Math. Stat. Learn. 3 (2020), 1–55
DOI 10.4171/MSL/21

Published by EMS Press

# The all-or-nothing phenomenon in sparse linear regression

Galen Reeves, Jiaming Xu and Ilias Zadik

**Abstract.** We study the problem of recovering a hidden binary $k$-sparse $p$-dimensional vector $\beta$ from $n$ noisy linear observations $Y = X\beta + W$, where $X_{ij}$ are i.i.d. $\mathcal{N}(0,1)$ and $W_i$ are i.i.d. $\mathcal{N}(0,\sigma^2)$. A closely related hypothesis testing problem is to distinguish the pair $(X, Y)$ generated from this structured model from a corresponding null model where $(X, Y)$ consist of purely independent Gaussian entries. In the low sparsity $k = o(\sqrt{p})$ and high signal-to-noise ratio $k/\sigma^2 \to \infty$ regime, we establish an "all-or-nothing" information-theoretic phase transition at a critical sample size $n^* = 2k\log(p/k)/\log(1 + k/\sigma^2)$, resolving a conjecture of Gamarnik and Zadik (2017). Specifically, we show that if $\liminf_{p\to\infty} n/n^* > 1$, then the maximum likelihood estimator almost perfectly recovers the hidden vector with high probability and moreover the true hypothesis can be detected with a vanishing error probability. Conversely, if $\limsup_{p\to\infty} n/n^* < 1$, then it becomes information-theoretically impossible even to recover an arbitrarily small but fixed fraction of the hidden vector support, or to test hypotheses strictly better than random guess.

Our proof of the impossibility result builds upon two key techniques, which could be of independent interest. First, we use a conditional second moment method to upper bound the Kullback–Leibler (KL) divergence between the structured and the null model. Second, inspired by the celebrated area theorem, we establish a lower bound to the minimum mean squared estimation error of the hidden vector in terms of the KL divergence between the two models.[1]

*Mathematics Subject Classification* (2020). 62J05, 94A15; 94A17, 60F10.

*Keywords.* Sparse regression, second moment method, area theorem.

## 1. Introduction

In this paper, we study the information-theoretic limits of the Gaussian sparse linear regression problem. Specifically, for $n, p, k \in \mathbb{N}$ with $k \le p$ and $\sigma^2 > 0$ we consider two independent matrices $X \in \mathbb{R}^{n\times p}$ and $W \in \mathbb{R}^{n\times 1}$ with $X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $W_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2)$, and observe

$$Y = X\beta + W, \tag{1}$$

[1] An extended abstract version of this work appeared at the Proceedings of the Conference on Learning Theory (COLT), 2019. http://proceedings.mlr.press/v99/reeves19a/reeves19a.pdf

where $\beta$ is assumed to be a random vector uniformly distributed on the set

$$\{v \in \{0,1\}^p : \|v\|_0 = k\}$$

and independent of $(X, W)$. The problem of interest is to recover $\beta$ given the knowledge of $X$ and $Y$. Our focus will be on identifying the minimal sample size $n$ for which the recovery is information-theoretically possible. We study the model in the high-dimensional regime, where $p \to +\infty$ and $n = n_p$, $k = k_p$, $\sigma^2 = \sigma_p^2$ are function of $p$, potentially growing to infinity when $p \to +\infty$.

The problem of recovering the support of a hidden sparse vector $\beta \in \mathbb{R}^p$ given noisy linear observations has been extensively analyzed in the literature, as it naturally arises in many contexts including subset regression, e.g., [31], signal denoising, e.g., [14], compressive sensing, e.g., [12, 16], information and coding theory, e.g., [25], as well as high dimensional statistics, e.g., [45,46]. The assumptions of Gaussianity of the entries of $(X, W)$ are standard in the literature. Furthermore, much of the literature (e.g., [1, 33, 47]) assumes a lower bound $\beta_{\min} > 0$ for the smallest magnitude of a nonzero entry of $\beta$, that is $\min_{i:\beta_i \neq 0} |\beta_i| \geq \beta_{\min}$, as otherwise identification of the support of the hidden vector is in principle impossible. In this paper we adopt a simplifying assumption by focusing only on binary vectors $\beta$, similar to other papers in the literature such as [1, 18] and [19]. In this case recovering the support of the vectors is equivalent to identifying the vector itself.

To judge the recovery performance we focus on the mean squared error (MSE). That is, given an estimator $\widehat{\beta}$ as a function of $(X, Y)$, define mean squared error as

$$\mathrm{MSE}(\widehat{\beta}) \triangleq \mathbb{E}\big[\|\widehat{\beta} - \beta\|^2\big],$$

where $\|v\|$ denotes the $\ell_2$ norm of a vector $v$. In our setting, one can simply choose $\widehat{\beta} = \mathbb{E}[\beta]$, which equals $\frac{k}{p}(1, 1, \ldots, 1)^\top$, and obtain a trivial

$$\mathrm{MSE}_0 = \mathbb{E}[\|\beta - \mathbb{E}[\beta]\|^2],$$

which equals $k\big(1 - \frac{k}{p}\big)$. We will adopt the following two natural notions of recovery, by comparing the MSE of an estimator $\widehat{\beta}$ to $\mathrm{MSE}_0$.

**Definition 1** (Strong and weak recovery). We say that $\widehat{\beta} = \widehat{\beta}(Y, X) \in \mathbb{R}^p$ achieves

- strong recovery if $\limsup_{p\to\infty} \mathrm{MSE}(\widehat{\beta})/\mathrm{MSE}_0 = 0$;
- weak recovery if $\limsup_{p\to\infty} \mathrm{MSE}(\widehat{\beta})/\mathrm{MSE}_0 < 1$.

The fundamental question of interest in this paper is when $n$ as a function of $(p, k, \sigma^2)$ is such that strong/weak recovery is information-theoretically possible. Here, and everywhere in this work, by information-theoretical possibility (respectively, impossibility) of weak/strong recovery we refer to the existence (respectively, absence) of a measurable, not necessarily binary-valued, estimator $\widehat{\beta}(Y, X)$ achieving weak/strong recovery.

The focus of this paper will be on sublinear sparsity levels, that is on $k = o(p)$. A great amount of literature has been devoted on the study of the problem in the linear regime where $n, k, \sigma = \Theta(p)$. One line of work has provided upper and lower bounds on the accuracy of support recovery as a function of the problem parameters, e.g., [1, 38, 39, 42]. Another line of work has derived explicit formulas for the minimum MSE (MMSE)

$$\mathbb{E}\big[\|\beta - \mathbb{E}[\beta \mid X, Y]\|^2\big].$$

These formulas were first obtained heuristically using the replica method from statistical physics [21, 43] and later proven rigorously in [9, 40]. However, to the best of our knowledge, none of the rigorous techniques of [9, 40] apply when $k = o(p)$. Although there has been significant work focusing directly on the sublinear sparsity regime, the identification of the exact information theoretic threshold of this fundamental statistical problem remains largely open (see Section 1.2 for a detailed discussion). Obtaining a tight characterization of the information-theoretic threshold is the main contribution of this work.

Towards identifying the information theoretic limits of recovering $\beta$, and out of independent interest, we also consider a closely related hypothesis testing problem, where the goal is to distinguish the pair $(X, Y)$ generated according to (1) from a model where both $X$ and $Y$ are independently generated. More specifically, given two independent matrices $X \in \mathbb{R}^{n\times p}$ and $W \in \mathbb{R}^{n\times 1}$ with $X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $W_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2)$, we define

$$Y \triangleq \lambda W, \tag{2}$$

where $\lambda > 0$ is a scaling parameter. We refer to the Gaussian linear regression model (1) as the planted model, denoted by $P = P(X, Y)$, and (2) as the null model denoted by $Q_\lambda = Q_\lambda(Y, X)$. We focus on characterizing the total variation distance $\mathrm{TV}(P, Q_\lambda)$ for various values of $\lambda$. One choice of particular interest is

$$\lambda = \sqrt{k/\sigma^2 + 1},$$

under which $\mathbb{E}[YY^\top] = (k + \sigma^2)\mathbf{I}$ in both the planted and null models.

Analogous to recovery, we adopt the following two natural notions of testing [3, 34].

**Definition 2** (Strong and weak detection). Fix two probability measures $P, Q$ on our observed data $(Y, X)$. We say a test statistic $\mathcal{T}(X, Y)$ with a threshold $\tau$ achieves

- strong detection if

$$\limsup_{p\to\infty} \big[P(\mathcal{T}(X,Y) < \tau) + Q(\mathcal{T}(X,Y) \geq \tau)\big] = 0,$$

- weak detection if

$$\limsup_{p\to\infty} \big[P(\mathcal{T}(X,Y) < \tau) + Q(\mathcal{T}(X,Y) \geq \tau)\big] < 1.$$

Note that strong detection asks for the test statistic to determine with high probability whether $(X, Y)$ is drawn from $P$ or $Q$, while weak detection, similar to weak recovery, only asks for the test statistic to strictly outperform the random guess. Recall that

$$\inf_{\mathcal{T},\tau} \big[ P(\mathcal{T}(X,Y) < \tau) + Q(\mathcal{T}(X,Y) \geq \tau) \big] = 1 - \mathrm{TV}(P,Q).$$

Thus, equivalently, strong detection is possible if and only if

$$\liminf_{p\to\infty} \mathrm{TV}(P,Q) = 1,$$

and weak detection is possible if and only if

$$\liminf_{p\to\infty} \mathrm{TV}(P,Q) > 0.$$

The fundamental question of interest is when $n$ as a function of $(p, k, \sigma^2)$ is such that strong/weak detection is information-theoretically possible. Here, and everywhere in this work, by information-theoretical possibility (respectively, impossibility) of weak/strong detection we refer to the existence (respectively, absence) of a measurable test statistic $\mathcal{T}(X, Y)$ and threshold $\tau$ achieving weak/strong detection.

**1.1. Contributions.** Of fundamental importance is the following sample size:

$$n^* \triangleq \frac{2k \log(p/k)}{\log(1 + k/\sigma^2)}. \tag{3}$$

We establish that $n^*$ is a sharp phase transition point for the recovery of $\beta$ when $k = o(\sqrt{p})$ and the signal to noise ratio $k/\sigma^2$ is growing to infinity with $p$. In particular, for an arbitrarily small but fixed constant $\epsilon > 0$, when $n < (1-\epsilon)n^*$, *weak recovery* is impossible, but when $n > (1+\epsilon)n^*$, *strong recovery* is possible. This implies that the rescaled MMSE undergoes a jump from 1 to 0 at $n^*$ samples up to a small window of size $\epsilon n$. We state this in the following theorem, which summarizes Theorems 2, 3, 4, and 5 from the main body of the paper.

**Theorem** (All-or-nothing phase transition). *Let $\delta \in \big(0, \frac{1}{2}\big)$ and $\epsilon \in (0,1)$ be two arbitrary but fixed constants. Then there exists a constant $C(\delta, \epsilon) > 0$ only depending on $\delta$ and $\epsilon$, such that if $k/\sigma^2 \geq C(\delta,\epsilon)$, then*

- *When $k \leq p^{\frac{1}{2}-\delta}$ and*
$$n < (1-\epsilon)n^*,$$
*both weak recovery of $\beta$ from $(Y, X) \sim P$ and weak detection between $P$ and $Q_{\lambda_0}$ are information-theoretically impossible, where*
$$\lambda_0 = \sqrt{k/\sigma^2 + 1}.$$

- *When $k = o(p)$ and*

$$n > (1 + \epsilon)n^*,$$

*both strong recovery of $\beta$ from $(Y, X) \sim P$ and strong detection between $P$ and $Q_\lambda$ are information-theoretically possible for any $\lambda > 0$.*

Our result establishes as a corollary a conjecture from [19] where the recovery problems is studied under the additional assumptions $\log k = o(\log p)$ and $k/\sigma^2 \to +\infty$ as $p \to +\infty$. In particular, it is predicted in [19] that the sharp all-or-nothing phase transition takes place at the sample size

$$n_{\text{conj}} = \frac{2k \log p}{\log(1 + 2k/\sigma^2)},$$

which coincides with the phase transition point $n^*$ defined in (3) asymptotically, that is, $n^*/n_{\text{conj}} \to 1$ as $p \to +\infty$, under the assumptions $\log k = o(\log p)$ and $k/\sigma^2 \to +\infty$.

Given that the hidden vector is binary-valued, one can naturally wonder whether a similar "all-or-nothing" phase transition for recovery takes place with respect to the Hamming error instead of the squared error used in Definition 1. In Appendix D.3 we present a short argument showing how the "all-or-nothing" phase transition for the mean squared error implies indeed the same sharp phase transition for the Hamming error.

We highlight that the positive part of our main result is achieved by the construction of an, in principle, super-polynomial-time computable optimal estimator (for recovery) and by an, in principle, super-polynomial-time computable optimal test statistic (for detection). While the present work is not focusing on such computational considerations, we encourage the interested reader to [18, 19] for the potential fundamental nature of such a computational bottleneck in the context of estimation.

Observe that when $\sigma$ is sufficiently small, the critical value of $n^*$ can be smaller than $k$. Hence, according to our main result, in this regime for sufficiently small $\epsilon > 0$, $n = \lfloor (1+\epsilon)n^* \rfloor < k$ samples suffice to achieve strong recovery. In the extreme case where $\sigma = 0$, $n^*$ trivializes to zero and we can directly argue that one sample suffices for strong recovery. In fact, for any $\beta \in \{0, 1\}^p$ and $Y_1 = \langle X_1, \beta \rangle$ for $X_1 \sim \mathcal{N}(0, \mathbf{I}_p)$, we can identify $\beta$ as the unique binary-valued solution of $Y_1 = \langle X_1, \beta \rangle$, almost surely with respect to the randomness of $X$ (see e.g., [20]). Note that here the binary assumption on $\beta$ is crucial. In contrast, if $\beta_i$ were real-valued such that either $\beta_i = 0$ or $|\beta_i| \geq \beta_{\min}$ for some $\beta_{\min} > 0$, $n = \Omega(k)$ samples would be necessary for strong recovery for any noise level.[2]

Finally, note that the first part of the above result focuses on $k \leq p^{1/2-\delta}$. It turns out that this is not a technical artifact and $k = o(p^{1/2})$ is needed for $n^*$ to be

[2]This can be seen by considering a genie-aided scenario where the locations of the nonzero entries of $\beta$ are revealed in advance; in this setting, the problem reduces to the standard linear system (defined with respect to an $n \times k$ submatrix of $X$) without sparsity constraints and $n \geq k$ samples are needed to invert this linear system.

the weak detection threshold. More details can be found in Appendix C. The sharp information-theoretic threshold for either detection or recovery is still open when $\Omega(p^{1/2}) \le k \le o(p)$.

**The phase transition role of $n^*$.** According to our main result, the rescaled minimum mean squared error of the problem, $\mathrm{MMSE}/\mathrm{MSE}_0$, exhibits a step behavior asymptotically. Loosely speaking, when $n < n^*$ it equals to one and when $n > n^*$ it equals to zero. We next intuitively explain why such a step behavior for sparse high dimensional regression occurs at $n^*$, using ideas related to *the area theorem*. The area theorem has been used in the channel coding literature to study the MAP decoding threshold [30] and the capacity-achieving codes [27]. The approach described below is similar to the one used previously for linear regression [40].

First let us observe that $n^*$ is asymptotically equal to the *ratio* of entropy $H(\beta) = \log \binom{p}{k}$ and Gaussian channel capacity $\frac{1}{2}\log(1+k/\sigma^2)$. We explore this relationship in the following way. Let

$$I_n \triangleq I(Y_1^n; X, \beta)$$

denote the mutual information between $\beta$ and $(Y_1^n; X)$ with a total of $n$ linear measurements and let

$$\mathrm{MMSE}_n \triangleq \mathbb{E}\big[\|\beta - \mathbb{E}[\beta \mid X, Y_1^n]\|^2\big]$$

denote the corresponding minimum MSE. Using the chain rule for the mutual information and the fact that mutual information in the Gaussian channel under a second moment constraint is maximized by the Gaussian input distribution, it follows that the increment of mutual information satisfies

$$I_{n+1} - I_n \le \frac{1}{2}\log\big(1 + \mathrm{MMSE}_n/\sigma^2\big).$$

See for example the second part of Lemma 15 in [40]. In particular, all the increments are between zero and $\frac{1}{2}\log(1+k/\sigma^2)$ and by telescopic summation for every $n$:

$$I_n \le \frac{n}{2}\log(1+k/\sigma^2), \tag{4}$$

with equality only if for all $m < n$, $\mathrm{MMSE}_m = k$. This is illustrated in Figure 1, where we plot $n$ against $I_{n+1} - I_n$.

Suppose now that we have established that strong recovery is achieved with $n = (1+o(1))n^*$ samples. Then strong recovery and standard identities connecting mutual information and entropy implies that $I_n = (1-o(1))H(\beta)$, and thus

$$I_n = (1-o(1))\frac{n^*}{2}\log(1+k/\sigma^2).$$

In conjunction with (4) and the upper bound on the mutual information increment in terms of the MMSE, this implies that $\mathrm{MMSE}_n = (1-o(1))k$ for $m = (1-o(1))n^*$. In other words, a strong recovery just above the critical threshold $n^*$ implies that weak recovery just below the threshold is impossible.

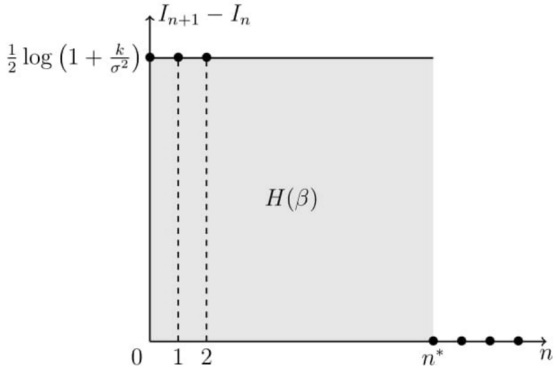


Figure 1. The phase transition diagram in Gaussian sparse linear regression. The $y$-axis is the increment of mutual information with one additional measurement. The area of gray region equals the entropy $H(\beta) \sim k \log(p/k)$.

**1.2. Comparison with related work.** The information-theoretic limits of high-dimensional sparse linear regression have been studied extensively and there is a vast literature of multiple decades of research. In this paper, we establish conditions for the all-or-nothing phenomenon for the model under Gaussian assumptions on $X, W$ and binary assumptions on the entries of $\beta^*$.

For the purposes of comparison, we note that multiple papers in the literature have studied the model in discussion under slightly different contexts compared to ours. For example, some of them have studied the model under the minimax setting instead of our Bayesian setting with a uniform prior; some have offered recovery guarantees which hold with high probability rather than in expectation, or hold under the $\ell_0$ error (Hamming distance) instead of the squared error. While we mention how the most relevant results compare to ours below, we point interested readers to Appendix D for a detailed discussion on how many of these different assumptions/notions of recovery compare with each other.

**Information-theoretic negative results for weak/strong recovery.** For the impossibility direction, previous work [1, Theorem 5.2] has established that as $p \to \infty$, achieving Hamming distortion $\mathbb{E}\big[\|\widehat{\beta} - \beta\|_0\big] \le d$ for any $d \in [0, k]$ is information-theoretically impossible if

$$n \le 2p \frac{h_2(k/p) - h_2(d/p)}{\log(1 + k/\sigma^2)},$$

where

$$h_2(\alpha) = -\alpha \log \alpha - (1-\alpha)\log(1-\alpha) \quad \text{for } \alpha \in [0,1]$$

is the binary entropy function. This converse result is proved via a simple rate-distortion argument (see, e.g., [48] for an exposition). In particular, given any

estimator $\widehat{\beta}(X,Y)$ with $\mathbb{E}\big[\|\widehat{\beta}-\beta\|_0\big] \le d$, we have

$$p\big(h_2(k/p)-h_2(d/p)\big) \le \inf_{\widetilde{\beta}:\mathbb{E}[\|\widetilde{\beta}-\beta\|_0]\le d} I(\widetilde{\beta};\beta)$$
$$\le I(\widehat{\beta};\beta) \le I(X,Y;\beta) \le \frac{n}{2}\log(1+k/\sigma^2).$$

Assuming $k = o(p)$, this result implies that under the Hamming distortion, if $n \le (1-o(1))n^*$, *strong* recovery, that is $d = o(k)$, is information-theoretically impossible and if $n = o(n^*)$, *weak* recovery, that is $d \le (1-\epsilon)k$ for an arbitrary $\epsilon \in (0,1)$, is impossible.

More recent work [42, Corollary 2] further quantified the fraction of support that can be recovered when $n < (1-\epsilon)n^*$ for some fixed constant $\epsilon > 0$. Specifically with $k = o(p)$ and any scaling of $k/\sigma^2$, if $n < (1-\epsilon)n^*$, then the fraction of the support of $\beta$ that can be recovered correctly is at most $1-\epsilon$ with high probability; thus strong recovery is impossible under Hamming distortion. We point interested readers to Appendix D for more details on how the different notions of strong/weak recovery in probability and the corresponding notion of strong/weak recovery in expectation are related.

Restricting to the Maximum Likelihood Estimator (MLE) performance of the problem, it is shown in [19] that under significantly small sparsity

$$k = O\big(\exp(\sqrt{\log p})\big) \quad \text{and} \quad k/\sigma^2 \to +\infty \quad \text{if } n \le (1-\epsilon)n^*,$$

the MLE not only fails to achieve strong recovery, but also fails to weakly recover the vector, that is recover correctly any positive constant fraction of the support.

