Computational and statistical challenges in high dimensional statistical models

by

Ilias Zadik

B.A., National Kapodistrean University of Athens (2013)
M.A.St., University of Cambridge (2014)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

Author .....
Sloan School of Management
July 6, 2019

Certified by .....
David Gamarnik
Professor of Operations Research
Thesis Supervisor

Accepted by .....
Patrick Jaillet
Dugald C. Jackson Professor
Co-Director, Operations Research Center

Computational and statistical challenges in high dimensional statistical models

by

Ilias Zadik

Submitted to the Sloan School of Management
on July 6, 2019, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

This thesis focuses on two long-studied high-dimensional statistical models, namely

(1) the high-dimensional linear regression (HDLR) model, where the goal is to recover a hidden vector of coefficients from noisy linear observations, and

(2) the planted clique (PC) model, where the goal is to recover a hidden community structure from a much larger observed network.

The following results are established.

First, under assumptions, we identify the exact statistical limit of the model, that is the minimum signal strength allowing a statistically accurate inference of the hidden vector. We couple this result with an all-or-nothing information theoretic (IT) phase transition. We prove that above the statistical limit, it is IT possible to almost-perfectly recover the hidden vector, while below the statistical limit, it is IT impossible to achieve non-trivial correlation with the hidden vector.

Second, we study the computational-statistical gap of the sparse HDLR model; The statistical limit of the model is significantly smaller than its apparent computational limit, which is the minimum signal strength required by known computationally-efficient methods to perform statistical inference. We propose an explanation of the gap by analyzing the Overlap Gap Property (OGP) for HDLR. The OGP is known to be linked with algorithmic hardness in the theory of average-case optimization. We prove that the OGP for HDLR appears, up-to-constants, simultaneously with the computational-statistical gap, suggesting the OGP is a fundamental source of algorithmic hardness for HDLR.

Third, we focus on noiseless HDLR. Here we do not assume sparsity, but we make a certain rationality assumption on the coefficients. In this case, we propose a polynomial-time recovery method based on the Lenstra-Lenstra-Lóvasz lattice basis reduction algorithm. We prove that the method obtains notable guarantees, as it recovers the hidden vector with using only one observation.

Finally, we study the computational-statistical gap of the PC model. Similar to HDLR, we analyze the presence of OGP for the PC model. We provide strong (first-moment) evidence

that again the OGP coincides with the model's computational-statistical gap. For this reason, we conjecture that the OGP provides a fundamental algorithmic barrier for PC as well, and potentially in a generic sense for high-dimensional statistical tasks.

Thesis Supervisor: David Gamarnik
Title: Professor of Operations Research

In loving memory of my father, Pavlos Zadik (Dec. 1947 - Sept. 2014).

Acknowledgments

First and foremost, I would like to thank my advisor David Gamarnik. I was very lucky to be advised by him during the years of my PhD. During my time at MIT, I learned plenty of things and by far the most I learned from David. Beyond introducing me to numerous fascinating research questions and results across probability theory, statistical inference, statistical physics and computer science, my collaboration with David has taught me the value of optimism, patience and devotion to research. I wish that in the future I will be as good and influential an advisor to my students, as David was to me.

I would also like to thank Jennifer Chayes and Christian Borgs, who mentored me during my internship at Microsoft Research New England in the summer of 2017. In my time there, Jennifer and Christian, together with Adam Smith from Boston University, introduced me to the notion of differential privacy and its connection with network and random graphs estimation. I would like to thank all three of them for tireless long and technical discussions on the theory of privacy and random graphs, which have significantly improved my understanding on these two fascinating research topics. Many techniques and ideas used in this thesis were influenced by these discussions.

Over my PhD, I had the pleasure to collaborate on research with more people. I would like to thank all of them, I have learned something from each one of you! Specifically, I would like to thank Galen Reeves and Jiaming Xu for many conversations on the limits of sparse regression which lead to the creation of the second chapter of the present thesis, Yury Polyanskiy and Christos Thrampoulidis for exploring with me various fundamental questions on information theory, Vasilis Syrgkanis and Lester Mackey for working together on a beautiful problem on the theory of orthogonal machine learning, Juan Pablo Vielma and Miles Lubin for numerous discussions on integer programming and convex geometry and Patricio Foncea and Andrew Zheng for running certain very useful simulations for the work presented in the fifth chapter of this thesis. A big thank you also to Guy Bresler and Yury Polyanskiy for multiple mathematical discussions over these years both during their reading groups at LIDS and IDSS and while I was their teaching assistant for the graduate class "Moden Discrete Probability"; I have learned a lot from both of you. I would like to further acknowledge Guy Bresler and Lester Mackey for serving as members of my PhD thesis committee, it has been a pleasure sharing with you the

material included in the present thesis.

Doing my PhD at the Operations Research Center (ORC) at MIT has been an extremely influential experience for me. The ORC offers a unique and excellent interdisciplinary academic environment, which has broadened my perspective on mathematics and their applicability across various disciplines of science. A special acknowledgment to Dimitris Bertsimas, both for introducing me to the ORC more than six year ago, when I was still an undergraduate student at the Mathematics department of the University of Athens, and for being my mentor from that point forward. Finally, I would also like to thank all the people at the ORC who make it the unique place it is!

I would also like to express my gratitude to all the professors at the Mathematics Department of the University of Athens, for showing me the beauty of mathematics over the four years that I was an undergraduate student there!

On a more personal level, I would like to dedicate the present thesis to the memory of my beloved father, Pavlos Zadik. My father passed away very suddenly during the first month of my PhD studies at MIT. His strong values on the pursuit and importance of knowledge will always provide for me a strong guidance in life. I would also like to deeply thank my mother Nelli and my sister Mary who have always been next to me; without them none of this would have been possible. Finally, I would like to thank my cousin Joseph and his lovely family for always offering a warm place in Boston when it was needed.

Last but not least, I would like to thank all the friends I have made over these years, it has been a lot of fun! A special thanks goes to my dear friends and roomates Manolis, Konstantinos and Billy for building a warm family environment, wherever we lived together over the last five years. This has been truly an amazing experience!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The problem of statistical inference is one of the most fundamental tasks in the field of statistics. The question it studies is the following: assuming one has access to a dataset consisting of samples drawn from an unknown data distribution, can they infer structural properties of the underlying distribution? One of the earliest recorded examples of statistical inference methods can be traced back at least to the early 1800's. In 1801, Gauss introduced and used the least squares method, a now popular statistical method, to infer the orbits of celestial bodies [Mar77] (as a remark, the least squares method was introduced independently by Legendre in 1805 [Sti81]). In that way, Gauss had major impact in astronomy, as he guided the astronomers of the time to successfully infer the orbit of the newly-then discovered asteroid Ceres [Mar77].

During the 19th and 20th century, statistical inference established its existence as a mathematical field of study with the fundamental work of the statisticians Galton, Neyman, Pearson, Fisher and Yule among others (see e.g. some of their fundamental works [Gal85], [Yul97], [Fis22], [NP33]). Furthermore, the field shows an extensive study of classical statistical inference models such as regression, classification and (more recently) network models (see the associated chapters in the book [HTF09] and references therein). One common characteristic in most of this classic work, is that the statistical models considered are assumed to have a relatively small number of features and the focus is on creating statistical estimators which achieve asymptotically optimal performance as the sample size becomes arbitrarily large ("grows to infinity"). A common example of such an asymptotic property is statistical consistency, where an estimator is named consistent if it converges to some "fixed" true value, as the sample size grows [HTF09].

However, in recent years, mostly due to the emergence of the Big Data paradigm, there has been an explosion on the available data which are actively used for various statistical inference tasks across disciplines of science [BBHL09], [CCL+08], [LDSP08], [QMP+12], [PZHS16], [CWD16]. For example, this has proven a revolutionary fact for many scientific fields from biology [BBHL09], [CCL+08] to electrical engineering [QMP+12], [PZHS16] to social sciences [CWD16]. Naturally, though, the "explosion" of the available data leads to the "explosion" of the feature size which should be taken into account in the "high-dimensional" statistical inference models. This implies that the feature size should grow together with the sample size to infinity. On top of this, in many high dimensional statistical application, such as genomics [BBHL09], [CCL+08] and radar imaging [LDSP08], the feature size is not only comparable with the number of samples, but significantly larger than it. This is exactly the opposite regime to the one which is classically analyzed in statistical inference. These reasons lead to the recent research field of *high dimensional statistical inference*.

The study of high-dimensional inference is inherently connected with computational questions. The computational challenge is rather evident; the statistical algorithms are now defined on input domains of a very large size and therefore, to produce meaningful outputs in reasonable time, their termination time guarantees should be scalable with respect to the (potentially massive) input's size. Note that, with high dimensional input, this is a non-trivial consideration as many "textbook" statistically optimal algorithms usually take the form of an, in principle non-convex optimization problem. A standard example is the paradigm of maximum likelihood estimation.

High-dimensionality also leads to multiple statistical and modeling challenges. An important challenge is with respect to the techniques that can be used in that setting: both the classical version of the Central Limit Theorem [Nag76] and the Student-t test [FHY07] have been proven to fail in high dimensional cases. A case in point, which is highly relevant to the results in this thesis, is a modeling challenge in high dimensional linear regression. Specifically, consider the linear regression setting where the statistician observes n noisy linear samples of a hidden vector $\beta^* \in \mathbb{R}^p$ of the form $Y = X\beta^* + W$ for $X \in \mathbb{R}^{n\times p}$ and $W \in \mathbb{R}^n$. Note that here p corresponds to the number of features. The goal is to infer the hidden vector β^* from the pair (Y, X). High-dimensionality translates simply to $n < p$ and $p \to +\infty$. Note that the

moment high-dimensionality is imposed, a non-identifiability issue arises: even in the extremely optimistic case for the statistician where $W = 0$, β^* is simply one out of the infinitely many solutions of the underdetermined linear system $Y = X\beta^*$. This, in principle, makes inference in high dimensional linear regression impossible. In particular, additional assumptions need to be added to the regression model. For example, one of the standard assumptions in the literature of high dimensional linear regression is that the vector β^* is sparse, that is most of its entries are equal to zero. Under the sparsity assumption, accurate inference indeed becomes possible for n much smaller than p (see [HTW15] and references therein).

It becomes rather clear from the above discussion that the study of high dimensional statistical models require a novel study with respect to both its computational and statistical limits. Towards this goal a large body of recent research has been devoted to identifying those limits for various high dimensional statistical models. For example, the following high dimensional models have been analyzed in the literature: the sparse PCA problem, submatrix localization, RIP certification, rank-1 submatrix detection, biclustering, high dimensional linear regression, the tensor PCA problem, Gaussian mixture clustering and the stochastic block model (see [WX18], [BPW18] for two recent surveys and references therein for each model). We start by explicitly stating how the statistical and computational limits are defined for a high dimensional statistical inference problem.

For the statistical limit, the focus is on understanding the sampling complexity (or minimax rates) of the high dimensional statistical models. That is the focus is on the following question,

The statistical question: *What is the minimum necessary "signal strength" to perform an accurate statistical inference?*

We call the answer to the question above, the *statistical limit* of the model. Notice that to define statistical limit we assume unbounded computational power for the statistical estimators. For the computational limits, the focus is on computationally efficient estimators. For the results in this thesis we interpet computationally-efficient algorithms as algorithms with termination time being polynomial in the input dimensions. We focus on:

The computational question: *What is the minimum necessary "signal strength" to perform an accurate and computationally efficient statistical inference?*

We call the answer to the question above, the *computational limit* of the model. For many of the mentioned models, the accurate identification of the statistical and computational limits are far from being well-understood.

Despite being far from a complete theory, an interesting phenomenon has been repeatedly observed in the study of high-dimensional statistical models; the statistical limit of the problem appears usually significantly below the smallest known computational limit that is,

$$\text{statistical limit} \ll \text{computational limit}.$$

This phenomenon is called *a computational-statistical gap* [WX18], [BPW18]. Examples of models where computational-statistical gaps appears include, but are not limited to: the high-dimensional linear regression problem, the planted independent set problem and the planted dense subgraphs problems in sparse Erdős-Rényi graphs, the planted clique problem in dense Erdős-Rényi graphs, the Gaussian bi-clustering problem, the sparse rank-1 submatrix problem, the tensor decomposition problem, the sparse PCA problem, the tensor PCA problem and the stochastic block model (see [BBH18] and references therein).

Computational-statistical gaps provide a decomposition of the parameters space into three (possibly empty) regimes;

- (*the information-theoretic impossible regime*) The regime where the "signal strength" is less than the statistical limit, making inference impossible.

- (*the algorithmically easy regime*) The regime where the "signal strength" is larger than the computational limit so that the inference task is possible and is achieved by computationally efficient methods.

- (*the apparent algorithmically hard regime*) The regime where the "signal strength" is in between the statistical limit and the computational limit, and therefore the inference task is statistically possible but no computationally efficient method is known to succeed.

Note that the existence (or non-triviality) of the hard regime is equivalent with the presence of a computational-statistical gap for the model.

Towards understanding computational-statistical gaps, and specifically identifying the fundamentally hard region for various inference problems, a couple of approaches have been considered. One of the approaches seeks to identify the algorithmic limit "from above", in the sense of identifying the fundamental limits in the statistical performance of various families of known computationally efficient algorithms. Some of the families that have been analyzed are (1) the Sum of Squares (SOS) hierarchy, which is a family of convex relaxation methods [Par00], [Las01] (2) the family of local algorithms inspired by the Belief Propagation with the celebrated example of Approximate Message Passing [DMM09], [DJM13]), (3) the family of statistical query algorithms [Kea98] and (4) several Markov Chain Monte Carlo algorithms such as Metropolis Hasting and Glauber Dynamics [LPW06]. Another approach offers an average-case complexity-theory point of view [BR13], [CLR17], [WBP16], [BBH18]. In this line of work, the hard regimes of the various inference problems are linked by showing that solving certain high dimensional statistical problems in their hard regime reduces in a polynomial time to solving other high dimensional statistical problems in their own hard regime.

In this thesis, we build on a third approach to understand computational-statistical gaps. We study the geometry of the parameter space (we also call it solution space geometry for reasons that are to become apparent) and investigate whether a geometrical phase transition occurs between the easy and the hard regime.

The geometric point of view we follow is motivated from the study of average-case optimization problems, that is combinatorial optimization problems under random input. These problems are known to exhibit *computational-existential gaps*; that is there exists a range of values of the objective function which on the one hand are achievable by some feasible solution but on the other hand no computationally efficient method is proven to succeed. The link with the geometry comes out of the observation that for several average-case optimization problems (and their close relatives, random constraint satisfaction problems) an inspiring connection have been drawn between the geometry of the space of feasible solutions and their algorithmic difficulty in the regime where the computational-existential gap appears (the conjectured hard regime). Specifically it has been repeatedly observed that the conjectured algorithmically hard regime for the problem coincides with the appearance of a certain disconnectivity property in the solution space called the *Overlap Gap Property (OGP)*, originated in spin glass theory. Furthermore, it has also been

seen that at the absence of this property very simple algorithms, such as greedy algorithms can exploit the smooth geometry and succeed. The definition of OGP is motivated by the concentration of the associated Gibbs measures [Tal10] for low enough temperature to the optimization problem, and concerns the geometry of the near (optimal) feasible solutions. We postpone the exact definition of OGP to later chapters of this thesis. The connection between the hard regime for the optimization problem and the presence of OGP in the feasible space was initially made in the study of the celebrated example of random k-SAT (independently by [MMZ05], [ACORT11]) but then has been established for other models such as maximum independent set in random graphs [GSa], [RV14].

Note that contrary to statistical inference models, in average-case optimization problems there is no "planted" structured to be inferred and the goal is solely to maximize an objective value among a set of feasible solutions. For this reason, one cannot immediately transfer the literature on the Overlap Gap Property from computational-existential gaps to computational-statistical gaps. Nevertheless, one goal of this thesis is to make this possible by appropriately defining and using the Overlap Gap Property notion to study the computational-statistical gaps. In particular we are interested in the following question,

Can the Overlap Gap Property phase transition explain the appearance of computational-statistical gaps in statistical inference?

The goal of this thesis is to present results for the computational-statistical gaps of two well-studied and fundamental statistical inference problems: the high dimensional linear regression model and the planted clique model.

1.1 The Models: Definitions and Inference Tasks

In this subsection we describe the two high-dimensional statistical models this thesis is focusing on. Our goal to study their computational-statistical gaps.

1.1.1 The High Dimensional Linear Regression Model

As explained in the introduction, fitting linear regression models to perform statistical inference has been the focus of a lot of research work over the last two centuries. Recently the study

of linear regression has seen a revival of interest from scientists, because of the new challenge of high dimensionality, with applications ranging from compressed sensing [CT05], [Don06] to biomedical imaging [BLH+14], [LDSP08] to sensor networks [QMP+12], [PZHS16] (see also three recent books on the topic [Wai19], [HTW15], [FR13]).

Our first model of study is *the high dimensional linear regression model* which is a simplified and long-studied mathematical version of high dimensional linear regression. Despite its simplicity, the analysis of the model has prompted various important algorithmic and statistical developments in the recent years, for example the development of the LASSO algorithm [HTW15] and multiple compressed sensing methods [FR13].

We study the high dimensional linear regression model in Chapters 2, 3, 4 and 5.

Setting Let $n, p \in \mathbb{N}$. Let

$$Y = X\beta^* + W \tag{1.1}$$

where X is a data $n \times p$ matrix, W is a $n \times 1$ noise vector, and β^* is the (unknown) $p \times 1$ vector of regression coefficients. We refer to n as the number of samples of the model and p as the number of features for the model.

Inference Task The inference task is to recover β^* from having access only to the data matrix X and the noisy linear observations Y. The goal is to identify the following two fundamental limits of this problem

- the minimum n so that statistically accurate inference of β^* is possible by using any estimator (statistical limit) and
- the minimum n so that statistically accurate inference of β^* is possible by using a computationally efficient estimator, that is an estimator with worst case termination time being polynomial in n, p (computational limit).

Gaussian Assumptions on X, W Unless otherwise mentioned, we study the problem in the average case where (X, W) are generated randomly where X has iid $\mathcal{N}(0, 1)$ entries and W has iid $\mathcal{N}(0, \sigma^2)$. Here and everywhere below by $\mathcal{N}(\mu, \sigma^2)$ we denote the normal distribution on the real line with mean μ and variance σ^2. The model has been studied extensively under these

assumptions in the literature, see for example [EACP11], [JBC17], [Wai09b], [Wai09a], [WWR10] and the references in [HTW15, Chapter 11].

Parameters Assumptions The focus is on the high dimensional setting where $n < p$ and both $p \to +\infty$. The recovery should occur with probability tending to one, with respect to the randomness of X, W, as $p \to +\infty$ (w.h.p.). For the whole thesis, we assume that $p \to +\infty$ and the parameters n, k, σ^2 are sequences indexed by p, n_p, k_p, σ_p^2. The parameters n, k, σ^2 are assumed to grow or not to infinity, depending on the specific context.

Structural Assumptions on β^* As mentioned in the Introduction, the high-dimensional regime is an, in principle, impossible regime for (exact) inference of β^* from (Y, X) ; the underlying linear system, even at the extreme case $\sigma = 0$, is underdetermined. For this reason, following a large line of research, we study the model under the additional structural assumptions on the vector β^*.

Depending on the chapter we make different structural assumption on the vector of coefficients β^*. We mention here the two most common assumptions throughout the different Chapters of this thesis. Unless otherwise specified, we study the high dimensional linear regression model under these assumptions.

First, we assume that the vector of coefficients β^* is k-sparse, that is the support size of β^* (i.e. the number of regression coefficients with non-zero value) equals to some positive integer parameter k which is usually taken much smaller than p. Sparsity is a well-established assumption in the statistics and compressed sensing literature (see for exaple, the books [HTW15, FR13]), with various applications for example in biomedical imaging [BLH$^+$14], [LDSP08] and sensor networks [QMP$^+$12], [PZHS16].

Second, we assume that the non-zero regression coefficients β_i^* are all equal with each other and (after rescaling) equal to one; that is we assume we assume a binary $\beta^* \in \{0, 1\}^p$. From a theoretical point of view, we consider this more restrictive case to make possible a wider technical development and a more precise mathematical theory. From an applied point of view, the case of binary and more generally discrete-valued β^* has received a large interest in the study of wireless communications and information-theory literature [HB98], [HV02], [BB99], [GZ18], [TZP19], [ZTP19]. Finally, recovering a binary vector is equivalent with recovering the support of the

vector (indices of non-zero coordinates), which is a fundamental question in the literature of the model [TWY12], [OWJ11],[RG13],[Geo12], [Zha93], [MB06a] with various applications such as in gene selection in genomics [HC08], [HG10], [HY09] and radar signal processing [Dud17], [XZB01], [CL99].

Now, we would like to emphasize that in most cases we do not assume a prior distribution on β^*; we simply assume that β^* is an arbitrary fixed, yet unknown, structured vector (e.g. a binary and k-sparse vector). This is the setting of interest in Chapters 3, 4 and 5 where the exact structural assumptions are explicitly described. The only time we assume a prior distribution is on Chapter 2 where we assume that β^* is chosen according to a uniform prior over the space of binary k-sparse vectors.

1.1.2 The Planted Clique Model

Inferring a hidden community structure in large complex networks has been the focus on multiple recent statistical applications from cognitive science (brain modeling) to web security (worm propagation) to biology (protein interactions) and natural language processing [GAM$^+$18, ZCZ$^+$09, PDFV05]

A simplified, yet long-studied, mathematical model for community detection is *the planted clique model*, first introduced in [Jer92]. Despite its simplicity the model has motivated a large body of algorithmic work and is considered one of the first and most well-studied models for which a computational-statistical gaps appears [WX18, BPW18].

The planted clique model is studied in Chapter 6.

Setting Let $n, k \in \mathbb{N}$ with $k \leq n$. The statistician observes an n-vertex undirected graph G sampled in two stages. In the first stage, the graph is sampled according to an Erdős-Rényi graph $G\left(n, \frac{1}{2}\right)$, that is there are n vertices and each undirected edges is placed independently with probability $\frac{1}{2}$. In the second stage, k out of the n vertices are chosen uniformly at random and all the edges between these k vertices are deterministically added (if they did not already exist due to the first stage sampling). We call the second stage chosen k-vertex subgraph the *planted clique* $\mathcal{PC}$.

Inference Task The inference task of interest is to recover $\mathcal{PC}$ from observing G. The computational-statistical gap relies upon identifying the minimum $k = k_n$ so that inference of $\mathcal{PC}$ is possible by using an arbitrary estimator (statistical limit) and the minimum $k = k_n$ so that inference of $\mathcal{PC}$ is possible by a computationally efficient estimator (computational limit). The statistical limit of the inference task is well-known in the literature to equal $k = k_n = 2\log_2 n$. For this reason, in this thesis we focus on the computational limit of the model.

Parameters Assumptions The focus is on the asymptotic high-dimensional setting where both $k = k_n, n \to +\infty$ and the recovery should hold with probability tending to one as $n \to +\infty$ (w.h.p.).

1.2 Notation

For the rest of the Introduction we require the following mathematical notation.

For $p \in (0, \infty), d \in \mathbb{N}$ and a vector $x \in \mathbb{R}^d$ we use its $\mathcal{L}_p$-norm, $\|x\|_p := (\sum_{i=1}^{p} |x_i|^p)^{\frac{1}{p}}$. For $p = \infty$ we use its infinity norm $\|x\|_\infty := \max_{i=1,\ldots,d} |x_i|$ and for $p = 0$, its 0-norm $\|x\|_0 = |\{i \in \{1, 2, \ldots, d\} | x_i \neq 0\}|$. We say that x is k-sparse if $\|x\|_0 = k$. We also define the support of x, $\text{Support}\,(x) := \{i \in \{1, 2, \ldots, d\} | x_i \neq 0\}$. For $k \in \mathbb{Z}_{>0}$ we adopt the notation $[k] := \{1, 2, \ldots, k\}$. Finally with the real function $\log : \mathbb{R}_{>0} \to \mathbb{R}$ we refer everywhere to the natural logarithm. For $\mu \in \mathbb{R}, \sigma^2 > 0$ we denote by $\mathcal{N}(\mu, \sigma^2)$ the normal distribution on the real line with mean μ and variance σ^2. We use standard asymptotic notation, e.g. for any real-valued sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, $a_n = \Theta(b_n)$ if there exists an absolute constant $c > 0$ such that $\frac{1}{c} \leq |\frac{a_n}{b_n}| \leq c$; $a_n = \Omega(b_n)$ or $b_n = O(a_n)$ if there exists an absolute constant $c > 0$ such that $|\frac{a_n}{b_n}| \geq c$; $a_n = \omega(b_n)$ or $b_n = o(a_n)$ if $\lim_n |\frac{a_n}{b_n}| = 0$.

1.3 Prior Work and Contribution per Chapter

First, we consider *the high dimensional linear regression problem* defined in Subsection 1.1.1. As explained in the Introduction, our main goal is to study the existence and properties of the computational-statistical gap of the problem. We study it under the distributional assumptions

that X has iid $\mathcal{N}(0,1)$ entries and W has iid $\mathcal{N}(0,\sigma^2)$ for some parameter σ^2. Furthermore we assume that β^* is a binary k-sparse vector.

The focus of most of this thesis is on *sublinear sparsity levels*, that is, using asymptotic notation, $k = o(p)$. Nevertheless, before diving into specific contributions it is worth pointing out that a great amount of literature has been devoted on the study of the computational-statistical gap in the linear regime where $n, k, \sigma = \Theta(p)$. One line of work has provided upper and lower bounds on the minimum MSE (MMSE) $\mathbb{E}\left[\|\beta^* - \mathbb{E}\left[\beta^* \mid X, Y\right]\|_2^2\right]$ as a function of the problem parameters, e.g. [ASZ10, RG12, RG13, SC17]. Here and everywhere below for a vector $v \in \mathbb{R}^p$ we denote by $\|v\|_2 \triangleq \sqrt{\sum_{i=1}^p v_i^2}$ it's ℓ_2 norm. Another line of work has derived explicit formulas for the MMSE. These formulas were first obtained heuristically using the replica method from statistical physics [Tan02, GV05] and later proven rigorously in [RP16, BDMK16]. Finally, another line of work has provided nearly-optimal computationally efficient methods in this setting using Approximate Message Passing [DMM09, DJM13]. However, to our best of knowledge, most of the techniques used in the proportional regime cannot be used to establish similar results when $k = o(p)$ (with notable exceptions such as [RGV17]). Although there has been significant work focusing also on the sublinear sparsity regime, the exact identification of both the computational and the statistical limits in the sublinear regime, remained an outstanding open problem prior to the results of this thesis. We provide below a brief and targeted literature review, postponing a detailed literature review at the beginning of each Chapter, and provide a high-level summary of our contributions.

The statistical limit of High-Dimensional Linear Regression

We start our study with the statistical limit of the problem. To identify the statistical limit we adopt a Bayesian perspective and assume β^* is chosen from the uniform prior over the binary k-sparse vectors that is uniform on the set $\{\beta^* \in \{0,1\}^p : \|\beta^*\|_0 = k\}$.

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

$$\mathrm{MSE}\left(\hat{\beta}\right) \triangleq \mathbb{E}\left[\|\hat{\beta} - \beta^*\|_2^2\right],$$

where $\|v\|$ denotes the ℓ_2 norm of a vector v. In our setting, one can simply choose $\hat{\beta} = \mathbb{E}[\beta^*]$, which equals $\frac{k}{p}(1, 1, \ldots, 1)^\top$, and obtain a trivial $\mathrm{MSE}_0 = \mathbb{E}[\|\beta^* - \mathbb{E}[\beta^*]\|_2^2]$, which equals $k\left(1 - \frac{k}{p}\right)$. We will adopt the following two natural notions of recovery, by comparing the MSE of an estimator $\hat{\beta}$ to MSE_0.

Definition 1.3.1 (Strong and weak recovery)**.** *We say that $\hat{\beta} = \hat{\beta}(Y, X) \in \mathbb{R}^p$ achieves*

- *strong recovery if* $\limsup_{p\to\infty} MSE\left(\hat{\beta}\right)/MSE_0 = 0$*;*
- *weak recovery if* $\limsup_{p\to\infty} MSE\left(\hat{\beta}\right)/MSE_0 < 1$.

A series of results studies the statistical, or information-theoretic as it is also called, limit of both the strong and weak recovery problems have appeared in the literature. A crucial value for the sample size appearing in all such results when $k = o(p)$ is

$$n_{\mathrm{info}} = \frac{2k \log \frac{p}{k}}{\log\left(\frac{k}{\sigma^2} + 1\right)}. \tag{1.2}$$

For the impossibility direction, previous work [ASZ10, Theorem 5.2], [SC17, Corollary 2] has established that when $n \leq (1 - o(1))\, n_{\mathrm{info}}$, *strong* recovery, is information-theoretically impossible and if $n = o(n_{\mathrm{info}})$, *weak* recovery is impossible. For the achievability direction, Rad in [Rad11] has proven that for any $k = o(p)$ and $\sigma^2 = \Theta(1)$, there exist some large enough constant $C > 0$ such that if $n > Cn_{\mathrm{info}}$ then strong recovery is possible with high probability. In a similar spirit, [AT10, Theorem 1.5] shows that when $k = o(p)$, $k/\sigma^2 = \Theta(1)$, and $n > C_{k/\sigma^2} n_{\mathrm{info}}$ for some large enough $C_{k/\sigma^2} > 0$, it is information theoretically possible to weakly recover the hidden vector.

The literature suggests that the statistical limit is of the order $\Theta(n_{\mathrm{info}})$, but (1) it is only established in very restrictive regimes for the scaling of k, σ^2, (2) the distinction between weak and strong recovery is rather unclear and finally (3) the identification of the exact constant in front of n_{info} seems not to be well understood. These considerations raises the main question of study in Chapter 2;

Question 1: *What is the exact statistical limit for strong/weak recovery of β^*?*

We answer this question and establish that n_{info} is the *exact* statistical limit of the problem

in a very strong sense. We prove that assuming $k = o\left(\sqrt{p}\right)$, for any positive constant $\epsilon > 0$ if the signal-to-noise ratio k/σ^2 bigger than a sufficiently large constant then,

- when $n < (1-\epsilon)\, n_{\text{info}}$ *weak* recovery is impossible, but
- when $n > (1+\epsilon)\, n_{\text{info}}$ *strong* recovery is possible.

This establishes an "all-or-nothing statistical phase transition". To the best of our knowledge this is the first time such a phase transition is established for a high dimensional inference model.

The computational limit of High-Dimensional Linear Regression

We now turn to the study of the computational-limit for the high-dimensional linear regression model. For these results no prior distribution is assumed on β^* and β^* is assumed to be an arbitrary but fixed binary k-sparse vector. Note that this is a weaker assumption, in the sense that any with high probability property established under such an assumption, immediately transfers to any prior distribution for β^*.

The optimal sample size appearing in the best known computationally efficient results is

$$n_{\text{alg}} = \left(2k + \sigma^2\right) \log p. \tag{1.3}$$

More specifically, a lot of the literature has analyzed the performance of LASSO, the ℓ_1- constrained quadratic program:

$$\min_{\beta \in \mathbb{R}^p} \{\|Y - X\beta\|_2^2 + \lambda_p \|\beta\|_1\}, \tag{1.4}$$

for a tuning parameter $\lambda_p > 0$ [Wai09b], [MB06b], [ZY06], [BRT09a]. It is established that as long as $\sigma = \Theta(1)$ and k grows with p, if $n > (1+\epsilon)\, n_{\text{alg}}$ for some fixed $\epsilon > 0$, then for appropriately tuned λ_p the optimal solution of LASSO strongly recovers β^* [Wai09b]. Furthermore, a greedy algorithm called Orthogonal Matching Pursuit has also proven to succeed with $(1+\epsilon)\, n_{\text{alg}}$ samples [CW11]. To the best of our knowledge, besides decades of research efforts, no tractable (polynomial-in-n, p, k termination time) algorithms is known outperform LASSO when $k = o(p)$, in the sense of achieving strong recovery of β^* with $n \leq n_{\text{alg}}$ samples. This suggest that the sample size n_{alg} could correspond to the computational limit of the problem.

At this point, we would like to compare the thresholds n_{alg} and the information-theoretic limit n_{info} given in (1.2). Assuming the signal to noise ratio k/σ^2 is sufficiently large, which is considered in almost all the above results, n_{info} is significantly smaller than n_{alg}. This reason gives rise to the computational-statistical gap which motivates studying the following question in Chapters 3, 4:

Question 2: *Is there a fundamental reason for the failure of computationally efficient methods when* $n \in [n_{\text{info}}, n_{\text{alg}}]$*?*

We offer a geometrical explanation for the gap by identifying n_{alg} as the, up-to-constants, Overlap Gap Property phase transition point of the model. Specifically, we consider the least squares optimization problem defined by the Maximum Likelihood Estimation (MLE) problem;

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

As explained in the Introduction, geometry and the Overlap Gap Property has played crucial role towards understanding several computational-existential gaps for average-case optimization problems. Note that (MLE) is an average-case optimization problem with random input (Y, X). We study it first with respect to optimality; we prove that as long as $n \geq (1+\epsilon)\, n_{\text{info}}$, under certain assumption on the parameters, the optimal solution of (MLE) equals β^*, up to negligible Hamming distance error. Hence (MLE) is an average-case optimization problem with optimal solution almost equal to β^*. Thus, in light of the discussion above, it is expected to be algorithmically hard to solve to optimality when $n < n_{\text{alg}}$. For this reason we appropriately define and study the presence of the Overlap Gap Property (OGP) in the solution space of (MLE) in the regime $n \in [n_{\text{info}}, n_{\text{alg}}]$ and $n > n_{\text{alg}}$. We say that OGP holds in the solution space of (MLE) if the nearly optimal solutions of (MLE) have only either high or low Hamming distance to β^*, and therefore the possible Hamming distances ("overlaps") to β^* exhibit gaps. We direct the reader to Chapter 3 for the the exact definition of OGP and more references. We show that for some constants $c, C > 0$,

- when $n < cn_{\text{alg}}$ *OGP holds* in the solution space of (MLE) (Chapter 3), and

- when $n > Cn_{\text{alg}}$ *OGP does not hold* in the solution space of (MLE) (Chapter 4).

This provides evidence that the high dimensional linear regression recovery problem corresponds to an algorithmically hard problem in the regime $n < cn_{\text{alg}}$. In Chapter 3 we support the hardness conjecture by establishing that the LASSO not only provably fails to recover exactly the support in the regime $n < cn_{\text{alg}}$, but in the same regime it fails to achieve a different notion of recovery called ℓ_2-stable recovery. In Chapter 4 besides establishing that OGP does not hold, we also perform a direct local analysis of the optimization problem (MLE) proving that (1) when $n > Cn_{\text{alg}}$ the optimization landscape is extremely "smooth" to the extent that the only local minimum (under the Hamming distance metric between the binary β's) is the global minimum β^* and (2) the success of a greedy local search algorithm with arbitrary initialization for recovering β^* when $n > Cn_{\text{alg}}$.

The Noiseless High-Dimensional Linear Regression

An admittedly extreme, yet broadly studied, case in the literature of high dimensional linear regression is when the noise level σ is extremely small, or even zero (also known as the noiseless regime). In this case, the statistical limit of the problem n_{info}, defined in (1.2), trivializes to zero. In particular, our results in Chapter 2 imply that one can strongly recover information-theoretically a sparse binary p-dimensional β^* with $n = 1$ sample, as $p \to +\infty$. Notice that in this regime the statistical-computational gap becomes even more profound: $n_{\text{info}} = 1$ and when $\sigma = 0$ according to (1.3) $n_{\text{alg}} = 2k \log p$ remains of the order of $k \log p$. Moreover Donoho in [Don06] establishes that Basis Pursuit, a well-studied Linear Program recovery mechanism, fails in the noiseless regime $\sigma = 0$ with $n \leq (1 - \epsilon)n_{\text{alg}}$ samples for any fixed $\epsilon > 0$. On top of this, our Overlap Gap Property phase transition result described below in Question 2, suggests that the geometry of the sparse binary vectors is rather complex at $\sigma = 0$ for any n with $n_{\text{info}} = 1 \leq n < n_{\text{alg}}$.

On the other hand, the absence of noise makes the model significantly simpler ; it simply is an underdetermined linear system. The linear structure allows the suggestion and rigorous analysis of various computationally efficient mechanisms, moving potentially beyond the standard algorithmic literature of the linear regression model which, as explained in the Introduction, usually is based on either local or convex relaxation methods. These considerations lead to the

following question which is investigated in Chapter 5:

Question 3: *Is there a way to achieve computationally-efficient recovery with* $n < n_{\text{alg}} = 2k \log p$ *samples, in the "extreme" noiseless regime* $\sigma = 0$*?*

We answer this question affirmatively by showing that computational efficient estimation is possible even when $n = 1$. We do this by proposing a novel computationally efficient method using the celebrated Lenstra-Lenstra-Lovasz (LLL) lattice basis reduction algorithm (LLL was introduced in [LLL82] for factoring polynomials with rational coefficients). We establish that our recovery method can provably recover the vector β^* with access only to one sample, $n = 1$, and $p \to +\infty$. This profoundly breaks the algorithmic barrier n_{alg} of the local search and convex relaxation method (e.g. LASSO) in the literature. Our proposed algorithm heavily relies on the integrality assumption on the regression coefficients, which trivially holds since β^* is assumed to be binary-valued. In particular, as opposed to Chapters 2, 3, 4, we do not expect the results in Chapter 5 to generalize to the real-valued case. Interestingly, though, our algorithm does not depend at all to the sparsity assumption on β^* to be successful; it works for any integer-valued β^*. We consider the independence of our proposed algorithm to the sparsity assumption a fundamental reason for its success. In that way the algorithm does not need to "navigate" in the complex landscape of the binary sparse vectors where Overlap Gap Property holds, avoiding the conjectured algorithmic barrier in this case.

The computational-statistical gap of the Planted Clique Model

We now proceed with discussing our results for *the planted clique problem* defined in Subsection 1.1.2. As said in the definition of the model, our goal is to provide an explanation for the computational-statistical gap of this model as well.

The statistical limit of the model is exactly known in the literature to be $k = k_n = 2 \log_2 n$ (see e.g. [Bol85]): if $k < (2 - \epsilon) \log_2 n$ then it is impossible to recover $\mathcal{PC}$, but if $k \geq (2 + \epsilon) \log_2 n$ the recovery of $\mathcal{PC}$ is possible by a brute-force algorithm. Here and everywhere below by $\log_2$ we refer to the logarithm function with base 2. It is further known that if $k \geq (2 + \epsilon) \log_2 n$, a relatively simple quasipolynomial-time algorithm, that is an algorithm with termination time $n^{O(\log n)}$, also recovers $\mathcal{PC}$ correctly (see the discussion in [FGR+17] and references therein). On

the other hand, recovering $\mathcal{PC}$ with a computationally-efficient (polynomial-in-n time) method appears much more challenging. A fundamental work [AKS98] proved that a polynomial-time algorithm based on spectral methods recovers $\mathcal{PC}$ when $k \geq c\sqrt{n}$ for any fixed $c > 0$ (see also [FR10], [DM], [DGGP14] and references therein.) Furthermore, in the regime $k/\sqrt{n} \to 0$, various computational barriers have been established for the success of certain classes of polynomial-time algorithms [BHK$^+$16], [Jer92], [FGR$^+$17]. Nevertheless, no general algorithmic barrier such as worst-case complexity-theoretic barriers has been proven for recovering $\mathcal{PC}$ when $k/\sqrt{n} \to 0$. The absence of polynomial-time algorithms together with the absence of a complexity-theory explanation in the regime where $k \geq (2+\epsilon)\log_2 n$ and $k/\sqrt{n} \to 0$ gives rise to arguably one of the most celebrated and well-studied computational-statistical gaps in the literature, known as the *planted clique problem.*

As described below Question 2, in Chapters 3, 4 we carefully define and establish an Overlap Gap Property phase transition result for the high dimensional linear regression problem. In that way we provide a possible explanation for the conjectured algorithmic hardness. This suggests the following question which we study in Chapter 6.

Question 4: *Is there an Overlap Gap Property phase transition explaining the conjectured algorithmic hardness of recovering $\mathcal{PC}$ when $k/\sqrt{n} \to 0$?*

We provide strong evidence that the answer to the above question is affirmative. We consider the landscape of the dense subgraphs of the observed graph, that is subgraphs with nearly maximal number of edges. We study their possible intersection sizes ("overlaps") with the planted clique. Using the first moment method, as an non-rigorous approximation technique, we provide evidence of a phase transition for the presence of Overlap Gap Property (OGP) exactly at the algorithmic threshold $k = \Theta(\sqrt{n})$. More specifically, we say that OGP happens in the landscape of dense subgraphs of the observed graph if any sufficiently dense subgraph has either high or low overlap with the planted clique. We direct the reader to the Introduction of Chapter 6 for the the exact definition of OGP. We provide evidence that

- when $k = o(\sqrt{n})$ *OGP holds* in the landscape of dense subgraphs, and
- when $k = \omega(\sqrt{n})$ *OGP does not hold* in the landscape of dense subgraphs.

We prove parts of the conjecture such as the presence of OGP when k is a small positive power of n by using a conditional second moment method. We expect the complete proof of the conjectured OGP phase transition to be a challenging but important part towards understanding the algorithmic difficulty of the planted clique problem.

1.4 Technical Contributions

Various technical results are established towards proving the results presented in Section 1.3. In this Section, we describe two of the key technical results obtained towards establishing two of the most important results presented in this thesis: the presence of the Overlap Gap Property for the high dimensional linear regression mode (Chapter 3) and for the planted clique model (Chapter 6). The results are of fundamental nature and can be phrased independently from the Overlap Gap Property or any statistical context, and are of potential independent mathematical interest.

The Gaussian Closest Vector Problem

In Chapter 3 and the study of the Overlap Gap Property for the high dimensional linear regression model the following random geometry question naturally arises.

Let $n, p \in \mathbb{N}$ and $\mathcal{B} := \{\beta \in \{0,1\}^p : \|\beta\|_0 = k\}$ the set of all binary k-sparse vectors in $\mathbb{R}^p$. Suppose $X \in \mathbb{R}^{n\times p}$ has i.i.d. $\mathcal{N}(0,1)$ entries and $Y \in \mathbb{R}^{n\times p}$ has i.i.d. $\mathcal{N}(0,\sigma^2)$ entries. We would like to understand the asymptotic behavior of the following minimization problem,

$$\min_{\beta\in\mathcal{B}} n^{-1/2}\|Y - X\beta\|_2 \tag{1.5}$$

subject to specific scaling of σ^2, k, n, p as they grow together to infinity. In words, (1.5) estimates how well some vector of the form $X\beta, \beta \in \mathcal{B}$ approximate in (rescaled) ℓ_2 error a target vector Y.

The focus is on $\sigma^2 = \Theta(k)$, which makes the per-coordinate variance of Y, which equals to σ^2, comparable with the per-coordinate variance of $X\beta$, which equals k. Studying the extrema of Gaussian processes, such as (1.5), has a vast literature in probability theory and the develop-

ment of fundamental tools such as Slepian's, Sudakov-Fernique and Gordon's inequalities [Ver18, Section 7.2], with multiple applications in spin glass theory [Tal10], compressed sensing [OTH13] and in information theory [ZTP19, TZP19]. The problem can also be motivated in statistics as an "overfitting test"; it corresponds to the fundamental limits of fitting a sparse binary linear model to an independent random vector of observations.

Now if $n = \Omega\left(k \log \frac{p}{k}\right)$ then a well-studied random matrix property, called the Restricted Isometry Property (RIP) is known to hold for the random matrix X. The RIP states that for any k-sparse vector v, it holds $n^{-1/2}\|Xv\|_2 \approx \|v\|_2$ with probability tending to one as $n, p, k \to +\infty$ (see e.g. [FR13, Chapter 6]). Here and everywhere in this Section, the approximation sign should be understood as equality up to a multiplicative constant. Now, if $n = \Omega\left(k \log \frac{p}{k}\right)$, using the RIP and $\sigma^2 = \Theta(k)$ it is a relatively straightforward exercise that

$$\min_{\beta \in \mathcal{B}} n^{-1/2}\|Y - X\beta\|_2 \approx \sqrt{k + \sigma^2}, \tag{1.6}$$

as $n, p, k \to +\infty$. Here $\sqrt{k + \sigma^2}$ simply corresponds to the variance per coordinate of any vector of the form $Y - X\beta$ for $\beta \in \mathcal{B}$.

