Proceedings of Machine Learning Research vol 65:1–39, 2017 30th Annual Conference on Learning Theory

High-Dimensional Regression with Binary Coefficients. Estimating Squared Error and a Phase Transition.

author names withheld

Editor: Under Review for COLT 2017

Abstract

We consider a sparse linear regression model $Y = X\beta^* + W$ where X is $n \times p$ matrix Gaussian i.i.d. entries, W is $n \times 1$ noise vector with i.i.d. mean zero Gaussian entries and standard deviation σ, and β^* is $p \times 1$ binary vector with support size (sparsity) k. Using a novel conditional second moment method we obtain a tight up to a multiplicative constant approximation of the optimal squared error $\min_\beta \|Y - X\beta\|_2$, where the minimization is over all k-sparse binary vectors β. The approximation reveals interesting structural properties of the underlying regression problem. In particular,

(a) We establish that $n^* = 2k \log p / \log(2k/\sigma^2 + 1)$ is a phase transition point with the following "all-or-nothing" property. When n exceeds n^*, $(2k)^{-1}\|\beta_2 - \beta^*\|_0 \approx 0$, and when n is below n^*, $(2k)^{-1}\|\beta_2 - \beta^*\|_0 \approx 1$, where β_2 is the optimal solution achieving the smallest squared error. As a corollary n^* is the asymptotic threshold for recovering β^* information theoretically. Note that n^* is asymptotically below the threshold $n_{\text{LASSO/CS}} = (2k + \sigma^2) \log p$, above which the LASSO and Compressive Sensing methods are able to recover β^*.

(b) We compute the squared error for an intermediate problem $\min_\beta \|Y - X\beta\|_2$ where minimization is restricted to vectors β with $\|\beta - \beta^*\|_0 = 2k\zeta$, for some fixed ratio $\zeta \in [0, 1]$. We show that a lower bound part $\Gamma(\zeta)$ of the estimate, which essentially corresponds to the estimate based on the first moment method, undergoes a phase transition at three different thresholds, namely $n_{\text{inf},1} = \sigma^2 \log p$, which is information theoretic bound for recovering β^* when $k = 1$ and σ is large, then at n^* and finally at $n_{\text{LASSO/CS}}$.

(c) We establish a certain Overlap Gap Property (OGP) on the space of all k-sparse binary vectors β when $n \le ck \log p$ for sufficiently small constant c. By drawing a connection with a similar OGP exhibited by many randomly generated constraint satisfaction problems and statistical physics models, we conjecture that OGP is the source of algorithmic hardness of solving the minimization problem $\min_\beta \|Y - X\beta\|_2$ in the regime $n < n_{\text{LASSO/CS}}$.

Keywords: Linear regression, High-dimensional inference, Second moment method, Phase transitions.

1. Introduction

We consider a high-dimensional linear regression model of the form $Y = X\beta^* + W$ where X is $n \times p$ matrix, W is $n \times 1$ noise vector, and β^* is $p \times 1$ vector of regression coefficients to be recovered from observing X and Y. A great body of literature is devoted to the problem of identifying the underlying regression vector β^*, assuming its support size (the number of coordinates with non-zero coefficients) k is sufficiently small. The support recovery problem has attracted a lot of attention in recent years, because it naturally arises in many contexts including signal denoising Chen et al.

(2001) and compressive sensing Candes and Tao (2005),Donoho (2006). In this paper we assume that X has i.i.d. standard normal entries, W has i.i.d normal entries with standard deviations σ, and β^* is a binary vector (all entries are either zero or one). The results in the existing literature discussed below are adopted to this setting.

A lot of work has been devoted in particular to finding computationally efficient ways for recovering the support of β^*. In the noiseless setting ($W = 0$), Donoho and Tanner (2006) show that the simple linear program (named Basis Pursuit): $\min ||\beta||_1$ subject to $Y = X\beta$, will have with high probability (w.h.p.) β^* as its optimal solution if $n \geq 2(1+\epsilon)k \log p$. Here and below $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the standard ℓ_1 and ℓ_2 norms, respectively: $\|x\|_1 = \sum_{1\leq i\leq p} |x_i|$ and $\|x\|_2 = \left(\sum_{1\leq i\leq p} x_i^2\right)^{\frac{1}{2}}$ for every $x \in \mathbb{R}^p$. In the noisy setting, sufficient and necessary conditions have been found so that the ℓ_1- constrained quadratic programming, also known as LASSO: $\min_{\beta\in\mathbb{R}^p}\{||Y-X\beta||_2^2+\lambda_p||\beta||_1\}$, for appropriately chosen $\lambda_p > 0$, recovers the correct support of β^*, Meinshausen and Bhlmann (2006),Wainwright (2009b),Zhao and Yu (2006). See also the recent book Foucart and Rauhut (2013). In particular, Wainwright (2009b) showed that if X is a Gaussian random matrix and W is a Gaussian noise vector with variance σ^2 such that $\frac{\sigma^2}{k} \to 0$, then for every arbitrarily small constant $\epsilon > 0$ and for $n > (1+\epsilon)(2k+\sigma^2)\log p$, the LASSO method recovers the support of β^* exactly w.h.p. At the same time given any $\epsilon > 0$, if $n < (1-\epsilon)(2k+\sigma^2)\log p$, then LASSO method provably fails to recover the support of β^* exactly, also w.h.p. Finally, orthogonal matching pursuit, a simple and popural greedy algorithm, has also been proven to work given again that σ^2 satisfies $\frac{\sigma^2}{k} \to 0$ and $n > (1+\epsilon)(2k+\sigma^2)\log p$, Fletcher and Rangan (2009). We note that the impact of σ^2 on this threshold value is asymptotically negligible when $\sigma^2/k \to 0$. It will be convenient for us to maintain it though. Thus we denote $(2k+\sigma^2)\log p$ by $n_{\text{LASSO/CS}}$. At the present time no tractable (polynomial time) algorithms are known for the support recovery when $n \leq n_{\text{LASSO/CS}}$.

On the complimentary direction, results regarding the information theoretic limits for the problem of support recovery have also been obtained Donoho and Tanner (2006),Wainwright (2009a),Wang et al. (2010). These papers are devoted to obtaining bounds on the minimum sampling size n so that the support recovery problem is solvable by any algorithmic methods, regardless of the algorithmic complexity, including for example the brute force method of exhaustive search. An easy corollary of Theorem 2 in Wainwright (2009a), when applied to our context below involving vectors β^* with binary values, shows that if $n < (1-\epsilon)\sigma^2\log p$, then for every support recovery algorithm, a binary vector β^* can be constructed in such a way that the underlying algorithm fails to recover β^* exactly, with probability at least $\frac{\epsilon}{2}$. Viewing the problem from the Gaussian channel perspective, vector Y can be viewed as a noisy encoding of β^* through the code book X and in our case the sparsity k becomes the strength of this Gaussian channel. Then when $k = 1$, the information theoretic limit of recovering the unit bit support of β^* is $\log p/\log(1+1/\sigma^2)$ which is $\sigma^2\log p$ asymptotically when σ is large. We let $n_{\text{inf},1} \triangleq \sigma^2\log p$. Subsequently, it was shown by Wang et al Wang et al. (2010) that the exact recovery of β^* is information theoretically impossible when n is smaller than $n^* \triangleq 2k\log p/\log(1+2k/\sigma^2)$, where n^* is the information theoretic limit of this Gaussian channel for general k. The critical threshold n^* will play a fundamental role in our paper. We note that the result above does not preclude the possibility of the existence of an algorithm which for $n < n^*$ recovers some portion of the support of β^* and this question is one of the motivating questions for the present work.

Based on the above discussion the regime $n \in [n_{\text{inf},1}, n_{\text{LASSO/CS}}]$ remains largely unexplored from the algorithmic perspective, and the present paper is devoted to studying this regime. Towards this goal, for the regression model $Y = X\beta^* + W$, we consider the corresponding maximum likelihood estimation problem:

$$
\begin{array}{lll}
(\Phi_2) & \min & n^{-\frac{1}{2}}\|Y - X\beta\|_2 \\
 & \text{s.t.} & \beta \in \{0,1\}^p \\
 & & \|\beta\|_0 = k,
\end{array}
$$

where $\|\beta\|_0$ is the sparsity of β. Namely, it is the cardinality of the set $\{i \in [p] \big| \beta_i \neq 0\}$. We denote by ϕ_2 its optimal value and by β_2 the unique optimal solution. As above, the matrix X is assumed to have i.i.d. standard normal entries, the elements of the noise vector W are assumed to have i.i.d. zero mean normal entries with variance σ^2, and the vector β^* is assumed to be binary k-sparse; $\|\beta^*\|_0 = k$. In particular, we assume that the sparsity k is known to the optimizer. The normality of the entries of X is not an essential assumption for our results, since the Central Limit Theorem based estimates can be easily used instead. We adopt however the normality assumption for simplicity. The normality of the entries of W is more crucial, since our large deviation estimates arising in the application of the conditional second moment method depend on this assumption. It is entirely possible though that similar results are derivable by applying the large deviations estimates for the underlying distribution of entries of Y in the general case.

We address two questions in this paper: (a) What is the value of the squared error estimator $\min_\beta \|Y - X\beta\|_2 = \|Y - X\beta_2\|_2$; and (b) how well does the optimal vector β_2 approximate the ground truth vector β^*?

Our problem setup, including the assumption that β^* is binary, has several motivations. From the application perspective, binary regression coefficients model naturally an idealistic version of the feature selection problem, where weights of the coefficients are not important. From the theoretical perspective, the gap between the information theoretic and algorithmic bounds is particularly profound when β^* is binary. Observe, for example, that in the noiseless setting ($W = 0$), given that β^* is binary, even one sample ($n = 1$) is sufficient to recover β^* by brute force search, whereas $n_{\text{LASSO/CS}} = (2k + \sigma^2)\log p$.

The optimization problem Φ_2 is naturally hard algorithmically since it involves a combinatorial constraint $\|\beta\|_0 = k$. At the same time, it can be cast as an integer programming optimization problem, and the advances in this area make such problems solvable in many practical settings Bertsimas and Parys. Thus the performance of the optimization problem Φ_2 is still of interest, even though formally, it is not proven to be a tractable algorithmic problem. Note however,that the algorithmic hardness of solving the minimization problem Φ_2 pertains to the worst case instances and does not apply to settings involving randomly generated data, such as our X and Y. In fact, one of the goals of this paper is to shed some light on possible sources of the apparent algorithmic hardness of this problem in the case when X and Y are indeed random.

Results

Towards the goals outlined above we obtain several structural results regarding the optimization problem Φ_2, its optimal value ϕ_2, and its optimal solution β_2. We introduce a new method of analysis based on a certain conditional second moment method. The method will be explained below in high level terms. Using this method we obtain a tight up to a multiplicative constant approximation of the squared error ϕ_2 w.h.p., as parameters p, n, k diverge to infinity, and $n \leq ck \log p$ for a small constant c. Some additional assumptions on p, n and k are needed and will be introduced in the statements of the results. The approximation enables us to reveal interesting structural properties of the underlying optimization problem Φ_2. In particular,

(a) We prove that $n^* = 2k \log p / \log(2k/\sigma^2 + 1)$ which was shown in Wang et al. (2010) to be an information theoretic lower bound for the exact recovery of β^* is the sharp phase transition point with the following "all-or-nothing" property. When n exceeds n^* asymptotically, $(2k)^{-1}\|\beta_2 - \beta^*\|_0 \approx 0$, and when n is asymptotically below n^*, $(2k)^{-1}\|\beta_2 - \beta^*\|_0 \approx 1$. Namely, when $n > n^*$ the recovery of β^* is achievable via solving Φ_2, whereas below n^* the optimization problem Φ_2 "misses" the ground truth vector β^* almost entirely. Since, as discussed above, when $n < n^*$, the recovery of β^* is impossible information theoretically, our result implies that n^* is indeed the information theoretic threshold for this problem. We recall that n^* exceeds asymptotically the asymptotic one-bit ($k = 1$) information theoretic threshold $n_{\text{inf},1} = \sigma^2 \log p$, and is asymptotically below the LASSO/Compressive Sensing threshold $n_{\text{LASSO/CS}} = (2k + \sigma^2) \log p$. We note also that our result improves upon the result of Wainwright Wainwright (2009a), who shows that the recovery of β^* is possible by the brute force search method, though only when n is of the order $O(k \log p)$.

(b) We consider an intermediate optimization problem $\min_\beta \|Y - X\beta\|_2$ when the minimization is restricted to vectors β with $\|\beta - \beta^*\|_0 = 2k\zeta$, for some fixed ratio $\zeta \in [0, 1]$. This is done towards a deeper understanding of the problem Φ_2. We show that the function

$$\Gamma(\zeta) \triangleq \left(2\zeta k + \sigma^2\right)^{\frac{1}{2}} \exp\left(-\frac{\zeta k \log p}{n}\right),$$

is, up to a multiplicative constant, a lower bound on this restricted optimization problem, and in the special case of $\zeta = 0$ and $\zeta = 1$, it is also an upper bound, up to a multiplicative constant. Since Γ is a log-concave function in ζ, returning to part (a) above, this implies that that the squared error of the original optimization problem Φ_2 is w.h.p. $\Gamma(0) = \sigma$ when $n > n^*$, and is w.h.p. $\Gamma(1) = \left(2k + \sigma^2\right)^{\frac{1}{2}} \exp\left(-\frac{k \log p}{n}\right)$ when $n < n^*$, both up to multiplicative constants. We further establish that the function Γ exhibits a phase transition property at all three important thresholds $n_{\text{inf},1}, n^*$ and $n_{\text{LASSO/CS}}$, described pictorially on Figures 1 in the next section. In particular, we prove that when $n > n_{\text{LASSO/CS}}$, $\Gamma(\zeta)$ is a strictly increasing function with minimum at $\zeta = 0$, and when $n < n_{\text{inf},1}$, it is a strictly decreasing function with minimum at $\zeta = 1$. When $n^* < n < n_{\text{LASSO/CS}}$, $\Gamma(\zeta)$ is non-monotonic and achieves the minimum value at $\zeta = 0$, and when $n_{\text{inf},1} < n < n^*$, $\Gamma(\zeta)$ is again non-monotonic and achieves the minimum value at $\zeta = 1$. In the critical case $n = n^*$, both $\zeta = 0$ and $\zeta = 1$ are minimum values of γ.

The results above suggest the following, albeit completely intuitive and heuristic picture, which is based on assuming that the function Γ provides an accurate approximation of the value of ϕ_2. When $n > n_{\text{LASSO/CS}}$, a closer overlap with the ground truth vector β^* allows for lower squared error value (Γ is increasing in ζ). In this case the convex relaxation based methods such as LASSO and Basis Pursuit and greedy algorithms like orthogonal matching pursuit succeed in identifying β^*.

When n is below $n_{\text{LASSO/CS}}$ but above n^*, the optimal solution β_2 of Φ_2 still approximately coincides with β^*, but in this case there is a proliferation of solutions which, while they achieve a sufficiently low squared error value, at the same time have very little overlap with β^*. Considering a cost value below the largest value of the function Γ, we obtain two groups of solutions: those with a "substantial" overlap with β^* and those with a "small" even zero overlap with β^*. This motivates looking at the so-called Overlap Gap Property discussed in (c) below.

When n is below n^*, there are solutions, and in particular the optimal solution β_2, which achieve better squared error value than even the ground truth β^*. This is exhibited by the fact that the minimum value of Γ is achieved at $\zeta = 1$. We are dealing here with the case of overfitting. We note that, while information theoretically it is impossible to precisely recover β^* in this regime, it is not clear whether in this case there exists any algorithm which can recover at least a portion of the support of β^*, algorithmic complexity aside. We leave it as an interesting open question.

When n is below the one-bit information theoretic lower bound $n_{\text{inf},1}$, the overfitting issue is even more profound. Moving further away from β^* allows for better and better squared error values (Γ is decreasing in ζ).