Our result (Theorem 3) establishes that the MLE performance is fundamental. Furthermore, in view of relation between MSE and Hamming distortion in Appendix D.3, our result improves upon the negative results in the literature by identifying a sharp threshold for weak recovery, showing that if $k = o(\sqrt{p})$, $k/\sigma^2 \to \infty$, and $n \le (1-\epsilon)n^*$, then *weak* recovery is information-theoretically impossible by any estimator $\widehat{\beta}(Y,X)$. In other words, no constant fraction of the support is recoverable under these assumptions. Note that while our result holds for the setting of binary vectors, the impossibility of weak recovery in this setting implies the impossibility of weak recovery, in a minimax sense, over any class of vectors that contains the set of all $k$-spare binary vectors. For example, our result implies that weak recovery is impossible in the setting where the nonzero values are bounded from below in magnitude (by a number less than or equal to one) but otherwise arbitrary.

**Information-theoretic positive results for weak/strong recovery.** In the positive direction, previous work [2, Theorem 1.5] shows that when $k = o(p)$, $k/\sigma^2 = \Theta(1)$, and $n > C_{k/\sigma^2} k\log(p-k)$ for some $C_{k/\sigma^2}$, it is information theoretically possible to weakly recover the hidden vector. Albeit very similar to our results, our positive

result (Theorem 4) identifies the explicit value of $C_{k/\sigma^2}$ for which both weak and strong recovery are possible, that is

$$C_{k/\sigma^2} = 2/\log(1 + k/\sigma^2)$$

for which $C_{k/\sigma^2} k \log(p/k) = n^*$.

In [19] it is shown that when $k = O(\exp(\sqrt{\log p}))$ and $k/\sigma^2 \to +\infty$ then if $n \geq (1+\epsilon)n^*$ for some fixed $\epsilon > 0$, *strong* recovery is achieved by the MLE of the problem. We improve upon this result with Theorem 4 by showing that when $n \geq (1+\epsilon)n^*$ for some fixed $\epsilon > 0$ and any $k = o(p)$, there exists a constant

$$C = C(\epsilon) > 0$$

such that if $k/\sigma^2 \geq C$, then the MLE achieves *strong* recovery. In particular, we significantly relax the assumption from [19] by showing that MLE achieves *strong* recovery with $(1+\epsilon)n^*$ samples for (1) any sparsity level $k$ up to $k = o(p)$ and (2) finite but large values of signal to noise ratio.

**Exact asymptotic characterization of MMSE for linear sparsity.** For both weak and strong recovery, the central object of interest is the MMSE

$$\mathbb{E}\big[\|\beta - \mathbb{E}[\beta \mid X, Y]\|^2\big]$$

and its asymptotic behavior. While the asymptotic behavior of the MMSE remains a challenging open problem when $k = o(p)$, it has been accurately understood when $k = \Theta(p)$ and $k/\sigma^2 = \Theta(1)$.

To be more specific, consider the asymptotic regime, where $k = \varepsilon p$, $\sigma^2 = k/\gamma$, and $n = \delta p$, for fixed positive constants $\varepsilon, \gamma, \delta$ as $p \to +\infty$. The asymptotic minimum mean-square error (MMSE) can be characterized explicitly in terms of $(\varepsilon, \gamma, \delta)$. This characterization was first obtained heuristically using the replica method from statistical physics [21,43] and later proven rigorously [9,40]. More specifically, for fixed $(\varepsilon, \gamma)$, let the asymptotic MMSE as a function of $\delta$ be defined by

$$\mathcal{M}_{\varepsilon,\gamma}(\delta) = \lim_{p\to\infty} \frac{\mathbb{E}\big[\|\beta - \mathbb{E}\big[\beta \mid X, Y\big]\|^2\big]}{\mathbb{E}\big[\|\beta - \mathbb{E}[\beta]\|^2\big]}.$$

The results in [9, 40] lead to an explicit formula for $\mathcal{M}_{\varepsilon,\gamma}(\delta)$. Furthermore, they show that for $\varepsilon \in (0,1)$ and all sufficiently large $\gamma \in (0,\infty)$, $\mathcal{M}_{\varepsilon,\gamma}(\delta)$ has a jump discontinuity as a function of $\delta$. The location of this discontinuity, denoted by $\delta^* = \delta^*(\varepsilon,\gamma)$, occurs at a value that is strictly greater than the threshold $n^*/p$.

Furthermore, at the discontinuity, the MMSE transitions from a value that is strictly less than the MMSE without any observations to a value that is strictly positive, i.e.,

$$\mathcal{M}_{\varepsilon,\gamma}(0) > \lim_{\delta\uparrow\delta^*} \mathcal{M}_{\varepsilon,\gamma}(\delta) > \lim_{\delta\downarrow\delta^*} \mathcal{M}_{\varepsilon,\gamma}(\delta) > 0.$$

To compare these formulas to the sub-linear sparsity studied in this paper, one can consider the limiting behavior of $\mathcal{M}_{\varepsilon,\gamma}(\delta)$ as $\varepsilon$ decreases to zero. Note that the comparison is qualitative, since in the work by [9, 40] the coefficients of $\beta$ are generated i.i.d. according to a Bernoulli$(k/p)$ distribution, while in this paper we consider $\beta$ to be chosen according to a uniform prior over the space of binary $k$-sparse vectors. Nevertheless, it can be verified that $\mathcal{M}_{\varepsilon,\gamma}(\delta)$ converges indeed to a step zero-one function as $\varepsilon \to 0$ and the jump discontinuity transfers to the critical value $n^*/p$ which makes the behavior consistent with the results in this paper. However, an important difference is that the results in this paper are derived directly under the scaling regime $k = o(p)$ whereas the derivation described above requires one to first take the asymptotic limit $p \to \infty$ for fixed $(\epsilon, \gamma)$ and then take $\epsilon \to 0$. Since the limits cannot interchange in any obvious way, the results in this paper cannot be derived as a consequence of the rigorous results in [9, 40]. Finally, it should be mentioned that taking the limit $\epsilon \to 0$ for the replica prediction suggests the step behavior for all values of signal-to-noise ratio $\gamma$ (see Figure 2). In this paper, the step behavior is rigorously proven in the high signal-to-noise ratio regime. The proof of the step behavior when the signal-to-noise ratio is low remains an open problem.

**Sparse superposition codes.** Constructing an algorithm for recovering a binary $k$-sparse $\beta$ from $(Y = X\beta + W, X)$ has received a lot of attention from a coding theory point of view. The reason is that such recovery corresponds naturally to a code for the memoryless additive Gaussian white noise (AWGN) channel with signal-to-noise ratio equal to $k/\sigma^2$. Specifically in this context achieving strong recovery of a uniformly chosen binary $k$-sparse $\beta$ with $(1+\epsilon)n^*$ samples, for arbitrary $\epsilon > 0$, corresponds exactly to capacity-achieving encoding-decoding mechanism of $\binom{p}{k} \sim (pe/k)^k$ messages through a AWGN channel. A recent line of work has analyzed a similar mechanism where $(p/k)^k$ messages are encoded through $k$-block-sparse vectors; that is the vector $\beta$ is designed to have at most one non-zero value in each $k$ block of entries indexed by

$$i\lfloor p/k \rfloor, i\lfloor p/k \rfloor + 1, \ldots, (i+1)\lfloor p/k \rfloor - 1 \quad \text{for } i = 0, 1, 2, \ldots, k-1.$$

It has been shown that by using various polynomial-time decoding mechanisms, such as adaptive successive decoding [25,26], a soft-decision iterative decoder [10,15] and finally Approximate Message Passing techniques [7, 8, 41], one can strongly recover the hidden $k$-block-sparse vector with $(1+\epsilon)n^*$ samples and achieve capacity. Their techniques are tailored to work for any $k = p^{1-c}$ with $c \in (0, 1)$ and also require the vector to have carefully chosen non-zero entries, that is the hidden vector is not assumed to simply be binary. In this work Theorem 4 establishes that under the simple assumption on $\beta$ being binary and arbitrarily (not block) $k$-sparse it suffices to make strong recovery possible with $(1+\epsilon)n^*$ samples when $k = o(p)$. Nevertheless, our decoding mechanism requires a search over the space of $k$-sparse binary vectors and therefore is not in principle polynomial-time. The design of a polynomial-time

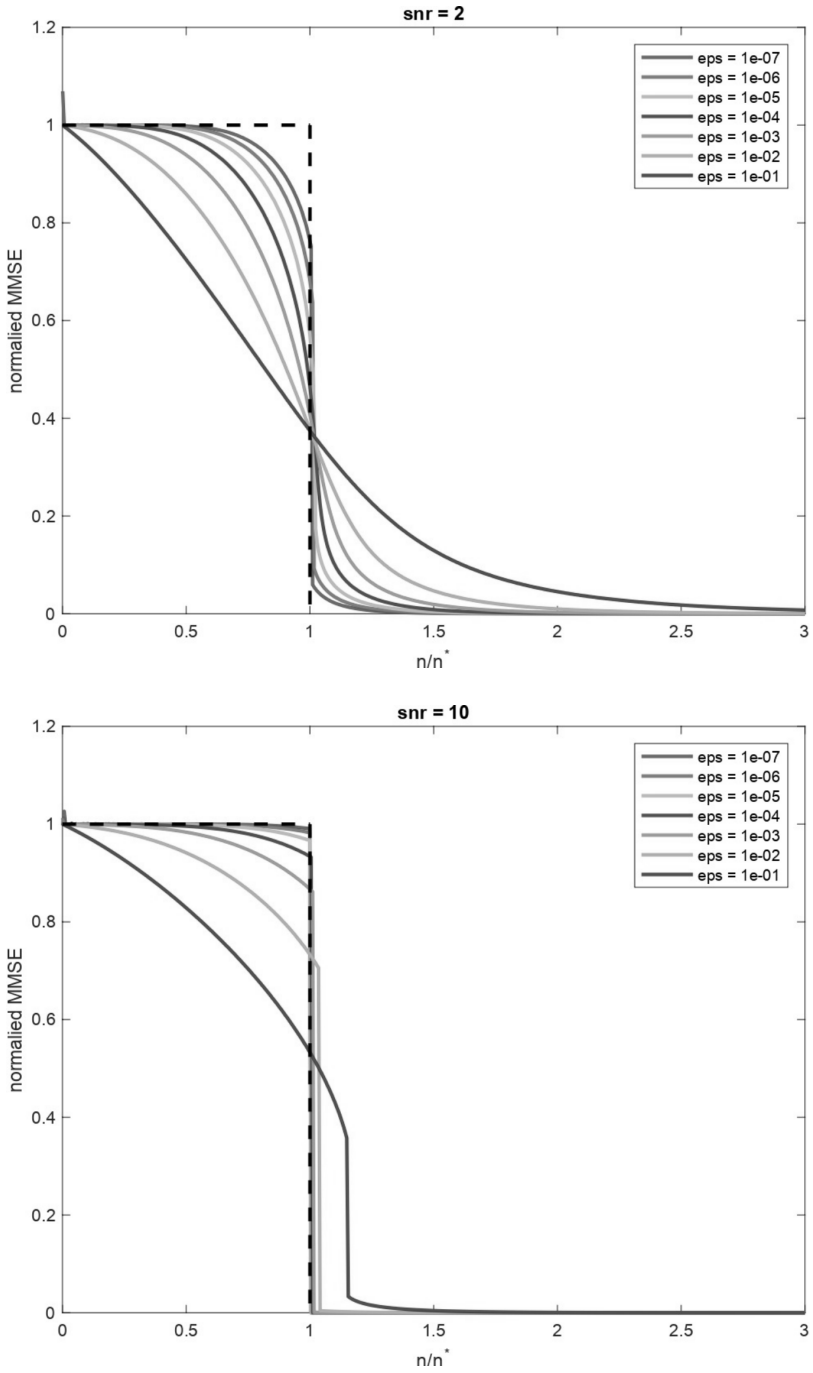


Figure 2. The limit of the replica-symmetric predicted MMSE $\mathcal{M}_{\varepsilon,\gamma}(\cdot)$ as $\epsilon \to 0$ for signal to noise ratio (snr) $\gamma$ equal to 2 (top curve) and equal to 10 (bottom curve).

recovery algorithm for this task and $(1+\epsilon)n^*$ samples remains largely an open problem (see [19]).

**Information-theoretic limits up to constant factors for exact recovery.** Although exact recovery is not our focus, we briefly mention some of the rich literature on the information-theoretic limits for the exact recovery of $\beta$, i.e., $\mathbb{P}\{\widehat{\beta} = \beta\} \to 1$ as $p \to \infty$ (see, e.g., [17,33,36,45,47] and the references therein). Note that if exact recovery is information-theoretically possible, then weak and strong recovery are also information-theoretically possible. As a consequence, the sample sizes required to be achieve exact recovery are in principle no smaller than $n^*$.

Specifically, it has been shown in [45, Theorem 1] that the maximum likelihood estimator achieves exact recovery if

$$n \geq \Omega\left( \log \binom{p-k}{k} + \sigma^2 \log(p-k) \right) \quad \text{and} \quad n - k \to +\infty.$$

Conversely,

$$n > \max\left\{ f_1(p,k), \dots, f_k(p,k), k \right\}$$

is shown in [47, Theorem 1] to be necessary for exact recovery, where

$$f_m(p,k) = 2\frac{\log\binom{p-k+m}{m} - 1}{\log\left(1 + \frac{m(p-k)}{p-k+m}/\sigma^2\right)}.$$

In the special regime where $k$ and $\sigma$ are fixed constants, it has been shown in [24, Theorem 1] that exact recovery is information-theoretically possible if and only if $n \geq (1+o(1))n^*$. Notice that this result achieves exact recovery for approximately $n^*$ sample size, but in this case of constant $k$ it can be easily seen that the information-theoretic limits of exact and strong recovery coincide.

Computationally, it has been shown in [46, Section IV-B] that LASSO achieves exact recovery in polynomial-time if $n \geq 2k\log(p-k)$. More recently, it is shown in [33, Theorem 3.2, Corollary 3.2] that exact recovery can be achieved in polynomial-time, provided that $k = o(p)$, $\sigma \geq \sqrt{3}$, and $n \geq \Omega\big(k \log \frac{ep}{k} + \sigma^2 \log p\big)$.

**Large noise regime.** With the exception of the positive result for strong detection (Theorem 5) the analysis in this paper requires the assumption $k/\sigma^2 \geq C$ for some sufficiently large constant $C > 0$. Hence, our results include the case where the signal-to-noise ratio $k/\sigma^2$ is diverging and also the case where it is fixed, provided that it is sufficiently large.

For comparison, we now make some remarks concerning the large-noise regime $k/\sigma^2 = o(1)$. The result in [33] addresses a more general setting than the one in our paper and specifies a certain initial estimator in the first step. If this initial estimator is replaced by the zero estimator, then their procedure reduces to the classical matched

filter followed by thresholding. However, it is unclear how Theorem III.2 of [33] directly yields the tight sufficient condition for strong recovery. For completeness, we provide a self-contained proof in the arXiv version of the paper, showing that the matched filter followed by thresholding can achieve strong recovery if $k = o(p)$, $k = o(\sigma^2)$, and $n \geq (1+\epsilon)n^*$. Our proof is based on adapting the arguments given in [11]. We note that a related argument shows that the condition $n \geq (1+\varepsilon)n^*$ is also sufficient for strong detection.

In summary, the positive results for both strong detection and strong recovery which appear in the present work also hold in the large-noise regime $k/\sigma^2 = o(1)$, and moreover these results can be achieved using computationally efficient methods. Whether the corresponding negative results established in this work also hold in the large noise setting $k/\sigma^2 = o(1)$ remains, to the best of our knowledge, still unknown. Note that the information theoretic arguments from [42, Corollary 2], which are discussed above, establish that $n > (1-\epsilon)n^*$ is a necessary condition for *strong* recovery, regardless of the noise level. However, the corresponding negative result for *weak* recovery becomes $n = o(n^*)$ and thus in the large noise regime $k/\sigma^2 = o(1)$ it appears that there is still a gap between the conditions for weak and strong recovery. Note that if $\beta$ were assumed to be i.i.d. Bern$(k/p)$, then one could use the genie-aided argument (see e.g., [39] and [33, Theorem 2.1]) where all the entries of $\beta$ but one are revealed, showing that weak recovery is impossible if $n \leq (1-\epsilon)n^*$ when $k = o(p)$ and $k = o(\sigma^2)$. However, in our setting with the fixed sparsity, the value of any entry $\beta_i$ is known once all the other entries of $\beta$ have been revealed and hence this genie-aided argument does not directly apply.

**1.3. Proof techniques.** In this section, we give an overview of our proof techniques. Given two probability distributions $P, Q$ with $P$ absolutely continuous to $Q$ and any convex function $f$ such that $f(1) = 0$, the $f$-divergence of $Q$ from $P$ is given by

$$D_f(P\|Q) \triangleq \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right].$$

Three choices of $f$ are of particular interest (see [35, Section 6] for details):

- The *Total variation distance* $\mathrm{TV}(P,Q)$: $f(x) = |x-1|/2$;
- The *Kullback–Leibler divergence* (a.k.a. relative entropy) $D(P\|Q)$: $f(x) = x\log x$;
- The $\chi^2$*-divergence* $\chi^2(P\|Q)$: $f(x) = (x-1)^2$.

Note that the $\chi^2$-divergence $\chi^2(P\|Q)$ is equal to the variance of the Radon–Nikodym derivative (likelihood ratio) $dP/dQ$ under $Q$ and hence

$$\chi^2(P\|Q) + 1 = \mathbb{E}_Q\left[\left(\frac{dP}{dQ}\right)^2\right] = \mathbb{E}_P\left[\frac{dP}{dQ}\right].$$

A key to our proof is the following chain of inequalities:

$$\mathrm{TV}(P, Q) \le \sqrt{2D(P\|Q)} \le \sqrt{2\log\left(\chi^2(P\|Q)+1\right)}, \tag{5}$$

where the first inequality is simply Pinsker's inequality, and the second inequality holds by Jensen's inequality:

$$D(P\|Q) = \mathbb{E}_P\left[\log\frac{dP}{dQ}\right] \le \log\left(\mathbb{E}_P\left[\frac{dP}{dQ}\right]\right) = \log\left(\chi^2(P\|Q)+1\right). \tag{6}$$

Recall that to show the weak detection between $P$ and $Q_\lambda$ is impossible, it is equivalent to proving that $\mathrm{TV}(P, Q_\lambda) = o(1)$. In view of (5) there is a natural strategy towards proving it: it suffices to prove that $\chi^2(P, Q_\lambda) = o(1)$, which amounts to showing the second moment

$$\mathbb{E}_Q\left[(dP/dQ_\lambda)^2\right] = 1 + o(1).$$

We prove that indeed if $n \le (1-o(1))n^*/2$ and $\lambda$ is appropriately chosen, then this second moment is indeed $1+o(1)$ (Theorem 1); however, if $n > n^*/2$, then it blows up to infinity. This is because even if potentially $\mathrm{TV}(P, Q_\lambda) = o(1)$, rare events can cause the second moment to explode and in particular (5) is far from being tight.

We are able to circumvent this difficulty by computing the second moment conditioned on an event $\mathcal{E}$, which rules out the catastrophic rare ones. In particular, we introduce the following conditioned planted model.

**Definition 3** (Conditioned planted model). Given a subset $\mathcal{E} \subset \mathbb{R}^{n\times p} \times \mathbb{R}^p$, define the conditioned planted model

$$P_{\mathcal{E}}(X, Y) = \frac{\mathbb{E}_\beta[P(X, Y \mid \beta)\mathbf{1}_{\{\mathcal{E}\}}(X, \beta)]}{\mathbb{P}\{\mathcal{E}\}}. \tag{7}$$

Using this notation we can write

$$P(X, Y) = (1-\varepsilon)P_{\mathcal{E}}(X, Y) + \varepsilon P_{\mathcal{E}^c}(X, Y),$$

where $\mathcal{E}^c$ denotes the complement of $\mathcal{E}$ and $\varepsilon = \mathbb{P}\{(X, \beta) \in \mathcal{E}^c\}$. By Jensen's inequality and the convexity of KL-divergence,

$$D(P\|Q_\lambda) \le (1-\varepsilon)D(P_{\mathcal{E}}\|Q_\lambda) + \varepsilon D(P_{\mathcal{E}^c}\|Q_\lambda). \tag{8}$$

Under an appropriately chosen $\mathcal{E}$, and $\lambda > 0$, our main impossibility of detection result (Theorem 2) shows that if $n \le (1+o(1))n^*$, then

$$\mathbb{E}_{Q_\lambda}\left[(dP_{\mathcal{E}}/dQ_\lambda)^2\right] = 1 + o(1),$$

or equivalently,

$$\chi^2(P_{\mathcal{E}}\|Q_\lambda) = o(1),$$

which immediately implies that $D(P_{\mathcal{E}}\|Q_\lambda) = o(1)$ and $\mathrm{TV}(P_{\mathcal{E}}, Q_\lambda) = o(1)$.

Finally, we argue that $\varepsilon$ converges to 0 sufficiently fast so that according to (8),

$$\mathrm{TV}(P, Q_\lambda) \le \mathrm{TV}(P_{\mathcal{E}}, Q) + o(1) = o(1)$$

and

$$D(P\|Q_\lambda) \le D(P_{\mathcal{E}}\|Q_\lambda) + o(1) = o(1).$$

We remark that this (conditional) second moment method for providing detection lower bound has been used in many high-dimensional inference problems (see e.g., [5, 6, 13, 22, 23, 32, 34, 44, 48] and references therein).

To further show weak recovery is impossible in the regime for sample size $n < n^*$ (Theorem 3), we establish a lower bound of MSE in terms of $D(P\|Q_\lambda)$ (Lemma 2) which implies that the minimum MSE needs to be $(1-o(1))k$ if $D(P\|Q_\lambda) = o(n)$. The key underpinning our lower bound proof is the area theorem [27, 30].