On the other hand, when $n = o\left(k \log \frac{p}{k}\right)$ X is known not to satisfy RIP and to the best of our knowledge no other tool provides tight results for the value of the optimization problem (1.5). In Chapter 3 we study the following question:

Question: *Which value does (1.5) concentrate on when $n = o\left(k \log \frac{p}{k}\right)$?*

We answer this question under the assumption that n satisfies $n \leq ck \log \frac{p}{k}$ for some small constant $c > 0$ but also $k \log k \leq n$. Notice that naturally restricts k to be at most $p^{\frac{c}{1+c}}$ for the small constant $c > 0$. We show under these assumptions that

$$\min_{\beta \in \mathcal{B}} n^{-1/2}\|Y - X\beta\|_2 \approx \sqrt{k + \sigma^2} \exp\left(-\frac{k \log \frac{p}{k}}{n}\right),$$

as $n, p, k \to +\infty$. Comparing with (1.6) we see that the behavior of (1.5) in the regime $n = o\left(k \log \frac{p}{k}\right)$ differs from the regime $n = \Omega\left(k \log \frac{p}{k}\right)$ by an exponential factor in $-\frac{k \log \frac{p}{k}}{n}$. The exact statement and proof of the above result can be found in Section 3.3 of Chapter 3.

The Densest Subgraph Problem in Erdős-Rényi random graphs

In Chapter 6 and the study of the Overlap Gap Property for the planted clique model the following random graph theory question naturally arises. Consider an n-vertex undirected graph G sampled from the Erdős-Rényi model $G\left(n, \frac{1}{2}\right)$, that is each edge appears independently with probability $\frac{1}{2}$. We would like to understand the concentration properties of the K-Densest subgraph problem,

$$d_{\mathrm{ER},K}(G) = \max_{C \subseteq V(G), |C|=K} |E[C]|, \tag{1.7}$$

where $V(G)$ denotes the set of vertices of G and for any $C \subseteq V(G)$, $E[C]$ denotes the set of induced edges in G between the vertices of C. In words, (1.7) is the maximum number of edges of a K-vertex subgraph of G.

The study of $d_{\mathrm{ER},K}(G)$ is an admittedly natural question in random graph theory which, to the best of our knowledge, remains not well-understood even for moderately large values of $K = K_n$. It should be noted that this is in sharp contrast with other combinatorial optimization questions in Erdős-Rényi graphs, such as the size of the maximum clique or the chromatic number, where tight results are known [Bol85, Chapter 11].

We describe briefly the literature of the problem. For small enough values of K, specifically $K < 2\log_2 n$, it is well-known that a clique of size K exists and therefore $d_{\mathrm{ER},K}(G) = \binom{K}{2}$ w.h.p. as $n \to +\infty$ [GM75]. On the other hand when $K = n$, trivially $d_{\mathrm{ER},K}(G)$ follows Binom $\left(\binom{K}{2}, \frac{1}{2}\right)$ and hence for any $\alpha_K \to +\infty$, $d_{\mathrm{ER},K}(G) = \frac{1}{2}\binom{K}{2} + O\left(K\alpha_K\right)$ w.h.p. as $n \to +\infty$. In Chapter 6 we study the following question:

Question: *How does $d_{\mathrm{ER},K}(G)$ behave asymptotically when $2\log_2 n \leq K = o(n)$?*

A recent result in the literature studies the case $K = C\log n$ for $C > 2$ [BBSV18] and establishes (it is an easy corollary of the main result of [BBSV18]),

$$d_{\mathrm{ER},K}(G) = h^{-1}\left(\log 2 - \frac{2\left(1 + o\left(1\right)\right)}{C}\right)\binom{k}{2}, \tag{1.8}$$

w.h.p. as $n \to +\infty$. Here log is natural logarithm and h^{-1} is the inverse of the (rescaled) binary entropy $h : [\frac{1}{2}, 0] \to [0, 1]$ is defined by $h(x) = -x\log x - (1-x)\log x$. Notice that $\lim_{C\to+\infty} h^{-1}\left(\log 2 - \frac{2(1+o(1))}{C}\right) = \frac{1}{2}$ which means that the result from [BBSV18] agrees with the

first order behavior of $d_{\mathrm{ER},K}(G)$ at " very large" K such as $K = n$.

In Chapter 6 we obtain results on the behavior of $d_{\mathrm{ER},K}(G)$ for any $K = n^C$, for $C \in (0, 1/2)$. Specifically in Theorem 6.2.10 we show that for any $K = n^C$ for $C \in (0, 1/2)$ there exists some positive constant $\beta = \beta(C) \in (0, \frac{3}{2})$ such that

$$d_{\mathrm{ER},K}(G) = h^{-1}\left(\log 2 - \frac{\log\binom{n}{K}}{\binom{K}{2}}\right)\binom{K}{2} - O\left(K^{\beta}\sqrt{\log n}\right) \tag{1.9}$$

w.h.p. as $n \to +\infty$.

First notice that as our result are established when K is a power n it does not apply in the logarithmic regime. Nevertheless, it is in agreement with the result of of [BBSV18] since for $K = C \log n$,

$$\frac{\log\binom{n}{K}}{\binom{K}{2}} = (1 + o(1))\frac{K \log\left(\frac{n}{K}\right)}{\frac{K^2}{2}} = (1 + o(1))\frac{2}{C},$$

that is the argument in h^{-1} of (1.9) converges to the argument in h^{-1} of (1.8) at this scaling.

Finally, by Taylor expanding h^{-1} around $\log 2$ and using (1.9) we can identify the second order behavior of $d_{\mathrm{ER},K}(G)$ for $K = n^C$, for $C \in (0, 1/2)$ to be,

$$d_{\mathrm{ER},K}(G) = \frac{1}{2}\binom{K}{2} + \frac{K^{\frac{3}{2}}\sqrt{\log\left(\frac{n}{K}\right)}}{2} + o\left(K^{\frac{3}{2}}\right),$$

w.h.p. as $n \to +\infty$.

The exact statements and proofs of the above results can be found in Chapter 6.

1.5 Organization and Bibliographic Information

Most results described in this thesis have already appeared in existing publications, which we briefly mention below.

Chapter 2 presents new results on the statistical limit of the high dimensional linear regression model. It presents an exact calculation of the statistical limit of the model, and reveals a strong "all-to-nothing" information-theoretic phase transition of the model. The results of this Chapter are included in the paper with title "The All-or-Nothing Phenomenon in Sparse Linear Regression" which is joint work with Galen Reeves and Jiaming Xu and appeared in the

Proceedings of the Conference on Learning Theory (COLT) 2019 [RXZ19].

Chapter 3 establishes the presence of the Overlap Gap Property in the conjectured hard regime for high dimensional linear regression. This is based on the paper "High dimensional linear regression with binary coefficients: Mean squared error and a phase transition" which is joint work with David Gamarnik and appeared in the Proceedings of the Conference on Learning Theory (COLT) 2017 [GZ17a].

Chapter 4 proves the absence of the Overlap Gap Property in the algorithmically easy regime for high dimensional linear regression. Moreover, it shows the success of a greedy Local Search method also in the easy regime. This is based on the paper "Sparse High Dimensional Linear Regression: Algorithmic Barrier and a Local Search Algorithm" which is joint work with David Gamarnik and is currently available as a preprint [GZ17b].

Chapter 5 offers a new computationally-efficient recovery method for noiseless high dimensional linear regression using lattice basis reduction. The algorithm provably infers the hidden vector even with access to only one sample. This is based on the paper "High dimensional linear regression using Lattice Basis Reduction" which is joint work with David Gamarnik and appeared in the Advances of Neural Information Processing Systems (NeurIPS) 2018 [GZ18].

Chapter 6 presents strong evidence that Overlap Gap Property for the planted clique model appears exactly at the conjectured algorithmic hard regime for the model. It also offers a proof of the presence of the Overlap Gap Property in a part of the hard regime. This is based on the paper "The Landscape of the Planted Clique Problem: Dense Subgraphs and the Overlap Gap Property" which is joint work with David Gamarnik and is currently available as a preprint [GZ19].

Other papers by the author over the course of his PhD that are not included in this thesis are [ZTP19, TZP19, LVZ17, MSZ18, LVZ18, BCSZ18].

Chapter 2

The Statistical Limit of High Dimensional Linear Regression. An All-or-Nothing Phase Transition.

2.1 Introduction

In this Chapter, we study the statistical, or information-theoretic, limits of the high-dimensional linear regression problem (defined in Subsection 1.1.1) under the following assumptions. For $n, p, k \in \mathbb{N}$ with $k \leq p$ and $\sigma^2 > 0$ we consider two independent matrices $X \in \mathbb{R}^{n \times p}$ and $W \in \mathbb{R}^{n \times 1}$ with $X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $W_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, and observe

$$Y = X\beta^* + W, \tag{2.1}$$

where β^* is assumed to be uniformly chosen at random from the set $\{v \in \{0,1\}^p : \|v\|_0 = k\}$ and independent of (X, W). The problem of interest is to recover β^* given the knowledge of X and Y. Our focus will be on identifying the minimal sample size n for which the recovery is information-theoretic possible.

The problem of recovering the support of a hidden sparse vector $\beta^* \in \mathbb{R}^p$ given noisy linear observations has been extensively analyzed in the literature, as it naturally arises in many contexts including subset regression, e.g. [CH90], signal denoising, e.g. [CDS01], compressive

sensing, e.g. [CT05], [Don06], information and coding theory, e.g. [JB12], as well as high dimensional statistics, e.g. [**?**, Wai09b]. The assumptions of Gaussianity of the entries of (X, W) are standard in the literature. Furthermore, much of the literature (e.g. [ASZ10], [NT18], [WWR10]) assumes a lower bound $\beta^*_{\min} > 0$ for the smallest magnitude of a nonzero entry of β^*, that is $\min_{i:\beta^*_i \neq 0} |\beta^*_i| \geq \beta^*_{\min}$, as otherwise identification of the support of the hidden vector is in principle impossible. In this Chapter we adopt a simplifying assumption by focusing only on binary vectors β^*, similar to other papers in the literature such as [ASZ10], [GZ17a] and [GZ17b]. In this case recovering the support of the vectors is equivalent to identifying the vector itself.

To judge the recovery performance we focus on the mean squared error (MSE) and specifically the notions of weak and strong recovery introduced in Definition 1.3.1. The fundamental question of interest in this Chapter is when n as a function of (p, k, σ^2) is such that strong/weak recovery is information-theoretically possible. Obtaining a tight characterization of the statistical limit, or the information-theoretic limit as it is also called, for these notions of recovery is the main contribution of the work described in this Chapter. Note that here and for rest of the Chapter, to avoid confusion of using them both, we stick with the term "information-theoretic limit" instead of "statistical limit".

Towards identifying the information theoretic limits of recovering β^*, and out of independent interest, we also consider a closely related hypothesis testing problem, where the goal is to distinguish the pair (X, Y) generated according to (2.1) from a model where both X and Y are independently generated. More specifically, given two independent matrices $X \in \mathbb{R}^{n\times p}$ and $W \in \mathbb{R}^{n\times 1}$ with $X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and $W_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, we define

$$Y \triangleq \lambda W, \tag{2.2}$$

where $\lambda > 0$ is a scaling parameter. We refer to the Gaussian linear regression model (2.1) as the planted model, denoted by $P = P(X, Y)$, and (2.2) as the null model denoted by $Q_\lambda = Q_\lambda(Y, X)$. We focus on characterizing the total variation distance $\mathrm{TV}(P, Q_\lambda)$ for various values of λ. One choice of particular interest is $\lambda = \sqrt{k/\sigma^2 + 1}$, under which $\mathbb{E}\left[YY^\top\right] = (k + \sigma^2)\mathbf{I}$ in both the planted and null models.

Analogous to recovery definitions 1.3.1, we adopt the following two natural notions of testing [PWB16, AKJ17].

Definition 2.1.1 (Strong and weak detection)**.** *Fix two probability measures $\mathbb{P}, \mathbb{Q}$ on our observed data (Y, X). We say a test statistic $\mathcal{T}(X, Y)$ with a threshold τ achieves*

- *strong detection if*

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

- *weak detection, if*

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

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

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

Thus equivalently, strong detection is possible if and only if $\liminf_{p\to\infty} \mathrm{TV}(\mathbb{P}, \mathbb{Q}) = 1$, and weak detection is possible if and only if $\liminf_{p\to\infty} \mathrm{TV}(\mathbb{P}, \mathbb{Q}) > 0$. The fundamental question of interest is when n as a function of (p, k, σ^2) is such that strong/weak detection is information-theoretically possible.

2.1.1 Contributions

Of fundamental importance is the following sample size:

$$n_{\mathrm{info}} \triangleq \frac{2k \log(p/k)}{\log(1 + k/\sigma^2)}. \tag{2.3}$$

We establish that n_{info} is a sharp phase transition point for the recovery of β^* when $k = o(\sqrt{p})$ and the signal to noise ratio k/σ^2 is above a sufficiently large constant. In particular, for an arbitrarily small but fixed constant $\epsilon > 0$, when $n < (1 - \epsilon)n_{\mathrm{info}}$, *weak recovery* is impossible, but when $n > (1 + \epsilon)n_{\mathrm{info}}$, *strong recovery* is possible. This implies that the rescaled MMSE undergoes a jump from 1 to 0 at n_{info} samples up to a small window of size ϵn.

We state this in the following Theorem, which summarizes the Theorems 2.2.3, 2.2.4, 2.2.5 and 2.2.6 from the main body of the Chapter.

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

- *When $k \leq p^{\frac{1}{2}-\delta}$ and*

$$n < (1 - \epsilon)\, n_{\text{info}},$$

  both weak recovery of β^* from $(Y, X) \sim P$ and weak detection between P and Q_{λ_0} are information-theoretically impossible, where $\lambda_0 = \sqrt{\frac{k}{\sigma^2} + 1}$.

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

$$n > (1 + \epsilon)\, n_{\text{info}},$$

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

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

Note that the theorem above assumes $\sigma > 0$. In the extreme case where $\sigma = 0$, n_{info} trivializes to zero and we can directly argue that one sample suffices for strong recovery. In fact, for any $\beta^* \in \{0, 1\}^p$ and $Y_1 = \langle X_1, \beta^* \rangle$ for $X_1 \sim \mathcal{N}(0, \mathbf{I}_p)$, we can identify β^* as the unique binary-valued solution of $Y_1 = \langle X_1, \beta^* \rangle$, almost surely with respect to the randomness of X (see e.g. [GZ18])

Note that the first part of the above result focuses on $k \leq p^{1/2-\delta}$. It turns out that this is not a technical artifact and $k = o\left(p^{1/2}\right)$ is needed for n_{info} to be the weak detection sample size threshold. More details can be found in 2.9. The sharp information-theoretic threshold for either detection or recovery is still open when $k = \Omega\left(p^{1/2}\right)$ and $k = o(p)$.

The phase transition role of n_{info} According to our main result, the rescaled minimum mean squared error of the problem, MMSE/MSE_0, exhibits a step behavior asymptotically. Loosely speaking, when $n < n_{\text{info}}$ it equals to one and when $n > n_{\text{info}}$ it equals to zero. We next intuitively explain why such a step behavior for sparse high dimensional regression occurs at n_{info}, using ideas

related to *the area theorem*. The area theorem has been used in the channel coding literature to study the MAP decoding threshold [MMU08] and the capacity-achieving codes [KKM+17]. The approach described below is similar to the one used previously for linear regression [RP16].

First let us observe that n_{info} is asymptotically equal to the *ratio* of entropy $H(\beta^*) = \log \binom{p}{k}$ and Gaussian channel capacity $\frac{1}{2}\log(1+k/\sigma^2)$. We explore this coincidence in the following way. Let $I_n \triangleq I(Y_1^n; X, \beta^*)$ denote the mutual information between β^* and $(Y_1^n; X)$ with a total of n linear measurements. Since the mutual information in the Gaussian channel under a second moment constraint is maximized by the Gaussian input distribution, it follows that the increment of mutual information $I_{n+1} - I_n \le \frac{1}{2}\log(1 + \text{MMSE}_n/\sigma^2)$, where MMSE_n denotes the minimum MSE with n measurements. In particular, all the increments are between zero and $\frac{1}{2}\log(1+k/\sigma^2)$ and by telescopic summation for any n:

$$I_n \le \frac{n}{2}\log(1 + k/\sigma^2), \tag{2.4}$$

with equality only if for all $m < n$, $\text{MMSE}_m = k$. This is illustrated in Figure 2-1 where we plot n against $I_{n+1} - I_n$.

Suppose now that we have established that strong recovery is achieved with $n_{\text{info}} = \frac{H(\beta^*)}{\frac{1}{2}\log(1+k/\sigma^2)}$ samples.

Then strong recovery and standard identities connecting mutual information and entropy implies that

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

In particular, (2.4) holds with equality, which means for all $n \le n_{\text{info}} - 1$, $\text{MMSE}_n = k$. In particular, for all $n < n_{\text{info}}$, weak recovery is impossible. This area theorem is the key underpinning our converse proof of the weak recovery.

2.1.2 Comparison with Related Work

The information-theoretic limits of high-dimensional sparse linear regression have been studied extensively and there is a vast literature of multiple decades of research. In this section we focus solely on the Gaussian and binary setting and furthermore on the results applying to high values of signal to noise ratio and sublinear sparsity.

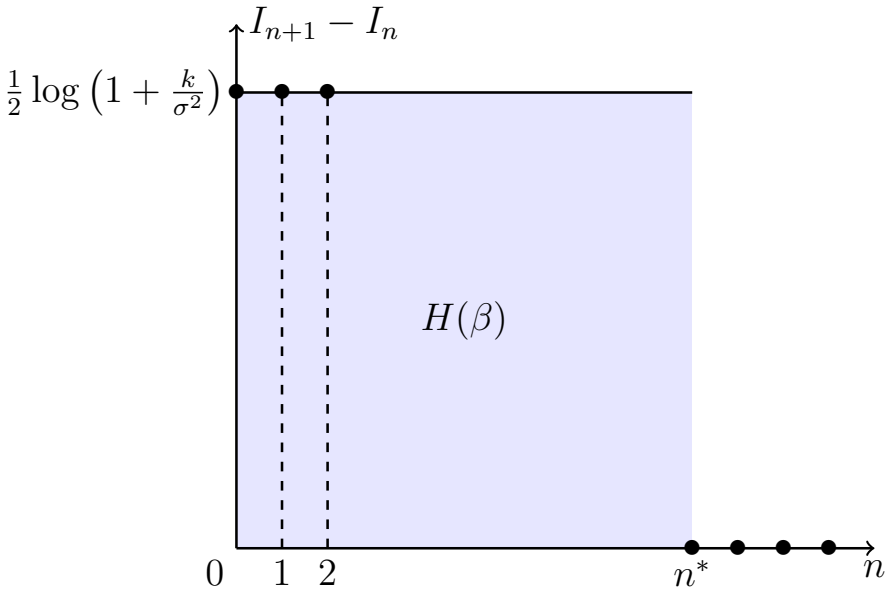


Figure 2-1: The phase transition diagram in Gaussian sparse linear regression. The y-axis is the increment of mutual information with one additional measurement. The area of blue region equals the entropy $H(\beta^*) \sim k\log(p/k)$. Here by n^* we denote the n_{info}.

Information-theoretic Negative Results for weak/strong recovery For the impossibility direction, previous work [ASZ10, Theorem 5.2] has established that as $p \to \infty$, achieving $\text{MSE}(\hat{\beta}^*) \le d$ for any $d \in [0, k]$ is information-theoretically impossible if

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

where $h_2(\alpha) = -\alpha\log\alpha - (1-\alpha)\log(1-\alpha)$ for $\alpha \in [0,1]$ is the binary entropy function. This converse result is proved via a simple rate-distortion argument (see, e.g. [WX18] for an exposition). In particular, given any estimator $\hat{\beta}^*(X, Y)$ with $\text{MSE}(\hat{\beta}^*) \le d$, we have

$$p\left(h_2(k/p) - h_2(d/p)\right) \le \inf_{\text{MSE}(\hat{\beta}^*)\le d} I(\tilde{\beta}^*; \beta^*) \le I(\hat{\beta}^*; \beta^*) \le I(X, Y; \beta^*) \le \frac{n}{2}\log\left(1 + k/\sigma^2\right).$$

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

More recent work [SC17, Corollary 2] further quantified the fraction of support that can be recovered when $n < (1-\epsilon)n_{\text{info}}$ for some fixed constant $\epsilon > 0$. Specifically with $k = o(p)$ and any

scaling of k/σ^2, if $n < (1-\epsilon)n_{\text{info}}$, then the fraction of the support of β^* that can be recovered correctly is at most $1-\epsilon$ with high probability; thus strong recovery is impossible.

Restricting to the Maximum Likelihood Estimator (MLE) performance of the problem, it is shown in [GZ17a] that under significantly small sparsity $k = O\left(\exp\left(\sqrt{\log p}\right)\right)$ and $k/\sigma^2 \to +\infty$, if $n \leq (1-\epsilon)n_{\text{info}}$, the MLE not only fails to achieve strong recovery, but also fails to weakly recover the vector, that is recover correctly any positive constant fraction of the support.

Our result (Theorem 2.2.4) establishes that the MLE performance is fundamental. It improves upon the negative results in the literature by identifying a sharp threshold for weak recovery, showing that if $k = o\left(\sqrt{p}\right)$, $k/\sigma^2 \geq C$ for some large constant $C > 0$, and $n \leq (1-\epsilon)\, n_{\text{info}}$, then *weak* recovery is information-theoretically impossible by any estimator $\hat{\beta}^*(Y, X)$. In other words, no constant fraction of the support is recoverable under these assumptions.

Information-theoretic Positive Results for weak/strong recovery In the positive direction, previous work [AT10, Theorem 1.5] shows that when $k = o(p)$, $k/\sigma^2 = \Theta(1)$, and $n > C_{k/\sigma^2} k \log(p-k)$ for some C_{k/σ^2}, it is information theoretically possible to weakly recover the hidden vector.

Albeit very similar to our results, our positive result (Theorem 2.2.5) identifies the explicit value of C_{k/σ^2} for which both weak and strong recovery are possible, that is $C_{k/\sigma^2} = 2/\log\left(1 + k/\sigma^2\right)$ for which $C_{k/\sigma^2} k \log(p/k) = n_{\text{info}}$.

In [GZ17a] it is shown that when $k = O\left(\exp\left(\sqrt{\log p}\right)\right)$ and $k/\sigma^2 \to +\infty$ then if $n \geq (1+\epsilon)n_{\text{info}}$ for some fixed $\epsilon > 0$, *strong* recovery is achieved by the MLE of the problem. We improve upon this result with Theorem 2.2.5 by showing that when $n \geq (1+\epsilon)n_{\text{info}}$ for some fixed $\epsilon > 0$ and any $k \leq cp$ for some $c > 0$, then there exists a constant $C > 0$ such that $k/\sigma^2 \geq C$ the MLE achieves *strong* recovery. In particular, we significantly relax the assumption from [GZ17a] by showing that MLE achieves *strong* recovery with $(1+\epsilon)n_{\text{info}}$ samples for (1) any sparsity level less than cp and (2) finite but large values of signal to noise ratio.

Exact asymptotic characterization of MMSE for linear sparsity For both weak and strong recovery, the central object of interest is the MMSE $\mathbb{E}\left[\|\beta^* - \mathbb{E}\left[\beta^* \mid X, Y\right]\|^2\right]$ and its asymptotic behavior. While the asymptotic behavior of the MMSE remains a challenging open problem when $k = o(p)$, it has been accurately understood when $k = \Theta(p)$ and $k/\sigma^2 = \Theta(1)$.

To be more specific, consider the asymptotic regime where $k = \varepsilon p$, $\sigma^2 = k/\gamma$, and $n = \delta p$, for fixed positive constants $\varepsilon, \gamma, \delta$ as $p \to +\infty$. The asymptotic minimum mean-square error (MMSE) can be characterized explicitly in terms of $(\varepsilon, \gamma, \delta)$.

This characterization was first obtained heuristically using the replica method from statistical physics [Tan02, GV05] and later proven rigorously [RP16, BDMK16]. More specifically, for fixed (ε, γ), let the asymptotic MMSE as a function of δ be defined by

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

The results in [RP16, BDMK16] lead to an explicit formula for $\mathcal{M}_{\varepsilon,\gamma}(\delta)$. Furthermore, they show that for $\varepsilon \in (0,1)$ and all sufficiently large $\gamma \in (0,\infty)$, $\mathcal{M}_{\varepsilon,\gamma}(\delta)$ has a jump discontinuity as a function of δ. The location of this discontinuity, denoted by $\delta^* = \delta^*(\varepsilon,\gamma)$, occurs at a value that is strictly greater than the threshold n_{info}/p.

Furthermore, at the the discontinuity, the MMSE transitions from a value that is strictly less than the MMSE without any observations to a value that is strictly positive, i.e., $\mathcal{M}_{\varepsilon,\gamma}(0) > \lim_{\delta\uparrow\delta^*} \mathcal{M}_{\varepsilon,\gamma}(\delta) > \lim_{\delta\downarrow\delta^*} \mathcal{M}_{\varepsilon,\gamma}(\delta) > 0$.

To compare these formulas to the sub-linear sparsity studied in this Chapter, one can consider the limiting behavior of $\mathcal{M}_{\varepsilon,\gamma}(\delta)$ as ε decreases to zero. It can be verified that $\mathcal{M}_{\varepsilon,\gamma}(\delta)$ converges indeed to a step zero-one function as $\varepsilon \to 0$ and the jump discontinuity transfers indeed to the critical value n_{info}/p which makes the behavior consistent with the results in this Chapter.

However, an important difference is that the results in this Chapter are derived directly under the scaling regime $k = o(p)$ whereas the derivation described above requires one to first take the asymptotic limit $p \to \infty$ for fixed (ϵ, γ) and then take $\epsilon \to 0$. Since the limits cannot interchange in any obvious way, the results in this Chapter cannot be derived as a consequence of the rigorous results in [RP16, BDMK16]. Finally, it should be mentioned that taking the limit $\epsilon \to 0$ for the replica prediction suggests the step behavior for all values of signal-to-noise ratio γ (see Figures 2-2, 2-3). In this Chapter, the step behavior is rigorously proven in the high signal-to-noise ratio regime. The proof of the step behavior when the signal-to-noise ratio is low remains an open problem.

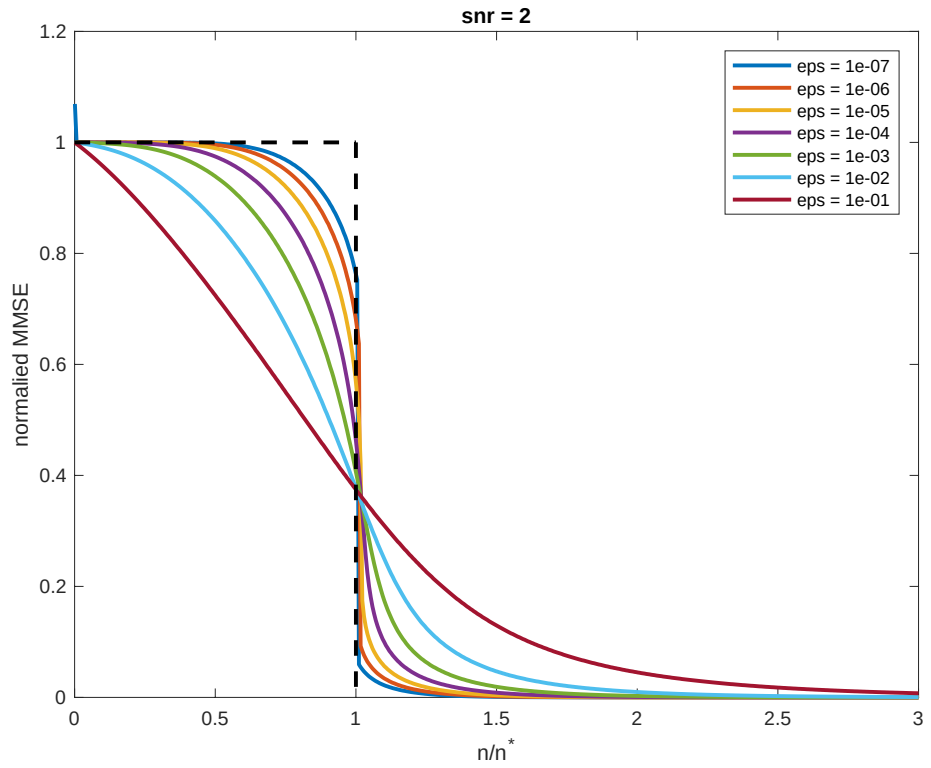


Figure 2-2: The limit of the replica-symmetric predicted MMSE $\mathcal{M}_{\varepsilon,\gamma}(\cdot)$ as $\epsilon \to 0$ for signal to noise ratio (snr) γ equal to 2. Here by n^* we denote the n_{info}.

Sparse Superposition Codes Constructing an algorithm for recovering a binary k-sparse β^* from $(Y = X\beta^* + W, X)$ receives a lot of attention from a coding theory point of view. The reason is that such recovery corresponds naturally to a code for the memoryless additive Gaussian white noise (AWGN) channel with signal-to-noise ratio equal to k/σ^2. Specifically in this context achieving strong recovery of a uniformly chosen binary k-sparse β^* with $(1+\epsilon)n_{\text{info}}$ samples, for arbitrary $\epsilon > 0$, corresponds exactly to capacity-achieving encoding-decoding mechanism of $\binom{p}{k} \sim (pe/k)^k$ messages through a AWGN channel. A recent line of work has analyzed a similar mechanism where $(p/k)^k$ messages are encoded through k-block-sparse vectors; that is the vector β^* is designed to have at most one non-zero value in each of k block of entries indexed by $i\lfloor p/k \rfloor, i\lfloor p/k \rfloor + 1, \cdots, (i+1)\lfloor p/k \rfloor - 1$ for $i = 0, 1, 2, \ldots, k-1$. It has shown that by using various polynomial-time decoding mechanisms, such as adaptive successive decoding [JB12], [JB14], a soft-decision iterative decoder [BC12], [Cho14] and finally Approximate Message Passing techniques [RGV17], one can strongly recover the hidden k-block-sparse vector with $(1+\epsilon)n_{\text{info}}$ samples and achieve capacity. Their techniques are tailored to work for any $k = p^{1-c}$

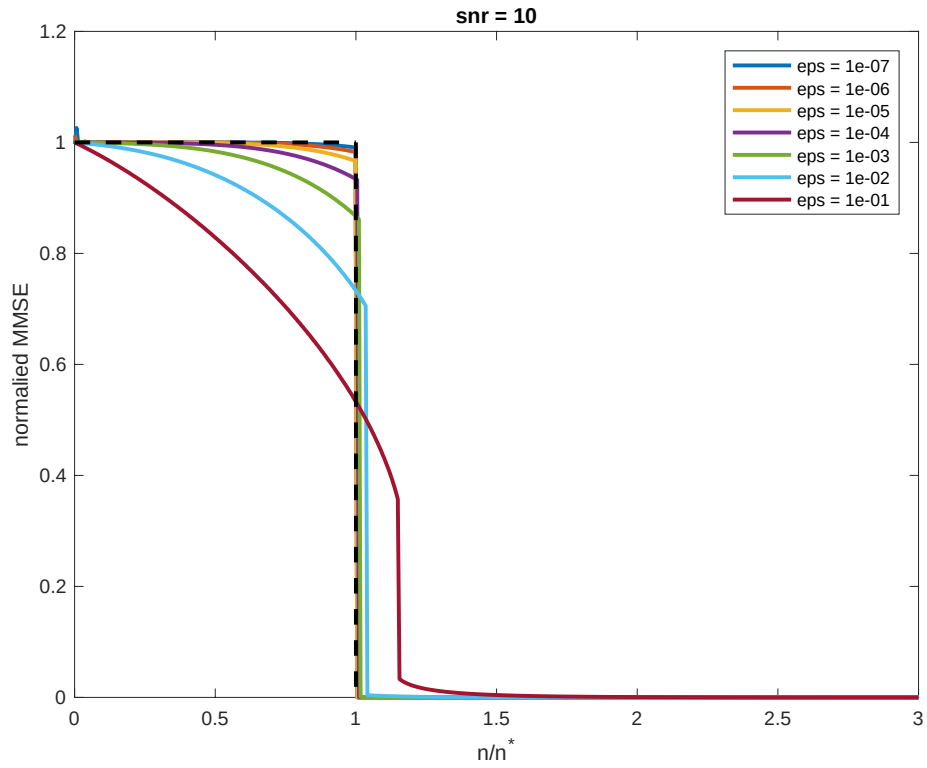


Figure 2-3: The limit of the replica-symmetric predicted MMSE $\mathcal{M}_{\varepsilon,\gamma}(\cdot)$ as $\epsilon \to 0$ for signal to noise ratio (snr) γ equal to 10. Here by n^* we denote the n_{info}.

with $c \in (0,1)$ and also require the vector to have carefully chosen non-zero entries, that is the hidden vector is not assumed to simply be binary. In this work Theorem 2.2.5 establishes that under the simple assumption on β^* being binary and arbitrarily (not block) k-sparse it suffices to make strong recovery possible with $(1+\epsilon)n_{\text{info}}$ samples when $k = o(p)$. Nevertheless, our decoding mechanism requires a search over the space of k-sparse binary vectors and therefore is not in principle polynomial-time. The design of a polynomial-time recovery algorithm for this task and $(1+\epsilon)n_{\text{info}}$ samples remains largely an open problem (see [GZ17a]).

Information-theoretic limits up to constant factors for exact recovery Although exact recovery is not our focus, we briefly mention some of the rich literature on the information-theoretic limits for the exact recovery of β^*, i.e., $\mathbb{P}\left\{\hat{\beta}^* = \beta^*\right\} \to 1$ as $p \to \infty$ (see, e.g. [**?**, FRG09, Rad11, WWR10, NT18] and the references therein). Clearly since exact recovery implies weak and strong recovery, the sample sizes required to be achieve exact recovery are in principle no smaller than n_{info}.

Specifically, it has been shown in [**?**, Theorem 1] that the maximum likelihood estimator achieves exact recovery if $n \geq \Omega\left(\log\binom{p-k}{k} + \sigma^2 \log(p-k)\right)$ and $n - k \to +\infty$. Conversely, $n > \max\{f_1(p,k), \ldots, f_k(p,k), k\}$ is shown in [WWR10, Theorem 1] to be necessary for exact recovery, where $f_m(p,k) = 2\frac{\log\binom{p-k+m}{m}-1}{\log\left(1+\frac{m(p-k)}{p-k+m}/\sigma^2\right)}$. In the special regime where k and σ are fixed constants, it has been shown in [JKR11, Theorem 1] that exact recovery is information-theoretically possible if and only if $n \geq (1+o(1))n_{\text{info}}$. Notice that this result achieves exact recovery for approximately n_{info} sample size, but in this case of constant k it can be easily seen that the two notions of exact and strong recovery coincide.

Computationally, it has been shown in [Wai09b, Section IV-B] that LASSO achieves exact recovery in polynomial-time if $n \geq 2k\log(p-k)$. More recently, it is shown in [NT18, Theorem 3.2, Corollary 3.2] that exact recovery can be achieved in polynomial-time, provided that $k = o(p)$, $\sigma \geq \sqrt{3}$, and $n \geq \Omega\left(k\log\frac{ep}{k} + \sigma^2\log p\right)$.

2.1.3 Proof Techniques

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

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

Three choices of f are of particular interests (See [PW15, Section 6] for details):

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

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

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

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

$$\mathrm{TV}(P,Q) \le \sqrt{2D(P||Q)} \le \sqrt{2\log\left(\chi^2(P\|Q)+1\right)}, \tag{2.5}$$

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

$$D(P\|Q) = \mathrm{Exp}_P\left[\log\frac{dP}{dQ}\right] \le \log\left(\mathrm{Exp}_P\left[\frac{dP}{dQ}\right]\right) = \log\left(\chi^2(P\|Q)+1\right). \tag{2.6}$$

Recall that to show the weak detection between P and Q_λ is impossible, it is equivalent to proving that $\mathrm{TV}\left(P,Q_\lambda\right) = o(1)$. In view of (2.5) there is a natural strategy towards proving it: it suffices to prove that $\chi^2(P,Q_\lambda) = o\left(1\right)$, which amounts to showing the second moment $\mathrm{Exp}_Q\left[(dP/dQ_\lambda)^2\right] = 1 + o\left(1\right)$. We prove that indeed if $n \le (1-o(1))\, n_{\mathrm{info}}/2$ and λ is appropriately chosen, then this second moment is indeed $1+o(1)$ (Theorem 2.2.1); however, if $n > n_{\mathrm{info}}/2$, then it blows up to infinity. This is because even if potentially $\mathrm{TV}(P,Q_\lambda) = o(1)$, rare events can cause the second moment to explode and in particular (2.5) is far from being tight.

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

Definition 2.1.2 (Conditioned planted model)**.** *Given a subset $\mathcal{E} \subset \mathbb{R}^{n\times p}\times\mathbb{R}^p$, define the conditioned planted model*

$$P_{\mathcal{E}}(X,Y) = \frac{\mathrm{Exp}_{\beta^*}\left[P(X,Y\mid\beta^*)\mathbf{1}_{\{\mathcal{E}\}}(X,\beta^*)\right]}{\mathbb{P}\left\{\mathcal{E}\right\}}. \tag{2.7}$$

Using this notation we can write

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

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

convexity of KL-divergence,

$$D(P||Q_\lambda) \le (1-\varepsilon)D(P_{\mathcal{E}}||Q_\lambda) + \varepsilon D(P_{\mathcal{E}^c}||Q_\lambda). \tag{2.8}$$

Under an appropriately chosen $\mathcal{E}$, and $\lambda > 0$, our main impossibility of detection result (Theorem 2.2.3) shows that if $n \le (1+o(1))n_{\text{info}}$, then $\text{Exp}_{Q_\lambda}[(dP_{\mathcal{E}}/dQ_\lambda)^2] = 1+o(1)$, or equivalently, $\chi^2(P_{\mathcal{E}}\|Q_\lambda) = o(1)$, which immediately implies that $D(P_{\mathcal{E}}\|Q_\lambda) = o(1)$ and $\text{TV}(P_{\mathcal{E}}, Q_\lambda) = o(1)$. Finally, we argue that ε converges to 0 sufficiently fast so that according to (2.8), $\text{TV}(P, Q_\lambda) \le \text{TV}(P_{\mathcal{E}}, Q) + o(1) = o(1)$ and $D(P\|Q_\lambda) \le D(P_{\mathcal{E}}\|Q_\lambda) + o(1) = o(1)$.

We remark that this (conditional) second moment method for providing detection lower bound has been used in many high-dimensional inference problems (see e.g. [MNS15, BMV$^+$18, BMNN16, PWB16, WX18] and references therein).

To further show weak recovery is impossible in the regime for sample size $n < n_{\text{info}}$ (Theorem 2.2.4), we establish a lower bound of MSE in terms of $D(P\|Q_\lambda)$ (Lemma 2.4.1) which implies that the minimum MSE needs to be $(1-o(1))\,k$ if $D(P\|Q_\lambda) = o(n)$. The key underpinning our lower bound proof is the area theorem [MMU08, KKM$^+$17].

2.1.4 Notation and Organization

Denote the identity matrix by $\mathbf{I}$. We let $\|X\|$ denote the spectral norm of a matrix X and $\|x\|$ denote the ℓ_2 norm of a vector x. For any positive integer n, let $[n] = \{1, \dots, n\}$. For any set $T \subset [n]$, let $|T|$ denote its cardinality and T^c denote its complement. We use standard big O notations, e.g., for any sequences $\{a_p\}$ and $\{b_p\}$, $a_p = \Theta(b_p)$ if there is an absolute constant $c > 0$ such that $1/c \le a_p/b_p \le c$; $a_p = \Omega(b_p)$ or $b_p = O(a_p)$ if there exists an absolute constant $c > 0$ such that $a_p/b_p \ge c$. We say a sequence of events $\mathcal{E}_p$ indexed by a positive integer p holds with high probability, if the probability of $\mathcal{E}_p$ converges to 1 as $p \to +\infty$. Without further specification, all the asymptotics are taken with respect to $p \to \infty$. All logarithms are natural and we use the convention $0 \log 0 = 0$. For two real numbers a and b, we use $a \vee b = \max\{a, b\}$ to denote the larger of a and b. For two vectors u, v of the same dimension, we use $\langle u, v\rangle$ denote their inner product. We use χ^2_n denote the standard chi-squared distribution with n degrees of freedom. For $n, m, k \in \mathbb{N}$ with $m \le k \le n$ and $m + k \le n$ we denote

by $\mathrm{Hyp}\,(n, m, k)$ the Hypergeometric distribution with parameters n, m, k and probability mass function $p(s) = \binom{m}{s}\binom{n-m}{k-s}/\binom{n}{k}, s \in [0, m] \cap \mathbb{Z}$.

The remainder of the Chapteris organized as follows. Section 2.2 presents the main results without proofs. Section 2.3 and Section 2.4 prove the negative results for detection and recovery, respectively. Section 2.5 proves the positive results for detection and recovery. We conclude the Chapter in Section 2.6, mentioning a few open problems. Auxiliary lemmata and miscellaneous details are left to rest Sections.

2.2 Main Results

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

2.2.1 Impossibility of Weak Detection with $n < n_{\mathrm{info}}$

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

Theorem 2.2.1. *Suppose $k \le p^{1/2-\delta}$ for a fixed constant $\delta > 0$ and $k/\sigma^2 \ge C$ for a sufficiently large constant C only depending on δ.*

If

$$n \le \frac{1}{2}\left(1 - \frac{\log\log(p/k)}{\log(p/k)}\right) n_{\mathrm{info}}, \tag{2.9}$$

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

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

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

The complete proof of the above Theorem can be found in Section 2.3.1. Nevertheless, let us provide here a short proof sketch. Using an explicit calculation, we first find that for any

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

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

where $S = \langle \beta^*, (\beta^*)' \rangle$ is the overlap between two independent copies $\beta^*, (\beta^*)'$ and follows a Hypergeometric distribution with parameters (p, k, k). Plugging in $\lambda = \lambda_0 = \sqrt{k/\sigma^2 + 1}$, we get that

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

Using this we show that if $n \leq (1 + o(1))\, n_{\mathrm{info}}/2$, then $\chi^2(P \| Q_{\lambda_0})$ is indeed $o\,(1)$, implying by (2.5) the impossibility result. However, if $n > n_{\mathrm{info}}/2$, then this χ^2-divergence can be proven to blow up to infinity, rendering the method based on (2.5) uninformative in this regime. To see this, by considering the event $S = k$ which happens with probability $1/\binom{p}{k}$, we get that

$$\chi^2(P \| Q_{\lambda_0}) \geq \frac{1}{\binom{p}{k}} \left[\left(1 - \frac{k}{k+\sigma^2}\right)^{-n} \right] - 1 = \exp\left(n \log\left(1 + \frac{k}{\sigma^2}\right) - \log \binom{p}{k} \right) - 1. \tag{2.10}$$

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

To be able to obtain tighter results and go all the way to n_{info} sample size, we resort to a *conditional* second moment method as explained in the proof techniques. Specifically we show that weak detection is impossible for any $n \leq (1-\alpha) n_{\mathrm{info}}$, for some $\alpha > 0$ that can be made to be arbitrarily small by increasing k/σ^2 and p/k. In particular, this improves on the direct calculation of the χ^2 distance by a multiplicative factor of 2 and shows that n_{info} is a sharp information theoretic threshold for weak detection between the planted model P and the null model Q_{λ_0}.

Before formally stating our main theorem, we specify the conditioning event $\mathcal{E}_{\gamma,\tau}$ which will be shown to hold with high probability in 2.8.1 under appropriate choices of γ and τ.