(c) Motivated by the results in the theory of spin glasses and the later results in the context of randomly generated constraint satisfaction problems, and in light of the evidence of the Overlap Gap Property (OGP) discussed above, we consider the solution space geometry of the problem Φ_2, as well as the restricted problem corresponding to the constraint $\|\beta - \beta^*\|_0 = 2\zeta k$. For many models of randomly generated constraint satisfaction problems such as random K-SAT, proper coloring of a sparse random graph, the problem of finding a largest independent subset of a sparse random graph, and many others, it has been conjectured and later established rigorously that solutions achieving near optimality, or solutions satisfying a set of randomly generated constraints, break down into clusters separated by cost barriers of a substantial size in some appropriate sense, Achlioptas et al. (2011), Achlioptas and Coja-Oghlan (2008), Montanari et al. (2011), Coja-Oghlan and Efthymiou (2011), Gamarnik and Sudan (a), Rahman and Virag (2014), Gamarnik and Sudan (b). As a result, these models indeed exhibit the OGP. For example, independent sets achieving near optimal size in sparse random graph exhibit the OGP in the following sense. The intersection of every two such independent sets is either at most some value τ_1 or at least some value $\tau_2 > \tau_1$. This and similar properties were used in Gamarnik and Sudan (a), Rahman and Virag (2014), Gamarnik and Sudan (b) and Coja-Oghlan et al. (2016) to establish a fundamental barriers on the power of so-called local algorithms for solving certain types of random constraint satisfaction problems. The OGP was later established in a setting other than constraint satisfaction problems on graphs, specifically in the context of finding a densest submatrix of a matrix with i.i.d. Gaussian entries Gamarnik and Li (2016).

The non-monotonicity of the function Γ for $n < n_{\text{LASSO/CS}}$ already suggests the presence of the OGP. Note that for any value r strictly below the maximum value $\max_{\zeta \in (0,1)} \Gamma(\zeta)$ we obtain the existence of two values $\zeta_1 < \zeta_2$, such that for every ζ with $\Gamma(\zeta) \leq r$, either $\zeta \leq \zeta_1$ or $\zeta \geq \zeta_2$. Namely, this property suggests that every binary vector achieving a cost at most r either has the overlap at most $\zeta_1 k$ with β^*, or the overlap at least $\zeta_2 k$ with β^*. Unfortunately, this is no more than a guess, since $\Gamma(\zeta)$ provides only a lower bound on the optimization cost. Nevertheless, we establish that the OGP provably takes place w.h.p. when $C\sigma^2 \log p \le n \le ck \log p$, for appropriately large constant C and appropriately small constant c. Our result takes advantage of the tight up to a multiplicative error estimates of the squared errors associated with the restricted optimization problem Φ_2 with the restricted $\|\beta - \beta^*\|_0 = 2k\zeta$, discussed earlier. It remains an intriguing open question to verify whether the optimization problem Φ_2 is indeed algorithmically intractable in this regime.

Methods

In order to obtain estimates of the squared error for the problem Φ_2 we use a novel conditional second moment method, which we now describe in high level terms. We begin with the following model which we call Pure Noise model, in which it is assumed that $\beta^* = 0$ and thus Y is simply a vector of i.i.d. zero mean Gaussian random variables with variance σ^2. For every value $t > 0$ we consider Z_t to be the number of solutions β such that $\|Y - X\beta\|_\infty \le t$, where $\|x\|_\infty = \max_i |x_i|$ is the max norm. It turns out that while $\|\cdot\|_\infty$ norm is easier to deal with, it provides sufficiently accurate estimates for the $\|\cdot\|_2$ norm that we care about. We compute the expected value of Z_t and find a critical value t^* such that for $t < t^*$ this expectation converges to zero. This provides a lower bound on the value of the optimization problem using the Markov inequality. We then consider the second moment of Z_t. In the naive form the second moment method would succeed if for $t > t^*$, $\mathbb{E}\left[Z_t^2\right]$ was close to $(\mathbb{E}[Z_t])^2$, as in this case the Paley-Zigmund inequality would give $\mathbb{P}(Z_t \ge 1) \ge (\mathbb{E}[Z_t])^2/\mathbb{E}\left[Z_t^2\right] \approx 1$ and thus t^* is the true value of the optimization problem under $\|\cdot\|_\infty$ norm. Unfortunately, the naive second moment estimation fails due to fluctuations of Y which alone is enough to create a substantial gap between the two moments of Z_t. Instead, we conditional the conditional first and second moment of Z_t, where we condition on Y. The conditional second moment involves computing large deviations estimates on a sequence of coupled bi-variate normal random variables, where the correlation corresponds to the overlaps of pairs β_1 and β_2 contributing to the second moment. A fairly detailed analysis of this large deviation estimate is obtained to arrive at the estimation of the ratio $\mathbb{E}\left[Z_t^2|Y\right] / (\mathbb{E}[Z_t|Y])^2$. We then essentially use the conditional version of the Paley-Zigmund inequality $\mathbb{P}(Z_t \ge 1|Y) \ge (\mathbb{E}[Z_t|Y])^2 / \mathbb{E}\left[Z_t^2|Y\right]$ to obtain the lower bound on $\mathbb{P}(Z_t \ge 1)$ after unconditioning on Y and this leads to an upper bound on Z_t. The application of the conditional second method requires some "sacrifice" of the constant factor multipliers.

Next we use the estimates from the Pure Noise model, for the original model involving β^* with $\|\beta^*\|_0 = k$, where we consider a restricted problem in which the optimization is conducted over the space of binary k-sparse vectors β which have a fixed intersection of the support with the support of β^*. In this form the problem is reduced to the Pure Noise problem. To the best of our knowledge, this is the first time the conditional second moment is used in the form described above, and this might be of independent interest. The estimation of the value of ϕ_2 for the Pure Noise and for the original model of interest with a fixed β^*, are at the heart of all of our other results discussed above.

The remainder of the paper is organized as follows. The description of the model, assumptions and the main results are found in the next section. We then proceed with a section which explains the main ideas behind the proofs. We end with a section for conclusion and open problems. Appendix A is devoted to the analysis of the Pure Noise model which is also defined in this appendix. Appendix B, C and D are devoted to proofs of our main results.

2. Model and the Main Results

We remind our model for convenience. Let $X \in \mathbb{R}^{n \times p}$ be an $n \times p$ matrix with i.i.d. standard normal entries, and $W \in \mathbb{R}^p$ be a vector with i.i.d. $N\left(0, \sigma^2\right)$ entries. We also assume that β^* is a $p \times 1$ binary vector with exactly k entries equal to unity (β^* is binary and k-sparse). For every binary vector $\beta \in \{0,1\}^p$ we let Support$(\beta) := \{i : \beta_i = 0\}$. Namely, $\beta_i = 1$ if $i \in$ Support(β) and $\beta_i = 0$ otherwise. We observe n noisy measurements $Y \in \mathbb{R}^n$ of the vector $\beta^* \in \mathbb{R}^p$ given by

$$Y = X\beta^* + W \in \mathbb{R}^n.$$

Throughout the paper we are interested in the high dimensional regime where p exceeds n and both diverge to infinity. Various assumptions on k, n, p are required for technical reasons and some of the assumptions vary from theorem to theorem. But almost everywhere we will be assuming that n is at least of the order $k \log k$ and at most a small constant multiplied with $k \log p$, and therefore k should be understood as being at most a small power of p. The results usually hold in the "with high probability" (w.h.p.) sense as k, n and p diverge to infinity, but for concreteness we usually explicitly say that k diverges to infinity. This automatically implies the same for p, since $p \geq k$, and for n since it is assumed to be at least of the order $O(k \log k)$.

In order to recover β^*, we consider the following constrained optimization problem

$$\begin{array}{lll}(\Phi_2) & \min & n^{-\frac{1}{2}}||Y - X\beta||_2 \\ & \text{s.t.} & \beta \in \{0,1\}^p \\ & & ||\beta||_0 = k.\end{array}$$

We denote by $\phi_2 = \phi_2(X, W)$ its optimal value and by β_2 its (unique) optimal solution. Note that the solution is indeed unique due to discreteness of β and continuity of the distribution of X and Y. Namely, the optimization problem Φ_2 chooses the k-sparse binary vector β such that $X\beta$ is as close to Y as possible, with respect to the ℓ_2 norm. Also note that since our noise vector, W, consists of i.i.d. Gaussian entries, β_2 is also the Maximum Likelihood Estimator of β^*.

Consider now the following restricted version of the problem Φ_2:

$$\begin{array}{lll}(\Phi_2(\ell)) & \min & n^{-\frac{1}{2}}||Y - X\beta||_2 \\ & \text{s.t.} & \beta \in \{0,1\}^p \\ & & ||\beta||_0 = k, ||\beta - \beta^*||_0 = 2l,\end{array}$$

where $\ell = 0, 1, 2, .., k$. For every fixed ℓ, denote by $\phi_2(\ell)$ the optimal value of $\Phi_2(\ell)$. $\Phi_2(\ell)$ is the problem of finding the k-sparse binary vector β, such that $X\beta$ is as close to Y as possible with respect to the ℓ_2 norm, but also subject to the restriction that the cardinality of the intersection of the supports of β and β^* is exactly $k - l$. Then $\phi_2 = \min_\ell \phi_2(\ell)$.

Consider for example the extreme cases $\ell = 0$ and $\ell = k$. For $\ell = 0$, the region that defines $\Phi_2(0)$ consists only of the vector β^*. On the other hand, for $\ell = k$, the region that defines $\Phi_2(k)$ consists of all k-sparse binary vectors β, whose common support with β^* is empty.

We are now ready to state our first main result.

Theorem 1 *Suppose $k \log k \leq Cn$ for some constant C for all k, n. Then* *(a) W.h.p. as k increases*

$$\phi_2(\ell) \geq e^{-\frac{3}{2}} \sqrt{2l + \sigma^2} \exp\left(-\frac{\ell \log p}{n}\right), \tag{1}$$

for all $0 \leq \ell \leq k$.

(b) Suppose further that $\sigma^2 \leq 2k$. Then for every sufficiently large constant D_0 if $n \leq k \log p/(3 \log D_0)$, then w.h.p. as k increases, the cardinality of the set

$$\left\{\beta : \|\beta - \beta^*\|_\infty = 2k,\ n^{-\frac{1}{2}} \|Y - X\beta\|_2 \leq D_0 \sqrt{2k + \sigma^2} \exp\left(-\frac{k \log p}{n}\right)\right\} \tag{2}$$

is at least $D_0^{\frac{n}{3}}$. In particular, this set is exponentially large in n.

The proof of this theorem is found in Appendix B and relies on the analysis for the Pure Noise model developed in the Appendix. The part (a) of the theorem above gives a lower bound on the optimal value of the optimization problem $\Phi_2(\ell)$ for all $\ell = 0, 1, \ldots, k$ w.h.p. For this part, as stated, we only need that $k \log k \leq Cn$ and k diverging to infinity. When $\ell = 0$ the value of $\phi(\ell)$ is just $n^{-\frac{1}{2}} \sqrt{\sum_{1 \leq i \leq n} W_i^2}$ which converges to σ by the Law of Large Numbers. Note that σ is also the value of $\sqrt{2l + \sigma^2} \exp\left(-\frac{\ell \log p}{n}\right)$ when $\ell = 0$. Thus the lower bound value in part (a) is tight up to a multiplicative constant when $\ell = 0$. Importantly, as the part (b) of the theorem shows, the lower bound value is also tight up to a multiplicative constant when $\ell = k$, as in this case not only vectors β achieving this bound exist, but the number of such vectors is exponentially large in n w.h.p. as k increases. This result will be instrumental for our "all-or-nothing" Theorem 3 below.

Now we will discuss some implications of Theorem 1. The expression $\left(2\ell + \sigma^2\right)^{\frac{1}{2}} \exp\left(-\frac{\ell \log p}{n}\right)$, appearing in the theorem above, motivates the following notation. Let the function $\Gamma : [0, 1] \to \mathbb{R}_+$ be defined by

$$\Gamma(\zeta) = \left(2\zeta k + \sigma^2\right)^{\frac{1}{2}} \exp\left(-\frac{\zeta k \log p}{n}\right). \tag{3}$$

Then the lower bound (1) can be rewritten as

$$\phi_2(\ell) \geq e^{-\frac{3}{2}} \Gamma(\ell/k).$$

A similar inequality applies to (2).

Let us make some immediate observations regarding the function Γ. It is a strictly log-concave function in $\zeta \in [0, 1]$:

$$\log \Gamma(\zeta) = \frac{1}{2} \log\left(2\zeta k + \sigma^2\right) - \zeta \frac{k \log p}{n}.$$

and hence

$$\min_{0 \leq \zeta \leq 1} \Gamma(\zeta) = \min\left(\Gamma(0), \Gamma(1)\right) = \min\left(\sigma, \sqrt{2k + \sigma^2} \exp\left(-\frac{k \log p}{n}\right)\right).$$

Now combining this observation with the results of Theorem 1 we obtain as a corollary a tight up to a multiplicative constant approximation of the value ϕ_2 of the optimization problem Φ_2.

Theorem 2 *Under the assumptions of parts (a) and (b) of Theorem 1, for every $\epsilon > 0$ and for every sufficiently large constant D_0 if $n \leq k \log p/(3 \log D_0)$, then w.h.p. as k increases,*

$$e^{-\frac{3}{2}} \min\left(\sigma, \sqrt{2k+\sigma^2} \exp\left(-\frac{k \log p}{n}\right)\right) \leq \phi_2 \ \leq \min\left((1+\epsilon)\sigma, D_0\sqrt{2k+\sigma^2} \exp\left(-\frac{k \log p}{n}\right)\right).$$

Proof By Theorem 1 we have that Φ_2 is at least

$$e^{-\frac{3}{2}} \min_{\zeta} \Gamma(\zeta) = e^{-\frac{3}{2}} \min\left(\Gamma(0), \Gamma(1)\right).$$

This establishes the lower bound. For the upper bound we have $\phi_2 \leq \min(\phi_2(0), \phi_2(k))$. By the Law of Large Numbers, $\phi_2(0)$ is at most $(1+\epsilon)\sigma$ w.h.p. as k (and therefore n) increases. The second part of Theorem1 gives provides the necessary bound on $\Phi_2(k)$. ■

As in the introduction, letting $n^* = \frac{2k \log p}{\log\left(\frac{2k}{\sigma^2}+1\right)}$, we conclude that $\min_\zeta \Gamma(\zeta) = \Gamma(1)$ when $n < n^*$ and $= \Gamma(0)$ when $n > n^*$, with the critical case $n = n^*$ (ignoring the integrality of n^*), giving $\Gamma(0) = \Gamma(1)$. This observation suggests the following **"all-or-nothing"** type behavior of the problem Φ_2, if Γ was an accurate estimate of the value of the optimization problem Φ_2. When $n > n^*$ the solution β_2 of the minimization problem Φ_2 is expected to coincide with the ground truth β^* since in this case $\zeta = 0$, which corresponds to $\ell = 0$, minimizes $\Gamma(\zeta)$. On the other hand, when $n < n^*$, the solution β_2 of the minimization problem Φ_2 is not even expected to have any common support with the ground truth β^*, as in this case $\zeta = 1$, which corresponds to $\ell = k$, minimizes $\Gamma(\zeta)$. Of course, this is nothing more than just a suggestion, since by Theorem 1, $\Gamma(\zeta)$ only provides a lower and upper bounds on the optimization problem Φ_2, which tight only up to a multiplicative constant. Nevertheless, we can turn this observation into a theorem, which is our second main result.

