# The All-or-Nothing Phenomenon in Sparse Linear Regression

**Galen Reeves** GALEN.REEVES@DUKE.EDU
*Department of ECE and Department of Statistical Science, Duke University*

**Jiaming Xu** JIAMINGXU.868@DUKE.EDU
*The Fuqua School of Business, Duke University*

**Ilias Zadik** IZADIK@MIT.EDU
*Operations Research Center, MIT*

**Editors:** Alina Beygelzimer and Daniel Hsu

## 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 $k/\sigma^2 = \Omega(1)$ regime, we establish an "All-or-Nothing" information-theoretic phase transition at a critical sample size $n^* = 2k\log(p/k)/\log\left(1+k/\sigma^2\right)$, resolving a conjecture of Gamarnik and Zadik (2017a). 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 $\liminf_{p\to\infty} n/n^* < 1$, 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]

**Keywords:** Sparse linear regression; conditional 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

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

where $\beta$ is assumed to be uniformly chosen at random from 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-theoretic possible.

1. Extended abstract. Full version appears as [arXiv reference, 1903.05046]

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. Chapman and Hall (1990), signal denoising, e.g. Chen et al. (2001), compressive sensing, e.g. Candes and Tao (2005), Donoho (2006), information and coding theory, e.g. Joseph and Barron (2012), as well as high dimensional statistics, e.g. Wainwright (2009b,a). The assumptions of Gaussianity of the entries of $(X, W)$ are standard in the literature. Furthermore, much of the literature (e.g. Aeron et al. (2010), Ndaoud and Tsybakov (2018), Wang et al. (2010)) 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 Aeron et al. (2010), Gamarnik and Zadik (2017a) and Gamarnik and Zadik (2017b). 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

$$\text{MSE}\left(\widehat{\beta}\right) \triangleq \mathbb{E}\left[\|\widehat{\beta} - \beta\|^2\right],$$

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 $\text{MSE}_0 = \mathbb{E}\left[\|\beta - \mathbb{E}[\beta]\|^2\right]$, which equals $k\left(1 - \frac{k}{p}\right)$. We will adopt the following two natural notions of recovery, by comparing the MSE of an estimator $\widehat{\beta}$ to $\text{MSE}_0$.

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

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

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.

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. Aeron et al. (2010); Reeves and Gastpar (2012, 2013); Scarlett and Cevher (2017). Another line of work has derived explicit formulas for the minimum MSE (MMSE) $\mathbb{E}\left[\|\beta - \mathbb{E}[\beta \mid X, Y]\|^2\right]$. These formulas were first obtained heuristically using the replica method from statistical physics Tanaka (2002); Guo and Verdú (2005) and later proven rigorously in Reeves and Pfister (2016); Barbier et al. (2016). However, to our best of knowledge, none of the rigorous techniques of Reeves and Pfister (2016); Barbier et al. (2016) 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 $\mathbb{E}\left[YY^\top\right] = (k + \sigma^2)\mathbf{I}$ in both the planted and null models.

Analogous to recovery, we adopt the following two natural notions of testing Perry et al. (2016); Alaoui et al. (2017).

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

- *strong detection if*

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

- *weak detection, if*

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

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

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

Thus equivalently, strong detection is possible if and only if $\liminf_{p\to\infty} \mathrm{TV}(\mathbb{P}, \mathbb{Q}) = 1$, and weak detection is possible if and only if $\liminf_{p\to\infty} \mathrm{TV}(\mathbb{P}, \mathbb{Q}) > 0$. 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.

### 1.1. Overview of the Main Results

Of fundamental importance is the following sample size:

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

In this work, 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 above a sufficiently large constant. We state our main contribution in the following Theorem, which summarizes all the main results stated in the arXiv version of the current paper Reeves et al. (2019).