Definition 2.2.2 (Conditioning event)**.** *Given $\gamma \geq 0$ and $\tau \in [0, k]$, define an event $\mathcal{E}_{\gamma,\tau} \subset$*

$\mathbb{R}^{n\times p}\times\mathbb{R}^p$ *as*

$$\mathcal{E}_{\gamma,\tau}=$$
$$\left\{(X,\beta^*)\,:\,\frac{\|X(\beta^*+(\beta^*)')\|^2}{\mathbb{E}\left[\|X(\beta^*+(\beta^*)')\|^2\right]}\le 2+\gamma,\ \forall(\beta^*)'\in\{0,1\}^p\ \textit{with}\ \|(\beta^*)'\|_0=k\ \textit{and}\ \langle(\beta^*)',\beta^*\rangle\ge\tau\right\}. \tag{2.11}$$

To understand the value of γ,τ in the definition of this event, notice that for each $\beta^*,(\beta^*)'$, from the definition of X, we have $X(\beta^*+(\beta^*)')\sim\mathcal{N}\left(0,2(k+s)\mathbf{I}_n\right)$ and therefore,

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

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

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

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

Next, we explain the intuition behind our choice of conditioning event $\mathcal{E}_{\gamma,\tau}$. Recall that in view of (2.10), $\chi^2(P\|Q_{\lambda_0})$ blows up to infinity when the overlap $\langle\beta^*,(\beta^*)'\rangle$ is equal to k. In fact, when the overlap $\langle\beta^*,(\beta^*)'\rangle=k$, $\|X(\beta^*+(\beta^*)')\|^2$ can be enormously large, causing $\chi^2(P\|Q_{\lambda_0})$ to explode. We rule out this catastrophic event by conditioning on $\mathcal{E}_{\gamma,\tau}$ which upper bounds $\|X(\beta^*+(\beta^*)')\|^2$ when the overlap $\langle\beta^*,(\beta^*)'\rangle$ is large (See (2.35) for the key step of upper bounding $\|X(\beta^*+(\beta^*)')\|^2$).

As a result, we are able to prove that the χ^2-divergence between the conditional planted model $P_{\mathcal{E}_{\gamma,\tau}}$ and the null model Q_{λ_0} for $\lambda_0=\sqrt{k/\sigma^2+1}$ is $o(1)$, which implies the following general impossibility of detection result.

Theorem 2.2.3. *Suppose $k\le p^{\frac{1}{2}-\delta}$ for an arbitrarily small fixed constant $\delta\in(0,\frac{1}{2})$ and $k/\sigma^2\ge C$ for a sufficiently large constant C only depending on δ. Assume $n\le(1-\alpha)\,n_{\text{info}}$ for $\alpha\in$*

$(0, 1/2]$ *such that*

$$\alpha = \frac{8}{\log(1 + k/\sigma^2)} \vee \frac{32 \log\log(p/k)}{\log(p/k)}. \tag{2.12}$$

Set

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

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

$$\chi^2\left(P_{\mathcal{E}_{\gamma,\tau}} \| Q_{\lambda_0}\right) = o(1). \tag{2.13}$$

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

The proof of the Theorem can be found in Section 2.3.2.

2.2.2 Impossibility of Weak Recovery with $n < n_{\text{info}}$

In this section we present our impossibility of recovery result. We do this using the impossibility of detection result established above. Specifically we first strengthen Theorem 2.2.3 and show that under the assumptions of Theorem 2.2.3, $D(P\|Q_{\lambda_0}) = o_p(1)$. Notice that this is not needed to conclude impossibility of detection, that is $TV(P, Q_{\lambda_0}) = o(1)$, but is needed here for establishing the impossibility of recovery result. As a second step, inspired by the celebrated area theorem, we establish (Lemma 2.4.1) a lower bound to the minimum MSE in terms of $D(P\|Q_{\lambda_0})$, which is potentially of independent interest. The lemma essentially quantifies the natural idea that if the data (Y, X) drawn from planted model are statistically close to the data (Y, X) drawn from null model then there are limitations on the performance of recovering the hidden vector β^* based on the data (Y, X) from the planted model. Interestingly the lemma itself does not require the hidden vector β^* to be binary or k-sparse but only to satisfy $\mathbb{E}\left[\|\beta^*\|_2^2\right] = k$. Combining the two steps allows us to conclude that the minimum MSE is $k(1 + o_p(1))$; hence the impossibility of weak recovery.

Theorem 2.2.4. *Suppose* $k \le p^{\frac{1}{2}-\delta}$ *for an arbitrarily small fixed constant* $\delta \in (0, \frac{1}{2})$ *and* $k/\sigma^2 \ge C$ *for a sufficiently large constant* C *only depending on* δ. *Let* $\lambda_0 = \sqrt{k/\sigma^2 + 1}$. *If*

$n \leq (1-\alpha)\, n_{\text{info}}$ *for* $\alpha \in (0, 1/2]$ *given in* (2.12)*, then it holds that*

$$D\left(P \| Q_{\lambda_0}\right) = o_p(1). \tag{2.14}$$

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

$$MSE\left(\hat{\beta}^*\right) = k\left(1 + o_p(1)\right). \tag{2.15}$$

The proof of the above Theorem can be found in Section 2.4.

2.2.3 Positive Result for Strong Recovery with $n > n_{\text{info}}$

This subsection and the next one are in the regime where $n > n_{\text{info}}$. In these regimes, in contrast to $n < n_{\text{info}}$ we establish that both strong recovery and strong detection are possible.

Towards recovering the vector β^*, we consider the Maximum Likelihood Estimator (MLE) of β^*:

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

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

Specifically, we establish the following result.

Theorem 2.2.5. *Suppose* $\log\log(p/k) \geq 1$*. If*

$$n \geq \left(1 + \frac{\log 2}{\log\left(1 + k/(2\sigma^2)\right)}\right)\left(1 + \frac{4\log\log(p/k)}{\log(p/k)}\right) n_{\text{info}}, \tag{2.16}$$

then

$$\mathbb{P}\left\{\|\hat{\beta}^* - \beta^*\|^2 \geq \frac{2k}{\log(p/k)}\right\} \leq \frac{e^2}{\log^2(p/k)(1 - e^{-1})}. \tag{2.17}$$

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

$$\frac{1}{k}\mathbb{E}\left[\left\|\hat{\beta}^* - \beta^*\right\|_2^2\right] = o_p(1), \tag{2.18}$$

i.e., MLE achieves strong recovery of β^*.

The proof of the above Theorem can be found in Section 2.5.1.

2.2.4 Positive Result for Strong Detection with $n > n_{\text{info}}$

In this subsection we establish that when $n > n_{\text{info}}$ strong detection is possible. To distinguish the planted model P and the null model Q_λ, we consider the test statistic:

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

Theorem 2.2.6. *Suppose*

$$\log n - \frac{2}{n} \log \binom{p}{k} \to +\infty \tag{2.19}$$

and

$$n \geq \frac{2 \log \binom{p}{k}}{\log(1 + k/\sigma^2) + \log(1 - \alpha)} \tag{2.20}$$

for an arbitrarily small but fixed constant $\alpha \in (0, 1)$. Then by letting $\tau = \frac{1}{(1-\alpha/2)(1+k/\sigma^2)}$, we have that

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

which achieves the strong detection between the planted model P and the null model Q_λ.

The proof of Theorem 2.2.6 can be found in Section 2.5.2.

We close this section with one remark, explaining the newly introduced condition (2.19).

Remark 2.2.7. *Recall that $n_{\text{info}} = 2k \log(p/k)/\log(1 + k/\sigma^2)$ and $\binom{p}{k} \leq (ep/k)^k$. Thus,*

$$\begin{aligned}
\log n_{\text{info}} - \frac{2}{n_{\text{info}}} \log \binom{p}{k} &\geq \log\left(\frac{2k \log(p/k)}{\log(1 + k/\sigma^2)}\right) - \frac{\log(ep/k)}{\log(p/k)} \log\left(1 + k/\sigma^2\right) \\
&\geq \log\left(k \log \frac{p}{k}\right) - \log\log\left(1 + k/\sigma^2\right) - \log\left(1 + k/\sigma^2\right) - \frac{\log\left(1 + k/\sigma^2\right)}{\log(p/k)}.
\end{aligned}$$

If $1 + k/\sigma^2 \leq \left(k \log \frac{p}{k}\right)^{1-\eta}$ for some fixed constant $\eta > 0$, then it follows from the last displayed

equation that

$$\log n_{\text{info}} - \frac{2}{n_{\text{info}}} \log \binom{p}{k} \geq \eta \log \left(k \log \frac{p}{k} \right) - \log \log \left(k \log \frac{p}{k} \right) - \frac{\log \left(k \log \frac{p}{k} \right)}{\log(p/k)}$$

which goes to $+\infty$ *as* $p \to +\infty$*; hence* n_{info} *satisfies (2.19).*

Therefore, assuming that $1 + k/\sigma^2 \leq \left(k \log \frac{p}{k} \right)^{1-\eta}$ *and* $n \geq (1+\epsilon) n_{\text{info}}$ *for some arbitrarily small constants* $\eta, \epsilon > 0$*, there exists a constant* $C = C(\epsilon) > 0$ *such that if* $k/\sigma^2 \geq C(\epsilon)$*, then the test statistic* $\mathcal{T}(X, Y)$ *achieves strong detection.*

2.3 Proof of Negative Results for Detection

2.3.1 Proof of Theorem 2.2.1

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

Proposition 2.3.1. *For any* $\lambda > \sqrt{k/\sigma^2 + 1/2}$*,*

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

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

$$\frac{P(X,Y)}{Q_\lambda(X,Y)} = \frac{P(Y|X)}{Q_\lambda(Y)} = \frac{\operatorname{Exp}_{\beta^*}[P(Y|X,\beta^*)]}{Q_\lambda(Y)}.$$

Therefore

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

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

$$\operatorname{Exp}_{Q_\lambda} \left[\left(\frac{P}{Q_\lambda} \right)^2 \right] = \operatorname{Exp}_{\beta^* \perp\!\!\!\perp (\beta^*)'} \operatorname{Exp}_X \operatorname{Exp}_Y \left[\frac{P(Y|X,\beta^*) P(Y|X,(\beta^*)')}{Q_\lambda^2(Y)} \right], \tag{2.21}$$

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

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

$$\begin{aligned}\frac{P(Y|X,\beta^*)}{Q_\lambda(Y)} &= \lambda^n \exp\left(-\frac{1}{2\sigma^2}\|Y - X\beta^*\|_2^2 + \frac{1}{2\lambda^2\sigma^2}\|Y\|_2^2\right) \\ &= \lambda^n \exp\left(-\frac{\lambda^2-1}{2\sigma^2\lambda^2}\|Y\|_2^2 + \frac{1}{\sigma^2}\langle Y, X\beta^*\rangle - \frac{1}{2\sigma^2}\|X\beta^*\|_2^2\right).\end{aligned}$$

Hence,

$$\begin{aligned}&\frac{P(Y|X,\beta^*)P(Y|X,(\beta^*)')}{Q_\lambda^2(Y)} \\ &= \lambda^{2n} \exp\left(-\frac{\lambda^2-1}{\sigma^2\lambda^2}\|Y\|_2^2 + \frac{1}{\sigma^2}\langle Y, X(\beta^* + (\beta^*)')\rangle - \frac{1}{2\sigma^2}\left(\|X\beta^*\|_2^2 + \|X(\beta^*)'\|_2^2\right)\right)\end{aligned}$$

which equals

$$\lambda^{2n} e^{-\frac{\lambda^2-1}{\sigma^2\lambda^2}\left\|Y - \frac{\lambda^2 X(\beta^*+(\beta^*)')}{2(\lambda^2-1)}\right\|_2^2 + \frac{\lambda^2\|X(\beta^*+(\beta^*)')\|_2^2}{4(\lambda^2-1)\sigma^2} - \frac{1}{2\sigma^2}\left(\|X\beta^*\|_2^2 + \|X(\beta^*)'\|_2^2\right)}.$$

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

$$\begin{aligned}&\mathrm{Exp}_Y\left[\exp\left(-\frac{\lambda^2-1}{\sigma^2\lambda^2}\left\|Y - \frac{\lambda^2 X(\beta^* + (\beta^*)')}{2(\lambda^2-1)}\right\|_2^2\right)\right] \\ &= \frac{1}{(2\lambda^2-1)^{n/2}} \exp\left(-\frac{\lambda^2\|X(\beta^* + (\beta^*)')\|_2^2}{4(2\lambda^2-1)(\lambda^2-1)\sigma^2}\right).\end{aligned}$$

Combining the last two displayed equations yields that

$$\begin{aligned}&\mathrm{Exp}_Y\left[\frac{P(Y|X,\beta^*)P(Y|X,(\beta^*)')}{Q_\lambda^2(Y)}\right] \\ &= \frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}} \exp\left\{\frac{1}{2\sigma^2(2\lambda^2-1)}\left((1-\lambda^2)\left(\|X\beta^*\|_2^2 + \|X(\beta^*)'\|_2^2\right) + 2\lambda^2\langle X\beta^*, X(\beta^*)'\rangle\right)\right\}.\end{aligned} \tag{2.22}$$

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

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

Then conditional on β^* and $(\beta^*)'$, Z_0, Z_1, Z_2 are mutually independent and

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

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

$$X\beta^* = Z_0 + Z_1 \text{ and } X(\beta^*)' = Z_0 + Z_2. \tag{2.23}$$

Let $Z = [Z_0, Z_1, Z_2]^t \in \mathbb{R}^{3n}$. Using (2.22) and (2.23) and elementary algebra we have

$$\text{Exp}_Y \left[\frac{P(Y|X, \beta^*) P(Y|X, (\beta^*)')}{Q_\lambda^2(Y)} \right] = \frac{\lambda^{2n}}{(2\lambda^2 - 1)^{n/2}} \exp\left\{ t Z^\top A Z \right\}, \tag{2.24}$$

where

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

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

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

Note that

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

It is straightforward to find that the eigenvalues of AV are 0 of multiplicity n, $k+s$ of multiplicity

n, and $(k-s)(1-2\lambda^2)$ of multiplicity n. Thus,

$$\det(\mathbf{I}_{3n} - 2tAV) = (1-2t(k+s))^n \left(1-2t(k-s)(1-2\lambda^2)\right)^n. \tag{2.25}$$

It follows from (2.24) that

$$\begin{aligned}\operatorname{Exp}_X \operatorname{Exp}_Y \left[\frac{P(Y|X,\beta^*)P(Y|X,(\beta^*)')}{Q_\lambda^2(Y)}\right] &= \frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}} \operatorname{Exp}_Z \left[e^{tZ^\top AZ}\right] \\ &= \frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}} \frac{1}{\sqrt{\det(\mathbf{I}_{3n}-2tAV)}},\end{aligned} \tag{2.26}$$

where the last equality holds if $t < \frac{1}{2(k+s)}$ and follows from the expression of MGF of a quadratic form of normal random variables, see, e.g., [Bal67, Lemma 2].

Combining (2.25) and (2.26) yields that if $t = \frac{1}{2\sigma^2(2\lambda^2-1)} < \frac{1}{2(k+s)}$,

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

Note that if $2\lambda^2-1 > \frac{2k}{\sigma^2}$, then $\frac{1}{2\sigma^2(2\lambda^2-1)} < \frac{1}{2(k+s)}$ for all $0 \le s \le k$. It follows from (2.21) that if $2\lambda^2-1 > \frac{2k}{\sigma^2}$, then

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

□

We establish also the following lemma.

Lemma 2.3.2. *Suppose $k \le p^{\frac{1}{2}-\delta}$ for an arbitrarily small fixed constant $\delta \in (0, \frac{1}{2})$ and $\frac{k}{\sigma^2} \ge C$ for a sufficiently large constant C only depending on δ. If n satisfies condition (2.9), then*

$$\operatorname{Exp}_{S\sim\mathrm{Hyp}(k,k,p)}\left[\left(1-\frac{S}{k+\sigma^2}\right)^{-n}\right] = 1 + o_p(1). \tag{2.27}$$

Proof. The lemma readily follows by combining 2.7.2 and 2.7.5 with $\alpha = \frac{\log\log(p/k)}{\log(p/k)}$ and $c = p^{-1/2-\delta}$. □

Proof of Theorem 2.2.1. Using Proposition 2.3.1 for $\lambda = \lambda_0$ satisfying $\lambda_0^2 = k/\sigma^2 + 1$ we have

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

Using now 2.3.2 we have $\chi^2(P\|Q_{\lambda_0}) = o(1)$. The chain of inequalities (2.5) concludes the proof of Theorem 2.2.1.

□

2.3.2 Proof of Theorem 2.2.3

Proof of Theorem 2.2.3. For notational simplicity we denote in this proof the probability measure Q_{λ_0} simply by Q and the event $\mathcal{E}_{\gamma,\tau}$ by $\mathcal{E}$.

We first show that (2.13) implies $D(P_{\mathcal{E}}\|Q) = o(1)$, $\mathrm{TV}(P_{\mathcal{E}}, Q) = o(1)$, and $\mathrm{TV}(P, Q) = o(1)$.

It follows from (2.5) that $D(P_{\mathcal{E}}\|Q) = o(1)$ and $\mathrm{TV}(P_{\mathcal{E}}, Q) = o(1)$. Observe that under our choice of τ and γ, 2.8.1 implies that

$$\mathbb{P}\{\mathcal{E}^c\} \le \exp\left(-\frac{n\gamma}{8}\right) = \exp\left(-\frac{\alpha k\log(p/k)}{8}\right) \le \exp\left(-4k\log\log(p/k)\right) = o_p(1). \tag{2.28}$$

Thus, in view of (2.8), we get that

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

Next we prove (2.13). We first carry calculations for any $\lambda > \sqrt{k/\sigma^2 + 1/2}$; we then restrict to $\lambda = \sqrt{k/\sigma^2 + 1}$. In view of (2.7), we have

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

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

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

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

$$\mathrm{Exp}_Q\left[\left(\frac{P_{\mathcal{E}}}{Q}\right)^2\right]$$

equals

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

It follows from (2.22) that

$$\begin{aligned}&\mathrm{Exp}_Y\left[\frac{P(Y|X,\beta^*)P(Y|X,(\beta^*)')}{Q^2(Y)}\right]\\ &= \frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}}\exp\left\{\frac{\|X(\beta^*+(\beta^*)')\|^2 - (2\lambda^2-1)\,\|X(\beta^*-(\beta^*)')\|^2}{4\sigma^2(2\lambda^2-1)}\right\}.\end{aligned}$$

Combining the last two displayed equation yields that

$$\begin{aligned}&\mathrm{Exp}_Q\left[\left(\frac{P_{\mathcal{E}}}{Q}\right)^2\right]\\ &= \frac{(1+o(1))\,\lambda^{2n}}{(2\lambda^2-1)^{n/2}}\,\mathrm{Exp}_{\beta^* \perp\!\!\!\perp (\beta^*)'}\,\mathrm{Exp}_X\left[e^{\frac{\|X(\beta^*+(\beta^*)')\|^2 - \left(2\lambda^2-1\right)\|X(\beta^*-(\beta^*)')\|^2}{4\sigma^2(2\lambda^2-1)}}\mathbf{1}_{\{\mathcal{E}\}}(X,\beta^*)\mathbf{1}_{\{\mathcal{E}\}}(X,(\beta^*)')\right].\end{aligned} \tag{2.29}$$

Next we break the right hand side of (2.29) into two disjoint parts depending on whether $\langle\beta^*,(\beta^*)'\rangle \le \tau$. We prove that the part where $\langle\beta^*,(\beta^*)'\rangle \le \tau$ is $1+o(1)$ and the part where $\langle\beta^*,(\beta^*)'\rangle > \tau$ is $o(1)$. Combining them we conclude the desired result.

Part 1: Note that

$$\begin{aligned}
&\mathrm{Exp}_X\left[e^{\frac{\|X(\beta^*+(\beta^*)')\|^2-\left(2\lambda^2-1\right)\|X(\beta^*-(\beta^*)')\|^2}{4\sigma^2(2\lambda^2-1)}}\mathbf{1}_{\{\mathcal{E}\}}(X,\beta^*)\mathbf{1}_{\{\mathcal{E}\}}(X,(\beta^*)')\right]\mathbf{1}_{\{\langle\beta^*,(\beta^*)'\rangle\le\tau\}}\\
&\le \mathrm{Exp}_X\left[e^{\frac{\|X(\beta^*+(\beta^*)')\|^2-\left(2\lambda^2-1\right)\|X(\beta^*-(\beta^*)')\|^2}{4\sigma^2(2\lambda^2-1)}}\right]\mathbf{1}_{\{\langle\beta^*,(\beta^*)'\rangle\le\tau\}}. \qquad (2.30)
\end{aligned}$$

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

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

and therefore $X(\beta^*+(\beta^*)')\sim\mathcal{N}(0,2(k+s)\mathbf{I}_n)$ is independent of $X(\beta^*-(\beta^*)')\sim\mathcal{N}(0,2(k-s)\mathbf{I}_n)$, for $s=\langle\beta^*,(\beta^*)'\rangle$. Therefore,

$$\begin{aligned}
&\mathrm{Exp}_X\left[\exp\left\{\frac{\|X(\beta^*+(\beta^*)')\|^2-(2\lambda^2-1)\,\|X(\beta^*-(\beta^*)')\|^2}{4\sigma^2(2\lambda^2-1)}\right\}\right]\\
&=\mathrm{Exp}_X\left[\exp\left\{\frac{\|X(\beta^*+(\beta^*)')\|^2}{4\sigma^2(2\lambda^2-1)}\right\}\right]\mathrm{Exp}_X\left[\exp\left\{-\frac{\|X(\beta^*-(\beta^*)')\|^2}{4\sigma^2}\right\}\right]\\
&=\left(1-\frac{(k+s)}{\sigma^2(2\lambda^2-1)}\right)^{-n/2}\left(1+\frac{(k-s)}{\sigma^2}\right)^{-n/2}, \qquad (2.31)
\end{aligned}$$

where the last equality holds if $\lambda>\sqrt{(k+s)/(2\sigma^2)+1/2}$ and follows from the fact that $\mathbb{E}_{Z\sim\chi^2(1)}\left[e^{-tZ}\right]=\frac{1}{\sqrt{1+2t}}$ for $t>-1/2$. Combining (2.30) and (2.31) yields that if $\lambda>\sqrt{k/\sigma^2+1/2}$, then

$$\begin{aligned}
&\frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}} \qquad (2.32)\\
&\times\mathrm{Exp}_{\beta^*\perp(\beta^*)'}\,\mathrm{Exp}_X\left[e^{\frac{\|X(\beta^*+(\beta^*)')\|^2-\left(2\lambda^2-1\right)\|X(\beta^*-(\beta^*)')\|^2}{4\sigma^2(2\lambda^2-1)}}\mathbf{1}_{\{\mathcal{E}\}}(X,\beta^*)\mathbf{1}_{\{\mathcal{E}\}}(X,(\beta^*)')\right]\mathbf{1}_{\{\langle\beta^*,(\beta^*)'\rangle\le\tau\}}\\
&\le\frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}}\,\mathrm{Exp}_{\beta^*\perp(\beta^*)'}\left[\left(1-\frac{(k+s)}{\sigma^2(2\lambda^2-1)}\right)^{-n/2}\left(1+\frac{(k-s)}{\sigma^2}\right)^{-n/2}\mathbf{1}_{\{s\le\tau\}}\right],
\end{aligned}$$

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

$$\begin{aligned}
&\frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}} \qquad (2.33)\\
&\times \mathrm{Exp}_{\beta^* \perp\!\!\!\perp (\beta^*)'} \mathrm{Exp}_X \left[e^{\frac{\|X(\beta^*+(\beta^*)')\|^2 - \left(2\lambda^2-1\right)\|X(\beta^*-(\beta^*)')\|^2}{4\sigma^2(2\lambda^2-1)}} \mathbf{1}_{\{\mathcal{E}\}}(X,\beta^*)\mathbf{1}_{\{\mathcal{E}\}}(X,(\beta^*)') \right] \mathbf{1}_{\{\langle \beta^*,(\beta^*)'\rangle \le \tau\}}\\
&\overset{(a)}{\le} \left(\frac{k}{\sigma^2}+1\right)^n \mathrm{Exp}_{S\sim \mathrm{Hyp}(p,k,k)} \left\{ \left(1+\frac{(k-S)}{\sigma^2}\right)^{-n} \mathbf{1}_{\{S\le\tau\}} \right\}\\
&= \mathrm{Exp}_{S\sim \mathrm{Hyp}(p,k,k)} \left\{ \left(1-\frac{S}{k+\sigma^2}\right)^{-n} \mathbf{1}_{\{S\le\tau\}} \right\}, \qquad (2.34)
\end{aligned}$$

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

Using Lemma 2.7.2 we conclude that under our assumptions, there exists a constant $C > 0$ depending only on $\delta > 0$ such that if $k/\sigma^2 \ge C$ then

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

concluding the Part 1.

Part 2: By the definiton of $\mathcal{E}$, since $\tau \le s = \langle \beta^*, (\beta^*)' \rangle \le k$,

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

Therefore,

$$\begin{aligned}
&\mathrm{Exp}_X \left[\exp\left\{ \frac{\|X(\beta^*+(\beta^*)')\|^2 - (2\lambda^2-1)\|X(\beta^*-(\beta^*)')\|_2^2}{4\sigma^2(2\lambda^2-1)} \right\} \mathbf{1}_{\{\mathcal{E}\}}(X,\beta^*)\mathbf{1}_{\{\mathcal{E}\}}(X,(\beta^*)') \right]\\
&\times \mathbf{1}_{\{\langle \beta^*,(\beta^*)'\rangle > \tau\}}\\
&\le \mathrm{Exp}_X \left[\exp\left\{ \frac{4nk(2+\gamma) - (2\lambda^2-1)\|X(\beta^*-(\beta^*)')\|_2^2}{4\sigma^2(2\lambda^2-1)} \right\} \right] \mathbf{1}_{\{\langle \beta^*,(\beta^*)'\rangle > \tau\}}\\
&= \exp\left\{ \frac{nk(2+\gamma)}{\sigma^2(2\lambda^2-1)} \right\} \left(1+\frac{(k-s)}{\sigma^2}\right)^{-n/2} \mathbf{1}_{\{\langle \beta^*,(\beta^*)'\rangle > \tau\}}, \qquad (2.35)
\end{aligned}$$

where the first inequality follows from the definition of event $\mathcal{E}$ and the last equality holds due

to (2.31). It follows that

$$
\begin{aligned}
&\frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}} \\
&\times \mathrm{Exp}_{\beta^* \perp (\beta^*)'} \left[\mathrm{Exp}_X \left[e^{\frac{\|X(\beta^*+(\beta^*)')\|^2 - \left(2\lambda^2-1\right)\|X(\beta^*-(\beta^*)')\|_2^2}{4\sigma^2(2\lambda^2-1)}} \mathbf{1}_{\{\mathcal{E}\}}(X,\beta^*)\mathbf{1}_{\{\mathcal{E}\}}(X,(\beta^*)') \right] \mathbf{1}_{\{\langle \beta^*,(\beta^*)'\rangle > \tau\}} \right] \\
&\leq \frac{\lambda^{2n}}{(2\lambda^2-1)^{n/2}} \exp\left\{ \frac{nk(2+\gamma)}{\sigma^2(2\lambda^2-1)} \right\} \mathrm{Exp}_{S\sim \mathrm{Hyp}(p,k,k)} \left[\left(1+\frac{(k-S)}{\sigma^2}\right)^{-n/2} \mathbf{1}_{\{S>\tau\}} \right] \\
&\overset{(a)}{\leq} \lambda^n e^{n(1+\gamma/2)} \mathrm{Exp}_{S\sim \mathrm{Hyp}(p,k,k)} \left[\left(1+\frac{(k-S)}{\sigma^2}\right)^{-n/2} \mathbf{1}_{\{S>\tau\}} \right] \\
&\overset{(b)}{=} e^{n(1+\gamma/2)} \mathrm{Exp}_{S\sim \mathrm{Hyp}(p,k,k)} \left[\left(1-\frac{S}{k+\sigma^2}\right)^{-n/2} \mathbf{1}_{\{S>\tau\}} \right], \qquad (2.36)
\end{aligned}
$$

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

Recall that $n \leq (1-\alpha)n_{\mathrm{info}}$. Then under our choice of α and τ, applying 2.7.5 with n being replaced by $n/2$, $c = p^{-1/2-\delta}$, we get that there exits a universal constant $C > 0$ such that if $k/\sigma^2 \geq C$ then

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

where (a) follows because under our choice of γ and α,

$$
n\left(1+\frac{\gamma}{2}\right) \leq n + \frac{1}{2}\alpha k \log\frac{p}{k} \leq n_{\mathrm{info}} + \frac{1}{2}\alpha k \log\frac{p}{k} \leq \frac{3}{4}\alpha k \log\frac{p}{k};
$$

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

Combing the bounds for Parts 1 and 2, we conclude

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

as desired.

□

2.4 Proof of Negative Results for Recovery

2.4.1 Lower Bound on MSE

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

Lemma 2.4.1. *Given an integer $n \geq 2$ and an integer $m \in \{1, \ldots, n-1\}$, let $\hat{\beta}^*$ be an estimator that is a function of X and the first m observations $(Y_1, \ldots, Y_m)$. Then,*

$$MSE\left(\hat{\beta}^*\right) \geq e^{-\frac{2}{n-m}D(P\|Q)}(\sigma^2 + k) - \sigma^2. \tag{2.37}$$

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

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

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

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

Furthermore, by definition,

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

Combining the last three displayed equations gives that

$$\begin{aligned} I(\beta^*; Y \mid X) &= \frac{n}{2}\log(\lambda^2) + \frac{1}{2}\mathbb{E}\left[\frac{\|Y\|_2^2}{\lambda^2\sigma^2} - \frac{\|Y - X\beta^*\|_2^2}{\sigma^2}\right] - D(P\|Q) \\ &= \frac{n}{2}\left[\log\left(\frac{\lambda^2}{1+k/\sigma^2}\right) + \frac{1+k/\sigma^2}{\lambda^2} - 1\right] + \frac{n}{2}\log(1+k/\sigma^2) - D(P\|Q) \\ &\geq \frac{n}{2}\log(1+k/\sigma^2) - D(P\|Q). \end{aligned} \tag{2.38}$$

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

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

$$I(\beta^*; Y \mid X) = I(\beta^*; Y_1^m \mid X) + I(\beta^*; Y_{m+1}^n \mid X, Y_1^m), \tag{2.39}$$

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

$$\begin{aligned} I(\beta^*; Y_1^m \mid X) &\leq \sum_{i=1}^m I(\beta^*; Y_i \mid X) \\ &\leq \frac{m}{2}\mathbb{E}\left[\log\left(\mathbb{E}\left[\|Y_1\|^2 \mid X\right]/\sigma^2\right)\right] \\ &\leq \frac{m}{2}\log\left(\mathbb{E}\left[\|Y_1\|^2\right]/\sigma^2\right) \\ &\leq \frac{m}{2}\log\left(1 + k/\sigma^2\right), \end{aligned} \tag{2.40}$$

and

$$\begin{aligned} I(\beta^*; Y_{m+1}^n \mid X, Y_1^m) &\leq \sum_{i=m+1}^n I(\beta^*; Y_i \mid X, Y_1^m) \\ &\leq \frac{n-m}{2}\log\left(\mathbb{E}\left[\|Y_{m+1} - \mathbb{E}\left[Y_{m+1} \mid X, Y_1^m\right]\|^2\right]/\sigma^2\right) \\ &\leq \frac{n-m}{2}\log\left(1 + \mathrm{MSE}(\hat{\beta}^*)/\sigma^2\right), \end{aligned} \tag{2.41}$$

where the last inequality holds due to

$$\mathbb{E}\left[\|Y_{m+1} - \mathbb{E}\left[Y_{m+1} \mid X, Y_1^n\right]\|^2\right] = \mathbb{E}\left[\|\beta^* - \mathbb{E}\left[\beta^* \mid Y_1^m, X\right]\|^2\right] + \sigma^2 \leq \mathrm{MSE}(\hat{\beta}^*) + \sigma^2.$$

Plugging inequalities (2.40) and (2.41) back into (2.39) leads to

$$I(\beta^*; Y \mid X) \leq \frac{m}{2}\log(1 + k/\sigma^2) + \frac{n-m}{2}\log(1 + \mathrm{MSE}(\hat{\beta}^*)/\sigma^2). \tag{2.42}$$

Comparing (2.42) with (2.38) and rearranging terms gives the stated result. □

2.4.2 Upper Bound on Relative Entropy via Conditioning

We now show how a conditioning argument can be used to upper bound the relative entropy. Recall that (2.8) implies

$$D(P||Q) \leq (1-\varepsilon)D(P_{\mathcal{E}}||Q) + \varepsilon D(P_{\mathcal{E}^c}||Q). \tag{2.43}$$

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

Lemma 2.4.2. *For any $\mathcal{E} \subset \mathbb{R}^p \times \mathbb{R}^{n\times p}$ we have*

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

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

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

Proof. Starting with the definition of the conditioned planted model in (2.7), we have

$$P_{\mathcal{E}^c}(X, Y) = \frac{\mathrm{Exp}_{\beta^*}\left[P(X, Y \mid \beta^*)\mathbf{1}_{\{\mathcal{E}^c\}}(X, \beta^*)\right]}{\mathbb{P}\{\mathcal{E}^c\}} = \frac{P(X)\,\mathrm{Exp}_{\beta^*}\left[P(Y \mid X, \beta^*)\mathbf{1}_{\{\mathcal{E}^c\}}(X, \beta^*)\right]}{\epsilon}$$

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

$$P_{\mathcal{E}}(X, Y) \leq \frac{P(X)\,\mathrm{Exp}_{\beta^*}\left[\mathbf{1}_{\{\mathcal{E}\}}(\beta^*, X)\right]}{\varepsilon(2\pi\sigma^2)^{n/2}} \leq \frac{P(X)}{\varepsilon(2\pi\sigma^2)^{n/2}}.$$

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

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

Multiplying both sides by ε leads to

$$\varepsilon D(P_{\mathcal{E}^c}||Q) \leq \varepsilon \log\frac{1}{\varepsilon} + \frac{\varepsilon\, n}{2}\log(\lambda^2) + \frac{\mathrm{Exp}\left[\|Y\|^2\mathbf{1}_{\{\mathcal{E}^c\}}(\beta^*, X)\right]}{2\lambda^2\sigma^2}$$

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

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

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

2.4.3 Proof of Theorem 2.2.4

We are ready to prove Theorem 2.2.4.

Proof of Theorem 2.2.4. First, we prove (2.14) under the theorem assumptions. Let $\mathcal{E}$ be $\mathcal{E}_{\gamma,\tau}$ with γ and τ given in Theorem 2.2.3. It follows from Theorem 2.2.3 that $D(P_{\mathcal{E}}\|Q_{\lambda_0}) = o_p(1)$. Moreover, it follows from 2.8.1 and $k = o\,(p)$ that

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

Thus we get from 2.4.2 that for $\lambda^2 = k/\sigma^2 + 1$ and

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

where the last equality holds due to $k = o(p)$ and $k/\sigma^2 \ge C$ for a sufficiently large constant C. In view of the upper bound in (2.43), we immediately get $D(P\|Q_{\lambda_0}) = o_p(1)$ as desired.

Next we prove (2.15). Note that if $\lfloor(1-\alpha)n_{\text{info}}\rfloor \le 1$, then (2.15) is trivially true. Hence, we assume $\lfloor(1-\alpha)n_{\text{info}}\rfloor \ge 2$ in the following. Applying Lemma 2.4.1 with $n = \lfloor(1-\alpha)n_{\text{info}}\rfloor$ and $m = \lfloor(1-\alpha)n_{\text{info}}\rfloor - 1$ yields that

$$\frac{\text{MSE}(\hat{\beta}^*)}{k} \ge \left(1+\frac{\sigma^2}{k}\right)\exp\left\{-2D(P||Q_{\lambda_0})\right\} - \frac{\sigma^2}{k} = 1 - o_p\left(1\right). \tag{2.44}$$

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

2.5 Proof of Positive Results for Recovery and Detection

In this section we state and prove the positive result.

2.5.1 Proof of Theorem 2.2.5

Towards proving Theorem 2.2.5, we need the following lemma.

Lemma 2.5.1. *Let $X \in \mathbb{R}^{n\times p}$ with i.i.d. $\mathcal{N}(0,1)$ entries and $W \sim N(0,\sigma^2 I_n)$. Furthermore, assume that $\beta^*, (\beta^*)' \in \{0,1\}^p$ are two k-sparse vectors with $\|\beta^* - (\beta^*)'\|^2 = 2\ell$ for some $\ell \in \{1,\dots,k\}$. Then*

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

Proof. Let $Q(x)$ be the complementary cumulative distribution function of the standard Gaussian distribution, that is for any $x \in \mathbb{R}$, $Q(x) = \mathbb{P}\left[Z \ge x\right]$ for $Z \sim \mathcal{N}(0,1)$. The Chernoff bound gives

$Q(x) \leq e^{-x^2/2}$ for all $x \geq 0$. Then

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

where (a) holds because conditioning on X, $\frac{-W^T X(\beta^* - (\beta^*)')}{\sigma\|X(\beta^* - (\beta^*)')\|} \sim \mathcal{N}(0,1)$; (b) holds due to $Q(x) \leq e^{-x^2/2}$; the last inequality follows from $\|X(\beta^* - (\beta^*)')\|_2^2/(2\ell) \sim \chi^2(n)$ and $\mathbb{E}_{Z\sim\chi^2(1)}\left[e^{-tZ}\right] = \frac{1}{\sqrt{1+2t}}$ for $t > 0$. □

We now proceed with the proof of Theorem 2.2.5.

Proof of Theorem 2.2.5. First, note that when $k = o(p)$, (2.18) readily follows from (2.17). In particular, observe that since $\hat{\beta}^*, \beta^* \in \{0,1\}^p$ are binary k-sparse vectors, it follows that $\|\hat{\beta}^* - \beta^*\|^2 \leq 2k$ and therefore

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

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

It remains to prove (2.17). Set for convenience

$$d \triangleq \left\lceil \frac{k}{\log(p/k)} \right\rceil. \tag{2.45}$$

By the definition of the MLE,

$$\|W + X\left(\beta^* - \hat{\beta^*}\right)\|^2 = \|Y - X\hat{\beta^*}\|^2 \leq \|Y - X\beta^*\|^2 = \|W\|^2.$$

Hence,

$$\begin{aligned}
&\left\{\|\hat{\beta^*} - \beta^*\|^2 \geq 2d\right\} \\
&= \cup_{\ell=d}^{k} \left\{\exists (\beta^*)' \in \{0,1\}^p : \|(\beta^*)'\|_0 = k, \|(\beta^*)' - \beta^*\|^2 = 2\ell, \|W + X(\beta^* - (\beta^*)')\|^2 \leq \|W\|^2\right\}.
\end{aligned}$$

By a union bound and Lemma 2.5.1, we have that

$$\begin{aligned}
\mathbb{P}\left\{\|\hat{\beta^*} - \beta^*\|^2 \geq 2d\right\} &\leq \sum_{\ell=d}^{k} \binom{k}{\ell}\binom{p-k}{\ell}\left(1 + \frac{\ell}{2\sigma^2}\right)^{-n/2} \overset{(a)}{\leq} \sum_{\ell=d}^{k}\left(\frac{ke}{\ell}\right)^{\ell}\left(\frac{pe}{\ell}\right)^{\ell}\left(1 + \frac{\ell}{2\sigma^2}\right)^{-n/2} \\
&\overset{(b)}{\leq} \sum_{\ell=d}^{k}\left(\frac{e^2pk}{d^2}\right)^{\ell}\left(1 + \frac{\ell}{2\sigma^2}\right)^{-n/2} \triangleq \sum_{\ell=d}^{k} \exp\left(h(\ell) - \ell\right), \qquad (2.46)
\end{aligned}$$

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

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

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

$$\max_{d \leq \ell \leq k} h(\ell) \leq \max\left\{h(d), h(k)\right\}. \qquad (2.47)$$

We proceed to upper bound $h(d)$ and $h(k)$. Note that

$$\left(1 + \frac{\log 2}{\log\left(1 + k/(2\sigma^2)\right)}\right)\log\left(1 + \frac{k}{2\sigma^2}\right) \geq \log\left(1 + k/\sigma^2\right). \qquad (2.48)$$

Thus, it follows from (2.16) that

$$n \geq \frac{\log\left(1 + k/\sigma^2\right)}{\log\left(1 + \frac{k}{2\sigma^2}\right)}\left(1 + \frac{4\log\log(p/k)}{\log(p/k)}\right) n_{\text{info}} = \frac{2k\log(p/k)}{\log\left(1 + \frac{k}{2\sigma^2}\right)}\left(1 + \frac{4\log\log(p/k)}{\log(p/k)}\right). \qquad (2.49)$$

Then we conclude that

$$\begin{aligned} h(k) &= -\frac{n}{2}\log\left(1+\frac{k}{2\sigma^2}\right)+k\log\left(\frac{e^3pk}{d^2}\right) \\ &\overset{(2.49)}{\leq} -k\log(p/k)-4k\log\log(p/k)+k\log\left(\frac{e^3pk}{d^2}\right) \\ &\overset{(2.45)}{\leq} -k\log(p/k)-4k\log\log(p/k)+k\log\left(\frac{e^3pk\log^2(p/k)}{k^2}\right) \\ &= -2k\log\log(p/k)+3k. \end{aligned} \tag{2.50}$$

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

$$\begin{aligned} h(d) &= -\frac{n}{2}\log\left(1+\frac{d}{2\sigma^2}\right)+d\log\left(\frac{e^3pk}{d^2}\right) \\ &\overset{(2.49)}{\leq} -\left(1+\frac{4\log\log(p/k)}{\log(p/k)}\right)\frac{k\log(p/k)}{\log\left(1+k/(2\sigma^2)\right)}\log\left(1+\frac{d}{2\sigma^2}\right)+d\log\left(\frac{e^3pk}{d^2}\right). \end{aligned} \tag{2.51}$$

Let

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

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

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

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

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

Combining the last displayed equation with (2.51) gives that

$$\begin{aligned} h(d) &\leq -\left(1+\frac{4\log\log(p/k)}{\log(p/k)}\right)d\log(p/k)+d\log\left(\frac{e^3pk}{d^2}\right) \\ &\overset{(2.45)}{\leq} -d\log(p/k)-4d\log\log(p/k)+d\log\left(\frac{e^3pk\log^2(p/k)}{k^2}\right) \\ &\leq -2d\log\log(p/k)+3d. \end{aligned}$$

Combining the last displayed equation with (6.57) and (2.47), we get that

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

Combining the last displayed equation with (2.46) yields that

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

where the last inequality holds under the assumption $\log\log(p/k) \ge 1$. This completes the proof of Theorem 2.2.5.

□

2.5.2 Proof of Theorem 2.2.6

Proof. Under the planted model, we have

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

Note that $\|W\|^2/\sigma^2 \sim \chi^2(n)$ and $\|W+X\beta^*\|^2/(k+\sigma^2) \sim \chi^2(n)$. It follows from the concentration inequality for chi-square distributions that

$$\mathbb{P}\left\{\|W\|^2 \ge \sigma^2\left(n + 2\sqrt{nt} + 2t\right)\right\} \le e^{-t},$$

and

$$\mathbb{P}\left\{\|W + X\beta^*\|^2 \le (k+\sigma^2)\left(n - 2\sqrt{nt}\right)\right\} \le e^{-t}.$$

Therefore, for any t_n such that $t_n \to +\infty$ as $n \to +\infty$,

$$P\left(\mathcal{T}(X,Y) \geq \frac{\sigma^2}{k+\sigma^2}\frac{n+2\sqrt{nt_n}+2t_n}{n-2\sqrt{nt_n}}\right) \to 0.$$

In particular, using for example $t_n = \log n = o(n)$ we have $\frac{n+2\sqrt{nt_n}+2t_n}{n-2\sqrt{nt_n}} = 1+o(1)$, we can easily conclude from the definition of τ that

$$P\left(\mathcal{T}(X,Y) \geq \tau\right) \to 0.$$

Meanwhile, under the the null model, we have

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

Note that W and X are independent; thus we condition on X in the sequel. We have

$$\begin{aligned}
&\mathbb{E}\left[\min_{(\beta^*)' \in \{0,1\}^p, \|(\beta^*)'\|_0 = k} \|\lambda W - X\beta^*\|^2\right] \\
&\geq \min_{z_1,\ldots,z_M \in \mathbb{R}^n} \mathbb{E}\left[\min_{m \in [M]} \|\lambda W - z_m\|^2\right] \\
&\geq \mathbb{E}\left[\|\lambda W\|^2\right] M^{-2/n} \\
&= n\lambda^2\sigma^2 M^{-2/n}, \qquad (2.52)
\end{aligned}$$

where $M = \binom{p}{k}$ and the last inequality holds because the distortion rate function $D(R) = \sigma^2 \exp(2R)$ provides a non-asymptotic lower bound on the distortion of an i.i.d. $\mathcal{N}(0,\sigma^2)$ source with rate $R = \frac{1}{n}\log M$ (See e.g. [CT06, Section 10.3.2]).

