbb{P}(\boldsymbol{X},y)} \sum_{0 \le \boldsymbol{\xi} \le \lfloor \frac{\boldsymbol{\alpha}_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2} \rfloor} \prod_{i=1}^{n} \\
&\cdot \binom{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2\xi_i} \frac{(2\xi_i)!}{\xi_i!} \left(\frac{-\frac{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}}{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}+1}}{2} \right)^{\xi_i} \left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1}} \right)^{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi_i} \\
&\cdot \left[\prod_{j=1}^{p} \left(\boldsymbol{\beta}_{1,j} z_i + \boldsymbol{\beta}_{2,j}(1-z_i)\right)^{\alpha_{i,j}} \left(\boldsymbol{X}\boldsymbol{\beta}_1 \odot \boldsymbol{z} + \boldsymbol{X}\boldsymbol{\beta}_2 \odot (1-\boldsymbol{z})\right)_i^{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi_i} \right],
\end{aligned}
$$

which by simplification and expansion then leads us to

$$
\begin{aligned}
&\langle \mathsf{L}, \tilde{H}_{\boldsymbol{\alpha}} \rangle \\
&= \left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1}} \right)^{|\boldsymbol{\alpha}_{\cdot,p+1}|} \frac{\boldsymbol{\alpha}_{p+1}!}{(\boldsymbol{\alpha}_{p+1} - |\boldsymbol{\alpha}_{\cdot,:p}|)!} \sum_{0 \le \boldsymbol{\xi} \le \lfloor \frac{\boldsymbol{\alpha}_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2} \rfloor} \\
&\cdot \left(\prod_{i=1}^{n} \binom{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2\xi_i} \frac{(2\xi_i)!}{\xi_i!} \left(\frac{-\|\boldsymbol{\beta}\|_2^2}{2} \right)^{\xi_i} \right) \\
&\cdot \mathop{\mathbb{E}}_{[\boldsymbol{X},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{z}]\sim\mathbb{P}(\boldsymbol{X},y)} \prod_{i=1}^{n} \left[\prod_{j=1}^{p} \left(\boldsymbol{\beta}_{1,j} z_i + \boldsymbol{\beta}_{2,j}(1-z_i)\right)^{\alpha_{i,j}} \left(\sum_{j=1}^{p} X_{i,j}\boldsymbol{\beta}_{1,j} z_i + \sum_{j=1}^{p} X_{i,j}\boldsymbol{\beta}_{2,j}(1-z_i) \right)^{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi_i} \right].
\end{aligned}
$$

Bringing out the expectation with respect to $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ we then obtain:

$$\begin{aligned}
&\langle \mathsf{L}, \tilde{H}_{\boldsymbol{\alpha}} \rangle \\
&= \left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1}} \right)^{|\boldsymbol{\alpha}_{\cdot,p+1}|} \frac{\boldsymbol{\alpha}_{p+1}!}{(\boldsymbol{\alpha}_{p+1} - |\boldsymbol{\alpha}_{\cdot,:p}|)!} \sum_{0 \le \boldsymbol{\xi} \le \lfloor \frac{\boldsymbol{\alpha}_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2} \rfloor} \mathop{\mathbb{E}}_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \left(\prod_{i=1}^n \prod_{j=1}^p (\boldsymbol{\beta}_{1,j} z_i + \boldsymbol{\beta}_{2,j}(1 - z_i))^{\alpha_{i,j}} \right) \\
&\cdot \prod_{i=1}^n \binom{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2\xi_i} \frac{(2\xi_i)!}{\xi_i!} \left(\frac{-\|\boldsymbol{\beta}\|_2^2}{2} \right)^{\xi_i} \mathop{\mathbb{E}}_{\boldsymbol{X}, \boldsymbol{z}} \left(\sum_{j=1}^p X_{i,j} \boldsymbol{\beta}_{1,j} z_i + \sum_{j=1}^p X_{i,j} \boldsymbol{\beta}_{2,j}(1 - z_i) \right)^{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi_i} \\
&= \left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1}} \right)^{|\boldsymbol{\alpha}_{\cdot,p+1}|} \frac{\boldsymbol{\alpha}_{p+1}!}{(\boldsymbol{\alpha}_{p+1} - |\boldsymbol{\alpha}_{\cdot,:p}|)!} \sum_{0 \le \boldsymbol{\xi} \le \lfloor \frac{\boldsymbol{\alpha}_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2} \rfloor} \mathop{\mathbb{E}}_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \left(\prod_{i=1}^n \prod_{j=1}^p (\boldsymbol{\beta}_{1,j} z_i + \boldsymbol{\beta}_{2,j}(1 - z_i))^{\alpha_{i,j}} \right) \\
&\cdot \prod_{i=1}^n \binom{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2\xi_i} \frac{(2\xi_i)!}{\xi_i!} \left(\frac{-\|\boldsymbol{\beta}\|_2^2}{2} \right)^{\xi_i} \mathop{\mathbb{E}}_{w \sim \mathcal{N}(0, \|\boldsymbol{\beta}\|_2^2)} w^{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi_i},
\end{aligned}$$

where $\sum_{j=1}^p X_{i,j} \boldsymbol{\beta}_{1,j} z_i + \sum_{j=1}^p X_{i,j} \boldsymbol{\beta}_{2,j}(1 - z_i) \sim \mathcal{N}(0, \|\boldsymbol{\beta}\|_2^2)$, both marginally and conditionally on $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{z}$, and hence is independent of $\prod_{i=1}^n \prod_{j=1}^p (\boldsymbol{\beta}_{1,j} z_i + \boldsymbol{\beta}_{2,j}(1 - z_i))^{\alpha_{i,j}}$ (since $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ are constrained to have norm $\|\boldsymbol{\beta}\|_2$ according to our prior). After switching the sum with the product, combining the known equation for Gaussian moments $\mathbb{E}_{w \sim \mathcal{N}(0, \|\boldsymbol{\beta}\|_2^2)} w^b = (b-1)!! \|\boldsymbol{\beta}\|_2^b \mathbb{1}_{\{b \text{ even}\}}$ with additional factorial simplifications, and applying Lemma 19, we obtain

$$\begin{aligned}
&\langle \mathsf{L}, \tilde{H}_{\boldsymbol{\alpha}} \rangle \\
&= \left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1}} \right)^{|\boldsymbol{\alpha}_{p+1}|} \frac{\boldsymbol{\alpha}_{p+1}!}{(\boldsymbol{\alpha}_{p+1} - |\boldsymbol{\alpha}_{\cdot,:p}|)!} \sum_{0 \le \boldsymbol{\xi} \le \lfloor \frac{\boldsymbol{\alpha}_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2} \rfloor} \\
&\cdot \left(\prod_{i=1}^n \binom{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2\xi_i} \frac{(2\xi_i)!}{\xi_i!} \left(\frac{-\|\boldsymbol{\beta}\|_2^2}{2} \right)^{\xi_i} \right) \mathop{\mathbb{E}}_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \left(\prod_{i=1}^n \prod_{j=1}^p (\boldsymbol{\beta}_{1,j} z_i + \boldsymbol{\beta}_{2,j}(1 - z_i))^{\alpha_{i,j}} \right) \\
&\cdot \left(\prod_{i=1}^n (\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi_i - 1)!! \cdot \|\boldsymbol{\beta}\|_2^{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi_i} \mathbb{1}_{\{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi_i \text{ even}\}} \right) \\
&= \left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1}} \right)^{|\boldsymbol{\alpha}_{p+1}|} \frac{\boldsymbol{\alpha}_{p+1}!}{(\boldsymbol{\alpha}_{p+1} - |\boldsymbol{\alpha}_{\cdot,:p}|)!} \|\boldsymbol{\beta}\|_2^{|\boldsymbol{\alpha}_{p+1}| - |\boldsymbol{\alpha}_{\cdot,:p}|} \\
&\cdot \prod_{i=1}^n \sum_{\xi}^{\lfloor \frac{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2} \rfloor} \binom{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}|}{2\xi} \frac{(2\xi)!}{\xi!} \left(\frac{-1}{2} \right)^{\xi} \\
&\cdot (\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi - 1)!! \cdot \mathbb{1}_{\{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| - 2\xi \text{ even}\}} \mathop{\mathbb{E}}_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \left(\prod_{i=1}^n \prod_{j=1}^p (\boldsymbol{\beta}_{1,j} z_i + \boldsymbol{\beta}_{2,j}(1 - z_i))^{\alpha_{i,j}} \right)
\end{aligned}$$

$$= \left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}+1}}\right)^{|\boldsymbol{\alpha}_{p+1}|} \frac{\boldsymbol{\alpha}_{p+1}!}{(\boldsymbol{\alpha}_{p+1}-|\boldsymbol{\alpha}_{\cdot,:p}|)!}\|\boldsymbol{\beta}\|_2^{|\boldsymbol{\alpha}_{p+1}|-|\boldsymbol{\alpha}_{\cdot,:p}|}$$
$$\cdot \prod_{i=1}^{n} \mathbb{1}_{\{\alpha_{i,p+1}-|\boldsymbol{\alpha}_{i,:p}|=0\}} \mathop{\mathbb{E}}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}\left[\prod_{i=1}^{n}\prod_{j=1}^{p}(\boldsymbol{\beta}_{1,j}z_i+\boldsymbol{\beta}_{2,j}(1-z_i))^{\alpha_{i,j}}\right]$$
$$= \left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}+1}}\right)^{|\boldsymbol{\alpha}_{p+1}|} \|\boldsymbol{\beta}\|_2^{|\boldsymbol{\alpha}_{p+1}|-|\boldsymbol{\alpha}_{\cdot,:p}|}\boldsymbol{\alpha}_{p+1}!\prod_{i=1}^{n}\mathbb{1}_{\{\alpha_{i,p+1}-|\boldsymbol{\alpha}_{\cdot,:p}|=0\}} \mathop{\mathbb{E}}_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}\left[\prod_{i=1}^{n}\prod_{j=1}^{p}(\boldsymbol{\beta}_{1,j}z_i+\boldsymbol{\beta}_{2,j}(1-z_i))^{\alpha_{i,j}}\right],$$

which leads to the desired result. ■

Lemma 22 *Let $(\boldsymbol{\beta}_1^{(1)},\boldsymbol{\beta}_2^{(1)})$ and $(\boldsymbol{\beta}_1^{(2)},\boldsymbol{\beta}_2^{(2)})$ be two independent copies of signals sampled from $\mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$, and likewise for $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$ sampled entrywise from Bernoulli(ϕ). We then have*

$$\chi^2_{\le D}(\mathbb{P}(\boldsymbol{X},\boldsymbol{y})\|\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y}))+1 = \mathop{\mathbb{E}}_{\substack{(\boldsymbol{\beta}_1^{(1)},\boldsymbol{\beta}_2^{(1)}),(\boldsymbol{\beta}_1^{(2)},\boldsymbol{\beta}_2^{(2)})\overset{i.i.d.}{\sim}\mathcal{P}\\ z^{(1)},z^{(2)}\overset{i.i.d.}{\sim}Ber(\phi)}} \sum_{\frac{d}{2}=0}^{\lfloor\frac{D}{2}\rfloor}\left(\frac{1}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right)^{\frac{d}{2}}$$
$$\cdot\sum_{|\boldsymbol{\alpha}_{p+1}|=\frac{d}{2}}\prod_{i=1}^{n}\langle\boldsymbol{\beta}_1^{(1)}z_i^{(1)}+\boldsymbol{\beta}_2^{(1)}(1-z_i^{(1)}),\boldsymbol{\beta}_1^{(2)}z_i^{(2)}+\boldsymbol{\beta}_2^{(2)}(1-z_i^{(2)})\rangle^{\alpha_{i,p+1}}.$$

Proof We begin the proof by applying Lemma 21 to obtain:

$$\chi^2_{\le D}(\mathbb{P}(\boldsymbol{X},\boldsymbol{y})\|\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y}))+1$$
$$=\sum_{0\le|\boldsymbol{\alpha}|\le D}\frac{1}{\boldsymbol{\alpha}!}\langle\mathsf{L},\tilde{H}_{\boldsymbol{\alpha}}\rangle^2$$
$$=\sum_{0\le|\boldsymbol{\alpha}|\le D}\frac{1}{\boldsymbol{\alpha}!}\left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}+1}}\right)^{2|\boldsymbol{\alpha}_{p+1}|}\|\boldsymbol{\beta}\|_2^{2|\boldsymbol{\alpha}_{p+1}|-2|\boldsymbol{\alpha}_{\cdot,:p}|}(\boldsymbol{\alpha}_{p+1}!)^2$$
$$\cdot\prod_{i=1}^{n}\mathbb{1}_{\{\alpha_{i,p+1}-|\boldsymbol{\alpha}_{\cdot,:p}|=0\}}\left(\mathbb{E}\prod_{i=1}^{n}\prod_{j=1}^{p}(\boldsymbol{\beta}_{1,j}z_i+\boldsymbol{\beta}_{2,j}(1-z_i))^{\alpha_{i,j}}\right)^2$$
$$=\sum_{d=0}^{D}\sum_{h=0}^{d}\sum_{|\boldsymbol{\alpha}_{p+1}|=h}\sum_{|\boldsymbol{\alpha}_{\cdot,:p}|=d-h}\frac{1}{\boldsymbol{\alpha}!}\left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}+1}}\right)^{2|\boldsymbol{\alpha}_{p+1}|}\|\boldsymbol{\beta}\|_2^{2|\boldsymbol{\alpha}_{p+1}|-2|\boldsymbol{\alpha}_{\cdot,:p}|}(\boldsymbol{\alpha}_{p+1}!)^2$$
$$\cdot\prod_{i=1}^{n}\mathbb{1}_{\{\alpha_{i,p+1}-|\boldsymbol{\alpha}_{\cdot,:p}|=0\}}\left(\mathbb{E}\prod_{i=1}^{n}\prod_{j=1}^{p}(\boldsymbol{\beta}_{1,j}z_i+\boldsymbol{\beta}_{2,j}(1-z_i))^{\alpha_{i,j}}\right)^2.$$

We next split a squared expectation into the expectation of the multiplication of two independent random variables: $(\mathbb{E}_w[w])^2=\mathbb{E}_{w^{(1)}}\left[w^{(1)}\right]\mathbb{E}_{w^{(2)}}\left[w^{(2)}\right]=\mathbb{E}_{w^{(1)},w^{(2)}}\left[w^{(1)}w^{(2)}\right]$, where we have chosen $w^{(1)}$ and $w^{(2)}$ to be two independent and identically distributed random variables. Continuing in this way, we obtain:

$$\begin{aligned}
&\chi^2_{\leq D}(\mathbb{P}(\boldsymbol{X}, \boldsymbol{y}) \| \mathbb{P}(\boldsymbol{X}) \otimes \mathbb{P}(\boldsymbol{y})) + 1 \\
&= \sum_{d=0}^{D} \sum_{h=0}^{d} \sum_{|\boldsymbol{\alpha}_{p+1}|=h} \sum_{|\boldsymbol{\alpha}_{\cdot,:p}|=d-h} \frac{1}{\boldsymbol{\alpha}!} \left(\frac{\frac{1}{\sigma}}{\sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} + 1}} \right)^{2|\boldsymbol{\alpha}_{p+1}|} \|\boldsymbol{\beta}\|_2^{2|\boldsymbol{\alpha}_{p+1}| - 2|\boldsymbol{\alpha}_{\cdot,:p}|} (\boldsymbol{\alpha}_{p+1}!)^2 \\
&\cdot \underbrace{\prod_{i=1}^{n} \mathbb{1}_{\{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| = 0\}}}_{\implies \sum_{i=1}^{n} \alpha_{i,p+1} - |\boldsymbol{\alpha}_{\boldsymbol{i},\boldsymbol{:p}}| = 0 \implies 2h - d = 0 \implies d \text{ even}} \\
&\cdot \mathop{\mathbb{E}}_{\substack{(\boldsymbol{\beta}_1^{(1)}, \boldsymbol{\beta}_2^{(1)}), (\boldsymbol{\beta}_1^{(2)}, \boldsymbol{\beta}_2^{(2)}) \overset{\text{i.i.d.}}{\sim} \mathcal{P} \\ z^{(1)}, z^{(2)} \overset{\text{i.i.d.}}{\sim} \text{Ber}(\phi)}} \prod_{i=1}^{n} \prod_{j=1}^{p} \left(\boldsymbol{\beta}_{1,j}^{(1)} z_i^{(1)} + \boldsymbol{\beta}_{2,j}^{(1)} (1 - z_i^{(1)}) \right)^{\alpha_{i,j}} \left(\boldsymbol{\beta}_{1,j}^{(2)} z_i^{(2)} + \boldsymbol{\beta}_{2,j}^{(2)} (1 - z_i^{(2)}) \right)^{\alpha_{i,j}},
\end{aligned}$$

which we simplify after noticing $\sum_{i=1}^{n} \alpha_{i,p+1} - |\boldsymbol{\alpha}_{\boldsymbol{i},\boldsymbol{:p}}| = 0$ implies d must be even,

$$\begin{aligned}
&\chi^2_{\leq D}(\mathbb{P}(\boldsymbol{X}, \boldsymbol{y}) \| \mathbb{P}(\boldsymbol{X}) \otimes \mathbb{P}(\boldsymbol{y})) + 1 \\
&= \mathbb{E} \sum_{\frac{d}{2}=0}^{\lfloor \frac{D}{2} \rfloor} \sum_{|\boldsymbol{\alpha}_{p+1}| = \frac{d}{2}} \sum_{|\boldsymbol{\alpha}_{\cdot,:p}| = \frac{d}{2}} \frac{1}{\boldsymbol{\alpha}_{p+1}! \cdot \boldsymbol{\alpha}_{\cdot,:p}!} \left(\frac{1}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2} \right)^{\frac{d}{2}} \|\boldsymbol{\beta}\|_2^0 (\boldsymbol{\alpha}_{p+1}!)^2 \\
&\quad \cdot \prod_{i=1}^{n} \mathbb{1}_{\{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| = 0\}} \\
&\quad \cdot \prod_{i=1}^{n} \prod_{j=1}^{p} \left(\boldsymbol{\beta}_{1,j}^{(1)} z_i^{(1)} + \boldsymbol{\beta}_{2,j}^{(1)} (1 - z_i^{(1)}) \right)^{\alpha_{i,j}} \left(\boldsymbol{\beta}_{1,j}^{(2)} z_i^{(2)} + \boldsymbol{\beta}_{2,j}^{(2)} (1 - z_i^{(2)}) \right)^{\alpha_{i,j}} \\
&= \mathbb{E} \sum_{\frac{d}{2}=0}^{\lfloor \frac{D}{2} \rfloor} \left(\frac{1}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2} \right)^{\frac{d}{2}} \\
&\quad \cdot \sum_{|\boldsymbol{\alpha}_{p+1}| = \frac{d}{2}} \frac{(\boldsymbol{\alpha}_{p+1}!)^2}{\boldsymbol{\alpha}_{p+1}!} \sum_{|\boldsymbol{\alpha}_{\cdot,:p}| = \frac{d}{2}} \frac{1}{\boldsymbol{\alpha}_{\cdot,:p}!} \prod_{i=1}^{n} \mathbb{1}_{\{\alpha_{i,p+1} - |\boldsymbol{\alpha}_{i,:p}| = 0\}} \\
&\quad \cdot \prod_{i=1}^{n} \prod_{j=1}^{p} \left(\boldsymbol{\beta}_{1,j}^{(1)} z_i^{(1)} + \boldsymbol{\beta}_{2,j}^{(1)} (1 - z_i^{(1)}) \right)^{\alpha_{i,j}} \left(\boldsymbol{\beta}_{1,j}^{(2)} z_i^{(2)} + \boldsymbol{\beta}_{2,j}^{(2)} (1 - z_i^{(2)}) \right)^{\alpha_{i,j}},
\end{aligned}$$

and can be re-ordered in order to more clearly apply the multinomial theorem:

$$\begin{aligned}
&\chi^2_{\le D}(\mathbb{P}(\boldsymbol{X},\boldsymbol{y})\|\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y}))+1\\
&=\mathbb{E}\sum_{\frac{d}{2}=0}^{\lfloor\frac{D}{2}\rfloor}\left(\frac{1}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right)^{\frac{d}{2}}\sum_{|\boldsymbol{\alpha}_{p+1}|=\frac{d}{2}}\boldsymbol{\alpha}_{p+1}!\\
&\quad\cdot\sum_{|\boldsymbol{\alpha}_{1,:p}|=\alpha_{1,p+1}}\frac{1}{\boldsymbol{\alpha}_{1,:p}!}\cdots\sum_{|\boldsymbol{\alpha}_{n,:p}|=\alpha_{n,p+1}}\frac{1}{\boldsymbol{\alpha}_{n,:p}!}\\
&\quad\prod_{i=1}^{n}\prod_{j=1}^{p}\left(\boldsymbol{\beta}^{(1)}_{1,j}z_i^{(1)}+\boldsymbol{\beta}^{(1)}_{2,j}(1-z_i^{(1)})\right)^{\alpha_{i,j}}\left(\boldsymbol{\beta}^{(2)}_{1,j}z_i^{(2)}+\boldsymbol{\beta}^{(2)}_{2,j}(1-z_i^{(2)})\right)^{\alpha_{i,j}}\\
&=\mathbb{E}\sum_{\frac{d}{2}=0}^{\lfloor\frac{D}{2}\rfloor}\left(\frac{1}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right)^{\frac{d}{2}}\sum_{|\boldsymbol{\alpha}_{p+1}|=\frac{d}{2}}\boldsymbol{\alpha}_{p+1}!\\
&\quad\cdot\sum_{|\boldsymbol{\alpha}_{1,:p}|=\alpha_{1,p+1}}\frac{\prod_{j=1}^{p}\left(\boldsymbol{\beta}^{(1)}_{1,j}z_i^{(1)}+\boldsymbol{\beta}^{(1)}_{2,j}(1-z_i^{(1)})\right)^{\alpha_{1,j}}\left(\boldsymbol{\beta}^{(2)}_{1,j}z_i^{(2)}+\boldsymbol{\beta}^{(2)}_{2,j}(1-z_i^{(2)})\right)^{\alpha_{1,j}}}{\boldsymbol{\alpha}_{1,:p}!}\\
&\quad\cdots\sum_{|\boldsymbol{\alpha}_{n,:p}|=\alpha_{n,p+1}}\frac{\prod_{j=1}^{p}\left(\boldsymbol{\beta}^{(1)}_{1,j}z_i^{(1)}+\boldsymbol{\beta}^{(1)}_{2,j}(1-z_i^{(1)})\right)^{\alpha_{n,j}}\left(\boldsymbol{\beta}^{(2)}_{1,j}z_i^{(2)}+\boldsymbol{\beta}^{(2)}_{2,j}(1-z_i^{(2)})\right)^{\alpha_{n,j}}}{\boldsymbol{\alpha}_{n,:p}!}.
\end{aligned}$$

We then apply the multinomial theorem to obtain the result:

$$\begin{aligned}
&\chi^2_{\le D}(\mathbb{P}(\boldsymbol{X},\boldsymbol{y})\|\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y}))+1\\
&=\mathbb{E}\sum_{\frac{d}{2}=0}^{\lfloor\frac{D}{2}\rfloor}\left(\frac{1}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right)^{\frac{d}{2}}\sum_{|\boldsymbol{\alpha}_{p+1}|=\frac{d}{2}}\prod_{i=1}^{n}\langle\boldsymbol{\beta}^{(1)}_1z_i^{(1)}+\boldsymbol{\beta}^{(1)}_2(1-z_i^{(1)}),\boldsymbol{\beta}^{(2)}_1z_i^{(2)}+\boldsymbol{\beta}^{(2)}_2(1-z_i^{(2)})\rangle^{\alpha_{i,p+1}}.
\end{aligned}$$

■

In the next two subsections, we prove the computational lower bounds for $\mathtt{SLR}-\mathtt{D}$ (Theorem 10 and $\mathtt{SB\text{-}MSLR}-\mathtt{D}$ (Theorem 4) by specializing Lemma 22.