Theorem 3 *Let $\epsilon > 0$ be arbitrary. Suppose $k \log k \leq Cn$, for some $C > 0$ for all k and n. Suppose furthermore that $k \to \infty$ and $\sigma^2/k \to 0$ as $k \to \infty$. If $n \geq (1+\epsilon)\, n^*$, then w.h.p. as k increases*

$$\frac{1}{2k}\|\beta_2 - \beta^*\|_0 \to 0.$$

On the other hand if $n \leq (1-\epsilon)\, n^*$, then w.h.p. as k increases

$$\frac{1}{2k}\|\beta_2 - \beta^*\|_0 \to 1.$$

The proof of Theorem 3 is found in Appendix C. The theorem above confirms the "all-or-nothing" type behavior of the optimization problem Φ_2, depending on how n compares with n^*. Recall that, according to Wang et al. (2010), n^* is an information theoretic lower bound for recovering β^* from X and Y *precisely*, and also for $n < n^*$ it does not rule out the possibility of recovering at least a fraction of bits of β^*. Our theorem however shows firstly that n^* is exactly the infortmation theoretic threshold for exact recovery and also that if $n < n^*$ the optimization problem Φ_2 fails to recover asymptotically any of the bits of β^*. We note also that the value of n^* is naturally larger than the corresponding threshold when $k = 1$, namely $2 \log p/ \log(1 + 2\sigma^{-2})$, which is asymptotically

$\sigma^2 \log p = n_{\inf,1}$. Interestingly, however the value of this weaker information theoretic bound also marks a phase transition point as we discuss in the proposition below.

As our result above shows, the recovery of β^* is possible by solving Φ_2 (say by running the integer programming problem) when $n > n^*$, even though efficient algorithms such as Compressive sensing techniques and LASSO algorithms are only known to work when $n \geq (2k+\sigma^2)\log p$. This suggests that the region $n \in [n^*, (2k+\sigma^2)\log p]$ might correspond to solvable but algorithmically hard regime for the problem of finding β^*. Studying the properties of the limiting curve $\Gamma(\zeta)$ we discover an intriguing link between its behavior and the three fundamental thresholds discussed above. Namely, the threshold $n_{\inf,1} = \sigma^2 \log p$, the threshold $n^* = \frac{2k}{\log\left(\frac{2k}{\sigma^2}+1\right)} \log p$, and finally the threshold $n_{\text{LASSO/CS}} = (2k+\sigma^2)\log p$. For the illustration of different cases outlined in the proposition above see Figure 1.

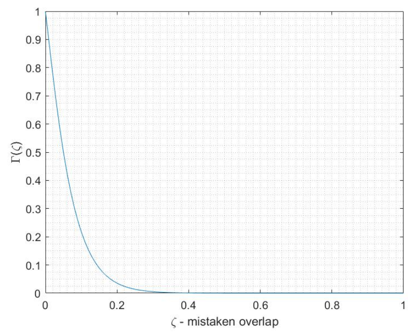
Proposition 4 *The function Γ satisfies the following properties.*

1. *When $n \leq \sigma^2 \log p$, Γ is a strictly decreasing function of ζ. (Figure 1(a)),*
2. *When $\sigma^2 \log p < n < n^*$, Γ is not monotonic and it attains its minimum at $\zeta = 1$. (Figure 1(b)),*
3. *When $n = n^*$, Γ is not monotonic and it attains its minimum at $\zeta = 0$ and $\zeta = 1$. (Figure 1(c))*
4. *When $n^* < n < (2k+\sigma^2)\log p$, Γ is not monotonic and it attains its minimum at $\zeta = 0$. (Figure 1(d))*
5. *When $n > \left(2k+\sigma^2\right)\log p$, Γ is a strictly increasing function of ζ. (Figure 1(d))*

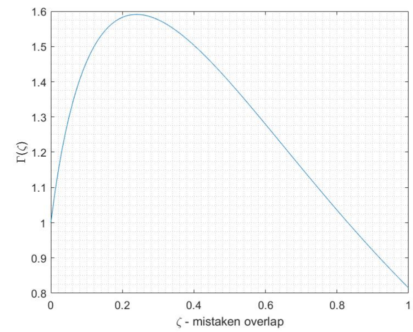
In particular, we see that both the bound $n_{\inf,1} = \sigma^2 \log p$, and $n_{\text{LASSO/CS}} = \left(2k+\sigma^2\right)\log p$ mark the phase transition change of (lack of) monotonicity property of the limiting curve Φ_2. We also summarize our findings in Table 1. The proof of this proposition is found in Appendix C.

To get an insight into possible reason for the apparent algorithmic hardness of the problem in the regime $n \in [n_{\inf,1}, n_{\text{LASSO/CS}}]$, we as well as to see whether the picture suggested by the curve Γ is actually accurate, we now turn to the geometry of the solution space of the problem Φ_2. We establish in particular, that the solutions β which are sufficiently "close" to optimality break into two separate clusters – those which are close in $\|\cdot\|_0$ norm to the optimal solution β_2, namely those which have a "large" overlap with β_2, and those which are far from it, namely those which have a "small" overlap with β_2. As discussed in Introduction, such an Overlap Gap Property (OGP) appears to mark the onset of algorithmic hardness for many randomly generated constraint satisfaction problems. Here we demonstrate its presence in the context of high dimensional regression problems.

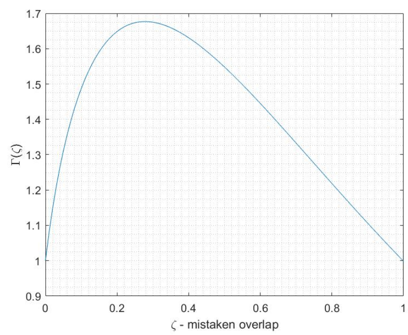
The presence of the OGP is indeed suggested by the lack of monotonicity of the limiting curve Γ when $\sigma^2 \log p < n < (2k+\sigma^2)\log p$. Indeed, in this case fixing any value γ strictly smaller than the maximum value of Γ, but larger than both $\Gamma(0)$ and $\Gamma(1)$, we see that set of overlaps ζ achieving value $\leq \gamma$ is disjoint union of two intervals of the form $[0, \zeta_1]$ and $[\zeta_2, 1]$ with $\zeta_1 < \zeta_2$. Of course, as before this is nothing but a suggestion, since the function Γ is only a lower bound on the objective value $\Phi_2(\ell)$ for $\zeta = \ell/k$. In the next theorem we establish that the OGP indeed takes



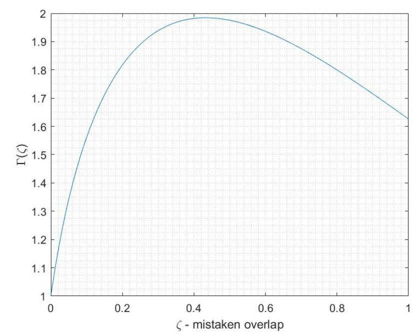
(a) The behavior of Γ if $n = 10 < \sigma^2 \log p$.



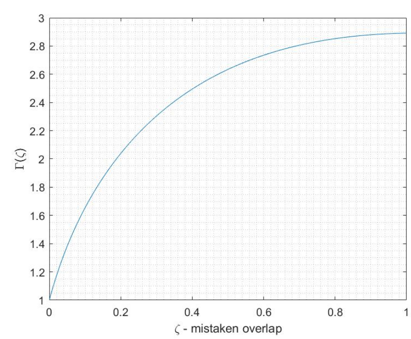
(b) The behavior of Γ for $\sigma^2 \log p < n = 120 < n^*$.



(c) The behavior of Γ for $n = 136 = n^*$.



(d) The behavior of Γ for $n^* < n = 200 < (2k + \sigma^2) \log p$.



(e) The behavior of Γ for $(2k + \sigma^2) \log p < n = 450$.

Figure 1: The five different phases of the function Γ as n grows. We consider the case when $p = 10^9, k = 10$ and $\sigma^2 = 1$. In this case $\lceil \sigma^2 \log p \rceil = 21, \lceil n^* \rceil = 137$ and $\lceil (2k + \sigma^2) \log p \rceil = 435$.

place, when the sampling size is bounded away by a constant from $n_{\text{inf},1}$ and $n_{\text{LASSO/CS}}$. Given any

$n < n_{\text{inf},1}$	Γ is monotonically decreasing
$n_{\text{inf},1} < n < n^*$	Γ is not monotonic and attains its minimum at $\zeta = 1$
$n^* < n < n_{\text{LASSO/CS}}$	Γ is not monotonic and attains its minimum at $\zeta = 0$
$n_{\text{LASSO/CS}} < n$	Γ is monotonically increasing

Table 1: The phase transition property of the limiting curve $\Gamma(\zeta)$

$r \geq 0$, let

$$S_r := \{\beta \in \{0,1\}^p : ||\beta||_0 = k, n^{-\frac{1}{2}}||Y - X\beta||_2 < r\}.$$

Theorem 5 (The Overlap Gap Property) *Suppose the assumptions of Theorem 1 hold. Suppose in addition $\sigma^2 \to +\infty$. For every sufficiently large constant D_0 there exist sequences $0 < \zeta_{1,k,n} < \zeta_{2,k,n} < 1$ satisfying*

$$\lim_{k \to \infty} k\left(\zeta_{2,k,n} - \zeta_{1,k,n}\right) = +\infty,$$

as $k \to \infty$, and such that if $r_k = D_0 \max\left(\Gamma(0), \Gamma(1)\right)$ and $\left(e^7 D_0^2 + 1\right)\sigma^2 \log p \leq n \leq k \log p/(3 \log D_0)$ then w.h.p. as k increases the following holds

(a) For every $\beta \in S_{r_k}$

$$(2k)^{-1}\|\beta - \beta^*\|_0 < \zeta_{1,k,n} \text{ or } (2k)^{-1}\|\beta - \beta^*\|_0 > \zeta_{2,k,n}.$$

(b) $\beta^* \in S_{r_k}$. In particular the set

$$S_{r_k} \cap \{\beta : (2k)^{-1}\|\beta - \beta^*\|_0 < \zeta_{1,k,n}\}$$

is non-empty.

(c) The cardinality of the set

$$|S_{r_k} \cap \{\beta : \|\beta - \beta^*\|_0\} = 2k\}|$$

is at least $D_0^{\frac{n}{3}}$. In particular the set $S_{r_k} \cap \{\beta : \|\beta - \beta^*\|_0\} = 2k\}$ has exponentially many in n elements.

The proof of Theorem 5 is found in Appendix D. The property $k\left(\zeta_{2,k,n} - \zeta_{1,k,n}\right) \to \infty$ in the statement of the theorem implies in particular that the difference $\left(\zeta_{2,k,n} - \zeta_{1,k,n}\right)$ grows faster than $1/k$ as k diverges, ensuring that for many overlap values ℓ, the ratio $2\ell/k$ falls within the interval $[\zeta_{1,k,n}, \zeta_{2,k,n}]$. Namely, the overlap gap interval is non-vacuous for all large enough k.

We note that the result above is established all the way down to n of the order $O(\sigma^2 \log p)$, even though its algorithmic significance below the information theoretic bound n^* is not clear. Nevertheless, this result might be of independent interest.

3. The Pure Noise Model and Outline of the Key Proofs

Our key step towards establishing Theorem 1, which implies the rest of our main results, is proving a, possibly of independent interest, property for a different regression model where $\beta^* = 0$ and the noise variance σ^2 satisfies $\sigma^2 \leq 3k$. In this subsection, we are going to state this new result, provide a sketch of it's proof and describe why it naturally implies Theorem 1.

3.1. The Pure Noise Model

We call the new model, the Pure Noise Model. To define it properly let $X \in \mathbb{R}^{n\times p}$ be an $n \times p$ matrix with i.i.d. standard normal entries, and $Y \in \mathbb{R}^n$ be a vector with i.i.d. $N\left(0,\sigma^2\right)$ entries. Y, X are independent. We study the optimal value ψ_2 of the following optimization problem:

$$
\begin{array}{lll}
(\Psi_2) & \min & n^{-\frac{1}{2}}||Y - X\beta||_2 \\
& \text{s.t.} & \beta \in \{0,1\}^p \\
& & ||\beta||_0 = k.
\end{array}
$$

That is, as claimed above we no longer have ground truth vector β^*, and instead we search for a binary k-sparse vector β which makes $X\beta$ as close to an independent vector Y as possible in $\|\cdot\|_2$ norm.

Our key result is a tight up-to-constants approximation of the value ψ_2 (under certain assumptions on the parameters n, p, k, σ^2):

$$
\psi_2 \sim \sqrt{k+\sigma^2}\exp\left(-\frac{k\log p}{n}\right), \tag{4}
$$

as n grows. An explicit statement of the result can be found in Theorem 6, in Appendix A.

3.2. Why Approximation 4 holds

Given any $t > 0$, let

$$
Z_t = \sum_{\beta \in \{0,1\}^p|,|\beta||_0 = k} \mathbf{1}\left(n^{-\frac{1}{2}}||Y - X\beta||_2 < t\right).
$$

Equivalent with approximation 4 is to show $t^* = \sqrt{k+\sigma^2}\exp\left(-\frac{k\log p}{n}\right)$ is -up to constants- the minimum t such that $Z_t \geq 1$ with high probability.

Using Markov inequality and standard χ^2-large deviation results we can deduce indeed that if $t < e^{-\frac{3}{2}}t^*$ then w.h.p. $Z_t = 0$ (for details see Appendix A.1.)

To prove the other direction, the natural step is to use Paley-Zigmund inequality $\mathbb{P}(Z_t \geq 1) \geq (\mathbb{E}[Z_t])^2/\mathbb{E}\left[Z_t^2\right]$ and show that if $t > Ct^*$ for some constant $C > 0$, then $(\mathbb{E}[Z_t])^2/\mathbb{E}\left[Z_t^2\right] \to 1$. Unfortunately, this limiting behavior for the ratio $(\mathbb{E}[Z_t])^2/\mathbb{E}\left[Z_t^2\right]$ is not true due to fluctuations of the random variable Y. Given this, we consider a novel conditional version of the second moment method according to which, it is enough to be shown that for

$$
Z_{t,\infty} = \sum_{\beta \in \{0,1\}^p|,|\beta||_0 = k} \mathbf{1}\left(||Y - X\beta||_\infty < t\right).
$$

the ratio $(\mathbb{E}[Z_{t,\infty}|Y])^2/\mathbb{E}\left[Z_{t,\infty}^2|Y\right]$ is close to 1, with high probability with respect to Y. The change from the ℓ_2 norm to the ℓ_∞ norm is there to make the calculations easier and it is shown not to weaken the approximation more than the level of constants (Details can be found in Appendix A.4.)

Showing that $(\mathbb{E}[Z_{t,\infty}|Y])^2/\mathbb{E}\left[Z_{t,\infty}^2|Y\right]$ is close to 1, with high probability with respect to Y, requires some effort and is the most technical part of the proof. We first expand the square $Z_{t,\infty}^2$ to get

$$\mathbb{E}\left[Z_{t,\infty}^2|Y\right]/(\mathbb{E}[Z_{t,\infty}|Y])^2 = \sum_{\ell=0}^{k} \frac{\binom{p}{k-l,k-l,l,p-2k+l}}{\binom{p}{k}^2} \prod_{i=1}^{n} \frac{q_{t,\frac{Y_i}{\sqrt{k}},\frac{l}{k}}}{p^2_{t,\frac{Y_i}{\sqrt{k}}}},$$

where for $t > 0, y \in \mathbb{R}, \rho \in [0,1)$ we define

$$p_{t,y} = \mathbb{P}\left(|Z - y| \leq t\right),$$

where Z is a standard normal random variable and

$$q_{t,y\rho} = \mathbb{P}\left(|Z_1 - y| \leq t, |Z_2 - y| \leq t\right),$$

where the random pair (Z_1, Z_2) follows a bivariate normal distribution with correlation ρ.

We then show that the summand corresponding to $\ell = 0$ tends to 1 and all the other tend to 0, both in probability with respect to Y. The latter is established using a lemma upper which upper bounds the fraction $\frac{q_{t,y,\rho}}{p^2_{t,y}}$ for any $t > 0, y \in \mathbb{R}, \rho \in [0,1)$ (Lemma 9 in Appendix A.2.) and using various large deviation type estimates. (The complete proof can be found in Appendix A.3.)