Define $f : \mathbb{R}^n \to \mathbb{R}$,

$$f(w) = \min_{(\beta^*)' \in \{0,1\}^p, \|(\beta^*)'\|_0 = k} \|\lambda w - X\beta^*\|$$

It follows that f is λ-Lipschitz and thus in view of the Gaussian concentration inequality for Lipschitz functions (see, e.g. [BLM13, Theorem 5.6]), we get that

$$\mathbb{P}\left\{|f(W) - \mathbb{E}\left[f(W)\right]| \geq t\right\} \leq 2\exp\left(-\frac{t^2}{2\lambda^2\sigma^2}\right). \qquad (2.53)$$

Thus

$$\begin{aligned}\mathsf{var}\left(f(W)\right) &= \mathbb{E}\left[\left(f(W) - \mathbb{E}\left[f(W)\right]\right)^2\right] \\ &= \int_0^\infty \mathbb{P}\left\{\left(f(W) - \mathbb{E}\left[f(W)\right]\right)^2 \geq t\right\} dt \\ &\leq \int_0^\infty 2\exp\left(-\frac{t}{2\lambda^2\sigma^2}\right) dt \\ &= 4\lambda^2\sigma^2.\end{aligned}$$

Combining the last displayed equation with (2.52) gives that

$$\mathbb{E}\left[f(W)\right] \geq \sqrt{\mathbb{E}\left[f^2(W)\right] - 4\lambda^2\sigma^2} \geq \lambda\sigma\sqrt{nM^{-2/n} - 4}.$$

Combining the last displayed equation with (2.53), we get that for any t_n such that

$$t_n \to +\infty$$

as $n \to +\infty$,

$$\mathbb{P}\left\{f(W) \leq \lambda\sigma\sqrt{nM^{-2/n} - 4} - \lambda\sigma t_n\right\} \to 0.$$

Also, it follows from the concentration inequality for chi-square distributions that

$$\mathbb{P}\left\{\|W\|^2 \geq \sigma^2\left(n + 2\sqrt{nt_n} + 2t_n\right)\right\} \to 0.$$

Thus, recalling that $T(X,Y) = f^2(W)/\|\lambda W\|^2$, we get that

$$Q\left(T(X,Y) \leq \frac{\left(\sqrt{nM^{-2/n} - 4} - t_n\right)^2}{\left(n + 2\sqrt{nt_n} + 2t_n\right)}\right) \to 0. \tag{2.54}$$

By assumption (2.20), there exists a positive constant $\alpha > 0$ such that

$$n \geq \frac{2\log M}{\log\left(1 + k/\sigma^2\right) + \log(1-\alpha)}.$$

It follows that

$$M^{2/n} \leq (1-\alpha)\left(1+k/\sigma^2\right).$$

Since

$$\tau = \frac{1}{(1-\alpha/2)\left(1+k/\sigma^2\right)}$$

we have

$$\tau < \frac{1}{(1-\alpha)\left(1+k/\sigma^2\right)} \leq M^{-2/n}.$$

By assumption (2.19), $nM^{-2/n} \to +\infty$. Hence, there exists a sequence of t_n such that $t_n \to +\infty$ and $t_n = o(\sqrt{n}M^{-1/n})$. In particular, for this choice of t_n, combining the above we have

$$\liminf_n \frac{\left(\sqrt{nM^{-2/n}-4}-t_n\right)^2}{\left(n+2\sqrt{nt_n}+2t_n\right)} > \tau.$$

Hence from (2.54) we can conclude

$$Q\left(T(X,Y) \leq \tau\right) \to 0.$$

Hence indeed,

$$P\left(T(X,Y) \geq \tau\right) + Q\left(T(X,Y) \leq \tau\right) \to 0,$$

which shows that $T(X,Y)$ with threshold τ indeed achieves the strong detection.

□

2.6 Conclusion and Open Problems

In this Chapter, we establish an *All-or-Nothing* information-theoretic phase transition for recovering a k-sparse vector $\beta^* \in \{0,1\}^p$ from n independent linear Gaussian measurements $Y = X\beta^* + W$ with noise variance σ^2. In particular, we show that the MMSE normalized by the trivial MSE jumps from 1 to 0 at a critical sample size $n_{\text{info}} = \frac{2k\log(p/k)}{\log(1+k/\sigma^2)}$ within a small window of size ϵn_{info}. The constant $\epsilon > 0$ can be made arbitrarily small by increasing the signal-to-noise ratio k/σ^2. Interestingly, the phase transition threshold n_{info} is asymptotically equal to the ratio of entropy $H(\beta^*)$ and the AWGN channel capacity $\frac{1}{2}\log\left(1+k/\sigma^2\right)$. Towards estab-

lishing this All-or-Northing phase transition, we also study a closely related hypothesis testing problem, where the goal is to distinguish this planted model P from a null model Q_λ where (X, Y) are independently generated and $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \lambda^2\sigma^2)$. When $\lambda = \lambda_0 = \sqrt{k/\sigma^2 + 1}$, we show that the sum of Type-I and Type-II testing errors also jumps from 1 to 0 at n_{info} within a small window of size ϵn_{info}.

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

For detection, as argued in 2.9, $k = o\left(p^{1/2}\right)$ is needed for n_{info} being the detection threshold, because weak detection is achieved for all $n = \Omega(n_{\text{info}})$ when $k = \Omega(p^{1/2})$, that is the weak detection threshold becomes $o(n_{\text{info}})$. The identification of the precise detection threshold when $\Omega(p^{1/2}) \le k \le o(p)$ is an interesting open problem.

For recovery, however, we believe that the recovery threshold still equals n_{info} when $\Omega\left(p^{1/2}\right) \le k \le o(p)$. To prove this, we propose to study the detection problem where both the (conditional) mean and the covariance are matched between the planted and null models. Specifically, let us consider a slightly modified null model Q with the matched conditional mean $\mathbb{E}_Q[Y|X] = \mathbb{E}_P[Y|X] = \frac{k}{p}X\mathbf{1}$ and the matched covariance $\mathbb{E}_Q\left[YY^\top\right] = \mathbb{E}_P\left[YY^\top\right]$, where $\mathbf{1}$ denotes the all-one vector. For example, if X, W are defined as before and $Y \triangleq \frac{k}{p}X\mathbf{1} + \lambda W$ with λ equal to $\sqrt{\frac{k}{\sigma^2} + 1 - \frac{k^2}{p}}$, then both the mean and covariance constraints are satisfied. It is an open problem whether this new null model is indistinguishable from the planted model P when $n \le (1-\epsilon)n_{\text{info}}$ and $\Omega\left(p^{1/2}\right) \le k \le o(p)$. If the answer is affirmative, then we may follow the analysis road map in this Chapter to further establish the impossibility of recovery.

Finally, another interesting question for future work is to understand the extent to which the All-or-Nothing phenomenon applies beyond the binary vectors setting or the Gaussian assumptions on (X, W). In this direction, some recent work [Ree17] has shown that under mild conditions on the distribution of β^*, the distance between the planted and null models can be bounded in term of "exponential moments" similar to the ones studied in 2.7.

2.7 Appendix A: Hypergeometric distribution and exponential moment bound

Throughout this section, we fix

$$\lambda^2 = k/\sigma^2 + 1, \quad \text{and} \quad \tau = k\left(1 - \frac{1}{\log^2 \lambda^2}\right). \tag{2.55}$$

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

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

for a given interval $[a, b]$. It turns out this "exponential" moment exhibit quantitatively different behavior in the following three different regimes of overlap S: small regime ($s \le \epsilon k$), intermediate regime ($\epsilon k < s \le \tau$), and large regime ($s \ge \tau$), where ϵ is given in (2.57).

In the sequel, we first prove 2.7.2, which focuses on the small and intermediate regimes under the assumption $n \le n_{\mathrm{info}}$. Then we prove 2.7.5, which focuses on the large regime under the assumption $n \le (1-\alpha)n_{\mathrm{info}}/2$ for $\alpha \in (0, 1/2)$.

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

Lemma 2.7.1. *Let $p, k \in \mathbb{N}$. Then for $S \sim \mathrm{Hyp}(p, k, k)$ and any $s \in [k]$,*

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

Proof. We have

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

□

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

Lemma 2.7.2. *Suppose $n \le n_{\mathrm{info}}$.*

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

$$\mathrm{Exp}_{S\sim\mathrm{Hyp}(p,k,k)}\left[\left(1-\frac{S}{k+\sigma^2}\right)^{-n}\mathbf{1}_{\{S\le\epsilon k\}}\right] = 1+o_p(1), \tag{2.56}$$

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

$$\epsilon = \epsilon_{k,p} = \frac{\log\log(p/k)}{2\log(p/k)}, \tag{2.57}$$

it holds that

$$\mathrm{Exp}_{S\sim\mathrm{Hyp}(p,k,k)}\left[\left(1-\frac{S}{k+\sigma^2}\right)^{-n}\mathbf{1}_{\{\epsilon k<S\le\tau\}}\right] = o_p(1), \tag{2.58}$$

Proof. Using Lemma 2.7.1,

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

Note that

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

where the last equality holds due to $k \le p^{1/2-\delta}$ for some constant $\delta \in (0, 1/2)$. Thus, to show (2.56) it suffices to show

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

and to show (2.58) it suffices to show

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

We first prove (2.56).

Proof of (2.56): Using the fact that $\binom{k}{s} \leq k^s$, we have

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

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

$$f(s) = -s \log \frac{p}{k^2} - n_{\text{info}} \log\left(1 - \frac{s}{k+\sigma^2}\right).$$

Claim 2.7.3. *Suppose $k \leq p^{1/2-\delta}$ for a constant $\delta \in (0, 1/2)$ and $\epsilon \leq 1/2$. There exists a constant $C_1 = C_1(\delta) > 0$, such that if $k/\sigma^2 \geq C_1$ then it holds that for any $s \in [1, \epsilon k]$, $f(s) \leq -\frac{1}{2} s \log \frac{p}{k^2}$.*

Proof of the Claim. Standard calculus implies that for $x \in (0,1)$, $\log(1-x) \geq -(1+x)x$. Hence, for $0 \leq x \leq \epsilon \leq 1/2$,

$$\log(1-x) \geq -(1+\epsilon)x. \tag{2.59}$$

Using this inequality it follows that for since for any $s \in [1, \epsilon k]$ $\frac{s}{k+\sigma^2} \leq \epsilon$, it also holds

$$f(s) \leq -s \log \frac{p}{k^2} + n_{\text{info}}(1+\epsilon)\frac{s}{k+\sigma^2} = s\left(-\log \frac{p}{k^2} + \frac{n(1+\epsilon)}{k+\sigma^2}\right) \leq -\frac{1}{2} s \log \frac{p}{k^2},$$

where the last inequality holds under the assumption that

$$n_{\text{info}} \leq \frac{(k+\sigma^2)\log \frac{p}{k^2}}{2(1+\epsilon)}.$$

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

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

which holds if and only if

$$\left[1 - \frac{4(1+\epsilon)}{(1+\sigma^2/k)\log(1+k/\sigma^2)}\right] \log \frac{p}{k} \geq \log k. \tag{2.60}$$

By assumption, $k \leq p^{1/2-\delta}$ for $\delta \in (0, \frac{1}{2})$. Hence, (2.60) is satisfied if

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

Since $\epsilon \leq 1/2$, there exists a constant $C_1 = C_1(\delta) > 0$ depending only on δ such that if

$$\frac{k}{\sigma^2} \geq C_1$$

then the last displayed equation is satisfied. This completes the proof of the claim. □

Using the above claim we conclude that

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

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

Next we prove (2.58). Again it suffices to prove (2.58) for $n = n_{\text{info}}$.

Proof of (2.58): Note that $\binom{k}{s} \leq 2^k$. Hence,

$$\begin{aligned}\sum_{s=\lceil \epsilon k \rceil}^{\lfloor \tau \rfloor} \binom{k}{s} \left(\frac{k}{p-k+1}\right)^s e^{-n_{\text{info}}\log\left(1-\frac{s}{k+\sigma^2}\right)} &\leq 2^k \sum_{s=\lceil \epsilon k \rceil}^{\lfloor \tau \rfloor} \left(\frac{k}{p-k+1}\right)^s e^{-n_{\text{info}}\log\left(1-\frac{s}{k+\sigma^2}\right)} \\ &= 2^k \sum_{s=\lceil \epsilon k \rceil}^{\lfloor \tau \rfloor} e^{-s\log\frac{p}{k} - n_{\text{info}}\log\left(1-\frac{s}{k+\sigma^2}\right) - s\log\frac{(p-k+1)}{p}}.\end{aligned}$$

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

$$g(s) \triangleq -s\log\frac{p}{k} - n_{\text{info}}\log\left(1 - \frac{s}{k+\sigma^2}\right). \tag{2.61}$$

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

Claim 2.7.4. *There exist a universal constant $C_2 > 0$ such that if $k/\sigma^2 \ge C_2$, then $g(\tau) \le -\frac{1}{2}k\log(p/k)$ and $g(\epsilon k) \le -\frac{\epsilon k}{2}\log\frac{p}{k}$.*

Proof of the Claim. We first upper bound $g(\tau)$.

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

where the last equality holds by plugging in the expressions of τ and n_{info}.

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

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

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

For $g(\epsilon k)$, applying (2.59), we get that

$$g(\epsilon k) = -\epsilon k \log\frac{p}{k} - n_{\text{info}} \log\left(1 - \frac{\epsilon k}{k+\sigma^2}\right) \le -\epsilon k \log\frac{p}{k} + \frac{n_{\text{info}}\epsilon k}{k+\sigma^2}(1+\epsilon)$$

which equals

$$\epsilon k \left(-\log\frac{p}{k} + \frac{n_{\text{info}}(1+\epsilon)}{k+\sigma^2}\right).$$

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

$$-\log\frac{p}{k} + \frac{n_{\text{info}}(1+\epsilon)}{k+\sigma^2} \le -\frac{1}{2}\log\frac{p}{k}$$

which holds if and only if

$$n_{\text{info}} = \frac{2k\log(p/k)}{\log(1+k/\sigma^2)} \le \frac{(k+\sigma^2)\log(p/k)}{2(1+\epsilon)}$$

or equivalently

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

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

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

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

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

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

□

Finally, we upper bound the "exponential" moment in the large overlap regime ($s \geq \tau$) where τ is defined in (2.55).

Lemma 2.7.5. *Suppose that $k \leq cp$ for $c \in (0,1)$ and $k/\sigma^2 \geq C$ for a sufficiently large universal constant C. If $n \leq \frac{1}{2}(1-\alpha)n_{\text{info}}$ for some $\alpha \leq 1/2$, then*

$$\text{Exp}_{S\sim\text{Hyp}(p,k,k)}\left[\left(1-\frac{S}{k+\sigma^2}\right)^{-n}\mathbf{1}_{\{S\geq\tau\}}\right] \leq \exp\left(-\alpha k\log\frac{p}{k}+\log\frac{2-c}{1-c}\right). \tag{2.62}$$

Proof. Using Lemma 2.7.1, we get that

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

where $g_n(s)$ is given by

$$g_n(s) \triangleq -s\log\frac{p}{k} - n\log\left(1-\frac{s}{k+\sigma^2}\right).$$

Note that $g_n(s)$ is convex in s for $\tau \leq s \leq k$. Hence, the maximum of $g_n(s)$ over $s \in [\tau, k]$ is achieved at either $s = \tau$ or $s = k$. In view of (2.61) and Claim 2.7.4, for all $n \leq n_{\mathrm{info}}$.

$$g_n(\tau) \leq g_{n_{\mathrm{info}}}(\tau) = g(\tau) \leq -\frac{1}{2}k\log\frac{p}{k}.$$

Thus it remains to upper bound $g_n(k)$.

Claim 2.7.6. *Assume $n \leq \frac{1}{2}(1-\alpha)n_{\mathrm{info}}$ for some $\alpha > 0$. Then $g_n(k) \leq -\alpha k\log(p/k)$.*

Proof of the Claim. For all $n \leq \frac{1}{2}(1-\alpha)n_{\mathrm{info}}$,

$$\begin{aligned}
g_n(k) &= -k\log\frac{p}{k} - n\log\left(1-\frac{k}{k+\sigma^2}\right) \\
&= -k\log\frac{p}{k} + \frac{1}{2}(1-\alpha)n_{\mathrm{info}}\log\left(1+\frac{k}{\sigma^2}\right) \\
&= -k\log\frac{p}{k} + (1-\alpha)k\log\left(\frac{p}{k}\right) \\
&= -\alpha k\log\frac{p}{k}.
\end{aligned}$$

□

In view of the above claim and the assumption that $\alpha \leq 1/2$, we conclude that for all

$n \le \frac{1}{2}(1-\alpha)n_{\text{info}}$,

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

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

2.8 Appendix B: Probability of the conditioning event

In this section, we upper bound the probability that the conditioning event $\mathcal{E}_{\gamma,\tau}$ defined in (2.11) does not happen.

Lemma 2.8.1. *Consider the set $\mathcal{E}_{\gamma,\tau}$ defined in (2.11). Let $\tau = k(1-\eta)$ for some $\eta \in [0,1]$. Then we have*

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

Furthermore, for

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

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

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

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

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

Observe also that the number of different $(\beta^*)'$ with $\langle \beta^*, (\beta^*)' \rangle \geq \tau$ is at most

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

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

$$\mathbb{P}\left\{(X, \beta^*) \in \mathcal{E}^c_{\gamma,\tau} \mid \beta^*\right\} \leq Q_{\chi^2_n}\left(n(2+\gamma)\right) \sum_{\ell=0}^{\lfloor \eta k \rfloor} \binom{k}{\ell} \binom{p-k}{\ell}, \tag{2.63}$$

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

For all $x > 0$, we have (see, e.g., [LM00, Lemma 1]:

$$Q_{\chi^2_n}\left(n(1+\sqrt{x}+x/2)\right) \leq e^{-\frac{nx}{4}}. \tag{2.64}$$

Noting that $\sqrt{\gamma} + \gamma/2 \leq 1 + \gamma$ for all $\gamma > 0$, we see that $Q_{\chi^2_n}(n(2+\gamma)) \leq \exp\{-n\gamma/4\}$.

Next, using the inequalities $\binom{a}{b} \leq (\frac{ae}{b})^b$ for $a, b \in \mathbb{Z}_{>0}$ with $a < b$, that $x \to x \log x$ decreases in $(0, \frac{1}{e})$, and $\sum_{i=0}^{d} \binom{m}{i} \leq (\frac{me}{d})^d$ for $d, m \in \mathbb{Z}_{>0}$ with $d < m$ (see, e.g., [Kum10]), we get that

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

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

For the second part, note that under our choice of η,

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

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

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

Combining the last two displayed equation yields that

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

This completes the proof of the lemma.

□

2.9 Appendix C: The reason why $k = o(p^{1/2})$ is needed for weak detection threshold n_{info}

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

Proposition 2.9.1. *Suppose*

$$\frac{nk^2}{p\left(k + \sigma^2 - k^2/p\right)} = \Omega_p(1). \tag{2.65}$$

Then weak detection is information-theoretically possible.

Remark 2.9.2. *If $k/\sigma^2 = \Omega_p(1)$ and k/p is bounded away from 1, then (2.66) is equivalent to*

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

Recall that

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

Therefore, if furthermore $k = \Omega_p(p^{1/2})$ and $\log(p/k) = \Omega_p\left(\log(1 + k/\sigma^2)\right)$,

then $n_{\text{info}}k/p = \Omega_p(1)$ *and hence weak detection is possible for all* $n = \Omega_p(n_{\text{info}})$.

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

$$\mathcal{T}(X, Y) = \left\langle Y, X\bar{\beta^*} \right\rangle;$$

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

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

Hence,

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

and

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

where $Q(x) = \int_x^\infty (2\pi)^{-1/2} \exp(-t^2/2)\, \text{ṭ}$ is the tail function of the standard Gaussian.

Therefore, as long as $\sqrt{\frac{k^2/p}{k+\sigma^2-k^2/p}}\|A\|$ does not converge to 0 in probability, then $P\left(\left\langle Y, X\bar{\beta^*} \right\rangle \leq 0\right) \leq 1/2 - \epsilon$ for some positive constant $\epsilon > 0$. Thus,

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

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

$$\frac{nk^2}{p\,(k + \sigma^2 - k^2/p)} = \Omega_p(1), \tag{2.66}$$

then weak detection is possible.

□

Chapter 3

The Computational-Statistical Gap for High Dimensional Regression. The Hard Regime.

3.1 Introduction

As mentioned in Chapter 1, this Chapter and Chapter 4 studies the computational-statistical gap for the high dimensional linear regression model. We study the model under the assumption described in Subsection 1.1.1. Specifically, we adopt the assumptions that $X \in \mathbb{R}^{n\times p}$ and $W \in \mathbb{R}^{n\times 1}$ are independent matrices with $X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $W_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2)$ for some $\sigma^2 > 0$, and finally β^* is an arbitrary but fixed binary k-sparse vector.

The goal of this Chapter and Chapter 4 is to study whether there is a fundamental explanation of the computational statistical gap exhibited by the model using the notion of the Overlap Gap Property (see Subsection 1.3). We start with providing more details and an extended literature review on the computational-statistical gap of the model.

We begin with the computational limit. A lot of work has been devoted in particular to finding computationally efficient ways for recovering the binary k-sparse β^* from noisy linear measurements $Y = X\beta^* + W$. Note that recovering β^* is equivalent with recovering its support. In the noiseless setting ($W = 0$), Donoho and Tanner show in [DT10] that the simple linear program: $\min \|\beta\|_1$ 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 binary k-sparse β^*, [MB06b],[Wai09b],[ZY06]. See also the recent book [FR13]. In particular, Wainwright [Wai09b] 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 based 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 the LASSO based method provably fails to recover the support of β^* exactly, also w.h.p. We note that the impact of σ^2 on this threshold is asymptotically negligible when $\sigma^2/k \to 0$. It will be convenient for us to keep it though and thus we denote $(2k+\sigma^2)\log p$ by n_{alg}. At the present time no tractable (polynomial time) algorithms are known for the support recovery when $n \leq n_{\text{alg}}$.

On the complimentary direction, results regarding the information theoretic limits for the problem of support recovery have also been obtained [DT10],[Wai09a],[WWR10], [RG12], [RG13], [SC15]. 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 [Wai09a], which follows from an appropriate use of Fano's inequality, when applied to our context below involving vectors β^* with binary values, yields one information-theoretic lower bound. it is shown 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 $\frac{\epsilon}{2}$. Interestingly, this lower bound value does not depend on the value of k. 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. Using the tight characterization of the Gaussian communication channel capacity (see e.g. Theorem 10.1.1. in [CT06]) 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 [WWR10] using similar ideas that the exact recovery of β^* is information theoretically impossible when n smaller than $2k\log p/\log(1+2k/\sigma^2)$, which is the information theoretic limit of this Gaussian channel for general k. Furthermore, Rad in [Rad11] showed that it is information-theoretically possible to recover exactly β^* with $Ck\log p/\log(1+2k/\sigma^2)$ samples for some sufficiently large constant $C>0$. Finally, in Chapter 2 we established that the threshold $n_{\text{info}} \triangleq 2k\log p/\log(1+2k/\sigma^2)$ is the exact statistical limit of the problem, when β^* is chosen from a uniform prior over the binary k-sparse vectors and k/σ^2 is sufficiently large. Notice that there is a negligible discrepancy between the value of n_{info} defined here and in Chapter 2 which is $2k\log p/\log(1+k/\sigma^2)$. The reason the discrepancy is negligible is because as k/σ^2 grows, which is the regime of interest, the ratio between the two thresholds $2k\log p/\log(1+2k/\sigma^2)$ and $2k\log p/\log(1+k/\sigma^2)$ converges to one. The critical threshold n_{info} will play a fundamental role in the results of this Chapter.

The regime $n\in[n_{\text{info}}, n_{\text{alg}}]$ remains largely unexplored from the algorithmic perspective and comprises what is known as a computational statistical gap (see the Introduction of the thesis - Chapter 1 - for more details on computational-statistical gaps) and the results presented in this Chapter are devoted to studying this gap. Specifically we would like to study the following question:

Is there a fundamental explanation for the computational-statistical gap when $n\in[n_{\text{info}}, n_{\text{alg}}]$?

Towards this goal, for the regression model $Y=X\beta^*+W$, we consider the corresponding maximum likelihood estimation problem:

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

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 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 Chapter: (1) What is the value of the squared error estimator $\min_{\beta\in\{0,1\}^p,\|\beta\|_0=k}\|Y-X\beta\|_2=\|Y-X\beta_2\|_2$; and (2) how well does the optimal vector β_2 approximate the ground truth vector β^*? As an outcome we seek to shed light on the algorithmic barriers in the regime $n\in[n_{\text{info}},n_{\text{alg}}]$.

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 $n_{\text{info}}=2k\log p/\log(2k/\sigma^2+1)$ which was shown in [WWR10] to be the information theoretic lower bound for the exact recovery of β^* is the phase transition point with the following "all-or-nothing" property. When n exceeds n_{info} asymptotically, $(2k)^{-1}\|\beta_2-\beta^*\|_0\approx 0$, and when n is asymptotically below n_{info}, $(2k)^{-1}\|\beta_2-\beta^*\|_0\approx 1$. Namely, when $n>n_{\text{info}}$ the recovery of β^* is achievable via solving Φ_2, whereas below n_{info} the optimization problem Φ_2 "misses" the ground truth vector β^* almost entirely. Since, as discussed above, when $n<n_{\text{info}}$, the recovery of β^* is impossible information theoretically, our result implies that n_{info} is indeed the information theoretic threshold for this problem. We recall that n_{info} 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 $n_{\text{alg}} = (2k + \sigma^2) \log p$. We note also that our result improves upon the result of Wainwright [Wai09a], who shows that the recovery of β^* is possible by the brute force search method, though only when n is of the order $O(k \log p)$.

Notice that this result does not compare immediately with the information-theoretic phase transition results of Chapter 2. The reason is that in Chapter 2 it is assumed that the vector β^* is chosen from a uniform prior over the binary k-sparse vectors, while this result holds for any fixed binary k-sparse vector β^*. Nevertheless, it establishes the same all-or-nothing behavior, but this time only for the performance of the MLE of the problem.

(b) We consider an intermediate optimization problem $\min_\beta n^{-\frac{1}{2}} \|Y - X\beta\|_2$ 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 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 $n > n_{\text{info}}$, and is w.h.p. $\Gamma(1) = (2k + \sigma^2)^{\frac{1}{2}} \exp\left(-\frac{k \log p}{n}\right)$ when $n < n_{\text{info}}$, both up to multiplicative constants. We further establish that the function Γ exhibits phase transition property at all three important thresholds $n_{\text{inf},1}, n_{\text{info}}$ and n_{alg}, described pictorially on Figures 6-2 in the next section. In particular, we prove that when $n > n_{\text{alg}}$, $\Gamma(\zeta)$ 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 $n_{\text{info}} < n < n_{\text{alg}}$, $\Gamma(\zeta)$ is non-monotonic and achieves the minimum value at $\zeta = 0$, and when $n_{\text{inf},1} < n < n_{\text{info}}$, $\Gamma(\zeta)$ is again non-monotonic and achieves the minimum value at $\zeta = 1$. In the critical case $n = n_{\text{info}}$, 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 $n > n_{\text{alg}}$, 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 Compressive Sensing succeed in identifying β^*. We conjecture that in this case even more straightforward, greedy type algorithms based on one step improvements might be able to recover β^*. At this stage, this remains a conjecture.

When n is below n_{alg} but above n_{info}, 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_{info}, 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. 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 ($k = 1$, large σ) information theoretic lower bound $n_{\text{inf},1}$, the overfitting situation 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 examples 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 solution satisfying a

set of randomly generated constraints, break down into clusters separated by cost barriers of a substantial size in some appropriate sense, [ACORT11, ACO08, MRT11, COE11, GSa, RV14, GSb]. As a result, these models indeed exhibit the OGP. For example, independent sets achieving near optimality 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 [GSa],[RV14] and [GSb] to establish a fundamental barriers on the power of so-called local algorithms for finding nearly largest independent sets. 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 [GL16].

The non-monotonicity of the function Γ for $n < n_{\text{alg}}$ 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 $C\sigma^2 \log p \leq n \leq ck \log p$, 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 $\|\beta - \beta^*\| = 2k\zeta$, discussed earlier. It remains an intriguing open question to verify whether the optimization problem Φ_2 is indeed algorithmically intractable in this regime.

(d) Finally, as an outcome of our geometric results for the lanscape of (Φ_2), we obtain negative results on the performance of the well-known ℓ_1-constrained quadratic optimization recovery scheme call LASSO when $n < cn_{\text{alg}}$. LASSO is defined as follows: let

$$\text{LASSO}_\lambda : \min_{\beta\in\mathbb{R}^p} n^{-1}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{3.1}$$

for appropriately chosen tuning parameter $\lambda > 0$. When $n > Cn_{\text{alg}}$, the optimal solution of

LASSO, and of it's closer relative linear program called Dantzig selector, has been shown to be approximating β^* up to the noise level [CT07, BRT09b, OTH13]. More specifically, an easy corollary of the seminar work by Bickel, Ritov and Tsybakov [BRT09b] applied to X with Gaussian iid entries implies that as long as $n \geq Cn_{\text{alg}} = Ck \log p$ for some sufficiently large constant $C > 0$ if $\lambda = A\sigma\sqrt{\log p/n}$ the optimal solution $\hat{\beta}_{\text{LASSO},\lambda}$ of LASSO_λ satisfies for some constant $c > 0$,

$$\|\hat{\beta} - \beta^*\|_2 \leq c\sigma \tag{3.2}$$

w.h.p. Condition (3.2) is known in the literature as ℓ_2-*stable recovery* of the vector β^*, w.h.p. Tighter results for the performance on LASSO and the constants c, C are established in the literature (see [OTH13] and references therein), yet they do not apply in the regime where the sparsity is sublinear to the feature size p, which as it is explained above, is the main focus of this work. Note that ℓ_2-stable recovery condition (3.2) is not comparable to support recovery, which was discuss previously on this chapter. Yet, interestingly, notice that even if the focus here is on ℓ_2-stable recovery, the best known computational limit for our setting where β^* is binary and k-sparse *remains of order* n_{alg} but the information-theoretic limit can be still established to be of order n_{info} [Rad11]. The extent to which LASSO can ℓ_2-stably recover the vector which fewer number of samples, remained before the present work, to the best of our knowledge, an open problem.

We establish that if $n^* \leq n < cn_{\text{alg}}$ for small enough $c > 0$ and β^* exactly k-sparse and binary, then for any

$$\lambda \geq \sigma\sqrt{\frac{1}{k}}\exp\left(-\frac{k\log p}{5n}\right)$$

the optimal solution of LASSO_λ fails to ℓ_2-stable recover β^* w.h.p More specifically, we show that under the assumptions described above the optimal solution of LASSO_λ, call it $\hat{\beta}_{\text{LASSO},\lambda}$, satisfies

$$\|\hat{\beta}_{\text{LASSO},\lambda} - \beta^*\|_2 \geq \sigma\exp\left(\frac{k\log p}{5n}\right) \tag{3.3}$$

w.h.p. Note that if $n = ck\log p$ the right hand side of (3.3) becomes $\exp\left(\frac{1}{c}\right)\sigma$ and in particular as $k\log p/n = c \to 0$, according to (3.3), the ratio $\|\hat{\beta}_{\text{LASSO}} - \beta^*\|_2/\sigma$ explodes to

infinity, indicating the failure of LASSO to ℓ_2-stably recover β^* in this regime.

Albeit our result does not apply for any arbitrarily small value of $\lambda > 0$ our result covers certain arguably important choices of λ in the literature of LASSO. Perhaps most importantly, our results covers the theoretically successful choice of the tuning parameter λ for LASSO_λ when $n \geq Cn_{\text{alg}}$ in [BRT09b] which, as explained above, shows that LASSO_λ with

$$\lambda = \lambda^* := A\sigma\sqrt{\log p/n}$$

for constant $A > 2\sqrt{2}$, ℓ_2-stably recovers β^*. Indeed, since in our case, we need c small enough, we have $n < k \log p$ and therefore

$$\lambda = \lambda^* \geq A\sigma\sqrt{\frac{1}{k}} > \sigma\sqrt{\frac{1}{k}}$$

which finally implies

$$\lambda = \lambda^* \geq \sigma\sqrt{\frac{1}{k}}\exp\left(-\frac{k\log p}{5n}\right).$$

An important feature of our result is that it is *quantitative*, as (3.3) gives a lower bound of how far the optimal solution of LASSO_λ is from β^* in the ℓ_2 norm. Interestingly, our lower bound depends exponentially on the ratio $k \log p/n$, implying a exponential rate of divergence from ℓ_2-stable recovery. Moreover, given the existing positive result of [BRT09b] for LASSO, our result confirms that $n_{\text{alg}} = k \log p$ is the *exact order of necessary number of samples* for LASSO_λ to ℓ_2-stably recover the ground truth vector β^*, when $\lambda \geq \sigma\sqrt{1/k}\exp(-k\log p/5n)$. Our result is therefore closed in spirit with the literature on LASSO in the context of support recovery where, as we discussed in the Introduction, a similar phase transition results is established around n_{alg} by Wainwright in [Wai09b].

In the specific case β^* is binary a natural modification of LASSO it is to add the box constraint $\beta \in [0,1]^p$ to the LASSO formulation. Such box constraints have been proven to improve the performance of LASSO in many cases, such as in signal processing applications [BTK$^+$17]. We show that in our case, our negative result for LASSO remains valid even

with the box constraint. Specifically, let us focus for any $\lambda > 0$ on

$$\text{LASSO(box)}_\lambda : \min_{\beta \in [0,1]^p} n^{-1}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1. \tag{3.4}$$

We show that if $n^* \leq n < cn_{\text{alg}}$ for small enough $c > 0$ and β^* is an exactly k-sparse binary vector, for any $\lambda \geq \sigma\sqrt{\frac{1}{k}}\exp\left(-\frac{k\log p}{5n}\right)$ the optimal solution of $\text{LASSO(box)}_\lambda$, call it $\hat{\beta}_{\text{LASSO(box)},\lambda}$, satisfies

$$\|\hat{\beta}_{\text{LASSO(box)},\lambda} - \beta^*\|_2 \geq \sigma\exp\left(\frac{k\log p}{5n}\right) \tag{3.5}$$

w.h.p., and therefore also fails to ℓ_2-stably recover β^* w.h.p.

3.1.1 Methods

In order to obtain estimates of the squared error for the problem Φ_2 we use a first and 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. In this model the interest is on estimating the quantity $\min_\beta \|Y - X\beta\|_2$ where β binary and k-sparse.

For every value $t > 0$ we consider the counting random variable Z_t equal to the number of k-sparse binary β such that $\|Y - X\beta\|_\infty \leq t$, where $\|x\|_\infty = \max_i |x_i|$ is the infinity norm. It turns out that while $\|\cdot\|_\infty$ norm estimates for the difference $Y - X\beta$ are easier to deal with, they provide sufficiently accurate information for the $\|\cdot\|_2$ norm of $Y - X\beta$ we originally care about; hence our focus on the former. We compute the expected value of Z_t and find a critical value t^* such that for $t < t^*$ this expectation converges to zero. Combining with Markov inequality we have $\mathbb{P}(Z_t \geq 1) \leq \mathbb{E}[Z_t] \to 0$ for all $t < t^*$ or $Z_t = 0$ w.h.p. for all $t < t^*$. In particular, t^* serves as a lower bound on $\min_\beta \|Y - X\beta\|_\infty$ where β binary and k-sparse. This technique of finding the lower bound t^* is known as the first moment method.

We then consider the second moment method for Z_t. In the naive form the second moment method would succeed if for $t > t^*$, $\mathbb{E}[Z_t^2]$ was close to $(\mathbb{E}[Z_t])^2$, as in this case the Paley-Zigmund inequality would give $\mathbb{P}(Z_t \geq 1) \geq \mathbb{E}[Z_t]^2 / \mathbb{E}[Z_t^2] \to 1$ and therefore t^* is also an upper bound for $\min_\beta \|Y - X\beta\|_\infty$. Unfortunately, the naive second moment estimation fails as it can be easily

checked that for t close to t^*, $\mathbb{E}[Z_t]^2 / \mathbb{E}[Z_t^2] \to 0$.

We consider an appropriate conditioning to make the second moment method work. We notice that the fluctuations of Y alone are enough to create a substantial gap between the two moments of Z_t. For this reason, we consider the conditional first and second moment of Z_t, where the conditioning is done on Y. The conditional second moment involves computing large deviations estimates on a sequence of coupled bi-variate normal random variables. A fairly detailed analysis of this large deviation estimate is obtained to arrive at the estimation of the ratio $\mathbb{E}[Z_t|Y]^2 / \mathbb{E}[Z_t^2|Y]$. We then employ the conditional version of the Paley-Zigmund inequality $\mathbb{P}(Z_t \geq 1|Y) \geq \mathbb{E}[Z_t|Y]^2 / \mathbb{E}[Z_t^2|Y]$ to obtain the lower bound $\mathbb{P}(Z_t \geq 1) \geq \mathbb{E}\big[\mathbb{E}[Z_t|Y]^2 / \mathbb{E}[Z_t^2|Y]\big]$ where expectation is taken over Y. Using the estimation on the lower bound we show that t^*, the first moment estimate, serves also as an upper bound for $\min_\beta \|Y - X\beta\|_\infty$, up to certain multiplicative constant factors.

To explain the success of the conditional technique notice that by tower property and Cauchy-Schwarz inequality

$$\mathbb{E}\left[\frac{\mathbb{E}[Z_t|Y]^2}{\mathbb{E}[Z_t^2|Y]}\right]\mathbb{E}\big[Z_t^2\big] = \mathbb{E}\left[\frac{\mathbb{E}[Z_t|Y]^2}{\mathbb{E}[Z_t^2|Y]}\right]\mathbb{E}\big[\mathbb{E}\big[Z_t^2|Y\big]\big] \geq \mathbb{E}[\mathbb{E}[Z_t|Y]]^2 = \mathbb{E}[Z_t]^2$$

which equivalently gives

$$\mathbb{E}\left[\frac{\mathbb{E}[Z_t|Y]^2}{\mathbb{E}[Z_t^2|Y]}\right] \geq \frac{\mathbb{E}[Z_t]^2}{\mathbb{E}[Z_t^2]}$$

certifying that the lower bound on $\mathbb{P}(Z_t \geq 1)$ obtained through conditioning dominates the one from the direct application of Paley-Zigmund inequality.

Next we use the estimates from the Pure Noise model, for the original model involving the binary β^* with $\|\beta^*\|_0 = k$. We consider the $2^k = 2^{|\text{Support}(\beta^*)|}$ restricted versions of the original problem of interest (Φ_2) in which the optimization is conducted over the space of binary k-sparse vectors β where the support of β is constrained to intersect the support of β^* in a specific way. In this form the problem can be reduced to the Pure Noise problem in a relative straightforward way (see Section 3.4 for the exact reduction). This reduction alongside with the first and second moment estimates for the Pure Noise model described above allows us to approximate the optimal value of the restricted problems, and in particular of (Φ_2) as well.

Note that conditional first and second moment methods have been used extensively in the

literature (e.g. see [MNS15, BMV$^+$18, BMNN16, PWB16, **?**, BBSV18] for recent examples) but it is a common understanding that the appropriate choice of conditioning does not follow a universal reasoning. To the best of our knowledge this is the first time the conditional second moment method is used in the form described above and this might be of independent interest.

Organization The remainder of the Chapter is organized as follows. The description of the model, assumptions and the main results are found in the next section. Section 3.3 is devoted to the analysis of the Pure Noise model which is also defined in this section. Sections 3.4, 3.5 and 3.6 are devoted to proofs of our main results. We conclude in the last section with some open questions and directions for future research.

3.2 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. $N(0, \sigma^2)$ 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 $\text{Support}(\beta) := \{i : \beta_i = 0\}$. Namely, $\beta_i = 1$ if $i \in \text{Support}(\beta)$ 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 Chapter 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 of the order $k \log p$. 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

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

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 $\mathbb{L}_2$ 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:

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

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 $k - \ell$. Then $\phi_2 = \min_\ell \phi_2(\ell)$.

Consider the extreme cases $\ell = 0$ and $\ell = k$, we see that 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 3.2.1. *Suppose $k \log k \leq Cn$ for some constant C for all k, n. Then*

(a) W.h.p. as k increases

$$\phi_2(\ell) \geq e^{-\frac{3}{2}}\sqrt{2\ell + \sigma^2} \exp\left(-\frac{\ell \log p}{n}\right), \tag{3.6}$$

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*

$$\left\{\beta \in \{0,1\}^p : \|\beta\|_0 = k, \|\beta - \beta^*\|_0 = 2k,\ n^{-\frac{1}{2}}\|Y - X\beta\|_2 \leq D_0\sqrt{2k+\sigma^2}\exp\left(-\frac{k\log p}{n}\right)\right\} \tag{3.7}$$

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 Section 3.4 and relies on the analysis for the Pure Noise model developed in the next section. 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 3.2.3 below.

Now we will discuss some implications of Theorem 3.2.1. The expression $(2\ell+\sigma^2)^{\frac{1}{2}}\exp\left(-\frac{\ell\log p}{n}\right)$, appearing in the theorem above, motivates the following notation. Let the function $\Gamma : [0,1] \to \mathbb{R}_+$ be defined by

$$\Gamma(\zeta) = \left(2\zeta k + \sigma^2\right)^{\frac{1}{2}}\exp\left(-\frac{\zeta k\log p}{n}\right). \tag{3.8}$$

Then the lower bound (3.6) can be rewritten as

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

A similar inequality applies to (3.7).

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

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

Now combining this observation with the results of Theorem 3.2.1 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 3.2.2. *Under the assumptions of parts (a) and (b) of Theorem3.2.1, for every $\epsilon > 0$ and for every sufficiently large constant D_0 if $n \le k\log p/(3\log D_0)$, then w.h.p. as k increases,*

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

Proof. By Theorem 3.2.1 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 $\phi_2 \le \min(\phi_2(0), \phi_2(k))$. 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 Theorem3.2.1 gives provides the necessary bound on $\phi_2(k)$. □

As in the introduction, letting $n_{\text{info}} = \frac{2k\log p}{\log\left(\frac{2k}{\sigma^2}+1\right)}$, we conclude that $\min_\zeta \Gamma(\zeta) = \Gamma(1)$ when $n < n_{\text{info}}$ and $= \Gamma(0)$ when $n > n_{\text{info}}$, with the critical case $n = n_{\text{info}}$ (ignoring the integrality of n_{info}), 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 $n > n_{\text{info}}$ 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 $n < n_{\text{info}}$, 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 3.2.1, $\Gamma(\zeta)$ 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 3.2.3. *Let $\epsilon > 0$ be arbitrary. Suppose $\max\{k, \frac{2k}{\sigma^2} + 1\} \leq \exp\left(\sqrt{C \log p}\right)$, 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 $n \geq (1+\epsilon)\, n_{\text{info}}$, then w.h.p. as k increases*

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

On the other hand if $\frac{1}{C} k \log k \leq n \leq (1-\epsilon)\, n_{\text{info}}$, then w.h.p. as k increases

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

The proof of Theorem 3.2.3 is found in Section 3.5. The theorem above confirms the "all-or-nothing" type behavior of the optimization problem Φ_2, depending on how n compares with n_{info}. Recall that, according to [WWR10], n_{info} is an information theoretic lower bound for recovering β^* from X and Y *precisely*, and also for $n < n_{\text{info}}$ it does not rule out the possibility of recovering at least a fraction of bits of β^*. Our theorem however shows firstly that n_{info} is exactly the infortmation theoretic threshold for exact recovery and also that if $n < n_{\text{info}}$ the optimization problem Φ_2 fails to recover asymptotically any of the bits of β^*. We note also that the value of n_{info} is naturally larger than the corresponding threshold when $k = 1$, namely $2 \log p / \log(1 + 2\sigma^{-2})$, which is asymptotically $\sigma^2 \log p = n_{\text{inf},1}$. Interestingly, however this value for n, which has appeared also, as explained in the Introduction as a 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 $n > n_{\text{info}}$, even though efficient algorithms such as compressive sensing and LASSO algorithms are only known to work when $n \geq (2k + \sigma^2) \log p$. This suggests that the region $n \in [n_{\text{info}}, (2k + \sigma^2) \log p]$ might correspond to solvable but algorithmically hard regime for the problem of finding β^*.