B.2. Special Case: $\mathtt{SLR}-\mathtt{D}$

Proof [Proof of Theorem 10] In Theorem 23 below, recalling that $\mathtt{SNR}=\|\boldsymbol{\beta}\|_2^2/\sigma^2$, we let $n=(1-\epsilon)(1-2\theta)\frac{\|\boldsymbol{\beta}\|_2^2}{\boldsymbol{\beta}\|_\infty^2}\frac{\mathtt{SNR}+1}{\mathtt{SNR}}\log p$. Then Theorem 23 implies that for all $D\le\frac{2\epsilon}{1-\epsilon}n$, we have $\chi^2_{\le D}(\mathbb{P}(\boldsymbol{X},\boldsymbol{y})\|\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y}))=O(1)$. Applying Conjecture 2 with $D=\frac{2\epsilon}{1-\epsilon}n$ and recalling $n=\omega(\log p)$ (by the assumptions of the theorem), we have that running time $\exp\left(\tilde{\Omega}(n)\right)$ is required. ■

Theorem 23 (General $\mathtt{SLR}-\mathtt{D}$ lower bound) *Consider the setting of* $\mathtt{SLR}-\mathtt{D}$ *(Definition 3 with* $\boldsymbol{\beta}_1=\boldsymbol{\beta}_2$*). Let* $\boldsymbol{\beta}\sim\mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$*. If* $k=O(p^\theta)\le\sqrt{p}$ *for some* $\theta\in(0,1/2]$*, then for any* $\epsilon\in(0,1)$*,*

$n \leq (1-\epsilon)(1-2\theta)\left(\frac{\|\boldsymbol{\beta}\|_2^2+\sigma^2}{\|\boldsymbol{\beta}\|_\infty^2}\right)\log p$ *and* $D \leq \frac{2\epsilon}{1-\epsilon}n$*, we have* $\chi^2_{\leq D}(\mathbb{P}(\boldsymbol{X},\boldsymbol{y})\|\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y})) = O(1)$.

Proof Let $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$ denote the support sets of $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$, respectively. We apply Lemma 22 to obtain:

$$\begin{aligned}
\chi^2_{\leq D}(\mathbb{P}(\boldsymbol{X},\boldsymbol{y})\|\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y}))+1 &= \mathbb{E}\sum_{\frac{d}{2}=0}^{\lfloor\frac{D}{2}\rfloor}\left(\frac{1}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right)^{\frac{d}{2}}\sum_{|\boldsymbol{\alpha}_{p+1}|=\frac{d}{2}}\prod_{i=1}^{n}\langle\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)}\rangle^{\alpha_{i,p+1}} \\
&= \mathbb{E}\sum_{\frac{d}{2}=0}^{\lfloor\frac{D}{2}\rfloor}\left(\frac{1}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right)^{\frac{d}{2}}\sum_{|\boldsymbol{\alpha}_{p+1}|=\frac{d}{2}}\langle\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)}\rangle^{\frac{d}{2}} \\
&\leq \mathbb{E}\sum_{\frac{d}{2}=0}^{\lfloor\frac{D}{2}\rfloor}\left(\frac{1}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right)^{\frac{d}{2}}\binom{\frac{D}{2}+n-1}{n-1}\langle\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)}\rangle^{\frac{d}{2}} \\
&\leq \mathbb{E}\sum_{\frac{d}{2}=0}^{\lfloor\frac{D}{2}\rfloor}\left(\frac{1}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right)^{\frac{d}{2}}\frac{\left(\frac{D}{2}+n\right)^{\frac{d}{2}}}{\frac{d}{2}!}\langle\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)}\rangle^{\frac{d}{2}} \\
&\leq \mathop{\mathbb{E}}_{\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)}\overset{\text{i.i.d.}}{\sim}\mathcal{P}}\exp\left(\frac{\frac{1}{\sigma^2}}{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}+1}\left(\frac{D}{2}+n\right)\langle\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)}\rangle\right) \\
&\leq \mathop{\mathbb{E}}_{\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)}\overset{\text{i.i.d.}}{\sim}\mathcal{P}}\exp\left(\frac{\langle\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)}\rangle}{\|\boldsymbol{\beta}\|_\infty^2}(1-2\theta)\log p\right).
\end{aligned}$$

We then apply Lemma 20 and notice that $\langle\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)}\rangle \leq \|\boldsymbol{\beta}\|_\infty^2|\mathcal{S}^{(1)}\cap\mathcal{S}^{(2)}|$ to obtain, for $p>4$ and $k \leq p$:

$$\begin{aligned}
\chi^2_{\leq D}(\mathbb{P}(\boldsymbol{X},\boldsymbol{y})\|\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y}))+1 &= \mathop{\mathbb{E}}_{\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)}\overset{\text{i.i.d.}}{\sim}\mathcal{P}}\exp\left(\frac{\langle\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)}\rangle}{\|\boldsymbol{\beta}\|_\infty^2}(1-2\theta)\log p\right) \\
&\leq \mathop{\mathbb{E}}_{\langle\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)}\rangle}\exp\left(|\mathcal{S}^{(1)}\cap\mathcal{S}^{(2)}|(1-2\theta)\log p\right) \\
&\leq \sum_{l=0}^{k}\frac{\binom{k}{l}\binom{p-k}{k-l}}{\binom{p}{k}}\exp\left(l(1-2\theta)\log p\right) \\
&\leq \sum_{l=0}^{k}\frac{4k!}{p^k}\frac{k^l}{l!}\frac{(p-k)^{k-l}}{(k-l)!}\exp\left(l(1-2\theta)\log p\right) \\
&\leq 4\sum_{l=0}^{k}\left(\frac{k^2}{p}\right)^l\exp\left(l(1-2\theta)\log p\right) \\
&= 4\sum_{l=0}^{k}\left(\frac{k^2}{p^{2\theta}}\right)^l = O(1).
\end{aligned}$$

■

B.3. Special Case: $\texttt{SB-MSLR}-\texttt{D}$

Proof [Proof of Theorem 4] This follows from Theorem 25 below. Choosing any sample size n such that $n \geq k$, $n = \omega(\log p)$, and $n = o((\|\boldsymbol{\beta}\|_2^2 + \sigma^2)^2/(\|\boldsymbol{\beta}\|_\infty^4 \log p))$, we have that $\chi^2_{\leq D}(\mathbb{P}(\boldsymbol{X},\boldsymbol{y})\|\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y})) = O(1)$ for $D = (\sqrt{2}-1)\min\left\{\frac{(\|\boldsymbol{\beta}\|_2^2+\sigma^2)^2}{n\|\boldsymbol{\beta}\|_\infty^4}, n\right\}$. We then invoke Conjecture 2 and use $\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2} = \texttt{SNR}$, and notice that for signals with bounded amplitude, we have $k\|\boldsymbol{\beta}\|_\infty/\sigma^2 \gtrsim \texttt{SNR} \gtrsim k\|\boldsymbol{\beta}\|_\infty/\sigma^2$. ■

Lemma 24 *For $\texttt{SB-MSLR}-\texttt{D}$, we have that:*

$$\chi^2_{\leq D}(\mathbb{P}(\boldsymbol{X},\boldsymbol{y})\|\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y})) + 1 \leq \mathop{\mathbb{E}}_{\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)}} \sum_{\frac{d}{4}=0}^{\lfloor\frac{D}{4}\rfloor} \left(\frac{1}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right)^{\frac{d}{2}} \frac{\left(\frac{D}{4}+n\right)^{\frac{d}{4}}}{\frac{d}{4}!} \langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\rangle^{\frac{d}{2}},$$

where $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}$ are two independent copies of the random variable $\boldsymbol{\beta} \overset{d}{=} \boldsymbol{\beta}_1$.

Proof We begin by applying the assumptions into Lemma 22 and applying independence of the z_i's. Notice that in the context of $\texttt{SB-MSLR}$, we have in particular that $\phi\boldsymbol{\beta}_1 + (1-\phi)\boldsymbol{\beta}_2 = 0$, and hence we plug in $\boldsymbol{\beta} := \boldsymbol{\beta}_1 = -\frac{\phi}{1-\phi}\boldsymbol{\beta}_2$.

$$\begin{aligned}
&\chi^2_{\leq D}(\mathbb{P}(\boldsymbol{X},\boldsymbol{y})\|\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y})) + 1 \\
&\leq \mathbb{E} \sum_{\frac{d}{2}=0}^{\lfloor\frac{D}{2}\rfloor} \left(\frac{1}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right)^{\frac{d}{2}} \sum_{|\boldsymbol{\alpha}_{p+1}|=\frac{d}{2}} \prod_{i=1}^{n} \langle \boldsymbol{\beta}_1^{(1)} z_i^{(1)} + \boldsymbol{\beta}_2^{(1)}(1-z_i^{(1)}), \boldsymbol{\beta}_1^{(2)} z_i^{(2)} + \boldsymbol{\beta}_2^{(2)}(1-z_i^{(2)})\rangle^{\alpha_{i,p+1}} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\beta}} \sum_{\frac{d}{2}=0}^{\lfloor\frac{D}{2}\rfloor} \left(\frac{1}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right)^{\frac{d}{2}} \sum_{|\boldsymbol{\alpha}_{p+1}|=\frac{d}{2}} \cdot \prod_{i=1}^{n} \Big(\phi^2 \langle \boldsymbol{\beta}_1^{(1)}, \boldsymbol{\beta}_1^{(2)}\rangle^{\alpha_{i,p+1}} + \phi(1-\phi)\langle \boldsymbol{\beta}_1^{(1)}, \boldsymbol{\beta}_2^{(2)}\rangle^{\alpha_{i,p+1}} \\
&\qquad + \phi(1-\phi)\langle \boldsymbol{\beta}_2^{(1)}, \boldsymbol{\beta}_1^{(2)}\rangle^{\alpha_{i,p+1}} + (1-\phi)^2 \langle \boldsymbol{\beta}_2^{(1)}, \boldsymbol{\beta}_1^{(2)}\rangle^{\alpha_{i,p+1}}\Big) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\beta}} \sum_{\frac{d}{2}=0}^{\lfloor\frac{D}{2}\rfloor} \left(\frac{1}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right)^{\frac{d}{2}} \sum_{|\boldsymbol{\alpha}_{p+1}|=\frac{d}{2}} \\
&\qquad \cdot \prod_{i=1}^{n} \left(\phi^2 + 2\phi(1-\phi)\left(-\frac{\phi}{1-\phi}\right)^{\alpha_{i,p+1}} + (1-\phi)^2\left(\frac{\phi}{1-\phi}\right)^{2\alpha_{i,p+1}}\right) \langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\rangle^{\alpha_{i,p+1}} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\beta}} \sum_{\frac{d}{2}=0}^{\lfloor\frac{D}{2}\rfloor} \left(\frac{1}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right)^{\frac{d}{2}} \sum_{|\boldsymbol{\alpha}_{p+1}|=\frac{d}{2}} \prod_{i=1}^{n} \left(\phi + (1-\phi)\left(-\frac{\phi}{1-\phi}\right)^{\alpha_{i,p+1}}\right)^2 \langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\rangle^{\alpha_{i,p+1}}.
\end{aligned}$$

We now notice that the term inside of the product equals zero for all $\alpha_{i,p+1}$ odd if and only if $\phi = 1/2$, which is the case for SB-MSLR. So we sum only over even terms to obtain:

$$\begin{aligned}
&\chi^2_{\leq D}(\mathbb{P}(\boldsymbol{X}, \boldsymbol{y}) \| \mathbb{P}(\boldsymbol{X}) \otimes \mathbb{P}(\boldsymbol{y})) + 1 \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\beta}} \sum_{\substack{\frac{d}{2}=0 \\ \text{even}}}^{\lfloor \frac{D}{2} \rfloor} \left(\frac{1}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2} \right)^{\frac{d}{2}} \sum_{\substack{|\boldsymbol{\alpha}_{p+1}| = \frac{d}{2} \\ \text{even}}} \prod_{i=1}^{n} \langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)} \rangle^{\alpha_{i,p+1}} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\beta}} \sum_{\frac{d}{4}=0}^{\lfloor \frac{D}{4} \rfloor} \left(\frac{1}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2} \right)^{\frac{d}{2}} \sum_{\substack{|\boldsymbol{\alpha}_{p+1}| = \frac{d}{2} \\ \text{even}}} \langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)} \rangle^{|\boldsymbol{\alpha}_{p+1}|} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\beta}} \sum_{\frac{d}{4}=0}^{\lfloor \frac{D}{4} \rfloor} \left(\frac{1}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2} \right)^{\frac{d}{2}} \binom{\frac{d}{4} + n - 1}{n-1} \langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)} \rangle^{\frac{d}{2}} \\
&\leq \mathop{\mathbb{E}}_{\boldsymbol{\beta}} \sum_{\frac{d}{4}=0}^{\lfloor \frac{D}{4} \rfloor} \left(\frac{1}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2} \right)^{\frac{d}{2}} \frac{\left(\frac{D}{4} + n\right)^{\frac{d}{4}}}{\frac{d}{4}!} \langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)} \rangle^{\frac{d}{2}}.
\end{aligned}$$

■

Theorem 25 (General SB-MSLR − D lower bound) *Consider the setting of* SB-MSLR − D *with joint prior* $\mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$. *If* $k \leq \sqrt{\frac{p}{e}}$, *and* $k \leq D \leq 2(\sqrt{2}-1)\min\left\{\frac{(\|\boldsymbol{\beta}\|_2^2+\sigma^2)^2}{n\|\boldsymbol{\beta}\|_\infty^4}, n\right\}$, *we have that* $\chi^2_{\leq D}(\mathbb{P}(\boldsymbol{X}, \boldsymbol{y}) \| \mathbb{P}(\boldsymbol{X}) \otimes \mathbb{P}(\boldsymbol{y})) = O(1)$.

Proof We first apply the result of Lemma 22 to obtain:

$$\begin{aligned}
\chi^2_{\leq D}(\mathbb{P}(\boldsymbol{X}, \boldsymbol{y}) \| \mathbb{P}(\boldsymbol{X}) \otimes \mathbb{P}(\boldsymbol{y})) + 1 &\leq \sum_{\frac{d}{4}=0}^{\frac{D}{4}} \left(\frac{1}{(\|\boldsymbol{\beta}\|_2^2 + \sigma^2)^2} \right)^{\frac{d}{4}} \frac{\left(\frac{D}{4} + n\right)^{\frac{d}{4}}}{\frac{d}{4}!} \mathop{\mathbb{E}}_{\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)} \overset{\text{i.i.d.}}{\sim} \mathcal{P}} \langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)} \rangle^{\frac{d}{2}} \\
&\leq \sum_{\frac{d}{4}=0}^{\frac{D}{4}} \left(\frac{1}{(\|\boldsymbol{\beta}\|_2^2 + \sigma^2)^2} \right)^{\frac{d}{4}} \frac{\left(\frac{D}{4} + n\right)^{\frac{d}{4}}}{\frac{d}{4}!} \mathop{\mathbb{E}}_{\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)} \overset{\text{i.i.d.}}{\sim} \mathcal{P}} \langle \mathtt{abs}(\boldsymbol{\beta}^{(1)}), \mathtt{abs}(\boldsymbol{\beta}^{(2)}) \rangle^{\frac{d}{2}},
\end{aligned}$$

where $\langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)} \rangle \leq \langle \mathtt{abs}(\boldsymbol{\beta}^{(1)}), \mathtt{abs}(\boldsymbol{\beta}^{(2)}) \rangle$, and we recall $\mathtt{abs}(\boldsymbol{\beta})$ denotes the entry-wise absolute value operation on the vector $\boldsymbol{\beta}$. Notice that by Lemma 20 we have, for $p > 4, k \leq \sqrt{p}$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)} \overset{\text{i.i.d.}}{\sim} \mathcal{P}} \langle \mathtt{abs}(\boldsymbol{\beta}^{(1)}), \mathtt{abs}(\boldsymbol{\beta}^{(2)}) \rangle^{\frac{d}{2}} &\leq \sum_{l=0}^{k} \frac{\binom{k}{l}\binom{p-k}{k-l}}{\binom{p}{k}} (l \|\boldsymbol{\beta}\|_\infty^2)^{\frac{d}{2}} \\
&\leq 4 \sum_{l=0}^{k} \left(\frac{k^2}{p} \right)^{l} (l \|\boldsymbol{\beta}\|_\infty^2)^{\frac{d}{2}}.
\end{aligned}$$

We then obtain:

$$\chi^2_{\leq D}(\mathbb{P}(\boldsymbol{X},\boldsymbol{y})\|\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y}))+1\leq 4\sum_{\frac{d}{4}=0}^{\frac{D}{4}}\left(\frac{1}{(\|\boldsymbol{\beta}\|_2^2+\sigma^2)^2}\right)^{\frac{d}{4}}\frac{\left(\frac{D}{4}+n\right)^{\frac{d}{4}}}{\frac{d}{4}!}\sum_{l=0}^{k}\left(\frac{k^2}{p}\right)^l(l\|\boldsymbol{\beta}\|_\infty^2)^{\frac{d}{2}}.$$

With the aim of bounding the right hand side, we enforce condition i): $(D/4+n)D \leq (\|\boldsymbol{\beta}\|_2^2+\sigma^2)^2/\|\boldsymbol{\beta}\|_\infty^4$. After switching sums, this yields

$$\begin{aligned}\chi^2_{\leq D}(\mathbb{P}(\boldsymbol{X},\boldsymbol{y})\|\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y}))+1 &\leq 4\sum_{l=0}^{k}\left(\frac{k^2}{p}\right)^l\sum_{\frac{d}{4}=0}^{\frac{D}{4}}\frac{\left(\frac{(\|\boldsymbol{\beta}\|_2^2+\sigma^2)^2\|\boldsymbol{\beta}\|_\infty^4}{D(\|\boldsymbol{\beta}\|_2^2+\sigma^2)^2\|\boldsymbol{\beta}\|_\infty^4}\right)^{\frac{d}{4}}}{\frac{d}{4}!}l^{\frac{d}{2}}\\ &\leq 4\sum_{l=0}^{k}\left(\frac{k^2}{p}\right)^l\exp\left(\frac{l^2}{D}\right).\end{aligned}$$

We now enforce condition ii): $k\leq D$ to obtain the result,

$$\begin{aligned}\chi^2_{\leq D}(\mathbb{P}(\boldsymbol{X},\boldsymbol{y})\|\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y}))+1 &\leq 4\sum_{l=0}^{k}\left(\frac{k^2}{p}\right)^l\exp(l)\\ &=4\sum_{l=0}^{k}\left(\frac{k^2e}{p}\right)^l=O(1).\end{aligned}$$

Note that conditions i) and ii) are satisfied for any $n>0$ and $k\leq D\leq 2(\sqrt{2}-1)\min\left\{\frac{(\|\boldsymbol{\beta}\|_2^2+\sigma^2)^2}{n\|\boldsymbol{\beta}\|_\infty^4},n\right\}$.
■

Appendix C. Proofs of Polynomial-Time Reductions

Consider signed support recovery in the MSLR problem, where we seek to recover the support of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, along with the along with the signs of their entries. Take $(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)\sim\mathcal{P}_{\|\beta\|_2}(\{-1,1\})$, and let $\mathcal{S}_1:=\operatorname{supp}(\boldsymbol{\beta}_1)=\{j\in[p]:\boldsymbol{\beta}_{1,j}\neq 0\}$, and $\mathcal{S}_2$ defined similarly for $\boldsymbol{\beta}_2$. We study the computational hardness of the problem as we vary two parameters of our joint signal distribution, the *overlap* ξ and the *signed overlap* τ respectively:

$$\xi=\frac{|\mathcal{S}_1\cap\mathcal{S}_2|}{k},\qquad \tau=\frac{\langle\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\rangle}{|\mathcal{S}_1\cap\mathcal{S}_2|},$$

that are of constant order, i.e., do not scale with respect to n,p,k. Previous work (Gamarnik and Zadik, 2022) studies exact support recovery in sparse linear regression and the computational hardness that arises from the overlap distribution of two identical copies of the signal. We extend the analysis by considering exact *signed* support recovery by varying the parameter τ, which measures the relative frequency of $+1$ and -1 entries with the same index. Note that for $\xi=1,\tau=1$ we have the usual SLR problem, for $\xi=1,\tau=-1$ we have the SB-MSLR regime, and importantly for $\phi=1/2,\tau\in(-1,1),\xi>0$ we have the PSB-MSLR regime. We denote $\mathtt{MSLR}_{\xi,\tau}$ and $\mathtt{MSLR}-\mathtt{D}_{\xi,\tau}$ as the $\texttt{MSLR}$ and $\texttt{MSLR}-\texttt{D}$ problems with the joint signal prior $\mathcal{P}_{\|\beta\|_2}(\{-1,1\})$ constrained to signals $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ with overlap and signed overlap ξ and τ respectively.