**1.4. Notation and organization.** Denote the identity matrix by $\mathbf{I}$. We let $\|X\|$ denote the spectral norm of a matrix $X$ and $\|x\|$ denote the $\ell_2$ norm of a vector $x$. For any positive integer $n$, let $[n] = \{1, \dots, n\}$. For any set $T \subset [n]$, let $|T|$ denote its cardinality and $T^c$ denote its complement. We use standard big $O$ notations, e.g., for any positive sequences $\{a_p\}$ and $\{b_p\}$, $a_p = \Theta(b_p)$ if there is an absolute constant $c > 0$ such that $1/c \le a_p/b_p \le c$; $a_p = \Omega(b_p)$ or $b_p = O(a_p)$ if there exists an absolute constant $c > 0$ such that $a_p/b_p \ge c$. We use standard little $o$ notations, e.g., for any positive sequences $\{a_p\}$ and $\{b_p\}$, $a_p = \omega(b_p)$ or $b_p = o(a_p)$ if $\lim a_p/b_p = +\infty$. We say a sequence of events $\mathcal{E}_p$ indexed by a positive integer $p$ holds with high probability, if the probability of $\mathcal{E}_p$ converges to 1 as $p \to +\infty$. Without further specification, all the asymptotics are taken with respect to $p \to \infty$. All logarithms are natural and we use the convention $0 \log 0 = 0$. For two real numbers $a$ and $b$, we use $a \vee b = \max\{a, b\}$ to denote the larger of $a$ and $b$. For two vectors $u, v$ of the same dimension, we use $\langle u, v\rangle$ denote their inner product. We use $\chi^2_n$ to denote the standard chi-squared distribution with $n$ degrees of freedom. For $n, m, k \in \mathbb{N}$ with $m \le k \le n$ and $m + k \le n$ we denote by $\mathrm{Hyp}(n, m, k)$ the Hypergeometric distribution with parameters $n, m, k$ and probability mass function $p(s) = \binom{m}{s}\binom{n-m}{k-s}/\binom{n}{k}, s \in [0, m] \cap \mathbb{Z}$.

The remainder of the paper is organized as follows. Section 2 presents the main results without proofs. Section 3 and Section 4 prove the negative results for detection and recovery, respectively. Section 5 proves the positive results for detection and recovery. We conclude the paper in Section 6, mentioning a few open problems. Auxiliary lemmata and miscellaneous details are left to appendices.

## 2. Main results

In this section we present our main results. The proofs are deferred to the following sections.

**2.1. Impossibility of weak detection with $n < n^*$.** Our first impossibility detection result is based on a direct calculation of the second moment between the planted model $P$ and the null model $Q_\lambda$. Specifically, we are able to show that weak detection between the two models is impossible, if $n \le (1-\alpha)n^*/2$ for some $\alpha = o_p(1)$ and $\lambda = \sqrt{k/\sigma^2+1}$.

**Theorem 1.** *Suppose $k \le p^{\frac{1}{2}-\delta}$ for a fixed constant $\delta > 0$ and $k/\sigma^2 \ge C$ for a sufficiently large constant $C$ only depending on $\delta$. If*

$$n \le \frac{1}{2}\left(1 - \frac{\log\log(p/k)}{\log(p/k)}\right)n^*, \tag{9}$$

*then for $\lambda_0 = \sqrt{k/\sigma^2+1}$, it holds that*

$$\chi^2(P\|Q_{\lambda_0}) = o(1)$$

*Furthermore, $D(P\|Q_{\lambda_0}) = o(1)$ and $\mathrm{TV}(P, Q_{\lambda_0}) = o(1)$.*

The complete proof of the above theorem can be found in Section 3.1. Nevertheless, let us provide here a short proof sketch. Using an explicit calculation, we first find that for any $\lambda > \sqrt{k/\sigma^2+1/2}$,

$$\chi^2(P\|Q_\lambda) = \lambda^{2n}\mathbb{E}_{S\sim\mathrm{Hyp}(p,k,k)}\left[\left(2\lambda^2 - 1 - \frac{k+S}{\sigma^2}\right)^{-n/2}\left(1+\frac{k-S}{\sigma^2}\right)^{-n/2}\right] - 1,$$

where $S = \langle\beta,\beta'\rangle$ is the overlap between two independent copies $\beta,\beta'$ and follows a hypergeometric distribution with parameters $(p,k,k)$. Plugging in

$$\lambda = \lambda_0 = \sqrt{k/\sigma^2+1},$$

we get that

$$\chi^2(P\|Q_{\lambda_0}) = \mathbb{E}_{S\sim\mathrm{Hyp}(p,k,k)}\left[\left(1-\frac{S}{k+\sigma^2}\right)^{-n}\right] - 1.$$

Using this we show that if $n \le (1+o(1))n^*/2$, then $\chi^2(P\|Q_{\lambda_0})$ is indeed $o(1)$, implying by (5) the impossibility result. However, if $n > n^*/2$, then this $\chi^2$-divergence can be proven to blow up to infinity, rendering the method based on (5) uninformative in this regime. To see this, by considering the event $S = k$ which happens with probability $1/\binom{p}{k}$, we get that

$$\chi^2(P\|Q_{\lambda_0}) \ge \frac{1}{\binom{p}{k}}\left[\left(1-\frac{k}{k+\sigma^2}\right)^{-n}\right] - 1 = \exp\left(n\log\left(1+\frac{k}{\sigma^2}\right) - \log\binom{p}{k}\right) - 1. \tag{10}$$

Recall that $n^*$ is asymptotically equal to $2\log\binom{p}{k}/\log\left(1+k/\sigma^2\right)$. Hence, if $n \ge n^*(1+\epsilon)/2$ for some constant $\epsilon > 0$, then $\chi^2(P\|Q_{\lambda_0}) \to +\infty$.

To be able to obtain tighter results and go all the way to $n^*$ sample size, we resort to a *conditional* second moment method as explained in the proof techniques. Specifically we show that weak detection is impossible for any $n \le (1-\alpha)n^*$, for some $\alpha > 0$ that can be made to be arbitrarily small by increasing $k/\sigma^2$ and $p/k$. In particular, this improves on the direct calculation of the $\chi^2$ distance by a multiplicative factor of 2 and shows that $n^*$ is a sharp information theoretic threshold for weak detection between the planted model $P$ and the null model $Q_{\lambda_0}$.

Before formally stating our main theorem, we specify the conditioning event $\mathcal{E}_{\gamma,\tau}$ which will be shown to hold with high probability in Lemma 8 under appropriate choices of $\gamma$ and $\tau$.

**Definition 4** (Conditioning event). Given $\gamma \ge 0$ and $\tau \in [0,k]$, define an event $\mathcal{E}_{\gamma,\tau} \subset \mathbb{R}^{n\times p} \times \mathbb{R}^p$ as

$$\mathcal{E}_{\gamma,\tau} = \Big\{(X,\beta) : \frac{\|X(\beta+\beta')\|^2}{\mathbb{E}[\|X(\beta+\beta')\|^2]} \le 2+\gamma, \\ \forall \beta' \in \{0,1\}^p \text{ with } \|\beta'\|_0 = k \text{ and } \langle \beta', \beta\rangle \ge \tau\Big\}. \tag{11}$$

To understand the value of $\gamma, \tau$ in the definition of this event, notice that for each $\beta, \beta'$, from the definition of $X$, we have $X(\beta+\beta') \sim \mathcal{N}(0, 2(k+s)\mathbf{I}_n)$, for $s := \langle \beta', \beta\rangle$, and therefore,

$$\frac{\|X(\beta+\beta')\|^2}{2(k+s)} \sim \chi^2_n.$$

Thus, by the concentration inequality of chi-squared distributions, the random variable

$$\frac{\|X(\beta+\beta')\|^2}{\mathbb{E}[\|X(\beta+\beta')\|^2]}$$

is expected to concentrate around 1 and thus is likely to be smaller than $2+\gamma$ for a relatively large $\gamma$. The parameter $\tau$ quantifies the set of $k$-sparse $\beta'$ for which we expect this relation to hold. Notice that $\langle \beta', \beta\rangle \ge \tau$ is equivalent with the Hamming-distance between $\beta$ and $\beta'$ to be equal to $2(k-\tau)$.

Next, we explain the intuition behind our choice of conditioning event $\mathcal{E}_{\gamma,\tau}$. Recall that in view of (10), $\chi^2(P \| Q_{\lambda_0})$ blows up to infinity when the overlap $\langle \beta, \beta'\rangle$ is equal to $k$. In fact, when the overlap $\langle \beta, \beta'\rangle = k$, $\|X(\beta+\beta')\|^2$ can be enormously large, causing $\chi^2(P \| Q_{\lambda_0})$ to explode. We rule out this catastrophic event by conditioning on $\mathcal{E}_{\gamma,\tau}$ which bounds from above $\|X(\beta+\beta')\|^2$ when the overlap $\langle \beta, \beta'\rangle$ is large; see (32) for the key step of bounding from above $\|X(\beta+\beta')\|^2$.

As a result, we are able to prove that the $\chi^2$-divergence between the conditional planted model $P_{\mathcal{E}_{\gamma,\tau}}$ and the null model $Q_{\lambda_0}$ for $\lambda_0 = \sqrt{k/\sigma^2+1}$ is $o(1)$, which implies the following general impossibility of detection result.

**Theorem 2.** *Suppose $k \le p^{\frac{1}{2}-\delta}$ for an arbitrarily small fixed constant $\delta \in \left(0, \frac{1}{2}\right)$ and $k/\sigma^2 \ge C$ for a sufficiently large constant $C$ only depending on $\delta$. Assume $n \le (1-\alpha)n^*$ for $\alpha \in (0, 1/2]$ such that*

$$\alpha = \frac{8}{\log(1+k/\sigma^2)} \vee \frac{32 \log\log(p/k)}{\log(p/k)}. \tag{12}$$

*Set*

$$\gamma = \frac{\alpha k \log(p/k)}{n} \quad \text{and} \quad \tau = k\left(1 - \frac{1}{\log^2(1+k/\sigma^2)}\right).$$

*Then for $\lambda_0 = \sqrt{k/\sigma^2+1}$,*

$$\chi^2(P_{\mathcal{E}_{\gamma,\tau}} \| Q_{\lambda_0}) = o(1). \tag{13}$$

*Furthermore,*

$$D(P_{\mathcal{E}_{\gamma,\tau}} \| Q_{\lambda_0}) = o(1), \quad \mathrm{TV}(P_{\mathcal{E}_{\gamma,\tau}}, Q_{\lambda_0}) = o(1), \quad \text{and} \quad \mathrm{TV}(P, Q_{\lambda_0}) = o(1).$$

The proof of the theorem can be found in Section 3.2.

**2.2. Impossibility of weak recovery with $n < n^*$.** In this section we present our impossibility of recovery result. We do this using the impossibility of detection result established above. Specifically we first strengthen Theorem 2 and show that under the assumptions of Theorem 2, $D(P \| Q_{\lambda_0}) = o_p(1)$. Notice that this is not needed to conclude impossibility of detection, that is $\mathrm{TV}(P, Q_{\lambda_0}) = o(1)$, but is needed here for establishing the impossibility of recovery result. As a second step, inspired by the celebrated area theorem, we establish (Lemma 2) a lower bound to the minimum MSE in terms of $D(P \| Q_{\lambda_0})$, which is potentially of independent interest. The lemma essentially quantifies the natural idea that if the data $(Y, X)$ drawn from planted model are statistically close to the data $(Y, X)$ drawn from null model then there are limitations on the performance of recovering the hidden vector $\beta$ based on the data $(Y, X)$ from the planted model. Interestingly the lemma itself does not require the hidden vector $\beta$ to be binary or $k$-sparse but only to satisfy $\mathbb{E}[\|\beta\|^2] = k$. Combining the two steps allows us to conclude that the minimum MSE is $k(1 + o_p(1))$; hence the impossibility of weak recovery.

**Theorem 3.** *Suppose $k \le p^{\frac{1}{2}-\delta}$ for an arbitrarily small fixed constant $\delta \in \left(0, \frac{1}{2}\right)$ and $k/\sigma^2 \ge C$ for a sufficiently large constant $C$ only depending on $\delta$. Let $\lambda_0 = \sqrt{k/\sigma^2+1}$. If $n \le (1-\alpha)n^*$ for $\alpha \in (0, 1/2]$ given in* (12)*, then it holds that*

$$D(P \| Q_{\lambda_0}) = o_p(1). \tag{14}$$

*Furthermore, if $n \le \lfloor (1-\alpha)n^* \rfloor - 1$, then for any estimator $\widehat{\beta}$ that is a function of $X$ and $Y$,*

$$\mathrm{MSE}(\widehat{\beta}) = k(1 + o_p(1)). \tag{15}$$

The proof of the above theorem can be found in Section 4.1.

**2.3. Positive result for strong recovery with $n > n^*$.** This subsection and the next one are in the regime where $n \geq (1+\epsilon)n^*$ for some $\epsilon > 0$. In this regime, we establish that both strong recovery and strong detection are possible.

Towards recovering the vector $\beta$, we consider the Maximum Likelihood Estimator (MLE) of $\beta$:

$$\widehat{\beta} = \arg \min_{\beta' \in \{0,1\}^p, \|\beta'\|_0 = k} \|Y - X\beta'\|^2.$$

We show that MLE achieves strong recovery of $\beta$ if $n \geq (1+\epsilon)n^*$ for an arbitrarily small but fixed constant $\epsilon$ whenever $k = o(p)$ and $k/\sigma^2 \geq C(\epsilon)$ for a sufficiently large constant $C(\epsilon) > 0$.

Specifically, we establish the following result.

**Theorem 4.** *Suppose* $\log\log(p/k) \geq 1$. *If*

$$n \geq \left(1 + \frac{4\log\log(p/k)}{\log(p/k)}\right) \frac{2k\log(p/k)}{\log(1 + k/2\sigma^2)}, \tag{16}$$

*then*

$$\mathbb{P}\left\{\|\widehat{\beta} - \beta\|^2 \geq \frac{2k}{\log(p/k)}\right\} \leq \frac{e^2}{\log^2(p/k)(1-e^{-1})}. \tag{17}$$

*Furthermore, if additionally* $k = o(p)$, *then*

$$\frac{1}{k}\mathbb{E}\big[\|\widehat{\beta} - \beta\|^2\big] = o_p(1), \tag{18}$$

*i.e., MLE achieves strong recovery of* $\beta$.

The proof of the above theorem can be found in Section 5.1.

**2.4. Positive result for strong detection with $n > n^*$.** In this subsection we establish that when $n \geq (1+\epsilon)n^*$ for some $\epsilon > 0$ strong detection is possible. To distinguish the planted model $P$ and the null model $Q_\lambda$, we consider the test statistic:

$$\mathcal{T}(X, Y) = \min_{\beta' \in \{0,1\}^p, \|\beta'\|_0 = k} \frac{\|Y - X\beta'\|^2}{\|Y\|^2}.$$

**Theorem 5.** *Suppose* $M = \binom{p}{k} \to +\infty$. *If*

$$n \geq \left(1 + \sqrt{\frac{\log\log M}{\log M}}\right) \frac{2\log M}{\log(1 + k/\sigma^2)}, \tag{19}$$

*then there exist proper choices of* $\tau$ *such that*

$$P(\mathcal{T}(X, Y) \geq \tau) + Q_\lambda(\mathcal{T}(X, Y) \leq \tau) = o(1),$$

*which achieves the strong detection between the planted model* $P$ *and the null model* $Q_\lambda$.

The proof of Theorem 5 can be found in Section 5.2.

## 3. Proof of negative results for detection

**3.1. Proof of Theorem 1.** We start with an explicit computation of the chi-squared divergence $\chi^2(P\|Q_\lambda)$.

**Proposition 1.** *For any* $\lambda > \sqrt{k/\sigma^2 + 1/2}$,

$$\chi^2(P\|Q_\lambda) = \lambda^{2n}\mathbb{E}_{S\sim\mathrm{Hyp}(p,k,k)}\left[\left(2\lambda^2 - 1 - \frac{k+S}{\sigma^2}\right)^{-n/2}\left(1 + \frac{k-S}{\sigma^2}\right)^{-n/2}\right] - 1.$$

*Proof.* Since the marginal distribution of $X$ is the same under the planted and null models, it follows that for any $\beta$,

$$\frac{P(X,Y)}{Q_\lambda(X,Y)} = \frac{P(Y|X)}{Q_\lambda(Y)} = \frac{\mathbb{E}_\beta[P(Y|X,\beta)]}{Q_\lambda(Y)}.$$

Therefore,

$$\left(\frac{P(X,Y)}{Q_\lambda(X,Y)}\right)^2 = \mathbb{E}_{\beta\perp\!\!\!\perp\beta'}\left[\frac{P(Y|X,\beta)P(Y|X,\beta')}{Q_\lambda^2(Y)}\right],$$

where $\beta \perp\!\!\!\perp \beta'$ denote two independent copies. By Fubini's theorem, we have

$$\mathbb{E}_{Q_\lambda}\left[\left(\frac{P}{Q_\lambda}\right)^2\right] = \mathbb{E}_{\beta\perp\!\!\!\perp\beta'}\mathbb{E}_X\mathbb{E}_Y\left[\frac{P(Y|X,\beta)P(Y|X,\beta')}{Q_\lambda^2(Y)}\right], \tag{20}$$

where $X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $Y_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\lambda^2\sigma^2)$.

Since in the planted model, conditional on $(X,\beta)$, $Y \sim \mathcal{N}(X\beta, \sigma^2\mathbf{I}_n)$. It follows that

$$\begin{aligned}\frac{P(Y|X,\beta)}{Q_\lambda(Y)} &= \lambda^n \exp\Big(-\frac{1}{2\sigma^2}\|Y - X\beta\|^2 + \frac{1}{2\lambda^2\sigma^2}\|Y\|^2\Big)\\ &= \lambda^n \exp\Big(-\frac{\lambda^2-1}{2\sigma^2\lambda^2}\|Y\|^2 + \frac{1}{\sigma^2}\langle Y, X\beta\rangle - \frac{1}{2\sigma^2}\|X\beta\|^2\Big).\end{aligned}$$

Hence,

$$\begin{aligned}&\frac{P(Y|X,\beta)P(Y|X,\beta')}{Q_\lambda^2(Y)}\\ &= \lambda^{2n}\exp\Big(-\frac{\lambda^2-1}{\sigma^2\lambda^2}\|Y\|^2 + \frac{1}{\sigma^2}\langle Y, X(\beta+\beta')\rangle - \frac{1}{2\sigma^2}\big(\|X\beta\|^2 + \|X\beta'\|^2\big)\Big)\\ &= \lambda^{2n}\exp\Big(-\frac{\lambda^2-1}{\sigma^2\lambda^2}\Big\|Y - \frac{\lambda^2 X(\beta+\beta')}{2(\lambda^2-1)}\Big\|^2 + \frac{\lambda^2\|X(\beta+\beta')\|^2}{4(\lambda^2-1)\sigma^2}\\ &\quad - \frac{1}{2\sigma^2}\big(\|X\beta\|^2 + \|X\beta'\|^2\big)\Big).\end{aligned}$$

Using the fact that $\mathbb{E}\big[e^{tZ^2}\big] = 1/\big(\sqrt{1-2t\sigma^2}\big)\, e^{\mu^2 t/(1-2t\sigma^2)}$ for $t < 1/2\sigma^2$ and $Z \sim \mathcal{N}(\mu, \sigma^2)$, we get that

$$\mathbb{E}_Y\bigg[\exp\bigg(-\frac{\lambda^2-1}{\sigma^2\lambda^2}\Big\|Y - \frac{\lambda^2 X(\beta+\beta')}{2(\lambda^2-1)}\Big\|^2\bigg)\bigg] = \frac{1}{(2\lambda^2-1)^{n/2}}\exp\bigg(-\frac{\lambda^2\,\|X(\beta+\beta')\|^2}{4(2\lambda^2-1)(\lambda^2-1)\sigma^2}\bigg).$$

Combining the last two displayed equations yields that

$$\mathbb{E}_Y\bigg[\frac{P(Y|X,\beta)P(Y|X,\beta')}{Q_\lambda^2(Y)}\bigg] = \frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}} \cdot \exp\bigg\{\frac{1}{2\sigma^2(2\lambda^2-1)}\Big((1-\lambda^2)\big(\|X\beta\|^2 + \big\|X\beta'\big\|^2\big) + 2\lambda^2\big\langle X\beta, X\beta'\big\rangle\Big)\bigg\}. \tag{21}$$

Let $T = \mathrm{supp}(\beta)$ and $T' = \mathrm{supp}(\beta')$. Let $X_i$ denote the $i$th column of $X$. Define

$$Z_0 = \sum_{i\in T\cap T'} X_i, \quad Z_1 = \sum_{i\in T\setminus T'} X_i, \quad Z_2 = \sum_{i\in T'\setminus T} X_i.$$

Then, conditional on $\beta$ and $\beta'$, $Z_0, Z_1, Z_2$ are mutually independent and

$$Z_0 \sim \mathcal{N}(0, s\mathbf{I}_n), \quad Z_1 \sim \mathcal{N}(0, (k-s)\mathbf{I}_n), \quad Z_2 \sim \mathcal{N}(0, (k-s)\mathbf{I}_n),$$

where $s = |T\cap T'| = \langle\beta, \beta'\rangle$. Moreover, $X\beta, X\beta'$ can be expressed as a function of $Z_0, Z_1, Z_2$ simply by

$$X\beta = Z_0 + Z_1 \quad \text{and} \quad X\beta' = Z_0 + Z_2. \tag{22}$$

Let $Z = [Z_0, Z_1, Z_2]^t \in \mathbb{R}^{3n}$. Using (21) and (22) and elementary algebra we have

$$\mathbb{E}_Y\bigg[\frac{P(Y|X,\beta)P(Y|X,\beta')}{Q_\lambda^2(Y)}\bigg] = \frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}}\exp\{tZ^\top AZ\}, \tag{23}$$

where

$$t = \frac{1}{2\sigma^2(2\lambda^2-1)}, \quad \text{and} \quad A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1-\lambda^2 & \lambda^2 \\ 1 & \lambda^2 & 1-\lambda^2 \end{bmatrix} \otimes \mathbf{I}_n \in \mathbb{R}^{3n\times 3n},$$

where by $A\otimes B$ we refer to the Kronecker product between two matrices $A$ and $B$. Note that $Z$ is a zero-mean Gaussian vector with covariance matrix

$$V = \mathsf{diag}\,\{s, k-s, k-s\} \otimes \mathbf{I}_n.$$

Note that

$$AV = \left( \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1-\lambda^2 & \lambda^2 \\ 1 & \lambda^2 & 1-\lambda^2 \end{bmatrix} \mathsf{diag}\{s, k-s, k-s\} \right) \otimes \mathbf{I}_n.$$

It is straightforward to find that the eigenvalues of $AV$ are 0 of multiplicity $n$, $k+s$ of multiplicity $n$, and $(k-s)(1-2\lambda^2)$ of multiplicity $n$. Thus,

$$\det(\mathbf{I}_{3n} - 2tAV) = \big(1-2t(k+s)\big)^n \big(1-2t(k-s)(1-2\lambda^2)\big)^n. \tag{24}$$

It follows from (23) that

$$\begin{aligned} \mathbb{E}_X \mathbb{E}_Y \left[ \frac{P(Y|X,\beta)P(Y|X,\beta')}{Q_\lambda^2(Y)} \right] &= \frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}} \mathbb{E}_Z \big[ e^{tZ^\top AZ} \big] \\ &= \frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}} \frac{1}{\sqrt{\det(\mathbf{I}_{3n} - 2tAV)}}, \end{aligned} \tag{25}$$

where the last equality holds if $t < 1/2(k+s)$ and follows from the expression of MGF of a quadratic form of normal random variables, see, e.g., [4, Lemma 2].