We turn to the study of the "limiting curve" $\Gamma(\zeta)$. Note that we refer to the curve $\Gamma = \Gamma(\zeta)$ as the limiting curve because (1) from Theorem 3.2.1 for $\zeta := \ell/k$ it provides a deterministic lower bound for the value of $\phi_2(\ell)$ (up to a multiplicative constant) but also an upper bound when $\ell = 0$ and $\ell = k$ (up to a multiplicative constant) and (2) from the example of Theorem 3.2.3 it appears that structural properties of the curve Γ, such as the behavior of its maximum argument, accurately suggest a similar behavior for the actual values of $\phi_2(\ell)$,

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 $n_{\text{inf},1} = \sigma^2 \log p$, the threshold $n_{\text{info}} = \frac{2k}{\log\left(\frac{2k}{\sigma^2}+1\right)} \log p$, and finally the threshold $n_{\text{alg}} = (2k + \sigma^2) \log p$. For the illustration of different cases outlined in the proposition above see Figure 6-2.

Proposition 3.2.4. *The function Γ satisfies the following properties.*

1. *When $n \leq \sigma^2 \log p$, Γ is a strictly decreasing function of ζ. (Figure 3-1(a)),*

2. *When $\sigma^2 \log p < n < n_{\text{info}}$, Γ is not monotonic and it attains its minimum at $\zeta = 1$. (Figure 3-1(b)),*

3. *When $n = n_{\text{info}}$, Γ is not monotonic and it attains its minimum at $\zeta = 0$ and $\zeta = 1$. (Figure 3-2(a))*

4. *When $n_{\text{info}} < n < (2k + \sigma^2) \log p$, Γ is not monotonic and it attains its minimum at $\zeta = 0$. (Figure 3-2(b))*

5. *When $n > (2k + \sigma^2) \log p$, Γ is a strictly increasing function of ζ. (Figure 3-3)*

In particular, we see that both the bound $n_{\text{inf},1} = \sigma^2 \log p$, and $n_{\text{alg}} = (2k + \sigma^2) \log p$ mark the phase transition change of (lack of) monotonicity property of the limiting curve Γ. We also summarize our findings in Table 3.1. The proof of this proposition is found in Section 3.5.

To study the apparent algorithmic hardness of the problem in the regime $n \in [n_{\text{inf},1}, n_{\text{alg}}]$, as well as to see whether the picture suggested by the curve Γ is actually accurate, we now study 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 in Φ_2, that is the β's which have objective value $\|Y - X\beta\|_2$ close to the optimal value ϕ_2, break into two separate clusters; namely those

$n < n_{\text{inf},1}$	Γ is monotonically decreasing
$n_{\text{inf},1} < n < n_{\text{info}}$	Γ is not monotonic and attains its minimum at $\zeta = 1$
$n_{\text{info}} < n < n_{\text{alg}}$	Γ is not monotonic and attains its minimum at $\zeta = 0$
$n_{\text{alg}} < n$	Γ is monotonically increasing

Table 3.1: The phase transition property of the limiting curve $\Gamma(\zeta)$

which have a "large" overlap with β^*, and those which are far from it, namely those which have a "small" overlap with β^*. 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.

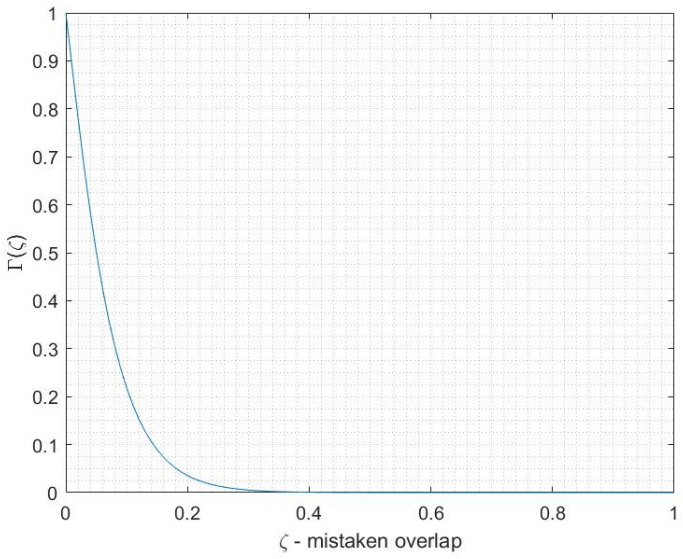
The presence of the OGP is indeed suggested by the lack of monotonicity of the limiting curve Γ when $\sigma^2 \log p < n < (2k + \sigma^2) \log p$. Indeed, in this case fixing any value γ strictly smaller than the maximum value of Γ, but larger than both $\Gamma(0)$ and $\Gamma(1)$, we see that set 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 place, in the case where n is between the information-theoretic threshold n_{info} and a constant multiple of n_{alg}. The case where n lies between $\sigma^2 \log p$ and n_{info} is discussed subsequent to the statement of the Theorem. Given any $r \geq 0$, let

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

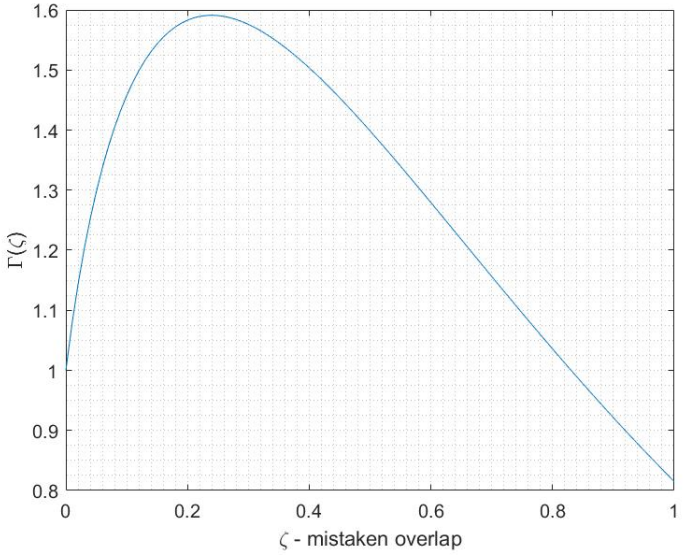
Theorem 3.2.5 (The Overlap Gap Property)**.** *Suppose the assumptions of Theorem 3.2.1 hold and for some $C > 0$, $k \log k \leq Cn$. For every sufficiently large constant D_0 there exist sequences $0 < \zeta_{1,k,n} < \zeta_{2,k,n} < 1$ satisfying*

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

as $k \to \infty$, and such that if $r_k = D_0 \max(\Gamma(0), \Gamma(1))$ and $n_{\text{info}} \leq n \leq k \log p/(3 \log D_0)$ then w.h.p. as k increases the following holds



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

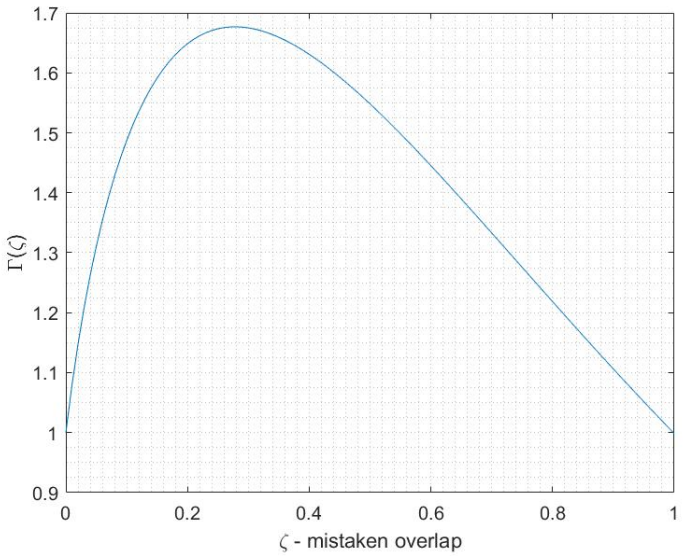


(b) The behavior of Γ for $\sigma^2 \log p < n = 120 < n_{\text{info}}$.

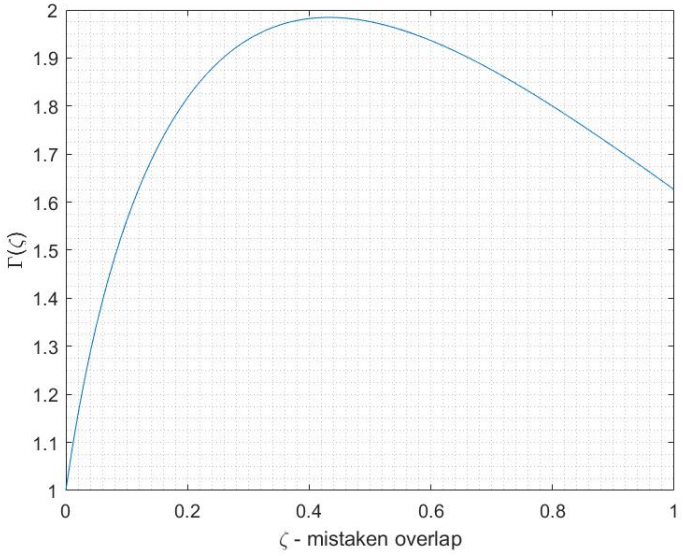
Figure 3-1: The first two different phases of the function Γ as n grows, where $n < n_{\text{info}}$. We consider the case when $p = 10^9, k = 10$ and $\sigma^2 = 1$. In this case $\lceil \sigma^2 \log p \rceil = 21, \lceil n_{\text{info}} \rceil = 137$ and $\lceil (2k + \sigma^2) \log p \rceil = 435$.

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

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



(a) The behavior of Γ for $n = 136 = n_{\text{info}}$.

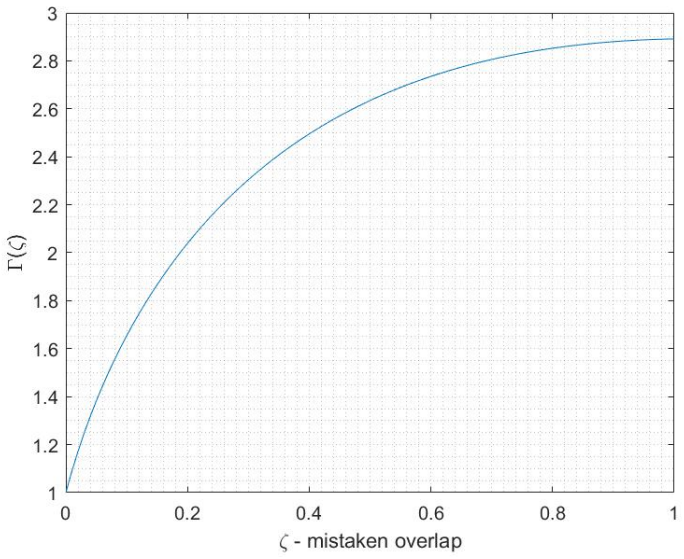


(b) The behavior of Γ for $n_{\text{info}} < n = 200 < (2k + \sigma^2) \log p$.

Figure 3-2: The middle two different phases of the function Γ as n grows where $n_{\text{info}} \leq n < n_{\text{alg}}$. We consider the case when $p = 10^9, k = 10$ and $\sigma^2 = 1$. In this case $\lceil \sigma^2 \log p \rceil = 21, \lceil n_{\text{info}} \rceil = 137$ and $\lceil (2k + \sigma^2) \log p \rceil = 435$.

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

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



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

Figure 3-3: The final phase of the function Γ as n grows where $n_{\text{alg}} \leq n$. We consider the case when $p = 10^9, k = 10$ and $\sigma^2 = 1$. In this case $\lceil \sigma^2 \log p \rceil = 21, \lceil n_{\text{info}} \rceil = 137$ and $\lceil (2k+\sigma^2)\log p \rceil = 435$.

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 3.2.5 is found in Section 3.6. The property $k\,(\zeta_{2,k,n} - \zeta_{1,k,n}) \to \infty$ in the statement of the theorem implies in particular that the difference $(\zeta_{2,k,n} - \zeta_{1,k,n})$ 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. Note that for k such that $\max\{k, \frac{2k}{\sigma^2} + 1\} \leq \exp\left(\sqrt{C \log p}\right)$ for large k it holds $\frac{1}{C} k \log k < n_{\text{info}}$ and in particular the result of Theorem 3.2.5 holds for all $n \in [n_{\text{info}}, k \log p/(3 \log D_0)]$ w.h.p. since the constraint $k \log k \leq Cn$ becomes redundant.

The study of Overlap Gap Property in the case where $\sigma^2 \log p < n < n_{\text{info}}$ does not have a clear

algorithmic value, since the problem becomes information-theoretic impossible. Nevertheleess, the first moment curve is also non-monotonic in that regime suggesting that the Overlap Gap Property still holds. Under the additional stringent assumption that $\sigma^2 \to +\infty$ as $k \to +\infty$ it can be established that Overlap Gap Property indeed appears in that regime. The proof follows by almost identical arguments with the proof of Theorem 3.2.5 by setting $\zeta_{1,k,n} := \frac{e^7 D_0^2 \sigma^2}{2k}$ and $\zeta_{2,k,n} := \frac{e^7 D_0^2 \sigma^2}{k}$.

We close this Section with stating a negative result on the popular recovery scheme called LASSO. As explained in the Introduction, besides support recovery, LASSO above n_{alg} samples is known to also ℓ_2-stably recover β^* (see (3.2). Albeit it is known that below n_{alg} samples, LASSO fails to recover the support of β^* [Wai09b], whether it can ℓ_2-recover β^* or not with less than n_{alg} samples remained an open problem prior to this work. We show that, as stated in the introduction, when n/n_{alg} is sufficiently small, for a wide range of tuning parameters λ LASSO_λ, *fails to ℓ_2-stably recover* the ground truth vector β^*. Our result applies for LASSO_λ *with and without* box constraints.

Furthermore, our result works for arbitrary choice of the tuning parameter λ as long as

$$\lambda \geq \frac{\sigma}{\sqrt{k}} \exp\left(-\frac{k \log p}{5n}\right). \tag{3.9}$$

Note that this range of possible λ's *include the standard optimal choice in the literature* of the tuning parameter $\lambda = A\sigma\sqrt{\log p/n}$ for constant $A > 2\sqrt{2}$ (see Introduction for details).

We present now the result.

Theorem 3.2.6. *Suppose that $\hat{C}\sigma^2 \leq k \leq \min\{1, \sigma^2\} \exp\left(C\sqrt{\log p}\right)$ for some constants $C, \hat{C} > 0$. Then, there exists a constant $c > 0$ such that the following holds. If $n^* \leq n \leq cn_{\text{alg}}$, $\beta^* \in \mathbb{R}^p$ is an exactly k-sparse binary vector, arbitrary choice of λ satisfying (3.9) and $\hat{\beta}_{\text{LASSO},\lambda}$, $\hat{\beta}_{\text{LASSO(box)},\lambda}$ are the optimal solutions of the formulations LASSO_λ and $\text{LASSO(box)}_\lambda$ respectively, then*

$$\min\left(\|\hat{\beta}_{\text{LASSO},\lambda} - \beta^*\|_2, \|\hat{\beta}_{\text{LASSO(box)},\lambda} - \beta^*\|_2\right) \geq \exp\left(\frac{k \log p}{5n}\right)\sigma,$$

w.h.p. as $k \to +\infty$.

Note that ℓ_2 stable recovery means finding a vector β such that $\|\beta - \beta^*\|_2 \leq C'\sigma$ for some constant $C' > 0$. The above theorem establishes that in the case of an exactly k-sparse and

binary β^*, when the samples size is less than $k \log p$ both the optimal solutions of LASSO_λ and $\text{LASSO(box)}_\lambda$ for any λ satisfying (3.9) fails to ℓ_2-stable recover the ground truth vector β^* by a multiplicative factor which is exponential on the ratio $\frac{k \log p}{n}$. In particular, coupled with the result from [BRT09b] this shows that $k \log p$ is the necessary and sufficient order of samples for which LASSO can ℓ_2-stable recover β^* for some $\lambda > 0$ satisfying (3.9).

3.3 The Pure Noise Model

In this subsection 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 3.2.1. 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. $N(0, \sigma^2)$ entries. Y, X are independent. We study the optimal value ψ_2 of the following optimization problem:

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

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. Note that (Ψ_2) can be cast also as a Gaussian Closest Vector Problem which estimates how well some vector of the form $X\beta$ where β is binary and k-sparse approximates in (rescaled) ℓ_2 error an independent target Gaussian vector Y.

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

Theorem 3.3.1. *The following holds for all n, p, k, σ:*

$$\mathbb{P}\left(\psi_2 \geq e^{-3/2}\sqrt{k+\sigma^2} \exp\left(-\frac{k \log p}{n}\right)\right) \geq 1 - e^{-n}. \tag{3.10}$$

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

$$\left\{\beta \in \{0,1\}^p : \|\beta\|_0 = k, n^{-\frac{1}{2}}\|Y - X\beta\|_2 \leq D_0\sqrt{k+\sigma^2}\exp\left(-\frac{k\log p}{n}\right)\right\}$$

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 section.

The proof of Theorem 3.3.1 is the subject of this section. 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

$$\left\{\beta \in \{0,1\}^p : \|\beta\|_0 = k, \|Y - X\beta\|_\infty \leq D_0\sqrt{k+\sigma^2}\exp\left(-\frac{k\log p}{n}\right)\right\}, \tag{3.11}$$

and establishing the same result for this set.

3.3.1 The Lower Bound. Proof of (3.10) of Theorem 3.3.1

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 $||Y - X\beta||_2^2 < t_0^2(k+\sigma^2)n$.

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. $N(0,\sigma^2)$ and $X_{i,j}$ are i.i.d. $N(0,1)$, then $||Y - X\beta||_2^2$ is distributed as $(k+\sigma^2)\sum_{i=1}^n Z_i^2$ where Z_i i.i.d. standard normal Gaussian, namely $(k+\sigma^2)$ 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

$$\mathbb{P}\left(\psi_2 < e^{-\frac{3}{2}}\binom{p}{k}^{-\frac{1}{n}}\sqrt{k+\sigma^2}\right) \leq \mathbb{E}\left[Z_{t_0\sqrt{k+\sigma^2}}\right] = \binom{p}{k}\mathbb{P}\left(\sum_{i=1}^{n} Z_i^2 \leq t_0^2 n\right). \tag{3.12}$$

Using standards large deviation theory estimates (see for example [SW95]), for the sum of n chi-square distributed random variables we obtain that for $t_0 \in (0,1)$,

$$\mathbb{P}\left(\sum_{i=1}^{n} Z_i^2 \leq n t_0^2\right) \leq \exp\left(nf\left(t_0\right)\right) \tag{3.13}$$

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

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

which implies

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

Hence using the above inequality, (3.13) and (3.12) 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 (3.10) is complete.

□

We now turn to proving the upper bound part of Theorem 3.3.1. We begin by establishing several preliminary results.

3.3.2 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

$$k^4 \leq p. \tag{3.14}$$

We establish the following two auxiliary lemmas.

Lemma 3.3.2. *If $m_1, m_2 \in \mathbb{N}$ with $m_1 \geq 4$ and $m_2 \leq \sqrt{m_1}$ then*

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

Proof. We have,

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

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 3.3.3. *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 $\log(1+x) \leq x$ implies $\log\left(\frac{1+\rho}{1-\rho}\right) \leq \frac{2\rho}{1-\rho}$. Hence,

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

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 $g'(\rho) \leq g'(0) = 0$ which implies that g is also decreasing. In particular for all $\rho \in [0, 1)$, $g(\rho) \leq g(0) = 0$. □

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

$$p_{t,y} := \mathbb{P}(|Z - y| \leq t). \tag{3.15}$$

Observe that

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

leading to

$$\log p_{t,y} \geq \log t - \frac{t^2}{2} - \frac{y^2}{2} + (1/2)\log(2/\pi). \tag{3.16}$$

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

$$q_{t,y,\rho} := \mathbb{P}(|Z_1 - y| \leq t, |Z_2 - y| \leq t), \tag{3.17}$$

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

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

Proof. We have

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

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

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

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

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

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

which is exactly:

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

This completes the proof of Lemma 3.3.4. □

3.3.3 Roadmap of the Upper Bound's proof

Recall, that our goal is to establish the required bound on the cardinality of the set (3.11) instead. Thus for every $s > 0$ we consider the counting random variable of interest,

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

Our goal is to establish that under our assumptions for sufficiently large constant $D_0 > 0$ and $s = D_0\sqrt{k+\sigma^2}\exp\left(-\frac{k\log p}{n}\right)$ it holds

$$Z_{s,\infty} \geq D_0^{\frac{n}{3}} \tag{3.18}$$

w.h.p. as $k \to +\infty$.

To establish this we use a conditional second moment method where the conditioning is happening on the "target" vector Y. We first show that the conditional first moment satisfies a similar property to (3.18); it holds

$$\mathbb{E}[Z_{s,\infty}|Y] \geq D_0^{\frac{n}{4}} \tag{3.19}$$

w.h.p. as $k \to +\infty$ (Lemma 3.3.8). This step follows from standard algebraic manipulations and an appropriate use of the Law of Large Numbers.

To establish (3.18) from (3.19) we study the conditional second moment $\mathbb{E}\left[Z_{s,\infty}^2|Y\right]$ as well and specifically the ratio the squared first moment,

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

where we have used for convenience $s = t\sqrt{k}$ for some t which throughout the proof of order $O(1)$. The second moment analysis is done in two parts. The first part is an observation; if $\Upsilon(Y)$ converges to 1 in expectation, then (3.19) implies (3.18). The proof of this part is based on the fact that for any probability measure and any positive random variable R using Chebyshev's inequality,

$$\mathbb{P}\left(R < \frac{\mathbb{E}[R]}{2}\right) \leq \mathbb{P}\left(|R - \mathbb{E}[R]\,| > \frac{\mathbb{E}[R]}{2}\right) \leq \frac{\mathbb{E}[R^2]}{\mathbb{E}[R]^2} - 1. \tag{3.20}$$

We then consider the conditional probability measure $\mathbb{P}$ on the random variable Y for our setting and apply the above inequality for $R = Z_{t\sqrt{k},\infty}$ to derive,

$$\mathbb{P}\left(Z_{t\sqrt{k},\infty} \geq \mathbb{E}\big[Z_{t\sqrt{k},\infty}|Y\big]\,|Y\right) \leq \Upsilon(Y) - 1 \tag{3.21}$$

and therefore

$$\mathbb{P}\left(Z_{t\sqrt{k},\infty} \geq \mathbb{E}\big[Z_{t\sqrt{k},\infty}|Y\big]\right) \leq \mathbb{E}_Y\{\Upsilon(Y) - 1\}. \tag{3.22}$$

The first part follows immediately from (3.22).

Unfortunately we cannot establish that $\Upsilon = \Upsilon(Y)$ converges to 1 in expectation due to a *lottery effect*; it turns out that Υ can take arbitrary large values but with negligible probability which make the expected value of Υ to explode. The second part is to show that $\min\{\Upsilon, 2\}$, the truncated version of Υ, indeed converges to 1 in expectation, as $k \to +\infty$. The exact statement of this part can be found in Proposition 3.3.5. Note that the argument with the Chebyshev's inequality described above can be easily adapted to work for the truncated version of Υ simply because the probability on the right hand side of (3.21) is upper bounded by 1 allowing to improve (3.22) to

$$\mathbb{P}\left(Z_{t\sqrt{k},\infty} \geq \mathbb{E}\big[Z_{t\sqrt{k},\infty}|Y\big]\right) \leq \mathbb{E}_Y\{\min\{\Upsilon - 1, 1\}\} = \mathbb{E}_Y\{\min\{\Upsilon, 2\} - 1\}. \tag{3.23}$$

Establishing Proposition 3.3.5 comprises the bulk of the proof and requires the use various concentration of measure inequalities and properties of the (uni-variate and bi-variate) Gaussian density function.

3.3.4 Conditional second moment bounds

We start this subsection with obtaining estimates on $\mathbb{E}\big[Z_{t\sqrt{k},\infty}|Y\big]$ and $\mathbb{E}\Big[Z^2_{t\sqrt{k},\infty}|Y\Big]$ for $t = O(1)$.

A direct calculation gives

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

where V is a standard normal random variable and $p_{t,y}$ was defined in (3.15). Similarly,

$$\mathbb{E}\Big[Z^2_{t\sqrt{k},\infty}|Y\Big] = \sum_{\ell=0}^{k}\binom{p}{k-\ell, k-\ell, \ell, p-2k+\ell}\prod_{i=1}^{n}\mathbb{P}\left(|Y_i - V_1^\ell| < t\sqrt{k}, |Y_i - V_2^\ell| < t\sqrt{k}\right),$$

where V_1^ℓ, V_2^ℓ are each $N(0,k)$ random variables with covariance l. In terms of $q_{t,y,\rho}$ defined in (3.17) we have for every l,

$$\mathbb{P}\left(|Y_i - V_1^\ell| < t\sqrt{k}, |Y_i - V_2^\ell| < t\sqrt{k}\right) = q_{t,\frac{Y_i}{\sqrt{k}},\frac{\ell}{k}}.$$

Hence,

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

We obtain

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

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 $\ell = 1, .., k$

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

By (3.14) we have $k^4 \le p$ implying $k \le \sqrt{p}$ and applying Lemma 3.3.2 we have

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

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

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

Hence we have

$$\Upsilon \le 1 + 4\sum_{\ell=1}^{k}\left(\frac{p}{k^2}\right)^{-\ell}\prod_{i=1}^{n}\frac{q_{t,\frac{Y_i}{\sqrt{k}},\frac{\ell}{k}}}{p^2_{t,\frac{Y_i}{\sqrt{k}}}}. \tag{3.24}$$

Our key result regarding the conditional second moment estimate and its ratio to the square of the conditional first moment estimate is the following proposition.

Proposition 3.3.5. *Suppose $k\log k \le Cn$ 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 $n \le \frac{k\log\left(\frac{p}{k^2}\right)}{2\log D}$ and $t = D\sqrt{1+\sigma^2}\left(\frac{p}{k^2}\right)^{-\frac{k}{n}}$ we have*

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

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

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

Observe that if $\Upsilon \ge 1 + \zeta^n$, then (3.24) implies that at least one of the summands of

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

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

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

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

$$\mathbb{E}_Y\left(\min\{1, \Upsilon - 1\}\right) \leq \zeta^n + \mathbb{P}\left(\min\{1, \Upsilon - 1\} \geq \zeta^n\right) \leq \zeta^n + \sum_{\rho = \frac{1}{k}, \frac{2}{k}, .., 1} \mathbb{P}\left(\Upsilon_\rho\right), \tag{3.25}$$

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 $\mathbb{P}\left(\Upsilon_\rho\right)$ 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 3.3.6. *For all $\rho \in (\rho_*, 1]$ and $\zeta \in (0, 1)$.*

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

Proof. Since $\rho > \rho_*$ then

$$-\left(1 - \rho\right) \frac{k \log\left(\frac{p}{k^2}\right)}{n} \geq -\frac{1}{3} \log D. \tag{3.26}$$

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

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

Applying (3.16) we obtain

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

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 (3.26)

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

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

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

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

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

as claimed. □

Lemma 3.3.7. *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 3.3.4 we have

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

Let

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

Applying Lemma 3.3.3 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 $f\left(\rho_*\right)$, and therefore

$$\begin{aligned}
\mathbb{P}\left(\Upsilon_\rho\right) &\leq \mathbb{P}\left(\sum_{i=1}^{n}\frac{Y_i^2}{kn} \geq \min\{f\left(\frac{1}{k}\right), f\left(\rho_*\right)\}\right) \\
&\leq \mathbb{P}\left(\sum_{i=1}^{n}\frac{Y_i^2}{kn} \geq f\left(\frac{1}{k}\right)\right) + \mathbb{P}\left(\sum_{i=1}^{n}\frac{Y_i^2}{kn} \geq f\left(\rho_*\right)\right). \qquad (3.27)
\end{aligned}$$

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 $\mathbb{E}[\exp\left(\theta Y_i^2/k\right)] = \frac{1}{\sqrt{1-2(\sigma^2/k)\theta}} < \infty$ 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

$$\mathbb{P}\left(\Upsilon_\rho\right) \leq 2^{\frac{n}{2}}\exp(-nf(1/k)/12) + 2^{\frac{n}{2}}\exp(-nf(\rho^*)/12). \tag{3.28}$$

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

$$2\log(4k)/n \leq 2C\log(4k/k). \tag{3.29}$$

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 $n \leq k\log(p/k^2)/(2\log D)$, the expression above is at least

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

for sufficiently large D. Combining with (3.29) we obtain that for sufficiently large D

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

Combining two bounds we obtain

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

□

We now return to the proof of Proposition 3.3.5. Combining the results of Lemma 3.3.6 and Lemma 3.3.7, and assuming $k \geq 6 \cdot 12 = 72$, we obtain that

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

for all $\rho \in [1/k, 1]$ and $\zeta \in (0,1)$. Recalling (3.25) we obtain

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

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 $n \geq (1/C)k\log k$, we obtain a bound of the form $1/k^c$ for some constant $c > 0$ as claimed. This completes the proof of Proposition 3.3.5.

□

3.3.5 The Upper Bound

Proof of Theorem 3.3.1. By an assumption of the theorem, we have $k^4 \leq p$. Thus

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

Then

$$n \leq \frac{k\log p}{2\log D_0} \leq \frac{k\log(p/k^2)}{\log D_0} = \frac{k\log(p/k^2)}{2\log D_0^{\frac{1}{2}}}. \tag{3.30}$$

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

$$\Big\{\beta \in \{0,1\}^p : \|\beta\|_0 = k, \|Y - X\beta\|_\infty \leq D_0\sqrt{k}\sqrt{1+\sigma^2/k}\exp\left(-\frac{k\log p}{n}\right)\Big\},$$

Recall that $k \leq \sigma^2 \leq 3k$. 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 $k\log k \leq Cn$, then for sufficiently large D_0,

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

and thus it suffices to obtain the claimed bound on $Z_{t_1\sqrt{k}}$. We note that by our bound (3.30)

$$\tau \leq D_0^{\frac{1}{2}}\sqrt{1+\sigma^2/k}/D_0 \leq 2/D_0^{\frac{1}{2}} \leq 1, \tag{3.31}$$

provided D_0 is sufficiently large. Let $D = D_0^{\frac{1}{2}}$. Then, by the definition of τ and by (3.30) the assumptions of Proposition 3.3.5 are satisfied for this choice of D and $t = \tau$.

Lemma 3.3.8. *The following bound holds with high probability with respect to Y as k increases*

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

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

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

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

$$\log\mathbb{E}\big[Z_{\tau\sqrt{k},\infty}|Y\big] = \log\binom{p}{k} + \sum_{i=1}^n \log p_{\tau,\frac{Y_i}{\sqrt{k}}}. \tag{3.32}$$

Applying (3.16), we have

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

Using

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

and $\tau \leq 1$, we obtain

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

Since by (3.14) we have $k \leq \sqrt{p}$, applying Lemma 3.3.2 we have $\frac{1}{n}\log\binom{p}{k} - \frac{k}{n}\log\left(\frac{p}{k^2}\right) \geq 0$. By Law of Large Numbers and since Y_i is distributed as $N(0,\sigma^2)$ with $k \leq \sigma^2 \leq 3k$, we have $n^{-1}\sum_{i=1}^{n}\frac{Y_i^2}{2k}$ converges to $\sigma^2/(2k) \leq 3/2$ as k and therefore n increases. Assuming D is sufficiently large we obtain that w.h.p. as k increases,

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

This concludes the proof of the lemma. □

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

$$Z_{\tau\sqrt{k},\infty} \geq \frac{1}{2}\mathbb{E}\big[Z_{\tau\sqrt{k},\infty}|Y\big]. \tag{3.33}$$

We have

$$\mathbb{P}\left(Z_{\tau\sqrt{k},\infty} < \frac{1}{2}\mathbb{E}\big[Z_{\tau\sqrt{k},\infty}|Y\big]\right) \leq \mathbb{P}\left(|Z_{\tau\sqrt{k},\infty} - \mathbb{E}\big[Z_{\tau\sqrt{k},\infty}|Y\big]| \geq \frac{1}{2}\mathbb{E}\big[Z_{\tau\sqrt{k},\infty}|Y\big]\right), \tag{3.34}$$

and applying Chebyshev's inequality we obtain,

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

Hence, taking expectation over Y we obtain,

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

We conclude

$$\mathbb{P}\left(Z_{\tau\sqrt{k},\infty} < \frac{1}{2}\mathbb{E}\big[Z_{\tau\sqrt{k},\infty}|Y\big]\right) \leq 4\mathbb{E}_Y\left[\min\left[\frac{\mathbb{E}\Big[Z^2_{\tau\sqrt{k},\infty}|Y\Big]}{\mathbb{E}\big[Z_{\tau\sqrt{k},\infty}|Y\big]^2} - 1, 1\right]\right]. \tag{3.35}$$

Applying Proposition 3.3.5 the assumptions of which have been verified as discussed above, we obtain

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

for some $c > 0$. This establishes the claim (3.33). Combining with Lemma 3.3.8, we conclude that w.h.p. as k increases

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

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

3.4 Proof of Theorem 3.2.1

In this section we prove Theorem 3.2.1. 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 3.2.1 we will also consider the following restriction of Φ_2. For any $S \subseteq$

Support (β^*) consider the optimization problem $(\Phi_2(S))$:

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

and set $\phi_2(S)$ its optimal value. Notice that for a binary k-sparse β with Support $(\beta) \cap$ 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 Support(β^*)
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 Support(β^*). Notice that $||\beta_1||_0 = k - |S|$.

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

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

We claim that the above problem is satisfying all the assumptions of Theorem 3.3.1 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 (3.10) of Theorem 3.3.1, 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 3.3.1 we conclude the optimal value $\phi_2\,(S)$ satisfies

$$\begin{aligned}&\mathbb{P}\left(\phi_2\,(S) \geq e^{-\frac{3}{2}}\sqrt{2\,(k-|S|)+\sigma^2}\exp\left(-\frac{(k-|S|)\log\left((p-k)\right)}{n}\right)\right)\\ &\quad\geq 1-\exp(-n).\end{aligned} \tag{3.36}$$

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 3.4.1. *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*

$$\left\{\beta\in\{0,1\}^p : \|\beta\|_0 = k, n^{-\frac{1}{2}}\|Y'-X'\beta\|_2 \leq D_0\sqrt{2k+\sigma^2}\exp\left(-\frac{k\log(p-k)}{n}\right)\right\}$$

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

Proof of Theorem 3.2.1. Applying the union bound and (3.36) we obtain

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

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 3.4.1. 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. □

3.5 The optimization problem Φ_2

In this section we give proofs of Proposition 3.2.4 and Theorem 3.2.3.

Proof of Proposition 3.2.4. 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 $\zeta^* = \frac{1}{2k \log p}(n - \sigma^2 \log p)$. 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 $n \in (\sigma^2 \log p, (2k + \sigma^2) \log p)$, respectively. The minimum value achieved by f, and its dependence on n_{info} was already established earlier. □

Proof of Theorem 3.2.3. 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: $n > (1+\epsilon)\, n_{\text{info}}$. 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 (3.8). From Theorem 3.2.1 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.:

$$e^{-\frac{3}{2}} \Gamma\left(\frac{\Lambda_p}{k}\right) \leq \phi_2(0). \tag{3.37}$$

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,

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

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

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

Combining this with (3.37) 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 $n > (1+\epsilon)\, n_{\text{info}}$, we obtain,

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

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

$$\frac{\Lambda_p}{k} \leq \frac{\sigma^2}{2k}\left(4e^3\left(\frac{2k}{\sigma^2}+1\right)^{\frac{1}{(1+\epsilon)}} - 1\right). \tag{3.38}$$

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 (3.38) 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 (3.38), $\frac{\Lambda_p}{k} \to 0$ w.h.p. as k increases, and the proof is complete in that case.

Case 2: $\frac{1}{C}k\log k < n < (1-\epsilon)\, n_{\text{info}}$. First we check that this regime for n is well-defined. Indeed the assumption $\max\{k, \frac{2k}{\sigma^2}+1\} \leq \exp\left(\sqrt{C\log p}\right)$ implies that it holds

$$n_{\text{info}} = \frac{2k\log p}{\log\left(\frac{2k}{\sigma^2}+1\right)} \geq \frac{2k\log p}{\sqrt{C\log p}} \geq \frac{2}{C}k\log k > \frac{1}{C}k\log k. \tag{3.39}$$

Now we need to show that w.h.p. as k increases

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

By the definition of Λ_p, $\phi_2(\Lambda_p) \leq \phi_2(1)$. Again applying Theorem 3.2.1 we have that w.h.p. as k increases it holds $\phi_2(\Lambda_p) \geq e^{-\frac{3}{2}}\Gamma\left(\frac{\Lambda_p}{k}\right)$. Combining the above two inequalities we obtain that w.h.p.,

$$e^{-\frac{3}{2}}\Gamma\left(\frac{\Lambda_p}{k}\right) \leq \phi_2(1). \tag{3.40}$$

Now we apply the second part of Theorem 3.2.1. Given any D_0 from part (b) of Theorem 3.2.1 and since $k/\sigma \to \infty$, we have that $\frac{1}{C}k\log k \leq n \leq (1-\epsilon)n_{\text{info}}$ furthermore then satisfies $\frac{1}{C}k\log k \leq n \leq k\log p/(3\log D_0)$ for all sufficiently large k. We obtain that w.h.p. as k increases

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

Using this in (3.40) and letting $c = 1/(e^{\frac{3}{2}}D_0)$ we obtain

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

namely,

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

and therefore

$$c^2\left(\frac{2\Lambda_p+\sigma^2}{2k+\sigma^2}\right) = c^2\left(\frac{\frac{2\Lambda_p}{\sigma^2}+1}{\frac{2k}{\sigma^2}+1}\right) \leq e^{\frac{2(\Lambda_p-k)\log p}{n}}. \tag{3.41}$$

Now using $n \leq (1-\epsilon)n_{\text{info}}$ and $\Lambda_p - k \leq 0$, we obtain

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

Combining the above with (3.41) we obtain that w.h.p.,

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

or w.h.p.,

$$c^2\left(\frac{2\Lambda_p}{\sigma^2}+1\right)\leq\left(\frac{2k}{\sigma^2}+1\right)^{-\frac{\epsilon}{1-\epsilon}+\frac{\Lambda_p}{k(1-\epsilon)}}. \tag{3.42}$$

from which we obtain a simpler bound

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

namely

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

or

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

Since by the assumption of the theorem we have $k/\sigma^2\to\infty$, we obtain that $\frac{\Lambda_p}{k}\geq\epsilon/2$ w.h.p. as $k\to\infty$. Now we reapply this bound for (3.42) and obtain that w.h.p.

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

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

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

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. □

3.6 The Overlap Gap Property

In this section we prove Theorem 3.2.5. We begin by establishing a certain property regarding the the limiting curve function Γ.

Lemma 3.6.1. *Under the assumption of Theorem 3.2.5, 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

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

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 3.2.4 and thus recall the definition of n_{info}.

Assume now $n_{\text{info}} \leq n < \frac{k\log p}{3\log D_0}$. We choose $\zeta_{1,k,n} = \frac{1}{5}$ and $\zeta_{2,k,n} = \frac{1}{4}$. Clearly $k\left(\zeta_{2,k,n} - \zeta_{1,k,n}\right) \to +\infty$. Since $n \geq n_{\text{info}}$ we know that $\Gamma(0) < \Gamma(1)$ and therefore it suffices to show

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

Using the log-concavitiy of Γ and squaring both side it suffices to establish

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

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

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

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

Now we return to the proof of Theorem 3.2.5.

Proof of Theorem 3.2.5. Choose $0 < \zeta'_{1,k,n} < \zeta'_{2,k,n} < 1$ from Lemma 3.6.1 and we set $r_k = D_0 \max\left(\Gamma(0), \Gamma(1)\right)$. 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 3.2.1 we obtain $\beta^* \in S_{r_k}$ 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 3.2.1.

We now establish part (a). Assume there exists a $\beta \in S_{r_k}$ 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

$$\phi_2 \left(k \left(1 - \zeta \right) \right) \leq r_k. \tag{3.43}$$

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 3.6.1 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 (3.43) contradicts the first part of Theorem 3.2.1. □

3.7 Proof of Theorem 3.2.6

3.7.1 Auxilary Lemmata

Lemma 3.7.1. *Fix any $C_1 > 0$. Any vector β that satisfies $\|\beta\|_1 \leq k - C_1 \sigma \sqrt{k}$ also satisfies $\|\beta - \beta^*\|_2 \geq C_1 \sigma$.*

Proof. Assume β satisfies $\|\beta - \beta^*\|_2 \leq C_1 \sigma$. We let S denote the support of β^*, and let $\beta_S \in \mathbb{R}^p$ be the vector which equals to β in the coordinates that correspond to S and is zero otherwise. We have by the triangle inequality and the Cauchy Schwartz inequality,

$$k - \|\beta_S\|_1 = \|\beta^*_S\|_1 - \|\beta_S\|_1 \leq \|\beta_S - \beta^*_S\|_1 \leq \sqrt{k} \| \left(\beta - \beta^*\right)_S \|_2 \leq \sqrt{k} \|\beta - \beta^*\|_2 \leq C_1 \sigma \sqrt{k},$$

which gives $k - C_1 \sigma \sqrt{k} \leq \|\beta_S\|_1 \leq \|\beta\|_1$. □

We also need the following immediate corollary of Theorem 3.3.1.

Corollary 1. *Let $Y' \in \mathbb{R}^n$ be a vector with i.i.d. normal entries with mean zero and abritrary variance $\mathrm{Var}(Y_1)$ and $X \in \mathbb{R}^{n\times p}$ be a matrix with iid standard Gaussian entries. Then for every $C > 0$ there exists $c_0 > 0$ such that if $c < c_0$ and for some integer k' it holds $k' \log k' \leq Cn$, $k' \leq \mathrm{Var}\left(Y_1'\right) \leq 3k'$, and $n \leq ck' \log p$, then there exists an exactly k'-sparse binary β such that*

$$n^{-\frac{1}{2}} \|Y - X\beta\|_2 \leq \exp\left(\frac{1}{2c}\right) \sqrt{k' + \mathrm{Var}\left(Y_1'\right)} \exp\left(-\frac{k' \log p}{n}\right)$$

w.h.p. as $k' \to \infty$.

Finally, we establish the following Lemma.

Lemma 3.7.2. *Under the assumptions of Theorem 3.2.6 there exists universal constants $c > 0$ such that the following holds. If $n^* \leq n \leq ck \log p$ then there exists $\alpha \in [0,1]^p$ with*

(1) $n^{-\frac{1}{2}} \|Y - X\alpha\|_2 \leq \sigma$

(2) $\|\alpha\|_1 = k - 2\exp\left(\frac{k \log p}{5n}\right) \sigma\sqrt{k}$,

w.h.p. as $k \to +\infty$.

Proof. Let

$$C_1 := \exp\left(\frac{k \log p}{5n}\right). \tag{3.44}$$

Let

$$\tilde{\lambda} := 1 - 4C_1\sqrt{\frac{\sigma^2}{k}}$$

and

$$A_{C_1} = \{\tilde{\lambda}\beta^* + (1-\tilde{\lambda})\beta | \beta \in \{0,1\}^p, \|\beta\|_0 = k/2, \mathrm{Support}\left(\beta\right) \cap \mathrm{Support}\left(\beta^*\right) = \emptyset\}.$$

A_{C_1} is the set of vectors of the form $\alpha := \tilde{\lambda}\beta^* + (1-\tilde{\lambda})\beta$ where β is exactly $\frac{k}{2}$-sparse binary with support disjoint from the support of β^*. Since by our assumption $n > n^*$ or equivalently

$$\frac{k \log p}{5n} < \frac{1}{10} \log\left(1 + \frac{2k}{\sigma^2}\right)$$

we conclude that for some $C' > 0$ large enough, if $C'\sigma^2 \leq k$ then

$$4C_1\sqrt{\frac{\sigma^2}{k}} = 4\exp\left(\frac{k\log p}{5n}\right)\sqrt{\frac{\sigma^2}{k}} < 4\left(1+\frac{2k}{\sigma^2}\right)^{\frac{1}{10}}\sqrt{\frac{\sigma^2}{k}} < 1.$$

In particular $\tilde{\lambda} > 0$ and thus $\tilde{\lambda} \in [0,1]$. Therefore $A_{C_1} \subset [0,1]^p$. It is straightforward to see also that all these vectors have ℓ_1 norm equal to $k\tilde{\lambda} + k(1-\tilde{\lambda})/2 = k(\tilde{\lambda}+1)/2$. But for our choice of $\tilde{\lambda}$ we have

$$k(\tilde{\lambda}+1)/2 = k - 2C_1\sigma\sqrt{k}$$

Therefore for all $\alpha \in A_{C_1}$ it holds $\|\alpha\|_1 = k - 2C_1\sigma\sqrt{k}$ and $\alpha \in [0,1]^p$. In particular, in order to prove our claim it is enough to find $\alpha \in A_{C_1}$ with $n^{-\frac{1}{2}}\|Y - X\alpha\|_2 \leq \sigma$.