Using these definitions, we first form in Lemma 26 a polynomial-time reduction from $\texttt{SB-MSLR}-\texttt{D}$ to exact signed support recovery in $\texttt{SB-MSLR}$ within the scaling regime of Theorem 4. Notice that $\texttt{MSLR}-\texttt{D}_{1,-1}$ is equivalent to the $\texttt{SB-MSLR}-\texttt{D}$ problem. Next, we prove a polynomial-time reduction from exact signed support recovery in $\texttt{SB-MSLR}$ to exact signed support recovery in $\texttt{MSLR}_{\xi,\tau}$ for $\tau \in (-1,1), \xi > 0$ ($\texttt{PSB-MSLR}$) within the scaling regime of Theorem 4, proving that if exact signed support recovery can be achieved in $\texttt{PSB-MSLR}$, then it can also be achieved in $\texttt{SB-MSLR}$.

Combining the two arguments above, we have that solving exact signed support recovery in $\texttt{PSB-MSLR}$ implies solving strong detection in $\texttt{SB-MSLR}-\texttt{D}$, which would contradict the implication in Theorem 4 that $\texttt{SB-MSLR}-\texttt{D}$ cannot be solved in polynomial time, resulting in Theorem 28. For more background on the logic of average-case reductions, we refer to (Brennan and Bresler, 2020b).

The reduction from $\texttt{SLR}-\texttt{D}$ to $\texttt{SLR}$ is nearly identical to that in Lemma 26 with the midly less restrictive condition that $\texttt{SNR} \geq 1$, and hence the proof is omitted.

Throughout the proofs, we use the following measure of recovery error for mixtures of linear regressions (Chen et al., 2014):

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

This error measure takes into account recovery of the two signals up to relabelling. For vectors a, b, $\hat{a}$, $\hat{b} \in \mathbb{R}^p$, define

$$\|(\hat{a}, \hat{b}) - (a,b)\|_\infty := \min\left\{\|\hat{a} - a\|_\infty + \|\hat{b} - b\|_\infty, \|\hat{b} - a\|_\infty + \|\hat{a} - b\|_\infty\right\}.$$

Notice that for $\epsilon \in [0,1)$ and signals $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \sim \mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\{-1,1\})$ we have that

$$\mathbb{P}\left[\rho((\hat{\boldsymbol{\beta}}_1,\, \hat{\boldsymbol{\beta}}_2), (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)) > \epsilon\right] \to 0 \iff \mathbb{P}\left[\|(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) - (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\|_\infty > \epsilon\right] \to 0,$$

Main Statements We begin by defining the parameter regimes of interest:

$$\begin{aligned}\mathcal{C}_1 = \Big\{&(p_i, n_i, k_i, \sigma_i)_{i=1}^\infty \subset \mathbb{N}^4 : \; p_i = \omega_i(1), k_i = o(\sqrt{p_i}), n_i = \omega(\max\{k_i, \log p_i\}), \\ &n_i = o\left((k_i + \sigma_i^2)^2 \cdot \frac{1}{\log p_i}\right)\Big\}.\end{aligned} \tag{7}$$

$$\begin{aligned}\mathcal{C}_2 = \Big\{&(p_i, n_i, k_i, \sigma_i)_{i=1}^\infty \subset \mathbb{N}^4 : \; p_i = \omega_i(1), k_i = o(\sqrt{p_i}), n_i = \omega(\max\{k_i, \log p_i\}), \\ &n_i \gtrsim \frac{k_i \log p_i}{\log(1 + \frac{k_i}{\sigma_i^2})}\Big\}.\end{aligned} \tag{8}$$

Notice that $\mathcal{C}_1, \mathcal{C}_2$ are both contained within the parameter regime where $\texttt{SB-MSLR}-\texttt{D}$ encounters a computational barrier, as per Theorem 4. The following lemmas consist of two sub-reductions which together give the reduction argument from $\texttt{SB-MSLR}-\texttt{D}$ to exact recovery in $\texttt{PSB-MSLR}$.

Lemma 26 *Let $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) {\sim} \mathcal{P}_{\|\beta\|_2}(\{-1,1\})$, $SNR = \omega(1)$. Given a sequence of parameters $\{(p_i, n_i, k_i, \sigma_i)\}_{i=1}^\infty$ in $\mathcal{C}_2$ for $\texttt{SB-MSLR}-\texttt{D}$ and $\texttt{SB-MSLR}$, if for any $\epsilon > 0$ there exists a randomized polynomial-time algorithm $\mathcal{A}$ for $\texttt{SB-MSLR}$ producing $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ with $\mathbb{P}\left[\|(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) - (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\|_\infty < \epsilon\right] \underset{(i\to\infty)}{\to} 1$, then there exists a randomized polynomial-time detection algorithm $\mathcal{A}'$ for $\texttt{SB-MSLR}-\texttt{D}$ with vanishing Type I+II errors as $i \to \infty$.*

The proof of Lemma 26 is given in Section C.1.

Lemma 27 *Fix signal priors to be $\mathcal{P}_{\|\beta\|_2}(\{-1,1\})$. For any sequence of parameters $\{(p_i', n_i', k_i', \sigma_i')\}_{i=1}^{\infty}$ in $\mathcal{C}_1$ for* `PSB-MSLR` *with solution $\boldsymbol{\beta}_1', \boldsymbol{\beta}_2'$ and problem instances $(\boldsymbol{X}', \boldsymbol{y}')$, there exists a sequence of parameters $\{(p_i, n_i, k_i, \sigma_i)\}_{i=1}^{\infty}$ in $\mathcal{C}_1$ for* `SB-MSLR` *with solution $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and problem instances $(\boldsymbol{X}, \boldsymbol{y})$ such that, for any randomized polynomial time algorithm $\mathcal{A}'$ for* `PSB-MSLR` *outputting $(\hat{\boldsymbol{\beta}}'_1, \hat{\boldsymbol{\beta}}'_2)$ with*

$$\mathbb{P}\left[\|(\hat{\boldsymbol{\beta}}'_1, \hat{\boldsymbol{\beta}}'_2) - (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')\|_\infty > 0\right] \to 0,$$

we can construct a second randomized polynomial time algorithm $\mathcal{A}$ for `PSB-MSLR` *outputting $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ such that*

$$\mathbb{P}\left[\|(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) - (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\|_\infty > 0\right] \to 0.$$

The proof of Lemma 27 is given in Section C.2.

Theorem 28 (Reduction from `SB-MSLR` − `D` to exact recovery in `PSB-MSLR`) *Consider the setting of* `PSB-MSLR` *(2) with joint signal prior $\mathcal{P}_{\|\beta\|_2}(\{-1,1\})$. Any randomized polynomial-time algorithm $\mathcal{A}$ solving* `PSB-MSLR` *within parameter regimes $\mathcal{C}_1 \cap \mathcal{C}_2$ and with $SNR = \omega(1)$ would contradict Theorem 4.*

Proof Suppose there exists a randomized polynomial-time algorithm $\mathcal{A}$ solving exact recovery in `PSB-MSLR` with signals in $\{-1,0,1\}^p$ and parameter regime contained in $\mathcal{C}_1$ defined in (7). Then by Lemma 27 we would have a randomized polynomial time algorithm $\mathcal{A}'$ solving exact recovery in `SB-MSLR` within this regime. By Lemma 26, we would then consequently have a polynomial-time algorithm solving `SB-MSLR` − `D` in the scaling regime $\mathcal{C}_1 \cap \mathcal{C}_2$, which is contained in the scaling regime of Theorem 4 and hence contradicts Theorem 4. ■

Remark 29 *Note that the lower bounds $\frac{k \log p}{\log(1+SNR)}$ in constraint $\mathcal{C}_1$ in (7) used in Theorem 28 are not restrictive as this is the information-theoretic minimal sample complexity for support recovery in* `SLR` *(Reeves et al., 2019; Gamarnik and Zadik, 2022; Wang et al., 2010).*

C.1. Reduction from `SB-MSLR` − `D` to `SB-MSLR`

We utilize a variant of a theorem in Gamarnik and Zadik (2017) to construct our reduction, Lemma 30. Consider the following optimization problem for $\sigma > 0$:

$$\begin{aligned} \psi := \min \quad & n^{-\frac{1}{2}}\|\sigma \boldsymbol{w} - \boldsymbol{X}\boldsymbol{\beta}_1 \odot \boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}_2 \odot (1-\boldsymbol{z})\|_2 \\ \text{s.t.} \quad & \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \{-1,0,1\}^p, \ \boldsymbol{z} \in \{0,1\}^n \\ & \|\boldsymbol{\beta}_1\|_0 = \|\boldsymbol{\beta}_2\|_0 = k, \end{aligned} \tag{9}$$

where $\boldsymbol{X} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, independent from $w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.

Lemma 30 *Let ψ be as defined in (9). For $\delta > 0$ we have:*

$$\mathbb{P}\left[\psi \geq e^{-(1+\delta)/2} \exp\left(-\frac{2k(\log p + 1)}{n}\right)\sqrt{k+\sigma^2}\right] \geq 1 - e^{-\frac{\delta}{2}n}.$$

The proof follows by nearly identical arguments as that of Theorem 3.1 in (Gamarnik and Zadik, 2017) and is hence omitted.

Proof [Proof of Lemma 26] Throughout the proof, we drop the i subscript in the parameters $(p_i, n_i, k_i, \sigma_i)$ for convenience. We refer to $\texttt{SB-MSLR}$ and $\texttt{SB-MSLR}-\texttt{D}$ as $\texttt{MSLR}_{\xi,\tau}$ and $\texttt{MSLR}-\texttt{D}_{\xi,\tau}$ respectively, with $\xi = 1, \tau = -1$. We take $\mathrm{P} := \mathbb{P}(\boldsymbol{X}, \boldsymbol{y})$ to represent the planted measure in the formulation of $\texttt{SLR}-\texttt{D}$, and $\mathrm{Q} := \mathbb{P}(\boldsymbol{X}) \otimes \mathbb{P}(\boldsymbol{y})$ to represent the null measure. We emphasize that, since $\boldsymbol{\beta} \sim \mathcal{P}_{\|\beta\|_2}(\{-1,1\})$, $\|\boldsymbol{\beta}\|_2^2$ and k are interchangeable. As prescribed in the statement of the lemma, suppose that $\mathbb{P}\left[\rho((\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2), (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)) < \epsilon\right] \underset{(i\to\infty)}{\to} 1$ for any $\epsilon > 0$.

Define the two following events under the planted hypothesis P:

$$\tilde{\Omega}_1 := \left\{ \{\hat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1, \hat{\boldsymbol{\beta}}_2 = \boldsymbol{\beta}_2\} \cup \{\hat{\boldsymbol{\beta}}_2 = \boldsymbol{\beta}_1, \hat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_2\} \right\},$$

and

$$\tilde{\Omega}_2 := \{|w_q| < |\sigma^{-1}\langle \boldsymbol{X}_q, \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\rangle + w_q|, \;\; \forall q \in [n]\}.$$

Note that by assumption, $\tilde{\Omega}_1$ occurs with probability $1 - o(1)$ under P. Indeed, we can choose $\epsilon < 1$ in the definition of our given algorithm $\mathcal{A}$ and since $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \{-1,0,1\}^p$ we obtain that $\mathbb{P}\left[\tilde{\Omega}_1\right] \to 1$. We first consider the planted hypothesis P. Let $\nu_q \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, (1-\xi\tau)\frac{k}{\sigma^2})$, and $g_q \sim \mathcal{N}(0,1)$, independent from each other and from w_q, for $q \in [n]$. In this case we have by symmetry that

$$\begin{aligned}
\mathrm{P}\left[\tilde{\Omega}_2^{\complement}\middle|\tilde{\Omega}_1\right] &= \mathrm{P}\left[\{|w_q| \geq |\sigma^{-1}\langle \boldsymbol{X}_q, \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\rangle + w_q|, \;\; \forall q \in [n]\}\right] \\
&= \int \mathrm{P}\left[\{|w_q| \geq |\nu_q + w_q|, \;\; \forall q \in [n]\}|w_q\right] \mathrm{P}[dw_q] \\
&= 2\int_0^\infty \mathrm{P}\left[\nu_q \in [-2w_q, 0], \;\; \forall q \in [n]\}|w_q\right] \mathrm{P}[dw_q] \\
&= 2\int_0^\infty \mathrm{P}\left[\{g_q \in [-2w_q/((1-\xi\tau)k/\sigma^2), 0], \;\; \forall q \in [n]\}\middle|w_q\right] \mathrm{P}[dw_q],
\end{aligned}$$

where we have

$$\mathrm{P}\left[\{g_q \in [-2w_q/((1-\xi\tau)k/\sigma^2), 0], \;\; \forall q \in [n]\}\middle|w_q\right] \to 0 \quad \text{as } k/\sigma^2 \to \infty,$$

and $\mathrm{P}\left[\{g \in [-2w_q/((1-\xi\tau)k/\sigma^2), 0], \;\; \forall q \in [n]\}\middle|w_q\right] \leq 1$, so we can apply the Dominated Convergence Theorem to obtain that

$$\mathrm{P}\left[\tilde{\Omega}_2^{\complement}\middle|\tilde{\Omega}_1\right] = 2\int_0^\infty \mathrm{P}\left[\{g_q \in [-2w_q/((1-\xi\tau)k/\sigma^2), 0], \;\; \forall q \in [n]\}\middle|w_q\right] \mathrm{P}[dw_q] \to 0 \text{ as } k/\sigma^2 \to \infty.$$

We therefore have that $\mathrm{P}\left[\tilde{\Omega}_2\middle|\tilde{\Omega}_1\right] = 1 - o(1)$, and hence $\mathrm{P}\left[\left(\tilde{\Omega}_1, \tilde{\Omega}_2\right)\right] = 1 - o(1)$.

Next, note that under the planted hypothesis P and in the joint event $(\tilde{\Omega}_1, \tilde{\Omega}_2)$ we have for indices q such that $z_q = 1$:

$$\begin{aligned}
|y_q - \frac{1}{\sigma}\langle \boldsymbol{X}_q, \hat{\boldsymbol{\beta}}_1\rangle| &= |y_q - \frac{1}{\sigma}\langle \boldsymbol{X}_q, \boldsymbol{\beta}_1\rangle| \\
&= |w_q| \\
&< |\frac{1}{\sigma}\langle \boldsymbol{X}_q, \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\rangle + w_q| \\
&= |y_q - \frac{1}{\sigma}\langle \boldsymbol{X}_q, \boldsymbol{\beta}_2\rangle| \\
&= |y_q - \frac{1}{\sigma}\langle \boldsymbol{X}_q, \hat{\boldsymbol{\beta}}_2\rangle|.
\end{aligned}$$

An analogous statement with $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ swapped holds for indices q such that $z_q = 0$. We therefore have that under $(\tilde{\Omega}_1, \tilde{\Omega}_2)$ we can exactly estimate $\boldsymbol{z}$ using the above thresholding procedure, and we call this exact estimate $\hat{\boldsymbol{z}}$. We then define our detection algorithm in this case:

$$\mathcal{A}'\left(\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{y}\end{bmatrix}\right) = \begin{cases}\mathtt{p}, & n^{-1/2}\|\boldsymbol{y} - \frac{1}{\sigma}\boldsymbol{X}\hat{\boldsymbol{\beta}}_1 \odot \hat{\boldsymbol{z}} - \frac{1}{\sigma}\boldsymbol{X}\hat{\boldsymbol{\beta}}_2 \odot (1-\hat{\boldsymbol{z}})\| \le \sqrt{5} \\ \mathtt{q}, & n^{-1/2}\|\boldsymbol{y} - \frac{1}{\sigma}\boldsymbol{X}\hat{\boldsymbol{\beta}}_1 \odot \hat{\boldsymbol{z}} - \frac{1}{\sigma}\boldsymbol{X}\hat{\boldsymbol{\beta}}_2 \odot (1-\hat{\boldsymbol{z}})\| > \sqrt{5}\end{cases}.$$

We will proceed to prove that $\mathcal{A}'$ has vanishing Type II error. Indeed, under P and under the high-probability event $(\tilde{\Omega}_1, \tilde{\Omega}_2)$ we have:

$$\begin{aligned}
&\|\boldsymbol{y} - \frac{1}{\sigma}\boldsymbol{X}\hat{\boldsymbol{\beta}}_1 \odot \hat{\boldsymbol{z}} - \frac{1}{\sigma}\boldsymbol{X}\hat{\boldsymbol{\beta}}_2 \odot (1-\hat{\boldsymbol{z}})\|_2 \\
&= \|\frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_1 \odot \boldsymbol{z} + \frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_2 \odot (1-\boldsymbol{z}) + \boldsymbol{w} - \frac{1}{\sigma}\boldsymbol{X}\hat{\boldsymbol{\beta}}_1 \odot \hat{\boldsymbol{z}} - \frac{1}{\sigma}\boldsymbol{X}\hat{\boldsymbol{\beta}}_2 \odot (1-\hat{\boldsymbol{z}})\|_2 \\
&\le \|\boldsymbol{w}\|_2 + \|\frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_1 \odot \boldsymbol{z} + \frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_2 \odot (1-\boldsymbol{z}) - \frac{1}{\sigma}\boldsymbol{X}\hat{\boldsymbol{\beta}}_1 \odot \hat{\boldsymbol{z}} - \frac{1}{\sigma}\boldsymbol{X}\hat{\boldsymbol{\beta}}_2 \odot (1-\hat{\boldsymbol{z}})\|_2 \\
&= \|\boldsymbol{w}\|_2.
\end{aligned}$$

We therefore have

$$\begin{aligned}
&\mathrm{P}\left[\mathcal{A}'\left(\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{y}\end{bmatrix}\right) = \mathtt{q}\right] \\
&= \mathrm{P}\left[\mathcal{A}'\left(\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{y}\end{bmatrix}\right) = \mathtt{q}\middle|(\tilde{\Omega}_1, \tilde{\Omega}_2)^{\complement}\right] \cdot \mathrm{P}\left[(\tilde{\Omega}_1, \tilde{\Omega}_2)^{\complement}\right] + \mathrm{P}\left[\left\{\mathcal{A}'\left(\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{y}\end{bmatrix}\right) = \mathtt{q}\right\} \cap (\tilde{\Omega}_1, \tilde{\Omega}_2)\right] \\
&\le \mathrm{P}\left[\mathcal{A}'\left(\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{y}\end{bmatrix}\right) = \mathtt{q}\middle|(\tilde{\Omega}_1, \tilde{\Omega}_2)^{\complement}\right] \cdot \mathrm{P}\left[(\tilde{\Omega}_1, \tilde{\Omega}_2)^{\complement}\right] + \mathrm{P}\left[\left\{\|\boldsymbol{w}\|_2 > \sqrt{5n}\right\} \cap (\tilde{\Omega}_1, \tilde{\Omega}_2)\right] \\
&\le 1 \cdot o(1) + \mathrm{P}\left[\left\{\|\boldsymbol{w}\|_2 > \sqrt{5n}\right\}\right] \\
&\le 1 \cdot o(1) + e^{-n} = o(1),
\end{aligned}$$

where the last inequality is obtained using a standard chi-square large deviation tail bounds (Example 2.11 of Wainwright (2019)):

$$\mathrm{P}\left[\|\boldsymbol{w}\|_2 < \sqrt{n + 2\sqrt{nt} + 2t}\right] \ge 1 - e^{-t},$$

and taking $t = \sqrt{n}$.

We now turn to showing that $\mathcal{A}'$ has vanishing Type I error. Under the null hypothesis Q we have by definition of ψ in Definition 9 that

$$\|\boldsymbol{y} - \frac{1}{\sigma}\boldsymbol{X}\hat{\boldsymbol{\beta}}_1 \odot \hat{\boldsymbol{z}} - \frac{1}{\sigma}\boldsymbol{X}\hat{\boldsymbol{\beta}}_2 \odot (1 - \hat{\boldsymbol{z}})\|_2 \geq \psi,$$

where we recall from Lemma 30 that

$$\mathrm{Q}\left[\psi \geq e^{-3/2} \exp\left(-\frac{2k(\log p + 1)}{n}\right)\sqrt{k + \sigma^2}\right] \geq 1 - e^{-n},$$

hence it would suffice to show that

$$e^{-3/2} \exp\left(-\frac{2k(\log p + 1)}{n}\right)\sqrt{k + \sigma^2} > \sqrt{5}.$$

In order to do so, choose $n^* = \frac{4k(\log p+1)}{\log(1+\frac{k}{\sigma^2}) - \log 5 - 3} = \Theta(\frac{4k \log p}{\log(1+\frac{k}{\sigma^2})})$ and notice that if the inequality holds for n^*, it must hold for all $n \geq n^*$ since the left hand side is increasing with n. We plug in n^* to obtain

$$e^{-3/2} \exp\left(-1/2 \log(1 + \frac{k}{\sigma^2}) + \log\sqrt{5} + 3/2\right)\sqrt{1 + 2\frac{k}{\sigma^2}} = \frac{\sqrt{1+2\frac{k}{\sigma^2}}}{\sqrt{1+\frac{k}{\sigma^2}}}\sqrt{5} > \sqrt{5},$$

and therefore we have that

$$\begin{aligned}
&\mathrm{Q}\left[\mathcal{A}'\left(\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{y}\end{bmatrix}\right) = \mathsf{q}\right] \\
&= \mathrm{Q}\left[\|\boldsymbol{y} - \frac{1}{\sigma}\boldsymbol{X}\hat{\boldsymbol{\beta}}_1 \odot \hat{\boldsymbol{z}} - \frac{1}{\sigma}\boldsymbol{X}\hat{\boldsymbol{\beta}}_2 \odot (1 - \hat{\boldsymbol{z}})\|_2 > \sqrt{5}\right] \\
&\geq \mathrm{Q}\left[\psi > \sqrt{5}\right] \\
&\geq \mathrm{Q}\left[\psi \geq e^{-3/2} \exp\left(-\frac{2k(\log p + 1)}{n}\right)\sqrt{k + \sigma^2}\right] \\
&\geq 1 - e^{-n}.
\end{aligned}$$

Importantly, note that n satisfies the constraints of $\mathcal{C}_2$.