3.3. Implications for Theorem 1

In order to see why bounds on ψ_2 help us prove bounds on $\phi_2(\ell), \ell = 0, 1, \ldots, k$ observe the following. For any $S \subseteq \text{Support}(\beta^*)$ consider the optimization problem $(\Phi_2(S))$:

$$\begin{array}{lll} (\Phi_2(S)) & \min & n^{-\frac{1}{2}}||Y - X\beta||_2 \\ & \text{s.t.} & \beta \in \{0,1\}^p \\ & & ||\beta||_0 = k, \text{Support}(\beta) \cap \text{Support}(\beta^*) = S, \end{array}$$

and let $\phi_2(S)$ be its optimal value. It is easy to see that $\phi_2(\ell) = \min_{S \subset [p], |S| = \ell} \phi_2(S)$. Notice now that for a binary k-sparse β with $\text{Support}(\beta) \cap \text{Support}(\beta^*) = S$ we have (details in Appendix B):

$$Y - X\beta = Y' - X'\beta_1,$$

where we define Y', X', β_1 as following:

1. $X' \in \mathbb{R}^{n \times (p-k)}$ to be the matrix which is X after deleting the columns corresponding to $\text{Support}(\beta^*)$
2. $Y' := \sum_{i \in \text{Support}(\beta^*) - S} X_i + W$
3. $\beta_1 \in \{0,1\}^{p-k}$ is obtained from β after deleting coordinates in $\text{Support}(\beta^*)$. Notice that $||\beta_1||_0 = k - |S|$.

Hence, solving $\Phi_2(S)$ can be written equivalently with respect to Y', X', β' as following,

$$\begin{array}{lll} (\Phi_2(S)) & \min & n^{-\frac{1}{2}}||Y' - X'\beta'||_2 \\ & \text{s.t.} & \beta' \in \{0,1\}^{p-k} \\ & & ||\beta'||_0 = k - |S|. \end{array}$$

This has the form of Ψ_2 for sparsity level $k-|S|$ and feature number $p-k$ which implies by analogy with 4 that under certain assumptions on $|S|$ w.h.p. $\phi_2(S) \sim \sqrt{2(k-|S|)+\sigma^2}\exp\left(-\frac{(k-|S|)\log(p-k)}{n}\right)$. This, based on our parameters assumptions, using standard methods help us to conclude that for all ℓ w.h.p.,

$$\phi_2(\ell) \gtrsim \sqrt{2(k-\ell)+\sigma^2}\exp\left(-\frac{(k-\ell)\log p}{n}\right) = \Gamma\left(1-\frac{\ell}{k}\right),$$

which is the first part of Theorem 1 and that for $\ell = 0$ and $\ell = k$ it holds as well w.h.p.

$$\phi_2(\ell) \lesssim \sqrt{2(k-\ell)+\sigma^2}\exp\left(-\frac{(k-\ell)\log p}{n}\right) = \Gamma\left(1-\frac{\ell}{k}\right),$$

which is the second part of Theorem 1. Details and the complete proof can be found in Appendix B.

4. Conclusions and Open Questions

Our paper prompts several new directions for research. Relaxing the assumption that regression coefficients are binary is a natural first step in extending the results of this paper. We believe that both the general picture and the main approach should prevail in a general setting where $\min_{i=1,2,\ldots,p}|\beta_i^*| \geq 1$. An appropriate discretization of the coefficients of the regression vector values might be a viable approach towards proving it. Furthermore, it would be interesting to see if the conditional second moment approach proposed in this paper can be used to obtain squared error associated with the relaxation of the problem such as LASSO and the Compressive Sensing methods.

An interesting question is to see as to what extent n^* is indeed the information theoretic limit for the problem of recovery of β^* in a strong sense. As per the results of Wang et al. (2010), the application of the Gaussian channel estimates imply that below this threshold the precise recovery of β^* is impossible information theoretically. However, it is not ruled out that it might be possible to recover at least a portion of the support of β^*. Our results show that the method based on minimizing the squared error is a poor help for this problem as the optimal solution β_2 misses the support almost completely. But it is not ruled out that some other method is capable of recovering at least some positive fraction of the support of β^*. We conjecture that this is not the case and that below n^* the recovery of β^* is impossible in the very strong sense that even obtaining a fraction of support β^* is not possible information theoretically. Similarly, motivated by the fact that $n_{\text{inf},1}$ is an asymptotic information theoretic lower bound when $k = 1$ and σ grows, it would be interesting to see if the recovery of any part of the support of β^* is possible when $n < n_{\text{inf},1}$.

Our results apply to the case when the sampling size n is essentially of the order $o(k\log p)$ (though a small constant in front of $k\log p$ is allowed). Obtaining estimates of the squared error for the regime between $o(k\log p)$ and the LASSO/Compressive Sensing threshold $n_{\text{LASSO/CS}} = (2k+\sigma^2)\log p$ is of interest. This appears to be a difficult regime, as in this case the gap between the conditional first and second moment widens as n approaches the order $O(k \log p)$. It is possible that non-rigorous methods of Replica Symmetry Breaking might be of help here to obtain at least good predictions for the answers. Such predictions are available in the regime when k, n and p are of the same order Bayati and Montanari (2011),Zheng et al. (2015).

Last but not the least, understanding the algorithmic complexity of the problem of finding β^* when n is between n^* and $n_{\text{LASSO/CS}}$ is of great interest. The Overlap Gap Property established in this paper suggests that the problem might indeed be algorithmically hard, though such formal hardness results are lacking even for random constraint satisfaction problems for which the OGP was known already for a long time. On the other hand, it is often observed that for random constraint satisfaction problems outside the regime where OGP takes place, even very naive local algorithms are successful. By drawing an analogy between this class of problems and the problems of high dimensional regression, it is possible that above say threshold $n_{\text{LASSO/CS}}$ some version of a local search algorithm is successful in recovering the regression vector β^*. Similarly, it would interesting to establish that the OGP ceases to exist above the threshold $n_{\mathrm{LASSO/CS}}$.

References

D. Achlioptas, A. Coja-Oghlan, and F. Ricci-Tersenghi. On the solution space geometry of random formulas. *Random Structures and Algorithms*, 38:251–268, 2011.

Dimitris Achlioptas and Amin Coja-Oghlan. Algorithmic barriers from phase transitions. In *Foundations of Computer Science, 2008. FOCS'08. IEEE 49th Annual IEEE Symposium on*, pages 793–802. IEEE, 2008.

Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

D. Bertsimas and Bart Van Parys. Sparse high dimensional regression: Exact scalable algorithms and phase transitions. *Preprint*.

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

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`.

A. Coja-Oghlan and C. Efthymiou. On independent sets in random graphs. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 136–144. SIAM, 2011.

Amin Coja-Oghlan, Amir Haqshenas, and Samuel Hetterich. Walksat stalls well below the satisfiability threshold. *arXiv preprint arXiv:1608.00346*, 2016.

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

David L Donoho and Jared Tanner. Counting faces of randomly-projected polytopes when then projection radically lowers dimension. 2006.

Alyson K. Fletcher and Sundeep Rangan. Orthogonal matching pursuit from noisy measurements: A new analysis. In *Proceedings of the 22Nd International Conference on Neural Information Processing Systems*, NIPS'09, pages 540–548, USA, 2009. Curran Associates Inc. ISBN 978-1-61567-911-9. URL http://dl.acm.org/citation.cfm?id=2984093.2984154.

Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Springer, 2013.

David Gamarnik and Quan Li. Finding a large submatrix of a gaussian random matrix. *arXiv preprint arXiv:1602.08529*, 2016.

David Gamarnik and Madhu Sudan. Limits of local algorithms over sparse random graphs. *Annals of Probability*. To appear, a.

David Gamarnik and Madhu Sudan. Performance of sequential local algorithms for the random nae-k-sat problem. *SIAM Journal on Computing*. To appear, b.

Nicolai Meinshausen and Peter Bhlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 06 2006. doi: 10.1214/009053606000000281. URL http://dx.doi.org/10.1214/009053606000000281.

Andrea Montanari, Ricardo Restrepo, and Prasad Tetali. Reconstruction and clustering in random constraint satisfaction problems. *SIAM Journal on Discrete Mathematics*, 25(2):771–808, 2011.

Mustazee Rahman and Balint Virag. Local algorithms for independent sets are half-optimal. *arXiv preprint arXiv:1402.0485*, 2014.

A. Shwartz and A. Weiss. *Large deviations for performance analysis*. Chapman and Hall, 1995.

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

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, 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.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, December 2006. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1248547.1248637.

Le Zheng, Arian Maleki, Xiaodong Wang, and Teng Long. Does $\backslash ell\_p$-minimization outperform $\backslash ell\_1$-minimization? *arXiv preprint arXiv:1501.03704*, 2015.

Appendix A. The Pure Noise Model

In this appendix we consider a modified model corresponding to the case $\beta^* = 0$, which we dub as pure noise model. This model serves as a technical building block towards proving Theorem 1. The model is described as follows.

The Pure Noise Model

Let $X \in \mathbb{R}^{n \times p}$ be an $n \times p$ matrix with i.i.d. standard normal entries, and $Y \in \mathbb{R}^n$ be a vector with i.i.d. $N\left(0, \sigma^2\right)$ entries. Y, X are independent. We study the optimal value ψ_2 of the following optimization problem:

$$\begin{array}{lll} (\Psi_2) & \min & n^{-\frac{1}{2}}||Y - X\beta||_2 \\ & \text{s.t.} & \beta \in \{0,1\}^p \\ & & ||\beta||_0 = k. \end{array}$$

That is, we no longer have ground truth vector β^*, and instead search for a vector β which makes $X\beta$ as close to an independent vector Y as possible in $\|\cdot\|_2$ norm.

We now state our main result for the pure noise model case.

Theorem 6 *The following holds for all n, p, k, σ:*

$$\mathbb{P}\left(\psi_2 \geq e^{-3/2}\sqrt{k+\sigma^2}\exp\left(-\frac{k\log p}{n}\right)\right) \geq 1 - e^{-n}. \tag{5}$$

Furthermore, for every $C > 0$ and every sufficiently large constant D_0, if $k\log k \leq Cn$, $k \leq \sigma^2 \leq 3k$, and $n \leq k\log p/(2\log D_0)$, the cardinality of the set

$$\left\{\beta : n^{-\frac{1}{2}}\|Y - X\beta\|_2 \leq D_0\sqrt{k+\sigma^2}\exp\left(-\frac{k\log p}{n}\right)\right\}$$

is at least $D_0^{\frac{n}{3}}$ w.h.p. as $k \to \infty$.

In the theorem above the value of the constant D_0 may depend on C (but does not depend on any other parameters, such as n, p or k). We note that in the second part of the theorem, our assumption $k \to \infty$ by our other assumptions also implies that both n and p diverge to infinity. The theorem above says that the value $\sqrt{k+\sigma^2}\exp\left(-\frac{k\log p}{n}\right)$ is the tight value of ψ_2 for the optimization problem Ψ_2, up to a multiplicative constant. Moreover, for the upper bound part, according to the second part of the theorem, the number of solutions achieving asymptotically this value is exponentially large in n. The assumption $k \leq \sigma^2 \leq 3k$ is adopted so that the result of the theorem is transferable to the original model where β^* is a k-sparse binary vector, in the way made precise in the following part of the appendix.

The proof of Theorem 6 is the subject of this part of the appendix. The lower bound is obtained by a simple moment argument. The upper bound is the part which consumes the bulk of the proof and will employ a certain conditional second moment method. Since for any $x \in \mathbb{R}^n$ we have $n^{-\frac{1}{2}}\|x\|_2 \leq \|x\|_\infty$, the result will be implied by looking instead at the cardinality of the set

$$\left\{\beta : \|Y - X\beta\|_\infty \leq D_0\sqrt{k+\sigma^2}\exp\left(-\frac{k\log p}{n}\right)\right\}, \tag{6}$$

and establishing the same result for this set.

A.1. The Lower Bound. Proof of (5) of Theorem 6

Proof Observe that $p^k \geq \binom{p}{k}$ implies $\exp\left(\frac{k\log p}{n}\right) \geq \binom{p}{k}^{\frac{1}{n}}$ and therefore

$$\mathbb{P}\left(\psi_2 \geq e^{-\frac{3}{2}} \exp\left(-\frac{k\log p}{n}\right)\sqrt{k+\sigma^2}\right) \geq \mathbb{P}\left(\psi_2 \geq e^{-\frac{3}{2}}\binom{p}{k}^{-\frac{1}{n}}\sqrt{k+\sigma^2}\right).$$

Thus it suffices to show

$$\mathbb{P}\left(\psi_2 \geq e^{-\frac{3}{2}}\binom{p}{k}^{-\frac{1}{n}}\sqrt{k+\sigma^2}\right) \geq 1-e^{-n}.$$

Given any $t>0$, let

$$\begin{aligned} Z_t &= |\{\beta \in \{0,1\}^p : |\beta||_0 = k, n^{-\frac{1}{2}}||Y-X\beta||_2 < t\}| \\ &= \sum_{\beta\in\{0,1\}^p|,|\beta||_0=k} \mathbf{1}\left(n^{-\frac{1}{2}}||Y-X\beta||_2 < t\right), \end{aligned}$$

$\mathbf{1}(A)$ denotes the indicator function applied to the event A. Let $t_0 := e^{-\frac{3}{2}}\binom{p}{k}^{-\frac{1}{n}}$. Observe that $t_0 \in (0,1)$. We have

$$\begin{aligned} \mathbb{P}\left(\psi_2 < e^{-\frac{3}{2}}\binom{p}{k}^{-\frac{1}{n}}\sqrt{k+\sigma^2}\right) &= \mathbb{P}\left(Z_{t_0\sqrt{k+\sigma^2}} \geq 1\right) \\ &\leq \mathbb{E}\left[Z_{t_0\sqrt{k+\sigma^2}}\right]. \end{aligned}$$

Now notice that $Z_{t_0\sqrt{k+\sigma^2}}$ is a sum of the $\binom{p}{k}$ indicator variables, each one of them referring to the event that a specific k-sparse binary β satisfies $n^{-\frac{1}{2}}||Y-X\beta||_2 < t_0\sqrt{k+\sigma^2}$ namely it satisfies $||Y-X\beta||_2^2 < t_0^2\left(k+\sigma^2\right)n$.

Furthermore, notice that for fixed $\beta \in \{0,1\}^p$ and k-sparse, $Y - X\beta = Y - \sum_{i\in S} X_i$ for $S \triangleq \text{Support}(\beta)$, where X_i is the i-th column of X. Hence since Y, X are independent, Y_i are i.i.d. $N\left(0,\sigma^2\right)$ and $X_{i,j}$ are i.i.d. $N(0,1)$, then $||Y-X\beta||_2^2$ is distributed as $\left(k+\sigma^2\right)\sum_{i=1}^n Z_i^2$ where Z_i i.i.d. standard normal Gaussian, namely $\left(k+\sigma^2\right)$ multiplied by a random variable with chi-squared distribution with n degrees of freedom. Hence for a fixed k-sparse $\beta \in \{0,1\}^p$, after rescaling, it holds

$$\mathbb{P}\left(||Y-X\beta||_2 n^{-\frac{1}{2}} < t_0\sqrt{k+\sigma^2}\right) = \mathbb{P}\left(\sum_{i=1}^n Z_i^2 \leq t_0^2 n\right).$$

Therefore

$$\begin{aligned} &\mathbb{E}\left[Z_{t_0\sqrt{k+\sigma^2}}\right] = \mathbb{E}\left[\sum_{\beta\in\{0,1\}^p|,|\beta||_0=k} 1\left(n^{-\frac{1}{2}}||Y-X\beta||_2 < t\right)\right] \\ &= \binom{p}{k}\mathbb{P}\left(||Y-X\beta||_2 n^{-\frac{1}{2}} < t_0\sqrt{k+\sigma^2}\right) \\ &= \binom{p}{k}\mathbb{P}\left(\sum_{i=1}^n Z_i^2 \leq t_0^2 n\right). \end{aligned}$$

We conclude

$$\mathbb{P}\left(\psi_2 < e^{-\frac{3}{2}}\binom{p}{k}^{-\frac{1}{n}}\sqrt{k+\sigma^2}\right) \leq \mathbb{E}\Big[Z_{t_0\sqrt{k+\sigma^2}}\Big] = \binom{p}{k}\mathbb{P}\left(\sum_{i=1}^{n} Z_i^2 \leq t_0^2 n\right). \tag{7}$$

Using standards large deviation theory estimates (see for example Shwartz and Weiss (1995)), for the sum of n chi-square distributed random variables we obtain that for $t_0 \in (0,1)$,

$$\mathbb{P}\left(\sum_{i=1}^{n} Z_i^2 \leq n t_0^2\right) \leq \exp\left(n f\left(t_0\right)\right) \tag{8}$$

with $f\left(t_0\right) \triangleq \frac{1-t_0^2+2\log(t_0)}{2}$.