**Theorem 3 (All-or-Nothing Phase Transition)** *Let $\delta \in (0, 1/2)$ and $\epsilon \in (0, 1)$ be two arbitrary but fixed constants. Then there exists a constant $C(\delta, \epsilon) > 0$ only depending only $\delta$ and $\epsilon$, such that if $k/\sigma^2 \geq C(\delta, \epsilon)$, then*

*(a) 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 $\lambda_0 = \sqrt{\frac{k}{\sigma^2} + 1}$.*

*(b) 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 $\lambda > 0$.*

*(†): strong detection requires an additional assumption $1 + k/\sigma^2 \leq (k \log(p/k))^{1-\eta}$ for some arbitrarily small but fixed constant $\eta > 0$.*

Our results establishes as a corollary a conjecture from (Gamarnik and Zadik (2017a)) 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 (Gamarnik and Zadik (2017a)) it is predicted that the sharp all-or-nothing phase transitions takes places at the sample size

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

Note that our result implies the conjecture because under the additional assumptions $\log k = o(\log p)$ and $k/\sigma^2 \to +\infty$ as $p \to +\infty$, the phase transition point $n^*$ defined in (3) can be straightforwardly checked to satisfy $n^*/n_{\text{conj}} \to 1$, as $p \to +\infty$.

Note that the theorem above assumes $\sigma > 0$. 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. Gamarnik and Zadik (2018))

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 $k = o\left(p^{1/2}\right)$ is needed for $n^*$ to be the weak detection sample size threshold. The sharp information-theoretic threshold for either detection or recovery is still open when $k = \Omega\left(p^{1/2}\right)$ and $k = o(p)$.

**The phase transition role of $n^*$** According to our main result, the rescaled minimum mean squared error of the problem, $\text{MMSE}/\text{MSE}_0$, 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* Méasson et al. (2008); Kudekar et al. (2017) The approach described below is similar to the one used previously for linear regression Reeves and Pfister (2016).

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 coincidence 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. Using the chain rule for the mutual information and that the 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 $I_{n+1} - I_n \le \frac{1}{2}\log(1 + \mathrm{MMSE}_n/\sigma^2)$, where $\mathrm{MMSE}_n$ denotes the minimum MSE with $n$ measurements (see for example the second part of Lemma 15 in Reeves and Pfister (2016) where the difference $\frac{1}{2}\log(1 + \mathrm{MMSE}_n/\sigma^2) - (I_{n+1} - I_n)$ is proven to be equal to the KL divergence between two distributions). In particular, all the increments are between zero and $\frac{1}{2}\log(1 + k/\sigma^2)$ and by telescopic summation for any $n$:

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

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

Suppose now that we have established that strong recovery is achieved with $n^* = \frac{H(\beta)}{\frac{1}{2}\log(1+k/\sigma^2)}$ samples. Then strong recovery and standard identities connecting mutual information and entropy implies that

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

In particular, (4) holds with equality, which means for all $n \le n^* - 1$, $\mathrm{MMSE}_n = k$. In particular, for all $n < n^*$, weak recovery is impossible. This area theorem is the key underpinning our converse proof of the weak recovery.

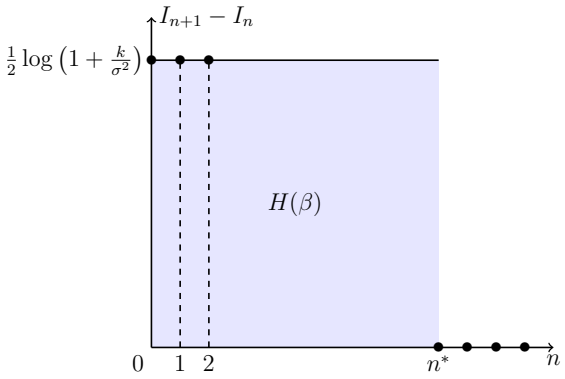


Figure 1: The phase transition diagram in Gaussian sparse linear regression. The $y$-axis is the increment of mutual information with one additional measurement. The area of blue region equals the entropy $H(\beta) \sim k\log(p/k)$.

### 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 section we focus solely on the Gaussian and binary setting and furthermore on the results applying to high values of signal-to-noise ratio and sublinear sparsity.