We need to show that for some $c > 0$, there exists w.h.p. a binary vector β which is exactly $k/2$ sparse, has disjoint support with β^* and also satisfies that

$$n^{-\frac{1}{2}}\|Y - X(\tilde{\lambda}\beta^* + (1-\tilde{\lambda})\beta)\|_2 \leq \sigma.$$

We notice the following equalities:

$$\begin{aligned}\|Y - X(\tilde{\lambda}\beta^* + (1-\tilde{\lambda})\beta)\|_2 &= \|X\beta^* + W - \tilde{\lambda}X\beta^* - \left(1-\tilde{\lambda}\right)X\beta\|_2 \\ &= (1-\tilde{\lambda})\|X\beta^* + \left(1-\tilde{\lambda}\right)^{-1}W - X\beta\|_2.\end{aligned}$$

Hence the condition we need to satisfy can be written equivalently as

$$n^{-\frac{1}{2}}\|X\beta^* + \left(1-\tilde{\lambda}\right)^{-1}W - X\beta\|_2 \leq \left(1-\tilde{\lambda}\right)^{-1}\sigma,$$

or equivalently

$$n^{-\frac{1}{2}}\|Y' - X\beta\|_2 \leq \frac{1}{4}\sqrt{k}\exp\left(-\frac{k\log p}{5n}\right),$$

where for the last equivalence we set $Y' := X\beta^* + (1-\lambda)^{-1}W$ and used the definition of $\tilde{\lambda}$ for the right hand side.

Now we apply Corollary 1 for Y' $X' \in \mathbb{R}^{n\times(p-k)}$, which is X after we deleted the k columns

corresponding to the support of β^*, and $k' = k/2$. We first check that the assumptions of the Theorem are satisfied. For all i, Y_i' are iid zero mean Gaussian with

$$\mathrm{Var}\,(Y_i') = k + \sigma^2\left(1-\tilde{\lambda}\right)^{-2} = k(1+\frac{1}{16}\exp\left(-\frac{2k\log p}{5n}\right)).$$

In particular for some constant $c_0 > 0$ if $n \le c_0 k \log p$ it holds

$$k' = \frac{k}{2} \le \mathrm{Var}\,(Y_i') \le 3k/2 = 3k'.$$

Finally we need $k'\log k' \le C'n$ for some $C' > 0$. For $k' = \frac{k}{2}$ it holds $k'\log k' \le k\log k$ and also as $\hat{C}\sigma^2 \le k \le \min\{1,\sigma^2\}\exp\left(C\sqrt{\log p}\right)$ it can be easily checked that for some constant $C' > 0$ it holds $k\log k \le C'\frac{2k\log p}{\log\left(\frac{2k}{\sigma^2}+1\right)} = C'n^*$. As we assume $n \ge n^*$ we get $k'\log k' \le C'n^* \le Cn$ as needed. Therefore all the conditions are satisfied.

Applying Corollary 1 we obtain that for some constant $c_1 > 0$ there exists w.h.p. an exactly $k/2$ sparse vector β with disjoint support with β^* and

$$n^{-\frac{1}{2}}\|Y' - X\beta\|_2 \le \exp\left(\frac{1}{2c_1}\right)\sqrt{k' + \mathrm{Var}\,(Y_i')}\exp\left(-\frac{k'\log(p-k)}{n}\right).$$

Plugging in the value for k' and using $\mathrm{Var}\,(Y_i') \le \frac{3}{2}k$ we conclude the w.h.p. existence of a binary $k/2$-sparse vector β with disjoint support with β^* and

$$n^{-\frac{1}{2}}\|Y' - X\beta\|_2 \le \exp\left(\frac{1}{2c_1}\right)\sqrt{2k}\exp\left(-\frac{k\log(p-k)}{2n}\right).$$

Finally we need to verify

$$\exp\left(\frac{1}{2c_1}\right)\sqrt{2k}\exp\left(-\frac{k\log(p-k)}{2n}\right) \le \frac{1}{4}\sqrt{k}\exp\left(-\frac{k\log p}{5n}\right).$$

We notice that as $k/\sqrt{p} \to 0$ as $k, p \to +\infty$, which is true since we assume $k \le \exp\left(C\sqrt{\log p}\right)$, we have

$$\exp\left(\frac{1}{2c_1}\right)\sqrt{2k}\exp\left(-\frac{k\log(p-k)}{2n}\right) \le \exp\left(\frac{1}{2c_1}\right)\sqrt{2k}\exp\left(-\frac{k\log p}{3n}\right), \text{ for large enough } k, p.$$

Hence we need to show

$$\exp\left(\frac{1}{2c_1}\right)\sqrt{2k}\exp\left(-\frac{k\log p}{3n}\right) \leq \frac{1}{4}\sqrt{k}\exp\left(-\frac{k\log p}{5n}\right).$$

or equivalently

$$\exp\left(\frac{1}{2c_1}\right)\sqrt{2} \leq \frac{1}{4}\exp\left(\frac{2k\log p}{15n}\right)$$

which is clearly satisfied if $n \leq c_3 k\log p$ for some constant $c_3 > 0$. Therefore choosing $c = \min\{c_1, c_3\}$ the proof of the claim and of the theorem is complete.

□

3.7.2 Proofs of Theorem 3.2.6

In this subsection we use the Lemmata from the previous subsections and prove the Theorem 3.2.6.

Proof of Theorem 3.2.6. Let

$$C_1 := \exp\left(\frac{k\log p}{5n}\right). \tag{3.45}$$

According the Lemma 3.7.1 it suffices to show that for C_1 given by (3.45),

$$\max\{\|\hat{\beta}_{\text{LASSO},\lambda}\|_1, \|\hat{\beta}_{\text{LASSO(box)},\lambda}\|_1 \leq k - C_1\sigma\sqrt{k}, \tag{3.46}$$

w.h.p. as $k \to +\infty$.

To show this, we notice that since $\hat{\beta}_{\text{LASSO},\lambda}$ and $\hat{\beta}_{\text{LASSO(box)},\lambda}$ are the optimal solutions to LASSO_λ and $\text{LASSO(box)}_\lambda$ respectively, they obtains objective value smaller then any other feasible solution. Note that α given in Lemma 3.7.2 is feasible for both quadratic optimization problems LASSO_λ and $\text{LASSO(box)}_\lambda$. Hence it holds almost surely,

$$\max_{v\in\{\hat{\beta}_{\text{LASSO},\lambda},\hat{\beta}_{\text{LASSO(box)},\lambda}\}}\{\frac{1}{n}\|Y - Xv\|_2^2 + \lambda_p\|v\|_1\} \leq \frac{1}{n}\|Y - X\alpha\|_2^2 + \lambda_p\|\alpha\|_1 \tag{3.47}$$

Hence we conclude that w.h.p. as $k \to +\infty$,

$$\begin{aligned}\lambda \max\{\|\hat{\beta}_{\text{LASSO},\lambda}\|_1, \|\hat{\beta}_{\text{LASSO(box)},\lambda}\|_1 &\leq \max_{v \in \{\hat{\beta}_{\text{LASSO},\lambda}, \hat{\beta}_{\text{LASSO(box)},\lambda}\}} \{\frac{1}{n}\|Y - Xv\|_2^2 + \lambda\|v\|_1\} \\ &\leq \frac{1}{n}\|Y - X\alpha\|_2^2 + \lambda\|\alpha\|_1 \text{ , using (3.47)} \\ &\leq \sigma^2 + \lambda\left(k - 2C_1\sqrt{k}\sigma\right) \text{, using Lemma 3.7.2}\end{aligned}$$

or by rearranging,

$$\lambda\left(k - C_1\sigma\sqrt{k} - \max\{\|\hat{\beta}_{\text{LASSO},\lambda}\|_1, \|\hat{\beta}_{\text{LASSO(box)},\lambda}\|_1\}\right) \geq \left(\lambda C_1\sqrt{k} - \sigma\right)\sigma. \tag{3.48}$$

By assumption on λ satisfying (3.9) we conclude from (3.45) that

$$\lambda C_1\sqrt{k} \geq \sigma.$$

Combining the last inequality we have that the right hand side of (3.48) is nonnegative, and therefore (3.48) implies that

$$k - C_1\sigma\sqrt{k} - \max\{\|\hat{\beta}_{\text{LASSO},\lambda}\|_1, \|\hat{\beta}_{\text{LASSO(box)}}\} \geq 0$$

holds w.h.p. as $k \to +\infty$ or equivalently (3.46) holds w.h.p. as $k \to +\infty$.

This completes the proof of the Theorem 3.2.6. □

3.8 Conclusion

In this Chapter, we study the hard regime $[n_{\text{info}}, n_{\text{alg}}]$ of the high dimensional linear regression model under Gaussian assumptions on X, W and β^* is an arbitrary fixed binary k-sparse vector. Under sufficiently low sparsity $\max\{k/\sigma^2 + 1, k\} \leq \exp\left(C\sqrt{\log p}\right)$ for some $C > 0$ and high signal-to-noise ratio $k/\sigma^2 \to +\infty$ we establish multiple results:

(1) We prove an all-or-nothing behavior for the statistical performance of the MLE of the problem. This is similar in spirit to, and in fact was a motivation for, the phase transition result established in Chapter 2. Yet the results are still different, as the result in Chapter 2

assumes a uniform prior on β^*, while the result in this Chapter applies to any fixed binary k-sparse β^*.

(2) We establish that the first moment curve $\Gamma(\zeta)$ undergoes monotonicity phase transitions exactly at the thresholds $n_{\text{inf},1}, n_{\text{info}}, n_{\text{alg}}$. This monotonicity behavior suggests an Overlap Gap Property phase transition in the high dimensional linear regression model exactly at the conjectured algorithmic threshold $n = n_{\text{alg}}$.

(3) We prove that Overlap Gap Property indeed appears in the model when $n < cn_{\text{alg}}$ for some small constant $c > 0$. This is based on a potentially new result on what we call in this Section as the Pure Noise model (Section 3.3), which could be of independent interest (see also Section 1.4 for a relevant discussion.) In the next Chapter we present a proof that Overlap Gap Property ceases to hold when $n > Cn_{\text{alg}}$ establishing rigorously the desired phase transition.

(4) We establish that the well-studied ℓ_1-constrained relaxation recovery scheme LASSO provably fails in the regime $n < cn_{\text{alg}}$ to ℓ_2-stably recover the vector β^*. With this result we provide support to the algorithmic hardness conjecture in the regime $n < n_{\text{alg}}$ not only for recovering the support of β^* but for other similar, yet not equivalent, recovery tasks of β^*.

Chapter 4

The Computational-Statistical Gap of High-Dimensional Linear Regression. The Easy Regime.

4.1 Introduction

In this Chapter we continue our study of the computational-statistical gap of the high dimensional linear regression model, which is initiated in Chapter 3. We remind the reader that we study the model described in Subsection 1.1.1, under the assumptions that $X \in \mathbb{R}^{n\times p}$ and $W \in \mathbb{R}^{n\times 1}$ are independent matrices with $X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $W_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2)$ for some $\sigma^2 > 0$, and finally β^* is an arbitrary but fixed binary k-sparse vector. In Chapter 2 it is established that $n = n_{\text{info}}$, defined in 1.2, is the exact statistical limit of the problem, while computationally efficient methods are only known to succeed when $n > n_{\text{alg}}$ where n_{alg} is defined in (1.3).

In Chapter 3, the compuational-statistical gap when $n \in [n_{\text{info}}, n_{\text{alg}}]$ is studied. It is established in Theorem 3.2.5 that the solution space of maximum likelihood estimation indeed exhibits the Overlap Gap Property (OGP), appropriately defined, when β^* is an arbitrary binary and k-sparse and $n < cn_{\text{alg}}$ for some small enough constant $c > 0$. For this reason, and drawing a correspondence with a large body of work in the literature for computational-existential gaps mentioned in Chapter 1, Theorem 3.2.5 provides evidence of algorithmic hardness for high dimensional linear regression when $n < cn_{\text{alg}}$.

On the other hand, in the literature it is conjectured that when OGP ceases to hold even simple greedy local search methods can exploit the smooth geometry and succeed (see for example the literature on the maximum independent set in Erdős-Rényi graphs [GSa],[RV14] and [GSb]). To the best of our knowledge, neither OGP has been proven to be absent, nor any simple local search algorithm is known to successfully work for high dimensional linear regression when $n > n_{\text{alg}}$.

In this Chapter we study the Overlap Gap Property for high dimensional linear regression when $n > n_{\text{alg}}$. In that regime the questions of interest are:

Does Overlap Gap Property hold when $n > n_{\text{alg}}$?

If not, is there a successful greedy local search method in that regime?

We answer both questions when $n > Cn_{\text{alg}}$ for some sufficiently large constant $C > 0$. Specifically we establish the following result.

Contribution

We establish that if $n \geq Cn_{\text{alg}}$ for some sufficiently large constant $C > 0$, then OGP indeed ceases to hold. We base this result on a direct local landscape analysos of the maximum likelihood estimation optimization problem. We show that in this algorithmically easy regime, the landscape is extremely smooth: all the local minima have identical support with the hidden vector β^*. Furthermore, we prove that for these values of n a very simple Local Search Algorithm can exploit the notably "smooth" local geometry of the solutions space and recover exactly the support of β^*. Interestingly, the termination time of the algorithm is proven to be independent of the feature size p.

One distinct attribute of the results of this Chapter is that *they generalize* much beyond the binary case for the values of β^* and the sublinear sparsity condition $k/p \to 0$, as $p \to +\infty$. We make this more precise with the following two bullet points.

(1) We show that the Local Search Algorithm (LSA) can be defined and provably work in the real-valued case for the k-sparse β^* under a constraint on its minimum value $|\beta^*|_{\min} = \min\{|\beta_i^*| : \beta_i^* \neq 0\} \geq 1$. We prove that LSA outputs a vector with the same support of β^*

which furthermore satisfies

$$\|\beta - \beta^*\|_2 \leq C_0 \sigma \tag{4.1}$$

for some constant $C_0 > 0$. The notion of recovery (4.1) is known in the literature as ℓ_2-stable recovery of the vector of β^*. (see e.g. [CRT06] [BD09] and references therein).

(2) All the results we present in this Chapter apply to any sparsity level $k \leq \frac{p}{3}$.

Finally, at a technical level, most of the results presented in this Chapter are based on the Restricted Isometry Property for the matrix X, the Hanson-Wright concentration inequality and a careful net argument, which could be of independent interest (see Section 4.3 for details).

Notation

For a matrix $A \in \mathbb{R}^{n \times n}$ we use its operator norm $\|A\| := \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}$, and its Frobenius norm $\|A\|_F := \left(\sum_{i,j} |a_{i,j}|^2\right)^{\frac{1}{2}}$. If $n, d \in \mathbb{N}$ and $A \in \mathbb{R}^{d \times p}$ by $A_i, i = 1, 2, \ldots, p$ we refer to the p columns of A. For $p \in (0, \infty), d \in \mathbb{N}$ and a vector $x \in \mathbb{R}^d$ we use its $\mathcal{L}_p$-norm, $\|x\|_p := (\sum_{i=1}^p |x_i|^p)^{\frac{1}{p}}$. For $p = \infty$ we use its infinity norm $\|x\|_\infty := \max_{i=1,\ldots,d} |x_i|$ and for $p = 0$, its 0-norm $\|x\|_0 = |\{i \in \{1, 2, \ldots, d\} | x_i \neq 0\}|$. We say that x is k-sparse if $\|x\|_0 \leq k$ and exactly k-sparse if $\|x\|_0 = k$. We also define the support of x, $\text{Support}\,(x) := \{i \in \{1, 2, \ldots, d\} | x_i \neq 0\}$. For $k \in \mathbb{Z}_{>0}$ we adopt the notation $[k] := \{1, 2, \ldots, k\}$. Finally with the real function $\log : \mathbb{R}_{>0} \to \mathbb{R}$ we refer everywhere to the natural logarithm.

Structure of the Chapter

The remained of the Chapter is structured as follows. The description of the model, assumptions and main results are found in the next section. Subsection 4.3 is devoted to the proof of the absence of the Overlap Gap Property and the success of the local search algorithm in the regime $n \geq C n_{\text{alg}}$.

4.2 Above n_{alg} samples: The Absence of OGP and the success of the Local Search Algorithm

Recall that according to (1.3), when β^* is binary-valued, $n_{\text{alg}} = (2k + \sigma^2) \log p$. We work in this Chapter under the assumption that $n \geq C n_{\text{alg}}$ for some sufficiently large $C > 0$. Furthermore, all of the results presented in this section are in the regime where the signal to noise ratio (SNR) k/σ^2 is at least a constant. In particular, this SNR assumption implies $n_{\text{alg}} = \Theta(k \log p)$. For these reasons, for simplicity and without loss of generality, from now on, in this Chapter we use the simplified notation $n_{\text{alg}} \triangleq k \log p$.

We establish the absence of the OGP in the case $n \geq C n_{\text{alg}} = Ck \log p$ for sufficiently large $C > 0$, w.h.p. For the same values of n we also propose a very simple Local Search Algorithm (LSA) for recovering β^* which provably succeeds w.h.p. In fact our results for OGP is an easy consequence of the success of LSA.

The Absence of OGP

We now state the definition of Overlap Gap Property (OGP) which generalizes the definition used in Chapter 3 where it focuses only the binary case for β^*.

Definition 4.2.1. *Fix an instance of X, W. The regression problem defined by (X, W, β^*) where a vector β^* is an exactly k-sparse vector with $|\beta^*|_{min} \geq 1$ satisfies the Overlap Gap Property (OGP) if there exists $r = r_{n,p,k,\sigma^2} > 0$ and constants $0 < \zeta_1 < \zeta_2 < 1$ such that*

(1) $\|Y - X\beta^*\|_2 < r$,

(2) There exists a k-sparse vector β with $\text{Support}(\beta) \cap \text{Support}(\beta^*) = \emptyset$ *and* $\|Y - X\beta\|_2 < r$, *and*

(3) If a k-sparse vector β satisfies $\|Y - X\beta\|_2 < r$ then either

$$|\text{Support}(\beta) \cap \text{Support}(\beta^*)| < \zeta_1 k$$

or

$$|\text{Support}(\beta) \cap \text{Support}(\beta^*)| > \zeta_2 k.$$

The OGP has a natural interpretation. It states that the k-sparse βs which achieve near optimal cost for the objective value $\|Y - X\beta\|_2$ split into two non-empty "well-separated" regions; the ones whose support is close with the support of β^* in the Hamming distance sense, and the ones whose support is far from the support of β^* in the Hamming distance sense, creating a "gap" for the vectors with supports in a "intermediate" Hamming distance.

In Chapter 3 the authors prove that under the assumption $\frac{1}{5}\sigma^2 \leq k \leq \min\{1, \sigma^2\} \exp\left(C\sqrt{\log p}\right)$ for some constant $C > 0$ if n satisfies $n_{\text{info}} < n \leq ck \log p$, for some sufficiently small constant $c > 0$, then the OGP restricted for binary vectors holds for some $r > 0$ and $\zeta_1 = \frac{1}{5}$ and $\zeta_2 = \frac{1}{4}$. Since OGP is associated with algorithmic hardness, it is naturally expected that OGP will not hold when $n \geq Ck \log p$ for some constant $C > 0$, which is the regime for n where efficient algorithms, such as LASSO, have been proven to work. We confirm this belief in the theorem below.

Theorem 4.2.2. *There exists $c, C > 0$ such that if $\sigma^2 \leq c \min\{k, \frac{\log p}{\log \log p}\}$, $n \geq Cn_{\text{alg}}$ the following holds. If the β^* is exactly k-sparse and satisfies $|\beta^*|_{\min} \geq 1$ then the regression problem (X, W, β^*) does not satisfy the OGP w.h.p. as $k \to +\infty$.*

We now give some intuition of how this result is derived. The proof is based on a lemma on the "local" behavior of the k-sparse βs with respect to the optimization problem

$$\begin{aligned} (\tilde{\Phi}_2) \quad \min \quad & \|Y - X\beta\|_2 \\ \text{s.t.} \quad & \|\beta\|_0 \leq k. \end{aligned}$$

We first give a natural definition of what a non-trivial local minimum is for $\tilde{\Phi}_2$.

Definition 4.2.3. *We define a k-sparse β to be a **non-trivial local minimum** for $\tilde{\Phi}_2$ if*

- $\text{Support}(\beta) \neq \text{Support}(\beta^*)$, *and*
- *if a k-sparse β_1 satisfies*

$$\max\{|\text{Support}(\beta) \setminus \text{Support}(\beta_1)|, |\text{Support}(\beta_1) \setminus \text{Support}(\beta)|\} \leq 1,$$

it must also satisfy

$$\|Y - X\beta_1\|_2 \geq \|Y - X\beta\|_2.$$

We continue with the observation that the presence of OGP deterministicaly implies the existence of a non-trivial local minimum for the problem $\tilde{\Phi}_2$.

Proposition 4.2.4. *Assume for some instance of X, W the regression problem (X, W, β^*) satisfies the Overlap Gap Property. Then for this instance of X, W there exists at least one non-trivial local minimum for $\tilde{\Phi}_2$.*

Proof. Assume that OGP holds for some values r, ζ_1, ζ_2. We choose β_1 the k-sparse vector β that minimizes $\|Y - X\beta\|_2$ under the condition $|\text{Support}(\beta) \cap \text{Support}(\beta^*)| \leq \zeta_1 k$. The existence of β_1 is guaranteed as the space of k-sparse vectors with $|\text{Support}(\beta) \cap \text{Support}(\beta^*)| \leq \zeta_1 k$ is closed under the Euclidean metric.

We claim this is a non-trivial local minimum. Notice that it suffices to prove that β_1 minimizes also $\|Y - X\beta\|_2$ under the more relaxed condition $|\text{Support}(\beta) \cap \text{Support}(\beta^*)| < \zeta_2 k$. Indeed then since $\zeta_1 k < \zeta_2 k$, β_1 will be the minimum over a region that contains its 2-neighborhood in the Hamming distance and as clearly the support of β_1 is not equal to the support of β^* we would be done.

Now to prove the claim consider a β with $\zeta_1 k < |\text{Support}(\beta) \cap \text{Support}(\beta^*)| < \zeta_2 k$. By the Overlap Gap Property we know that it must hold $\|Y - X\beta\|_2 > r$. Furthermore again by the Overlap Gap Property we know there is a β' with $|\text{Support}(\beta') \cap \text{Support}(\beta^*)| = 0 < \zeta_1 k$ for which it holds $\|Y - X\beta'\|_2 < r$. But by the definition of β_1 it must also hold $\|Y - X\beta_1\|_2 \leq \|Y - X\beta'\|_2 < r$ which combined with $\|Y - X\beta\|_2 > r$ implies $\|Y - X\beta_1\|_2 < \|Y - X\beta\|_2$. Since the β was arbitrary with $\zeta_1 k < |\text{Support}(\beta) \cap \text{Support}(\beta^*)| < \zeta_2 k$ the proof of the Proposition is complete. □

Now in light of the Proposition above, we know that a way to negate OGP is to prove the absence of non-trivial local minima for $\tilde{\Phi}_2$. We prove that indeed if $n \geq Ck \log p$ for some universal $C > 0$ our regression model does not have non-trivial local minima for $\tilde{\Phi}_2$ w.h.p. and in particular OGP does not hold in this regime w.h.p., as claimed. We state this as a separate result as it could be of independent interest.

Theorem 4.2.5. *There exists $c, C > 0$ such that if $\sigma^2 \leq c \min\{k, \frac{\log p}{\log \log p}\}$, $n \geq C n_{\text{alg}}$ such that the following is true. If the β^* is exactly k-sparse and satisfies $|\beta^*|_{\min} \geq 1$ then the optimization problem $(\tilde{\Phi}_2)$ has no non-trivial local minima w.h.p. as $k \to +\infty$.*

The complete proofs of both Theorem 4.2.2 and Theorem 4.2.5 are presented in Section 4.

Success of Local Search

As stated in the introduction, in parallel to many results for random constrained satisfaction problems, the disappearance of OGP suggests the existence of a very simple algorithm succeeding in recovering β^*, usually exploiting the smooth local structure. Here, we present a result that reveals a similar picture. A natural implication of the absence of non-trivial local minima property is the success w.h.p. of the following very simple local search algorithm. Start with any vector β_0 which is k-sparse and then iteratively conduct "local" minimization among all β's with support of Hamming distance at most two away from the support of our current vector.

We now state this algorithm formally. Let $e_i \in \mathbb{R}^p, i = 1, 2, \ldots, p$ be the standard basis vectors of $\mathbb{R}^p$.

Local Search Algorithm (LSA)

0. Input: A k-sparse vector β with support S.
1. For all $i \in S$ and $j \in [p]$ compute $\text{err}_i(j) = \min_q \|Y - X\beta + \beta_i X_i - qX_j\|_2$.
2. Find $(i_1, j_1) = \text{argmin}_{i \in S, j \in [p]} \text{err}_i(j)$ and $q_1 := \text{argmin}_{q \in \mathbb{R}} \|Y - X\beta + \beta_{i_1} X_{i_1} - qX_{j_1}\|_2$.
3. If $\|Y - X\beta + \beta_{i_1} X_{i_1} - q_1 X_{j_1}\|_2 < \|Y - X\beta\|_2$, update the vector β to $\beta - \beta_{i_1} e_{i_1} + q e_{j_1}$, the set S to the support of the new β and go to step 1. Otherwise terminate and output β.

For the performance of the algorithm we establish the following result.

Theorem 4.2.6. *There exist $c, C > 0$ so that if $\beta^* \in \mathbb{R}^p$ is an exactly k-sparse vector, $n \geq Cn_{\text{alg}}$ and $\sigma^2 \leq c|\beta^*|^2_{min} \min\{\frac{\log p}{\log\log p}, k\}$ then the algorithm LSA with an arbitrary k-sparse vector β_0 as input terminates in at most $\frac{4k\|Y - X\beta_0\|_2^2}{\sigma^2 n}$ iterations with a vector $\hat{\beta}$ such that*

(1) $\text{Support}\left(\hat{\beta}\right) = \text{Support}(\beta^*)$ *and*

(2) $\|\hat{\beta} - \beta^*\|_2 \leq \sigma$,

w.h.p. as $k \to +\infty$.

The complete proof of Theorem 4.2.6 is presented in subsection 4.3.3. Various auxiliary lemmas are established in the Subsections in between.

4.3 LSA Algorithm and the Absence of the OGP

4.3.1 Preliminaries

We introduce the notion of a super-support of a finite dimensional real vector.

Definition 4.3.1. *Let $d \in \mathbb{N}$. We call a set $\emptyset \neq S \subseteq [d]$ a **super-support** of a vector $x \in \mathbb{R}^d$ if* $\text{Support}\,(x) \subseteq S$.

We also need the definition and some basic properties of the Restricted Isometry Property (RIP).

Definition 4.3.2. *Let $n, k, p \in \mathbb{N}$ with $k \leq p$. We say that a matrix $X \in \mathbb{R}^{n \times p}$ satisfies the k-**Restricted Isometry Property (k-RIP)** with restricted isometric constant $\delta_k \in (0,1)$ if for every vector $\beta \in \mathbb{R}^p$ which is k-sparse it holds*

$$(1-\delta_k)\|\beta\|_2^2 n \leq \|X\beta\|_2^2 \leq (1+\delta_k)\|\beta\|_2^2 n.$$

A proof of the following theorem can be found in [BDDW08].

Theorem 4.3.3. *[BDDW08] Let $n, k, p \in \mathbb{N}$ with $k \leq p$. Suppose $X \in \mathbb{R}^{n \times p}$ has i.i.d. standard Gaussian entries. Then for every $\delta > 0$ there exists a constant $C = C_\delta > 0$ such that if $n \geq Ck \log p$ then X satisfies the k-RIP with restricted isometric constant $\delta_k < \delta$ w.h.p.*

We need the following properties of RIP.

Proposition 4.3.4. *Let $n, k, p \in \mathbb{N}$ with $k \leq p$. Suppose $X \in \mathbb{R}^{n \times p}$ satisfies the k-RIP with restricted isometric constant $\delta_k \in (0,1)$. Then for any $v, w \in \mathbb{R}^p$ which are k-sparse,*

(1)

$$|(Xv)^T(Xw)| \leq (1+\delta_k)\|v\|_2\|w\|_2 n \leq 2\|v\|_2\|w\|_2 n.$$

(2) If v, w have a common super-support of size k then

$$\|Xw\|_2^2 + 4\|v-w\|_2\|w\|_2 n + 2\|v-w\|_2^2 n \geq \|Xv\|_2 \geq \|Xw\|_2^2 - 4\|v-w\|_2\|w\|_2 n.$$

(3) If v, w have disjoint supports and a common super-support of size k then

$$|(Xv)^T(Xw)| \leq \delta_k \left(\|v\|_2^2 + \|w\|_2^2\right) n.$$

Proof. The first part follows from the Cauchy-Schwarz inequality and the definiton of k-RIP applied to the vectors v, w. For the second part we write $Xv = X(w + (v - w))$, and we have

$$\|Xv\|_2^2 = \|Xw\|_2^2 + 2\left(X(v-w)\right)^T (Xw) + \|X(v-w)\|_2^2.$$

Since v, w have a common super-support of size k, the vectors $v - w, w$ are k-sparse vectors. Hence from the first part we have

$$-2\|v-w\|_2\|w\|_2 n \leq |X(v-w)^T Xw| \leq 2\|v-w\|_2\|w\|_2 n$$

$$0 \leq \|X(v-w)\|_2^2 \leq 2\|v-w\|_2^2 n.$$

Applying these inequalities to the last equality, the proof follows.

For the third part since v, w are k-sparse and have a common super-support of size k the vectors $v+w$ and $v-w$ are k-sparse vectors. Hence by k-RIP and that v, w have disjoint supports we obtain

$$\|X(v+w)\|_2^2 \leq (1+\delta_k)\|v+w\|_2^2 n = (1+\delta_k)\left(\|v\|_2^2 + \|w\|_2^2\right) n$$

and similarly

$$\|X(v-w)\|_2^2 \geq (1-\delta_k)\left(\|v\|_2^2 + \|w\|_2^2\right) n.$$

Hence

$$\begin{aligned}
|(Xv)^T(Xw)| &= |\frac{1}{4}\left[\|X(v+w)\|_2^2 - \|X(v-w)\|_2^2\right]| \\
&\leq \frac{1}{4}|(1+\delta_k)\left(\|v\|_2^2 + \|w\|_2^2\right) n - (1-\delta_k)\left(\|v\|_2^2 + \|w\|_2^2\right) n| \\
&\leq \delta_k \left(\|v\|_2^2 + \|w\|_2^2\right) n,
\end{aligned}$$

as required. □

Finally, we need the so-called Hanson-Wright inequality.

Theorem 4.3.5 (Hanson-Wright inequality, [HW71])**.** *There exists a constant $d > 0$ such that the following holds. Let $n \in \mathbb{N}, A \in \mathbb{R}^{n\times n}$ and $t \geq 0$. Then for a vector $X \in \mathbb{R}^n$ with i.i.d. standard Gaussian components*

$$\mathbb{P}\left(|X^tAX - \mathbb{E}\big[X^tAX\big]| > t\right) \leq 2\exp\left[-d\min\left(\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|}\right)\right].$$

4.3.2 Study of the Local Structure of $(\tilde{\Phi}_2)$

We start by introducing the notion of an α-deviating local minimum (α-DLM).

Definition 4.3.6. *Let $n, p \in \mathbb{N}$,$\alpha \in (0,1), X \in \mathbb{R}^{n\times p}$ and $\emptyset \neq S_1, S_2, S_3 \subseteq [p]$. A triplet of vectors (a, b, c) with $a, b, c \in \mathbb{R}^p$ is called an α-**deviating local minimum (α-D.L.M.)** with respect to S_1, S_2, S_3 and to the matrix X if the following are satisfied:*

- *The sets S_1, S_2, S_3 are pairwise disjoint and the vectors a, b, c have super-supports S_1, S_2, S_3 respectively.*
- *$|S_1| = |S_2|$ and $|S_1| + |S_2| + |S_3| \leq 3k$.*
- *For all $i \in S_1$ and $j \in S_2$*

$$\|\left(Xa - a_iX_i\right) + \left(Xb - b_jX_j\right) + Xc\|_2^2 \geq \|Xa + Xb + Xc\|_2^2 - \alpha\left(\frac{\|a\|_2^2}{|S_1|} + \frac{\|b\|_2^2}{|S_2|}\right)n. \quad (4.2)$$

Remark 4.3.7. *In several cases in what follows we call a triplet (a, b, c) an α-**DLM** with respect to a matrix X without explicitly referring to their corresponding super-sets S_1, S_2, S_3 but we do always assume their existence.*

We first establish the following algebraic claim for the DLM property.

Claim 4.3.8. *Let $n, p, k \in \mathbb{N}$ with $k \leq \frac{1}{3}p$. Suppose a matrix $X \in \mathbb{R}^{n\times p}$ satisfies the $3k$-RIP for some isometric constant $\delta_{3k} \in (0,1)$ and that for some $\alpha \in (0,1)$ a triplet (a, b, c) is an α-D.L.M. with respect to X. Then*

$$\|X\left(a + b\right)\|_2^2 + 2(Xc)^T\left(X(a + b)\right) \leq \left(\alpha + 4\delta_{3k}\right)\left(\|a\|_2^2 + \|b\|_2^2\right)n.$$

Proof. Let S_1, S_2, S_3 the super-sets of the vectors a, b, c with respect to which the triplet (a, b, c) is an α-DLM. Set $m := |S_1| = |S_2|$. Based on the definition of an α-DLM by expanding the squared norm in the left hand side of (4.2) we have that $\forall i \in S_1, j \in S_2$ it holds

$$a_i^2\|X_i\|_2^2 + b_j^2\|X_j\|_2^2 + 2a_ib_jX_i^TX_j - 2\left(Xa + Xb + Xc\right)^T\left(a_iX_i + b_jX_j\right)$$

is at leasy

$$-\alpha\left(\frac{\|a\|_2^2}{m} + \frac{\|b\|_2^2}{m}\right)n.$$

Summing over all $i \in S_1, j \in S_2$ we obtain

$$\sum_{i\in S_1, j\in S_2}\left[a_i^2\|X_i\|_2^2 + b_j^2\|X_j\|_2^2 + 2a_ib_jX_i^TX_j - 2\left(Xa + Xb + Xc\right)^T\left(a_iX_i + b_jX_j\right)\right]$$

is at least

$$-m\alpha\left(\|a\|_2^2 + \|b\|_2^2\right)n$$

which equivalently gives

$$m\sum_{i\in S_1} a_i^2\|X_i\|_2^2 + m\sum_{j\in S_2} b_j^2\|X_j\|_2^2 + 2(Xa)^T(Xb) - 2m\left(Xa + Xb + Xc\right)^T\left(Xa + Xb\right)$$

is at least

$$-m\alpha\left(\|a\|_2^2 + \|b\|_2^2\right)n$$

which now after rearranging and multiplying with $-\frac{1}{m}$ implies that the quantity

$$\|X\left(a+b\right)\|_2^2 + 2(Xc)^T\left(X(a+b)\right) + 2\underbrace{\left(1 - \frac{1}{m}\right)(Xa)^T(Xb)}_{S}$$
$$+ \underbrace{\left[\|Xa\|_2^2 - \sum_{i\in S_1} a_i^2\|X_i\|_2^2\right] + \left[\|Xb\|_2^2 - \sum_{j\in S_2} b_j^2\|X_j\|_2^2\right]}_{T}$$

is at most $\alpha\left(\|a\|_2^2 + \|b\|_2^2\right)n$. To finish the proof it suffices to establish that S, T are both bounded from below by $-2\delta_{3k}\left(\|a\|_2^2 + \|b\|_2^2\right)n$. We start with bounding S. The vectors a, b have disjoint

supports which sizes sum up to at most $3k$. In particular, the union of their supports is a common super-support of them of size at most $3k$. Hence we can apply part (3) of Proposition 4.3.4 to get

$$S = 2\left(1-\frac{1}{m}\right)(Xa)^T(Xb) \geq -2\delta_{3k}\left(1-\frac{1}{m}\right)\left(\|a\|_2^2+\|b\|_2^2\right)n \geq -2\delta_{3k}\left(\|a\|_2^2+\|b\|_2^2\right)n.$$

For T it suffices to prove that $\left[\|Xa\|_2^2 - \sum_{i\in S_1} a_i^2\|X_i\|_2^2\right] \geq -2\delta_{3k}\|a\|_2^2 n$ and since the same will hold for b by symmetry, by summing the inequalities we will be done. Note that as a and all the standard basis vectors are $3k$-sparse vectors by $3k$-RIP for X we have $\|Xa\|_2^2 \geq (1-\delta_{3k})\|a\|_2^2 n$ and secondly $\|X_i\|_2^2 \leq (1+\delta_{3k})n$, for all $i \in [p]$. Combining we obtain

$$\left[\|Xa\|_2^2 - \sum_{i\in S_1} a_i^2\|X_i\|_2^2\right] \geq \left[(1-\delta_{3k})\|a\|_2^2 n - (1+\delta_{3k})\sum_{i\in S_1} a_i^2 n\right] = -2\delta_{3k}\|a\|_2^2 n.$$

The proof is complete. □

We now establish two properties for D.L.M. triplets.

Proposition 4.3.9. *Let $n, p, k \in \mathbb{N}$ with $k \leq \frac{1}{3}p$. Suppose that $X \in \mathbb{R}^{n\times p}$ satisfies the $3k$-RIP with restricted isometric constant $\delta_{3k} < \frac{1}{12}$. Then there is no $\frac{1}{4}$-D.L.M. triplet (a,b,c) with respect to the matrix X with $\|a\|_2^2 + \|b\|_2^2 \geq \frac{1}{4}\|c\|_2^2$.*

Proof. By Lemma 4.3.8 any $\frac{1}{4}$-D.L.M. triplet satisfies

$$\|X(a+b)\|_2^2 + 2(Xc)^T(X(a+b)) \leq \left(\frac{1}{4}+4\delta_{3k}\right)\left(\|a\|_2^2+\|b\|_2^2\right)n.$$

But using the $3k$-R.I.P. for X and that a, b, c have disjoint supports with sizes summing up to at most $3k$ we get the following two inequalities from Proposition (4.3.4);

- $\|X(a+b)\|_2^2 \geq (1-\delta_{3k})\left(\|a+b\|_2^2\right)n = (1-\delta_{3k})\left(\|a\|_2^2+\|b\|_2^2\right)n$, since $a+b$ is $3k$-sparse and a, b have disjoint supports.
- $(Xc)^T(X(a+b)) \geq -\delta_{3k}\left(\|c\|_2^2+\|a\|_2^2+\|b\|_2^2\right)$, from Proposition 4.3.4 (3).

We obtain

$$(1-\delta_{3k})\left(\|a\|_2^2+\|b\|_2^2\right) - 2\delta_{3k}\left(\|c\|_2^2+\|a\|_2^2+\|b\|_2^2\right)$$

is at most $\left(\frac{1}{4}+4\delta_{3k}\right)\left(\|a\|_2^2+\|b\|_2^2\right)$. But now, this inequality can be equivalently written as

$$\left(\frac{3}{4}-7\delta_{3k}\right)\left(\|a\|_2^2+\|b\|_2^2\right)\leq\delta_{3k}\|c\|_2^2. \tag{4.3}$$

Now we use that for $\delta_{3k}<\frac{1}{12}$ it holds $\frac{3}{4}-7\delta_{3k}>2\delta_{3k}$. Using this in (4.3) we conclude that $\sqrt{\|a\|_2^2+\|b\|_2^2}<\frac{1}{2}\|c\|_2$ and the proof of the proposition is complete.

□

The second property we want is the following.

Proposition 4.3.10. *Let $n,p,k\in\mathbb{N}$ with $k\leq\frac{1}{3}p$. Suppose $X\in\mathbb{R}^{n\times p}$ has i.i.d. $N(0,1)$ entries. There exists constants $c_1,C_1>0$ such that if $n\geq C_1k\log p$ then w.h.p. there is no $\frac{1}{4}$-D.L.M. triplet (a,b,c) with respect to the some sets $\emptyset\neq S_1,S_2,S_3\subset[p]$ and the matrix X such that the following conditions are satisfied.*

(1) $|a|_{\min}:=\min\{|a_i|:a_i\neq0\}\geq1$.

(2) $S_1\cup S_3=[k]\cup\{p\}$, $p\in S_3$ *and* $S_1=\text{Support}(a)$.

(3) $\|a\|_2^2+\|b\|_2^2+\|c\|_2^2\leq c_1\min\{\frac{\log p}{\log(\log p)},k\}$.

Proof. We first choose $C_1>0$ large enough based on Theorem 4.3.3 so that $n\geq C_1k\log p$ implies that X satisfies the $3k$-RIP with $\delta_{3k}<\frac{1}{16}$ w.h.p. In particular all the probability calculations below will be conditioned on this high-probability event.

We start with a lemma for bounding the probability that a specific triplet (a,b,c) is an $\frac{1}{2}$-D.L.M. triplet with respect to X.

Lemma 4.3.11. *There exists a $c_0>0$ such that for any fixed triplet (a,b,c) with $a\neq0$,*

$$\mathbb{P}\left((a,b,c)\text{ is a }\frac{1}{2}\text{-D.L.M. triplet}\right)\leq2\exp\left(-c_0n\min\{1,\frac{\|a\|_2^2+\|b\|_2^2}{\|c\|_2^2}\}\right),$$

where for the case $c=0$ we abuse the notation by defining $\frac{1}{0}:=+\infty$.