Since $f\left(t_0\right) < \frac{1}{2} + \log t_0$, and as we recall $t_0 = e^{-\frac{3}{2}}\binom{p}{k}^{-\frac{1}{n}} < 1$ we obtain,

$$f\left(t_0\right) < -1 - \frac{1}{n}\log\binom{p}{k},$$

which implies

$$\exp\left(n f\left(t_0\right)\right) < \exp\left(-n\right)\binom{p}{k}^{-1},$$

which implies

$$\binom{p}{k}\exp\left(n f\left(t_0\right)\right) < \exp\left(-n\right).$$

Hence using the above inequality, (8) and (7) we get

$$\mathbb{P}\left(\psi_2 < e^{-\frac{3}{2}}\binom{p}{k}^{-\frac{1}{n}}\sqrt{k+\sigma^2}\right) \leq \exp\left(-n\right),$$

and the proof of (5) is complete.

■

We now turn to proving the upper bound part of Theorem 6. We begin by establishing several preliminary results.

A.2. Preliminaries

We first observe that $k\log k \leq Cn$ and $n \leq k\log p/(2\log D_0)$, implies $\log k \leq C\log p/(2\log D_0)$. In particular, for D_0 sufficiently large

$$k^4 \leq p. \tag{9}$$

We establish the following two auxiliary lemmas.

Lemma 7 *If $m_1, m_2 \in \mathbb{N}$ with $m_1 \geq 4$ and $m_2 \leq \sqrt{m_1}$ then*

$$\binom{m_1}{m_2} \geq \frac{m_1^{m_2}}{4 m_2!}.$$

Proof We have,

$$\binom{m_1}{m_2} \geq \frac{m_1^{m_2}}{4 m_2!}$$

holds if an only if

$$\prod_{i=1}^{m_2-1} \left(1 - \frac{i}{m_1}\right) \geq \frac{1}{4}.$$

Now $m_2 \leq \sqrt{m_1}$ implies

$$\begin{aligned}\prod_{i=1}^{m_2-1} \left(1 - \frac{i}{m_1}\right) &\geq \prod_{i=1}^{\lfloor \sqrt{m_1} \rfloor} \left(1 - \frac{i}{m_1}\right) \\ &\geq \left(1 - \frac{1}{\sqrt{m_1}}\right)^{\sqrt{m_1}},\end{aligned}$$

It is easy to verify that $x \geq 2$ implies $\left(1 - \frac{1}{x}\right)^x \geq \frac{1}{4}$. This completes the proof.

■

Lemma 8 *The function $f : [0, 1) \to \mathbb{R}$ defined by*

$$f(\rho) := \frac{1}{\rho} \log\left(\frac{1-\rho}{1+\rho}\right),$$

for $\rho \in [0, 1)$ is concave.

Proof The second derivative of f equals

$$\frac{2\left(-4\rho^3 + \left(\rho^2 - 1\right)^2 \log\left(\frac{1-\rho}{1+\rho}\right) + 2\rho\right)}{\rho^3 \left(1 - \rho^2\right)^2}.$$

Hence, it suffices to prove that the function $g : [0, 1) \to \mathbb{R}$ defined by

$$g(\rho) := -4\rho^3 + \left(\rho^2 - 1\right)^2 \log\left(\frac{1-\rho}{1+\rho}\right) + 2\rho$$

is non-positive. But for $\rho \in [0, 1)$

$$g'(\rho) = 4\rho\left(1 - \rho^2\right) \log\left(\frac{1+\rho}{1-\rho}\right) - 10\rho^2 \text{ and } g''(\rho) = 4\left(\left(1 - 3\rho^2\right) \log\left(\frac{1+\rho}{1-\rho}\right) - 3\rho\right).$$

We claim the second derivate of g is always negative. If $1-3\rho^2<0$, then $g''(\rho)<0$ is clearly negative. Now suppose $1-3\rho^2>0$. The inequality $\log(1+x)\le x$ implies $\log\left(\frac{1+\rho}{1-\rho}\right)\le\frac{2\rho}{1-\rho}$. Hence,

$$g''(\rho)\le 4\left(\frac{2\rho}{1-\rho}\left(1-3\rho^2\right)-3\rho\right)=4\rho\frac{3\rho-6\rho^2-1}{1-\rho}<0,$$

where the last inequality follows from the fact that $3\rho-6\rho^2-1<0$ for all $\rho\in\mathbb{R}$.

Therefore g is concave and therefore $g'(\rho)\le g'(0)=0$ which implies that g is also decreasing. In particular for all $\rho\in[0,1)$, $g(\rho)\le g(0)=0$. ■

For any $t>0, y\in\mathbb{R}$ and a standard Gaussian random variable Z we let

$$p_{t,y}:=\mathbb{P}\left(|Z-y|\le t\right). \tag{10}$$

Observe that

$$p_{t,y}=\int_{[-t,t]}\frac{1}{\sqrt{2\pi}}e^{-\frac{(y+x)^2}{2}}dx\ge\sqrt{\frac{2}{\pi}}te^{-\frac{y^2+t^2}{2}},$$

leading to

$$\log p_{t,y}\ge\log t-\frac{t^2}{2}-\frac{y^2}{2}+(1/2)\log(2/\pi). \tag{11}$$

Similarly, for any $t>0, y\in\mathbb{R}, \rho\in[0,1]$ we let

$$q_{t,y,\rho}:=\mathbb{P}\left(|Z_1-y|\le t, |Z_2-y|\le t\right), \tag{12}$$

where the random pair (Z_1,Z_2) follows a bivariate normal distribution with correlation ρ. In particular, $q_{t,y,0}=p_{t,y}^2$ and $q_{t,y,1}=p_{t,y}$. We now state and prove a lemma which provides an upper bound on the ratio $\frac{q_{t,y,\rho}}{p_{t,y}^2}$, for any $\rho\in[0,1)$.

Lemma 9 *For any $t>0, y\in\mathbb{R}, \rho\in[0,1)$,*

$$\frac{q_{t,y,\rho}}{p_{t,y}^2}\le\sqrt{\frac{1+\rho}{1-\rho}}e^{\rho y^2}.$$

Proof We have

$$\begin{aligned}
q_{t,y,\rho}&=\frac{1}{2\pi\sqrt{1-\rho^2}}\int_{[y-t,y+t]^2}\exp\left(-\frac{x^2+z^2-2\rho xz}{2\left(1-\rho^2\right)}\right)dxdz\\
&=\frac{1}{2\pi\sqrt{1-\rho^2}}\int_{[y-t,y+t]^2}\exp\left(-\frac{(x-\rho z)^2}{2\left(1-\rho^2\right)}-\frac{z^2}{2}\right)dxdz\\
&\le\frac{1}{2\pi\sqrt{1-\rho^2}}\int_{[y-t,y+t]}\exp\left(-\frac{x_2^2}{2}\right)dx_2\int_{[y(1-\rho)-t(1+\rho),y(1-\rho)+t(1+\rho)]}\exp\left(-\frac{x_1^2}{2\left(1-\rho^2\right)}\right)dx_1,
\end{aligned}$$

where in the inequality we have introduced the change of variables $(x_1, x_2) = (x - \rho z, z)$ and upper bounded the transformed domain by

$$[y(1-\rho) - t(1+\rho), y(1-\rho) + t(1+\rho)] \times [y-t, y+t].$$

Introducing another change of variable $x_1 = x_3(1+\rho) + y(1-\rho)$, the expression on the right-hand side of the inequality above becomes

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{[y-t,y+t]} \exp\left(-\frac{x_2^2}{2}\right) dx_2 \, (1+\rho) \int_{[-t,t]} \exp\left(-\frac{(x_3(1+\rho)+y(1-\rho))^2}{2(1-\rho^2)}\right) dx_3,$$

$$\begin{aligned}
&= \exp\left(-\frac{y^2(1-\rho)}{2(1+\rho)}\right) \frac{1}{2\pi}\sqrt{\frac{1+\rho}{1-\rho}} \int_{[y-t,y+t]} \exp\left(-\frac{x_2^2}{2}\right) dx_2 \times \\
&\times \int_{[-t,t]} \exp\left(-\frac{x_3^2(1+\rho)^2 + 2x_3 y(1-\rho^2)}{2(1-\rho^2)}\right) dx_3 \\
&\le \exp\left(-\frac{y^2(1-\rho)}{2(1+\rho)}\right) \frac{1}{2\pi}\sqrt{\frac{1+\rho}{1-\rho}} \int_{[y-t,y+t]} \exp\left(-\frac{x_2^2}{2}\right) dx_2 \int_{[-t,t]} \exp\left(-\frac{x_3^2}{2} + x_3 y\right) dx_3 \\
&= \exp\left(\frac{y^2\rho}{1+\rho}\right) \frac{1}{2\pi}\sqrt{\frac{1+\rho}{1-\rho}} \int_{[y-t,y+t]} \exp\left(-\frac{x_2^2}{2}\right) dx_2 \int_{[-t,t]} \exp\left(-\frac{(x_3+y)^2}{2}\right) dx_3 \\
&= \exp\left(\frac{y^2\rho}{1+\rho}\right) \frac{1}{2\pi}\sqrt{\frac{1+\rho}{1-\rho}} \left(\int_{[y-t,y+t]} \exp\left(-\frac{x_2^2}{2}\right) dx_2\right)^2,
\end{aligned}$$

which is exactly:

$$\exp\left(\frac{y^2\rho}{1+\rho}\right)\sqrt{\frac{1+\rho}{1-\rho}} p_{t,y}^2 \le \exp\left(y^2\rho\right)\sqrt{\frac{1+\rho}{1-\rho}} p_{t,y}^2$$

This completes the proof of Lemma 9. ■

A.3. Conditional second moment bounds

Recall, that our goal is to establish the required bound on the cardinality of the set (6) instead. Thus for every $s > 0$ let

$$Z_{s,\infty} = |\{\beta \in \{0,1\}^p : \|\beta\|_0 = k, ||Y - X\beta||_\infty < s\}|.$$

For the purposes of our proof the relevant scaling of s is of the form $s = t\sqrt{k}$ where t is constant. Then our next step is obtaining estimates on $\mathbb{E}\left[Z_{t\sqrt{k},\infty}|Y\right]$ and $\mathbb{E}\left[Z^2_{t\sqrt{k},\infty}|Y\right]$ for constant t. Let

$$\Upsilon = \Upsilon(Y) \triangleq \frac{\mathbb{E}\left[Z^2_{t\sqrt{k}}|Y\right]}{\mathbb{E}\left[Z_{t\sqrt{k}}|Y\right]^2}.$$

Our goal ,for reasons that will become clear in the next subsection, is obtaining an upper bound on $\mathbb{E}_Y[1-1/\Upsilon]$. A direct calculation gives

$$\mathbb{E}\Big[Z_{t\sqrt{k},\infty}|Y\Big]=\binom{p}{k}\prod_{i=1}^{n}\mathbb{P}\left(|\frac{Y_i}{\sqrt{k}}-X|<t\right)=\binom{p}{k}\prod_{i=1}^{n}p_{t,\frac{Y_i}{\sqrt{k}}},$$

where X is a standard normal random variable and $p_{t,y}$ was defined in (10). Similarly,

$$\mathbb{E}\Big[Z^2_{t\sqrt{k},\infty}|Y\Big]=\sum_{\ell=0}^{k}\binom{p}{k-l,k-l,l,p-2k+l}\prod_{i=1}^{n}\mathbb{P}\left(|Y_i-X_1^l|<t\sqrt{k},|Y_i-X_2^l|<t\sqrt{k}\right),$$

where X_1^l,X_2^l are each $N(0,k)$ random variables with covariance l. In terms of $q_{t,y,\rho}$ defined in (12) we have for every l,

$$\mathbb{P}\left(|Y_i-X_1^l|<t\sqrt{k},|Y_i-X_2^l|<t\sqrt{k}\right)=q_{t,\frac{Y_i}{\sqrt{k}},\frac{l}{k}}.$$

Hence,

$$\mathbb{E}\Big[Z^2_{t\sqrt{k+\sigma^2},\infty}|Y\Big]=\sum_{\ell=0}^{k}\binom{p}{k-l,k-l,l,p-2k+l}\prod_{i=1}^{n}q_{t,\frac{Y_i}{\sqrt{k}},\frac{l}{k}}.$$

We obtain

$$\Upsilon=\Upsilon(Y)=\sum_{\ell=0}^{k}\frac{\binom{p}{k-l,k-l,l,p-2k+l}}{\binom{p}{k}^2}\prod_{i=1}^{n}\frac{q_{t,\frac{Y_i}{\sqrt{k}},\frac{l}{k}}}{p^2_{t,\frac{Y_i}{\sqrt{k}}}}.$$

Now for $\ell=0$ and all $i=1,2,...,n$ we have $q_{t,\frac{Y_i}{\sqrt{k}},0}=p^2_{t,\frac{Y_i}{\sqrt{k}}}$ a.s. and therefore the first term of this sum equals $\frac{\binom{p}{k,k,p-2k}}{\binom{p}{k}^2}\leq 1$.

We now analyze terms corresponding to $\ell\geq 1$. We have for all $\ell=1,..,k$

$$\binom{k}{l}\leq\frac{k^l}{l!}\leq k^l,\binom{p-k}{k-l}\leq\frac{(p-k)^{k-l}}{(k-l)!}.$$

By (9) we have $k^4\leq p$ implying $k\leq\sqrt{p}$ and applying Lemma 7 we have

$$\binom{p}{k}\geq\frac{p^k}{4k!}.$$

Combining the above we get that for every $\ell=1,...,k$ it holds:

$$\frac{\binom{p}{k-l,k-l,l,p-2k+l}}{\binom{p}{k}^2}=\binom{k}{l}\frac{\binom{p-k}{k-l}}{\binom{p}{k}}\leq k^l\frac{(p-k)^{k-l}}{(k-l)!}\frac{4k!}{p^k}\leq 4\left(\frac{p}{k^2}\right)^{-l}.$$

Hence we have

$$\Upsilon\leq 1+4\sum_{\ell=1}^{k}\left(\frac{p}{k^2}\right)^{-l}\prod_{i=1}^{n}\frac{q_{t,\frac{Y_i}{\sqrt{k}},\frac{l}{k}}}{p^2_{t,\frac{Y_i}{\sqrt{k}}}}. \tag{13}$$

A simple inequality $x + \frac{1}{x} \geq 2, x > 0$ implies that almost surely, $\Upsilon - 1 \geq 1 - \frac{1}{\Upsilon}$. At the same time $1 - \frac{1}{\Upsilon} \leq 1$. Hence a.s.

$$1 - \frac{1}{\Upsilon} \leq \min\{1, \Upsilon - 1\},$$

which implies

$$\mathbb{E}_Y \left(1 - \frac{1}{\Upsilon}\right) \leq \mathbb{E}_Y \left(\min\{1, \Upsilon - 1\}\right).$$

Our key result regarding the conditional second moment estimate and its ratio to the square of the conditional first moment estimate (namely Υ^{-1}) is the following proposition.