**Information-theoretic Negative Results for weak/strong recovery** For the impossibility direction, previous work (Aeron et al., 2010, Theorem 5.2) has established that as $p \to \infty$, achieving $\mathrm{MSE}(\widehat{\beta}) \leq d$ for any $d \in [0, k]$ is information-theoretically impossible if

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

where $h_2(\alpha) = -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$ for $\alpha \in [0,1]$ is the binary entropy function. This converse result is proved via a simple rate-distortion argument (see, e.g. Wu and Xu (2018) for an exposition). In particular, given any estimator $\widehat{\beta}(X, Y)$ with $\mathrm{MSE}(\widehat{\beta}) \leq d$, we have

$$p\left(h_2(k/p) - h_2(d/p)\right) \leq \inf_{\mathrm{MSE}(\widetilde{\beta}) \leq d} I(\widetilde{\beta}; \beta) \leq I(\widehat{\beta}; \beta) \leq I(X, Y; \beta) \leq \frac{n}{2} \log\left(1 + k/\sigma^2\right).$$

Notice that since $k = o(p)$ the result implies that if $n \leq (1 - o(1))\, n^*$, *strong* recovery, that is $d = o(k)$, is information-theoretically impossible and if $n = o(n^*)$, *weak* recovery, that is $d \leq (1-\epsilon)k$ for an arbitrary $\epsilon \in (0,1)$, is impossible.

More recent work (Scarlett and Cevher, 2017, 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.

Restricting to the Maximum Likelihood Estimator (MLE) performance of the problem, it is shown in Gamarnik and Zadik (2017a) that under significantly small sparsity $k = O\left(\exp\left(\sqrt{\log p}\right)\right)$ and $k/\sigma^2 \to +\infty$, if $n \leq (1-\epsilon)n^*$, 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 (part (a) of Theorem 3) establishes that the MLE performance is fundamental. It improves upon the negative results in the literature by identifying a sharp threshold for weak recovery, showing that if $k = o\left(\sqrt{p}\right)$, $k/\sigma^2 \geq C$ for some large constant $C > 0$, and $n \leq (1-\epsilon)\, n^*$, 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.

**Information-theoretic Positive Results for weak/strong recovery** In the positive direction, previous work (Akcakaya and Tarokh, 2010, 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 (part (b) of Theorem 3) identifies the explicit value of $C_{k/\sigma^2}$ for which both weak and strong recovery are possible, that is $C_{k/\sigma^2} = 2/\log\left(1 + k/\sigma^2\right)$ for which $C_{k/\sigma^2} k \log(p/k) = n^*$.

In Gamarnik and Zadik (2017a) it is shown that when $k = O\left(\exp\left(\sqrt{\log p}\right)\right)$ 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 our result mentioned in part (b) of Theorem 3 by showing that when $n \geq (1+\epsilon)n^*$ for some fixed $\epsilon > 0$ and any $k \leq cp$ for some $c > 0$, then there exists a constant $C > 0$ such that $k/\sigma^2 \geq C$ the MLE achieves *strong* recovery. In particular, we significantly relax the assumption from Gamarnik and Zadik (2017a) by showing that MLE achieves *strong* recovery with $(1+\epsilon)n^*$ samples for (1) any sparsity level less than $cp$ 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 $\mathbb{E}\left[\|\beta - \mathbb{E}\left[\beta \mid X, Y\right]\|^2\right]$ 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 Tanaka (2002); Guo and Verdú (2005) and later proven rigorously Reeves and Pfister (2016); Barbier et al. (2016). More specifically, for fixed $(\varepsilon, \gamma)$, let the asymptotic MMSE as a function of $\delta$ be defined by

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

The results in Reeves and Pfister (2016); Barbier et al. (2016) 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 $\varepsilon = k/p$ decreases to zero. Note that the comparison is qualitative in the following sense; in the work by Reeves and Pfister (2016); Barbier et al. (2016) the coefficients of $\beta$ are generated i.i.d. according to a $\mathrm{Bernoulli}\,(k/p)$ distribution, while in this paper we consider $\beta$ to be chosen according to a uniform prior over the 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 indeed 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 Reeves and Pfister (2016); Barbier et al. (2016). 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 the current work, we rigorously establish the step behavior 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 $(Y = X\beta + W, X)$ receives 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 of $k$ block of entries indexed by $i\lfloor p/k \rfloor, i\lfloor p/k \rfloor +$