■

C.2. Reduction from SB-MSLR to PSB-MSLR

Proof [Proof of Lemma 27] First, recall that the PSB-MSLR regime implies,

$$\phi = 1/2 \text{ and } \boldsymbol{\beta}_{1,j} = -\boldsymbol{\beta}_{2,j} \text{ for } j \in J \subseteq \mathrm{supp}(\boldsymbol{\beta}_1) \cap \mathrm{supp}(\boldsymbol{\beta}_2) \text{ with } C_1 k \leq |J| \leq C_2 k,$$

for some constants $1 \geq C_1, C_2 > 0$. Without loss of generality, we can take $|J| = Ck$ for some constant $0 < C \leq 1$, and all constants that follow can be lower bounded or upper bounded accordingly. In light of this, PSB-MSLR corresponds to $\mathtt{MSLR}_{\xi,\tau}$ for some $\tau \in (-1, 1)$ and some $\xi > 0$, where we recall that ξ, τ are of constant order, i.e., do not scale with respect to n, p, k.

For brevity, we denote $\mathfrak{P}' :=$ PSB-MSLR and $\mathfrak{P} :=$ SB-MSLR. Let $c = 1 - \frac{\tau\cdot\xi+1}{2} \in (0,1)$ denote the proportion of matching non-zero entries with opposite sign between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ (intuitively, this corresponds to the "hard" portion of the signal), and note that it is fixed. Note that this follows since $\tau \cdot \xi = \frac{\langle \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \rangle}{k}$ for $\xi > 0$.

Given a sequence of parameters $\{p_i', n_i', k_i', \sigma_i'\}_{i=1}^\infty \subseteq \mathcal{C}_1$ for $\mathfrak{P}'$, consider the sequence of parameters $\{(p_i, n_i, k_i, \sigma_i)\}_{i=1}^\infty = \{(cp_i', n_i', ck_i', \sigma_i')\}_{i=1}^\infty$ for $\mathfrak{P}$. Notice that $\{(p_i, n_i, k_i, \sigma_i)\}_{i=1}^\infty \subseteq \mathcal{C}_1$ since (dropping the subscript i notation for convenience):

- $k = ck' = o(cC\sqrt{p'}) = o(c^{3/2}C\sqrt{p}) = o(\sqrt{p})$
- $n = n' = o\left((k' + (\sigma')^2)^2 \cdot \frac{1}{\log p'}\right) = o\left((k+\sigma^2)^2 \cdot \frac{1}{\log p}\right)$,
- $n = n' = \omega(\max\{k', \log p'\}) = \omega(\max\{k, \log p\})$,

and hence the parameter regimes of $\mathfrak{P}$ are also contained in $\mathcal{C}_1$. For $i \in \mathbb{N}$, let $J = (\boldsymbol{X}, \boldsymbol{y})$ denote an instance of $\mathfrak{P}$ with parameters $(p_i, n_i, k_i, \sigma_i)$, where we recall:

$$\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} \\ \frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_1 \odot \boldsymbol{z} + \frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_2 \odot (1 - \boldsymbol{z}) + \boldsymbol{w} \end{bmatrix}.$$

We now want to show that, given a sequence of parameters $\{p_i', n_i', k_i'\}_{i=1}^\infty \subseteq \mathcal{C}_1$ for $\mathfrak{P}'$ and a randomized polynomial-time algorithm $\mathcal{A}'$ solving it, we can construct a randomized polynomial-time algorithm $\mathcal{A}$ solving $\mathfrak{P}$ along the above parameter sequence $\{p_i, n_i, k_i\}_{i=1}^\infty \subseteq \mathcal{C}_1$. We will construct our desired algorithm $\mathcal{A}$ by composing $\mathcal{A}'$ with a pre- and post-processing step. Indeed, we let $\mathcal{A} = \mathcal{B} \circ \mathcal{A}' \circ \mathcal{D}$, where we define $\mathcal{B}$ and $\mathcal{D}$ below.

First, let RS denote the random variable that reshuffles entries of a given size p vector or the columns of a given p-column matrix according to a uniform shuffling of the index set $[p]$. Let IRS denote the random variable which inverts this reshuffling process on a vector or matrix, such that IRS $\circ$ RS is the identity operation. Let $\mathbb{1}$ denote the all-ones vector of any size (to be inferred from context). Let $\bar{\mathbb{1}}_1$ and $\bar{\mathbb{1}}_2$ denote two independent copies of a $(1-c)k'$ sparse vector in $\{0, 1, -1\}^{(1-c)p'}$. We now define our pre- and post-processing procedures Algorithms 1, 2, which can be seen to run in randomized polynomial time with respect to p.

Algorithm 1: $\mathcal{D}(\boldsymbol{X}, \boldsymbol{y})$, where $(\boldsymbol{X}, \boldsymbol{y})$ is an instance of $\mathfrak{P}(p, n, k, \sigma)$
Data: $\boldsymbol{V} \in \mathbb{R}^{n \times (1-c)/c\, p}$ with columns $\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_n)$
$\tilde{\boldsymbol{y}} \leftarrow \boldsymbol{y} + \frac{1}{\sigma}\boldsymbol{V}\mathbb{1}$
$\tilde{\boldsymbol{X}} \leftarrow \mathtt{RS}\begin{bmatrix}\boldsymbol{V} & \boldsymbol{X}\end{bmatrix}$
return $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}})$

Algorithm 2: $\mathcal{B}(\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2)$, where $(\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2)$ are both in $\{0, 1, -1\}^{p'}$
$\begin{bmatrix} \bar{\mathbb{1}}_1 \\ \boldsymbol{\beta}_1 \end{bmatrix} \leftarrow \mathtt{IRS}(\tilde{\boldsymbol{\beta}}_1)$
$\begin{bmatrix} \bar{\mathbb{1}}_2 \\ \boldsymbol{\beta}_2 \end{bmatrix} \leftarrow \mathtt{IRS}(\tilde{\boldsymbol{\beta}}_2)$
return $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$

Proposition 31 *For $i \in \mathbb{N}$, $(\boldsymbol{X}, \boldsymbol{y})$ an instance of $\mathfrak{P}(p_i, n_i, k_i, \sigma_i)$, and $(\boldsymbol{X}', \boldsymbol{y}')$ an instance of $\mathfrak{P}'(p_i', n_i', k_i')$, we have that $\mathcal{D}(\boldsymbol{X}, \boldsymbol{y}) \overset{d}{=} (\boldsymbol{X}', \boldsymbol{y}')$.*

Proof Let $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}}) := \mathcal{D}(\boldsymbol{X}, \boldsymbol{y})$. First note that $\tilde{\boldsymbol{X}} \overset{d}{=} \boldsymbol{X}'$ since $\begin{bmatrix}\boldsymbol{V} & \boldsymbol{X}\end{bmatrix}$ has dimensions $n \times (p + \frac{1-c}{c}p) = n' \times (cp' + (1-c)p') = n' \times p'$. Next, note that we can decompose $\tilde{\boldsymbol{y}}$ as follows:

$$\begin{aligned}
\tilde{\boldsymbol{y}} &= \boldsymbol{y} + \frac{1}{\sigma}\boldsymbol{V}\mathbb{1} \\
&= \frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_1 \odot \boldsymbol{z} + \frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_2 \odot (1 - \boldsymbol{z}) + \frac{1}{\sigma}\boldsymbol{V}\mathbb{1} + w \\
&= \frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_1 \odot \boldsymbol{z} + \frac{1}{\sigma}\boldsymbol{X}\boldsymbol{\beta}_2 \odot (1 - \boldsymbol{z}) + \frac{1}{\sigma}\boldsymbol{V}\mathbb{1} \odot \boldsymbol{z} + \frac{1}{\sigma}\boldsymbol{V}\mathbb{1} \odot (1 - \boldsymbol{z}) + w \\
&= \frac{1}{\sigma}(\texttt{RS}\begin{bmatrix}\boldsymbol{V} & \boldsymbol{X}\end{bmatrix}) \left(\texttt{RS}\begin{bmatrix}\bar{\mathbb{1}}_1 \\ \boldsymbol{\beta}_1\end{bmatrix}\right) \odot \boldsymbol{z} + \frac{1}{\sigma}\left(\texttt{RS}\begin{bmatrix}\boldsymbol{V} & \boldsymbol{X}\end{bmatrix}\right)\left(\texttt{RS}\begin{bmatrix}\bar{\mathbb{1}}_2 \\ \boldsymbol{\beta}_2\end{bmatrix}\right) \odot (1 - \boldsymbol{z}) + w
\end{aligned}$$

and hence $\tilde{\boldsymbol{y}}$ is the output of a $\texttt{MSLR}$ model with size $cp' + (1-c)p' = p'$ signals $\left(\texttt{RS}\begin{bmatrix}\bar{\mathbb{1}}_1 \\ \boldsymbol{\beta}_1\end{bmatrix}, \texttt{RS}\begin{bmatrix}\bar{\mathbb{1}}_2 \\ \boldsymbol{\beta}_2\end{bmatrix}\right)$ containing ck' non-zero opposing sign entries from $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $(1-c)k'$ remaining non-zero entries from appending $\bar{\mathbb{1}}$, for a total support of size k'. Additionally, these signals are linearly transformed by a design matrix that is i.i.d Gaussian and $n' \times p'$ as mentioned above. The first and second point together imply that $\tilde{\boldsymbol{y}}$ is the output of a $\texttt{MSLR}_{\xi,\tau}$ model and $\mathcal{D}(\boldsymbol{X}, \boldsymbol{y}) = (\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}}) \overset{d}{=} (\boldsymbol{X}', \boldsymbol{y}')$. ∎

Following Proposition 31, all that is left to show is that $\mathbb{P}[\|\mathcal{B} \circ \mathcal{A}' \circ \mathcal{D}(\boldsymbol{X}, \boldsymbol{y}) - (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\|_\infty > 0] \to 0$. Indeed, the following steps hold due to Proposition 31 and the definition of $\mathcal{B}$:

$$\begin{aligned}
&\|\mathcal{A}'(\boldsymbol{X}', \boldsymbol{y}') - (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')\|_\infty \\
&\overset{d}{=} \left\|\mathcal{A}'(\texttt{RS}(\begin{bmatrix}\boldsymbol{V} & \boldsymbol{X}\end{bmatrix}), \boldsymbol{y} + \frac{1}{\sigma}\boldsymbol{V}\mathbb{1}) - \left(\texttt{RS}\begin{bmatrix}\bar{\mathbb{1}}_1 \\ \boldsymbol{\beta}_1\end{bmatrix}, \texttt{RS}\begin{bmatrix}\bar{\mathbb{1}}_2 \\ \boldsymbol{\beta}_2\end{bmatrix}\right)\right\|_\infty \\
&= \left\|\mathcal{A}' \circ \mathcal{D}(\boldsymbol{X}, \boldsymbol{y}) - \left(\texttt{RS}\begin{bmatrix}\bar{\mathbb{1}}_1 \\ \boldsymbol{\beta}_1\end{bmatrix}, \texttt{RS}\begin{bmatrix}\bar{\mathbb{1}}_2 \\ \boldsymbol{\beta}_2\end{bmatrix}\right)\right\|_\infty \\
&\geq \left\|\mathcal{B} \circ \mathcal{A}' \circ \mathcal{D}(\boldsymbol{X}, \boldsymbol{y}) - \mathcal{B}\left(\texttt{RS}\begin{bmatrix}\bar{\mathbb{1}}_1 \\ \boldsymbol{\beta}_1\end{bmatrix}, \texttt{RS}\begin{bmatrix}\bar{\mathbb{1}}_2 \\ \boldsymbol{\beta}_2\end{bmatrix}\right)\right\|_\infty \\
&= \|\mathcal{A}(\boldsymbol{X}, \boldsymbol{y}) - (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\|_\infty,
\end{aligned}$$

and hence:

$$\mathbb{P}\left[\|\mathcal{A}'(\boldsymbol{X}', \boldsymbol{y}') - (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')\|_\infty > 0\right] \to 0 \implies \mathbb{P}\left[\|\mathcal{A}(\boldsymbol{X}, \boldsymbol{y}) - (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\|_\infty > 0\right] \to 0,$$

where $\mathcal{A}$ runs in randomized polynomial time with respect to the input size p since it is a composition of randomized polynomial time procedures with respect to p. ∎

C.3. Reduction from SB-MSLR − D to SPR

In this section, we present a polynomial-time reduction from strong detection in noiseless SB-MSLR − D to exact support recovery in SPR , for signals with non-zero entries in $\{-1,0,1\}^p$. More specifically, we consider the following *symmetric* SLR problem.

Definition 32 (S − SLR) *For $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{w} \in \mathbb{R}^n$, and $\boldsymbol{z} \in \mathbb{R}^n$, consider the model:*

$$\boldsymbol{y} = g\left(\boldsymbol{X}\boldsymbol{\beta}\right) + \boldsymbol{w},$$

where $\odot$ denotes element-wise product between vectors, $X_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $w_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$, $\boldsymbol{\beta} \in \mathbb{R}^p$ each k-sparse, and $g : \mathbb{R}^n \to \mathbb{R}^n$ a separable entry-wise even function ($g_i(x) = g_i(-x)$ for $i \in [n]$ and $x \in \mathbb{R}$). Given $(\boldsymbol{X}, \boldsymbol{y})$ the objective is to estimate $\boldsymbol{\beta}$.

We note that setting $g_i(x) = |x|$ and $g_i(x) = x^2$ yields two standard formulations of the phase retrieval problem with sparse signals (SPR) (Candès et al., 2015; Liu et al., 2021). We seek to show hardness within the parameter scaling regime of Theorem 4 in the noiseless case ($\frac{\mathrm{SNR}+1}{\mathrm{SNR}} = 1$) , and hence we prove a reduction within the constraint set $\mathcal{C}_3$:

$$\mathcal{C}_3 = \Big\{(p_i, n_i, k_i, \sigma_i)_{i=1}^{\infty} \subset \mathbb{N}^4 : \exists C \in \mathbb{R}_+ \text{ s.t. } p_i = \omega_i(1), k_i = o(\sqrt{p_i}), , \\ n_i = \omega(\max\{k_i, \log p_i\})\, n_i = o\left(k_i^2 \cdot \frac{1}{\log p_i}\right)\Big\}. \tag{10}$$

We note that $\mathcal{C}_3$ is just $\mathcal{C}_1$ in (7), but with the stricter constraint $n_i = o\left(k_i^2 \cdot \frac{1}{\log p_i}\right)$. We begin with a lemma, which we use to initiate our reduction in Theorem 34.

Lemma 33 *Fix signal priors to be $\mathcal{P}_{\|\beta\|_2}(\{-1,1\})$. For any sequence of parameters $\{(p_i', n_i', k_i', \sigma_i')\}_{i=1}^{\infty}$ in $\mathcal{C}_3$ for S − SLR with solution $\boldsymbol{\beta}'$ and problem instances $(\boldsymbol{X}', \boldsymbol{y}')$, there exists a sequence of parameters $\{(p_i, n_i, k_i, \sigma_i = 0)\}_{i=1}^{\infty}$ in $\mathcal{C}_3$ for SB-MSLR (noiseless) with solution $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and problem instances $(\boldsymbol{X}, \boldsymbol{y})$ such that, for any randomized polynomial time algorithm $\mathcal{A}'$ for S − SLR producing $\hat{\boldsymbol{\beta}}'$ with*

$$\mathbb{P}\left[\|\hat{\boldsymbol{\beta}}' - \boldsymbol{\beta}\|_\infty > 0\right] \to 0,$$

we can construct a second randomized polynomial time algorithm $\mathcal{A}$ for SB-MSLR outputting $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ such that

$$\mathbb{P}\left[\|(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) - (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\|_\infty > 0\right] \to 0.$$

Proof We drop the subscript i notation for convenience. For brevity, let $\mathfrak{P}' :=$ S − SLR, and $\mathfrak{P} :=$ SB-MSLR. Given a sequence of parameters $\{p_i', n_i', k_i', \sigma_i'\}_{i=1}^{\infty}$ in $\mathcal{C}_3$ for $\mathfrak{P}'$, consider the sequence of parameters $\{(p_i, n_i, k_i, \sigma_i)\}_{i=1}^{\infty} = \{(p_i', n_i', k_i', 0)\}_{i=1}^{\infty}$ for $\mathfrak{P}$. Notice that $\{(p_i, n_i, k_i, \sigma_i)\}_{i=1}^{\infty}$ in $\mathcal{C}_3$. Let $(\boldsymbol{X}, \boldsymbol{y})$ be a problem instance of $\mathfrak{P}$ with parameters $\{(p_i, n_i, k_i, \sigma_i)\}_{i=1}^{\infty}$, and recall that for noiseless SB-MSLR the observation is of the form $(\boldsymbol{X}, \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} \odot (2\boldsymbol{z} - 1))$. We apply a preprocessing step to construct $\tilde{\boldsymbol{y}} := g(\boldsymbol{y}) + \boldsymbol{w} \in \mathbb{R}^n$ were $\boldsymbol{w} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, done in randomized polynomial time. Notice that $(\boldsymbol{X}, \tilde{\boldsymbol{y}})$ is now an instance of $\mathfrak{P}'$, by virtue of g being symmetric with respect to sign flips. We then run algorithm $\mathcal{A}'$ on $(\boldsymbol{X}, \tilde{\boldsymbol{y}})$ to obtain $\hat{\boldsymbol{\beta}}'$ with $\mathbb{P}\left[\|\hat{\boldsymbol{\beta}}' - \boldsymbol{\beta}\|_\infty > 0\right] \to 0$ as per the problem statement. Without loss of generality setting $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) = (\hat{\boldsymbol{\beta}}, -\hat{\boldsymbol{\beta}})$, we have an algorithm $\mathcal{A}$ which yields $\mathbb{P}\left[\|(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) - (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\|_\infty > 0\right] \to 0.$ ■

Theorem 34 (Reduction from SB-MSLR − D to exact recovery in SPR) *Consider the setting of* SPR *with joint signal prior $\mathcal{P}_{\|\beta\|_2}(\{-1,1\})$. Any randomized polynomial-time algorithm $\mathcal{A}$ solving* SPR *in parameter regime $\mathcal{C}_2 \cap \mathcal{C}_3$ and $\texttt{SNR} = \omega(1)$ would contradict Theorem 4.*

Proof We first reduce SB-MSLR − D to SB-MSLR within the constraint set $\mathcal{C}_2$ (8), using Lemma 26. We then reduce noiseless SB-MSLR to SPR using Lemma 33 within the constraint set $\mathcal{C}_3$ (10) by choosing $g(x) = |x|$ or $g(x) = x^2$ depending on the precise definition of SPR . Throughout, we have let $\texttt{SNR} = \omega(1)$ to satisfy Lemma 26. Suppose there exists a randomized polynomial-time algorithm $\mathcal{A}$ solving exact recovery in SPR with signals with non-zero entries in $\{-1,0,1\}^p$, i.e. $\mathbb{P}\left[\|\hat{\boldsymbol{\beta}}' - \boldsymbol{\beta}\|_\infty > 0\right] \to 0$. Then by the aforementioned chain of reductions we would have a randomized polynomial time algorithm $\mathcal{A}'$ solving strong detection in SB-MSLR − D with signals with non-zero entries in $\{-1,0,1\}^p$ in the scaling regime $\mathcal{C}_2 \cap \mathcal{C}_3$, which is included in the parameter regime stated in Theorem 4 and would hence contradict Theorem 4. ■

For completeness, we also include a reduction from SB-MSLR − D to a detection variant of SPR in Theorem 36.

Definition 35 (Detection Variant SPR − D) *For $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\sigma > 0$, and $\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)} \in \mathbb{R}^n$, consider the following hypothesis testing problem:*

$$\mathbb{P}(\boldsymbol{X}) \otimes \mathbb{P}(\boldsymbol{y}) : \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} \\ \sqrt{\frac{\|\boldsymbol{\beta}\|_2^2}{\sigma^2}}|\boldsymbol{w}_1| + \boldsymbol{w}_2 \end{bmatrix}$$

$$\mathbb{P}(\boldsymbol{X}, \boldsymbol{y}) : \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} \\ \frac{1}{\sigma}|\boldsymbol{X}\boldsymbol{\beta}| + \boldsymbol{w} \end{bmatrix}$$

where $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \sim \mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$, and $X_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $w_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $z_i \overset{i.i.d.}{\sim}$ Bernoulli(ϕ). The task is to construct a function f which strongly distinguishes $\mathbb{P}(\boldsymbol{X}) \otimes \mathbb{P}(\boldsymbol{y})$ from $\mathbb{P}(\boldsymbol{X}, \boldsymbol{y})$.