Proposition 10 *Suppose $k \log k \leq Cn$ for all k and n for some constant $C > 0$. Then for all sufficiently large constants $D > 0$ there exists $c > 0$ such that for $n \leq \frac{k \log\left(\frac{p}{k^2}\right)}{2 \log D}$ and $t = D\sqrt{1+\sigma^2}\left(\frac{p}{k^2}\right)^{-\frac{k}{n}}$ we have*

$$\mathbb{E}_Y \left(\min\{1, \Upsilon - 1\}\right) \leq \frac{1}{k^c}.$$

Proof Fix a parameter $\zeta \in (0,1)$ which will be optimized later. We have,

$$\begin{aligned}
\mathbb{E}_Y \left(\min\{1, \Upsilon - 1\}\right) &= \mathbb{E}_Y \left(\min\{1, \Upsilon - 1\}\mathbf{1}\left(\min\{1, \Upsilon - 1\} \geq \zeta^n\right)\right) \\
&\quad + \mathbb{E}_Y \left(\min\{1, \Upsilon - 1\}\mathbf{1}\left(\min\{1, \Upsilon - 1\} \leq \zeta^n\right)\right) \\
&\leq \mathbb{P}\left(\min\{1, \Upsilon - 1\} \geq \zeta^n\right) + \zeta^n.
\end{aligned}$$

Observe that if $\Upsilon \geq 1 + \zeta^n$, then (13) implies that at least one of the summands of

$$\sum_{\ell=1}^{k} 4\left(\frac{p}{k^2}\right)^{-l} \prod_{i=1}^{n} \frac{q_{t, \frac{Y_i}{\sqrt{k}}, l}}{p^2_{t, \frac{Y_i}{\sqrt{k}}}}$$

for $\ell = 1, 2.., k$ should be at least $\frac{\zeta^n}{k}$. Hence applying the union bound,

$$\begin{aligned}
\mathbb{P}\left(\min\{1, \Upsilon - 1\} \geq \zeta^n\right) &\leq \mathbb{P}\left(\Upsilon \geq 1 + \zeta^n\right) \\
&\leq \mathbb{P}\left(\bigcup_{\ell=1}^{k} \{4\left(\frac{p}{k^2}\right)^{-l} \prod_{i=1}^{n} \frac{q_{t, \frac{Y_i}{\sqrt{k}}, \frac{l}{k}}}{p^2_{t, \frac{Y_i}{\sqrt{k}}}} \geq \frac{\zeta^n}{k}\}\right) \\
&\leq \sum_{\ell=1}^{k} \mathbb{P}\left(4\left(\frac{p}{k^2}\right)^{-l} \prod_{i=1}^{n} \frac{q_{t, \frac{Y_i}{\sqrt{k}}, \frac{l}{k}}}{p^2_{t, \frac{Y_i}{\sqrt{k}}}} \geq \frac{\zeta^n}{k}\right)
\end{aligned}$$

Introducing parameter $\rho = \frac{l}{k}$ we obtain

$$\mathbb{E}_Y \left(\min\{1, \Upsilon - 1\}\right) \leq \zeta^n + \mathbb{P}\left(\min\{1, \Upsilon - 1\} \geq \zeta^n\right) \leq \zeta^n + \sum_{\rho = \frac{1}{k}, \frac{2}{k}, .., 1} \mathbb{P}\left(\Upsilon_\rho\right), \tag{14}$$

where for all $\rho = \frac{1}{k}, .., \frac{k-1}{k}, \frac{k}{k}$ we define

$$\Upsilon_\rho \triangleq \left\{ 4\left(\frac{p}{k^2}\right)^{-\rho k} \prod_{i=1}^{n} \frac{q_{t,\frac{Y_i}{\sqrt{k}},\rho}}{p^2_{t,\frac{Y_i}{\sqrt{k}}}} \geq \frac{\zeta^n}{k} \right\}.$$

Next we obtain an upper bound on $\mathbb{P}(\Upsilon_\rho)$ for any $\rho \in (0,1]$ as a function of ζ. Set

$$\rho_* := 1 - \frac{n \log D}{3k \log(p/k^2)}.$$

The cases $\rho \leq \rho_*$ and $\rho > \rho_*$ will be considered separately.

Lemma 11 *For all $\rho \in (\rho_*, 1]$ and $\zeta \in (0,1)$.*

$$\mathbb{P}(\Upsilon_\rho) \leq 2^n \left(D^{-\frac{1}{18}} z^{-\frac{1}{6}} \right)^n.$$

Proof Since $\rho > \rho_*$ then

$$-(1-\rho)\frac{k \log\left(\frac{p}{k^2}\right)}{n} \geq -\frac{1}{3}\log D. \tag{15}$$

Now we have $q_{t,\frac{Y_i}{\sqrt{k}},\rho} \leq p_{t,\frac{Y_i}{\sqrt{k}}}$ which implies $\frac{q_{t,\frac{Y_i}{\sqrt{k}},\rho}}{p^2_{t,\frac{Y_i}{\sqrt{k}}}} \leq p^{-1}_{t,\frac{Y_i}{\sqrt{k}}}$, which after taking logarithms and dividing both the sides by n gives

$$\mathbb{P}(\Upsilon_\rho) \leq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} -\log p_{t,\frac{Y_i}{\sqrt{k}}} \geq \log\zeta - \frac{\log 4k}{n} + \rho\frac{k\log\frac{p}{k^2}}{n} \right).$$

Applying (11) we obtain

$$\mathbb{P}(\Upsilon_\rho) \leq \mathbb{P}\left(-\log t + \frac{t^2}{2} + \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i^2}{2k} + (1/2)\log(2/\pi) \geq \log\zeta - \frac{\log 4k}{n} + \rho\frac{k\log\frac{p}{k^2}}{n} \right),$$

Recall that $t = D\sqrt{1+\sigma^2}\left(\frac{p}{k^2}\right)^{-\frac{k}{n}}$, namely $\log t \geq \log D - \frac{k}{n}\log\left(\frac{p}{k^2}\right)$ and thus applying (15)

$$\begin{aligned}\log t + \rho\frac{k\log\frac{p}{k^2}}{n} &\geq -(1-\rho)\frac{k\log\frac{p}{k^2}}{n} + \log D \\ &\geq \frac{2}{3}\log D.\end{aligned}$$

By the bound on n, we have $t \leq D\sqrt{1+\sigma^2}/D^2 \leq 2/D \leq 1$ for sufficiently large D. The same applies to $t^2/2$. Also since $k\log k \leq Cn$ then $\log(4k)/n \leq C/k + \log 4/(k\log k)$. Then for sufficiently large D we obtain

$$\begin{aligned}\mathbb{P}(\Upsilon_\rho) &\leq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i^2}{2k} \geq \log\zeta + (1/3)\log D\right) \\ &= \mathbb{P}\left(\exp\left(\frac{1}{6}\sum_{i=1}^{n}\frac{Y_i^2}{2k}\right) \geq \zeta^{\frac{n}{6}} D^{\frac{n}{18}}\right) \\ &\leq \frac{1}{\zeta^{\frac{n}{6}} D^{\frac{n}{18}}}\left(\mathbb{E}\left[\exp\left(\frac{Y_1^2}{12k}\right)\right]\right)^n\end{aligned}$$

Recall that since Y_1 has distribution $N(0,\sigma^2)$ and $\sigma^2 \leq 3k$ then

$$\mathbb{E}\left[\exp\left(\frac{Y_1^2}{12k}\right)\right] = \frac{1}{\sqrt{1-2\sigma^2/(12k)}} \leq \sqrt{2}.$$

We obtain a bound

$$\mathbb{P}\left(\Upsilon_\rho\right) \leq 2^n \left(D^{-\frac{1}{18}} z^{-\frac{1}{6}}\right)^n,$$

as claimed. ∎

Lemma 12 *For all $\rho \in [\frac{1}{k}, \rho_*]$ and $\zeta \in (0,1)$.*

$$\mathbb{P}\left(\Upsilon_\rho\right) \leq 4^n \left(D^{\frac{1}{2}} \zeta^k\right)^{-n/12}.$$

Proof Applying Lemma 9 we have

$$\begin{aligned}
\mathbb{P}\left(\Upsilon_\rho\right) &= \mathbb{P}\left(4\left(\frac{p}{k^2}\right)^{-\rho k} \prod_{i=1}^n \frac{q_{t,\frac{Y_i}{\sqrt{k}},\rho}}{p^2_{t,\frac{Y_i}{\sqrt{k}}}} \geq \frac{\zeta^n}{k}\right) \\
&\leq \mathbb{P}\left(4\left(\frac{p}{k^2}\right)^{-\rho k} \prod_{i=1}^n \left(\sqrt{\frac{1+\rho}{1-\rho}} \exp\left(\rho \frac{Y_i}{\sqrt{k}}^2\right)\right) \geq \frac{\zeta^n}{k}\right) \\
&= \mathbb{P}\left(\rho \sum_{i=1}^n \frac{Y_i^2}{kn} \geq \log\zeta - \frac{\log 4k}{n} + \frac{1}{2}\log\left(\frac{1-\rho}{1+\rho}\right) + \frac{\rho k \log\left(\frac{p}{k^2}\right)}{n}\right) \\
&= \mathbb{P}\left(\sum_{i=1}^n \frac{Y_i^2}{kn} \geq \rho^{-1}\log\zeta - \rho^{-1}\frac{\log 4k}{n} + \frac{1}{2\rho}\log\left(\frac{1-\rho}{1+\rho}\right) + \frac{k \log\left(\frac{p}{k^2}\right)}{n}\right).
\end{aligned}$$

Let

$$f(\rho) = \rho^{-1}\log\zeta - \rho^{-1}\frac{\log 4k}{n} + \frac{1}{2\rho}\log\left(\frac{1-\rho}{1+\rho}\right) + \frac{k\log\left(\frac{p}{k^2}\right)}{n}.$$

Applying Lemma 8 and that $\zeta < 1$ we can see that the function f is concave. This implies that the minimum value of f for $\rho \in [\frac{1}{k}, \rho_*]$ is either $f\left(\frac{1}{k}\right)$ or $f(\rho_*)$, and therefore

$$\begin{aligned}
\mathbb{P}\left(\Upsilon_\rho\right) &\leq \mathbb{P}\left(\sum_{i=1}^n \frac{Y_i^2}{kn} \geq \min\{f\left(\frac{1}{k}\right), f(\rho_*)\}\right) \\
&\leq \mathbb{P}\left(\sum_{i=1}^n \frac{Y_i^2}{kn} \geq f\left(\frac{1}{k}\right)\right) + \mathbb{P}\left(\sum_{i=1}^n \frac{Y_i^2}{kn} \geq f(\rho_*)\right). \qquad (16)
\end{aligned}$$

Now we apply a standard Chernoff type bound on $\mathbb{P}\left(\sum_{i=1}^n \frac{Y_i^2}{k} \geq nw\right)$ for $w \in \mathbb{R}$. We have $\mathbb{E}\left[\exp\left(\theta Y_i^2/k\right)\right] = \frac{1}{\sqrt{1-2(\sigma^2/k)\theta}} < \infty$ if $\theta < \frac{1}{2\sigma^2/k}$. Since in our case $1 \leq \mathbb{E}\left[\frac{Y_i^2}{k}\right] = \sigma^2/k \leq 3$, to obtain a finite bound we set $\theta = \frac{1}{12} < \frac{1}{6}$ and obtain

$$\mathbb{E}\left[\exp\left(\frac{Y_i^2}{12k}\right)\right] = \frac{1}{\sqrt{1-\frac{\sigma^2}{6k}}} \leq \sqrt{2}.$$

Therefore, we obtain

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{n}\frac{Y_i^2}{k} \geq nw\right) &\leq \exp\left(-n\frac{w}{12}\right)\left(\mathbb{E}\left[\exp\left(\frac{Y_i^2}{12k}\right)\right]\right)^n \\ &\leq 2^{\frac{n}{2}}\exp(-nw/12).\end{aligned}$$

We obtain

$$\mathbb{P}\left(\Upsilon_\rho\right) \leq 2^{\frac{n}{2}}\exp(-nf(1/k)/12) + 2^{\frac{n}{2}}\exp(-nf(\rho^*)/12). \tag{17}$$

Now we obtain bounds on $f\left(\frac{1}{k}\right)$ and $f\left(\rho_*\right)$. We have

$$f\left(\frac{1}{k}\right) = k\log\zeta - \frac{k\log 4k}{n} + \frac{k}{2}\log\left(\frac{1-\frac{1}{k}}{1+\frac{1}{k}}\right) + \frac{k\log\left(\frac{p}{k^2}\right)}{n}.$$

We have by our assumption $k\log k \leq Cn$ that $k\log(4k)/n \leq Ck\log(4k)/(k\log k)$. The sequence $\frac{k}{2}\log\left(\frac{1-\frac{1}{k}}{1+\frac{1}{k}}\right)$ is bounded by a universal constant for $k \geq 2$. Finally, we have $n \leq k\log(p/k^2)/(2\log D)$. Thus for sufficiently large D,

$$f\left(\frac{1}{k}\right) \geq k\log\zeta + \log D,$$

implying

$$2^{\frac{n}{2}}\exp(-nf(1/k)/12) \leq 2^{\frac{n}{2}}\left(D\zeta^k\right)^{-n/12}.$$

Now we will bound $f\left(\rho_*\right)$. We have

$$f\left(\rho^*\right) = (1/\rho^*)\log\zeta - (1/\rho^*)\frac{\log 4k}{n} + \frac{1}{2\rho^*}\log\left(\frac{1-\rho^*}{1+\rho^*}\right) + \frac{k\log\left(\frac{p}{k^2}\right)}{n}.$$

Applying upper bound on n, we have $\rho_* > 1/2$. Then $-1/(2\rho^*)\log(1+\rho^*) \geq -\log 2$. We obtain

$$f\left(\rho^*\right) = 2\log\zeta - 2\frac{\log 4k}{n} + \log\left(1-\rho^*\right) + \frac{k\log\left(\frac{p}{k^2}\right)}{n}.$$

We have again

$$2\log(4k)/n \leq 2C\log(4k/k). \tag{18}$$

Applying the value of ρ^* we have

$$\log\left(1-\rho^*\right) + \frac{k\log\left(\frac{p}{k^2}\right)}{n} = -\log\left(\frac{3k\log(p/k^2)}{n\log D}\right) + \frac{k\log\left(\frac{p}{k^2}\right)}{n}.$$

Consider

$$-\log\left(\frac{3k\log(p/k^2)}{\log D}\right)+\log n+\frac{k\log\left(\frac{p}{k^2}\right)}{n}.$$

For every $a>0$, the function $\log x+a/x$ is a decreasing on $x\in(0,a]$ and thus, applying the bound $n\le k\log(p/k^2)/(2\log D)$, the expression above is at least

$$\begin{aligned}-\log\left(\frac{3k\log(p/k^2)}{\log D}\right)+\log\left(k\log(p/k^2)/(2\log D)\right)+2\log D&=-\log 3-\log 2+2\log D\\&\ge(3/2)\log D,\end{aligned}$$

for sufficiently large D. Combining with (18) we obtain that for sufficiently large D

$$f(\rho^*)\ge 2\log\zeta+\log D,$$

Combining two bounds we obtain

$$\begin{aligned}\mathbb{P}\left(\Upsilon_\rho\right)&\le 2^{\frac{n}{2}}\left(D\zeta^k\right)^{-\frac{n}{12}}+2^{\frac{n}{2}}\left(D\zeta^2\right)^{-n/12}\\&\le 2^{\frac{n}{2}+1}\left(D\zeta^k\right)^{-\frac{n}{12}}.\end{aligned}$$

■

We now return to the proof of Proposition 10. Combining the results of Lemma 11 and Lemma 12, and assuming $k\ge 6\cdot 12=72$, we obtain that

$$\begin{aligned}\mathbb{P}\left(\Upsilon_\rho\right)&\le 2^n\left(D^{\frac{1}{18}}\zeta^6\right)^{-n}+2^{\frac{n}{2}+1}\left(D\zeta^k\right)^{-n/12}\\&\le 2^{n+1}\left(D^{\frac{1}{2}}\zeta^k\right)^{-n/12}\end{aligned}$$

for all $\rho\in[1/k,1]$ and $\zeta\in(0,1)$. Recalling (14) we obtain

$$\mathbb{E}_Y\left(\min\{1,\Upsilon-1\}\right)\le\zeta^n+(2k)2^n\left(D^{\frac{1}{2}}\zeta^k\right)^{-n/12}.$$

Let $D_1\triangleq D^{\frac{1}{2}}/2^{12}$ and rewrite the bound above as

$$\zeta^n+(2k)\left(D_1\zeta^k\right)^{-n/12}.$$

Assume D is large enough so that $D_1>1$ and let $\zeta=1/D_1^{\frac{1}{2k}}<1$. We obtain a bound

$$D_1^{-\frac{n}{2k}}+(2k)D_1^{-n/24}.$$

Finally since $n\ge(1/C)k\log k$, we obtain a bound of the form $1/k^c$ for some constant $c>0$ as claimed. This completes the proof of Proposition 10.