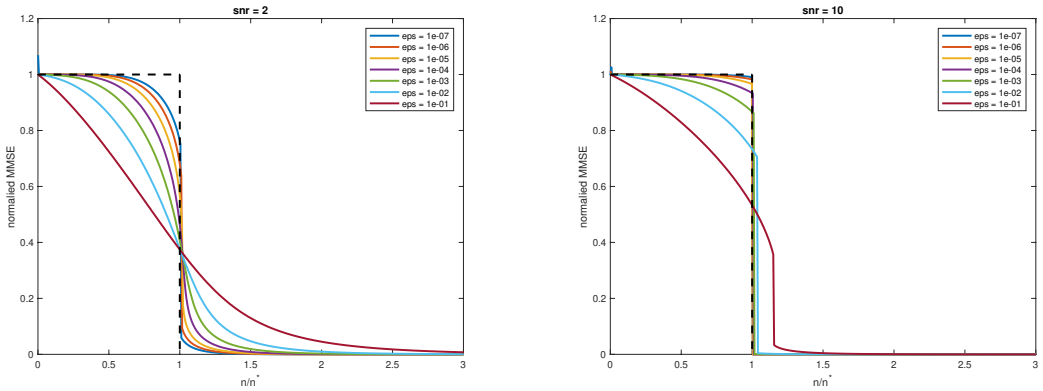


Figure 2: The limit of the replica-symmetric predicted MMSE $\mathcal{M}_{\varepsilon,\gamma}(\cdot)$ as $\epsilon \to 0$ for signal-to-noise ratio (snr) $\gamma$ equal to 2 (left curve) and equal to 10 (right curve).

$1, \cdots, (i+1)\lfloor p/k \rfloor - 1$ for $i = 0, 1, 2, \ldots, k-1$. It has shown that by using various polynomial-time decoding mechanisms, such as adaptive successive decoding Joseph and Barron (2012), Joseph and Barron (2014), a soft-decision iterative decoder Barron and Cho (2012), Cho (2014) and finally Approximate Message Passing techniques Barbier and Krzakala (2014, 2017); Rush et al. (2017), 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, part (b) of Theorem 3 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 recovery algorithm for this task and $(1+\epsilon)n^*$ samples remains largely an open problem (see Gamarnik and Zadik (2017a)).

**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., $\mathbb{P}\left\{\widehat{\beta} = \beta\right\} \to 1$ as $p \to \infty$ (see, e.g. Wainwright (2009b); Fletcher et al. (2009); Rad (2011); Wang et al. (2010); Ndaoud and Tsybakov (2018) and the references therein). Clearly since exact recovery implies weak and strong recovery, the sample sizes required to be achieve exact recovery are in principle no smaller than $n^*$.

Specifically, it has been shown in (Wainwright, 2009b, Theorem 1) that the maximum likelihood estimator achieves exact recovery if $n \geq \Omega\left(\log\binom{p-k}{k} + \sigma^2 \log(p-k)\right)$ and $n - k \to +\infty$. Conversely, $n > \max\{f_1(p,k), \ldots, f_k(p,k), k\}$ is shown in (Wang et al., 2010, 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 (Jin et al., 2011, 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 two notions of exact and strong recovery coincide.

Computationally, it has been shown in (Wainwright, 2009a, Section IV-B) that LASSO achieves exact recovery in polynomial-time if $n \geq 2k\log(p-k)$. More recently, it is shown in (Ndaoud and Tsybakov, 2018, 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 $n \geq \Omega\left(k \log \frac{ep}{k} + \sigma^2 \log p\right)$.

## 2. 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 $\frac{1}{2}\log\left(1 + k/\sigma^2\right)$. Towards establishing this All-or-Northing 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 $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \lambda^2\sigma^2\right)$. 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 $\Omega\left(p^{1/2}\right) \leq k \leq o\left(p\right)$, two open problems for the identification of the detection and the recovery thresholds, respectively.