Theorem 36 *Fix signal priors to be $\mathcal{P}_{\|\beta\|_2}(\{-1,1\})$. For any sequence of parameters $\{(p_i', n_i', k_i', \sigma_i')\}_{i=1}^\infty$ in $\mathcal{C}_3$ for* SPR − D *with signal $\boldsymbol{\beta}'$ and problem instances $(\boldsymbol{X}', \boldsymbol{y}')$, there exists a sequence of parameters $\{(p_i, n_i, k_i, \sigma_i = 0)\}_{i=1}^\infty \subseteq \mathcal{C}_3$ for* SB-MSLR − D *(noiseless) with signals $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and problem instances $(\boldsymbol{X}, \boldsymbol{y})$ such that, for any randomized polynomial time algorithm $\mathcal{A}'$ solving strong detection in $\mathfrak{P}'$, we can construct a second randomized polynomial time algorithm $\mathcal{A}$ for solving strong detection in $\mathfrak{P}$.*

Proof We drop the subscript i notation for convenience. For brevity, let $\mathfrak{P}' :=$ SPR − D, and $\mathfrak{P} :=$ SB-MSLR − D. Given a sequence of parameters $\{p_i', n_i', k_i', \sigma_i'\}_{i=1}^\infty$ in $\mathcal{C}_3$ for $\mathfrak{P}'$, consider the sequence of parameters $\{(p_i, n_i, k_i, \sigma_i)\}_{i=1}^\infty = \{(p_i', n_i', k_i', 0)\}_{i=1}^\infty$ for $\mathfrak{P}$. Notice that $\{(p_i, n_i, k_i, \sigma_i)\}_{i=1}^\infty$ is in $\mathcal{C}_3$. Let $(\boldsymbol{X}, \boldsymbol{y})$ be a problem instance of $\mathfrak{P}$ with parameters $\{(p_i, n_i, k_i, \sigma_i)\}_{i=1}^\infty$, and recall that in noiseless SB-MSLR − D the observation in the alternative hypothesis is of the form $(\boldsymbol{X}, \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} \odot (2\boldsymbol{z} - 1))$. We apply a preprocessing step to construct $\tilde{\boldsymbol{y}} := |\boldsymbol{y}| + \boldsymbol{w} \in \mathbb{R}^n$ were $\boldsymbol{w} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, done in randomized polynomial time. Notice that under both hypotheses, $(\boldsymbol{X}, \tilde{\boldsymbol{y}})$ is an instance of $\mathfrak{P}'$, by virtue of g being symmetric with respect to sign flips (even). We can then run algorithm $\mathcal{A}'$ on $(\boldsymbol{X}, \tilde{\boldsymbol{y}})$ to solve the hypothesis testing problem of $\mathfrak{P}$ on $(\boldsymbol{X}, \boldsymbol{y})$. ■

Appendix D. Proofs for efficient algorithms

D.1. CORR for support recovery in MSLR

Theorem 37 (CORR achieves joint support recovery in MSLR) *Consider the general setting of* MSLR*,* $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \sim \mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$*. Let* $\epsilon \in (0,1)$ *be the one used in* CORR *. Then provided*

$$n \geq \frac{32(1+\epsilon)}{\min\{\phi^2 \boldsymbol{\beta}_{min}^2, (1-\phi)^2 \boldsymbol{\beta}_{min}^2, \langle \boldsymbol{\beta} \rangle_{min}^2\}} \frac{k(SNR+1)}{SNR} \log 2p,$$

we have that CORR *outputs the exact joint support of signals* $\boldsymbol{\beta}_1$ *and* $\boldsymbol{\beta}_2$ *with probability at least* $1 - c_1(\frac{k}{p} + ke^{-c_2 n} + \frac{k}{n} + \frac{1}{p^{c_2}})$ *for constants* $c_1, c_2 > 0$*.*

Proof Let $\delta > 0$ to be chosen later, and for two sets A and B denote the symmetric difference $A \Delta B := (A \setminus B) \cup (B \setminus A)$. Let $\tau := \sqrt{2(1+\epsilon/2)\log 2p}$ and define the error event

$$\mathcal{E} := \cup_{j \in \mathcal{S}_1 \cup \mathcal{S}_2} \left\{ \left| \frac{\langle \boldsymbol{X}_j, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| < \tau \right\} \cup \left\{ \max_{q \in (\mathcal{S}_1 \cup \mathcal{S}_2)^{\complement}} \left| \frac{\langle \boldsymbol{X}_q, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| \geq \tau \right\}.$$

From here we partition the set of indices $j \in \mathcal{S}_1 \cup \mathcal{S}_2$ into two sets,

$$J_a := \{j \in [p] : j \in \mathcal{S}_1 \cap \mathcal{S}_2\}, \qquad J_b := \{j \in [p] : j \in \mathcal{S}_1 \Delta \mathcal{S}_2\},$$

with respect to which we perform a union bound:

$$\begin{aligned}
\mathbb{P}[\mathcal{E}] &\leq \mathbb{P}\left[\cup_{j \in \mathcal{S}_1 \cup \mathcal{S}_2} \left\{ \left| \frac{\langle \boldsymbol{X}_j, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| < \tau \right\} \right] + \mathbb{P}\left[\max_{q \in (\mathcal{S}_1 \cup \mathcal{S}_2)^{\complement}} \left| \frac{\langle \boldsymbol{X}_q, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| \geq \tau \right] \\
&\leq \mathbb{P}\left[\cup_{j_a \in J_a} \left\{ \left| \frac{\langle \boldsymbol{X}_{j_a}, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| < \tau \right\} \right] + \mathbb{P}\left[\cup_{j_b \in J_b} \left\{ \left| \frac{\langle \boldsymbol{X}_{j_b}, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| < \tau \right\} \right] + \mathbb{P}\left[\max_{q \in (\mathcal{S}_1 \cup \mathcal{S}_2)^{\complement}} \left| \frac{\langle \boldsymbol{X}_q, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| \geq \tau \right] \\
&\leq 2k\mathbb{P}\left[\left| \frac{\langle \boldsymbol{X}_{j_a}, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| < \tau \right] + 2k\mathbb{P}\left[\left\{ \left| \frac{\langle \boldsymbol{X}_{j_b}, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| < \tau \right\} \right] + \mathbb{P}\left[\max_{q \in (\mathcal{S}_1 \cup \mathcal{S}_2)^{\complement}} \left| \frac{\langle \boldsymbol{X}_q, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| \geq \tau \right], \quad (11)
\end{aligned}$$

where $j_a \in J_a$, $j_b \in J_b$, and $q \in (\mathcal{S}_1 \cup \mathcal{S}_2)^{\complement}$. Analyzing the last term, we note that for $q \notin \mathcal{S}_1 \cap \mathcal{S}_2$ we have $\frac{\langle \boldsymbol{X}_q, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. We therefore have,

$$\begin{aligned}
&\mathbb{P}\left[\max_{q \in (\mathcal{S}_1 \cup \mathcal{S}_2)^{\complement}} \left| \frac{\langle \boldsymbol{X}_q, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| \geq \sqrt{2(1+\epsilon/2)\log 2p} \right] \\
&\overset{i)}{\leq} \mathbb{P}\left[\max_{q \in (\mathcal{S}_1 \cup \mathcal{S}_2)^{\complement}} \left| \frac{\langle \boldsymbol{X}_q, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| \geq \sqrt{2\log 2p} + \frac{\epsilon}{2\sqrt{8}} \sqrt{\log 2p} \right] \overset{ii)}{\leq} (2p)^{-\frac{1}{16}\left(\frac{\epsilon}{2}\right)^2},
\end{aligned}$$

where $i)$ follows from $\sqrt{2(1+\epsilon/2)} \geq \sqrt{2} + \frac{\epsilon}{2\sqrt{8}}$, and $ii)$ from a tail bound on the maximum of standard Gaussians (see $\mathbb{P}[\Omega_2^{\complement}(t)]$ in Lemma 45). Applying Lemmas 39 and 38 respectively (and choosing $\delta > 0$ small enough to satisfy these) to the first two terms in (11) we obtain that, for some constants $c_1, c_2 > 0$,

$$\mathbb{P}[\mathcal{E}] \leq c_1 \left(\frac{k}{p} + ke^{-c_2 n} + \frac{k}{n} + \frac{1}{p^{c_2}} \right).$$

■

The principal concentration lemmas.

Lemma 38 (General Bound for $j \in \mathcal{S}_1 \Delta \mathcal{S}_2$) *Consider the setting of* `MSLR`*. Let* $j^* \in \mathcal{S}_1 \Delta \mathcal{S}_2$*. Then if* $n \geq \frac{32(1+\epsilon)}{\min\{\phi^2, (1-\phi)^2\} \boldsymbol{\beta}_{min}^2} (\|\boldsymbol{\beta}\|_2^2 + \sigma^2) \log 2p$ *for any* $\epsilon \in (0,1)$ *we obtain that*

$$\mathbb{P}\left[\left|\frac{\langle \boldsymbol{X}_{j^*}, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \leq \sqrt{2(1+\epsilon/2)\log 2p}\right] \leq \frac{1}{p} + 4e^{-\frac{\delta^2 n}{8}} + \frac{39}{\delta^2 n}.$$

Proof Let $g_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ (independent from $\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2$) for $i \in [n]$ and $\delta > 0$ to be chosen later. Without loss of generality, we can assume $j^* \in \mathcal{S}_1 \setminus \mathcal{S}_2$. We begin by considering fixed $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{y}$ and $\boldsymbol{z}$ vectors, and hence the initial randomness of interest lies in the design matrix $\boldsymbol{X}$. Define

- $\sigma_i^{(1)} := \frac{y_i}{\|\boldsymbol{y}\|_2}\sqrt{1 - \frac{\boldsymbol{\beta}_{1,j^*}^2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}}$,
- $\sigma_i^{(2)} := \frac{y_i}{\|\boldsymbol{y}\|_2}$,
- $\mu_i^{(1)} := \frac{y_i^2}{\|\boldsymbol{y}\|_2} \frac{\boldsymbol{\beta}_{1,j^*}}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}$,
- $\mu_i^{(2)} := 0$,

and notice that for $\tau \in \mathbb{R}$,

$$\mathbb{P}\left[X_{j^*,i} \leq \tau | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] = \mathbb{P}\left[z_i\left(\frac{y_i \cdot \boldsymbol{\beta}_{1,j^*}}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2} + \sqrt{1 - \frac{\boldsymbol{\beta}_{1,j^*}^2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}} g_i\right) + (1 - z_i) g_i \leq \tau \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, y_i, z_i\right],$$

which follows from Lemma 40. We then have that

$$\mathbb{P}\left[X_{j^*,i}\frac{y_i}{\|\boldsymbol{y}\|_2} \leq \tau \middle| \boldsymbol{y}, \boldsymbol{z}\right] = \mathbb{P}\left[z_i(\mu_i^{(1)} + \sigma_i^{(1)} g_i) + (1 - z_i)(\mu_i^{(2)} + \sigma_i^{(2)} g_i) \leq \tau \middle| \boldsymbol{y}, \boldsymbol{z}\right],$$

from which it follows that for $\tau > 0$,

$$\begin{aligned}
&\mathbb{P}\left[\left|\frac{\langle \boldsymbol{X}_{j^*}, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \le \tau \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] \\
&= \mathbb{P}\left[\left|\sum_{i=1}^n z_i(\mu_i^{(1)} + \sigma_i^{(1)} g_i) + (1-z_i)(\mu_i^{(2)} + \sigma_i^{(2)} g_i)\right| \le \tau \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] \\
&= \mathbb{P}\left[\left|\sum_{i=1}^n (z_i\mu^{(1)} + (1-z_i)\mu^{(2)}) + (z_i\sigma^{(1)} + (1-z_i)\sigma^{(2)}) g_i\right| \le \tau \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] \\
&\overset{i)}{=} \mathbb{P}\left[\left|\sum_{i=1}^n (z_i\mu^{(1)} + (1-z_i)\mu^{(2)}) + g_1\left(\sum_{i=1}^n (z_i\sigma_i^{(1)} + (1-z_i)\sigma_i^{(2)})^2\right)^{\frac{1}{2}}\right| \le \tau \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] \\
&= \mathbb{P}\left[\left\{\sum_{i=1}^n (z_i\mu^{(1)} + (1-z_i)\mu^{(2)}) + g_1\left(\sum_{i=1}^n (z_i\sigma_i^{(1)} + (1-z_i)\sigma_i^{(2)})^2\right)^{\frac{1}{2}} \le \tau\right\}\right. \\
&\qquad \left. \cap \left\{-\tau \le \sum_{i=1}^n (z_i\mu^{(1)} + (1-z_i)\mu^{(2)}) + g_1\left(\sum_{i=1}^n (z_i\sigma_i^{(1)} + (1-z_i)\sigma_i^{(2)})^2\right)^{\frac{1}{2}}\right\} \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] \\
&\overset{ii)}{\le} \mathbb{P}\left[g_1\left(\sum_{i=1}^n (z_i\sigma_i^{(1)} + (1-z_i)\sigma_i^{(2)})^2\right)^{\frac{1}{2}} \le \tau - \left|\sum_{i=1}^n (z_i\mu^{(1)} + (1-z_i)\mu^{(2)})\right| \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] \\
&= \mathbb{P}\left[g_1\left(\sum_{i=1}^n (z_i\sigma_i^{(1)} + (1-z_i)\sigma_i^{(2)})^2\right)^{\frac{1}{2}} \le \tau - \left|\sum_{i=1}^n \frac{y_i^2/\|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2} \cdot z_i \boldsymbol{\beta}_{1,j^*}\right| \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] \\
&=: \mathbb{P}\left[A | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right], \qquad (12)
\end{aligned}$$

where $i)$ holds by the closure of Gaussian random variables under finite sum, and $ii)$ holds by symmetry of the Gaussian g_1 and since $\mathbb{P}[B_1 \cap B_2] \le \min_{i \in \{1,2\}} \mathbb{P}[B_i]$.

Using the high probability events $\Omega_1(\delta), \Omega_4(\delta)$ defined in Lemma 45, we have

$$\begin{aligned}
&\mathbb{P}\left[\left|\frac{\langle \boldsymbol{X}_{j^*}, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \le \tau\right] = \int_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \int_{y,z} \mathbb{P}\left[\left|\frac{\langle X_{j^*,i}, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \le \tau \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] d\mathbb{P}[\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, z] \\
&\le \int_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \int_{\Omega_1(\delta) \cap \Omega_4(\delta)} \mathbb{P}\left[\left|\frac{\langle X_{j^*,i}, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \le \tau \middle| \boldsymbol{y}, \boldsymbol{z}\right] d\mathbb{P}[\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, z] \\
&\quad + \int_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \mathbb{P}\left[(\Omega_1(\delta) \cap \Omega_4(\delta))^{\complement} \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right] d\mathbb{P}[\boldsymbol{\beta}_1, \boldsymbol{\beta}_2] \\
&= \int_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \int_{\Omega_1(\delta) \cap \Omega_4(\delta)} \mathbb{P}\left[A | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] d\mathbb{P}[\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, z] + \mathbb{P}\left[(\Omega_1(\delta) \cap \Omega_4(\delta))^{\complement}\right] \\
&\le \int_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \int_{\Omega_1(\delta) \cap \Omega_4(\delta)} \mathbb{P}\left[A | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] d\mathbb{P}[\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, z] + \mathbb{P}\left[\Omega_1^{\complement}(\delta)\right] + \mathbb{P}\left[\Omega_4^{\complement}(\delta)\right]. \qquad (13)
\end{aligned}$$

Now setting $\tau = \sqrt{2(1+\epsilon/2)\log 2p}$ and $n \geq \frac{32(1+\epsilon)}{\phi^2 \boldsymbol{\beta}^2_{\texttt{min}}}(\|\boldsymbol{\beta}\|_2^2 + \sigma^2)\log p$ we obtain that for $(y, z) \in \Omega_1(\delta) \cap \Omega_4(\delta)$:

$$
\begin{aligned}
&\mathbb{P}\left[A | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] \\
&= \mathbb{P}\left[g_1 \left(\sum_{i=1}^n (z_i \sigma_i^{(1)} + (1-z_i)\sigma_i^{(2)})^2 \right)^{\frac{1}{2}} \leq \tau - \left| \sum_{i=1}^n \frac{y_i^2 / \|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2} \cdot z_i \boldsymbol{\beta}_{1,j^*} \right| \Bigg| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z} \right] \\
&\leq \mathbb{P}\Bigg[g_1 \left(\sum_{i=1}^n (z_i \sigma_i^{(1)} + (1-z_i)\sigma_i^{(2)})^2 \right)^{\frac{1}{2}} \leq \sqrt{2(1+\epsilon/2)\log 2p} \\
&\qquad\qquad - \left| \frac{\|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2} \phi \boldsymbol{\beta}_{1,j^*} - \sqrt{\frac{2(3+\delta)\log 2p}{(1-\delta)(1+\sigma^2/\|\boldsymbol{\beta}\|_2^2)}} \right| \Bigg| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z} \Bigg] \\
&\leq \mathbb{P}\Bigg[g_1 \left(\sum_{i=1}^n (z_i \sigma_i^{(1)} + (1-z_i)\sigma_i^{(2)})^2 \right)^{\frac{1}{2}} \leq \sqrt{2(1+\epsilon/2)\log 2p} \\
&\qquad\qquad - \left| \sqrt{\frac{n(1-\delta)}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}} \phi \boldsymbol{\beta}_{1,j^*} - \sqrt{\frac{2(3+\delta)\log 2p}{(1-\delta)(1+\sigma^2/\|\boldsymbol{\beta}\|_2^2)}} \right| \Bigg| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z} \Bigg] \\
&\leq \mathbb{P}\Bigg[g_1 \left(\sum_{i=1}^n (z_i \sigma_i^{(1)} + (1-z_i)\sigma_i^{(2)})^2 \right)^{\frac{1}{2}} \leq \sqrt{2(1+\epsilon/2)\log 2p} \\
&\qquad\qquad - \left| \sqrt{32(1+\epsilon)(1-\delta)\log 2p} - \sqrt{\frac{2(3+\delta)\log 2p}{(1-\delta)(1+\sigma^2/\|\boldsymbol{\beta}\|_2^2)}} \right| \Bigg| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z} \Bigg],
\end{aligned} \tag{14}
$$

where for small enough $\delta > 0$, we have

$$
\sqrt{2(1+\epsilon/2)} - \left| \sqrt{32(1+\epsilon)(1-\delta)} - \sqrt{2(3+\delta)/(1-\delta)(1+\sigma^2/\|\boldsymbol{\beta}\|_2^2)} \right| \leq -\sqrt{2}. \tag{15}
$$

Hence we obtain that the right hand side of the inequality in (14) is negative, leading us to the following inequality for $(y, z) \in \Omega_1(\delta) \cap \Omega_4(\delta)$,

$$
\begin{aligned}
\mathbb{P}\left[A | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] &\leq \mathbb{P}\left[g_1 \left(\sum_{i=1}^n (z_i \sigma_i^{(1)} + (1-z_i)\sigma_i^{(2)})^2 \right)^{1/2} \leq -\sqrt{2\log 2p} \Bigg| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z} \right] \\
&\leq \exp\left(-\frac{2\log 2p}{2\sum_{i=1}^n (z_i \sigma_i^{(1)} + (1-z_i)\sigma_i^{(2)})^2} \right) \\
&\overset{i)}{\leq} \exp\left(-\frac{2\log 2p}{2\sum_{i=1}^n \frac{y_i^2}{\|\boldsymbol{y}\|_2^2}} \right) \leq \frac{1}{2p},
\end{aligned} \tag{16}
$$

where $i)$ follows almost surely from the fact that $z_i \in \{0, 1\}$ and $\left(1 - \frac{\boldsymbol{\beta}_{1,j^*}^2}{\|\boldsymbol{\beta}_1\|_2^2 + \sigma^2}\right) \leq 1$. We generalize to the case where $j \in \mathcal{S}_2 \setminus \mathcal{S}_1$, and hence consider the analogous result with ϕ and $\boldsymbol{\beta}_1$

replaced with $(1-\phi)$ and $\boldsymbol{\beta}_2$. Putting it all together in (13) using Lemma 45 and choosing $\delta > 0$ small enough to satisfy (15), we obtain that for $n \geq \frac{32}{\min\{\phi^2,(1-\phi)^2\}\boldsymbol{\beta}_{\texttt{min}}^2}(1+\epsilon)(\|\boldsymbol{\beta}\|_2^2+\sigma^2)\log 2p$,

$$\mathbb{P}\left[\left|\frac{\langle \boldsymbol{X}_{j^*}, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \leq \sqrt{2(1+\epsilon/2)\log 2p}\right] \leq \frac{1}{p} + 4e^{-\frac{\delta^2 n}{8}} + \frac{39}{\delta^2 n}. \tag{17}$$

■

Lemma 39 (General Bound for $j \in \mathcal{S}_1 \cap \mathcal{S}_2$) *Consider the setting of* MSLR. *Let $j^* \in \mathcal{S}_1 \cap \mathcal{S}_2$. Then if $n \geq \frac{32(1+\epsilon)}{(\phi\beta_{1,j^*}+(1-\phi)\beta_{2,j^*})^2}(\|\boldsymbol{\beta}\|_2^2+\sigma^2)\log 2p$ for any $\epsilon \in (0,1)$ we obtain that*

$$\mathbb{P}\left[\left|\frac{\langle \boldsymbol{X}_{j^*}, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \leq \sqrt{2(1+\epsilon/2)\log 2p}\right] \leq \frac{1}{p} + 4e^{-\frac{\delta^2 n}{8}} + \frac{39}{\delta^2 n}.$$

Proof The proof is along the same lines as that of Lemma 38, with the main difference being in the conditional means and variances of $X_{j^*,i}$, for $i \in [n]$. As before, let $g_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ (independent from $y, z, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2$) for $i \in [n]$ and $\delta > 0$ to be chosen later. Define