■

A.4. The Upper Bound

Proof [Proof of Theorem 6] By an assumption of the theorem, we have $k^4 \le p$. Thus

$$k \log p \le 2k \log(p/k^2).$$

Then

$$n \le \frac{k \log p}{2 \log D_0} \le \frac{k \log(p/k^2)}{\log D_0} = \frac{k \log(p/k^2)}{2 \log D_0^{\frac{1}{2}}}. \tag{19}$$

Our goal is to obtain a lower bound on the cardinality of the set

$$\left\{ \beta : \|Y - X\beta\|_\infty \le D_0 \sqrt{k} \sqrt{1 + \sigma^2/k} \exp\left(-\frac{k \log p}{n}\right) \right\},$$

Recall that $k \le \sigma^2 \le 3k$. Letting

$$t_0 = D_0 \sqrt{1 + \sigma^2/k} \exp\left(-\frac{k \log p}{n}\right),$$

our goal is then obtaining a lower bound on $Z_{t_0\sqrt{k}}$. Since $k \log k \le Cn$, then for sufficiently large D_0,

$$t_0 \ge D_0^{\frac{1}{2}} \sqrt{1 + \sigma^2/k} \exp\left(-\frac{k \log(p/k^2)}{n}\right) \triangleq \tau,$$

and thus it suffices to obtain the claimed bound on $Z_{t_1\sqrt{k}}$. We note that by our bound (19)

$$\tau \le D_0^{\frac{1}{2}} \sqrt{1 + \sigma^2/k}/D_0 \le 2/D_0^{\frac{1}{2}} \le 1, \tag{20}$$

provided D_0 is sufficiently large. Let $D = D_0^{\frac{1}{2}}$. Then, by the definition of τ and by (19) the assumptions of Proposition 10 are satisfied for this choice of D and $t = \tau$.

Lemma 13 *The following bound holds with high probability with respect to Y as k increases*

$$n^{-1} \log \mathbb{E}\left[Z_{\tau\sqrt{k},\infty} | Y\right] \ge (1/2) \log D.$$

Proof As before for $Y = (Y_1, \ldots, Y_n)$,

$$\mathbb{E}\left[Z_{\tau\sqrt{k},\infty} | Y\right] = \binom{p}{k} \prod_{i=1}^{n} \mathbb{P}\left(|\frac{Y_i}{\sqrt{k}} - X| < t | Y\right) = \binom{p}{k} \prod_{i=1}^{n} p_{\tau, \frac{Y_i}{\sqrt{k}}},$$

where X is the standard normal random variable. Taking logarithms,

$$\log \mathbb{E}\left[Z_{\tau\sqrt{k},\infty} | Y\right] = \log \binom{p}{k} + \sum_{i=1}^{n} \log p_{\tau, \frac{Y_i}{\sqrt{k}}}. \tag{21}$$

Applying (11), we have

$$n^{-1}\log\mathbb{E}\Big[Z_{\tau\sqrt{k},\infty}|Y\Big] \ge n^{-1}\log\binom{p}{k} + \log\tau - \frac{\tau^2}{2} + (1/2)\log(2/\pi) - n^{-1}\sum_{i=1}^{n}\frac{Y_i^2}{2k}$$

Using

$$\tau \ge D\exp\left(-\frac{k\log(p/k^2)}{n}\right),$$

and $\tau \le 1$, we obtain

$$n^{-1}\log\mathbb{E}\Big[Z_{\tau\sqrt{k},\infty}|Y\Big] \ge n^{-1}\log\binom{p}{k} + \log D - \frac{k\log(p/k^2)}{n} - \frac{1}{2} + (1/2)\log(2/\pi) - n^{-1}\sum_{i=1}^{n}\frac{Y_i^2}{2k}$$

Since by (9) we have $k \le \sqrt{p}$, applying Lemma 7 we have $\frac{1}{n}\log\binom{p}{k} - \frac{k}{n}\log\left(\frac{p}{k^2}\right) \ge 0$. By Law of Large Numbers and since Y_i is distributed as $N(0,\sigma^2)$ with $k \le \sigma^2 \le 3k$, we have $n^{-1}\sum_{i=1}^{n}\frac{Y_i^2}{2k}$ converges to $\sigma^2/(2k) \le 3/2$ as k and therefore n increases. Assuming D is sufficiently large we obtain that w.h.p. as k increases,

$$n^{-1}\log\mathbb{E}\Big[Z_{\tau\sqrt{k},\infty}|Y\Big] \ge (1/2)\log D.$$

This concludes the proof of the lemma. ∎

Now we claim that w.h.p. as k increases,

$$Z_{\tau\sqrt{k},\infty} \ge \frac{1}{2}\mathbb{E}\Big[Z_{\tau\sqrt{k},\infty}|Y\Big]. \tag{22}$$

We have

$$\mathbb{P}\left(Z_{\tau\sqrt{k},\infty} < \frac{1}{2}\mathbb{E}\Big[Z_{\tau\sqrt{k},\infty}|Y\Big]\right) \le \mathbb{P}\left(|Z_{\tau\sqrt{k},\infty} - \mathbb{E}\Big[Z_{\tau\sqrt{k},\infty}|Y\Big]| \ge \frac{1}{2}\mathbb{E}\Big[Z_{\tau\sqrt{k},\infty}|Y\Big]\right), \tag{23}$$

and applying Chebyshev's inequality we obtain,

$$\mathbb{P}\left(|Z_{\tau\sqrt{k},\infty} - \mathbb{E}\Big[Z_{\tau\sqrt{k},\infty}|Y\Big]| \ge \frac{1}{2}\mathbb{E}\Big[Z_{\tau\sqrt{k},\infty}|Y\Big]\,|Y\right) \le 4\min\left[\frac{\mathbb{E}\Big[Z^2_{\tau\sqrt{k},\infty}|Y\Big]}{\mathbb{E}\Big[Z_{\tau\sqrt{k},\infty}|Y\Big]^2} - 1, 1\right].$$

Hence, taking expectation over Y we obtain,

$$\mathbb{P}\left(|Z_{\tau\sqrt{k},\infty} - \mathbb{E}\Big[Z_{\tau\sqrt{k},\infty}|Y\Big]| \ge \frac{1}{2}\mathbb{E}\Big[Z_{\tau\sqrt{k},\infty}|Y\Big]\right) \le 4\mathbb{E}_Y\left[\min\left[\frac{\mathbb{E}\Big[Z^2_{\tau\sqrt{k},\infty}|Y\Big]}{\mathbb{E}\Big[Z_{\tau\sqrt{k},\infty}|Y\Big]^2} - 1, 1\right]\right].$$

We conclude

$$\mathbb{P}\left(Z_{\tau\sqrt{k},\infty} < \frac{1}{2}\mathbb{E}\Big[Z_{\tau\sqrt{k},\infty}|Y\Big]\right) \le 4\mathbb{E}_Y\left[\min\left[\frac{\mathbb{E}\Big[Z^2_{\tau\sqrt{k},\infty}|Y\Big]}{\mathbb{E}\Big[Z_{\tau\sqrt{k},\infty}|Y\Big]^2} - 1, 1\right]\right]. \tag{24}$$

Applying Proposition 10 the assumptions of which have been verified as discussed above, we obtain

$$\begin{aligned}\mathbb{P}\left(Z_{\tau\sqrt{k},\infty} < \frac{1}{2}\mathbb{E}\Big[Z_{\tau\sqrt{k},\infty}|Y\Big]\right) &\le \mathbb{E}[\min\{1, \Upsilon - 1\}|Y] \\ &\le k^{-c},\end{aligned}$$

for some $c > 0$. This establishes the claim (22). Combining with Lemma 13, we conclude that w.h.p. as k increases

$$\begin{aligned}n^{-1}\log Z_{\tau\sqrt{k},\infty} &\ge n^{-1}\log \mathbb{E}\Big[Z_{\tau\sqrt{k},\infty}|Y\Big] - \log 2/n \\ &\ge (1/2)\log D - \log 2/n.\end{aligned}$$

Since n satisfying $Cn \ge k\log k$ increases as k increases, we conclude that w.h.p. as k increases $Z_{\tau\sqrt{k},\infty} \ge D^{\frac{n}{3}}$. This concludes the proof of the theorem. ■

Appendix B. Proof of Theorem 1

In this part of the appendix we prove Theorem 1. The proof is based on a reduction scheme to the simpler optimization problem Ψ_2 which is analyzed in the previous section.

To prove Theorem 1 we will also consider the following restriction of Φ_2. For any $S \subseteq \text{Support}(\beta^*)$ consider the optimization problem $(\Phi_2(S))$:

$$\begin{array}{lll}(\Phi_2(S)) & \min & n^{-\frac{1}{2}}||Y - X\beta||_2 \\ & \text{s.t.} & \beta \in \{0,1\}^p \\ & & ||\beta||_0 = k, \text{Support}(\beta) \cap \text{Support}(\beta^*) = S,\end{array}$$

and set $\phi_2(S)$ its optimal value. Notice that for a binary k-sparse β with $\text{Support}(\beta) \cap \text{Support}(\beta^*) = S$ we have:

$$\begin{aligned}Y - X\beta &= X\beta^* + W - X\beta \\ &= \sum_{i\in\text{Support}(\beta^*)} X_i + W - \sum_{i\in\text{Support}(\beta)} X_i \\ &= \sum_{i\in\text{Support}(\beta^*)-S} X_i + W - \sum_{i\in\text{Supp}(\beta)-S} X_i \\ &= Y' - X'\beta_1,\end{aligned}$$

where we have defined Y', X', β_1 as following:

1. $X' \in \mathbb{R}^{n\times(p-k)}$ to be the matrix which is X after deleting the columns corresponding to $\text{Support}(\beta^*)$
2. $Y' := \sum_{i\in\text{Support}(\beta^*)-S} X_i + W$
3. $\beta_1 \in \{0,1\}^{p-k}$ is obtained from β after deleting coordinates in $\text{Support}(\beta^*)$. Notice that $||\beta_1||_0 = k - |S|$.

Hence, solving $\Phi_2(S)$ can be written equivalently with respect to Y', X', β' as following,

$$\begin{array}{lll}(\Phi_2(S)) & \min & n^{-\frac{1}{2}}||Y' - X'\beta'||_2 \\ & \text{s.t.} & \beta' \in \{0,1\}^{p-k} \\ & & ||\beta'||_0 = k - |S|.\end{array}$$

We claim that the above problem is satisfying all the assumptions of Theorem 6 except for one of the assumptions which we discuss below. Indeed, Y', X' are independent since they are functions of disjoint parts of X, X' has standard Gaussian i.i.d. elements, $Y' = \sum_{i \in \text{Support}(\beta^*)-S} X_i + W$ has iid Gaussian elements with zero mean and variance $(k - |S|) + \sigma^2$, and the sparsity of β' is $k - |S|$. The only difference is that the ratio between the variance $(k - |S|) + \sigma^2$ and the sparsity $k - |S|$ is no longer necessarily upper bounded by 3, since this holds if and only if $\sigma^2 \leq 2(k - |S|)$, which does not hold necessarily, though it does hold in the special case $S = \emptyset$, provided $\sigma^2 \leq 2k$. Despite the absence of this assumption for general S we can still apply the lower bound (5) of Theorem 6, since the restriction on the relative value of the standard deviation of Y_i and other restrictions on p, n, k were needed only for the upper bound. Hence, applying the first part of Theorem 6 we conclude the optimal value $\phi_2(S)$ satisfies

$$\begin{aligned}&\mathbb{P}\left(\phi_2(S) \geq e^{-\frac{3}{2}}\sqrt{2(k-|S|)+\sigma^2}\exp\left(-\frac{(k-|S|)\log((p-k))}{n}\right)\right) \\ &\geq 1 - \exp(-n). \qquad (25)\end{aligned}$$

Also applying the second part of this theorem to the special case $S = \emptyset$ we obtain the following corollary for the case $\sigma^2 \leq 2k$.

Corollary 14 *Suppose $\sigma^2 \leq 2k$. For every $C > 0$ and every sufficiently large constant D_0, if $k \log k \leq Cn$, and $n \leq k\log(p-k)/(2\log D_0)$, the cardinality of the set*

$$\left\{\beta : n^{-\frac{1}{2}}\|Y' - X'\beta\|_2 \leq D_0\sqrt{2k+\sigma^2}\exp\left(-\frac{k\log(p-k)}{n}\right)\right\}$$

is at least $D_0^{\frac{n}{3}}$ w.h.p. as $k \to \infty$.

Proof [Proof of Theorem 1] Applying the union bound and (25) we obtain

$$\begin{aligned}&\mathbb{P}\left(\phi_2(\ell) \geq e^{-\frac{3}{2}}\sqrt{2\ell+\sigma^2}\exp\left(-\frac{\ell\log(p-k)}{n}\right), \ \forall\, 0 \leq \ell \leq k\right) \\ &\geq 1 - \sum_{0\leq \ell\leq k}\binom{k}{\ell}\exp(-n) \\ &\geq 1 - 2^k\exp(-n).\end{aligned}$$

Since $k\log k \leq Cn$, we have $2^k\exp(-n) \to 0$ as k increases. Replacing $p-k$ by a larger value p in the exponent we complete the proof of part (a) of the theorem.

We now establish the second part of the theorem. It follows almost immediately from Corollary 14. Since $k\log k \leq Cn$, the bound $n \leq k\log p/(3\log D_0)$ implies $\log k \leq C\log p/(3\log D_0)$

and in particular $k\log(p-k) = k\log p - O(\frac{k^2}{p})$ and $\frac{k^2}{p}$ converges to zero as k increases, provided D_0 is sufficiently large. Then we obtain $n \leq \exp(-k\log(p-k)/(2\log 2D_0))$ for all sufficiently large k. By a similar reason we may now replace $\exp(-k\log(p-k))$ by $\exp(-k\log p)$ in the upper bound on $n^{-\frac{1}{2}}\|Y' - X'\beta\|_2$ using the extra factor 2 in front of D_0. This completes the proof of the second part of the theorem. ■

Appendix C. The optimization problem Φ_2

In this part of the Appendix we give proofs of Proposition 4 and Theorem 3.

Proof [Proof of Proposition 4]

It is enough to study $f = \log\Gamma$ with respect to monotonicity. We compute the derivative for every $\zeta \in [0,1]$,

$$f'(\zeta) = -\frac{k\log p}{n} + \frac{k}{2\zeta k + \sigma^2} = -\frac{k}{n(2\zeta k + \sigma^2)}\left(\log p\left(2\zeta k + \sigma^2\right) - n\right).$$

Clearly, f' is strictly decreasing in ζ and $f'(\zeta) = 0$ has a unique solution $\zeta^* = \frac{1}{2k\log p}\left(n - \sigma^2\log p\right)$. Using the strictly decreasing property of f' and the fact that it has a unique root, we conclude that for $\zeta < \zeta^*$, $f'(\zeta) > 0$, and for $\zeta > \zeta^*$, $f'(\zeta) < 0$. As a result, if $\zeta^* \leq 0$ then f is a decreasing function on $[0,1]$, if $\zeta^* \geq 1$ f is an increasing function on $[0,1]$, and if $\zeta^* \in (0,1)$ then f is non monotonic. These cases are translated to the cases $n \leq \sigma^2\log p$, $n \geq (2k+\sigma^2)\log p$ and $n \in \left(\sigma^2\log p, (2k+\sigma^2)\log p\right)$, respectively. The minimum value achieved by f, and its dependence on n^* was already established earlier. ■

Proof [Proof of Theorem 3] We set

$$\Lambda_p \triangleq \text{argmin}_{\ell=0,1,..,k}\phi_2(\ell),$$

and we remind the reader that $\text{argmin}_{\ell=0,1,..,k}\phi_2(\ell) = k - |\text{Support}(\beta_2) \cap \text{Support}(\beta^*)|$.