For detection, as established in the Appendix of the arXiv version of the current paper Reeves et al. (2019), $k = o\left(p^{1/2}\right)$ is needed for $n^*$ being the detection threshold, because weak detection is achieved for all $n = \Omega\left(n^*\right)$ when $k = \Omega(p^{1/2})$, that is the weak detection threshold becomes $o\left(n^*\right)$. The identification of the precise detection threshold when $\Omega(p^{1/2}) \leq k \leq o\left(p\right)$ is an interesting open problem.

For recovery, however, we believe that the recovery threshold still equals $n^*$ when $\Omega\left(p^{1/2}\right) \leq k \leq o(p)$. 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 $\mathbb{E}_Q\left[Y|X\right] = \mathbb{E}_P\left[Y|X\right] = \frac{k}{p}X\mathbf{1}$ and the matched covariance $\mathbb{E}_Q\left[YY^\top\right] = \mathbb{E}_P\left[YY^\top\right]$, 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 $\sqrt{\frac{k}{\sigma^2} + 1 - \frac{k^2}{p}}$, 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 $n \leq (1-\epsilon)\, n^*$ and $\Omega\left(p^{1/2}\right) \leq k \leq o(p)$. 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 Reeves (2017) 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 used for the proof of the main results which can be found in the arXiv version of the current paper Reeves et al. (2019).

## Acknowledgment

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

## References

Shuchin Aeron, Venkatesh Saligrama, and Manqi Zhao. Information theoretic bounds for compressed sensing. *IEEE Transactions on Information Theory*, 56(10):5111–5130, October 2010. doi: 10.1109/TIT.2010.2059891.

Mehmet Akcakaya and Vahid Tarokh. Shannon-theoretic limits on noisy compressive sampling. *IEEE Transactions on Information Theory*, 56(1):492–504, December 2010. doi: 10.1109/TIT.2009.2034796.

Ahmed El Alaoui, Florent Krzakala, and Michael I Jordan. Finite size corrections and likelihood ratio fluctuations in the spiked Wigner model. *arXiv preprint arXiv:1710.02903*, 2017.

J. Barbier and F. Krzakala. Replica analysis and approximate message passing decoder for superposition codes. In *2014 IEEE International Symposium on Information Theory*, pages 1494–1498, June 2014. doi: 10.1109/ISIT.2014.6875082.

J. Barbier and F. Krzakala. Approximate message-passing decoder and capacity achieving sparse superposition codes. *IEEE Transactions on Information Theory*, 63(8):4894–4927, Aug 2017. ISSN 0018-9448. doi: 10.1109/TIT.2017.2713833.

Jean Barbier, Mohamad Dia, Nicolas Macris, and Florent Krzakala. The mutual information in random linear estimation. In *Proceedings of the Allerton Conference on Communication, Control, and Computing*, Monticello, IL, 2016.

A. R. Barron and S. Cho. High-rate sparse superposition codes with iteratively optimal estimates. *Proc. IEEE Int. Symp. Inf. Theory*, 2012.

Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

Alan Miller. Chapman and Hall. Subset selection in regression. *Chapman and Hall*, 1990.

Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, January 2001. ISSN 0036-1445. doi: 10.1137/S003614450037906X. URL http://dx.doi.org/10.1137/S003614450037906X.

S. Cho. High-dimensional regression with random design, including sparse superposition codes. *Ph.D. dissertation, Dept. Statist., Yale Univ., New Haven, CT, USA*, 2014.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Alyson K. Fletcher, Sundeep Rangan, and Vivek K Goyal. Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Transactions on Information Theory*, 55(12):5758–5772, November 2009. doi: 10.1109/TIT.2009.2032726.

David Gamarnik and Ilias Zadik. High dimensional linear regression with binary coefficients: Mean squared error and a phase transition. *Conference on Learning Theory (COLT)*, 2017a. URL https://arxiv.org/abs/1701.04455.

David Gamarnik and Ilias Zadik. Sparse high dimensional linear regression: Algorithmic barrier and a local search algorithm. *arXiv Preprint*, 2017b. URL https://arxiv.org/abs/1711.04952.