Combining (24) and (25) yields that if $t = 1/2\sigma^2(2\lambda^2-1) < 1/2(k+s)$, then

$$\begin{aligned} &\mathbb{E}_X \mathbb{E}_Y \left[ \frac{P(Y|X,\beta)P(Y|X,\beta')}{Q_\lambda^2(Y)} \right] \\ &\quad = \frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}} \left(1 - \frac{k+s}{\sigma^2(2\lambda^2-1)}\right)^{-n/2} \left(1 + \frac{k-s}{\sigma^2}\right)^{-n/2} \\ &\quad = \lambda^{2n} \left(2\lambda^2 - 1 - \frac{k+s}{\sigma^2}\right)^{-n/2} \left(1 + \frac{k-s}{\sigma^2}\right)^{-n/2}. \end{aligned}$$

Note that if $2\lambda^2 - 1 > \frac{2k}{\sigma^2}$, then

$$\frac{1}{2\sigma^2(2\lambda^2-1)} < \frac{1}{2(k+s)}$$

for all $0 \le s \le k$. It follows from (20) that if $2\lambda^2 - 1 > 2k/\sigma^2$, then

$$\mathbb{E}_{Q_\lambda} \left[ \left( \frac{P}{Q_\lambda} \right)^2 \right] = \lambda^{2n} \mathbb{E}_{S \sim \mathrm{Hyp}(p,k,k)} \left[ \left( 2\lambda^2 - 1 - \frac{k+S}{\sigma^2} \right)^{-n/2} \left( 1 + \frac{k-S}{\sigma^2} \right)^{-n/2} \right].$$

□

We establish also the following lemma.

**Lemma 1.** *Suppose $k \le p^{\frac{1}{2}-\delta}$ for an arbitrarily small fixed constant $\delta \in \big(0, \frac{1}{2}\big)$ and $k/\sigma^2 \ge C$ for a sufficiently large constant $C$ only depending on $\delta$. If $n$ satisfies condition* (9)*, then*

$$\mathbb{E}_{S \sim \mathrm{Hyp}(k,k,p)} \left[ \left( 1 - \frac{S}{k+\sigma^2} \right)^{-n} \right] = 1 + o_p(1). \tag{26}$$

*Proof.* The lemma readily follows by combining Lemma 6 and Lemma 7 with

$$\alpha = \frac{\log\log(p/k)}{\log(p/k)} \quad \text{and} \quad c = p^{-1/2-\delta}. \qquad \square$$

*Proof of Theorem 1.* Using Proposition 1 for $\lambda = \lambda_0$ satisfying $\lambda_0^2 = k/\sigma^2 + 1$ we have

$$\chi^2(P\|Q_{\lambda_0}) = \mathbb{E}_{S\sim\mathrm{Hyp}(p,k,k)}\left[\left(1 - \frac{S}{k+\sigma^2}\right)^{-n}\right] - 1.$$

Using now Lemma 1 we have $\chi^2(P\|Q_{\lambda_0}) = o(1)$. The chain of inequalities (5) concludes the proof of Theorem 1. $\square$

### 3.2. Proof of Theorem 2.

*Proof.* For notational simplicity we denote in this proof the probability measure $Q_{\lambda_0}$ simply by $Q$ and the event $\mathcal{E}_{\gamma,\tau}$ by $\mathcal{E}$.

We first show that (13) implies $D(P_{\mathcal{E}}\|Q) = o(1)$, $\mathrm{TV}(P_{\mathcal{E}}, Q) = o(1)$, and $\mathrm{TV}(P, Q) = o(1)$.

It follows from (5) that $D(P_{\mathcal{E}}\|Q) = o(1)$ and $\mathrm{TV}(P_{\mathcal{E}}, Q) = o(1)$. Observe that under our choice of $\tau$ and $\gamma$, Lemma 8 implies that

$$\begin{aligned}\mathbb{P}\{\mathcal{E}^c\} \le \exp\left(-\frac{n\gamma}{8}\right) &= \exp\left(-\frac{\alpha k\log(p/k)}{8}\right)\\ &\le \exp\big(-4k\log\log(p/k)\big) = o_p(1). \qquad (27)\end{aligned}$$

Thus, in view of (8), we get that

$$\begin{aligned}\mathrm{TV}(P, Q) &\le \big(1 - \mathbb{P}\{\mathcal{E}^c\}\big)\mathrm{TV}(P_{\mathcal{E}}, Q) + \mathbb{P}\{\mathcal{E}^c\}\mathrm{TV}(P_{\mathcal{E}^c}, Q)\\ &\le \mathrm{TV}(P_{\mathcal{E}}, Q) + \mathbb{P}\{\mathcal{E}^c\} = o(1).\end{aligned}$$

Next we prove (13). We first consider the case where $\lambda$ satisfies $\lambda > \sqrt{k/\sigma^2 + 1/2}$; we then restrict to $\lambda = \sqrt{k/\sigma^2 + 1}$. In view of (7), we have

$$\begin{aligned}\frac{P_{\mathcal{E}}(X,Y)}{Q(X,Y)} &= \frac{1}{Q(Y)Q(X)}\mathbb{E}_\beta\left[\frac{P(X)P(Y|X,\beta)\mathbf{1}_{\{\mathcal{E}\}}(X,\beta)}{\mathbb{P}\{\mathcal{E}\}}\right]\\ &= \mathbb{E}_\beta\left[\frac{P(Y|X,\beta)\mathbf{1}_{\{\mathcal{E}\}}(X,\beta)}{Q(Y)\mathbb{P}\{\mathcal{E}\}}\right],\end{aligned}$$

where the last equality holds because $P(X) = Q(X)$. Hence,

$$\left(\frac{P_{\mathcal{E}}(X,Y)}{Q(X,Y)}\right)^2 = \mathbb{E}_{\beta\perp\beta'}\left[\frac{P(Y|X,\beta)P(Y|X,\beta')\mathbf{1}_{\{\mathcal{E}\}}(X,\beta)\mathbf{1}_{\{\mathcal{E}\}}(X,\beta')}{Q^2(Y)\mathbb{P}^2\{\mathcal{E}\}}\right],$$

where $\beta'$ is an independent copy of $\beta$. Recall $\mathbb{P}\{\mathcal{E}\} = 1 - o(1)$. Therefore,

$$\mathbb{E}_Q\left[\left(\frac{P_{\mathcal{E}}}{Q}\right)^2\right] = (1+o(1)) \cdot \mathbb{E}_{\beta \perp\!\!\!\perp \beta'} \mathbb{E}_X\left[\mathbb{E}_Y\left[\frac{P(Y|X,\beta)P(Y|X,\beta')}{Q^2(Y)}\right]\mathbf{1}_{\{\mathcal{E}\}}(X,\beta)\mathbf{1}_{\{\mathcal{E}\}}(X,\beta')\right].$$

It follows from (21) that

$$\mathbb{E}_Y\left[\frac{P(Y|X,\beta)P(Y|X,\beta')}{Q^2(Y)}\right] = \frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}} \cdot \exp\left\{\frac{\|X(\beta+\beta')\|^2 - (2\lambda^2-1)\|X(\beta-\beta')\|^2}{4\sigma^2(2\lambda^2-1)}\right\}.$$

Combining the last two displayed equations yields that

$$\mathbb{E}_Q\left[\left(\frac{P_{\mathcal{E}}}{Q}\right)^2\right] = \frac{(1+o(1))\lambda^{2n}}{(2\lambda^2-1)^{n/2}} \cdot \mathbb{E}_{\beta \perp\!\!\!\perp \beta'} \mathbb{E}_X\left[e^{\frac{\|X(\beta+\beta')\|^2 - (2\lambda^2-1)\|X(\beta-\beta')\|^2}{4\sigma^2(2\lambda^2-1)}}\mathbf{1}_{\{\mathcal{E}\}}(X,\beta)\mathbf{1}_{\{\mathcal{E}\}}(X,\beta')\right]. \tag{28}$$

Next we break the right hand side of (28) into two disjoint parts depending on whether $\langle \beta, \beta' \rangle \le \tau$. We prove that the part where $\langle \beta, \beta' \rangle \le \tau$ is $1 + o(1)$ and the part where $\langle \beta, \beta' \rangle > \tau$ is $o(1)$. Combining them we conclude the desired result.

*Part 1.* Note that

$$\mathbb{E}_X\left[\exp\left\{\frac{\|X(\beta+\beta')\|^2 - (2\lambda^2-1)\|X(\beta-\beta')\|^2}{4\sigma^2(2\lambda^2-1)}\right\}\mathbf{1}_{\{\mathcal{E}\}}(X,\beta)\mathbf{1}_{\{\mathcal{E}\}}(X,\beta')\right] \cdot \mathbf{1}_{\{\langle \beta,\beta' \rangle \le \tau\}} \le \mathbb{E}_X\left[\exp\left\{\frac{\|X(\beta+\beta')\|^2 - (2\lambda^2-1)\|X(\beta-\beta')\|^2}{4\sigma^2(2\lambda^2-1)}\right\}\right]\mathbf{1}_{\{\langle \beta,\beta' \rangle \le \tau\}}. \tag{29}$$

Since $\langle \beta+\beta', \beta-\beta' \rangle = 0$ and $X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, conditional on $(\beta, \beta')$,

$$\mathrm{Cov}\big(X(\beta+\beta'), X(\beta-\beta')\big) = 0,$$

and therefore,

$$X(\beta+\beta') \sim \mathcal{N}(0, 2(k+s)\mathbf{I}_n)$$

is independent of

$$X(\beta-\beta') \sim \mathcal{N}(0, 2(k-s)\mathbf{I}_n)$$

for $s = \langle \beta, \beta' \rangle$. Therefore,

$$\mathbb{E}_X\left[\exp\left\{\frac{\|X(\beta+\beta')\|^2 - (2\lambda^2-1)\|X(\beta-\beta')\|^2}{4\sigma^2(2\lambda^2-1)}\right\}\right]$$

$$= \mathbb{E}_X\left[\exp\left\{\frac{\|X(\beta+\beta')\|^2}{4\sigma^2(2\lambda^2-1)}\right\}\right]\mathbb{E}_X\left[\exp\left\{-\frac{\|X(\beta-\beta')\|^2}{4\sigma^2}\right\}\right]$$
$$= \left(1-\frac{(k+s)}{\sigma^2(2\lambda^2-1)}\right)^{-n/2}\left(1+\frac{(k-s)}{\sigma^2}\right)^{-n/2}, \tag{30}$$

where the last equality holds if $\lambda > \sqrt{(k+s)/(2\sigma^2)+1/2}$ and follows from the fact that

$$\mathbb{E}_{Z\sim\chi^2(1)}\big[e^{-tZ}\big] = \frac{1}{\sqrt{1+2t}}$$

for $t > -1/2$. Combining (29) and (30) yields that if $\lambda > \sqrt{k/\sigma^2+1/2}$, then

$$\frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}} \cdot \mathbb{E}_{\beta\perp\beta'}\mathbb{E}_X\left[e^{\frac{\|X(\beta+\beta')\|^2-(2\lambda^2-1)\|X(\beta-\beta')\|^2}{4\sigma^2(2\lambda^2-1)}}\mathbf{1}_{\{\mathcal{E}\}}(X,\beta)\mathbf{1}_{\{\mathcal{E}\}}(X,\beta')\right]\mathbf{1}_{\{\langle\beta,\beta'\rangle\le\tau\}}$$
$$\le \frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}}\mathbb{E}_{\beta\perp\beta'}\left[\left(1-\frac{(k+s)}{\sigma^2(2\lambda^2-1)}\right)^{-n/2}\left(1+\frac{(k-s)}{\sigma^2}\right)^{-n/2}\mathbf{1}_{\{s\le\tau\}}\right].$$

In particular, by plugging in $\lambda = \sqrt{k/\sigma^2+1}$, we get that

$$\frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}} \cdot \mathbb{E}_{\beta\perp\beta'}\mathbb{E}_X\left[e^{\frac{\|X(\beta+\beta')\|^2-(2\lambda^2-1)\|X(\beta-\beta')\|^2}{4\sigma^2(2\lambda^2-1)}}\mathbf{1}_{\{\mathcal{E}\}}(X,\beta)\mathbf{1}_{\{\mathcal{E}\}}(X,\beta')\right]\mathbf{1}_{\{\langle\beta,\beta'\rangle\le\tau\}}$$
$$\overset{(a)}{\le}\left(\frac{k}{\sigma^2}+1\right)^n\mathbb{E}_{S\sim\mathrm{Hyp}(p,k,k)}\left\{\left(1+\frac{(k-S)}{\sigma^2}\right)^{-n}\mathbf{1}_{\{S\le\tau\}}\right\}$$
$$= \mathbb{E}_{S\sim\mathrm{Hyp}(p,k,k)}\left\{\left(1-\frac{S}{k+\sigma^2}\right)^{-n}\mathbf{1}_{\{S\le\tau\}}\right\}, \tag{31}$$

where (a) holds by noticing that $s = \langle\beta,\beta'\rangle$ follows a Hypergeometric distribution with parameters $(p,k,k)$ as the dot product of two uniformly at random chosen binary $k$-sparse vectors.

Using Lemma 6 we conclude that under our assumptions, there exists a constant $C > 0$ depending only on $\delta > 0$ such that if $k/\sigma^2 \ge C$ then

$$\mathbb{E}_{S\sim\mathrm{Hyp}(p,k,k)}\left\{\left(1-\frac{S}{k+\sigma^2}\right)^{-n}\mathbf{1}_{\{S\le\tau\}}\right\} = 1+o(1),$$

concluding the Part 1.

*Part 2.* By the definition of $\mathcal{E}$, since $\tau \le s = \langle\beta,\beta'\rangle \le k$,

$$\|X(\beta+\beta')\|^2 \le \mathbb{E}_X\big[\|X(\beta+\beta')\|^2\big](2+\gamma) = 2n(k+s)(2+\gamma) \le 4nk(2+\gamma).$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_X\left[\exp\left\{\frac{\|X(\beta+\beta')\|^2-(2\lambda^2-1)\|X(\beta-\beta')\|^2}{4\sigma^2(2\lambda^2-1)}\right\}\mathbf{1}_{\{\mathcal{E}\}}(X,\beta)\mathbf{1}_{\{\mathcal{E}\}}(X,\beta')\right]\\
&\qquad\cdot\mathbf{1}_{\{\langle\beta,\beta'\rangle>\tau\}}\\
&\quad\le\mathbb{E}_X\left[\exp\left\{\frac{4nk(2+\gamma)-(2\lambda^2-1)\|X(\beta-\beta')\|^2}{4\sigma^2(2\lambda^2-1)}\right\}\right]\mathbf{1}_{\{\langle\beta,\beta'\rangle>\tau\}}\\
&\quad=\exp\left\{\frac{nk(2+\gamma)}{\sigma^2(2\lambda^2-1)}\right\}\left(1+\frac{(k-s)}{\sigma^2}\right)^{-n/2}\mathbf{1}_{\{\langle\beta,\beta'\rangle>\tau\}},
\end{aligned}\tag{32}$$

where the first inequality follows from the definition of event $\mathcal{E}$ and the last equality holds due to (30). It follows that

$$\begin{aligned}
&\frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}}\\
&\qquad\cdot\mathbb{E}_{\beta\perp\beta'}\left[\mathbb{E}_X\left[e^{\frac{\|X(\beta+\beta')\|^2-(2\lambda^2-1)\|X(\beta-\beta')\|^2}{4\sigma^2(2\lambda^2-1)}}\mathbf{1}_{\{\mathcal{E}\}}(X,\beta)\mathbf{1}_{\{\mathcal{E}\}}(X,\beta')\right]\mathbf{1}_{\{\langle\beta,\beta'\rangle>\tau\}}\right]\\
&\le\frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}}\exp\left\{\frac{nk(2+\gamma)}{\sigma^2(2\lambda^2-1)}\right\}\mathbb{E}_{S\sim\mathrm{Hyp}(p,k,k)}\left[\left(1+\frac{(k-S)}{\sigma^2}\right)^{-n/2}\mathbf{1}_{\{S>\tau\}}\right]\\
&\overset{(a)}{\le}\lambda^n e^{n(1+\gamma/2)}\mathbb{E}_{S\sim\mathrm{Hyp}(p,k,k)}\left[\left(1+\frac{(k-S)}{\sigma^2}\right)^{-n/2}\mathbf{1}_{\{S>\tau\}}\right]\\
&\overset{(b)}{=}e^{n(1+\gamma/2)}\mathbb{E}_{S\sim\mathrm{Hyp}(p,k,k)}\left[\left(1-\frac{S}{k+\sigma^2}\right)^{-n/2}\mathbf{1}_{\{S>\tau\}}\right],
\end{aligned}\tag{33}$$

where (a) follows due to $2\lambda^2-1\ge\lambda^2$ and $2\lambda^2-1\ge 2k/\sigma^2$; (b) follows by plugging in $\lambda^2=k/\sigma^2+1$.

Recall that $n\le(1-\alpha)n^*$. Then under our choice of $\alpha$ and $\tau$, applying Lemma 7 with $n$ being replaced by $n/2$, $c=p^{-1/2-\delta}$, we get that there exists a universal constant $C>0$ such that if $k/\sigma^2\ge C$ then

$$\begin{aligned}
&e^{n(1+\gamma/2)}\mathbb{E}_{S\sim\mathrm{Hyp}(p,k,k)}\left[\left(1-\frac{S}{k+\sigma^2}\right)^{-n/2}\mathbf{1}_{\{S>\tau\}}\right]\\
&\qquad\le\exp\left(-\alpha k\log(p/k)+\log((2-c)/(1-c))+n\left(1+\frac{\gamma}{2}\right)\right)\\
&\qquad\overset{(a)}{=}\exp\left(-\frac{1}{4}\alpha k\log(p/k)+\log((2-c)/(1-c))\right)\\
&\qquad\overset{(b)}{\le}\exp\big(-8k\log\log(p/k)+\log((2-c)/(1-c))\big)=o_p(1),
\end{aligned}$$

where (a) follows because under our choice of $\gamma$ and $\alpha$,

$$n\left(1+\frac{\gamma}{2}\right) \le n + \frac{1}{2}\alpha k \log(p/k) \le n^* + \frac{1}{2}\alpha k \log(p/k) \le \frac{3}{4}\alpha k \log(p/k);$$

(b) holds due to $\alpha k \log(p/k) \ge 32k \log\log(p/k)$.

Combing the bounds for Parts 1 and 2, we conclude

$$\chi^2(P_{\mathcal{E}} \| Q) = \mathbb{E}_Q\left[\left(\frac{P_{\mathcal{E}}}{Q}\right)^2\right] - 1 = o(1),$$

as desired. □

## 4. Proof of negative results for recovery

**4.1. Lower bound on MSE.** Our first result provides a connection between the relative entropy $D(P\|Q_\lambda)$ and the MSE of an estimator that depends only a subset of the observations. This bound is general in the sense that it holds for any distribution on $\beta$ with $\mathbb{E}[\|\beta\|^2] = k$. For ease of notation, we write $Q_\lambda$ as $Q$ whenever the context is clear.