Case 1: $n > (1+\epsilon)n^*$. Showing $\|\beta_2 - \beta^*\|_0/k \to 0$ as k increases is equivalent to showing

$$\frac{\Lambda_p}{k} \to 0,$$

w.h.p. as k increases. By the definition of Λ_p we have:

$$\phi_2(\Lambda_p) \leq \phi_2(0).$$

Recall the definition of function Γ from (3). From Theorem 1 we have that w.h.p. as k increases that $\phi_2(\Lambda_p) \geq e^{-\frac{3}{2}}\Gamma\left(\frac{\Lambda_p}{k}\right)$. Combining the above two inequalities we derive that w.h.p.:

$$e^{-\frac{3}{2}}\Gamma\left(\frac{\Lambda_p}{k}\right) \leq \phi_2(0). \tag{26}$$

Now from $Y = X\beta^* + W$ we have

$$\phi_2(0) = n^{-\frac{1}{2}}||Y - X\beta^*||_2 = n^{-\frac{1}{2}}||W||_2.$$

Hence,

$$\frac{1}{\sigma^2}\phi_2^2\left(0\right) = \frac{1}{\sigma^2}n^{-1}||W||_2^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i}{\sigma}\right)^2,$$

where W_i are i.i.d. $N\left(0,\sigma^2\right)$. But by the Law of Large Numbers, w.h.p. $\frac{1}{\sigma^2}\phi_2^2\left(0\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i}{\sigma}\right)^2$ is less than $4\mathbb{E}\left[\left(\frac{W_i}{\sigma}\right)^2\right] = 4$. Hence, since $\Gamma\left(0\right) = \sigma$, this means that w.h.p. as k (and therefore n) increases it holds:

$$\phi_2\left(0\right) \leq 2\sigma = 2\Gamma\left(0\right).$$

Combining this with (26) we get that w.h.p. as k increases

$$e^{-\frac{3}{2}}\Gamma\left(\frac{\Lambda_p}{k}\right) \leq 2\sigma,$$

or equivalently

$$e^{-\frac{3}{2}}\sqrt{2\Lambda_p+\sigma^2}e^{-\frac{\Lambda_p\log p}{n}} \leq 2\sigma,$$

which we rewrite as

$$e^{-\frac{3}{2}}\sqrt{\frac{2\Lambda_p}{\sigma^2}+1} \leq 2e^{\frac{\Lambda_p\log p}{n}}.$$

Now applying $n > \left(1+\epsilon\right)n^*$, we obtain,

$$2e^{\frac{\Lambda_p\log p}{n}} < 2e^{\frac{\Lambda_p\log p}{n^*(1+\epsilon)}} = 2\left(\frac{2k}{\sigma^2}+1\right)^{\frac{\Lambda_p}{2(1+\epsilon)k}}.$$

But $\Lambda_p \leq k$, and therefore

$$2\left(\frac{2k}{\sigma^2}+1\right)^{\frac{\Lambda_p}{2(1+\epsilon)k}} \leq 2\left(\frac{2k}{\sigma^2}+1\right)^{\frac{1}{2(1+\epsilon)}}.$$

Combining we obtain that w.h.p. as k increases,

$$e^{-\frac{3}{2}}\sqrt{\frac{2\Lambda_p}{\sigma^2}+1} \leq 2\left(\frac{2k}{\sigma^2}+1\right)^{\frac{1}{2(1+\epsilon)}},$$

which after squaring and rearranging gives w.h.p.,

$$\frac{2\Lambda_p}{\sigma^2} \leq 4e^3\left(\frac{2k}{\sigma^2}+1\right)^{\frac{1}{(1+\epsilon)}} - 1,$$

which we further rewrite as

$$\frac{\Lambda_p}{k} \leq \frac{\sigma^2}{2k}\left(4e^3\left(\frac{2k}{\sigma^2}+1\right)^{\frac{1}{(1+\epsilon)}} - 1\right). \tag{27}$$

We claim that this upper bound tends to zero, as $k \to +\infty$. Indeed, let $x_k = \frac{k}{\sigma^2}$. By the assumption of the theorem $x_k \to +\infty$. But the right-hand side of (27) can be upper bounded by a constant multiple of $x_k^{-1} x_k^{\frac{1}{1+\epsilon}} = x_k^{-\frac{\epsilon}{1+\epsilon}}$, which converges to zero as k increases. Therefore from (27), $\frac{\Lambda_p}{k} \to 0$ w.h.p. as k increases, and the proof is complete in that case.

Case 2: $n < (1-\epsilon)\, n^*$. We need to show that w.h.p. as k increases

$$\frac{\Lambda_p}{k} \to 1.$$

By the definition of Λ_p, $\phi_2(\Lambda_p) \le \phi_2(1)$. Again applying Theorem 1 we have that w.h.p. as k increases it holds $\phi_2(\Lambda_p) \ge e^{-\frac{3}{2}} \Gamma\left(\frac{\Lambda_p}{k}\right)$. Combining the above two inequalities we obtain that w.h.p.,

$$e^{-\frac{3}{2}} \Gamma\left(\frac{\Lambda_p}{k}\right) \le \phi_2(1). \tag{28}$$

Now we apply the second part of Theorem 1. Given any D_0 from part (b) of Theorem 1 and since $k/\sigma \to \infty$, we have that $n \le (1-\epsilon) n^*$ furthermore then satisfies $n \le k \log p / (3 \log D_0)$ for all sufficiently large k. We obtain that w.h.p. as k increases

$$\phi_2(1) \le D_0 \Gamma(1).$$

Using this in (28) and letting $c = 1/(e^{\frac{3}{2}} D_0)$ we obtain

$$c\Gamma\left(\frac{\Lambda_p}{k}\right) \le \Gamma(1),$$

namely,

$$c\sqrt{\frac{2\Lambda_p}{\sigma^2}+1}e^{-\frac{\Lambda_p \log p}{n}} \le \sqrt{\frac{2k}{\sigma^2}+1}e^{-\frac{k \log p}{n}},$$

and therefore

$$c^2\left(\frac{2\Lambda_p+\sigma^2}{2k+\sigma^2}\right) = c^2\left(\frac{\frac{2\Lambda_p}{\sigma^2}+1}{\frac{2k}{\sigma^2}+1}\right) \le e^{\frac{2(\Lambda_p-k)\log p}{n}}. \tag{29}$$

Now using $n \le (1-\epsilon)\, n^*$ and $\Lambda_p - k \le 0$, we obtain

$$e^{\frac{2(\Lambda_p-k)\log p}{n}} \le e^{\frac{2(\Lambda_p-k)\log p}{(1-\epsilon)n^*}} = \left(\frac{2k}{\sigma^2}+1\right)^{-\frac{k-\Lambda_p}{k(1-\epsilon)}}.$$

Combining the above with (29) we obtain that w.h.p.,

$$c^2\left(\frac{2\Lambda_p+\sigma^2}{2k+\sigma^2}\right) \le \left(\frac{2k}{\sigma^2}+1\right)^{-\frac{k-\Lambda_p}{k(1-\epsilon)}},$$

or w.h.p.,

$$c^2\left(\frac{2\Lambda_p}{\sigma^2}+1\right) \le \left(\frac{2k}{\sigma^2}+1\right)^{-\frac{\epsilon}{1-\epsilon}+\frac{\Lambda_p}{k(1-\epsilon)}}. \tag{30}$$

from which we obtain a simpler bound

$$c^2 \le \left(\frac{2k}{\sigma^2}+1\right)^{-\frac{\epsilon}{1-\epsilon}+\frac{\Lambda_p}{k(1-\epsilon)}},$$

namely

$$2\log c \le \left(-\frac{\epsilon}{1-\epsilon}+\frac{\Lambda_p}{k\,(1-\epsilon)}\right)\log\left(\frac{2k}{\sigma^2}+1\right)$$

or

$$\frac{2\log c}{\log\left(\frac{2k}{\sigma^2}+1\right)}(1-\epsilon)+\epsilon \le \frac{\Lambda_p}{k}.$$

Since by the assumption of the theorem we have $k/\sigma^2 \to \infty$, we obtain that $\frac{\Lambda_p}{k} \ge \epsilon/2$ w.h.p. as $k \to \infty$. Now we reapply this bound for (30) and obtain that w.h.p.

$$c^2\left(\frac{\epsilon k}{\sigma^2}+1\right) \le \left(\frac{2k}{\sigma^2}+1\right)^{-\frac{\epsilon}{1-\epsilon}+\frac{\Lambda_p}{k(1-\epsilon)}}.$$

Taking logarithm of both sides, we obtain that w.h.p.

$$(1-\epsilon)\log^{-1}\left(\frac{2k}{\sigma^2}+1\right)\left(\log\left(\frac{\epsilon k}{\sigma^2}+1\right)+2\log c\right)+\epsilon \le \frac{\Lambda_p}{k}.$$

Now again since $k/\sigma^2 \to \infty$, it is easy to see that the ratio of two logarithms approaches unity as k increases, and thus the limit of the left-hand side is $1-\epsilon+\epsilon=1$ in the limit. Thus Λ_p/k approaches unity in the limit w.h.p. as k increases. This completes the proof. ■

Appendix D. The Overlap Gap Property

In this part of the Appendix we prove Theorem 5. We begin by establishing a certain property regarding the the limiting curve function Γ.

Lemma 15 *Under the assumption of Theorem 5, there exist sequences* $0 < \zeta_{1,k,n} < \zeta_{2,k,n} < 1$ *such that* $\lim_k k\left(\zeta_{2,k,n}-\zeta_{1,k,n}\right) = +\infty$ *and such that for all sufficiently large* k

$$\inf_{\zeta\in\left(\zeta_{1,k,n},\zeta_{2,k,n}\right)} \min\left(\frac{\Gamma(\zeta)}{\Gamma(0)},\frac{\Gamma(\zeta)}{\Gamma(1)}\right) \ge e^3 D_0.$$

Proof Recall that $\Gamma(0)=\sigma$ and $\Gamma(1)=\sqrt{2k+\sigma^2}\exp\left(-\frac{k\log p}{n}\right)$. We will rely on the results of Proposition 4 and thus recall the definition of n^*. The proof proceeds by considering two cases.

Assume first $\left(e^7D_0^2+1\right)\sigma^2\log p \le n \le n^*$. We choose $\zeta_{1,k,n}=\frac{e^7D_0^2\sigma^2}{2k}, \zeta_{2,k,n}=\frac{e^7D_0^2\sigma^2}{k}$. Since by assumption $\sigma^2 \to +\infty$, then

$$k\left(\zeta_{2,k,n}-\zeta_{1,k,n}\right)=\frac{1}{2}e^7D_0^2\sigma^2 \to +\infty.$$

Now since $n < n^*$, and therefore by Proposition 4, $\Gamma(0) > \Gamma(1)$, it suffices to show $\frac{\Gamma(\zeta)}{\Gamma(0)} \ge e^3D_0$ for all $\zeta \in \left(\zeta_{1,k,n},\zeta_{2,k,n}\right)$. Since Γ is log-concave, it is sufficient to show that

$$\min\left(\frac{\Gamma\left(\zeta_{1,k,n}\right)}{\Gamma(0)},\frac{\Gamma\left(\zeta_{2,k,n}\right)}{\Gamma(0)}\right) \ge e^3D_0.$$

Plugging in the values $\Gamma\left(\zeta_{1,k,n}\right), \Gamma\left(\zeta_{2,k,n}\right)$, we have

$$\begin{aligned}
\min\left(\left(\frac{\Gamma\left(\zeta_{1,k,n}\right)}{\Gamma\left(0\right)}\right)^2, \left(\frac{\Gamma\left(\zeta_{2,k,n}\right)}{\Gamma\left(0\right)}\right)^2\right) &= \min\left(\left(e^7 D_0^2+1\right) e^{-\frac{e^7 D_0^2 \sigma^2 \log p}{2n}}, \left(2e^7 D_0^2+1\right) e^{-\frac{e^7 D_0^2 \sigma^2 \log p}{n}}\right) \\
&\geq \left(e^7 D_0^2+1\right) e^{-\frac{e^7 D_0^2 \sigma^2 \log p}{n}} \\
&\geq \left(e^7 D_0^2+1\right) e^{-1} \\
&\geq e^6 D_0^2,
\end{aligned}$$

using our assumption that $n \geq e^7 D_0^2 \sigma^2 \log p$.

Assume now $n^* \leq n < \frac{k \log p}{3 \log D_0}$. In this case we choose $\zeta_{1,k,n} = \frac{1}{5}$ and $\zeta_{2,k,n} = \frac{1}{4}$. Then $k\left(\zeta_{2,k,n} - \zeta_{1,k,n}\right) \to +\infty$. Since $n \geq n^*$ it suffices to show

$$\min\left(\left(\frac{\Gamma\left(\zeta_{1,k,n}\right)}{\Gamma\left(1\right)}\right)^2, \left(\frac{\Gamma\left(\zeta_{2,k,n}\right)}{\Gamma\left(1\right)}\right)^2\right) > e^6 D_0^2.$$

But since $n < k \log p / (3 \log D_0)$ have

$$\begin{aligned}
\min\left(\left(\frac{\Gamma\left(\zeta_{1,k,n}\right)}{\Gamma\left(1\right)}\right)^2, \left(\frac{\Gamma\left(\zeta_{2,k,n}\right)}{\Gamma\left(1\right)}\right)^2\right) &= \min\left(\frac{\frac{2k}{5}+\sigma^2}{2k+\sigma^2} e^{\frac{4k \log p}{5n}}, \frac{\frac{3k}{4}+\sigma^2}{2k+\sigma^2} e^{\frac{3k \log p}{4n}}\right) \\
&\geq \min\left(\frac{1}{4} D_0^{\frac{12}{5}}, \frac{2}{3} D_0^{\frac{9}{4}}\right) \\
&> e^6 D_0^2,
\end{aligned}$$

for all sufficiently large D_0. This completes the proof of the lemma. ■

Now we return to the proof of Theorem 5.

Proof [Proof of Theorem 5] Choose $0 < \zeta'_{1,k,n} < \zeta'_{2,k,n} < 1$ from Lemma 15 and set $r_k = D_0 \max\left(\Gamma\left(0\right), \Gamma\left(1\right)\right)$. We will now prove that for this value of r_k and $\zeta_{1,k,n} = 1 - \zeta'_{2,k,n}, \zeta_{2,k,n} = 1 - \zeta'_{1,k,n}$, the set S_{r_k} satisfies the claim of the theorem. Applying the second part of Theorem 1 we obtain $\beta^* \in S_{r_p}$ since $n^{-\frac{1}{2}} ||Y - X\beta^*||_2 = n^{-\frac{1}{2}} \sqrt{\sum_i W_i^2}$ which by the Law of Large Numbers is w.h.p. at most $2\sigma = 2\Gamma(0) < r_k$, provided D_0 is sufficiently large. This establishes (b). We also note that (c) follows immediately from Theorem 1.

We now establish part (a). Assume there exists a $\beta \in S_{r_p}$ with overlap $\zeta \in (\zeta_{1,k,n}, \zeta_{2,k,n})$. This implies that the optimal value of the optimization problem $\Phi_2(\ell)$ satisfies

$$\phi_2\left(k\left(1-\zeta\right)\right) \leq r_k. \tag{31}$$

Now $1 - \zeta \in (1 - \zeta_{2,k,n}, 1 - \zeta_{1,k,n}) = \left(\zeta'_{1,k,n}, \zeta'_{2,k,n}\right)$ and Lemma 15 imply

$$e^3 D_0 \max\{\Gamma\left(0\right), \Gamma\left(1\right)\} \leq \Gamma\left(1-\zeta\right).$$

We obtain

$$r_k \leq e^{-3} \Gamma\left(1-\zeta\right),$$

which combined with (31) contradicts the first part of Theorem 1. ■