Proof. We prove only the case $c\neq0$. The case $c=0$ is similar. Assume a fixed triplet (a,b,c) is an $\frac{1}{2}$-DLM. Using Claim 4.3.8 we have that it holds

$$\|X\left(a+b\right)\|_2^2+2(Xc)^T\left(X(a+b)\right)\leq\left(\frac{1}{2}+4\delta_{3k}\right)\left(\|a\|_2^2+\|b\|_2^2\right)n.$$

We set $X_1 = X\left(\frac{a+b}{\sqrt{\|a\|_2^2+\|b\|_2^2}}\right)$ and $W_1 = X\left(\frac{c}{\|c\|_2}\right)$ and notice that X_1, W_1 have independent $N(0,1)$ entries because a, b, c have disjoint supports. The last inequality can be expressed with respect to X_1, W_1 as,

$$\|X_1\|_2^2 + 2\frac{\|c\|_2}{\sqrt{\|a\|_2^2+\|b\|_2^2}} W_1 X_1 \leq \left(\frac{1}{2} + 4\delta_{3k}\right) n.$$

Now we introduce matrix notation. For I_n the $n \times n$ identity matrix we set

$$A := \begin{bmatrix} I_n & \frac{\|c\|_2}{\sqrt{\|a\|_2^2+\|b\|_2^2}} I_n \\ \frac{\|c\|_2}{\sqrt{\|a\|_2^2+\|b\|_2^2}} I_n & 0_n \end{bmatrix}$$

and V be the $2n$ vector obtained by concatenating X_1, W_1, that is $V := (X_1, W_1)^t$. Then the last inequality can be rewritten with respect to the matrix notation as

$$V^t A V \leq \left(\frac{1}{2} + 4\delta_{3k}\right) n.$$

We now bound the probability of this inequality. First note that since V is a vector with iid standard Gaussian elements it holds that $\mathbb{E}[V^t A V] = \text{trace}\,(A) = n$. Hence,

$$\begin{aligned}
&\mathbb{P}\left(V^t A V \leq \left(\frac{1}{2} + 4\delta_{3k}\right) n\right) \\
&\leq \mathbb{P}\left(|V^t A V - \mathbb{E}\big[V^t A V\big]| \geq (\frac{1}{2} - 4\delta_{3k}) n\right), \text{ using } \mathbb{E}\big[V^t A V\big] = n, \\
&\leq \mathbb{P}\left(|V^t A V - \mathbb{E}\big[V^t A V\big]| \geq \frac{n}{4}\right), \text{ using that } \delta_{3k} < \frac{1}{16} \text{ implies } \frac{1}{2} - 4\delta_{3k} > \frac{1}{4}.
\end{aligned}$$

Now we apply Hanson-Wright inequality, so we need to estimate the Frobenious norm and the spectral norm of the matrix A. We have

$$\|A\|_F^2 \leq 3n\|A\|_\infty^2 \leq 3\max\{1, \frac{\|c\|_2^2}{\|a\|_2^2 + \|b\|_2^2}\} n. \tag{4.4}$$

Now using that A can be represented as the Kronecker product

$$A = \begin{bmatrix} 1 & \frac{\|c\|_2}{\sqrt{\|a\|_2^2+\|b\|_2^2}} \\ \frac{\|c\|_2}{\sqrt{\|a\|_2^2+\|b\|_2^2}} & 0 \end{bmatrix} \otimes I_n$$

we obtain that the maximal eigenavalue of A is the maximal eigenvalue of the 2×2 first product term of the Kronecker product. In particular from this it can be easily checked that,

$$\|A\| \leq 2 \max\{1, \sqrt{\frac{\|c\|_2^2}{\|a\|_2^2 + \|b\|_2^2}}\}. \tag{4.5}$$

Now from Hanson-Wright inequality we have for some constant $d > 0$,

$$\mathbb{P}\left(|V^t A V - \mathbb{E}\big[V^t A V\big]| \geq \frac{1}{4}n\right) \leq 2 \exp\left[-d \min\left(\frac{\frac{1}{16}n^2}{\|A\|_{\mathrm{F}}^2}, \frac{\frac{1}{4}n}{\|A\|}\right)\right] \tag{4.6}$$

Using (4.4), (4.5) and noticing that $\max\{1, \sqrt{\frac{\|c\|_2^2}{\|a\|_2^2+\|b\|_2^2}}\} \leq \max\{1, \frac{\|c\|_2^2}{\|a\|_2^2+\|b\|_2^2}\}$ we obtain that for the constant $c_0 := \frac{1}{48}d$ it holds

$$d \min\left(\frac{\frac{1}{16}n^2}{\|A\|_{\mathrm{F}}^2}, \frac{\frac{1}{4}n}{\|A\|}\right) \geq c_0 n \min\{1, \frac{\|a\|_2^2 + \|b\|_2^2}{\|c\|_2^2}\}$$

and therefore using (4.6) the proof is complete in this case. □

Now we proceed with the proof of the proposition. We define the following sets parametrized by $r, \tilde{c} > 0$ and $\alpha \in (0, 1)$

$$B_{r,\tilde{c}} := \{(a,b,c)\big|a,b,c \in \mathbb{R}^p, \|a\|_0 + \|b\|_0 + \|c\|_0 \leq 2k+1, \|a\|_2^2 + \|b\|_2^2 + \|c\|_2^2 \leq r^2, |a|_{\min} \geq \tilde{c}\}$$

and

$D_{\alpha,r,\tilde{c}}$ equal to

$\{(a,b,c) \in B_{r,\tilde{c}}\big|(a,b,c)$ is α-D.L.M. with correspondning super-supports satisfying the assumption (2) of the Proposition 4.3.10 $\}$

We call a triplet of sets $\emptyset \neq S_1, S_2, S_3 \subseteq [p]$ *good* if

- S_1, S_2, S_3 are pair-wise disjoint
- $|S_1| = |S_2|$, $p \in S_3$ and $S_1 \cup S_3 = [k] \cup \{p\}$

For $\alpha \in \mathbb{R}$ and $S \subseteq \mathbb{R}$ we define the set

$$S - \alpha := \{s - \alpha | s \in S\}.$$

For $i = 1, 2, 3$ we set $P_i := \{(i-1)p + 1, (i-1)p + 2, \ldots, ip\}$. Notice that the sets P_1, P_2, P_3 partition $[3p]$. We define the following family of subsets of $[3p]$,

$$\mathcal{T} := \{T \subset [3p] | \text{ the triplet } T \cap P_1, T \cap P_2 - p, T \cap P_3 - 2p \text{ is good}\}.$$

It is easy to see that $\mathcal{T} \subset \{T \subset [3p] | |T| \leq 2k + 1\}$. Furthermore for any $T \in \mathcal{T}$ we define

$$B_{r,\tilde{c}}(T) := \{(a, b, c) \in B_{r,\tilde{c}} \big| \text{Support}\,((a, b, c)) \subseteq T, T \cap P_1 = \text{Support}\,(a)\}$$

and

$D_{\alpha,r,\tilde{c}}(T)$ equal to
$\{(a, b, c) \in B_{r,\tilde{c}}(T) \big| (a, b, c)$ is α-D.L.M. with respect to $T \cap P_1, T \cap P_2 - p, T \cap P_3 - 2p\}$.

We claim that

$$D_{\frac{1}{4},r,1} = \bigcup_{T \in \mathcal{T}} D_{\frac{1}{4},r,1}(T). \tag{4.7}$$

For the one direction, if $A = (a, b, c) \in D_{\frac{1}{4},r,1}(T)$ for some $T \in \mathcal{T}$ then (a, b, c) is α-DLM with corresponding super-supports $T \cap P_1, T \cap P_2 - p, T \cap P_3 - 2p$ which can be easily checked that they satisfy assumption (2) of the Proposition 4.3.10 based on our assumptions. For the other direction if $A \in D_{\frac{1}{4},r,1}$ is an α-DLM with respect to S_1, S_2, S_3 satisfying the assumption (2) of the Proposition, it can be easily verified that for the set $T = S_1 \cup (S_2 + p) \cup (S_3 + 2p)$ it holds $T \in \mathcal{T}$ and furthermore $A \in D_{\frac{1}{4},r,1}(T)$.

Now to prove the proposition it suffices to prove that there exists $c_1, C_1 > 0$ such that if $n \geq C_1 k \log p$ and $r = \sqrt{c_1 \min\{\frac{\log p}{\log\log p}, k\}}$ then

$$\lim_{k\to+\infty} \mathbb{P}\left(D_{\frac{1}{4},r,1} \neq \emptyset\right) = 0.$$

Using the equation (4.7) for $\alpha = \frac{1}{4}$ and $\tilde{c} = 1$ and the union bound it suffices to be shown that for some $c_1, C_1 > 0$ if $n \geq C_1 k \log p$ and $r = \sqrt{c_1 \min\{\frac{\log p}{\log\log p}, k\}}$ then

$$\lim_{k\to+\infty} \sum_{T\in\mathcal{T}} \mathbb{P}\left(D_{\frac{1}{4},r,1}(T) \neq \emptyset\right) = 0.$$

We now state and prove the following packing lemma.

Lemma 4.3.12. *There exists $C_2 > 0$ such that for any $r > 0, \delta \in (0,1)$ and $T \in \mathcal{T}$ we can find $Q_{r,1-\delta}(T) \subseteq B_{r,1-\delta}(T)$ with the following two properties*

- $|Q_{r,1-\delta}(T)| \leq C_2 \left(\frac{12r}{\delta}\right)^{2k+1}$.
- *For any $p \in B_{r,1}(T)$ there exists $q \in Q_{r,1-\delta}(T)$ with $\|p-q\|_2 \leq \delta$.*

Proof. Fix $r > 0, \delta \in (0,1)$ and $T \in \mathcal{T}$. Since $T \subset [3p]$ and $|T| \leq 2k+1$ using standard packing arguments (see for example [BDDW08]) there exists universal constant $C_2 > 0$ and a set

$$Q'_{r,1-\delta}(T) \subset B_r(T) := \{(a,b,c)\big|a,b,c \in \mathbb{R}^p, \text{Support}\left((a,b,c)\right) \subseteq T, \|a\|_2^2 + \|b\|_2^2 + \|c\|_2^2 \leq r^2\}$$

with the properties that $|Q'_{r,1-\delta}(T)| \leq C_2 \left(\frac{12r}{\delta}\right)^{2k+1}$ and that for any $p \in B_r(T)$ there exists $q \in Q'_{r,1-\delta}(T)$ with $\|p-q\|_2 \leq \delta$.

To complete the proof we define

$$Q_{r,1-\delta}(T) = Q'_{r,1-\delta}(T) \cap B_{r,1-\delta}(T).$$

As $Q_{r,1-\delta}(T) \subseteq Q'_{r,1-\delta}(T)$ it also holds

$$|Q_{r,1-\delta}(T)| \leq |Q'_{r,1-\delta}(T)| \leq C_2 \left(\frac{12r}{\delta}\right)^{2k+1}.$$

For the other property let $p = (a, b, c) \in B_{r,1}(T)$. Since $B_{r,1}(T) \subseteq B_r(T)$ there exist $q = (l, m, n) \in Q'_{r,1-\delta}(T)$ with $\|p - q\|_2 \leq \delta$. We claim that $q \in B_{r,1-\delta}(T)$ which completes the proof. It suffices to establish $|l|_{\min} \geq 1-\delta$ and that Support $(l) = T \cap P_1$. We know $\|a-l\|_\infty \leq \|a-l\|_2 \leq \|p-q\|_2 \leq \delta$. Therefore since for al $i \in T \cap P_1$, $|a_i| \geq 1$ we get that for all $i \in T \cap P_1$, $|l_i| \geq 1 - \delta$. Since $T \cap P_1$ was assumed to be a super-support of l this implies both Support $(l) = T \cap P_1$ and $|l|_{\min} \geq 1 - \delta$.

□

Claim 4.3.13. *Consider the sets $\{Q_{r,1-\delta}(T)\}_{T\in\mathcal{T}}$ from Lemma (4.3.12) defined for some $r > 0$ and $0 < \delta \leq \min\{\frac{1}{50r}, \frac{1}{5}\}$. If X satisfies the $3k$-RIP with $\delta_{3k} \in (0, 1)$ then for any $T \in \mathcal{T}$ such that $D_{\frac{1}{4},r,1}(T) \neq \emptyset$, we have $Q_{r,1-\delta}(T) \cap D_{\frac{1}{2},r,\frac{1}{2}}(T) \neq \emptyset$.*

Proof. To prove the claim, we consider an element $A = (a, b, c) \in D_{\frac{1}{4},r,1}(T)$. Note that since $A \in D_{\frac{1}{4},r,1}(T) \subseteq B_{r,1}(T) \subset B_{r,1-\delta}(T)$ the definition of $Q_{r,1-\delta}(T)$ implies that for some $L = (l, m, g) \in Q_{r,1-\delta}(T)$ it holds $\|A - L\|_2 \leq \delta$. To complete the proof we show that $L \in D_{\frac{1}{2},r,\frac{1}{2}}(T)$.

Notice that from the definition of the sets $Q_{r,1-\delta}(T), D_{\frac{1}{4},r,1}(T)$, the vectors a, l share the set $S_1 = T \cap P_1$ as a common super-support and furthermore the vectors b, m share the set $S_2 = T \cap P_2$ as a common super-support. Since $A \in D_{\frac{1}{4},r,1}(T)$ we know firstly $S_1 = \text{Support}(a)$, secondly for any $i \in S_1 = \text{Support}(a)$, $|a_i| \geq 1$ and finally that for any $i \in S_1$ and $j \in S_2$

$$\|\left(Xa - a_iX_i + Xb - b_jX_j\right) + Xc\|_2^2 \geq \|X(a + b + c)\|_2^2 - \frac{1}{4}\left(\frac{\|a\|_2^2}{|S_1|} + \frac{\|b\|_2^2}{|S_2|}\right)n. \tag{4.8}$$

To prove $L \in D_{\frac{1}{2},r,\frac{1}{2}}(T)$ it suffices to prove now firstly that $S_1 = \text{Support}(l)$, secondly for any $i \in \text{Support}(l)$, $|l_i| \geq \frac{1}{2}$ and finally that for every $i \in S_1$ and $j \in S_2$

$$\|\left(Xl - l_iX_i + Xm - m_jX_j\right) + Xg\|_2^2 \geq \|X(l + m + g)\|_2^2 - \frac{1}{2}\left(\frac{\|l\|_2^2}{|S_1|} + \frac{\|m\|_2^2}{|S_2|}\right)n. \tag{4.9}$$

We start with the first two properties. This is a similar calculation as in the proof of Lemma 4.3.12. We know $\|a - l\|_2 \leq \|A - L\|_2 \leq \delta < \frac{1}{2}$. In particular, $\|a - l\|_\infty \leq \frac{1}{2}$. But we know that $S_1 = \text{Support}(a)$ and $|a|_{\min} \geq 1$. These together imply that for all $i \in S_1$, $|l_i| \geq \frac{1}{2}$. Since S_1 is a super-support of l we conclude that indeed $S_1 = \text{Support}(l)$ and that for any $i \in \text{Support}(l)$, $|l_i| \geq \frac{1}{2}$ as required. Now we prove the third property and use Proposition 4.3.4. By part (2) of

this proposition we know that since X satisfies the $3k$-RIP for some restricted isometric constant $\delta_{3k} < 1$, any two vectors v, w which share a common super-support of size at most $3k$ satisfy

$$\|Xw\|_2^2 + 4\|v-w\|_2\|w\|_2 n + 2\|v-w\|_2^2 n \geq \|Xv\|_2 \geq \|Xw\|_2^2 - 4\|v-w\|_2\|w\|_2 n \tag{4.10}$$

For our convenience for the calculations that follow we set for all $i \in S_1$ and $j \in S_2$, $A_{i,j} := A - a_i e_i - b_j e_j$ and $L_{i,j} := L - l_i e_i - m_j e_j$, where by $\{e_i\}_{i\in[3p]}$ we denote the standard basis vectors of $\mathbb{R}^{3p}$. In words for all $i \in S_1$ and $j \in S_2$ we set $A_{i,j}$ the vector A after we set zero its i and j coordinates and similarly we define $L_{i,j}$. Now fix $i \in S_1$, $j \in S_2$. Then we have by directly applying (4.10) for the two pairs $v = L_{i,j}$ and $w = A_{i,j}$ and $v = L, w = A$ that

$$\|X(A_{i,j})\|_2^2 \leq \|X(L_{i,j})\|_2^2 + 4\|L_{i,j} - A_{i,j}\|_2\|A_{i,j}\|_2 n + 2\|L_{i,j} - A_{i,j}\|_2^2 n$$

and

$$\|X(A)\|_2^2 \geq \|X(L)\|_2^2 - 4\|A-L\|_2\|L\|_2 n,$$

Hence $\|X(A_{i,j})\|_2^2 - \|X(A)\|_2^2$ is at most

$$\|X(L_{i,j})\|_2^2 + 4\|L_{i,j} - A_{i,j}\|_2\|A_{i,j}\|_2 n + 2\|L_{i,j} - A_{i,j}\|_2^2 n - \|X(L)\|_2^2 + 4\|A-L\|_2\|A\|_2 n.$$

But using the easy observations

$$\|A_{i,j} - L_{i,j}\|_2 \leq \|A-L\|_2 \leq \delta$$

and

$$\|A_{i,j}\|_2 \leq \|A\|_2 \leq r$$

we get that the last quantity can be upper bounded by $\|XL_{i,j}\|_2^2 - \|XL\|_2^2 + (8\delta r + 2\delta^2)n$. Therefore combining the last steps we have established

$$\|X(A_{i,j})\|_2^2 - \|X(A)\|_2^2 \leq \|XL_{i,j}\|_2^2 - \|XL\|_2^2 + (8\delta r + 2\delta^2)n.$$

But we know that by our assumptions $\|X(A_{i,j})\|_2^2 - \|X(A)\|_2^2 \geq -\frac{1}{4}\left(\frac{\|a\|_2^2}{|S_1|} + \frac{\|b\|_2^2}{|S_2|}\right) n$. Therefore

$$\|XL_{i,j}\|_2^2 - \|XL\|_2^2 \geq -\frac{1}{4}\left(\frac{\|a\|_2^2}{|S_1|} + \frac{\|b\|_2^2}{|S_2|}\right) n - (8\delta r + 2\delta^2)n.$$

So to prove (4.9) it suffices to be proven that

$$-\frac{1}{4}\left(\frac{\|a\|_2^2}{|S_1|} + \frac{\|b\|_2^2}{|S_2|}\right) n - (8\delta r + 2\delta^2)n \geq -\frac{1}{2}\left(\frac{\|l\|_2^2}{|S_1|} + \frac{\|m\|_2^2}{|S_2|}\right) n. \tag{4.11}$$

Note that $\|A\|_2 \leq r, \|L\|_2 \leq r, \|A - L\|_2 \leq \delta$ implies $\|a\|_2^2 - \|l\|_2^2 \leq 2\delta r$ and $\|b\|_2^2 - \|m\|_2^2 \leq 2\delta r$. Hence from the definition of A, L and since $|S_1| = |S_2| \geq 1$ it holds,

$$\frac{1}{2}\left(\frac{\|a\|_2^2}{|S_1|} + \frac{\|b\|_2^2}{|S_2|}\right) n - \frac{1}{2}\left(\frac{\|l\|_2^2}{|S_1|} + \frac{\|m\|_2^2}{|S_2|}\right) n \leq 2\delta r n.$$

In particular it holds

$$-\frac{1}{2}\left(\frac{\|a\|_2^2}{|S_1|} + \frac{\|b\|_2^2}{|S_2|}\right) n \geq -\frac{1}{2}\left(\frac{\|l\|_2^2}{|S_1|} + \frac{\|m\|_2^2}{|S_2|}\right) n - 2\delta r n.$$

Hence using the last inequality we can immediately derive (4.11) provided that

$$\frac{1}{4}\left(\frac{\|a\|_2^2}{|S_1|} + \frac{\|b\|_2^2}{|S_2|}\right) n \geq 2\delta r n + (8\delta r + 2\delta^2)n = (10\delta r + 2\delta^2)n.$$

But now since $a_i^2 \geq 1$ for all $i \in S_1$, $\frac{\|a\|_2^2}{|S_1|} \geq 1$ and therefore

$$\frac{1}{4}\left(\frac{\|a\|_2^2}{|S_1|} + \frac{\|b\|_2^2}{|S_2|}\right) n \geq \frac{1}{4} n.$$

so it suffices that $2\delta^2 + 10\delta r \leq \frac{1}{4}$. It can be easily checked to be true if $\delta \leq \min\{\frac{1}{50r}, \frac{1}{5}\}$. The proof of the claim is complete. □

To prove the proposition we need to show that for some $c_1, C_1 > 0$ if $n \geq C_1 k \log p$, $r = \sqrt{c_1 \min\{\frac{\log p}{\log\log p}, k\}}$ and $\delta = \frac{1}{60r}$ then for the appropriately defined sets $\{Q_{r,1-\delta}(T)\}_{T\in\mathcal{T}}$ it holds

$$\lim_{k\to+\infty} \sum_{T\in\mathcal{T}} \mathbb{P}\left(|Q_{r,1-\delta}(T) \cap D_{\frac{1}{2},r,\frac{1}{2}}(T)| \geq 1\right) = 0.$$

But by Markov inequality for all such $T \in \mathcal{T}$,

$$\mathbb{P}\left(|Q_{r,1-\delta}(T) \cap D_{\frac{1}{2},r,\frac{1}{2}}| \geq 1\right) \leq \mathbb{E}\Big[|Q_{r,1-\delta}(T) \cap D_{\frac{1}{2},r,\frac{1}{2}}|\Big].$$

Furthermore for all $T \in \mathcal{T}$, $1 \leq |T \cap P_2| \leq k$. By the Markov inequality and summing over the possible values of $|T \cap P_2|$ for $T \in \mathcal{T}$, it suffices to show that for some $c_1, C_1 > 0$ if $n \geq C_1 k \log p$ and $r = \sqrt{c_1 \min\{\frac{\log p}{\log\log p}, k\}}$ then,

$$\lim_{k\to+\infty} \sum_{m=1}^{k} \sum_{T\in\mathcal{T},|T\cap P_2|=m} \mathbb{E}\left(|Q_{r,1-\delta}(T) \cap D_{\frac{1}{2},r,\frac{1}{2}}(T)|\right) = 0 \tag{4.12}$$

Fix $m \in [k]$ and a set $T \in \mathcal{T}$ with $|T \cap P_2| = m$. Then for any $A = (a, b, c) \in Q_{r,1-\delta}(T) \cap D_{\frac{1}{2},r,\frac{1}{2}}(T)$, since $D_{\frac{1}{2},r,\frac{1}{2}}(T) \subseteq B_{r,\frac{1}{2}}(T)$, we have $|a|_{\min} \geq \frac{1}{2}$ and $\|a\|_2^2 + \|b\|_2^2 + \|c\|_2^2 \leq r^2$. Based on the definition of $D_{\frac{1}{2},r,\frac{1}{2}}(T)$, we also have $|\text{Support}(a)| = |S_1| = |S_2| = |T \cap P_2| = m$. Hence, $\|a\|_2^2 \geq |a|_{\min}^2 m \geq \frac{1}{4}m$ and $\|c\|_2^2 \leq \|a\|_2^2 + \|b\|_2^2 + \|c\|_2^2 \leq r^2$. By Lemma 4.3.11 we know that for any triplet $A = (a, b, c)$, $\mathbb{P}\left(A \in D_{\frac{1}{2},r,\frac{1}{2}}(T)\right) \leq \exp\left(-c_0 n \min\{1, \frac{\|a\|_2^2+\|b\|_2^2}{\|c\|_2^2}\}\right)$. Hence using the above inequalities we can conclude that for any such $A = (a, b, c) \in Q_{r,1-\delta}(T)$ it holds

$$\mathbb{P}\left(A \in D_{\frac{1}{2},r,\frac{1}{2}}(T)\right) \leq 2\exp\left(-\frac{1}{4}c_0 n \min\{1, \frac{m}{r^2}\}\right) \tag{4.13}$$

Linearity of expectation, the above bound and the cardinality assumption on $Q_{r,1-\delta}(T)$ imply

$$\mathbb{E}\Big[|Q_{r,1-\delta}(T) \cap D_{\frac{1}{2},r,\frac{1}{2}}(T)|\Big] \leq 2|Q_{r,1-\delta}(T)| \exp\left(-\frac{1}{4}c_0 n \min\{1, \frac{m}{r^2}\}\right) \tag{4.14}$$

$$\leq 2C_2 \left(\frac{12r}{\delta}\right)^{2k+1} \exp\left(-\frac{1}{4}c_0 n \min\{1, \frac{m}{r^2}\}\right). \tag{4.15}$$

We now count the number of possible $T \in \mathcal{T}$ with $|T \cap P_2| = m$. Recall that any $T \subseteq [3p]$ satisfies $T \in \mathcal{T}$ if and only if the triplet of sets $T \cap P_1, T \cap P_2 - p, T \cap P_3 - 2p$ is a *good* triplet. That is if and only if

(1) $T\cap P_1, T\cap P_2-p, T\cap P_3-2p$ are pairwise disjoint sets and $|T\cap P_1| = |T\cap P_2-p| = |T\cap P_2| = m$

(2) $p \in T \cap P_3 - 2p$

(3) $(T \cap P_1) \cup (T \cap P_3 - 2p) = [k] \cup \{p\}$

Since a set $T \subseteq [3p]$ is completely characterized by the intersections with P_1, P_2, P_3, it suffices to count the number of triplets of sets $T \cap P_i$, $i = 1, 2, 3$ satisfying the three above conditions. Now conditions (1),(3) imply that $T \cap P_3$ is completely characterized by $T \cap P_1$. Furthermore by checking conditions $(1), (2), (3)$ we know that $T \cap P_1$ is an arbitrary subset of $[k]$ of size m. Hence we have $\binom{k}{m}$ choices for both the sets $T \cap P_1$ and $T \cap P_3$. Finally for the set $T \cap P_2$ we only have that it needs to satisfy $|T \cap P_2| = m$. Hence for $T \cap P_2$ we have $\binom{p}{m}$ choices, giving in total that the number of sets $T \in \mathcal{T}$ with $|T \cap P_2| = m$ equals to $\binom{k}{m}\binom{p}{m}$. Hence,

$$\sum_{T \in \mathcal{T}, |T \cap P_2| = m} \mathbb{E}\left(|Q_{r,1-\delta}(T) \cap D_{\frac{1}{2}}(T)|\right) \leq 2\binom{k}{m}\binom{p}{m} C_2 \left(\frac{12r}{\delta}\right)^{2k+1} \exp\left(-\frac{1}{4} c_0 n \min\{1, \frac{m}{r^2}\}\right).$$

Summing over all $m = 1, 2, \ldots, k$ and using the bounds $\binom{k}{m} \leq 2^k, \binom{p}{m} \leq p^m$ we conclude that

$$\sum_{m=1}^{k} \sum_{T \in \mathcal{T}, |T \cap P| = m} \mathbb{E}\left(|Q_{r,1-\delta}(T) \cap D_{\frac{1}{2}, r, \frac{1}{2}}(T)|\right)$$

is at most

$$2C_3 k 2^k \max_{m=1,\ldots,k} \left[p^m \left(\frac{12r}{\delta}\right)^{2k+1} \exp\left(-\frac{1}{4} c_0 n \min\{1, \frac{m}{r^2}\}\right)\right].$$

Therefore it suffices to show that for some $c_1, C_1 > 0$ if $n \geq C_1 k \log p$, $r = \sqrt{c_1 \min\{\frac{\log p}{\log\log p}, k\}}$ and $\delta = \frac{1}{60r}$ then

$$\lim_{k \to \infty} k 2^k \max_{m=1,\ldots,k} \left[p^m \left(\frac{12r}{\delta}\right)^{2k+1} \exp\left(-\frac{1}{4} c_0 n \min\{1, \frac{m}{r^2}\}\right)\right] = 0.$$

Since this is an increasing quantity in n and in $\frac{1}{\delta}$ we plug in $n = \frac{4}{c_0} C_1 k \log p$ and $\delta = \frac{1}{60r}$ (since $r \to +\infty$) and after taking logarithms it suffices to be proven that for C_1 large enough but constant and $c_1 > 0$ small enough but constant, if $r = \sqrt{c_1 \min\{\frac{\log p}{\log\log p}, k\}}$ then

$$\max_{m=1,\ldots,k} \left[m \log p + (2k+1) \log\left(1000 r^2\right) - C_1 k \log p \min\{1, \frac{m}{r^2}\}\right] + k \log 2 + \log k \to -\infty.$$

We consider the two cases: when $m \leq r^2$ and when $m \geq r^2$. Suppose $m \geq r^2$, that is $\min\{1, \frac{m}{r^2}\} = 1$. We choose c_1 small enough so that $1000r^2 \leq k \leq p$ and therefore

$$\begin{aligned}
&\max_{k \geq m \geq r^2} \left[m \log p + (2k+1) \log\left(1000r^2\right) - C_1 k \log p \min\{1, \frac{m}{r^2}\} \right] + k \log 2 + \log k \\
&= \max_{k \geq m \geq r^2} \left[m \log p + (2k+1) \log\left(1000r^2\right) - C_1 k \log p \right] + k \log 2 + \log k \\
&\leq -(C_1 - 4) k \log p + k \log 2 + \log k, \text{ since } m \log p + (2k+1) \log\left(1000r^2\right) \leq 4k \log p, \\
&\leq -(C_1 - 5) k \log p,
\end{aligned}$$

which if $C_1 > 6$ clearly diverges to $-\infty$ as $k \to +\infty$.

Now suppose $m \leq r^2$, that is when $\min\{1, \frac{m}{r^2}\} = \frac{m}{r^2}$. We have

$$\begin{aligned}
&\max_{1 \leq m \leq r^2} \left[m \log p + (2k+1) \log\left(1000r^2\right) - C_1 k \log p \min\{1, \frac{m}{r^2}\} \right] + k \log 2 + \log k \\
&= \max_{1 \leq m \leq r^2} \left[m \log p + (2k+1) \log\left(1000r^2\right) - C_1 k \log p \frac{m}{r^2} \right] + k \log 2 + \log k.
\end{aligned}$$

We write

$$\begin{aligned}
&m \log p + (2k+1) \log\left(1000r^2\right) - C_1 k \log p \frac{m}{r^2} \\
&= m \log p - \frac{C_1}{2} k \log p \cdot \frac{m}{r^2} + (2k+1) \log\left(1000r^2\right) - \frac{C_1}{2} k \log p \cdot \frac{m}{r^2}.
\end{aligned}$$

But now for $c_1 < 1$ we have $r^2 \leq k$ and therefore

$$m \log p - \frac{C_1}{2} k \log p \cdot \frac{1}{4} \frac{m}{r^2} \leq (1 - \frac{C_1}{2}) m \log p \leq -2 \log p \tag{4.16}$$

for $C_1 \geq 6$. Now we will bound the second summand. Again assuming $C_1 > 6$ and using that $m \geq 1$ we have

$$(2k+1) \log\left(1000r^2\right) - \frac{C_1}{2} k \log p \cdot \frac{m}{r^2} \leq 3k \left(\log\left(1000r^2\right) - \frac{1}{4r^2} \log p \right) \tag{4.17}$$

Now we claim that the right hand side of the above inequalty is at most $-3k$, given c_1 small enough, as $k \to +\infty$. It suffices to prove that if $r \leq \sqrt{c_1 \frac{\log p}{\log \log p}}$ for some $c_1 > 0$ small enough

then $\log(1000r^2) - \frac{1}{4r^2}\log p \leq -1$ or equivalently $r^2\log(1000r^2) + r^2 \leq \frac{1}{4}\log p$. But notice that the left hand side of the last inequality is increasing in r and it can be easily checked that if $r^2 = \frac{1}{5}\frac{\log p}{\log\log p}$ then $\frac{r^2\log\left(1000r^2\right)+r^2}{\log p}$ tends in the limit (as p grows to infinity) to $\frac{1}{5}$ which is less than $\frac{1}{4}$. Therefore if $c_1 < \frac{1}{5}$ the inequality becomes true for large enough p for this value of r and my monotonicity for all smaller values of r as well. Now combining (4.16) and (4.17) we conclude that for small enough $c_1 > 0$ and large enough $C_1 > 0$ that

$$\begin{aligned}
&\max_{1\leq m\leq 4r^2}\left[m\log p + (2k+1)\log\left(1000r^2\right) - C_1 k\log p\frac{1}{4}\frac{m}{r^2}\right] + k\log 2 + \log k \\
&\leq -2\log p - 3k + k\log 2 + \log k \\
&\leq -(3-2\log 2)k + \log k \to -\infty, \text{ as } n,p,k \to +\infty
\end{aligned}$$

which completes the proof.

□

4.3.3 Proof of Theorems 4.2.2, 4.2.5 and 4.2.6

We first prove Theorem 4.2.6 and then we show how it implies Theorems 4.2.2 and 4.2.5.

Proof of Theorem 4.2.6. Let X' be an $n\times(p+1)$ matrix such that for all $i\in[n], j\in[p]$ it holds $X'_{i,j} = X_{i,j}$ and for $i\in[n], j = p+1$, $X'_{i,p+1} := \frac{1}{\sigma}W_i$. In words, we create X' by augmenting X with the rescaled $\frac{1}{\sigma}W$ as an extra column. Note that X' has iid standard Gaussian entries and furthermore $Y = X\beta^* + W = X'\begin{bmatrix}\beta^*\\ \sigma\end{bmatrix}$.

Notice that the performance of our algorithm is invariant with respect to rescaling of the quantities $Y, \beta^*, \sigma, \beta_0$ by a scalar. In particular by rescaling $Y = X\beta^* + W$ with $\frac{1}{|\beta^*|_{\min}}$ we can replace Y by $\frac{Y}{|\beta^*|_{\min}}$, β^* with $\frac{\beta^*}{|\beta^*|_{\min}}$, σ^2 by $\frac{\sigma^2}{|\beta^*|^2_{\min}}$ and finally β_0 by $\frac{\beta_0}{|\beta^*|^2_{\min}}$ and thus we may assume for our proof that $|\beta^*|_{\min} = 1$. Notice that in this case our desired upper bound on the running time remains $4k\frac{\|Y-X\beta_0\|_2^2}{\sigma^2 n}$ and our assumptions on the variance of the noise is now simply $\sigma^2 \leq c\min\{\frac{\log p}{\log\log p}, k\}$ for some $c > 0$.

Recall that the desired output of the algorithm are vectors $\hat{\beta}$ satisfying the following termination conditions.

Termination Conditions:

(TC1) Support $\left(\hat{\beta}\right) = \text{Support}\,(\beta^*)$ and,

(TC2) $\|\hat{\beta} - \beta^*\|_2 \leq \sigma$.

We start with the following deterministic claim.

Claim 4.3.14. *Assume that the algorithm LSA has the following property. For any k-sparse β which violates at least one of (TC1),(TC2) we have $\|Y - X\beta'\|_2^2 \leq \|Y - X\beta\|_2^2 - \frac{\sigma^2}{4k}n$, where β' is obtained from β in one iteration of the LSA. Then the algorithm LSA terminates for any k-sparse vector β_0 as input in at most $4k\frac{\|Y - X\beta_0\|_2^2}{\sigma^2 n}$ iterations with an output vector β satisfying both conditions $(TC1), (TC2)$.*

Proof. The property clearly implies that for the algorithm to terminate it needs to satisfy both conditions $(TC1), (TC2)$. Hence we need to bound only the termination time appropriately. But since at every iteration that the algorithm does not terminate the quantity $\|Y - X\beta\|_2^2$ decreases by at least $\frac{\sigma^2}{4k}n$, the result follows.

□

For any vector $v \in \mathbb{R}^p$ and $\emptyset \neq A \subseteq [p]$ we denote by $v_A \in \mathbb{R}^p$ the p-dimensional real vector such that $(v_A)_i = v_i$ for $i \in A$ and $(v_A)_i = 0$ for $i \notin A$. Furthermore we set $v_\emptyset = 0_{p\times 1}$ for any vector v. Without the loss of generality from now on we assume Support $(\beta^*) = [k]$. Following the Claim 4.3.14 and our discussion, in order to prove Theorem 4.2.6 it suffices to prove that there exists $c, C > 0$ such that w.h.p. there is no k-sparse β that violates at least one of (TC1),(TC2) and furthermore satisfies that $\|Y - X\beta'\|_2^2 \geq \|Y - X\beta\|_2^2 - \frac{\sigma^2}{4k}n$, where β' is obtained from β in one iteration of the LSA.

Suppose the existence of such a β. We first choose $C > 0$ large enough so that X' satisfies the $3k$-RIP with $\delta_{3k} < \frac{1}{12}$. The existence of this $C > 0$ is guaranteed by Theorem 4.3.3. Denote by T a super support of β, that satisfies $|T| = k$ and $T \cap [k] = \text{Support}\,(\beta) \cap [k]$. The existence of T is guaranteed as $|\text{Support}\,(\beta)\,| \leq k$ and $k \leq \frac{p}{3}$. Note that in particular that (TC1) is satisfied if and only iff Support $(\beta) = [k]$ if and only if $T = [k]$. We know that for all $i \in [p]$, $j \in T$ and $q \in \mathbb{R}$,

$$\|Y - X\beta + \beta_j X_j - qX_i\|_2^2 \geq \|Y - X\beta\|_2 - \frac{\sigma^2}{4k}n$$

or equivalently,

$$\|X\beta^* + W - X\beta + \beta_j X_j - qX_i\|_2^2 \geq \|X\beta^* + W - X\beta\|_2 - \frac{\sigma^2}{4k}n, \forall i \in [p], j \in T, q \in \mathbb{R}. \quad (4.18)$$

Consider the triplets $(a, b, c), (d, e, g) \in \mathbb{R}^{p+1} \times \mathbb{R}^{p+1} \times \mathbb{R}^{p+1}$, where

$$a := \begin{bmatrix} \beta^*_{[k]\setminus T} \\ 0 \end{bmatrix}, b := \begin{bmatrix} -\beta_{T\setminus[k]} \\ 0 \end{bmatrix}, c := \begin{bmatrix} (\beta^* - \beta)_{[k]\cap T} \\ \sigma \end{bmatrix}$$

and

$$d := \begin{bmatrix} (\beta^* - \beta)_{[k]\cap T} \\ 0 \end{bmatrix}, f := \begin{bmatrix} 0_{p\times 1} \\ 0 \end{bmatrix}, g := \begin{bmatrix} (\beta^*)_{[k]\setminus T} - (\beta)_{T\setminus[k]} \\ \sigma \end{bmatrix}.$$

Lemma 4.3.15. *Assume that $\|(\beta - \beta^*)_{[k]\cap T}\|_2^2 \geq \sigma^2$. Then the inequalities (4.18) imply that the triplet (d, f, g) is $\frac{1}{4}$-DLM with respect to the matrix X'.*

Proof. We use the relation (4.18) and we choose $i = j \in [k] \cap T$, and $q = \beta_i^*$ to get that

$$\|X\beta^* + W - X\beta + (\beta_i - \beta_i^*)X_i\|_2^2 \geq \|X\beta^* + W - X\beta\|_2 - \frac{\sigma^2}{4k}n, \text{ for all } i \in [k] \cap T.$$

But now notice that with respect to $X' \in \mathbb{R}^{n\times(p+1)}$ and the vectors d, f, g defined above this condition can be written as

$$\|X'd + X'f + X'g - d_i X'_i\|_2^2 \geq \|X'(d + f + g)\|_2^2 - \frac{\sigma^2}{4k}n, \text{ for all } i \in [k] \cap T. \quad (4.19)$$

But based on our assumptions we have

$$\frac{\|d\|_2^2 + \|f\|_2^2}{|[k] \cap T|} = \frac{\|(\beta - \beta^*)_{[k]\cap T}\|_2^2}{|[k] \cap T|} \geq \frac{\sigma^2}{|[k] \cap T|} \geq \frac{\sigma^2}{k}$$

which combined with the inequality above gives,

$$\| (X'd - d_i X_i') + X'f + X'g\|_2^2 \geq \|X'd + X'f + X'g\|_2^2 - \frac{1}{4}\frac{\|d\|_2^2 + \|f\|_2^2}{|[k]\cap T|}n, \text{ for all } i \in [k]\cap T, \tag{4.20}$$

which by definition since $f = 0$ says that (d, f, g) is a $\frac{1}{4}$-DLM triplet with respect to $[k]\cap T$, U and Support(g), where U is an arbitrary set of cardinality $|[k]\cap T|$ which is disjoint from $[k]\cap T$ and Support(g). □

Recall that β does not satisfy at least one of (TC1) and (TC2). We now consider different cases with respect to that.

Case 1: $T = [k]$ but $\|\beta - \beta^*\|_2^2 > \sigma^2$.

In that case $\|(\beta - \beta^*)_{[k]\cap T}\|_2^2 \geq \sigma^2$, because $T = [k]$. In particular, from Claim 4.3.15 we know that (d, f, g) is a $\frac{1}{4}$-DLM triplet with respect to the matrix X'. From Lemma 4.3.9 since we assume that X' satisfies the $3k$-RIP with $\delta_{3k} < \frac{1}{12}$ w.h.p. we know that for (d, f, g) to be a $\frac{1}{4}$-DLM triplet it needs to satisfy

$$\|d\|_2^2 + \|f\|_2^2 < \frac{1}{4}\|g\|_2^2, \text{ w.h.p.}$$

which equivalently means

$$\| (\beta - \beta^*)_{[k]\cap T} \|_2^2 < \frac{1}{4}\left(\|\beta^*_{[k]\setminus T}\|_2^2 + \|\beta_{T\setminus[k]}\|_2^2 + \sigma^2\right) \text{ w.h.p.}$$

or equivalently as $T = [k]$

$$\|\beta - \beta^*\|_2^2 < \frac{\sigma^2}{4}\text{w.h.p.}$$

This is a contradiction with our assumption on β that $\|\beta - \beta^*\|_2^2 > \sigma^2$. Therefore indeed this case leads w.h.p. to a contradiction and the proof in this case is complete.

Case 2: $T \neq [k]$.

We start by proving that in this case if we choose $c < 1$ then the inequalities (4.18) imply deterministically that (a, b, c) is an $\frac{1}{4}$-DLM triplet with respect to $[k]\setminus T$, $T\setminus[k]$ and $([k]\cap T)\cup$

$\{p+1\}$ and the matrix X'. For $i \in [k] \setminus T$, $j \in T \setminus [k]$ and $q = \beta^*_j$ (4.18) implies

$$\|X\beta^* + W - X\beta + \beta_j X_j - \beta^*_i X_i\|_2^2 \geq \|X\beta^* + W - X\beta\|_2 - \frac{\sigma^2}{4k}n, \text{ for all } i \in [k] \setminus T, j \in T \setminus [k].$$

But now notice that with respect to $X' \in \mathbb{R}^{n \times (p+1)}$, and the vectors a, b, c defined above, this condition can be written as

$$\|X'a + X'b + X'c - a_i X'_i - b_j X'_j\|_2^2 \geq \|X'(a+b+c)\|_2^2 - \frac{\sigma^2 n}{4k}, \tag{4.21}$$

$$\text{for all } i \in [k] \setminus T, j \in T \setminus [k] \tag{4.22}$$

Furthermore since the non-zero elements of a are non-zero elements of β^* we know $|a|_{\min} \geq 1$. In particular for all $i \in [k] \setminus T$ it holds $a_i^2 \geq 1$ and therefore for $m = |[k] \setminus T|$ it holds $\frac{\|a\|_2^2 + \|b\|_2^2}{m} \geq |a|_{\min} \geq 1$. Therefore the inequality above implies

$$\|X'a + X'b + X'c - a_i X'_i - b_j X'_j\|_2^2 \geq \|X'a + X'b + X'c\|_2^2 - \frac{\sigma^2 n}{4k}\left(\frac{\|a\|_2^2 + \|b\|_2^2}{m}\right), \tag{4.23}$$

$$\text{for all } i \in [k] \setminus T, j \in T \setminus [k] \tag{4.24}$$

Finally, since we are assuming $c < 1$ we have $\sigma^2 \leq k$ and therefore

$$\|\left(X'a - a_i X'_i\right) + \left(X'b - b_j X'_j\right) + X'c\|_2^2 \geq \|X'a + X'b + X'c\|_2^2 - \frac{n}{4}\left(\frac{\|a\|_2^2 + \|b\|_2^2}{m}\right), \tag{4.25}$$

$$\text{for all } i \in [k] \setminus T, j \in T \setminus [k] \tag{4.26}$$

which since $m = k - |[k] \cap T| = |[k] \setminus T| = |T \setminus [k]|$ is exactly the property that (a, b, c) is $\frac{1}{4}$-DLM with respect to the sets $[k] \setminus T$, $T \setminus [k]$ and $([k] \cap T) \cap \{p+1\}$ and the matrix X'. Since we assume that X' satisfies the $3k$-RIP with $\delta_{3k} < \frac{1}{12}$ we conclude from Proposition 4.3.9 that $\|a\|_2^2 + \|b\|_2^2 \leq \frac{1}{4}\|c\|_2^2$ or equivalently,

$$\|\beta^*_{[k]\setminus T}\|_2^2 + \|\beta_{T \setminus [k]}\|_2^2 \leq \frac{1}{4}\left(\|(\beta - \beta^*)_{[k]\cap T}\|_2^2 + \sigma^2\right). \tag{4.27}$$

Now we apply Proposition 4.3.10 for the $\frac{1}{4}$-DLM triplet (a, b, c) with respect to $S_1 := [k] \setminus T$, $S_2 := T \setminus [k]$ and $S_3 := ([k] \cap T) \cup \{p+1\}$. Let $c_1, C_1 > 0$ the corresponding constants of the

proposition. We choose our C to satisfy $C > C_1$ so that the hypothesis of the Proposition 4.3.10 applies for any $\frac{1}{4}$-DLM triplet with respect to our matrix X'. In particular since (a, b, c) is a $\frac{1}{4}$-DLM triplet we know that it should not satisfy one of the conditions w.h.p. We have that $|a|_{\min} \geq 1$ and it is easy to check that $S_1 \cup S_3 = [k] \cup \{p+1\}$, $p+1 \in S_3$ and $S_1 = \text{Support}(a)$. Therefore from the conclusion of Proposition 4.3.10 it must be true that the triplet (a, b, c) must violate the third condition, that is

$$c_1 \min\{\frac{\log p}{\log\log p}, k\} \leq \|a\|_2^2 + \|b\|_2^2 + \|c\|_2^2, \text{ w.h.p.}$$

or equivalently

$$c_1 \min\{\frac{\log p}{\log\log p}, k\} \leq \|\beta^*_{[k]\setminus T}\|_2^2 + \|\beta_{T\setminus[k]}\|_2^2 + \|(\beta - \beta^*)_{T\cap[k]}\|_2^2 + \sigma^2,$$

Applying inequality (4.27) with the last inequality we conclude

$$c_1 \min\{\frac{\log p}{\log\log p}, k\} \leq \frac{1}{4}(\|(\beta - \beta^*)_{[k]\cap T}\|_2^2 + \sigma^2) + \|(\beta - \beta^*)_{T\cap[k]}\|_2^2 + \sigma^2,$$

or equivalently

$$\frac{4}{5}c_1 \min\{\frac{\log p}{\log\log p}, k\} - \sigma^2 \leq \|(\beta - \beta^*)_{[k]\cap T}\|_2^2,$$

Choosing our constant $c > 0$ to satisfy $c < \frac{2}{5}c_1$, we can assume $2\sigma^2 < \frac{4}{5}c_1 \min\{\frac{\log p}{\log\log p}, k\}$ and therefore the last inequality implies

$$\sigma^2 \leq \| (\beta - \beta^*)_{[k]\cap T} \|_2^2, \tag{4.28}$$

This by Lemma 4.3.15 implies that (d, f, g) is also an $\frac{1}{4}$-DLM triplet. In particular from Proposition 4.3.9 we have

$$\|d\|_2^2 + \|f\|_2^2 < \frac{1}{4}\|g\|_2^2,$$

which equivalently means

$$\| (\beta - \beta^*)_{[k]\cap T} \|_2^2 < \frac{1}{4}\left(\|\beta^*_{[k]\setminus T}\|_2^2 + \|\beta_{T\setminus[k]}\|_2^2 + \sigma^2\right).$$

Using (4.27) the above inequality implies w.h.p.

$$\| (\beta - \beta^*)_{[k]\cap T} \|_2^2 < \frac{1}{4} \left(1/4(\|(\beta - \beta^*)_{[k]\cap T}\|_2^2 + \sigma^2) + \sigma^2\right)$$

which implies

$$\| (\beta - \beta^*)_{[k]\cap T} \|_2^2 < \frac{1}{3}\sigma^2,$$

a contradiction with the inequality (4.28).