**Lemma 2.** *Given an integer $n \ge 2$ and an integer $m \in \{1, \dots, n-1\}$, let $\widehat{\beta}$ be an estimator that is a function of $X$ and the first $m$ observations $(Y_1, \dots, Y_m)$. Then,*

$$\mathrm{MSE}(\widehat{\beta}) \ge e^{-\frac{2}{n-m} D(P\|Q)}(\sigma^2 + k) - \sigma^2. \tag{34}$$

*Proof.* The conditional mutual information $I(\beta; Y \mid X)$ can be rewritten as

$$\begin{aligned} I(\beta; Y \mid X) &= \mathbb{E}_{(\beta,X,Y)\sim P}\left[\log \frac{P(Y|X,\beta)}{P(Y|X)}\right] \\ &= \mathbb{E}_{(\beta,X,Y)\sim P}\left[\log \frac{P(Y|X,\beta)}{Q(Y)}\right] + \mathbb{E}_{(X,Y)\sim P}\left[\log \frac{Q(Y)}{P(Y|X)}\right], \end{aligned}$$

where $(\beta, X, Y) \sim P$ denotes that $(\beta, X, Y)$ are generated according to the planted model. Plugging in the expression of $P(Y|X,\beta)$ and $Q(Y)$, we get that

$$\mathbb{E}_{(\beta,X,Y)\sim P}\left[\log \frac{P(Y|X,\beta)}{Q(Y)}\right] = \frac{n}{2}\log(\lambda^2) + \frac{1}{2}\mathbb{E}\left[\frac{\|Y\|_2^2}{\lambda^2\sigma^2} - \frac{\|Y - X\beta\|_2^2}{\sigma^2}\right].$$

Furthermore, by definition,

$$\mathbb{E}_{(X,Y)\sim P}\left[\log \frac{Q(Y)}{P(Y|X)}\right] = -D(P\|Q).$$

Combining the last three displayed equations gives that

$$\begin{aligned} I(\beta; Y \mid X) &= \frac{n}{2}\log(\lambda^2) + \frac{1}{2}\mathbb{E}\left[\frac{\|Y\|_2^2}{\lambda^2\sigma^2} - \frac{\|Y - X\beta\|_2^2}{\sigma^2}\right] - D(P\|Q) \\ &= \frac{n}{2}\left[\log\left(\frac{\lambda^2}{1+k/\sigma^2}\right) + \frac{1+k/\sigma^2}{\lambda^2} - 1\right] + \frac{n}{2}\log(1+k/\sigma^2) - D(P\|Q) \\ &\geq \frac{n}{2}\log(1+k/\sigma^2) - D(P\|Q), \end{aligned} \tag{35}$$

where the inequality follows from the fact that $\log(u) + 1/u - 1 \geq 0$ for all $u > 0$.

To proceed, we will now provide an upper bound on $I(\beta; Y \mid X)$ in terms of the MSE. Starting with the chain rule for mutual information, we have

$$I(\beta; Y \mid X) = I(\beta; Y_1^m \mid X) + I(\beta; Y_{m+1}^n \mid X, Y_1^m), \tag{36}$$

where we have used the shorthand notation $Y_i^j = (Y_i, \ldots, Y_j)$. Next, we use the fact that mutual information in the Gaussian channel under a second moment constraint is maximized by the Gaussian input distribution. Hence,

$$\begin{aligned} I(\beta; Y_1^m \mid X) &\leq \sum_{i=1}^m I(\beta; Y_i \mid X) \\ &\leq \frac{m}{2}\mathbb{E}\big[\log\big(\mathbb{E}\big[\|Y_1\|^2 \mid X\big]/\sigma^2\big)\big] \\ &\leq \frac{m}{2}\log\big(\mathbb{E}\big[\|Y_1\|^2\big]/\sigma^2\big) \\ &\leq \frac{m}{2}\log(1+k/\sigma^2), \end{aligned} \tag{37}$$

and

$$\begin{aligned} I(\beta; Y_{m+1}^n \mid X, Y_1^m) &\leq \sum_{i=m+1}^n I(\beta; Y_i \mid X, Y_1^m) \\ &\leq \frac{n-m}{2}\log\big(\mathbb{E}\big[\|Y_{m+1} - \mathbb{E}\big[Y_{m+1} \mid X, Y_1^m\big]\|^2\big]/\sigma^2\big) \\ &\leq \frac{n-m}{2}\log\big(1 + \mathrm{MSE}(\widehat{\beta})/\sigma^2\big), \end{aligned} \tag{38}$$

where the last inequality holds due to

$$\begin{aligned} \mathbb{E}\big[\|Y_{m+1} - \mathbb{E}\big[Y_{m+1} \mid X, Y_1^m\big]\|^2\big] &= \mathbb{E}\big[\|\beta - \mathbb{E}\big[\beta \mid Y_1^m, X\big]\|^2\big] + \sigma^2 \\ &\leq \mathrm{MSE}(\widehat{\beta}) + \sigma^2. \end{aligned}$$

Plugging inequalities (37) and (38) back into (36) leads to

$$I(\beta; Y \mid X) \leq \frac{m}{2}\log(1+k/\sigma^2) + \frac{n-m}{2}\log(1+\mathrm{MSE}(\widehat{\beta})/\sigma^2). \tag{39}$$

Comparing (39) with (35) and rearranging terms gives the stated result. □

**4.2. Upper bound on relative entropy via conditioning.** We now show how a conditioning argument can be used to bound from above the relative entropy. Recall that (8) implies

$$D(P\,\|\,Q) \le (1-\varepsilon)D(P_{\mathcal{E}}\,\|\,Q) + \varepsilon D(P_{\mathcal{E}^c}\,\|\,Q). \tag{40}$$

The next result provides an upper bound on the second term on the right-hand side.

**Lemma 3.** *For any $\mathcal{E} \subset \mathbb{R}^p \times \mathbb{R}^{n\times p}$ we have*

$$\varepsilon D(P_{\mathcal{E}^c}\,\|\,Q) \le 2\sqrt{\varepsilon} + \frac{\varepsilon n}{2}\log(\lambda^2) + \frac{\sqrt{\varepsilon}\,n(1+k/\sigma^2)}{\lambda^2},$$

*where $\varepsilon = \mathbb{P}\{(X,\beta) \in \mathcal{E}^c\}$. In particular, if $\lambda^2 = 1 + k/\sigma^2$, then*

$$\varepsilon D(P_{\mathcal{E}^c}\,\|\,Q) \le \frac{\varepsilon n}{2}\log(1+k/\sigma^2) + \sqrt{\varepsilon}(2+n).$$

*Proof.* Starting with the definition of the conditioned planted model in (7), we have

$$P_{\mathcal{E}^c}(X,Y) = \frac{\mathbb{E}_\beta\big[P(X,Y\mid\beta)\mathbf{1}_{\{\mathcal{E}^c\}}(X,\beta)\big]}{\mathbb{P}\{\mathcal{E}^c\}} = \frac{P(X)\mathbb{E}_\beta\big[P(Y\mid X,\beta)\mathbf{1}_{\{\mathcal{E}^c\}}(X,\beta)\big]}{\varepsilon}$$

Recall that $W_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2)$. It follows that $P(Y\mid\beta,X) \le (2\pi\sigma^2)^{-n/2}$, and thus

$$P_{\mathcal{E}^c}(X,Y) \le \frac{P(X)\mathbb{E}_\beta\big[\mathbf{1}_{\{\mathcal{E}^c\}}(\beta,X)\big]}{\varepsilon(2\pi\sigma^2)^{n/2}} \le \frac{P(X)}{\varepsilon(2\pi\sigma^2)^{n/2}}.$$

Therefore, recalling that $Q(X,Y) = P(X)Q(Y)$, we have

$$\begin{aligned} D(P_{\mathcal{E}^c}\,\|\,Q) &= \mathbb{E}_{P_{\mathcal{E}^c}}\left[\log\frac{P_{\mathcal{E}^c}(X,Y)}{P(X)Q(Y)}\right] \\ &\le \mathbb{E}_{P_{\mathcal{E}^c}}\left[\log\frac{1}{\varepsilon\,(2\pi\sigma^2)^{n/2}Q(Y)}\right] \\ &= \log(1/\varepsilon) + \frac{n}{2}\log(\lambda^2) + \frac{\mathbb{E}\big[\|Y\|^2 \mid (X,\beta)\in\mathcal{E}^c\big]}{2\lambda^2\sigma^2} \end{aligned}$$

Multiplying both sides by $\varepsilon$ leads to

$$\varepsilon D(P_{\mathcal{E}^c}\,\|\,Q) \le \varepsilon\log(1/\varepsilon) + \frac{\varepsilon\, n}{2}\log(\lambda^2) + \frac{\mathbb{E}\big[\|Y\|^2\mathbf{1}_{\{\mathcal{E}^c\}}(\beta,X)\big]}{2\lambda^2\sigma^2}$$

The first term on the right-hand side satisfies $\varepsilon\log(1/\varepsilon) \le 2\sqrt{\varepsilon}$. Furthermore, by the Cauchy–Schwarz inequality,

$$\mathbb{E}\big[\|Y\|^2\mathbf{1}_{\{\mathcal{E}^c\}}(\beta,X)\big] \le \sqrt{\mathbb{E}\big[\mathbf{1}_{\{\mathcal{E}^c\}}(X,\beta)\big]\,\mathbb{E}\big[\|Y\|^4\big]} = \sqrt{\varepsilon n(2+n)}(k+\sigma^2),$$

where we have used the fact that $\|Y\|^2/(k+\sigma^2)$ has a chi-squared distribution with $n$ degrees of freedom. Combining the last two displayed inequalities and using the inequality $n+2 \le 3n$ leads to the stated result. □

### 4.3. Proof of Theorem 3.

*Proof.* First, we prove (14) under the theorem assumptions. Let $\mathcal{E}$ be $\mathcal{E}_{\gamma,\tau}$ with $\gamma$ and $\tau$ given in Theorem 2. It follows from Theorem 2 that $D(P_{\mathcal{E}} \| Q_{\lambda_0}) = o_p(1)$. Moreover, it follows from Lemma 8 and $k = o(p)$ that

$$\varepsilon = \mathbb{P}\{\mathcal{E}^c\} \le e^{-4k \log\log(p/k)}.$$

Thus, we get from Lemma 3 that for $\lambda^2 = k/\sigma^2 + 1$

$$\begin{aligned}
\varepsilon D(P_{\mathcal{E}^c} \| Q_{\lambda_0}) &\le \frac{\varepsilon n}{2} \log(1 + k/\sigma^2) + \sqrt{\varepsilon}\,(2 + n) \\
&\le \frac{\varepsilon n^*}{2} \log(1 + k/\sigma^2) + \sqrt{\varepsilon}\,(2 + n^*) \\
&\le e^{-4k \log\log(p/k)} \big(k \log(p/k)\big) + 2e^{-2k \log\log(p/k)} \left(1 + \frac{k \log(p/k)}{\log(1 + k/\sigma^2)}\right) \\
&= o_p(1),
\end{aligned}$$

where the last equality holds due to $k = o(p)$ and $k/\sigma^2 \ge C$ for a sufficiently large constant $C$. In view of the upper bound in (40), we immediately get

$$D(P \| Q_{\lambda_0}) = o_p(1)$$

as desired.

Next we prove (15). Note that if $\lfloor (1-\alpha)n^* \rfloor \le 1$, then (15) is trivially true. Hence, we assume $\lfloor (1-\alpha)n^* \rfloor \ge 2$ in the following. Applying Lemma 2 with

$$n = \lfloor (1-\alpha)n^* \rfloor \quad \text{and} \quad m = \lfloor (1-\alpha)n^* \rfloor - 1$$

yields that

$$\frac{\mathrm{MSE}(\widehat{\beta})}{k} \ge \left(1 + \frac{\sigma^2}{k}\right) \exp\big\{-2D(P \| Q_{\lambda_0})\big\} - \frac{\sigma^2}{k} = 1 - o_p(1), \tag{41}$$

where the last equality holds because $D(P \| Q_{\lambda_0}) = o_p(1)$ and $k/\sigma^2 \ge C$ for a constant $C$. □

## 5. Proof of positive results for recovery and detection

In this section we prove the positive result.

**5.1. Proof of Theorem 4.** Towards proving Theorem 4, we need the following lemma.

**Lemma 4.** *Let $X \in \mathbb{R}^{n\times p}$ with i.i.d. $\mathcal{N}(0,1)$ entries and $W \sim N(0,\sigma^2 I_n)$. Furthermore, assume that $\beta, \beta' \in \{0,1\}^p$ are two $k$-sparse vectors with $\|\beta - \beta'\|^2 = 2\ell$ for some $\ell \in \{1,\dots,k\}$. Then,*

$$\mathbb{P}\big\{\|W + X(\beta-\beta')\|^2 \le \|W\|^2\big\} \le \left(1 + \frac{\ell}{2\sigma^2}\right)^{-n/2}.$$

*Proof.* Let $Q(x)$ be the complementary cumulative distribution function of the standard Gaussian distribution, that is for any $x \in \mathbb{R}$, $Q(x) = \mathbb{P}[Z \ge x]$ for $Z \sim \mathcal{N}(0,1)$. The Chernoff bound gives $Q(x) \le e^{-x^2/2}$ for all $x \ge 0$. Then,

$$\begin{aligned}
\mathbb{P}\big\{\|W + X(\beta-\beta')\|^2 \le \|W\|^2\big\} &= \mathbb{P}\big\{2W^T X(\beta-\beta') + \|X(\beta-\beta')\|^2 \le 0\big\} \\
&= \mathbb{P}\left\{\frac{-W^T X(\beta-\beta')}{\sigma\|X(\beta-\beta')\|} \ge \frac{\|X(\beta-\beta')\|}{2\sigma}\right\} \\
&\overset{(a)}{=} \mathbb{E}\left[Q\left(\frac{\|X(\beta-\beta')\|}{2\sigma}\right)\right] \\
&\overset{(b)}{\le} \mathbb{E}\left[\exp\left(-\frac{\|X(\beta-\beta')\|^2}{8\sigma^2}\right)\right] \\
&\le \left(1 + \frac{\ell}{2\sigma^2}\right)^{-n/2},
\end{aligned}$$

where (a) holds because conditioning on $X$, $\frac{-W^T X(\beta-\beta')}{\sigma\|X(\beta-\beta')\|} \sim \mathcal{N}(0,1)$; (b) holds due to $Q(x) \le e^{-x^2/2}$; the last inequality follows from $\|X(\beta-\beta')\|^2/(2\ell) \sim \chi^2(n)$ and $\mathbb{E}_{Z\sim\chi^2(1)}\big[e^{-tZ}\big] = \frac{1}{\sqrt{1+2t}}$ for $t > 0$. □

We now proceed with the proof of Theorem 4.

*Proof of Theorem 4.* First, note that when $k = o(p)$, (18) readily follows from (17). In particular, observe that since $\widehat{\beta}, \beta \in \{0,1\}^p$ are binary $k$-sparse vectors, it follows that $\|\widehat{\beta} - \beta\|^2 \le 2k$, and therefore

$$\begin{aligned}
\frac{1}{k}\mathrm{MSE}(\widehat{\beta}) &= \frac{1}{k}\mathbb{E}\big[\|\widehat{\beta} - \beta\|^2\big] \\
&\le \frac{2}{\log(p/k)} + 2\mathbb{P}\left[\|\widehat{\beta} - \beta\|^2 \ge \frac{2k}{\log(p/k)}\right] \\
&\le \frac{2}{\log(p/k)} + \frac{2e^2}{\log^2(p/k)(1-e^{-1})},
\end{aligned}$$

which is $o_p(1)$ when $k = o(p)$.

It remains to prove (17). Set for convenience

$$d \triangleq \frac{k}{\log(p/k)}. \tag{42}$$

By the definition of the MLE,

$$\|W + X(\beta - \widehat{\beta})\|^2 = \|Y - X\widehat{\beta}\|^2 \le \|Y - X\beta\|^2 = \|W\|^2.$$

Hence,

$$\{\|\widehat{\beta} - \beta\|^2 \ge 2d\} = \cup_{\ell \ge d}^{k} \{\exists \beta' \in \{0,1\}^p : \|\beta'\|_0 = k, \|\beta' - \beta\|^2 = 2\ell, \\ \|W + X(\beta - \beta')\|^2 \le \|W\|^2\}.$$

By a union bound and Lemma 4, we have that

$$\begin{aligned}
\mathbb{P}\{\|\widehat{\beta} - \beta\|^2 \ge 2d\} &\le \sum_{\ell \ge d}^{k} \binom{k}{\ell}\binom{p-k}{\ell}\left(1 + \frac{\ell}{2\sigma^2}\right)^{-n/2} \\
&\overset{(a)}{\le} \sum_{\ell \ge d}^{k} \left(\frac{ke}{\ell}\right)^{\ell}\left(\frac{pe}{\ell}\right)^{\ell}\left(1 + \frac{\ell}{2\sigma^2}\right)^{-n/2} \\
&\overset{(b)}{\le} \sum_{\ell \ge d}^{k} \left(\frac{e^2 pk}{d^2}\right)^{\ell}\left(1 + \frac{\ell}{2\sigma^2}\right)^{-n/2} \\
&\triangleq \sum_{\ell \ge d}^{k} \exp(h(\ell) - \ell), \qquad (43)
\end{aligned}$$

where (a) holds due to $\binom{m_1}{m_2} \le (em_1/m_2)^{m_2}$; (b) holds due to $\ell \ge d$; and

$$h(x) \triangleq -\frac{n}{2}\log\left(1 + x/2\sigma^2\right) + x\log\left(e^3 pk/d^2\right).$$

Note that $h(x)$ is convex in $x$; hence the maximum of $h(\ell)$ for $\ell \in [d, k]$ is achieved at either $\ell = d$ or $\ell = k$, i.e.,

$$\max_{d \le \ell \le k} h(\ell) \le \max\{h(d), h(k)\}. \qquad (44)$$

We proceed to bound from above $h(d)$ and $h(k)$. Note that by assumption (16),

$$n \ge \frac{2k\log(p/k)}{\log\left(1 + k/2\sigma^2\right)}\left(1 + \frac{4\log\log(p/k)}{\log(p/k)}\right). \qquad (45)$$

Then we conclude that

$$\begin{aligned}
h(k) &= -\frac{n}{2}\log\left(1 + k/2\sigma^2\right) + k\log\left(e^3 pk/d^2\right) \\
&\overset{(45)}{\le} -k\log(p/k) - 4k\log\log(p/k) + k\log\left(e^3 pk/d^2\right) \\
&\overset{(42)}{\le} -k\log(p/k) - 4k\log\log(p/k) + k\log\left(\frac{e^3 pk\log^2(p/k)}{k^2}\right) \\
&= -2k\log\log(p/k) + 3k. \qquad (46)
\end{aligned}$$

Analogously, we can bound from above $h(d)$ as follows:

$$\begin{aligned} h(d) &= -\frac{n}{2}\log\left(1+d/2\sigma^2\right) + d\log\left(e^3 pk/d^2\right) \\ &\overset{(45)}{\leq} -\left(1+\frac{4\log\log(p/k)}{\log(p/k)}\right)\frac{k\log(p/k)}{\log(1+k/(2\sigma^2))}\log\left(1+d/2\sigma^2\right) \\ &\quad + d\log\left(e^3 pk/d^2\right). \end{aligned} \tag{47}$$

Let

$$q(x) \triangleq \log\left(1+x/2\sigma^2\right) - \frac{x}{k}\log\left(1+k/2\sigma^2\right).$$

Note that $q(x)$ is concave in $x$, $q(0)=0$, and $q(k)=0$. Thus,

$$\min_{0\leq x\leq k} q(x) \geq \min\left\{q(0), q(k)\right\} \geq 0.$$

Hence, $q(d)\geq 0$, i.e.,

$$k\log\left(1+d/2\sigma^2\right) \geq d\log\left(1+k/2\sigma^2\right).$$

Combining the last displayed equation with (47) gives that

$$\begin{aligned} h(d) &\leq -\left(1+\frac{4\log\log(p/k)}{\log(p/k)}\right)d\log(p/k) + d\log\left(e^3 pk/d^2\right) \\ &\overset{(42)}{\leq} -d\log(p/k) - 4d\log\log(p/k) + d\log\left(\frac{e^3 pk\log^2(p/k)}{k^2}\right) \\ &\leq -2d\log\log(p/k) + 3d. \end{aligned}$$

Combining the last displayed equation with (46) and (44), we get that

$$\max_{d\leq \ell\leq k} h(\ell) \leq -2d\log\log(p/k) + 3d.$$

Combining the last displayed equation with (43) yields that

$$\begin{aligned} \mathbb{P}\left\{\|\widehat{\beta}-\beta\|^2 \geq 2d\right\} &\leq e^{-2d\log\log(p/k)+3d}\sum_{\ell\geq d}^{k} e^{-\ell} \\ &\leq e^{-2d\log\log(p/k)+3d}\,\frac{e^{-d}}{1-e^{-1}} \\ &\leq e^{-2\log\log(p/k)}\,\frac{e^2}{1-e^{-1}} \\ &= \frac{e^2}{(1-e^{-1})\log^2(p/k)}, \end{aligned}$$

where the last inequality holds under the assumption $\log\log(p/k)\geq 1$. This completes the proof of Theorem 4. □

### 5.2. Proof of Theorem 5.

*Proof.* We start with the observation that under the null model, we have for any $\tau > 0$.

$$Q(\mathcal{T}(X,Y) \le \tau) \le M\tau^{n/2}. \tag{48}$$

To prove this, notice that under the null model it holds,

$$\mathcal{T}(X,Y) = \frac{\min_{\beta\in\{0,1\}^p, \|\beta\|_0 = k} \|\lambda W - X\beta\|^2}{\|\lambda W\|^2}.$$

Using the union bound, we have for any $\tau > 0$,

$$\begin{aligned} Q(\mathcal{T}(X,Y) \le \tau) &\le \sum_{\beta\in\{0,1\}^p, \|\beta\|_0 = k} \mathbb{P}\left\{\frac{\|\lambda W - X\beta\|^2}{\|\lambda W\|^2} \le \tau\right\} \\ &\le M \sup_{\beta\in\mathbb{R}^p} \mathbb{P}\left\{\frac{\|\lambda W - X\beta\|^2}{\|\lambda W\|^2} \le \tau\right\}, \end{aligned}$$

where $M = \binom{p}{k}$. Conditional on the event $\{W = w\}$, it follows that $\|\lambda w - X\beta\|^2/\|\beta\|^2$ has a noncentral chi-square distribution with $n$ degrees of freedom and noncentrality parameter $\lambda^2\|w\|^2/\|\beta\|^2$. Thus,

$$\begin{aligned} \sup_{\beta\in\mathbb{R}^p} \mathbb{P}\left\{\frac{\|\lambda W - X\beta\|^2}{\|\lambda W\|^2} \le \tau\right\} &\le \mathbb{E}\left[\sup_{\beta\in\mathbb{R}^p} \mathbb{P}\left\{\frac{\|\lambda W - X\beta\|^2}{\|\lambda W\|^2} \le \tau \mid \|W\|\right\}\right] \\ &= \sup_{u\ge 0} \mathbb{P}\{\chi^2_{\mathrm{NC}}(n,u) \le u\tau\}, \end{aligned} \tag{49}$$

where $\chi^2_{\mathrm{NC}}(n,u)$ denotes a noncentral chi-square variable with $n$ degrees of freedom and noncentrality parameter $u$. For all $\theta \ge 0$ the Chernoff bound yields

$$\begin{aligned} \mathbb{P}\{\chi^2_{\mathrm{NC}}(n,u) \le u\tau\} &\le \mathbb{E}\left[\exp\left(\frac{\theta u\tau}{2} - \frac{\theta}{2}\chi^2_{\mathrm{NC}}(n,u)\right)\right] \\ &= \exp\left(\frac{\theta u\tau}{2} - \frac{\theta}{2}\frac{u}{1+\theta} - \frac{n}{2}\log(1+\theta)\right). \end{aligned}$$

Evaluating the above displayed equation with $\theta = \max\{1/\tau - 1, 0\}$ leads to

$$\mathbb{P}\{\chi^2_{\mathrm{NC}}(n,u) \le u\tau\} \le \tau^{n/2}.$$

Combining it with (49) yields that for any $\tau > 0$ (48) holds.