David Gamarnik and Ilias Zadik. High dimensional linear regression using lattice basis reduction. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

Dongning Guo and Sergio Verdú. Randomly spread CDMA: Asymptotics via statistical physics. *IEEE Transactions on Information Theory*, 51(6):1983–2010, June 2005.

Yuzhe Jin, Young-Han Kim, and Bhaskar D Rao. Limits on support recovery of sparse signals via multiple-access communication techniques. *IEEE Transactions on Information Theory*, 57(12): 7877–7892, 2011.

A. Joseph and A. R. Barron. Fast sparse superposition codes have near exponential error probability for r ¡ c,. *IEEE Trans. Inf. Theory, vol. 60, no. 2, pp. 919–942*, 2014.

Antony Joseph and Andrew R. Barron. Least sqaures superposition codes of moderate dictionary-size are reliable at rates up to capacity. *IEEE Transactions on Information Theory*, 2012.

Shrinivas Kudekar, Santhosh Kumar, Marco Mondelli, Henry D Pfister, Eren Şaşoğlu, and Rüdiger L Urbanke. Reed–muller codes achieve capacity on erasure channels. *IEEE Transactions on Information Theory*, 63(7):4298–4316, 2017.

Cyril Méasson, Andrea Montanari, and Rüdiger Urbanke. Maxwell construction: The hidden bridge between iterative and maximum a posteriori decoding. *IEEE Transactions on Information Theory*, 54(12):5277–5307, 2008.

Mohamed Ndaoud and Alexandre B Tsybakov. Optimal variable selection and adaptive noisy compressed sensing. *arXiv preprint arXiv:1809.03145*, 2018.

Amelia Perry, Alexander S. Wein, and Afonso S. Bandeira. Statistical limits of spiked tensor models. arXiv:1612.07728, Dec. 2016.

K. Rahnama Rad. Nearly sharp sufficient conditions on exact sparsity pattern recovery. *IEEE Transactions on Information Theory*, 57(7):4672–4679, July 2011. ISSN 0018-9448. doi: 10.1109/TIT.2011.2145670.

Galen Reeves. Conditional central limit theorems for Gaussian projections. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 3055–3059, Aachen, Germany, June 2017.

Galen Reeves and Michael Gastpar. The sampling rate-distortion tradeoff for sparsity pattern recovery in compressed sensing. *IEEE Transactions on Information Theory*, 58(5):3065–3092, May 2012. doi: 10.1109/TIT.2012.2184848.

Galen Reeves and Michael Gastpar. Approximate sparsity pattern recovery: Information-theoretic lower bounds. *IEEE Transactions on Information Theory*, 59(6):3451–3465, June 2013. doi: 10.1109/TIT.2013.2253852.

Galen Reeves and Henry D. Pfister. The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 665 – 669, Barcelona, Spain, July 2016. doi: 10.1109/ISIT.2016.7541382. arXiv. Available: https://arxiv.org/abs/1607.02524.

Galen Reeves, Jiaming Xu, and Ilias Zadik. The all-or-nothing phenomenon in sparse linear regression. *arXiv Preprint arXiv:1903.05046*, 2019.

C. Rush, A. Greig, and R. Venkataramanan. Capacity-achieving sparse superposition codes via approximate message passing decoding. *IEEE Trans. Inf. Theory, vol. 63, pp. 1476–1500*, 2017.

Jonathan Scarlett and Volkan Cevher. Limits on support recovery with probabilistic models: An information-theoretic framework. *IEEE Transactions on Information Theory*, 63(1):593–620, September 2017. doi: 10.1109/TIT.2016.2606605.

T. Tanaka. A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors. *IEEE Transactions on Information Theory*, 48(11):2888–2910, November 2002.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5): 2183–2202, 2009a.

Martin J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741, December 2009b.

Wei Wang, Martin J Wainwright, and Kannan Ramchandran. Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *Information Theory, IEEE Transactions on*, 56(6):2967–2979, 2010.

Yihong Wu and Jiaming Xu. Statistical problems with planted structures: Information-theoretical and computational limits. *arXiv preprint arXiv:1806.00118*, 2018.