- $\sigma_i^{(1)} := \frac{y_i}{\|\boldsymbol{y}\|_2}\sqrt{1-\frac{\boldsymbol{\beta}_{1,j^*}^2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}}$,
- $\sigma_i^{(2)} := \frac{y_i}{\|\boldsymbol{y}\|_2}\sqrt{1-\frac{\boldsymbol{\beta}_{2,j^*}^2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}}$,
- $\mu_i^{(1)} := \frac{y_i^2}{\|\boldsymbol{y}\|_2}\frac{\boldsymbol{\beta}_{1,j^*}}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}$,
- $\mu_i^{(2)} := \frac{y_i^2}{\|\boldsymbol{y}\|_2}\frac{\boldsymbol{\beta}_{2,j^*}}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}$,

and notice that for $\tau \in \mathbb{R}$,

$$\begin{aligned}
&\mathbb{P}\left[X_{j^*,i} \leq \tau | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] \\
&= \mathbb{P}\left[z_i\left(\frac{y_i\cdot\boldsymbol{\beta}_{1,j^*}}{\|\boldsymbol{\beta}\|_2^2+\sigma^2} + \sqrt{1-\frac{\boldsymbol{\beta}_{1,j^*}^2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}}g_i\right)\right. \\
&\qquad \left. +(1-z_i)\left(\frac{y_i\cdot\boldsymbol{\beta}_{2,j^*}}{\|\boldsymbol{\beta}\|_2^2+\sigma^2} + \sqrt{1-\frac{\boldsymbol{\beta}_{2,j^*}^2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}}g_i\right) \leq \tau \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, y_i, z_i\right],
\end{aligned}$$

which follows from Lemma 40. We then have that

$$\mathbb{P}\left[X_{j^*,i}\frac{y_i}{\|\boldsymbol{y}\|_2} \leq \tau \middle| \boldsymbol{y}, \boldsymbol{z}\right] = \mathbb{P}\left[z_i(\mu_i^{(1)}+\sigma_i^{(1)}g_i) + (1-z_i)(\mu_i^{(2)}+\sigma_i^{(2)}g_i) \leq \tau \middle| \boldsymbol{y}, \boldsymbol{z}\right].$$

Then, using the same steps as in (12), we obtain that for for $\tau > 0$:

$$\begin{aligned}
&\mathbb{P}\left[\left|\frac{\langle \boldsymbol{X}_{j^*}, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \leq \tau \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] \\
&= \mathbb{P}\left[g_1\left(\sum_{i=1}^n (z_i\sigma_i^{(1)} + (1-z_i)\sigma_i^{(2)})^2\right)^{\frac{1}{2}} \leq \tau - \left|\sum_{i=1}^n \frac{y_i^2/\|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\cdot(z_i\boldsymbol{\beta}_{1,j^*} + (1-z_i)\boldsymbol{\beta}_{2,j^*})\right| \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] \\
&=: \mathbb{P}\left[A | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right].
\end{aligned} \tag{18}$$

Using the high probability events $\Omega_1(\delta), \Omega_3(\delta)$ defined in Lemma 45, by the same arguments as in (13) we have

$$\mathbb{P}\left[\left|\frac{\langle \boldsymbol{X}_{j^*}, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \le \tau\right] \le \int_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} \int_{\Omega_1(\delta)\cap\Omega_3(\delta)} \mathbb{P}\left[A|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] d\mathbb{P}[\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, z] + \mathbb{P}\left[\Omega_1^{\complement}(\delta)\right] + \mathbb{P}\left[\Omega_3^{\complement}(\delta)\right], \tag{19}$$

Now setting $\tau = \sqrt{2(1+\epsilon/2)\log 2p}$ and $n \ge \frac{32}{\langle \boldsymbol{\beta}\rangle^2_{\min}}(1+\epsilon)(\|\boldsymbol{\beta}\|_2^2 + \sigma^2)\log 2p$ we obtain that for $(y, z) \in \Omega_1(\delta) \cap \Omega_3(\delta)$,

$$\begin{aligned}
&\mathbb{P}\left[A|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] \\
&= \mathbb{P}\left[g_1\left(\sum_{i=1}^n (z_i\sigma_i^{(1)} + (1-z_i)\sigma_i^{(2)})^2\right)^{\frac{1}{2}} \le \tau - \left|\sum_{i=1}^n \frac{y_i^2/\|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\cdot(z_i\boldsymbol{\beta}_{1,j^*} + (1-z_i)\boldsymbol{\beta}_{2,j^*})\right| \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] \\
&\le \mathbb{P}\left[g_1\left(\sum_{i=1}^n (z_i\sigma_i^{(1)} + (1-z_i)\sigma_i^{(2)})^2\right)^{\frac{1}{2}} \le \sqrt{2(1+\epsilon/2)\log 2p}\right. \\
&\qquad\qquad \left. - \left|\frac{\|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}(\phi\boldsymbol{\beta}_{1,j^*} + (1-\phi)\boldsymbol{\beta}_{2,j^*}) - \sqrt{\frac{2(3+\delta)\log 2p}{(1-\delta)(1+\sigma^2/\|\boldsymbol{\beta}\|_2^2)}}\right| \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] \\
&\le \mathbb{P}\left[g_1\left(\sum_{i=1}^n (z_i\sigma_i^{(1)} + (1-z_i)\sigma_i^{(2)})^2\right)^{\frac{1}{2}} \le \sqrt{2(1+\epsilon/2)\log 2p}\right. \\
&\qquad\qquad \left. - \left|\sqrt{\frac{n(1-\delta)}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}}(\phi\boldsymbol{\beta}_{1,j^*} + (1-\phi)\boldsymbol{\beta}_{2,j^*}) - \sqrt{\frac{2(3+\delta)\log 2p}{(1-\delta)(1+\sigma^2/\|\boldsymbol{\beta}\|_2^2)}}\right| \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] \\
&\le \mathbb{P}\left[g_1\left(\sum_{i=1}^n (z_i\sigma_i^{(1)} + (1-z_i)\sigma_i^{(2)})^2\right)^{\frac{1}{2}} \le \sqrt{2(1+\epsilon/2)\log 2p}\right. \\
&\qquad\qquad \left. - \left|\sqrt{32(1+\epsilon)(1-\delta)\log 2p} - \sqrt{\frac{2(3+\delta)\log 2p}{(1-\delta)(1+\sigma^2/\|\boldsymbol{\beta}\|_2^2)}}\right| \middle| \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right],
\end{aligned} \tag{20}$$

where

$$\sqrt{2(1+\epsilon/2)} - \left|\sqrt{32(1+\epsilon)(1-\delta)} - \sqrt{2(3+\delta)/(1-\delta)(1+\sigma^2/\|\boldsymbol{\beta}\|_2^2)}\right| \le -\sqrt{2}, \tag{21}$$

for small enough $\delta > 0$. Hence we obtain that the right hand side of (20) is negative, leading us to the following inequality for $(y, z) \in \Omega_1(\delta) \cap \Omega_3(\delta)$, obtained via the same steps as (16):

$$\mathbb{P}\left[A|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{y}, \boldsymbol{z}\right] \le \frac{1}{2p}.$$

Putting it all together in (19) using Lemma 45 and choosing $\delta > 0$ small enough to satisfy (20) we obtain that for $n \geq \frac{32}{\langle \boldsymbol{\beta} \rangle_{\mathtt{min}}^2}(1+\epsilon)(\|\boldsymbol{\beta}\|_2^2 + \sigma^2)\log 2p$,

$$\mathbb{P}\left[\left|\frac{\langle \boldsymbol{X}_{j^*}, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \leq \sqrt{2(1+\epsilon/2)\log 2p}\right] \leq \frac{1}{p} + 4e^{-\frac{\delta^2 n}{8}} + \frac{39}{\delta^2 n}.$$

■

The conditioning lemmas.

Lemma 40 (General conditioning lemma) *Consider the setting of* MSLR.

For $j^* \in \mathcal{S}_1 \cap \mathcal{S}_2$ and $i \in [n]$ it holds that:

$$X_{j^*,i}|(y_i, \boldsymbol{\beta}_1, z_i = 1) \sim \mathcal{N}\left(\frac{y_i \cdot \boldsymbol{\beta}_{1,j^*}}{\|\boldsymbol{\beta}_1\|_2^2 + \sigma^2}, \left(1 - \frac{\boldsymbol{\beta}_{1,j^*}^2}{\|\boldsymbol{\beta}_1\|_2^2 + \sigma^2}\right)\right), \tag{22}$$

$$X_{j^*,i}|(y_i, \boldsymbol{\beta}_2, z_i = 0) \sim \mathcal{N}\left(\frac{y_i \cdot \boldsymbol{\beta}_{2,j^*}}{\|\boldsymbol{\beta}_1\|_2^2 + \sigma^2}, \left(1 - \frac{\boldsymbol{\beta}_{2,j^*}^2}{\|\boldsymbol{\beta}_1\|_2^2 + \sigma^2}\right)\right). \tag{23}$$

For $j^* \in \mathcal{S}_1 \Delta \mathcal{S}_2$ (without loss of generality $j^* \in \mathcal{S}_1 \setminus \mathcal{S}_2$) and $i \in [n]$ it holds that

$$X_{j^*,i}|(y_i, \boldsymbol{\beta}_1, z_i = 1) \sim \mathcal{N}\left(\frac{y_i \cdot \boldsymbol{\beta}_{1,j^*}}{\|\boldsymbol{\beta}_1\|_2^2 + \sigma^2}, \left(1 - \frac{\boldsymbol{\beta}_{1,j^*}^2}{\|\boldsymbol{\beta}_1\|_2^2 + \sigma^2}\right)\right), \tag{24}$$

$$X_{j^*,i}|(y_i, z_i = 0) \sim \mathcal{N}(0, 1) \tag{25}$$

Proof To prove (22) and (24), recall that

$$y_i = z_i(\boldsymbol{X}\boldsymbol{\beta}_1)_i + (1 - z_i)(\boldsymbol{X}\boldsymbol{\beta}_2)_i + w_i,$$

and hence given $z_i = 1$ we have

$$y_i = \sum_{j \neq j^* \in \mathcal{S}_1} \boldsymbol{\beta}_{1,j} X_{j,i} + \boldsymbol{\beta}_{1,j^*} X_{j^*,i} + w_i,$$

implying that

$$X_{j^*,i} \,|\, (y_i, \boldsymbol{\beta}_1, z_i = 1) \;=\; X_{j^*,i} \left| \left(\boldsymbol{\beta}_1, y_i = \sum_{j \neq j^* \in \mathcal{S}_1} \boldsymbol{\beta}_{1,j} X_{j,i} + \boldsymbol{\beta}_{1,j^*} X_{j^*,i} + w_i\right)\right. .$$

Therefore, applying Corollary 43, we obtain that $X_{j^*,i}|(y_i, \boldsymbol{\beta}_1, z_i = 1) \sim \mathcal{N}\left(\frac{y_i \boldsymbol{\beta}_{1,j^*}}{\|\boldsymbol{\beta}_1\|_2^2 + \sigma^2}, \left(1 - \frac{\boldsymbol{\beta}_{1,j^*}^2}{\|\boldsymbol{\beta}_1\|_2^2 + \sigma^2}\right)\right)$. The result in (23) is proved in the same way, but replacing $\boldsymbol{\beta}_1$ with $\boldsymbol{\beta}_2$.

For (25), notice that given $z_i = 0$ we have

$$y_i = \sum_{j \in \mathcal{S}_2} \boldsymbol{\beta}_{2,j} X_{j,i} + w_i,$$

which is independent of $X_{j^*,i}$ by definition, and hence $X_{j^*,i}|(y_i, z_i = 0)$ has the same distribution as $X_{j^*,i} \sim \mathcal{N}(0, 1)$. ■

Lemma 41 (General conditioning lemma) *Let* $a \sim \mathcal{N}\left(0, \Sigma_{k\times k}\right)$*, and* $b \in \mathbb{R}^k$ *a fixed vector. Then:*

$$a \Big| \left(\sum_{j=1}^{k} b_j a_j = \eta\right) \sim \mathcal{N}\left(\eta v, B\Sigma_{k\times k}B^T\right)$$

where, letting $\mathbb{1}$ *denote the all-ones vector,*

$$v = \frac{1}{b^T\tilde{\Sigma}\mathbb{1}}\tilde{\Sigma}\mathbb{1}, \quad B = I_{k\times k} - vb^T, \quad \tilde{\Sigma} = \mathbb{E}\left[a(a\odot b)^T\right].$$

Proof Let B be a deterministic matrix, and $\eta := \sum_{j=1}^{k} b_j a_j$. Then (Ba, η) is jointly normal. We will construct a fixed matrix B and fixed vector v such that

- Ba is independent from η
- $a = Ba + \eta v$.

If the above holds, then by independence we have the required result that $a|\eta \sim \mathcal{N}\left(\eta v, B\Sigma B^T\right)$. In order for the first point to hold, their covariances must be zero, implying

$$\mathbb{E}\left[Ba\eta\right] = \mathbb{E}\left[Ba(a\odot b)^T\mathbb{1}\right] = B\tilde{\Sigma}\mathbb{1} = 0.$$

Meanwhile, the second point is satisfied by choosing $B = I - vb^T$. Combining these two facts, we obtain the result. ■

Corollary 42 *Consider the setting of* SLR*, let* $j^* \in \mathcal{S}$*, and* $i \in [n]$*. Then it holds that*

$$X_{j^*,i} \Bigg| \left(\sum_{j\in\mathcal{S}} X_{j,i} + w_i = \eta\right) \sim \mathcal{N}\left(\frac{\eta}{k+\sigma^2}, \left(1 - \frac{1}{k+\sigma^2}\right)\right)$$

Proof Apply Lemma 41 conditioning on the sum $\sum_{j\in\mathcal{S}} X_{j,i} + w_i =: \eta$, where we recall $\left[\boldsymbol{X}_{\mathcal{S}}\ w_i\right] \sim \mathcal{N}\left(0, \Sigma\right)$ with

$$\Sigma = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \in \mathbb{R}^{(k+1)\times(k+1)}$$

yielding

$$v = \frac{1}{k+\sigma^2}\begin{bmatrix} 1 \\ 1 \\ \vdots \\ \sigma^2 \end{bmatrix}$$

and

$$B = \begin{bmatrix} (1 - \frac{1}{k+\sigma^2}) & -\frac{1}{k+\sigma^2} & \cdots & -\frac{1}{k+\sigma^2} \\ -\frac{1}{k+\sigma^2} & (1 - \frac{1}{k+\sigma^2}) & \cdots & -\frac{1}{k+\sigma^2} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{k+\sigma^2} & -\frac{1}{k+\sigma^2} & \cdots & (1 - \frac{1}{k+\sigma^2}) \end{bmatrix}.$$

Noticing that $(B\Sigma B^T)_{j^*,j^*} = 1 - \frac{1}{k+\sigma^2}$, we obtain the result. ■

Corollary 43 *Consider the setting of* SLR*, and let $b \in \mathbb{R}^p$ be fixed. Let $j^* \in \mathcal{S}$, and $i \in [n]$. Then it holds that*

$$X_{j^*,i} \Bigg| \left(\sum_{j \in \mathcal{S}} b_j X_{j,i} + w_i = \eta \right) \sim \mathcal{N}\left(\frac{\eta \cdot b_{j^*}}{\|b\|_2^2 + \sigma^2}, \left(1 - \frac{b_{j^*}^2}{\|b\|_2^2 + \sigma^2} \right) \right)$$

Proof The result is obtained by applying Lemma 41 conditioning on the sum $\sum_{j \in \mathcal{S}} b_j X_{j,i} + w_i = \eta$. We then have that

$$\tilde{\Sigma} = \mathbb{E}\left[\begin{bmatrix} X_{1,i} \\ X_{2,i} \\ \vdots \\ w_i \end{bmatrix} \begin{bmatrix} b_1 X_{1,i} \\ b_2 X_{2,i} \\ \vdots \\ w_i \end{bmatrix}^T \right] = \begin{bmatrix} b_1 & 0 & \cdots & 0 \\ 0 & b_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \in \mathbb{R}^{(k+1)\times(k+1)}$$

yielding

$$v = \frac{1}{\|b\|_2^2 + \sigma^2} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ \sigma^2 \end{bmatrix}$$

and

$$B = \begin{bmatrix} (1 - \frac{b_1^2}{\|b\|_2^2+\sigma^2}) & -\frac{b_1 b_2}{\|b\|_2^2+\sigma^2} & \cdots & -\frac{b_1}{\|b\|_2^2+\sigma^2} \\ -\frac{b_2 b_1}{\|b\|_2^2+\sigma^2} & (1 - \frac{b_2^2}{\|b\|_2^2+\sigma^2}) & \cdots & -\frac{b_2}{\|b\|_2^2+\sigma^2} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\sigma^2 b_1}{\|b\|_2^2+\sigma^2} & -\frac{\sigma^2 b_2}{\|b\|_2^2+\sigma^2} & \cdots & (1 - \frac{\sigma^2}{\|b\|_2^2+\sigma^2}) \end{bmatrix}.$$

Noticing that $(B\Sigma B^T)_{j^*,j^*} = 1 - \frac{b_{j^*}^2}{\|b\|_2^2+\sigma^2}$, we obtain the result. ■

Lemma 44 *Consider the setting of* MSLR *as in Definition 1. Then y and z are independent.*

Proof Let $g \sim \mathcal{N}(0,1)$, and A an event in the sigma algebra. By Bayes rule, for every $i \in [n]$ it holds that

$$\begin{aligned} \mathbb{P}[z_i = 1 | y_i \in A] &= \frac{\mathbb{P}[y_i \in A | z_i = 1]\, \mathbb{P}[z_i = 1]}{\mathbb{P}[y]} \\ &= \frac{\mathbb{P}[y_i \in A | z_i = 1]\, \mathbb{P}[z_i = 1]}{\mathbb{P}[y_i \in A | z_i = 1]\, \mathbb{P}[z_i = 1] + \mathbb{P}[y_i \in A | z_i = 0]\, \mathbb{P}[z_i = 0]} \\ &= \frac{\mathbb{P}[\sqrt{\|\boldsymbol{\beta}\|_2^2 + \sigma^2} g \in A]\phi}{\mathbb{P}[\sqrt{\|\boldsymbol{\beta}\|_2^2 + \sigma^2} g \in A]\phi + \mathbb{P}[\sqrt{\|\boldsymbol{\beta}\|_2^2 + \sigma^2} g \in A](1-\phi)} \\ &= \phi \\ &= \mathbb{P}[z_i = 1], \end{aligned}$$

and the analogous result holds for the case $z_i = 0$. The result then follows by recalling that y_i and z_i are i.i.d. across the i indices. ■

High-probability events and Concentration inequalities

Lemma 45 (High-probability events) *Consider the setting of* MSLR. *Let* $g_q \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ *for* $q \in [p]$ *and* $\delta \in (0,1)$, $t > 0, j^* \in [p]$. *Define the following events that will be necessary for the analysis of* CORR *on* MSLR*:*

$$\Omega_1(\delta) := \left\{(\|\boldsymbol{\beta}\|_2^2 + \sigma^2)n(1-\delta) \le \|\boldsymbol{y}\|^2 \le (\|\boldsymbol{\beta}\|_2^2 + \sigma^2)n(1+\delta)\right\},$$

$$\Omega_2(t) := \left\{\max_{q\in[p]} |g_q| \le \sqrt{2\log 2p} + \sqrt{2t\log 2p}\right\},$$

$$\Omega_3(\delta) := \left\{\left|\sum_{i=1}^n (z_i\boldsymbol{\beta}_{1,j^*} + (1-z_i)\boldsymbol{\beta}_{2,j^*})\frac{y_i^2/\|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2} - (\phi\boldsymbol{\beta}_{1,j^*} + (1-\phi)\boldsymbol{\beta}_{2,j^*})\frac{\|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right| < \sqrt{\frac{2(3+\delta)\log(2p)}{(1-\delta)(1+\sigma^2/\|\boldsymbol{\beta}\|_2^2)}}\right\},$$

$$\Omega_4(\delta) := \left\{\left|\sum_{i=1}^n z_i\boldsymbol{\beta}_{1,j^*}\frac{y_i^2/\|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2} - \phi\boldsymbol{\beta}_{1,j^*}\frac{\|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right| < \sqrt{\frac{2(3+\delta)\log(2p)}{(1-\delta)(1+\sigma^2/\|\boldsymbol{\beta}\|_2^2)}}\right\},$$

$$\Omega_5(\delta) := \left\{\left|\|\boldsymbol{y}\|_4^4 - 3n(\|\boldsymbol{\beta}\|_2^2+\sigma^2)^2\right| < \delta n(\|\boldsymbol{\beta}\|_2^2+\sigma^2)^2\right\}.$$

Then, the above events all occur with high probability. Specifically,

- $\mathbb{P}\left[\Omega_1^{\complement}(\delta)\right] \le 2e^{-\frac{\delta^2 n}{8}}$ *as per Example 2.11 in Wainwright (2019) noting that* $y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \|\boldsymbol{\beta}\|_2^2 + \sigma^2)$ *both marginally and conditionally on z (similarly for the setting of* SLR*).*
- $\mathbb{P}\left[\Omega_2^{\complement}(t)\right] \le \frac{1}{(2p)^t}$ *as per Lemma 5.2 in van Handel (2014).*
- $\mathbb{P}\left[\Omega_3^{\complement}(\delta)\middle|\Omega_1(\delta),\Omega_5(\delta)\right] \le \frac{1}{2p}$ *as per Lemma 47 and consequently* $\mathbb{P}\left[\Omega_3^{\complement}(\delta)\right] \le \frac{1}{2p} + 2e^{-\frac{\delta^2 n}{8}} + \frac{39}{\delta^2 n}$.
- $\mathbb{P}\left[\Omega_4^{\complement}(\delta)\middle|\Omega_1(\delta),\Omega_5(\delta)\right] \le \frac{1}{2p}$ *as per Lemma 47 with* $\boldsymbol{\beta}_{2,j^*} = 0$ *and consequently* $\mathbb{P}\left[\Omega_4^{\complement}(\delta)\right] \le \frac{1}{2p} + 2e^{-\frac{\delta^2 n}{8}} + \frac{39}{\delta^2 n}$.
- $\mathbb{P}\left[\Omega_5^{\complement}(\delta)\right] \le \frac{39}{\delta^2 n}$ *as per Lemma 46.*

Lemma 46 (Fast ℓ_4 norm concentration) *Let* $\boldsymbol{y} \in \mathbb{R}^n$ *with* $y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2)$ *for* $i \in [n]$. *Then for* $\delta > 0$ *we have,*

$$\mathbb{P}\left[\left|\|\boldsymbol{y}\|_4 - 3n(\sigma^2)^2\right| \ge \delta n(\sigma^2)^2\right] \le \frac{39}{\delta^2 n}.$$