We divide the rest of the proof into two cases depending on whether $n^*$ diverges or not. Let $M = \binom{p}{k}$ and $s = k/\sigma^2$.

*Case 1*: $\log(1+s) = o(\log M)$. In this case, $n^* = \omega(1)$. Under the planted model, we have

$$\mathcal{T}(X,Y) \le \frac{\|W\|^2}{\|W + X\beta\|^2}.$$

For any $\tau > 0$, introducing $Z \triangleq \tau\|W + X\beta\|^2/\sigma^2 - \|W\|^2/\sigma^2$, we observe

$$\frac{\|W\|^2}{\|W + X\beta\|^2} \geq \tau \quad \iff \quad Z \leq 0.$$

Recalling that $W/\sigma$ and $X\beta/\sqrt{k}$ are independent standard Gaussian vectors, we rewrite $Z$ as

$$Z = \begin{bmatrix} W/\sigma \\ X\beta/\sqrt{k} \end{bmatrix} \Psi \begin{bmatrix} W/\sigma & X\beta/\sqrt{k} \end{bmatrix}, \quad \text{where } \Psi \triangleq \begin{bmatrix} \tau - 1 & \tau\sqrt{s} \\ \tau\sqrt{s} & \tau s \end{bmatrix}.$$

It follows that $Z$ is equal in distribution to the random variable $\lambda_1 A + \lambda_2 B$, where $A$, $B$ are i.i.d. chi-squared random variables $\chi^2(n)$ with $n$ degrees of freedom, and $\lambda_1$ and $\lambda_2$ are the eigenvalues of $\Psi$. Hence, if $\tau > 1/(1 + s)$, then

$$\mathbb{E}[Z] = n \operatorname{Tr}(\Psi) = n[(1 + s)\tau - 1] > 0$$

and by Chebyshev's inequality,

$$\mathbb{P}\{Z \leq 0\} \leq \frac{\mathsf{var}(Z)}{\mathbb{E}[Z]^2} = \frac{2 \operatorname{Tr}(\Psi^2)}{n \operatorname{Tr}(\Psi)^2}.$$

Noting that

$$\operatorname{Tr}(\Psi^2) = 1 - 2\tau + \tau^2 + 2s\tau^2 + s^2\tau^2 = [(1 + s)\tau - 1]^2 + 2s\tau = \operatorname{Tr}(\Psi)^2 + 2s\tau$$

and combining the above displays leads to

$$P(\mathcal{T}(X, Y) \geq \tau) = \mathbb{P}\{Z \leq 0\} \leq \frac{2}{n}\left(1 + \frac{2s\tau}{[(1 + s)\tau - 1]^2}\right) \quad \text{for } \tau > 1/(1 + s). \tag{50}$$

In particular, choose

$$\tau = \frac{1 + \sqrt{st/(1 + s)n}}{1 + s} \tag{51}$$

for some choice of $t > 0$ such that $t = \omega(1)$. We get

$$\begin{aligned} P(\mathcal{T}(X, Y) \geq \tau) &\leq \frac{2}{n}\left(1 + 2s\frac{1 + \sqrt{st/(1 + s)n}}{1 + s}\frac{(1 + s)n}{st}\right) \\ &= \frac{2}{n} + \frac{4}{t} + 4\sqrt{s/(1 + s)nt} = o(1), \end{aligned}$$

where the last equality holds as $n \geq n^* = \omega(1)$ and $t = \omega(1)$.

Meanwhile, under the null model, plugging the choice of $\tau$ as per (51) in (48), we have

$$\begin{aligned}\log Q(\mathcal{T}(X,Y) \le \tau) &\overset{(a)}{\le} \log M - \frac{n}{2}\log(1+s) + \frac{1}{2}\sqrt{stn/(1+s)} \\ &\overset{(b)}{\le} \log M - \frac{n}{2}\log(1+s) + \frac{\sqrt{tn\log(1+s)}}{2} \\ &\overset{(c)}{\le} -\frac{\epsilon n \log(1+s)}{2(1+\epsilon)} + \frac{\sqrt{tn\log(1+s)}}{2} \overset{(d)}{\to} -\infty;\end{aligned}$$

where (a) holds due to $\log(1+x) \le x$; (b) holds due to $s/(1+s) \le \log(1+s)$; (c) holds by assumption that

$$n \ge (1+\varepsilon)\frac{2\log M}{\log(1+s)}, \quad \text{where } \epsilon = \sqrt{\frac{\log\log M}{\log M}};$$

(d) holds by choosing $\omega(1) \le t \le o(\log\log M)$ and noting that

$$n\log(1+s) \ge 2\log M.$$

In conclusion, we get that $P(T(X,Y) \ge \tau) + Q(T(X,Y) \le \tau) \to 0$.

*Case 2*: $\log(1+s) = \Omega(\log M)$. In this case, $s = \omega(1)$ and $n^* = O(1)$. Under the planted model, letting $A, B \sim \mathcal{N}(0, \mathbf{I}_n)$ be independent standard Gaussian vectors, we have

$$\frac{(1+s)\|W\|^2}{\|W + X\beta\|^2} \overset{\text{dist}}{=} \frac{(1+s)\|A\|^2}{\|A + \sqrt{s}B\|^2} = \frac{\|A\|^2}{\left\|\sqrt{\frac{1}{1+s}}A + \sqrt{\frac{s}{1+s}}B\right\|^2} \overset{\text{a.s.}}{\longrightarrow} \frac{\|A\|^2}{\|B\|^2}. \tag{52}$$

Note that the ratio $\|A\|^2/\|B\|^2$ has a beta prime distribution, and satisfies

$$\mathbb{P}\left\{\|A\|^2/\|B\|^2 \ge t\right\} = o(1) \quad \text{for } t = \omega(1).$$

Let

$$\tau = \frac{t}{1+s}.$$

As a consequence,

$$P(T(X,Y) \ge \tau) \le P\left(\frac{(1+s)\|W\|^2}{\|W + X\beta\|^2} \ge t\right) = \mathbb{P}\left\{\frac{\|A\|^2}{\|B\|^2} \ge t\right\} + o(1) = o(1).$$

Under the null model, it follows from (48) that

$$\begin{aligned}\log Q(\mathcal{T}(X,Y) \le \tau) &\le \log M - \frac{n}{2}\log(1+s) + \frac{n}{2}\log t \\ &\le -\frac{\epsilon n \log(1+s)}{2(1+\epsilon)} + \frac{n}{2}\log t \to -\infty,\end{aligned}$$

where the second inequality holds under the assumption that

$$n \log(1+s) \geq 2(1+\epsilon) \log M,$$

and the last equality holds by choosing $\omega(1) \leq \log t \leq o(\sqrt{\log M(\log\log M)})$ and noting that

$$\epsilon \log(1+s) = \Omega(\epsilon \log M) = \Omega(\sqrt{\log M(\log\log M)}).$$

In conclusion, we also get that $P(T(X,Y) \geq \tau) + Q(T(X,Y) \leq \tau) \to 0$. □

## 6. Conclusion and future work

In this paper, we establish an *all-or-nothing* information-theoretic phase transition for recovering a $k$-sparse vector $\beta \in \{0,1\}^p$ from $n$ independent linear Gaussian measurements $Y = X\beta + W$ with noise variance $\sigma^2$. In particular, we show that the MMSE normalized by the trivial MSE jumps from 1 to 0 at a critical sample size

$$n^* = \frac{2k \log(p/k)}{\log(1 + k/\sigma^2)}$$

within a small window of size $\epsilon n^*$. The constant $\epsilon > 0$ can be made arbitrarily small by increasing the signal-to-noise ratio $k/\sigma^2$. Interestingly, the phase transition threshold $n^*$ is asymptotically equal to the ratio of entropy $H(\beta)$ and the AWGN channel capacity $\frac{1}{2}\log(1 + k/\sigma^2)$. Towards establishing this all-or-nothing phase transition, we also study a closely related hypothesis testing problem, where the goal is to distinguish this planted model $P$ from a null model $Q_\lambda$ where $(X,Y)$ are independently generated and $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \lambda^2\sigma^2)$. When $\lambda = \lambda_0 = \sqrt{k/\sigma^2 + 1}$, we show that the sum of Type-I and Type-II testing errors also jumps from 1 to 0 at $n^*$ within a small window of size $\epsilon n^*$.

Our impossibility results for $n \leq (1-\epsilon)n^*$ apply under a crucial assumption that $k \leq p^{1/2-\delta}$ for some arbitrarily small but fixed constant $\delta > 0$. This naturally implies for $\Omega(p^{1/2}) \leq k \leq o(p)$, two open problems for the identification of the detection and the recovery thresholds, respectively.

For detection, as argued in Appendix C, $k = o(p^{1/2})$ is needed for $n^*$ being the detection threshold, because weak detection is achieved for all $n = \Omega(n^*)$ when $k = \Omega(p^{1/2})$, that is the weak detection threshold becomes $o(n^*)$. The identification of the precise detection threshold when $\Omega(p^{1/2}) \leq k \leq o(p)$ is an interesting open problem.

For recovery, however, we believe that the recovery threshold still equals $n^*$ when $\Omega(p^{1/2}) \leq k \leq o(p)$. To prove this, we propose to study the detection problem where both the (conditional) mean and the covariance are matched between

the planted and null models. Specifically, let us consider a slightly modified null model $Q$ with the matched conditional mean

$$\mathbb{E}_Q[Y|X] = \mathbb{E}_P[Y|X] = \frac{k}{p}X\mathbf{1}$$

and the matched covariance

$$\mathbb{E}_Q[YY^\top] = \mathbb{E}_P[YY^\top],$$

where $\mathbf{1}$ denotes the all-one vector. For example, if $X, W$ are defined as before and

$$Y \triangleq \frac{k}{p}X\mathbf{1} + \lambda W$$

with $\lambda$ equal to $\sqrt{k/\sigma^2 + 1 - k^2/p}$, then both the mean and covariance constraints are satisfied. It is an open problem whether this new null model is indistinguishable from the planted model $P$ when $n \le (1-\epsilon)n^*$ and $\Omega(p^{1/2}) \le k \le o(p)$. If the answer is affirmative, then we may follow the analysis road map in this paper to further establish the impossibility of recovery.

Finally, another interesting question for future work is to understand the extent to which the all-or-nothing phenomenon applies beyond the binary vectors setting or the Gaussian assumptions on $(X, W)$. In this direction, some recent work [37] has shown that under mild conditions on the distribution of $\beta$, the distance between the planted and null models can be bounded in term of "exponential moments" similar to the ones studied in Appendix A.

**Acknowledgements.** We thank the anonymous reviewers for their constructive feedback which improved the presentation of the results and also for pointing out the deeper connections with the works [11] and [33].

G. Reeves is supported by the NSF Grants CCF-1718494 and CCF-1750362. J. Xu is supported by the NSF Grants IIS-1838124, CCF-1850743, and CCF-1856424.

## A. Hypergeometric distribution and exponential moment bound

Throughout this subsection, we fix

$$\lambda^2 = k/\sigma^2 + 1 \quad \text{and} \quad \tau = k\left(1 - \frac{1}{\log^2 \lambda^2}\right). \tag{53}$$

The main focus of this subsection is to give tight characterization of the following "exponential" moment:

$$\mathbb{E}_{S\sim\mathrm{Hyp}(p,k,k)}\left[\left(1 - \frac{S}{k+\sigma^2}\right)^{-n}\mathbf{1}_{\{S\in[a,b]\}}\right]$$

for a given interval $[a, b]$. It turns out this "exponential" moment exhibit quantitatively different behavior in the following three different regimes of overlap $S$: small regime ($s \le \epsilon k$), intermediate regime ($\epsilon k < s \le \tau$), and large regime ($s \ge \tau$), where $\epsilon$ is given in (55).

In the sequel, we first prove Lemma 6, which focuses on the small and intermediate regimes under the assumption $n \le n^*$. Then we prove Lemma 7, which focuses on the large regime under the assumption $n \le (1-\alpha)n^*/2$ for $\alpha \in (0, 1/2)$.

We start with a simple lemma, bounding the probability mass of an hypergeometric distribution.

**Lemma 5.** *Let $p, k \in \mathbb{N}$. Then for $S \sim \mathrm{Hyp}(p, k, k)$ and any $s \in [k]$,*

$$\mathbb{P}(S = s) \le \binom{k}{s}\left(\frac{k}{p-k+1}\right)^s.$$

*Proof.* We have

$$\begin{aligned}\mathbb{P}(S = s) = \binom{k}{s}\frac{\binom{p-k}{k-s}}{\binom{p}{k}} &\le \binom{k}{s}\frac{\binom{p}{k-s}}{\binom{p}{k}} \\ &= \binom{k}{s}\frac{(p-k)!(k)!}{(p-k+s)!(k-s)!} \le \binom{k}{s}\left(\frac{k}{p-k+1}\right)^s. \ \square\end{aligned}$$

Next, we bound from above the "exponential" moment in the small overlap regime ($s \le \epsilon k$), and the intermediate overlap regime ($\epsilon k < s \le \tau$).

**Lemma 6.** *Suppose $n \le n^*$.*

- *If $k \le p^{\frac{1}{2}-\delta}$ for an arbitrarily small but fixed constant $\delta \in \left(0, \frac{1}{2}\right)$ and $k/\sigma^2 \ge C(\delta)$ for a sufficiently large constant $C(\delta)$ only depending on $\delta$, then for any $0 \le \epsilon \le \frac{1}{2}$,*

$$\mathbb{E}_{S\sim\mathrm{Hyp}(p,k,k)}\left[\left(1-\frac{S}{k+\sigma^2}\right)^{-n}\mathbf{1}_{\{S\le\epsilon k\}}\right] = 1 + o_p(1), \tag{54}$$

- *If $k = o(p)$ and $k/\sigma^2 \ge C$ for a sufficiently large universal constant $C$, then for*

$$\epsilon = \epsilon_{k,p} = \frac{\log\log(p/k)}{2\log(p/k)}, \tag{55}$$

*it holds that*

$$\mathbb{E}_{S\sim\mathrm{Hyp}(p,k,k)}\left[\left(1-\frac{S}{k+\sigma^2}\right)^{-n}\mathbf{1}_{\{\epsilon k<S\le\tau\}}\right] = o_p(1), \tag{56}$$

*Proof.* Using Lemma 5,

$$\mathbb{E}_{S\sim\mathrm{Hyp}(p,k,k)}\left[\left(1-\frac{S}{k+\sigma^2}\right)^{-n}\mathbf{1}_{\{S\le\tau\}}\right] = \mathbb{P}\{S=0\}+\sum_{s=1}^{\lfloor\tau\rfloor}\binom{k}{s}\left(\frac{k}{p-k+1}\right)^s e^{-n\log(1-s/(k+\sigma^2))}.$$

Note that

$$\mathbb{P}\{S=0\}=\frac{\binom{p-k}{k}}{\binom{p}{k}}\ge\left(1-\frac{k}{p}\right)^k\ge 1-k^2/p=1+o_p(1),$$

where the last equality holds due to $k\le p^{1/2-\delta}$ for some constant $\delta\in(0,1/2)$. Thus, to show (54) it suffices to show

$$\sum_{s=1}^{\lfloor\epsilon k\rfloor}\binom{k}{s}\left(\frac{k}{p-k+1}\right)^s e^{-n^*\log(1-s/(k+\sigma^2))}=o_p(1),$$

and to show (56) it suffices to show

$$\sum_{s=\lceil\epsilon k\rceil}^{\lfloor\tau\rfloor}\binom{k}{s}\left(\frac{k}{p-k+1}\right)^s e^{-n^*\log(1-s/(k+\sigma^2))}=o_p(1),$$

We first prove (54).

**Proof of (54).** Using the fact that $\binom{k}{s}\le k^s$, we have

$$\begin{aligned}&\sum_{s=1}^{\lfloor\epsilon k\rfloor}\binom{k}{s}\left(\frac{k}{p-k+1}\right)^s e^{-n^*\log(1-s/(k+\sigma^2))}\\ &\le\sum_{s=1}^{\lfloor\epsilon k\rfloor}k^s\left(\frac{k}{p-k+1}\right)^s e^{-n^*\log(1-s/(k+\sigma^2))}\\ &=\sum_{s=1}^{\lfloor\epsilon k\rfloor}e^{-s\log\frac{p-k+1}{k^2}-n^*\log(1-s/(k+\sigma^2))}=\sum_{s=1}^{\lfloor\epsilon k\rfloor}e^{f(s)-s\log((p-k+1/p)},\end{aligned}$$

where for $s\in[1,\epsilon k]$ let the real-valued function $f$ be given by

$$f(s)=-s\log(p/k^2)-n^*\log\left(1-s/(k+\sigma^2)\right).$$

**Claim 1.** Suppose $k\le p^{\frac{1}{2}-\delta}$ for a constant $\delta\in\left(0,\frac{1}{2}\right)$ and $\epsilon\le\frac{1}{2}$. There exists a constant $C_1=C_1(\delta)>0$, such that if $k/\sigma^2\ge C_1$ then it holds that for any $s\in[1,\epsilon k]$,

$$f(s)\le-\frac{1}{2}s\log(p/k^2).$$

*Proof of the claim.* Standard calculus implies that for $x \in (0,1)$,

$$\log(1-x) \geq -(1+x)x.$$

Hence, for $0 \leq x \leq \epsilon \leq 1/2$,

$$\log(1-x) \geq -(1+\epsilon)x. \tag{57}$$

Using this inequality it follows that since for any $s \in [1, \epsilon k]$,

$$\frac{s}{k+\sigma^2} \leq \epsilon,$$

it also holds that

$$\begin{aligned} f(s) &\leq -s\log(p/k^2) + n^*(1+\epsilon)\frac{s}{k+\sigma^2} \\ &= s\bigg(-\log(p/k^2) + \frac{n(1+\epsilon)}{k+\sigma^2}\bigg) \leq -\frac{1}{2}s\log(p/k^2), \end{aligned}$$

where the last inequality holds under the assumption that

$$n^* \leq \frac{(k+\sigma^2)\log(p/k^2)}{2(1+\epsilon)}.$$

Recall that $n^* = \frac{2k\log(p/k)}{\log(1+k/\sigma^2)}$. Hence, it suffices to show that

$$\frac{2k\log(p/k)}{\log(1+k/\sigma^2)} \leq \frac{(k+\sigma^2)\log(p/k^2)}{2(1+\epsilon)},$$

which holds if and only if

$$\bigg[1 - \frac{4(1+\epsilon)}{(1+\sigma^2/k)\log(1+k/\sigma^2)}\bigg]\log(p/k) \geq \log k. \tag{58}$$

By assumption, $k \leq p^{1/2-\delta}$ for $\delta \in \big(0, \frac{1}{2}\big)$. Hence, (58) is satisfied if

$$1 - \frac{4(1+\epsilon)}{(1+\sigma^2/k)\log(1+k/\sigma^2)} \geq \frac{1/2-\delta}{1/2+\delta}.$$

Since $\epsilon \leq \frac{1}{2}$, there exists a constant $C_1 = C_1(\delta) > 0$ depending only on $\delta$ such that if $k/\sigma^2 \geq C_1$ then the last displayed equation is satisfied. This completes the proof of the claim. □

Using the above claim we conclude that

$$\begin{aligned} \sum_{s=1}^{\lfloor \epsilon k \rfloor} e^{f(s) - s\log((p-k+1)/p)} &\leq \sum_{s=1}^{\lfloor \epsilon k \rfloor} e^{-\frac{1}{2}s(\log(p/k^2) + 2\log((p-k+1)/p))} \\ &\leq \frac{e^{-\frac{1}{2}\log((p-k+1)^2/pk^2)}}{1 - e^{-\frac{1}{2}\log((p-k+1)^2/pk^2)}} = o_p(1), \end{aligned}$$

where the last equality holds since $k \leq p^{\frac{1}{2}-\delta}$.

Next we prove (56). Again it suffices to prove (56) for $n = n^*$.

**Proof of (56).** Note that $\binom{k}{s} \le 2^k$. Hence,

$$\begin{aligned}
&\sum_{s=\lceil \epsilon k \rceil}^{\lfloor \tau \rfloor} \binom{k}{s} \left(\frac{k}{p-k+1}\right)^s e^{-n^* \log(1-s/(k+\sigma^2))} \\
&\quad \le 2^k \sum_{s=\lceil \epsilon k \rceil}^{\lfloor \tau \rfloor} \left(\frac{k}{p-k+1}\right)^s e^{-n^* \log(1-s/(k+\sigma^2))} \\
&\quad = 2^k \sum_{s=\lceil \epsilon k \rceil}^{\lfloor \tau \rfloor} e^{-s \log(p/k) - n^* \log(1-s/(k+\sigma^2)) - s \log((p-k+1)/p)}.
\end{aligned}$$

Define for $s \in [0, k]$, the function $g$ given by

$$g(s) \triangleq -s \log(p/k) - n^* \log\big(1 - s/(k+\sigma^2)\big). \tag{59}$$

The function $g$ is convex in $s$ for $\epsilon k \le s \le \tau$, as the addition of two convex functions. Hence, the maximum of $g(s)$ over $s \in [\epsilon k, \tau]$ is achieved at either $s = \epsilon k$ or $s = \tau$. Thus it suffices to bound from above $g(\epsilon k)$ and $g(\tau)$.

**Claim 2.** There exist a universal constant $C_2 > 0$ such that if $k/\sigma^2 \ge C_2$, then $g(\tau) \le -\frac{1}{2} k \log(p/k)$ and $g(\epsilon k) \le -\frac{\epsilon k}{2} \log(p/k)$.

*Proof of the claim.* We first bound from above $g(\tau)$.

$$\begin{aligned}
g(\tau) &\le -\tau \log(p/k) - n^* \log\big(1 - \tau/k\big) \\
&= -\left(1 - \frac{1}{\log^2 \lambda^2}\right) k \log(p/k) + \frac{4k \log(p/k) \log\log(\lambda^2)}{\log(\lambda^2)},
\end{aligned}$$

where the last equality holds by plugging in the expressions of $\tau$ and $n^*$.

Recall that $\lambda^2 = 1 + k/\sigma^2$. Hence, there exists a universal constant $C_2 > 0$ such that if $k/\sigma^2 \ge C_2$, then

$$-\left(1 - \frac{1}{\log^2 \lambda^2}\right) k \log(p/k) + \frac{4k \log(p/k) \log\log(\lambda^2)}{\log(\lambda^2)} \le -\frac{1}{2} k \log(p/k).$$

Combining the last two displayed equations yields that $g(\tau) \le -\frac{1}{2} k \log(p/k)$.

For $g(\epsilon k)$, applying (57), we get that

$$\begin{aligned}
g(\epsilon k) &= -\epsilon k \log(p/k) - n^* \log\left(1 - \frac{\epsilon k}{k+\sigma^2}\right) \\
&\le -\epsilon k \log(p/k) + \frac{n^* \epsilon k}{k+\sigma^2}(1+\epsilon) = \epsilon k \left( -\log(p/k) + \frac{n^*(1+\epsilon)}{k+\sigma^2} \right).
\end{aligned}$$

Note that we can conclude $g(\epsilon k) \le -\frac{\epsilon k}{2}\log(p/k)$ if

$$-\log(p/k) + \frac{n^*(1+\epsilon)}{k+\sigma^2} \le -\frac{1}{2}\log(p/k),$$

which holds if and only if

$$n^* = \frac{2k\log(p/k)}{\log(1+k/\sigma^2)} \le \frac{(k+\sigma^2)\log(p/k)}{2(1+\epsilon)},$$

or equivalently

$$\frac{4(1+\epsilon)}{(1+\sigma^2/k)\log(1+k/\sigma^2)} \le 1.$$

Note that there exists a universal constant $C_2 > 0$ such that if $k/\sigma^2 \ge C_2$ then the last displayed inequality is satisfied and hence $g(\epsilon k) \le -\frac{\epsilon k}{2}\log(p/k)$ where the last inequality holds by choosing $C_2$ sufficiently large. □

Using the above claim we now have that if $k/\sigma^2 \ge C_2$,

$$\begin{aligned}
&\sum_{s=\lceil \epsilon k\rceil}^{\lfloor \tau\rfloor} \binom{k}{s}\left(\frac{k}{p-k+1}\right)^s e^{-n^*\log(1-s/(k+\sigma^2))} \\
&\quad\le 2^k \sum_{s=\lceil \epsilon k\rceil}^{\lfloor \tau\rfloor} e^{g(s)-s\log((p-k+1)/p)} \\
&\quad\le e^{k\log 2+\log k-\frac{\epsilon k}{2}\log(p/k)-k\log((p-k+1)/p)} = o_p(1),
\end{aligned}$$

where the last equality holds due to $\log k \le k$, $k = o(p)$, and that

$$\frac{\epsilon k}{2}\log(p/k) = -\frac{k}{4}\frac{\log\log(p/k)}{\log(p/k)}\log(p/k) = -\frac{k}{4}\log\log(p/k). \qquad \square$$

Finally, we bound from above the "exponential" moment in the large overlap regime ($s \ge \tau$) where $\tau$ is defined in (53).

**Lemma 7.** *Suppose that $k \le cp$ for $c \in (0,1)$ and $k/\sigma^2 \ge C$ for a sufficiently large universal constant $C$. If $n \le \frac{1}{2}(1-\alpha)n^*$ for some $\alpha \le \frac{1}{2}$, then*

$$\begin{aligned}
&\mathbb{E}_{S\sim \mathrm{Hyp}(p,k,k)}\left[\left(1-\frac{S}{k+\sigma^2}\right)^{-n}\mathbf{1}_{\{S\ge\tau\}}\right] \\
&\qquad\le \exp\big(-\alpha k\log(p/k) + \log((2-c)/(1-c))\big). \qquad (60)
\end{aligned}$$

*Proof.* Using Lemma 5, we get that

$$\begin{aligned}
&\mathbb{E}_{S\sim \mathrm{Hyp}(p,k,k)}\left[\left(1-\frac{S}{k+\sigma^2}\right)^{-n}\mathbf{1}_{\{S\ge\tau\}}\right] \\
&\quad\le \sum_{s=\lfloor \tau\rfloor}^{k}\binom{k}{s}\left(\frac{k}{p-k+1}\right)^s e^{-n\log(1-s/(k+\sigma^2))}
\end{aligned}$$

$$\leq \sum_{s=\lfloor \tau \rfloor}^{k} \binom{k}{s} e^{-s\log(p/k) - n\log(1-s/(k+\sigma^2)) - s\log((p-k+1)/p)}$$
$$= \sum_{s=\lfloor \tau \rfloor}^{k} \binom{k}{s} e^{g_n(s) - s\log((p-k+1)/p)},$$

where $g_n(s)$ is given by

$$g_n(s) \triangleq -s\log(p/k) - n\log\big(1 - s/(k+\sigma^2)\big).$$

Note that $g_n(s)$ is convex in $s$ for $\tau \leq s \leq k$. Hence, the maximum of $g_n(s)$ over $s \in [\tau, k]$ is achieved at either $s = \tau$ or $s = k$. In view of (59) and Claim 2, for all $n \leq n^*$.

$$g_n(\tau) \leq g_{n^*}(\tau) = g(\tau) \leq -\frac{1}{2}k\log(p/k).$$

Thus, it remains to bound from above $g_n(k)$.

**Claim 3.** Assume $n \leq \frac{1}{2}(1-\alpha)n^*$ for some $\alpha > 0$. Then $g_n(k) \leq -\alpha k\log(p/k)$.

*Proof of the claim.* For all $n \leq \frac{1}{2}(1-\alpha)n^*$,

$$\begin{aligned} g_n(k) &= -k\log(p/k) - n\log\big(1 - k/(k+\sigma^2)\big) \\ &= -k\log(p/k) + \frac{1}{2}(1-\alpha)n^*\log\big(1 + k/\sigma^2\big)\big) \\ &= -k\log(p/k) + (1-\alpha)k\log(p/k) = -\alpha k\log(p/k). \end{aligned}$$ □

In view of the above claim and the assumption that $\alpha \leq 1/2$, we conclude that for all $n \leq \frac{1}{2}(1-\alpha)n^*$,

$$\begin{aligned} \mathbb{E}_{S\sim \mathrm{Hyp}(p,k,k)}\left[\left(1 - \frac{S}{k+\sigma^2}\right)^{-n} \mathbf{1}_{\{S\geq\tau\}}\right] &\leq \sum_{k=\lfloor\tau\rfloor}^{k} \binom{k}{s} e^{-\alpha k\log(p/k) - s\log((p-k+1)/p)} \\ &\leq e^{-\alpha k\log(p/k)} \sum_{s=0}^{k} \binom{k}{s} \left(\frac{p}{p-k+1}\right)^{s} \\ &\leq e^{-\alpha k\log(p/k)} \left(1 + \frac{p}{p-k+1}\right)^{k} \\ &\leq e^{-\alpha k\log(p/k) + k\log((2-c)/(1-c))}, \end{aligned}$$

where the last equality holds due to the assumption $k \leq cp$. □

## B. Probability of the conditioning event

In this section, we bound from above the probability that the conditioning event does not happen.

**Lemma 8.** *Consider the set $\mathcal{E}_{\gamma,\tau}$ defined in* (11). *Let $\tau = k(1-\eta)$ for some $\eta \in [0,1]$. Then we have*

$$\mathbb{P}\{(X,\beta) \in \mathcal{E}^c_{\gamma,\tau}\} \le \exp\left\{-\frac{n\gamma}{4} + \eta k \log\left(e^2 p/\eta^2 k\right)\right\}.$$

*Furthermore, for*

$$\eta = \frac{1}{\log^2(1+k/\sigma^2)} \quad \textit{and} \quad \gamma \ge \frac{k\log(p/k)}{n\log(1+k/\sigma^2)} \vee \frac{k}{n}$$

*there exists a universal constant $C > 0$ such that if $k/\sigma^2 \ge C$, then*

$$\mathbb{P}\{(X,\beta) \in \mathcal{E}^c_{\gamma,\tau}\} \le \exp\left\{-\frac{n\gamma}{8}\right\}.$$

*Proof.* Fix $\beta$ to be a $k$-sparse binary vector in $\{0,1\}^p$. Let $\beta'$ denote another $k$-sparse binary vector and $s = \langle \beta, \beta' \rangle$. We have $X(\beta+\beta') \sim \mathcal{N}(0, 2(k+s)\mathbf{I}_n)$ and therefore

$$\frac{\|X(\beta+\beta')\|^2}{2(k+s)} \sim \chi^2_n.$$

Observe also that the number of different $\beta'$ with $\langle \beta, \beta' \rangle \ge \tau$ is at most

$$\sum_{\ell=0}^{\lfloor \eta k \rfloor} \binom{k}{\ell}\binom{p-k}{\ell}$$

by counting on the different choices of positions of the entries where $\beta'$ differ from $\beta$. Combining the two observations it follows from the union bound that

$$\mathbb{P}\left\{(X,\beta) \in \mathcal{E}^c_{\gamma,\tau} \mid \beta\right\} \le Q_{\chi^2_n}(n(2+\gamma)) \sum_{\ell=0}^{\lfloor \eta k \rfloor} \binom{k}{\ell}\binom{p-k}{\ell}, \tag{61}$$

where $Q_{\chi^2_n}(x)$ is the tail function of the chi-square distribution.

For all $x > 0$, we have (see, e.g., [29, Lemma 1]:

$$Q_{\chi^2_n}\big(n(1+\sqrt{x}+x/2)\big) \le \exp(-nx/4). \tag{62}$$

Noting that $\sqrt{\gamma} + \gamma/2 \le 1 + \gamma$ for all $\gamma > 0$, we see that

$$Q_{\chi^2_n}\big(n(2+\gamma)\big) \le \exp\{-n\gamma/4\}.$$

Next, using the inequalities $\binom{a}{b} \le (ae/b)^b$ for $a, b \in \mathbb{Z}_{>0}$ with $a < b$, $x \to x \log x$ decreases in $\left(0, \frac{1}{e}\right)$, and $\sum_{i=0}^{d} \binom{m}{i} \le (me/d)^d$ for $d, m \in \mathbb{Z}_{>0}$ with $d < m$ (see, e.g., [28]), we get that

$$\sum_{\ell=0}^{\lfloor \eta k \rfloor} \binom{k}{\ell}\binom{p-k}{\ell} \le \sum_{\ell=0}^{\lfloor \eta k \rfloor} \left(\frac{ek}{\ell}\right)^{\ell}\binom{p-k}{\ell} \le \left(\frac{e}{\eta}\right)^{\eta k} \sum_{\ell=0}^{\lfloor \eta k \rfloor} \binom{p-k}{\ell} \le \left(\frac{e^2 p}{\eta^2 k}\right)^{\eta k}.$$

Combining the above expressions completes the first part of the proof of the lemma.

For the second part, note that under our choice of $\eta$,

$$-\frac{n\gamma}{4} + \eta k \log(e^2 p/\eta^2 k) = -\frac{n\gamma}{4} + \frac{k\big(\log(p/k) + 4\log\log(1 + k/\sigma^2) + 2\big)}{\log^2(1 + k/\sigma^2)}.$$

Under the choice of $\gamma$, there exists a universal constant $C > 0$ such that if $k/\sigma^2 \ge C$, then

$$\frac{n\gamma}{16} \ge \frac{k \log(p/k)}{\log^2(1 + k/\sigma^2)},$$
$$\frac{n\gamma}{16} \ge \frac{k\big(4\log\log(1 + k/\sigma^2) + 2\big)}{\log^2(1 + k/\sigma^2)}.$$

Combining the last two displayed equation yields that

$$-\frac{n\gamma}{4} + \eta k \log(e^2 p/\eta^2 k) \le -\frac{n\gamma}{8}.$$

This completes the proof of the lemma. □

### C. The reason $k = o(p^{1/2})$ is needed for weak detection threshold $n^*$

This section shows that weak detection between the planted model $P$ and the null model $Q_\lambda$ is possible for any choice of $\lambda > 0$ and for all $n = \Omega_p(n^*)$, if $k = \Omega_p(p^{1/2})$, $k/\sigma^2 = \Omega_p(1)$, and $\log(p/k) = \Omega_p(\log(1 + k/\sigma^2))$. In particular, we show the following proposition.

**Proposition 2.** *Suppose*

$$\frac{nk^2}{p(k + \sigma^2 - k^2/p)} = \Omega_p(1). \tag{63}$$

*Then weak detection is information-theoretically possible.*

**Remark 1.** If $k/\sigma^2 = \Omega_p(1)$ and $k/p$ is bounded away from 1, then (63) is equivalent to

$$\frac{nk}{p} = \Omega_p(1).$$

Recall that

$$n^* = \frac{2k\log(p/k)}{\log(1+k/\sigma^2)}.$$

Therefore, if furthermore $k = \Omega_p(p^{1/2})$ and $\log(p/k) = \Omega_p(\log(1+k/\sigma^2))$, then $n^*k/p = \Omega_p(1)$, and hence weak detection is possible for all $n = \Omega_p(n^*)$.

*Proof.* Let $\overline{\beta} = \mathbb{E}[\beta]$ and consider the test statistic

$$\mathcal{T}(X,Y) = \langle Y, X\overline{\beta}\rangle;$$

we declare planted model if $\mathcal{T}(X,Y) \geq 0$ and null model otherwise. Let $A, B$ be independent $n$-dimensional standard Gaussian vectors. Then we have that

$$(X\overline{\beta}, Y) \stackrel{d}{=} \begin{cases} \left(\sqrt{k^2/p}\, A,\ \sqrt{k^2/p}\, A + \sqrt{k+\sigma^2-k^2/p}\, B\right) & \text{if } (X,Y)\sim P, \\ \left(\sqrt{k^2/p}\, A,\ \lambda\sigma B\right) & \text{if } (X,Y)\sim Q_\lambda. \end{cases}$$

Hence,

$$Q_\lambda\left(\langle Y, X\overline{\beta}\rangle \leq 0\right) = \frac{1}{2},$$

and

$$P\left(\langle Y, X\overline{\beta}\rangle \leq 0\right) = \mathbb{E}\left[Q\left(\sqrt{\frac{k^2/p}{k+\sigma^2-k^2/p}}\|A\|\right)\right],$$

where

$$Q(x) = \int_x^\infty (2\pi)^{-1/2}\exp(-t^2/2)\,\mathrm{d}t$$

is the tail function of the standard Gaussian. Therefore, as long as

$$\sqrt{\frac{k^2/p}{k+\sigma^2-k^2/p}}\|A\|$$

does not converge to 0 in probability, then $P\left(\langle Y, X\overline{\beta}\rangle \leq 0\right) \leq 1/2 - \epsilon$ for some positive constant $\epsilon > 0$. Thus,

$$P\left(\langle Y, X\overline{\beta}\rangle < 0\right) + Q_\lambda\left(\langle Y, X\overline{\beta}\rangle \geq 0\right) \leq 1-\epsilon;$$

hence, weak detection is possible. Since $\|A\|_2^2 \sim \chi_n^2$ highly concentrates on $n$, it follows that if

$$\frac{nk^2}{p(k+\sigma^2-k^2/p)} = \Omega_p(1), \tag{64}$$

then weak detection is possible. □

## D. Prior distribution and other types of recovery

**D.1. The uniform prior is the least favorable prior.** All the recovery results of the present work are under the assumption of $\beta$ being a binary $k$-sparse vector chosen uniformly at random. One can naturally wonder how much restrictive is this assumption. In this section we establish that in our setting the uniform prior is the least favorable prior, or, in other words, the MMSE under the uniform prior matches the minimax risk. The result follows by appropriately exploiting the fact that the joint distribution of the columns of $X$ is permutation invariant.

More formally, assume the uniform prior on $\beta$ and the model generated according to (1). Now let $\mathcal{L}(X, Y \mid \beta)$ denote the law of $(X, Y)$ conditional on $\beta$. Then the following "invariance" holds:

$$\mathcal{L}(X, Y \mid \beta) = \mathcal{L}(X\Pi^\top, Y \mid \Pi\beta) \text{ for any } p \times p \text{ permutation matrix } \Pi\,. \tag{65}$$

This follows from the straightforward identities

$$\mathcal{L}(X \mid \beta) = \mathcal{L}(X) = \mathcal{L}(X\Pi^\top) = \mathcal{L}(X\Pi^\top \mid \beta)$$

and

$$\mathcal{L}(Y \mid X, \beta) = \mathcal{L}(Y \mid X\Pi^\top, \Pi\beta).$$

This invariance allows us to establish the claimed result.

**Proposition 3.** *Given a binary $k$-sparse $p$-dimensional vector $\beta$, let $Y, X, W$ generated according to* (1)*. Then it holds that*

$$\inf_{\widehat{\beta}} \sup_{\beta\in\{0,1\}^p:\|\beta\|_0=k} \mathbb{E}\big[\|\beta - \widehat{\beta}\|^2\big] = \inf_{\widehat{\beta}} \mathbb{E}_{\beta\sim\mathrm{Unif}}\big[\mathbb{E}\big[\|\beta - \widehat{\beta}\|^2\big]\big],$$

*where the infimum is taken over all measurable estimators $\widehat{\beta} = \widehat{\beta}(Y, X)$ and by* Unif *we denote the uniform distribution over binary $k$-sparse $p$-dimensional vectors.*

Proposition 3 shows that the minimax risk is equal to the MMSE under the uniform prior. Hence, our main result on the sharp "all-or-nothing" phenomenon for the MMSE under the uniform prior holds for the minimax risk as well.

We now proceed with the proof of Proposition 3.

*Proof of Proposition 3.* Let $\beta^*$ be an arbitrary binary $k$-sparse $p$-dimensional vector and $\Pi$ be an arbitrary $p \times p$ permutation matrix. It holds

$$\begin{aligned}
\mathbb{E}\big[\|\beta - \mathbb{E}[\beta \mid X, Y]\|^2 \mid \beta = \Pi\beta^*\big] &= \mathbb{E}\big[\|\Pi\beta^* - \mathbb{E}[\beta \mid X, Y]\|^2 \mid \beta = \Pi\beta^*\big] \\
&\overset{(a)}{=} \mathbb{E}[\|\beta^* - \mathbb{E}\big[\Pi^\top\beta \mid X, Y\big]\|^2 \mid \beta = \Pi\beta^*] \\
&\overset{(b)}{=} \mathbb{E}[\|\beta^* - \mathbb{E}\big[\beta \mid X\Pi^\top, Y\big]\|^2 \mid \beta = \Pi\beta^*] \\
&\overset{(c)}{=} \mathbb{E}[\|\beta^* - \mathbb{E}\big[\beta \mid X, Y\big]\|^2 \mid \beta = \beta^*] \\
&= \mathbb{E}\big[\|\beta - \mathbb{E}\big[\beta \mid X, Y\big]\|^2 \mid \beta = \beta^*\big],
\end{aligned}$$

where (a) follows from $\Pi^\top \Pi = I$, (b) follows from

$$\mathcal{L}(\beta \mid X\Pi^\top, Y) = \mathcal{L}(\Pi^\top \beta \mid X, Y)$$

due to (65), and (c) follows from (65). The last displayed equation immediately implies that

$$\sup_{\beta \in \{0,1\}^p : \|\beta\|_0 = k} \mathbb{E}\big[\|\beta - \mathbb{E}\big[\beta \mid X, Y\big]\|^2\big] = \mathbb{E}_{\beta \sim \mathrm{Unif}}\big[\mathbb{E}\big[\|\beta - \mathbb{E}\big[\beta \mid X, Y\big]\|^2\big]\big]$$

Therefore, for any measurable estimators $\widehat{\beta} = \widehat{\beta}(Y, X)$,

$$\begin{aligned}\inf_{\widehat{\beta}} \sup_{\beta \in \{0,1\}^p : \|\beta\|_0 = k} \mathbb{E}\big[\|\beta - \widehat{\beta}\|^2\big] &\le \mathbb{E}_{\beta \sim \mathrm{Unif}}\big[\mathbb{E}\big[\|\beta - \mathbb{E}\big[\beta \mid X, Y\big]\|^2\big]\big] \\ &= \inf_{\widehat{\beta}} \mathbb{E}_{\beta \sim \mathrm{Unif}}\big[\mathbb{E}\big[\|\beta - \widehat{\beta}\|^2\big]\big],\end{aligned}$$

where the last equality holds because $\mathbb{E}[\beta \mid X, Y]$ is the MMSE estimator. On the contrary, using the elementary fact that the average of a finite set of real numbers is upper bounded by the maximum of these numbers, it follows that

$$\inf_{\widehat{\beta}} \mathbb{E}_{\beta \sim \mathrm{Unif}}\big[\mathbb{E}\big[\|\beta - \widehat{\beta}\|^2\big]\big] \le \inf_{\widehat{\beta}} \sup_{\beta \in \{0,1\}^p : \|\beta\|_0 = k} \mathbb{E}\big[\|\beta - \widehat{\beta}\|^2\big].$$

Combining the two last displayed equations yields the desired identity. □

**D.2. Recovery in expectation and in probability.** All the recovery results of the present work are stated in terms of squared error in expectation. One can naturally wonder whether we could get similar guarantees/impossibility results in probability. We start with a definition.