Proof The proof follows from a standard application of Chebyshev's inequality. We first recall that the p^{th} centered Gaussian moment is given by $(\text{variance})^{\frac{p}{2}}(p-1)!!$, and consequently by independence we have that $\mathbb{E}\left[\|\boldsymbol{y}\|_4^4\right] = 3n(\sigma^2)^2$. We then proceed with Chebyshev's inequality (see Boucheron et al. (2013)) for $t > 0$:

$$\begin{aligned}\mathbb{P}\left[\left|\|\boldsymbol{y}\|_4^4 - \mathbb{E}\left[\|\boldsymbol{y}\|_4^4\right]\right| \geq t\right] &\leq \frac{n\text{Var}(y_1^4)}{t^2} \\ &= \frac{n\left(\mathbb{E}y_1^8 - \left(\mathbb{E}y_1^4\right)^2\right)}{t^2} \\ &= \frac{n(\sigma^2)^4((7-1)!! - 9)}{t^2} = \frac{39n(\sigma^2)^4}{t^2}.\end{aligned}$$

We set $t = n(\sigma^2)^2$ to obtain

$$\mathbb{P}\left[\left|\|\boldsymbol{y}\|_4^4 - \mathbb{E}\left[\|\boldsymbol{y}\|_4^4\right]\right| \geq n(\sigma^2)^2\right] \leq \frac{39n(k+\sigma^2)^4}{n^2(\sigma^2)^4} = \frac{39}{n}.$$

■

Lemma 47 *Consider the setting of* MSLR*. For $t > 0, \delta > 0, j^* \in [p]$ we have that*

$$\begin{aligned}&\mathbb{P}\left[\left|\sum_{i=1}^{n}(z_i\boldsymbol{\beta}_{1,j^*} + (1-z_i)\boldsymbol{\beta}_{2,j^*})\frac{y_i^2/\|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2} - (\phi\boldsymbol{\beta}_{1,j^*} + (1-\phi)\boldsymbol{\beta}_{2,j^*})\frac{\|\boldsymbol{y}\|_2^2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right| \geq t \middle| \Omega_1(\delta), \Omega_5(\delta)\right] \\ &\leq \exp\left(-\frac{(1-\delta)}{2(3+\delta)}(1+\sigma^2/\|\boldsymbol{\beta}\|_2^2)t^2\right)\end{aligned}$$

Proof Consider $y \in \Omega_1(\delta) \cap \Omega_5(\delta)$ in Lemma 45. We first apply Hoeffding's inequality (see Boucheron et al. (2013)), then the definitions of $\Omega_1(\delta)$ and $\Omega_5(\delta)$ to obtain,

$$\mathbb{P}\left[\left|\sum_{i=1}^{n}(z_i\boldsymbol{\beta}_{1,j^*}+(1-z_i)\boldsymbol{\beta}_{2,j^*})\frac{y_i^2/\|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}-(\phi\boldsymbol{\beta}_{1,j^*}+(1-\phi)\boldsymbol{\beta}_{2,j^*})\frac{\|\boldsymbol{y}\|_2^2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right|\geq t\middle|\boldsymbol{y}\right]$$
$$\leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(\boldsymbol{\beta}_{1,j^*}-\boldsymbol{\beta}_{2,j^*})^2\frac{y_i^4/\|\boldsymbol{y}\|_2^2}{(\|\boldsymbol{\beta}\|_2^2+\sigma^2)^2}}\right)$$
$$= \exp\left(-\frac{2t^2}{(\boldsymbol{\beta}_{1,j^*}-\boldsymbol{\beta}_{2,j^*})^2\frac{\|\boldsymbol{y}\|_4^4}{\|\boldsymbol{y}\|_2^2}\frac{1}{(\|\boldsymbol{\beta}\|_2^2+\sigma^2)^2}}\right)$$
$$\leq \exp\left(-\frac{2t^2}{(\boldsymbol{\beta}_{1,j^*}-\boldsymbol{\beta}_{2,j^*})^2\frac{3n(\|\boldsymbol{\beta}\|_2^2+\sigma^2)^2+\delta n(\|\boldsymbol{\beta}\|_2^2+\sigma^2)^2}{\|\boldsymbol{y}\|_2^2}\frac{1}{(\|\boldsymbol{\beta}\|_2^2+\sigma^2)^2}}\right)$$
$$\leq \exp\left(-\frac{2t^2}{(\boldsymbol{\beta}_{1,j^*}-\boldsymbol{\beta}_{2,j^*})^2\frac{3n(\|\boldsymbol{\beta}\|_2^2+\sigma^2)^2+\delta n(\|\boldsymbol{\beta}\|_2^2+\sigma^2)^2}{n(\|\boldsymbol{\beta}\|_2^2+\sigma^2)(1-\delta)}\frac{1}{(\|\boldsymbol{\beta}\|_2^2+\sigma^2)^2}}\right)$$
$$= \exp\left(-\frac{2(\|\boldsymbol{\beta}\|_2^2+\sigma^2)t^2}{(\boldsymbol{\beta}_{1,j^*}-\boldsymbol{\beta}_{2,j^*})^2\frac{3+\delta}{1-\delta}}\right)$$
$$\stackrel{i)}{\leq} \exp\left(-\frac{2(\|\boldsymbol{\beta}\|_2^2+\sigma^2)t^2}{4\|\boldsymbol{\beta}\|_2^2\frac{3+\delta}{1-\delta}}\right)$$
$$= \exp\left(-\frac{(1-\delta)}{2(3+\delta)}(1+\sigma^2/\|\boldsymbol{\beta}\|_2^2)t^2\right),$$

where $i)$ follows from $(\boldsymbol{\beta}_{1,j^*}-\boldsymbol{\beta}_{2,j^*})^2 \leq \|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_2^2$ and the triangle inequality. Applying the law of total probability to the above, we obtain the result. ■

D.2. CORR for MSLR − D

Proof [Proof of Theorem 7] We first note that, since $(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)\sim\mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$, the two signals have equal norm. Hence, the complement of the SB-MSLR regime can be equivalently expressed through the condition $\phi\boldsymbol{\beta}_1+(1-\phi)\boldsymbol{\beta}_2\neq 0$.

Let $\texttt{CORR}(\boldsymbol{X},\boldsymbol{y})$ denote the output of running CORR on inputs $\boldsymbol{X},\boldsymbol{y}$. Consider the test function

$$g\left(\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{y}\end{bmatrix}\right) := \begin{cases}\texttt{p} & \texttt{CORR}(\boldsymbol{X},\boldsymbol{y})\neq\emptyset\\ \texttt{q} & \texttt{CORR}(\boldsymbol{X},\boldsymbol{y})=\emptyset\end{cases}.$$

Let $\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{y}\end{bmatrix}\sim\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y})$. Recall that CORR outputs the following set

$$\texttt{CORR}(\boldsymbol{X},\boldsymbol{y}) = \left\{j\in[p]:\left|\frac{\langle\boldsymbol{X}_j,\boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right|\geq\sqrt{2(1+\epsilon/2)\log 2p}\right\}.$$

As in the proof of Theorem 37, we note that

$$\mathbb{P}\left[\max_{q\in[p]}\left|\frac{\langle \boldsymbol{X}_q, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \geq \sqrt{2(1+\epsilon/2)\log 2p}\right] \leq \mathbb{P}\left[\max_{q\in[p]}\left|\frac{\langle \boldsymbol{X}_q, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \geq \sqrt{2\log 2p} + \frac{\epsilon}{2\sqrt{8}}\sqrt{\log 2p}\right]. \tag{26}$$

Noting that for $\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{y}\end{bmatrix} \sim \mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y})$ we have that $\frac{\langle \boldsymbol{X}_q, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, we apply Lemma 45 to (26) and obtain that

$$\mathbb{P}\left[\max_{q\in[p]}\left|\frac{\langle \boldsymbol{X}_j, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \geq \sqrt{2(1+\epsilon/2)\log 2p}\right] \leq (2p)^{-\frac{1}{8}\left(\frac{\epsilon}{2}\right)^2},$$

and hence we have that, under $\mathbb{P}(\boldsymbol{X})\otimes\mathbb{P}(\boldsymbol{y})$, $g\left(\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{y}\end{bmatrix}\right) = \mathsf{q}$ with probability $1-o(1)$.

Conversely, let $\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{y}\end{bmatrix} \sim \mathbb{P}(\boldsymbol{X},\boldsymbol{y})$. Let $J \subseteq \text{supp}(\boldsymbol{\beta}_1)\cup\text{supp}(\boldsymbol{\beta}_2)$ such that $\phi\boldsymbol{\beta}_{1,j}+(1-\phi)\boldsymbol{\beta}_{2,j} \neq 0$ for $j \in J$. From Lemma 39, we then have that,

$$\begin{aligned}
\mathbb{P}\left[|\texttt{CORR}(\boldsymbol{X},\boldsymbol{y})| = 0\right] &= \mathbb{P}\left[\cap_{j\in[p]}\left\{\left|\frac{\langle \boldsymbol{X}_j, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \leq \sqrt{2(1+\epsilon/2)\log 2p}\right\}\right]\\
&\leq \mathbb{P}\left[\cap_{j\in J}\left\{\left|\frac{\langle \boldsymbol{X}_j, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \leq \sqrt{2(1+\epsilon/2)\log 2p}\right\}\right]\\
&\leq \mathbb{P}\left[\left|\frac{\langle \boldsymbol{X}_{J_1}, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \leq \sqrt{2(1+\epsilon/2)\log 2p}\right]\\
&\leq \left(\frac{1}{p} + 6e^{\frac{\delta^2 n}{w}} + \frac{39}{\delta^2 n}\right) = o(1),
\end{aligned}$$

and hence we have that, under $\mathbb{P}(\boldsymbol{X},\boldsymbol{y})$, $g\left(\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{y}\end{bmatrix}\right) = \mathsf{p}$ with probability $1-o(1)$. ■

D.3. Recovery algorithms for MSLR

General recovery algorithm for MSLR A recovery algorithm for MSLR in the noiseless and balanced regimes is given in Theorem 8. We measure the recovery error in MSLR as in Chen et al. (2014),

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

where $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\beta}}_2)$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)$.

Proof [Proof of Theorem 8] In the case of $\sigma = 0, \phi \neq 1/2$ (noiseless), we first apply CORR and then the Alternating Minimization (AM) algorithm of Yi et al. (2014). Theorem 48 shows that this succeeds in the regime of interest.

In the case of $\phi = 1/2$ (balanced), $\texttt{SNR} = \Omega(k)$, we first apply CORR and then the algorithm of Chen et al. (2014). Theorem 56 proves that in the high SNR regime $\texttt{SNR} = \Omega(1)$, we have that

$\rho(\hat{\theta}, \theta) = \Theta\left(\sigma\sqrt{\frac{k}{n}}\right)$. In order for exact recovery to be achieved for all allowable finite n, we would require $\sqrt{\frac{\sigma^2 k}{(\|\boldsymbol{\beta}\|_2^2+\sigma^2)\log p}} \to 0$, which is satisfied for $\mathtt{SNR} = \|\boldsymbol{\beta}\|_2^2/\sigma^2 = \Omega(k)$, implying a non-vanishing signal-to-noise ratio with respect to the support set. This is satisfied by hypothesis, and hence CORR together with Algorithm 6 succeeds in solving asymptotically exact recovery (up to relabeling of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$) in the regime of interest. ■

Recovery algorithm for noiseless, unbalanced MSLR In this section we will show that the CORR algorithm can be used to reduce the MSLR problem to a dense problem where $n, k = \Theta(p)$ and the signal is not assumed to be sublinearly sparse, where state-of-the-art algorithms for this dense mixed linear regression case can then infer $\boldsymbol{\beta}_1$ from $\boldsymbol{\beta}_2$. In the noiseless case with $\phi \neq 1/2$, we can apply CORR together with the existing polynomial-time Alternating Minimization (AM) algorithm from Yi et al. (2014) to fully solve MSLR in what is a constant number of steps. We recall that Theorem 37 only provided guarantees on the support recovery of the joint signal, whereas now with the execution of the AM algorithm on the reduced joint support set one can fully infer $\boldsymbol{\beta}_1$ from $\boldsymbol{\beta}_2$. Define the mixture proportions,

$$\frac{n_1}{n} := \frac{\sum_{i=1}^n z_i}{n}$$
$$\frac{n_2}{n} := \frac{\sum_{i=1}^n (1 - z_i)}{n}.$$

We begin by stating the theorem.

Theorem 48 (Success of CORR+ AM on noiseless MSLR) *Consider the general setting of* MSLR *with parameters $p, n, k, \sigma = 0$, $\phi \neq 1/2$, $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \sim \mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$. Suppose*

$$n \geq \frac{32}{\min\{\phi^2 \boldsymbol{\beta}_{min}^2, (1-\phi)^2 \boldsymbol{\beta}_{min}^2, \langle \boldsymbol{\beta} \rangle_{min}^2\}} (1+\epsilon)\|\boldsymbol{\beta}\|_2^2 \log 2p$$

for $\epsilon \in (0,1)$ used in CORR *. Then with probability at least $1 - c_1(\frac{k}{p} + ke^{-c_2 n} + \frac{k}{n} + \frac{1}{p^{c_2}})$ for constants $c_1, c_2 > 0$, the output $\hat{\theta} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ of* CORR+ *Algorithm 3 + Algorithm 5 satisfies*

$$\rho(\hat{\theta}, \theta) = 0.$$

Proof [Proof of Theorem 48] By Theorem 37, we can recover the joint support set (at most of size $2k$) of signals $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ with probability at least $1 - c_1(\frac{k}{p} + ke^{-c_2 n} + \frac{k}{n} + \frac{1}{p^{c_2}})$ by running CORR.

After running CORR and identifying the $< 2k$ joint support set indices, we restrict the regression problem to these indices by removing all other columns from the design matrix $\boldsymbol{X}$ (as these do not influence the output y since they do not correspond to support indices of $\boldsymbol{\beta}_1$ or $\boldsymbol{\beta}_2$). We are then tasked with solving a two-component mixtures of regressions problem with n samples and signals of dimension between k and $2k$ (importantly, the dimension is no longer p).

From this point, the idea is to spectrally initialize $(\boldsymbol{\beta}_1^{(0)}, \boldsymbol{\beta}_2^{(0)})$ using Algorithm 3 for which Proposition 50 provides guarantees, pass $(\boldsymbol{\beta}_1^{(0)}, \boldsymbol{\beta}_2^{(0)})$ into Algorithm 5 for which Theorem 51 provides guarantees on geometric error decay given this initialization, and run Algorithm 5 for a finite number of iterations guaranteed by Proposition 52.

The condition on sample size n of Proposition 50 is met, since it is assumed that $n \gtrsim \|\boldsymbol{\beta}\|_2^2 \log 2p \geq \boldsymbol{\beta}_{\mathtt{min}}^2 k \log 2p$ and $k \log 2p = \omega(k \log^2 k)$. The condition of Theorem 51 is met by the result of Proposition 50. The condition of Proposition 52 is met by running Algorithm 5 (resampling) with $O(k \log^2 k)$ samples (see Remark 53).

What remains to show is that $n_1 \neq n_2$, as this is required by Remark 49. Recall $n_1 = \sum_{i=1}^n z_i$ and $n_2 = \sum_{i=1}^n (1 - z_i)$ and z_i are independent Bernoulli(ϕ). Without loss of generality assuming $\phi < 1/2$, there exists a $\delta \in (0, 1)$ such that,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n z_i \geq 1/2\right] \leq \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n z_i \geq (1+\delta)\phi\right] \leq e^{-\Theta(n)},$$

after applying a standard Chernoff bound. This high probability statement can be absorbed into the $1 - c_1(\frac{k}{p} + ke^{-c_2 n} + \frac{k}{n} + \frac{1}{p^{c_2}})$ high probability statement provided by CORR , choosing adjusted constants $c_1, c_2 > 0$.

We hence conclude that running CORR followed by Algorithm 3 followed by Algorithm 5 we obtain $\rho(\theta^{(t)}, \theta) = 0$ in finite t with probability at least $1 - c_1(\frac{k}{p} + ke^{-c_2 n} + \frac{k}{n} + \frac{1}{p^{c_2}})$. ■

Their initialization algorithm is based on the positive semidefinite matrix:

$$M := \frac{1}{n}\sum_{i=1}^n y_i^2 \boldsymbol{x}_i \otimes \boldsymbol{x}_i$$

which serves as an unbiased estimator of a matrix whose two largest eigenvectors span the space spanned by $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$.

Remark 49 *It is stated in Fan et al. (2018) that, when the mixture frequencies are equal to each other ($n_1 = n_2$), the top two eigenvectors of $\mathbb{E}M$ will not be $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$. Hence, their algorithms only work for the case $\phi \neq 1/2$ and $\sigma = 0$ (noiseless).*

Outside of the case $\phi = 1/2$, when the mixture proportions are known, an approximation of $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ can be computed in closed form through Algorithm 3, where

$$sign(b) = \begin{cases} 1, & b = 1 \\ -1, & b = 2. \end{cases}$$

In what follows, we state the iterative algorithms proposed in Fan et al. (2018) and their guarantees.

Algorithm 3: Initialization with proportion information
Data: Input: n_1, n_2, samples $\{(y_i, \boldsymbol{x}_i), i = 1, 2, ..., n\}$
$M \leftarrow \frac{1}{N}\sum_{i=1}^N y_i^2 \boldsymbol{x}_i \otimes \boldsymbol{x}_i$
Compute top 2 eigenvectors and eigenvalues $(v_b, \lambda_b), b = 1, 2$ of $(M - I)/2$
Compute $\boldsymbol{\beta}_b^{(0)} = \sqrt{\frac{1-\Delta_b}{2}} v_b + sign(b)\sqrt{\frac{1+\Delta_b}{2}} v_{-b}$, where $\Delta_b = \frac{(\lambda_b - \lambda_{-b})^2 + n_b^2 - n_{-b}^2}{2(\lambda_{-b} - \lambda_b)n_b}, \; b = 1, 2$
return $\boldsymbol{\beta}_1^{(0)}, \boldsymbol{\beta}_2^{(0)}$

Algorithm 4: AM

Data: Initial $\boldsymbol{\beta}_1^{(0)}, \boldsymbol{\beta}_2^{(0)}$, # iterations t_0, samples $\{(y_i, \boldsymbol{x}_i), i = 1, 2, ..., N\}$

for $t = 0, \cdots, t_0 - 1$ **do**
- $J_1, J_2 \leftarrow \emptyset$
- **for** $i = 1, 2, \cdots, N$ **do**
  - **if** $\left| y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_1^{(t)} \rangle \right| < \left| y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}_2^{(t)} \rangle \right|$ **then**
    - $J_1 \leftarrow J_1 \cup \{i\}$
  - **else**
    - $J_2 \leftarrow J_2 \cup \{i\}$
  - **end**
- **end**
- $\boldsymbol{\beta}_1^{(t+1)} \leftarrow \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \|\boldsymbol{y}_{J_1} - \boldsymbol{X}_{J_1} \boldsymbol{\beta}\|_2$
- $\boldsymbol{\beta}_2^{(t+1)} \leftarrow \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \|\boldsymbol{y}_{J_2} - \boldsymbol{X}_{J_2} \boldsymbol{\beta}\|_2$

end

return $\boldsymbol{\beta}_1^{(t_0)}, \boldsymbol{\beta}_2^{(t_0)}$

Proposition 50 *(Yi et al., 2014) Consider the initialization method in Algorithm 3. Given any constant $\widehat{c} < 1/2$, with probablity at least $1 - \frac{1}{p^2}$, the approach produces an initialization $(\boldsymbol{\beta}_1^{(0)}, \boldsymbol{\beta}_2^{(0)})$ satisfying*

$$\rho(\theta^{(0)}, \theta) \ \leq \ \widehat{c} \min\{n_1/n, n_2/n\} \, \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2,$$

if

$$n \geq c_1 \left(\frac{1}{\widetilde{\widetilde{\delta}}} \right)^2 p \log^2 p.$$

Here c_1 is a constant that depends on $\widehat{c}$. And

$$\sqrt{\widetilde{\delta}} = \widehat{c} \sqrt{\min\{n_1/n, n_2/n\}}^3 \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2 (\sqrt{1 - \kappa}) \kappa,$$

where $\kappa = \sqrt{1 - 4(1 - \langle \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \rangle^2) \frac{n_1}{n} \frac{n_2}{n}}$.