**Definition 5** (Strong and weak recovery in probability). We say that $\widehat{\beta} = \widehat{\beta}(Y, X) \in \mathbb{R}^p$ achieves:

- Strong recovery in probability if for any $\epsilon > 0$,

$$\limsup_p \mathbb{P}\big(\|\widehat{\beta}(Y, X) - \beta\|^2 / k > \epsilon\big) = 0;$$

- Weak recovery in probability if for some $\epsilon > 0$,

$$\limsup_p \mathbb{P}\big(\|\widehat{\beta}(Y, X) - \beta\|^2 / k > 1 - \epsilon\big) = 0.$$

Recall the corresponding definition of strong/weak recovery in the MSE sense (Definition 1). It is straightforward to verify that strong recovery with respect to MSE in expectation implies strong recovery in probability, by Markov's inequality. We establish here that the followings also hold.

**Proposition 4.** *Let $p \in \mathbb{N}$ and three fixed sequences of $p$, $n = n_p$, $k = k_p$, $\sigma^2 = \sigma_p^2$. Then the following hold:*

- *Strong recovery (with respect to MSE) is equivalent with strong recovery in probability;*
- *Weak recovery (with respect to MSE) is implied by weak recovery in probability.*

*Proof.* As already discussed, by Markov's inequality, it is straightforward to conclude that strong recovery with respect to MSE implies strong recovery in probability. We now prove that strong/weak recovery in probability imply strong/weak recovery with respect to MSE.

Suppose that for some $\varepsilon \in (0,1)$, $\mathbb{P}\{\frac{1}{k}\|\widehat{\beta} - \beta\|^2 > \varepsilon\}$ converges to zero. Weak recovery implies the existence of such an $\varepsilon \in (0,1)$ while strong recovery implies that one can take any such arbitrarily small $\varepsilon > 0$. Recall that under our assumptions $\frac{1}{k}\|\beta\|^2 = 1$ is bounded. We now define

$$\widehat{\beta}' = \begin{cases} \widehat{\beta} & \text{for } \frac{1}{k}\|\widehat{\beta}\|^2 \le 4, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathcal{E} := \{\frac{1}{k}\|\widehat{\beta} - \beta\| \le \varepsilon\}$. By the triangle inequality,

$$\frac{1}{k}\|\widehat{\beta}\|^2 \le \frac{1}{k}\big(\|\beta\| + \|\beta - \widehat{\beta}\|\big)^2 \le 2 + \frac{2}{k}\|\widehat{\beta} - \beta\|^2,$$

and thus in the event $\mathcal{E}$, since $\varepsilon < 1$, we can conclude that $\frac{1}{k}\|\widehat{\beta}\|^2 \le 4$, and therefore $\widehat{\beta}' = \widehat{\beta}$. Hence, it holds that

$$\frac{1}{k}\mathbb{E}\big[\|\widehat{\beta}' - \beta\|^2\big] = \frac{1}{k}\mathbb{E}\big[\|\widehat{\beta} - \beta\|^2 1_{\mathcal{E}}\big] + \frac{1}{k}\mathbb{E}\big[\|\widehat{\beta}' - \beta\|^2 1_{\mathcal{E}^c}\big] \tag{66}$$

$$\le \varepsilon + 10\mathbb{P}\big\{\tfrac{1}{k}\|\widehat{\beta} - \beta\| \ge \varepsilon\big\}, \tag{67}$$

where the second line follows from the definitions of $\mathcal{E}$ as well as the bound

$$\frac{1}{k}\|\beta - \widehat{\beta}'\|^2 \le \frac{2}{k}\|\beta\|^2 + \frac{2}{k}\|\widehat{\beta}'\|^2 \le 10.$$

The assumption of weak recovery (and of course strong recovery) in probability implies that the second term in (67) converges to zero. Hence, under weak recovery we conclude

$$\limsup_{p \to +\infty} \frac{1}{k}\mathbb{E}\big[\|\widehat{\beta}' - \beta\|^2\big] \le \varepsilon. \tag{68}$$

Since $\varepsilon \in (0,1)$ we conclude the weak recovery with respect to MSE. Now, under the assumption of strong recovery in probability we can take $\varepsilon > 0$ arbitrarily small, and therefore we conclude the strong recovery with respect to MSE. □

**D.3. The all-or-nothing phenomenon for Hamming error.** In this section we formally show that one of our main results, the all-or-nothing phenomenon for strong/weak recovery of $\beta$, holds also if one considers recovering $\beta$ with respect to the Hamming error instead of the MSE.

We start with appropriate definitions. For $\beta$ chosen uniformly at random from the set of binary $k$-sparse vectors, $(X, Y)$ given by (1) and an estimator $\widehat{\beta}$ which is a function of $(X, Y)$, we define the Hamming error (HE) as

$$\mathrm{HE}(\widehat{\beta}) \triangleq \mathbb{E}\big[\|\widehat{\beta} - \beta\|_0\big],$$

where $\|v\|_0$ denotes the $\ell_0$ norm of a vector $v$, that is the cardinality of the support of $v$. In our setting, one can simply choose $\widehat{\beta} = 0$, and obtain (the optimal when $k < p/2$) trivial Hamming error $\mathrm{HE}_0 = \mathbb{E}\big[\|\beta - 0\|_0\big] = k$. We will adopt the following two natural notions of recovery, by comparing the HE of an estimator $\widehat{\beta}$ to $\mathrm{HE}_0$.

**Definition 6** (HE-strong and weak recovery). We say that $\widehat{\beta} = \widehat{\beta}(Y, X) \in \mathbb{R}^p$ achieves:

- HE-strong recovery if $\limsup_{p\to\infty} \mathrm{HE}(\widehat{\beta})/\mathrm{HE}_0 = 0$;
- HE-weak recovery if $\limsup_{p\to\infty} \mathrm{HE}(\widehat{\beta})/\mathrm{HE}_0 < 1$.

Recall the definitions of strong and weak recovery (with respect to MSE) we used in the main body of the paper (Definition 1). We establish the following proposition relating the different notions of recovery.

**Proposition 5.** *Let $p \in \mathbb{N}$ and three fixed sequences of $p$, $n = n_p$, $k = k_p$, $\sigma^2 = \sigma_p^2$. Then the following hold:*

- *Strong recovery (with respect to MSE) is equivalent with* HE-*strong recovery;*
- *Weak recovery (with respect to MSE) is implied by* HE-*weak recovery.*

Note that, under appropriate assumptions, the all-or-nothing phenomenon for the Hamming error follows directly from our main result and Proposition 5. Indeed, according to our main result assuming $k \le p^{\frac{1}{2}-\delta}$ for some $\delta \in (0, 1)$, and $k/\sigma^2 \to +\infty$ as $p \to +\infty$ for any $\epsilon > 0$, if $n > (1+\epsilon)n^*$ strong recovery (with respect to MSE) is possible, while if $n < (1-\epsilon)n^*$ weak recovery (with respect to MSE) is impossible. Using Proposition 5 we immediately conclude that if $n > (1+\epsilon)n^*$ HE-strong recovery is possible, while if $n < (1-\epsilon)n^*$ HE-weak recovery is impossible, as we wanted.

More formally, by combining Theorems 3, 4 and Proposition 5 one immediately obtains the following.

**Corollary 1** (All-or-nothing phase transition for Hamming error). *Let $\delta \in \big(0, \frac{1}{2}\big)$ and $\epsilon \in (0, 1)$ be two arbitrary but fixed constants. Then there exists a constant $C(\delta, \epsilon) > 0$ only depending on $\delta$ and $\epsilon$, such that if $k/\sigma^2 \ge C(\delta, \epsilon)$, then*

- HE-*weak recovery of $\beta$ from $(Y, X) \sim P$ is information-theoretically impossible when*
$$k \leq p^{\frac{1}{2}-\delta} \quad \textit{and} \quad n < (1-\epsilon)n^*;$$
- HE-*strong recovery of $\beta$ from $(Y, X) \sim P$ is information-theoretically possible when*
$$k = o(p) \quad \textit{and} \quad n > (1+\epsilon)n^*.$$

We now conclude the section with proving Proposition 5.

*Proof of Proposition 5.* Let us fix any signal $\beta \in \{0,1\}^p$ and estimator $\widehat{\beta} = \widehat{\beta}(X, Y) \in \mathbb{R}^p$. Denote by $[\widehat{\beta}]$ the Euclidean projection of $\widehat{\beta}$ on $\{0,1\}^p$ and by $\widehat{\beta}^{\text{box}}$ the Euclidean projection of $\widehat{\beta}$ onto $[0,1]^p$ (solving "ties" arbitrarily).

First observe the elementary deterministic inequalities,

$$\|\widehat{\beta} - \beta\|_0 \geq \|[\widehat{\beta}] - \beta\|_0 = \|[\widehat{\beta}] - \beta\|^2.$$

As an immediate corollary, HE-strong (respectively, weak) recovery implies strong (respectively, weak) recovery with respect to MSE.

Now, observe that using that any projection operator is a contraction and elementary deterministic inequalities,

$$\|\widehat{\beta} - \beta\|^2 \geq \|\widehat{\beta}^{\text{box}} - \beta\|^2 = \sum_{i=1}^{p} \left(\widehat{\beta}_i^{\text{box}} - \beta_i\right)^2 \geq \sum_{i=1}^{p} \frac{1}{4} |[\widehat{\beta}_i] - \beta_i| = \frac{1}{4}\|[\widehat{\beta}] - \beta\|_0.$$

As an immediate corollary strong recovery with respect to MSE implies HE-strong recovery. This completes the proof. □

## References

[1] S. Aeron, V. Saligrama, and M. Zhao, Information theoretic bounds for compressed sensing. *IEEE Trans. Inform. Theory* **56** (2010), no. 10, 5111–5130 Zbl 1366.94179 MR 2808668

[2] M. Akçakaya and V. Tarokh, Shannon-theoretic limits on noisy compressive sampling. *IEEE Trans. Inform. Theory* **56** (2010), no. 1, 492–504 Zbl 1366.94214 MR 2589459

[3] A. E. Alaoui, F. Krzakala, and M. I. Jordan, Finite size corrections and likelihood ratio fluctuations in the spiked Wigner model, 2017 arXiv:1710.02903

[4] B. Baldessari, The distribution of a quadratic form of normal random variables. *Ann. Math. Statist.* **38** (1967), 1700–1704 Zbl 0155.27301 MR 219158

[5] J. Banks, C. Moore, J. Neeman, and P. Netrapalli, Information-theoretic thresholds for community detection in sparse networks. In *29th Annual Conference on Learning Theory*, pp. 383–416, Proceedings of Machine Learning Research 49, PMLR, 2016

[6] J. Banks, C. Moore, R. Vershynin, N. Verzelen, and J. Xu, Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization. *IEEE Trans. Inform. Theory* **64** (2018), no. 7, 4872–4994 Zbl 1401.94065 MR 3819345

[7] J. Barbier and F. Krzakala, Replica analysis and approximate message passing decoder for superposition codes. In *2014 IEEE International Symposium on Information Theory*, pp. 1494–1498, IEEE, 2014

[8] J. Barbier and F. Krzakala, Approximate message-passing decoder and capacity achieving sparse superposition codes. *IEEE Trans. Inform. Theory* **63** (2017), no. 8, 4894–4927 Zbl 1372.94387 MR 3683543

[9] J. Barbier, N. Macris, M. Dia, and F. Krzakala, Mutual information and optimality of approximate message-passing in random linear estimation. *IEEE Trans. Inf. Theory* **66** (2020), no. 7, 4270–4303

[10] A. R. Barron and S. Cho, High-rate sparse superposition codes with iteratively optimal estimates. In *2012 IEEE International Symposium on Information Theory*, pp. 120–124, IEEE, 2012

[11] C. Butucea, M. Ndaoud, N. A. Stepanova, and A. B. Tsybakov, Variable selection with Hamming loss. *Ann. Statist.* **46** (2018), no. 5, 1837–1875 Zbl 1414.62126 MR 3845003

[12] E. J. Candes and T. Tao, Decoding by linear programming. *IEEE Trans. Inform. Theory* **51** (2005), no. 12, 4203–4215 Zbl 1264.94121 MR 2243152

[13] A. Carpentier and N. Verzelen, Optimal sparsity testing in linear regression model. *Bernoulli* **27** (2021), no. 2, 727–750 Zbl 7370687 MR 4255213

[14] S. S. Chen, D. L. Donoho, and M. A. Saunders, Atomic decomposition by basis pursuit. *SIAM Rev.* **43** (2001), no. 1, 129–159 Zbl 0979.94010 MR 1854649

[15] S. Cho, High-dimensional regression with random design, including sparse superposition codes. PhD thesis, Yale, New Haven, CT, 2014 MR 3251174

[16] D. L. Donoho, Compressed sensing. *IEEE Trans. Inform. Theory* **52** (2006), no. 4, 1289–1306 Zbl 1288.94016 MR 2241189

[17] A. K. Fletcher, S. Rangan, and V. K. Goyal, Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Trans. Inform. Theory* **55** (2009), no. 12, 5758–5772 Zbl 1367.94090 MR 2597192

[18] D. Gamarnik and I. Zadik, Sparse high dimensional linear regression: Algorithmic barrier and a local search algorithm, 2017 arXiv:1711.04952

[19] D. Gamarnik and I. Zadik, High dimensional linear regression with binary coefficients: Mean squared error and a phase transition. In *Proceedings of the 2017 Conference on Learning Theory*, pp. 948–953, Proceedings of Machine Learning Research, PMLR, 2017

[20] D. Gamarnik and I. Zadik, High dimensional linear regression using lattice basis reduction. In *Advances in Neural Information Processing Systems 31*, Curran Associates, Inc., 2018

[21] D. Guo and S. Verdú, Randomly spread CDMA: asymptotics via statistical physics. *IEEE Trans. Inform. Theory* **51** (2005), no. 6, 1983–2010 Zbl 1309.94005 MR 2235278

[22] Y. I. Ingster and I. A. Suslina, *Nonparametric goodness-of-fit testing under Gaussian models*. Lect. Notes Stat. 169, Springer, New York, 2003 Zbl 1013.62049 MR 1991446

[23] Y. I. Ingster, A. B. Tsybakov, and N. Verzelen, Detection boundary in sparse regression. *Electron. J. Stat.* **4** (2010), 1476–1526 Zbl 1329.62314 MR 2747131

[24] Y. Jin, Y.-H. Kim, and B. D. Rao, Limits on support recovery of sparse signals via multiple-access communication techniques. *IEEE Trans. Inform. Theory* **57** (2011), no. 12, 7877–7892 Zbl 1365.94076 MR 2895366

[25] A. Joseph and A. R. Barron, Least squares superposition codes of moderate dictionary size are reliable at rates up to capacity. *IEEE Trans. Inform. Theory* **58** (2012), no. 5, 2541–2557 Zbl 1365.94019 MR 2952500

[26] A. Joseph and A. R. Barron, Fast sparse superposition codes have near exponential error probability for $R < \mathcal{C}$. *IEEE Trans. Inform. Theory* **60** (2014), no. 2, 919–942 Zbl 1364.94402 MR 3164953

[27] S. Kudekar, S. Kumar, M. Mondelli, H. D. Pfister, E. Şaşoğlu, and R. L. Urbanke, Reed-Muller codes achieve capacity on erasure channels. *IEEE Trans. Inform. Theory* **63** (2017), no. 7, 4298–4316 Zbl 1370.94389 MR 3666961

[28] N. Kumar, Bounding the volume of hamming balls, August 2010 `https://cstheory.wordpress.com/2010/08/13/bounding-the-volume-of-hamming-balls/`

[29] B. Laurent and P. Massart, Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** (2000), no. 5, 1302–1338 Zbl 1105.62328 MR 1805785

[30] C. Méasson, A. Montanari, and R. Urbanke, Maxwell construction: the hidden bridge between iterative and maximum *a posteriori* decoding. *IEEE Trans. Inform. Theory* **54** (2008), no. 12, 5277–5307 Zbl 1319.94027 MR 2590511

[31] A. Miller, *Subset selection in regression*. Monogr. Statist. Appl. Probab. 40, Chapman and Hall, Ltd., London, 1990 Zbl 0702.62057 MR 1072361

[32] E. Mossel, J. Neeman, and A. Sly, Reconstruction and estimation in the planted partition model. *Probab. Theory Related Fields* **162** (2015), no. 3-4, 431–461 Zbl 1320.05113 MR 3383334

[33] M. Ndaoud and A. B. Tsybakov, Optimal variable selection and adaptive noisy compressed sensing. *IEEE Trans. Inform. Theory* **66** (2020), no. 4, 2517–2532 Zbl 1448.94081 MR 4087700

[34] A. Perry, A. S. Wein, and A. S. Bandeira, Statistical limits of spiked tensor models. *Ann. Inst. Henri Poincaré Probab. Stat.* **56** (2020), no. 1, 230–264 Zbl 1439.62073 MR 4058987

[35] Y. Polyanskiy and Y. Wu, Lecture notes on information theory, February 2015 `http://people.lids.mit.edu/yp/homepage/data/itlectures_v4.pdf`

[36] K. R. Rad, Nearly sharp sufficient conditions on exact sparsity pattern recovery. *IEEE Trans. Inform. Theory* **57** (2011), no. 7, 4672–4679 Zbl 1365.62203 MR 2840483

[37] G. Reeves, Conditional central limit theorems for Gaussian projections. In *2017 IEEE International Symposium on Information Theory*, pp. 3055–3059, IEEE, 2017

[38] G. Reeves and M. Gastpar, The sampling rate-distortion tradeoff for sparsity pattern recovery in compressed sensing. *IEEE Trans. Inform. Theory* **58** (2012), no. 5, 3065–3092 Zbl 1365.94184 MR 2952533

[39] G. Reeves and M. C. Gastpar, Approximate sparsity pattern recovery: information-theoretic lower bounds. *IEEE Trans. Inform. Theory* **59** (2013), no. 6, 3451–3465 Zbl 1364.94250 MR 3061258

[40] G. Reeves and H. D. Pfister, The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact. In *2016 IEEE International Symposium on Information Theory*, pp. 665–669, IEEE, 2016

[41] C. Rush, A. Greig, and R. Venkataramanan, Capacity-achieving sparse superposition codes via approximate message passing decoding. *IEEE Trans. Inform. Theory* **63** (2017), no. 3, 1476–1500 Zbl 1366.94609 MR 3625975

[42] J. Scarlett and V. Cevher, Limits on support recovery with probabilistic models: an information-theoretic framework. *IEEE Trans. Inform. Theory* **63** (2017), no. 1, 593–620 Zbl 1359.94282 MR 3599962

[43] T. Tanaka, A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors. *IEEE Trans. Inform. Theory* **48** (2002), no. 11, 2888–2910 Zbl 1062.94527 MR 1945581

[44] N. Verzelen, Minimax risks for sparse regressions: ultra-high dimensional phenomenons. *Electron. J. Stat.* **6** (2012), 38–90 Zbl 1334.62120 MR 2879672

[45] M. J. Wainwright, Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory* **55** (2009), no. 12, 5728–5741 Zbl 1367.94106 MR 2597190

[46] M. J. Wainwright, Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** (2009), no. 5, 2183–2202 Zbl 1367.62220 MR 2729873

[47] W. Wang, M. J. Wainwright, and K. Ramchandran, Information-theoretic limits on sparse signal recovery: dense versus sparse measurement matrices. *IEEE Trans. Inform. Theory* **56** (2010), no. 6, 2967–2979 Zbl 1366.94130 MR 2683451

[48] Y. Wu and J. Xu, Statistical problems with planted structures: Information-theoretical and computational limits, 2018 arXiv:1806.00118

Received 28 July 2019; revised 24 August 2021

G. Reeves, Department of Electrical and Computer Engineering and
Department of Statistical Science, Duke University, Durham, NC 27708, USA
E-mail: galen.reeves@duke.edu

J. Xu, Fuqua School of Business, Duke University, Durham, NC 27708, USA
E-mail: jiamingxu.868@duke.edu

I. Zadik, Operations Research Center, Massachusetts Institute of Technology,
Cambridge, MA 02142, USA
E-mail: izadik@mit.edu