Algorithm 5: AM with resampling

Data: Initial $\boldsymbol{\beta}_1^{(0)}, \boldsymbol{\beta}_2^{(0)}$, # iterations t_0, samples $\{(y_i, \boldsymbol{x}_i), i = 1, 2, ..., N\}$

Partition the samples $\{(y_i, \boldsymbol{x}_i)\}$ into t_0 disjoint sets: $\mathcal{S}_1, ..., \mathcal{S}_{t_0}$

for $t = 1, \cdots, t_0$ **do**
- Use $\mathcal{S}_t$ to run Algorithm 4 initialized with $(\boldsymbol{\beta}_1^{(t-1)}, \boldsymbol{\beta}_2^{(t-1)})$ and returning $(\boldsymbol{\beta}_1^t, \boldsymbol{\beta}_2^t)$

end

return $\boldsymbol{\beta}_1^{(t_0)}, \boldsymbol{\beta}_2^{(t_0)}$

Theorem 51 *Yi et al. (2014) Consider one iteration in Algorithm 5. For fixed $(\boldsymbol{\beta}_1^{(t-1)}, \boldsymbol{\beta}_2^{(t-1)})$, there exist absolute constants $\widetilde{c}, c_1, c_2$ such that if*

$$\rho(\theta^{(t-1)}, \theta) \ \leq \ \widetilde{c} \min\{n_1/n, n_2/n\} \, \|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|_2,$$

and if the number of samples in that iteration satisfies

$$|\mathcal{S}_t| \;\geq\; \left(\frac{c_1}{\min\{n_1/n, n_2/n\}}\right) p,$$

then with probability greater than $1-\exp(-c_2 p)$ *we have a geometric decrease in the error at the next stage, i.e.*

$$\rho(\theta^{(t)}, \theta) \leq \frac{1}{2}\rho(\theta^{(t-1)}, \theta)$$

Proposition 52 *(Exact Recovery) Yi et al. (2014) There exist absolute constants* c_1, c_2 *such that if*

$$\rho(\theta^{(t-1)}, \theta) \leq \frac{c_1}{p^2}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2$$

and

$$\frac{1}{\min\{n_1/n, n_2/n\}} p < |\mathcal{S}_t| < c_2 p,$$

then with probability greater than $1 - \frac{1}{p}$,

$$\rho(\theta^{(t)}, \theta) = 0.$$

Remark 53 *It is easy to see, and remarked in Fan et al. (2018) (pp. 11-12, above and below Proposition 4) that, running Algorithm 5 with guarantees given in Theorem 51, one would require* $O(p\log^2 p)$ *samples to obtain* $\rho(\theta^{(t-1)}, \theta) \leq \frac{c_1}{p^2}\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_2^*\|_2$ *for the constant* $c_1 > 0$ *suitable for Proposition 52.*

Algorithm for `MSLR` in the balanced case Consider the case of Mixtures of Linear Regressions (`MLR`) as in Chen et al. (2014), where $n, k = \Theta(p)$, and σ^2 is known. Further, consider the balanced regime where $\phi = 1/2$, but $\boldsymbol{\beta}_{1,j} \neq -\boldsymbol{\beta}_{2,j}$ for any $j \in \mathcal{S}_1 \cap \mathcal{S}_2$ ($\langle \boldsymbol{\beta} \rangle^2_{\mathtt{min}} \neq 0$). We claim that we can solve the aforementioned `MSLR` problem in this regime by recovering the joint support of $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ using `CORR` , and then running the algorithm in Chen et al. (2014) for general mixed linear regression. This latter Algorithm 6 is outlined below. As motivated in (Chen et al., 2014), the

Algorithm 6: Estimate $\boldsymbol{\beta}$'s (Chen et al., 2014)
Data: $(\boldsymbol{X}, \boldsymbol{y}) \in \mathbb{R}^{n\times p} \times \mathbb{R}^n$
Let $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}}) := \arg\min_{\boldsymbol{K},\boldsymbol{g}} \sum_{i=1}^n \left(-\left\langle \boldsymbol{x}_i\boldsymbol{x}_i^\top, \boldsymbol{K}\right\rangle + 2y_i\left\langle \boldsymbol{x}_i, \boldsymbol{g}\right\rangle - y_i^2 + \sigma^2\right)^2 + \lambda\left\|\boldsymbol{K}\right\|_*$
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 $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{g}} \pm \sqrt{\hat{\lambda}}\hat{\boldsymbol{v}}$
return $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$

algorithm performs a convex penalized least squares optimization to determine matrix and vector $(\hat{\boldsymbol{K}}, \hat{\boldsymbol{g}})$, $\hat{\boldsymbol{g}}$ being a naive estimate of $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and the leading eigenvector-eigenvalue of $\hat{\boldsymbol{J}}$ a necessary correction. We define the value $\alpha := \frac{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2}{\|\boldsymbol{\beta}_1\|_2^2 + \|\boldsymbol{\beta}_2\|_2^2}$.

Definition 54 *(Chen et al., 2014) Let* $n_1 = \{i \in [n] : z_i = 1\}$ *denote the number of samples obtained from* $\boldsymbol{\beta}_1$*, and* n_2 *the analogous for* $\boldsymbol{\beta}_2$*. We define the following regularity conditions, required for our further proofs:*

1. *$\boldsymbol{X}$ is an i.i.d. standard Gaussian matrix,*
2. $\alpha \geq c_3$,
3. $\min\{n_1, n_2\} \geq c_4 p$,
4. $\lambda = \Theta(\sigma(\|\boldsymbol{\beta}_1\|_2 + \|\boldsymbol{\beta}_2\|_2 + \sigma)\sqrt{np}\log^3 n)$,
5. $n \geq c_3 p \log^8 n$,
6. $|n_1 - n_2| = O\left(\sqrt{n \log n}\right)$,

for some constants $0 < c_3 < 2$ and c_4.

Theorem 55 *Chen et al. (2014) Suppose the conditions in Definition 54 hold. There exist constants $c_1, c_2, c_4 > 0$ such that with probability at least $1 - c_1 n^{-c_2}$, the output $\hat{\theta} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ of* `Algorithm 3` *satisfies*

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

The result in Theorem 55 implies that, in the high-snr regime $\|\boldsymbol{\beta}\|_2^2/\sigma^2 = \Omega(1)$, we have that $\rho(\hat{\theta}, \theta) = \Theta(\sigma\sqrt{\frac{p}{n}})$. This holds since $\|\boldsymbol{\beta}_1\|_2 + \|\boldsymbol{\beta}_2\|_2 = 2\sqrt{k}$.

Theorem 56 (Success of `CORR` + `Algorithm` 6 on `MSLR`) *Consider the general setting of* `MSLR` *with parameters $p, n, k, \phi = 1/2$, $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \sim \mathcal{P}_{\|\boldsymbol{\beta}\|_2}(\mathcal{D})$. Suppose the conditions of Definition 54 hold. There exist constants $c_1, c_2, c_4 > 0$ such that, provided*

$$n \geq \frac{32}{\min\{\phi^2 \boldsymbol{\beta}_{min}^2, (1-\phi)^2 \boldsymbol{\beta}_{min}^2, \langle \boldsymbol{\beta} \rangle_{min}^2\}} (1 + \epsilon)(\|\boldsymbol{\beta}\|_2^2 + \sigma^2) \log 2p$$

for $\epsilon \in (0, 1)$ used in `CORR` *, with probability at least $1 - c_1(\frac{k}{p} + ke^{-c_2 n} + \frac{k}{n} + \frac{1}{p^{c_2}})$, the output $\hat{\theta} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ of* `CORR` + `Algorithm` 6 *satisfies*

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

Proof [Proof of Theorem 56] By Theorem 37, we can recover the joint support set (at most of size $2k$) of signals $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ with probability at least $1 - c_1(\frac{k}{p} + ke^{-c_2 n} + \frac{k}{n} + \frac{1}{p^{c_2}})$ by running `CORR`.

After running `CORR` and identifying the $< 2k$ joint support set indices, we restrict the regression problem to these indices by removing all other columns from the design matrix $\boldsymbol{X}$ (as these do not influence the output y since they do not correspond to support indices of $\boldsymbol{\beta}_1$ or $\boldsymbol{\beta}_2$). We are then tasked with solving a two-component mixtures of regressions problem with n samples and signals of dimension between k and $2k$. We run `Algorithm 3` on this simplified regression problem, to obtain $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$.

What remains is to show that the assumptions of Theorem 55 hold with p replaced by $2k$. Indeed, condition 1. holds since without loss of generality we can assume $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$, otherwise the problem setting would be that of SLR. Condition 3., 4., 5. hold by the definition of the MSLR problem, and by freedom with respect to λ and c_3.

Condition 2. holds with probability at least $1 - \exp(-\Theta(n))$, indeed by Hoeffding's inequality (Boucheron et al., 2013) we have that

$$\begin{aligned}\mathbb{P}\left[n_1 < 2c_4 k\right] &= \mathbb{P}\left[\sum_{i=1}^{n} z_i < 2c_4 k\right] \\ &= \mathbb{P}\left[\sum_{i=1}^{n}(z_i - \phi) < 2c_4 k - \phi n\right] \\ &= \mathbb{P}\left[\sum_{i=1}^{n}(-z_i + \phi) \geq \phi n - 2c_4 k\right] \\ &\leq \exp\left(-\Theta(n)\right),\end{aligned}$$

where $\phi n - 2c_4 k$ is positive for some $c_4 > 0$ since by assumption $n \geq \frac{32(1+\epsilon)}{\min\{(2\phi-1)^2, \phi^2, (1-\phi)^2\}}(1+\epsilon)(k+\sigma^2)\log p$ and $(k+\sigma^2)\log p = \omega(k)$, analogously for n_2.

In the case $\phi = 1/2$ (so that $2\phi - 1 = 0$), we have that condition 6. holds with probability $\exp(-\Theta(\log n))$ by Hoeffding's inequality,

$$\begin{aligned}&\mathbb{P}\left[\left|\sum_{i=1}^{n} z_i - \sum_{i=1}^{n}(1 - z_i)\right| \geq \Theta(\sqrt{n \log n})\right] \\ &\leq \mathbb{P}\left[\sum_{i=1}^{n}(2z_i - 1) \geq \Theta(\sqrt{n \log n})\right] + \mathbb{P}\left[\sum_{i=1}^{n}(-2z_i + 1) \geq \Theta(\sqrt{n \log n})\right] \\ &\leq 2\exp\left(-\frac{2\Theta(n \log n)}{2n}\right) = \exp\left(-\Theta(\log n)\right).\end{aligned}$$

The above events occur with probability at least $1 - \exp(-\Theta(\log n))$, and the event that CORR succeeds occurs with probability $1 - c_1(\frac{k}{p} + ke^{-c_2 n} + \frac{k}{n} + \frac{1}{p^{c_2}})$. These two events occur together with probability at least $1 - c_1(\frac{k}{p} + ke^{-c_2 n} + \frac{k}{n} + \frac{1}{p^{c_2}})$ for adjusted constants $c_1, c_2 > 0$. Applying Theorem 55, we obtain the result. ■

D.4. CORR for signed support recovery in SLR

Proof [Proof of Theorem 11] The proof proceeds similarly as that of Theorem 7. Let CORR$(\boldsymbol{X}, \boldsymbol{y})$ denote the output of running CORR on inputs $\boldsymbol{X}, \boldsymbol{y}$. Consider the test function

$$g\left(\begin{bmatrix}\boldsymbol{X} \\ \boldsymbol{y}\end{bmatrix}\right) := \begin{cases}\mathtt{p} & \mathtt{CORR}(\boldsymbol{X}, \boldsymbol{y}) \neq \emptyset \\ \mathtt{q} & \mathtt{CORR}(\boldsymbol{X}, \boldsymbol{y}) = \emptyset\end{cases}.$$

Let $\begin{bmatrix}\boldsymbol{X} \\ \boldsymbol{y}\end{bmatrix} \sim \mathbb{P}(\boldsymbol{X}) \otimes \mathbb{P}(\boldsymbol{y})$. Recall that CORR outputs the following set

$$\texttt{CORR}(\boldsymbol{X}, \boldsymbol{y}) = \left\{ j \in [p] : \left| \frac{\langle \boldsymbol{X}_j, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| \geq \sqrt{2(1+\epsilon/2)\log 2p} \right\}.$$

As in the proof of Theorem 12, we note that for $\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{y} \end{bmatrix} \sim \mathbb{P}(\boldsymbol{X}) \otimes \mathbb{P}(\boldsymbol{y})$ we have $\frac{\langle \boldsymbol{X}_q, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Applying Lemma 45 to (27) we obtain,

$$\mathbb{P}\left[\max_{q \in [p]} \left| \frac{\langle \boldsymbol{X}_q, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| \geq \sqrt{2(1+\epsilon/2)\log 2p} \right] \leq \mathbb{P}\left[\max_{q \in [p]} \left| \frac{\langle \boldsymbol{X}_q, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| \geq \sqrt{2\log 2p} + \frac{\epsilon}{2\sqrt{8}}\sqrt{\log 2p} \right], \tag{27}$$

$$\leq p^{-\frac{1}{16}\left(\frac{\epsilon}{2}\right)^2} = o(1),$$

and hence we have that, under $\mathbb{P}(\boldsymbol{X}) \otimes \mathbb{P}(\boldsymbol{y})$, $g\left(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{y} \end{bmatrix}\right) = \mathsf{q}$ with probability $1 - o(1)$.

Conversely, let $\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{y} \end{bmatrix} \sim \mathbb{P}(\boldsymbol{X}, \boldsymbol{y})$. We then apply Theorem 12 to deduce that $\texttt{CORR}(\boldsymbol{X}, \boldsymbol{y}) = \text{supp}(\boldsymbol{\beta}) \neq \emptyset$ with probability at least $1 - \left(\frac{k}{p} + 2ke^{-c_2 n} + \frac{1}{p^{c_2}}\right)$ for some constant $c_2 > 0$. Hence, under $\mathbb{P}(\boldsymbol{X}, \boldsymbol{y})$, $g\left(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{y} \end{bmatrix}\right) = \mathsf{p}$ with probability $1 - o(1)$. ■

In what follows, we consider the modified CORR algorithm in (3) for estimating the *signed* support of $\boldsymbol{\beta}$.

Proof [Proof of Theorem 12] Let $\mathcal{S}$ denote the support set of $\boldsymbol{\beta}$. Define the error event

$$\mathcal{E} = \cup_{j \in \mathcal{S}} \left\{ \left| \frac{\langle \boldsymbol{X}_j, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| < \sqrt{2(1+\epsilon/2)\log 2p} \right\} \cup \left\{ \max_{q \in \mathcal{S}^\complement} \left| \frac{\langle \boldsymbol{X}_q, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| \geq \sqrt{2(1+\epsilon/2)\log 2p} \right\}.$$

The theorem claim follows by demonstrating that $\mathbb{P}[\mathcal{E}] = o(1)$. With this in mind, we perform a union bound

$$\begin{aligned}\mathbb{P}[\mathcal{E}] &\leq k\mathbb{P}\left[\left| \frac{\langle \boldsymbol{X}_{j^*}, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| < \sqrt{2(1+\epsilon/2)\log 2p} \right] + \mathbb{P}\left[\max_{q \in \mathcal{S}^\complement} \left| \frac{\langle \boldsymbol{X}_q, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| \geq \sqrt{2(1+\epsilon/2)\log 2p} \right] \\ &:= k\nu_1 + \nu_2\end{aligned}$$

where $j^* \in \mathcal{S}$. We first focus on ν_2, where we notice $\frac{\langle \boldsymbol{X}_q, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ for $q \in \mathcal{S}^\complement$. Applying Lemma 45, we deduce that

$$\begin{aligned}\nu_2 &= \mathbb{P}\left[\max_{q \in \mathcal{S}^\complement} \left| \frac{\langle \boldsymbol{X}_q, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| \geq \sqrt{2(1+\epsilon/2)\log 2p} \right] \\ &\leq \mathbb{P}\left[\max_{q \in \mathcal{S}^\complement} \left| \frac{\langle \boldsymbol{X}_q, \boldsymbol{y} \rangle}{\|\boldsymbol{y}\|_2} \right| \geq \sqrt{2\log 2p} + \frac{\epsilon}{2\sqrt{8}}\sqrt{\log 2p} \right] \\ &\leq (2p)^{-\frac{1}{16}\left(\frac{\epsilon}{2}\right)^2},\end{aligned}$$

since $\sqrt{2(1+\epsilon/2)} \geq \sqrt{2} + \frac{\epsilon}{2\sqrt{8}}$. Setting $n \geq \frac{8(1+\epsilon)}{\boldsymbol{\beta}_{\min}^2}(\|\boldsymbol{\beta}\|_2^2 + \sigma^2)\log 2p$ for some $\epsilon \in (0,1)$, the bound for ν_1 follows from applying Lemma 57 below:

$$\nu_1 \leq \frac{1}{2p} + 2e^{-\frac{\delta^2 n}{8}}.$$

Putting it all together, we obtain that $\mathbb{P}[\mathcal{E}] \leq \frac{k}{2p} + 2ke^{-c_2 n} + \frac{1}{p^{c_2}}$, for some constant $c_2 > 0$. ■

Lemma 57 (Concentration bound for SLR) *Consider the setting of* SLR *for* $j^* \in \mathcal{S}$. *Then for* $n \geq \frac{8(1+\epsilon)}{\boldsymbol{\beta}_{\min}^2}(\|\boldsymbol{\beta}\|_2^2 + \sigma^2)\log 2p$ *we have*

$$\mathbb{P}\left[\left|\frac{\langle \boldsymbol{X}_{j^*}, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \leq \sqrt{2(1+\epsilon/2)\log 2p}\right] \leq \frac{1}{p} + 2e^{-\frac{\delta^2 n}{8}}$$

Proof Let $\nu := \mathbb{P}\left[\left|\frac{\langle \boldsymbol{X}_{j^*}, \boldsymbol{y}\rangle}{\|\boldsymbol{y}\|_2}\right| \leq \sqrt{2(1+\epsilon/2)\log 2p}\right]$ and $\delta > 0$ to be chosen later. We begin by conditioning on the event $\Omega_1 := \Omega_1(\delta)$ from Lemma 45, and apply Lemma 41 (for the case $z_i = 1$) denoting $g_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ for $i \in [n]$ as independent (also from y, z) unit normal Gaussians, obtaining

$$\begin{aligned}
\nu &= \int \mathbb{P}\left[\left|\frac{\boldsymbol{X}_{j^*}^T \xi}{\|\xi\|_2}\right| < \sqrt{2(1+\epsilon/2)\log 2p} \;\middle|\; \boldsymbol{y}\right] d\mathbb{P}[y] \\
&= \int \mathbb{P}\left[\left|\sum_{i=1}^n g_i \frac{y_i}{\|\boldsymbol{y}\|_2}\sqrt{1 - \frac{\boldsymbol{\beta}_{j^*}^2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}} + \frac{\boldsymbol{\beta}_{j^*} \cdot y_i^2/\|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}\right| < \sqrt{2(1+\epsilon/2)\log 2p} \;\middle|\; \boldsymbol{y}\right] d\mathbb{P}[y] \\
&= \int \mathbb{P}\left[\left|g_1\sqrt{1 - \frac{\boldsymbol{\beta}_{j^*}^2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}} + \frac{\boldsymbol{\beta}_{j^*} \cdot \|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}\right| < \sqrt{2(1+\epsilon/2)\log 2p} \;\middle|\; \boldsymbol{y}\right] d\mathbb{P}[y] \\
&\leq \mathbb{P}\left[\left|g_1\sqrt{1 - \frac{\boldsymbol{\beta}_{j^*}^2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}} + \frac{\boldsymbol{\beta}_{j^*} \cdot \|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}\right| < \sqrt{2(1+\epsilon/2)\log 2p} \;\middle|\; \Omega_1\right] + \mathbb{P}\left[\Omega_1^{\complement}\right] \\
&= \mathbb{P}\Bigg[\left\{g_1\sqrt{1 - \frac{\boldsymbol{\beta}_{j^*}^2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}} + \frac{\boldsymbol{\beta}_{j^*} \cdot \|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2} < \sqrt{2(1+\epsilon/2)\log 2p}\right\} \\
&\qquad \cap \left\{-\sqrt{2(1+\epsilon/2)\log 2p} \leq g_1\sqrt{1 - \frac{\boldsymbol{\beta}_{j^*}^2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}} + \frac{\boldsymbol{\beta}_{j^*} \cdot \|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}\right\} \;\Bigg|\; \Omega_1\Bigg] + \mathbb{P}\left[\Omega_1^{\complement}\right] \\
&\leq \mathbb{P}\left[g_1\sqrt{1 - \frac{\boldsymbol{\beta}_{j^*}^2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}} < \sqrt{2(1+\epsilon/2)\log 2p} - \left|\frac{\boldsymbol{\beta}_{j^*} \cdot \|\boldsymbol{y}\|_2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}\right| \;\middle|\; \Omega_1\right] + \mathbb{P}\left[\Omega_1^{\complement}\right] \\
&\leq \mathbb{P}\left[g_1\sqrt{1 - \frac{\boldsymbol{\beta}_{j^*}^2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}} < \sqrt{2(1+\epsilon/2)\log 2p} - \left|\frac{\boldsymbol{\beta}_{j^*} \cdot \sqrt{n(\|\boldsymbol{\beta}\|_2^2 + \sigma^2)(1-\delta)}}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}\right| \;\middle|\; \Omega_1\right] + \mathbb{P}\left[\Omega_1^{\complement}\right] \\
&= \mathbb{P}\left[g_1\sqrt{1 - \frac{\boldsymbol{\beta}_{j^*}^2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}} < \sqrt{2(1+\epsilon/2)\log 2p} - \left|\boldsymbol{\beta}_{j^*}\sqrt{\frac{n(1-\delta)}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}}\right| \;\middle|\; \Omega_1\right] + \mathbb{P}\left[\Omega_1^{\complement}\right].
\end{aligned}$$

Now setting $n \geq \frac{8(1+\epsilon)}{\boldsymbol{\beta}^2_{\min}}(\|\boldsymbol{\beta}\|_2^2 + \sigma^2)\log 2p$ and applying standard sub-Gaussian bounds (see Wainwright (2019)) we obtain

$$\nu \leq \mathbb{P}\left[g\sqrt{1 - \frac{\boldsymbol{\beta}_{j^*}^2}{\|\boldsymbol{\beta}\|_2^2 + \sigma^2}} < (\sqrt{2(1+\epsilon/2)} - \sqrt{8(1+\epsilon)(1-\delta)})\sqrt{\log 2p}\right] + \mathbb{P}\left[\Omega_1^{\complement}\right]$$
$$\leq \exp\left(-\frac{(\sqrt{2(1+\epsilon/2)} - \sqrt{8(1+\epsilon)(1-\delta)})^2 \log 2p}{2\left(1 - \frac{\boldsymbol{\beta}_{j^*}^2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right)}\right) + 2e^{-\frac{\delta^2 n}{8}},$$

where

$$\sqrt{2(1+\epsilon/2)} - \sqrt{8(1+\epsilon)(1-\delta)} < -\sqrt{2} \tag{28}$$

for $\delta > 0$ small enough. Hence, choosing δ to satisfy (28) above, we obtain that for k large enough,

$$\nu \leq \exp\left(-\frac{2\log 2p}{2\left(1 - \frac{\boldsymbol{\beta}_{j^*}^2}{\|\boldsymbol{\beta}\|_2^2+\sigma^2}\right)}\right) + 2e^{-\frac{\delta^2 n}{8}}$$
$$\leq \exp\left(-\frac{2\log 2p}{2}\right) + 2e^{-\frac{\delta^2 n}{8}}$$
$$\leq \frac{1}{2p} + 2e^{-\frac{\delta^2 n}{8}}.$$

■