□

Proof of Theorem 4.2.2 and Theorem 4.2.5. Given Proposition 4.2.4 we only need to establish Theorem 4.2.5 to establish both of the Theorems, that is we only need to prove that there is no non-trivial local minimum for $(\tilde{\Phi}_2)$ w.h.p. We choose constants $c, C > 0$ so that the conclusion of Theorem 4.2.6 is valid. Suppose the existence of a k-sparse vector β which is a non-trivial local minimum for $(\tilde{\Phi}_2)$, that is it satisfies the following conditions (a),(b);

(a) $\text{Support}\,(\beta) \neq \text{Support}\,(\beta^*)$, and

(b) if a k-sparse β_1 satisfies

$$\max\{|\text{Support}\,(\beta) \setminus \text{Support}\,(\beta_1)\,|, |\text{Support}\,(\beta_1) \setminus \text{Support}\,(\beta)\,|\} \leq 1,$$

it must also satisfy

$$\|Y - X\beta_1\|_2 \geq \|Y - X\beta\|_2.$$

We feed now β as an input for the algorithm (LSA). From condition (b) we know that the algorithm will terminate immediately without updating the vector. But from Theorem 4.2.6 we know that the output of LSA with arbitrary k-sparse vector as input will output a vector satisfying conditions $(1), (2)$ of Theorem 4.2.6 w.h.p. In particular, since β was the output of LSA with input itself, it should satisfy condition (1) w.h.p., that is $\text{Support}\,(\beta) = \text{Support}\,(\beta^*)$, w.h.p. which contradicts the definition of β (condition (a)). Therefore w.h.p. there does not exist a non-trivial local minimum for $(\tilde{\Phi}_2)$. This completes the proof. □

4.4 Conclusion

In this Chapter, we continue our study of the high dimensional linear regression model under Gaussian assumptions on X, W and sparsity assumptions on β^*. In contrast to Chapters 2, 3, this Chapter does not assume the vector β^* is binary-valued, and real values are allowed for the entries of β^*. Our focus is on the "easy" regime, that is $n > n_{\text{alg}}$ where computationally efficient methods such as LASSO are known to provably recover the support of the vector β^*.

When $n > Cn_{\text{alg}}$ for some sufficiently large constant $C > 0$, we show that the Overlap Gap Property indeed ceases to hold. This confirms, up to the multiplicative constant $C > 0$, the behavior suggested by the first moment curve analysis, presented in Chapter 3. To establish this we perform a direct local analysis of the maximum likelihood estimation optimization problem of the model $(\tilde{\Phi}_2)$. We show that the landscape of the optimization problem is extremely smooth at the easy regime: when $n > Cn_{\text{alg}}$ all the local minima have identical support with β^*. Finally, we show that this can be exploited by a greedy local search algorithms which successfully works in termination time which is, in principle, independent of the growing feature size p.

Chapter 5

The Noiseless High Dimensional Linear Regression. A Lattice Basis Reduction Optimal Algorithm.

5.1 Introduction

We consider the following high-dimensional linear regression model. Consider n samples of a vector $\beta^* \in \mathbb{R}^p$ in a vector form $Y = X\beta^* + W$ for some $X \in \mathbb{R}^{n\times p}$ and $W \in \mathbb{R}^n$. Given the knowledge of Y and X the goal is to infer β^* using an efficient algorithm and the minimum number n of samples possible. Throughout the Chapter we call p the number of features, X the measurement matrix and W the noise vector.

This Chapter is devoted to the study of the high dimensional linear regression model but under significantly different assumptions compared to the Chapters 2, 3 and 4. For this reason, we motivate and carefully define the assumptions on $X, W\beta^*$ from scratch. Most results in the literature and ourselves in the previous Chapters study the high dimensional linear regression model under *sparsity assumption* on β^*, which refers to β^* having only a limited number of non-zero entries compared to its dimension [Don06], [CRT06], [FR13]. This allows valid inference of β^* with much less samples than feautres. During the past decades, the sparsity assumption led to a fascinating line of research in statistics and compressed sensing, which established, among other results, that several polynomial-time algorithms, such as Basis Pursuit Denoising Scheme

and LASSO, can efficiently recover a sparse β^* with number of samples much smaller than the number of features [CRT06], [Wai09b], [FR13]. For example, as we mentioned in the Introduction and in previous Chapters, it is established that if β^* is constrained to have at most $k \leq p$ non-zero entries, X has iid $N(0,1)$ entries, W has iid $N(0,\sigma^2)$ entries for $\sigma^2 = O(k)$, and n is of the order $k \log\left(\frac{p}{k}\right)$, then both of the mentioned algorithms can recover β^*, up to the level of the noise. Different structural assumptions than sparsity have also been considered in the literature. For example, a recent result [BJPD17] makes the assumption that β^* lies near the range of an L-Lipschitz generative model $G : \mathbb{R}^k \to \mathbb{R}^p$ and it proposes an algorithm which succeeds with $n = O(k \log L)$ samples.

A downside of all of the above results is that they provide no computationally efficient guarantee in the case n is much smaller than $k \log\left(\frac{p}{k}\right)$. Consider for example the case where the components of a sparse β^* are binary-valued, and X, W follow the Gaussian assumptions described above. Then as dicsussed in the Chapter 3 the statistical limit of the model is $n = n_{\text{info}} = 2k \log(p/k)/\log\left(k/\sigma^2 + 1\right)$. Supposing that σ is sufficiently small, it is a straightforward argument that n_{info} trivialized to zero and therefore when $n = 1$, β^* is recoverable from $Y = \langle X, \beta^* \rangle + W$. This can also be verified by a brute-force method which finds β^* directly, as β^* is the only binary k-sparse vector which can satisfy $Y = \langle X, \beta^* \rangle + W$ with probability tending to one as p goes to infinity (whp). On the other hand, for sparse and binary-valued β^*, the Basis Pursuit method in the noiseless case [DT10] and the Basis Pursuit Denoising Scheme in the noisy case [GZ17b] have been proven to fail to recover a binary β^* with $n = o(k \log\left(\frac{p}{k}\right))$ samples. Furthermore, LASSO has been proven to fail to recover a vector with the same support of β^*, with $n = o(k \log p)$ samples [Wai09b]. This failure to capture the complexity of the problem accurately enough for small sample sizes also lead to an algorithmic hardness conjecture for the regime $n = o(k \log\left(\frac{p}{k}\right))$ [GZ17a], [GZ17b] which is described in Chapters 3, 4. While this conjecture still stands in the general case, as we show in this Chapter, in the special case where β^* is rational-valued and the magnitude of the noise W is sufficiently small, the statistical computational gap can be closed and β^* can be recovered even when $n = 1$.

The structural assumption we impose on β^* is that its entries are rational numbers with denominator equal to some fixed positive integer value $Q \in \mathbb{Z}_{>0}$, something we refer to as the *Q-rationality assumption*. Note that for any Q, this assumption is trivially satisfied by the

binary-valued β^* which was discussed above. The 1-rationality assumption corresponds to β^* having integer entries, which is well-motivated in practise. For example, this assumption appears frequently in the study of global navigation satellite systems (GPS) and communications [HB98], [HV02], [BB99], [Bor11]. In the first reference the authors propose a mixed linear/integer model of the form $Y = Ax + Bz + W$ where z is an integer valued vector corresponding to integer multiples of certain wavelength. Several examples corresponding to regression models with integer valued regression coefficients and zero noise (though not always in the same model) are also discussed in the book [FR13]. In particular one application is the so-called Single-Pixel camera. In this model a vector β corresponds to color intensities of an image for different pixels and thus takes discrete values. The model assumes no noise, which is one of the assumptions we adopt in our model, though the corresponding regression matrix has i.i.d. $+1/-1$ Bernoulli entries, as opposed to a continuous distribution we assume. Two other applications involving noiseless regression models found in the same reference are MRI imaging and Radar detection.

A large body of literature on noiseless regression type models is a series of results on phase retrieval. Here the coefficients of the regression vector β^* and the entries of the regression matrix X are complex valued, but the observation vector $Y = X\beta^*$ is only observed through absolute values. This model has many applications, including crystallography, see [CESV15]. The aforementioned work provides many references to phase retrieval model including the cases when the entries of β^* have a finite support. We believe that our method can also be extended so that to model the case where the entries of the regression vector have a finite support, even if irrationally valued, and the entries of Y are only observed through their magnitude. In other words, we expect that the method presented in this Chapter applies to the phase retrieval problem at least in some of the cases and this is one of the current directions we are exploring.

Noiseless regression model with integer valued regression coefficients were also important in the theoretical development of compressive sensing methods. Specifically, Donoho [Don06] and Donoho and Tanner [DT05],[DT10],[DT09] consider a noiseless regression model of the form AB where A is a random (say Gaussian) matrix and B is the unit cube $[0,1]^p$. One of the goals of these results was to count number of extreme points of the projected polytope AB in order to explain the effectiveness of the linear programming based methods. The extreme points of this polytope can only appear as projections of extreme points of B which are all length-p binary

vector, namely one deals with noiseless regression model with binary coefficients – an important special case of the model we consider in this Chapter.

In the Bayesian setting, where the ground truth β^* is sampled according to a discrete distribution [DJM13] proposes a low-complexity algorithm which provably recovers β^* with $n = o(p)$ samples. This algorithm uses the technique of approximate message passing (AMP) and is motivated by ideas from statistical physics [KMS$^+$12]. Even though the result from [DJM13] applies to the general discrete case for β^*, it requires the matrix X to be spatially coupled, a property that in particular does not hold for X with iid standard Gaussian entries. Furthermore the required sample size for the algorithm to work is only guaranteed to be sublinear in p, a sample size potentially much bigger than the information-theoretic limit for recovery under sufficiently small noise ($n = 1$). In the present Chapter, where β^* satisfies the Q-rationality assumption, we propose a polynomial-time algorithm which applies for a large class of continuous distributions for the iid entries of X, including the normal distribution, and provably works even when $n = 1$.

The algorithm we propose is inspired by the algorithm introduced in [LO85] which solves, in polynomial time, a certain version of the so-called Subset-Sum problem. To be more specific, consider the following NP-hard algorithmic problem. Given $p \in \mathbb{Z}_{>0}$ and $y, x_1, x_2, \ldots, x_p \in \mathbb{Z}_{>0}$ the goal is to find a $\emptyset \neq S \subset [p]$ with $y = \sum_{i \in S} x_i$ when at least one such set S is assumed to exist. Over 30 years ago, this problem received a lot of attention in the field of cryptography, based on the belief that the problem would be hard to solve in many "real" instances. This would imply that several already built public key cryptosystems, called knapsack public key cryptosystems, could be considered safe from attacks [Lem79], [MH78]. This belief though was proven wrong by several works in the early 80s, see for example [Sha82]. Motivated by this line of research, Lagarias and Odlyzko in [LO85], and a year later Frieze in [Fri86], using a cleaner and shorter argument, proved the same surprising fact: if $x_1, x_2, \ldots, x_p$ follow an iid uniform distribution on $[2^{\frac{1}{2}(1+\epsilon)p^2}] := \{1, 2, 3, \ldots, 2^{\frac{1}{2}(1+\epsilon)p^2}\}$ for some $\epsilon > 0$ then there exists a polynomial-in-p time algorithm which solves the subset-sum problem whp as $p \to +\infty$. In other words, even though the problem is NP-hard in the worst-case, assuming a quadratic-in-p number of bits for the coordinates of x, the algorithmic complexity of the typical such problem is polynomial in p. The successful efficient algorithm is based on an elegant application of a seminal algorithm in the computational study of lattices called the Lenstra-Lenstra-Lovasz (LLL) algorithm, introduced

in [LLL82]. This algorithm receives as an input a basis $\{b_1, \ldots, b_m\} \subset \mathbb{Z}^m$ of a full-dimensional lattice $\mathcal{L}$ and returns in time polynomial in m and $\max_{i=1,2,\ldots,m} \log \|b_i\|_\infty$ a non-zero vector $\hat{z}$ in the lattice, such that $\|\hat{z}\|_2 \leq 2^{\frac{m}{2}} \|z\|_2$, for all $z \in \mathcal{L} \setminus \{0\}$.

Besides its significance in cryptography, the result of [LO85] and [Fri86] enjoys an interesting linear regression interpretation as well. One can show that under the iid uniform in $[2^{\frac{1}{2}(1+\epsilon)p^2}]$ assumption for $x_1, x_2, \ldots, x_p$, there exists exactly one set S with $y = \sum_{i \in S} x_i$ whp as p tends to infinity. Therefore if β^* is the indicator vector of this unique set S, that is $\beta_i^* = 1(i \in S)$ for $i = 1, 2, \ldots, p$, we have that $y = \sum_i x_i \beta_i^* = \langle x, \beta^* \rangle$ where $x := (x_1, x_2, \ldots, x_p)$. Furthermore using only the knowledge of y, x as input to the Lagarias-Odlyzko algorithm we obtain a polynomial in p time algorithm which recovers exactly β^* whp as $p \to +\infty$. Written in this form, and given our earlier discussion on high-dimensional linear regression, this statement is equivalent to the statement that the noiseless high-dimensional linear regression problem with binary β^* and X generated with iid elements from Unif$[2^{\frac{1}{2}(1+\epsilon)p^2}]$ is polynomial-time solvable even with one sample ($n = 1$), whp as p grows to infinity. The main focus of this Chapter is to extend this result to β^* satisfying the Q-rationality assumption, continuous distributions on the iid entries of X and non-trivial noise levels.

Summary of the Results

We propose a polynomial time algorithm for high-dimensional linear regression problem and establish a general result for its performance. We show that if the entries of $X \in \mathbb{R}^{n \times p}$ are iid from an arbitrary continuous distribution with bounded density and finite expected value, β^* satisfies the Q-rationality assumption, $\|\beta^*\|_\infty \leq R$ for some $R > 0$, and W is either an adversarial vector with infinity norm at most σ or has iid mean-zero entries with variance at most σ^2, then under some explicitly stated assumption on the parameters n, p, σ, R, Q our algorithm recovers exactly the vector β^* in time which is polynomial in $n, p, \log(\frac{1}{\sigma}), \log R, \log Q$, whp as p tends to infinity. As a corollary, we show that for any Q and R our algorithm can infer correctly β^*, when σ is at most exponential in $-(p^2/2 + (2+p)\log(QR))$, even from one observation ($n = 1$). We show that for general n our algorithm can tolerate noise level σ which is exponential in $-((2n+p)^2/2n + (2+p/n)\log(QR))$. We complement our results with the information-theoretic limits of our problem. We show that in the case of Gaussian white noise

W, a noise level which is exponential in $-\frac{p}{n}\log(QR)$, which is essentially the second part of our upper bound, cannot be tolerated. This allows us to conclude that in the regime $n = o(p/\log p)$ and $RQ = 2^{\omega(p)}$ our algorithm tolerates the optimal information theoretic level of noise.

The algorithm we propose receives as input real-valued data Y, X but importantly it truncates in the first step the data by keeping the first N bits after zero of every entry. In particular, this allows the algorithm to perform only **finite-precision** artihmetic operations. Here N is a parameter of our algorithm chosen by the algorithm designer. For our recovery results it is chosen to be polynomial in p and $\log(\frac{1}{\sigma})$.

A crucial step towards our main result is the extension of the Lagarias-Odlyzko algorithm [LO85], [Fri86] to not necessarily binary, integer vectors $\beta^* \in \mathbb{Z}^p$, for measurement matrix $X \in \mathbb{Z}^{n\times p}$ with iid entries not necessarily from the uniform distribution, and finally, for non-zero noise vector W. As in [LO85] and [Fri86], the algorithm we construct depends crucially on building an appropriate lattice and applying the LLL algorithm on it. There is though an important additional step in the algorithm presented in the present Chapter compared with the algorithm in [LO85] and [Fri86]. The latter algorithm is proven to recover a non-zero integer multiple $\lambda\beta^*$ of the underlying binary vector β^*. Then since β^* is known to be binary, the exact recovery becomes a matter of renormalizing out the factor λ from every non-zero coordinate. On the other hand, even if we establish in our case the corresponding result and recover a non-zero integer multiple of β^* whp, this last renormalizing step would be impossible as the ground truth vector is not assumed to be binary. We address this issue as follows. First we notice that the renormalization step remains valid if the greatest common divisor of the elements of β^* is 1. Under this assumption from any non-zero integer multiple of β^*, $\lambda\beta^*$ we can obtain the vector itself by observing that the greatest common divisor of $\lambda\beta^*$ equals to λ, and computing λ by using for instance the Euclid's algorithm. We then generalize our recovery guarantee to arbitrary β^*. We do this by first translating implicitly the vector β^* with a random integer vector Z via translating our observations $Y = X\beta^* + W$ by XZ to obtain $Y + XZ = X(\beta^* + Z) + W$. We then prove that the elements of $\beta^* + Z$ have greatest common divisor equal to unity with probability tending to one. This last step is based on an analytic number theory argument which slightly extends a beautiful result from probabilistic number theory (see for example, Theorem 332 in [HW75]) according to which $\lim_{m\to+\infty} \mathbb{P}_{P,Q\sim \mathrm{Unif}\{1,2,\ldots,m\}, P\perp\!\!\!\perp Q}\left[\gcd(P,Q) = 1\right] = \frac{6}{\pi^2}$, where

$P \perp\!\!\!\perp Q$ refers to P, Q being independent random variables. This result is not of clear origin in the literature, but possibly it is attributed to Chebyshev, as mentioned in [EL85]. A key implication of this result for us is the fact that the limit above is strictly positive.

Definitions and Notation

Let $\mathbb{Z}^*$ denote $\mathbb{Z} \setminus \{0\}$. For $k \in \mathbb{Z}_{>0}$ we set $[k] := \{1, 2, \ldots, k\}$. For a vector $x \in \mathbb{R}^d$ we define $\mathrm{Diag}_{d\times d}(x) \in \mathbb{R}^{d\times d}$ to be the diagonal matrix with $\mathrm{Diag}_{d\times d}(x)_{ii} = x_i,$ for $i \in [d]$. For $1 \leq p < \infty$ by $\mathcal{L}_p$ we refer to the standard p-norm notation for finite dimensionall real vectors. Given two vectors $x, y \in \mathbb{R}^d$ the Euclidean inner product notation is denoted by $\langle x, y\rangle := \sum_{i=1}^d x_i y_i$. By $\log : \mathbb{R}_{>0} \to \mathbb{R}$ we refer the logarithm with base 2. The lattice $\mathcal{L} \subseteq \mathbb{Z}^k$ generated by a set of linearly independent $b_1, \ldots, b_k \in \mathbb{Z}^k$ is defined as $\{\sum_{i=1}^k z_i b_i | z_1, z_2, \ldots, z_k \in \mathbb{Z}\}$. Throughout the Chapter we use the standard asymptotic notation, o, O, Θ, Ω for comparing the growth of two real-valued sequences $a_n, b_n, n \in \mathbb{Z}_{>0}$.Finally, we say that a sequence of events $\{A_p\}_{p\in\mathbb{N}}$ holds with high probability (whp) as $p \to +\infty$ if $\lim_{p\to+\infty} \mathbb{P}(A_p) = 1$.

5.2 Main Results

5.2.1 Extended Lagarias-Odlyzko algorithm

Let $n, p, R \in \mathbb{Z}_{>0}$. Given $X \in \mathbb{Z}^{n\times p}, \beta^* \in (\mathbb{Z} \cap [-R, R])^p$ and $W \in \mathbb{Z}^n$, set $Y = X\beta^* + W$. From the knowledge of Y, X the goal is to infer exactly β^*. For this task we propose the following algorithm which is an extension of the algorithm in [LO85] and [Fri86]. For realistic purposes the values of $R, \|W\|_\infty$ is not assumed to be known exactly. As a result, the following algorithm, besides Y, X, receives as an input a number $\hat{R} \in \mathbb{Z}_{>0}$ which is an estimated upper bound in absolute value for the entries of β^* and a number $\hat{W} \in \mathbb{Z}_{>0}$ which is an estimated upper bound in absolute value for the entries of W.

We explain here informally the steps of the (ELO) algorithm and briefly sketch the motivation behind each one of them. In the first and second steps the algorithm translates Y by XZ where Z is a random vector with iid elements chosen uniformly from $\{\hat{R}+1, \hat{R}+2, \ldots, 2\hat{R}+\log p\}$. In that way β^* is translated implicitly to $\beta = \beta^* + Z$ because $Y_1 = Y + XZ = X(\beta^* + Z) + W$. As

Algorithm 1 Extended Lagarias-Odlyzko (ELO) Algorithm

Input: $(Y, X, \hat{R}, \hat{W})$, $Y \in \mathbb{Z}^n, X \in \mathbb{Z}^{n\times p}$, $\hat{R}, \hat{W} \in \mathbb{Z}_{>0}$.

Output: $\hat{\beta^*}$ an estimate of β^*

1 Generate a random vector $Z \in \{\hat{R}+1, \hat{R}+2, \ldots, 2\hat{R}+\log p\}^p$ with iid entries uniform in $\{\hat{R}+1, \hat{R}+2, \ldots, 2\hat{R}+\log p\}$

2 Set $Y_1 = Y + XZ$.

3 For each $i = 1, 2, \ldots, n$, if $|(Y_1)_i| < 3$ set $(Y_2)_i = 3$ and otherwise set $(Y_2)_i = (Y_1)_i$.

4 Set $m = 2^{n+\lceil \frac{p}{2} \rceil + 3} p \left(\hat{R} \lceil \sqrt{p} \rceil + \hat{W} \lceil \sqrt{n} \rceil \right)$.

5 Output $\hat{z} \in \mathbb{R}^{2n+p}$ from running the LLL basis reduction algorithm on the lattice generated by the columns of the following $(2n+p) \times (2n+p)$ integer-valued matrix,

$$A_m := \begin{bmatrix} mX & -m\text{Diag}_{n\times n}(Y_2) & mI_{n\times n} \\ I_{p\times p} & 0_{p\times n} & 0_{p\times n} \\ 0_{n\times p} & 0_{n\times n} & I_{n\times n} \end{bmatrix} \tag{5.1}$$

6 Compute $g = \gcd(\hat{z}_{n+1}, \hat{z}_{n+2}, \ldots, \hat{z}_{n+p})$, using the Euclid's algorithm.

7 If $g \neq 0$, output $\hat{\beta^*} = \frac{1}{g}(\hat{z}_{n+1}, \hat{z}_{n+2}, \ldots, \hat{z}_{n+p})^t - Z$. Otherwise, output $\hat{\beta^*} = 0_{p\times 1}$.

we will establish using a number theoretic argument, $\gcd(\beta) = 1$ whp as $p \to +\infty$ with respect to the randomness of Z, even though this is not necessarily the case for the original β^*. This is an essential requirement for our technique to exactly recover β^* and steps six and seven to be meaningful. In the third step the algorithm gets rid of the significantly small observations. The minor but necessary modification of the noise level affects the observations in a negligible way.

The fourth and fifth steps of the algorithm provide a basis for a specific lattice in $2n+p$ dimensions. The lattice is built with the knowledge of the input and Y_2, the modified Y. The algorithm in step five calls the LLL basis reduction algorithm to run for the columns of A_m as initial basis for the lattice. The fact that Y has been modified to be non-zero on every coordinate is essential here so that A_m is full-rank and the LLL basis reduction algorithm, defined in [LLL82], can be applied,. This application of the LLL basis reduction algorithm is similar to the one used in [Fri86] with one important modification. In order to deal here with multiple equations and non-zero noise, we use $2n+p$ dimensions instead of $1+p$ in [Fri86]. Following though a similar strategy as in [Fri86], it can be established that the $n+1$ to $n+p$ coordinates of the output of the algorithm, $\hat{z} \in \mathbb{Z}^{2n+p}$, correspond to a vector which is a non-zero integer multiple of β, say $\lambda\beta$ for $\lambda \in \mathbb{Z}^*$, w.h.p. as $p \to +\infty$.

The proof of the above result is an important part in the analysis of the algorithm and it is

heavily based on the fact that the matrix A_m, which generates the lattice, has its first n rows multiplied by the "large enough" and appropriately chosen integer m which is defined in step four. It can be shown that this property of A_m implies that any vector z in the lattice with "small enough" $\mathcal{L}_2$ norm necessarily satisfies $(z_{n+1}, z_{n+2}, \ldots, z_{n+p}) = \lambda\beta$ for some $\lambda \in \mathbb{Z}^*$ whp as $p \to +\infty$. In particular, using that $\hat{z}$ is guaranteed to satisfy $\|\hat{z}\|_2 \leq 2^{\frac{2n+p}{2}}\|z\|_2$ for all non-zero z in the lattice, it can be derived that $\hat{z}$ has a "small enough" $\mathcal{L}_2$ norm and therefore indeed satisfies the desired property whp as $p \to +\infty$. Assuming now the validity of the $\gcd(\beta) = 1$ property, step six finds in polynomial time this unknown integer λ that corresponds to $\hat{z}$, because $\gcd(\hat{z}_{n+1}, \hat{z}_{n+2}, \ldots, \hat{z}_{n+p}) = \gcd(\lambda\beta) = \lambda$. Finally step seven scales out λ from every coordinate and then subtracts the known random vector Z, to output exactly β^*.

Of course the above is based on an informal reasoning. Formally we establish the following result.

Theorem 5.2.1. *Suppose*

(1) $X \in \mathbb{Z}^{n\times p}$ *is a matrix with iid entries generated according to a distribution* $\mathcal{D}$ *on* $\mathbb{Z}$ *which for some* $N \in \mathbb{Z}_{>0}$ *and constants* $C, c > 0$*, assigns at most* $\frac{c}{2^N}$ *probability on each element of* $\mathbb{Z}$ *and satisfies* $\mathbb{E}[|V|] \leq C2^N$*, for* $V \stackrel{d}{=} \mathcal{D}$*;*

(2) $\beta^* \in (\mathbb{Z} \cap [-R, R])^p$*,* $W \in \mathbb{Z}^n$*;*

(3) $Y = X\beta^* + W$*.*

Suppose furthermore that $\hat{R} \geq R$ *and*

$$N \geq \frac{1}{2n}(2n+p)\left[2n + p + 10\log\left(\hat{R}\sqrt{p} + (\|W\|_\infty + 1)\sqrt{n}\right)\right] + 6\log\left((1+c)\,np\right). \tag{5.2}$$

For any $\hat{W} \geq \|W\|_\infty$ *the algorithm ELO with input* $(Y, X, \hat{R}, \hat{W})$ *outputs* ***exactly*** β^* *w.p.* $1 - O\left(\frac{1}{np}\right)$ *(whp as* $p \to +\infty$*) and terminates in time at most polynomial in* $n, p, N, \log\hat{R}$ *and* $\log\hat{W}$*.*

We defer the proof to Section 5.4.

Remark 5.2.2. *In the statement of Theorem 5.2.1 the only parameters that are assumed to grow to infinity are* p *and whichever other parameters among* $n, R, \|W\|_\infty, N$ *are implied to grow to infinity because of (5.2). Note in particular that* n *can remain bounded, including the case* $n = 1$*, if* N *grows fast enough.*

Remark 5.2.3. *It can be easily checked that the assumptions of Theorem 5.2.1 are satisfied for* $n = 1$, $N = (1+\epsilon)\frac{p^2}{2}$, $R = 1$, $\mathcal{D} = \text{Unif}\{1, 2, 3, \ldots, 2^{(1+\epsilon)\frac{p^2}{2}}\}$ *and* $W = 0$. *Under these assumptions, the Theorem's implication is a generalization of the result from [LO85] and [Fri86] to the case* $\beta^* \in \{-1, 0, 1\}^p$.

5.2.2 Applications to High-Dimensional Linear Regression

The Model

We first define the Q-rationality assumption.

Definition 5.2.4. *Let* $p, Q \in \mathbb{Z}_{>0}$. *We say that a vector* $\beta \in \mathbb{R}^p$ *satisfies the* ***Q-rationality assumption*** *if for all* $i \in [p]$, $\beta_i^* = \frac{K_i}{Q}$, *for some* $K_i \in \mathbb{Z}$.

The high-dimensional linear regression model we are considering is as follows.

Assumptions 1. *Let* $n, p, Q \in \mathbb{Z}_{>0}$ *and* $R, \sigma, c > 0$. *Suppose*

(1) measurement matrix $X \in \mathbb{R}^{n \times p}$ *with iid entries generated according to a continuous distribution* $\mathcal{C}$ *which has density* f *with* $\|f\|_\infty \leq c$ *and satisfies* $\mathbb{E}[|V|] < +\infty$, *where* $V \stackrel{d}{=} \mathcal{C}$;

(2) ground truth vector β^* *satisfies* $\beta^* \in [-R, R]^p$ *and the* Q*-rationality assumption;*

(3) $Y = X\beta^* + W$ *for some noise vector* $W \in \mathbb{R}^n$. *It is assumed that either* $\|W\|_\infty \leq \sigma$ *or* W *has iid entries with mean zero and variance at most* σ^2, *depending on the context.*

Objective: Based on the knowledge of Y and X the goal is to recover β^* using an efficient algorithm and using the smallest number n of samples possible. The recovery should occur with high probability (w.h.p), as p diverges to infinity.

The Lattice-Based Regression (LBR) Algorithm

As mentioned in the Introduction, we propose an algorithm to solve the regression problem, which we call the Lattice-Based Regression (LBR) algorithm. The exact knowledge of $Q, R, \|W\|_\infty$ is not assumed. Instead the algorithm receives as an input, additional to Y and X, $\hat{Q} \in \mathbb{Z}_{>0}$ which is an estimated multiple of Q, $\hat{R} \in \mathbb{Z}_{>0}$ which is an estimated upper bound in absolute value

for the entries of β^* and $\hat{W} \in \mathbb{R}_{>0}$ which is an estimated upper bound in absolute value for the entries of the noise vector W. Furthermore an integer number $N \in \mathbb{Z}_{>0}$ is given to the algorithm as an input, which, as we will explain, corresponds to a truncation in the data in the first step of the algorithm. Given $x \in \mathbb{R}$ and $N \in \mathbb{Z}_{>0}$ let $x_N = \text{sign}(x)\frac{\lfloor 2^N|x| \rfloor}{2^N}$, which corresponds to the operation of keeping the first N bits after zero of a real number x.

Algorithm 2 Lattice Based Regression (LBR) Algorithm

Input: $(Y, X, N, \hat{Q}, \hat{R}, \hat{W})$, $Y \in \mathbb{Z}^n$, $X \in \mathbb{Z}^{n\times p}$ and $N, \hat{Q}, \hat{R}, \hat{W} \in \mathbb{Z}_{>0}$.
Output: $\hat{\beta}^*$ an estimate of β^*

8 Set $Y_N = ((Y_i)_N)_{i\in[n]}$ and $X_N = ((X_{ij})_N)_{i\in[n],j\in[p]}$.

9 Set $(\hat{\beta}_1)^*$ to be the output of the ELO algorithm with input:

$$\left(2^N\hat{Q}Y_N, 2^N X_N, \hat{Q}\hat{R}, 2\hat{Q}\left(2^N\hat{W} + \hat{R}p\right)\right).$$

10 Output $\hat{\beta}^* = \frac{1}{\hat{Q}}(\hat{\beta}_1)^*$.

We now explain informally the steps of the algorithm. In the first step, the algorithm truncates each entry of Y and X by keeping only its first N bits after zero, for some $N \in \mathbb{Z}_{>0}$. This in particular allows to perform finite-precision operations and to call the ELO algorithm in the next step which is designed for integer input. In the second step, the algorithm naturally scales up the truncated data to integer values, that is it scales Y_N by $2^N\hat{Q}$ and X_N by 2^N. The reason for the additional multiplication of the observation vector Y by $\hat{Q}$ is necessary to make sure the ground truth vector β^* can be treated as integer-valued. To see this notice that $Y = X\beta^* + W$ and Y_N, X_N being "close" to Y, X imply

$$2^N\hat{Q}Y_N = 2^N X_N(\hat{Q}\beta^*) + \text{ "extra noise terms"} + 2^N\hat{Q}W.$$

Therefore, assuming the control of the magnitude of the extra noise terms, by using the Q-rationality assumption and that $\hat{Q}$ is estimated to be a multiple of Q, the new ground truth vector becomes $\hat{Q}\beta^*$ which is integer-valued. The final step of the algorithm consist of rescaling now the output of Step 2, to an output which is estimated to be the original β^*. In the next subsection, we turn this discussion into a provable recovery guarantee.

Recovery Guarantees for the LBR algorithm

We state now our main result, explicitly stating the assumptions on the parameters, under which the LBR algorithm recovers **exactly** β^* from bounded but **adversarial noise** W.

Theorem 5.2.5.A. *Under Assumption 1 and assuming $W \in [-\sigma, \sigma]^n$ for some $\sigma \geq 0$, the following holds. Suppose $\hat{Q}$ is a multiple of Q, $\hat{R} \geq R$ and*

$$N > \frac{1}{2}(2n+p)\left(2n+p+10\log\hat{Q}+10\log\left(2^N\sigma+\hat{R}p\right)+20\log(3(1+c)np)\right). \tag{5.3}$$

For any $\hat{W} \geq \sigma$, the LBR algorithm with input $(Y, X, N, \hat{Q}, \hat{R}, \hat{W})$ terminates with $\hat{\beta}^* = \beta^*$ w.p. $1 - O\left(\frac{1}{np}\right)$ (whp as $p \to +\infty$) and in time polynomial in $n, p, N, \log\hat{R}, \log\hat{W}$ and $\log\hat{Q}$.

Applying Theorem 5.2.5.A we establish the following result handling **random noise** W.

Theorem 5.2.5.B. *Under Assumption 1 and assuming $W \in \mathbb{R}^n$ is a vector with iid entries generating according to an, independent from X, distribution $\mathcal{W}$ on $\mathbb{R}$ with mean zero and variance at most σ^2 for some $\sigma \geq 0$ the following holds. Suppose that $\hat{Q}$ is a multiple of Q, $\hat{R} \geq R$, and*

$$N > \frac{1}{2}(2n+p)\left(2n+p+10\log\hat{Q}+10\log\left(2^N\sqrt{np}\sigma+\hat{R}p\right)+20\log(3(1+c)np)\right). \tag{5.4}$$

For any $\hat{W} \geq \sqrt{np}\sigma$ the LBR algorithm with input $(Y, X, N, \hat{Q}, \hat{R}, \hat{W})$ terminates with $\hat{\beta}^* = \beta^*$ w.p. $1 - O\left(\frac{1}{np}\right)$ (whp as $p \to +\infty$) and in time polynomial in $n, p, N, \log\hat{R}, \log\hat{W}$ and $\log\hat{Q}$.

Both proofs of Theorems 5.2.5.A and 5.2.5.B are deferred to Section 5.5.

Noise tolerance of the LBR algorithm

The assumptions (5.2) and (5.4) might make it hard to build an intuition for the truncation level the LBR algorithm provably works. For this reason, in this subsection we *simplify it* and state a Proposition explicitly mentioning the optimal truncation level and hence characterizing the optimal level of noise that the LBR algorithm can tolerate with n samples.

First note that in the statements of Theorem 5.2.5.A and Theorem 5.2.5.B the only parameters that are assumed to grow are p and whichever other parameter is implied to grow because of (5.2) and (5.4). Therefore, importantly, n does not necessarily grow to infinity, if for example

$N, \frac{1}{\sigma}$ grow appropriately with p. That means that Theorem 5.2.5.A and Theorem 5.2.5.B imply non-trivial guarantees for *arbitrary sample size* n. The proposition below shows that if σ is at most exponential in $-(1+\epsilon)\left[\frac{(p+2n)^2}{2n} + (2+\frac{p}{n})\log(RQ)\right]$ for some $\epsilon > 0$, then for appropriately chosen truncation level N the LBR algorithm recovers exactly the vector β^* with n samples. In particular, with one sample ($n = 1$) LBR algorithm tolerates noise level up to exponential in $-(1+\epsilon)\left[p^2/2 + (2+p)\log(QR)\right]$ for some $\epsilon > 0$. On the other hand, if $n = \Theta(p)$ and $\log(RQ) = o(p)$, the LBR algorithm tolerates noise level up to exponential in $-O(p)$.

Proposition 5.2.6. *Under Assumption 1 and assuming $W \in \mathbb{R}^n$ is a vector with iid entries generating according to an, independent from X, distribution $\mathcal{W}$ on $\mathbb{R}$ with mean zero and variance at most σ^2 for some $\sigma \geq 0$, the following holds.*

Suppose $p \geq \frac{300}{\epsilon}\log\left(\frac{300}{(1+c)\epsilon}\right)$ and for some $\epsilon > 0$, $\sigma \leq 2^{-(1+\epsilon)\left[\frac{(p+2n)^2}{2n}+(2+\frac{p}{n})\log(RQ)\right]}$. Then the LBR algorithm with

- *input Y, X, $\hat{Q} = Q$, $\hat{R} = R$ and $\hat{W}_\infty = 1$ and*
- *truncation level N satisfying $\log\left(\frac{1}{\sigma}\right) \geq N \geq (1+\epsilon)\left[\frac{(p+2n)^2}{2n} + (2+\frac{p}{n})\log(RQ)\right]$,*

terminates with $\hat{\beta}^* = \beta^*$ w.p. $1 - O\left(\frac{1}{np}\right)$ (whp as $p \to +\infty$) and in time polynomial in $n, p, N, \log\hat{R}, \log\hat{W}$ and $\log\hat{Q}$.

The proof of Proposition 5.2.6 is deferred to Section 5.6.

It is worth noticing that in the noisy case ($\sigma > 0$) the above Proposition requires the truncation level N to be upper bounded by $\log(\frac{1}{\sigma})$, which implies the seemingly counter-intuitive conclusion that revealing more bits of the data after some point can "hurt" the performance of the recovery mechanism. Note that this is actually justified because of the presence of adversarial noise of magnitute σ. In particular, handling an arbitrary noise of absolute value at most of the order σ implies that the only bits of each observation that are certainly unaffected by the noise are the first $\log\left(\frac{1}{\sigma}\right)$ bits. Any bit in a later position could have potentially changed because of the noise. This correct middle ground for the truncation level N appears to be necessary also in the analysis of the synthetic experiments with the LBR algorithm (see Section 5.3).

Information Theoretic Bounds

In this subsection, we discuss the maximum noise that can be tolerated information-theoretically in recovering a $\beta^* \in [-R, R]^p$ satisfying the Q-rationality assumption. We establish that under Gaussian white noise, any successful recovery mechanism can tolerate noise level at most exponentially small in $-\left[p \log\left(QR\right)/n\right]$.

Proposition 5.2.7. *Suppose that $X \in \mathbb{R}^{n\times p}$ is a vector with iid entries following a continuous distribution $\mathcal{D}$ with $\mathbb{E}[|V|] < +\infty$, where $V \overset{d}{=} \mathcal{D}$, $\beta^* \in [-R, R]^p$ satisfies the Q-rationality assumption, $W \in \mathbb{R}^n$ has iid $N(0, \sigma^2)$ entries and $Y = X\beta^* + W$. Suppose furthermore that $\sigma > R(np)^3 \left(2^{\frac{2p\log(2QR+1)}{n}} - 1\right)^{-\frac{1}{2}}$. Then there is **no** mechanism which, whp as $p \to +\infty$, recovers **exactly** β^* with knowledge of Y, X, Q, R, σ. That is, for any function $\hat{\beta}^* = \hat{\beta}^*\left(Y, X, Q, R, \sigma\right)$ we have*

$$\limsup_{p\to+\infty} \mathbb{P}\left(\hat{\beta}^* = \beta^*\right) < 1.$$

The proof of Proposition 5.2.7 is deferred to Section 5.6.

Sharp Optimality of the LBR Algorithm

Using Propositions 5.2.6 and 5.2.7 the following **sharp** result is established.

Proposition 5.2.8. *Under Assumptions 1 where $W \in \mathbb{R}^n$ is a vector with iid $N(0, \sigma^2)$ entries the following holds. Suppose that $n = o\left(\frac{p}{\log p}\right)$ and $RQ = 2^{\omega(p)}$. Then for $\sigma_0 := 2^{-\frac{p\log(RQ)}{n}}$ and $\epsilon > 0$:*

- *if $\sigma > \sigma_0^{1-\epsilon}$,then the w.h.p. exact recovery of β^* from the knowledge of Y, X, Q, R, σ is impossible.*
- *if $\sigma < \sigma_0^{1+\epsilon}$, then the w.h.p. exact recovery of β^* from the knowledge of Y, X, Q, R, σ is possible by the LBR algorithm.*

The proof of Proposition 5.2.8 is deferred to Section 5.6.

5.3 Synthetic Experiments

In this section we present an experimental analysis of the ELO and LBR algorithms.

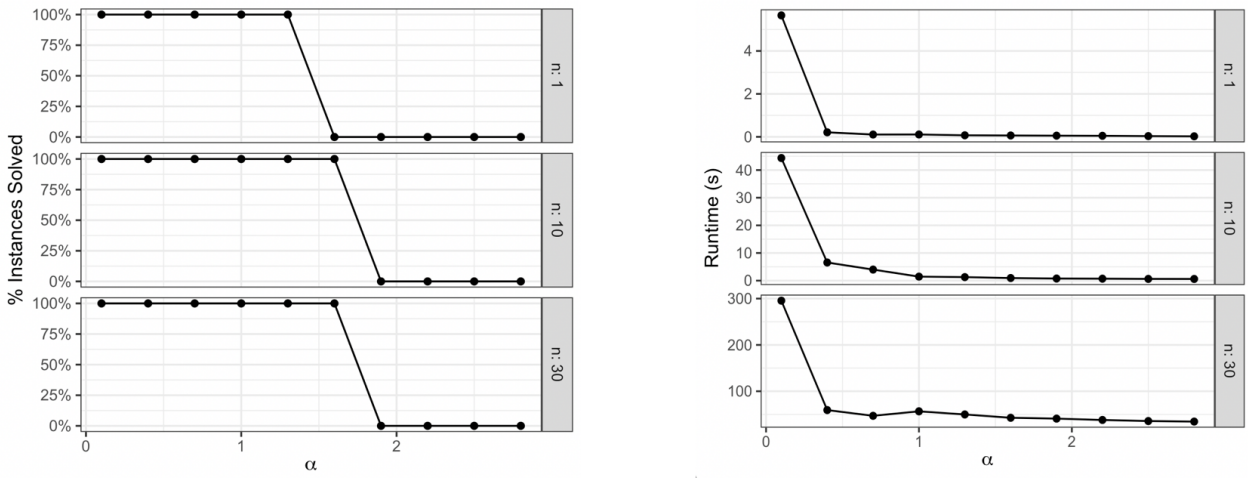


Figure 5-1: Average performance and runtime of ELO over 20 instances with $p = 30$ features and $n = 1, 10, 30$ samples.

ELO algorithm: We focus on $p = 30$ features sample sizes $n = 1, n = 10$ and $n = 30$, $R = 100$ and zero-noise $W = 0$. Each entry of β^* is iid Unif $(\{1, 2, \ldots, R = 100\})$. For 10 values of $\alpha \in (0, 3)$, specifically $\alpha \in \{0.25, 0.5, 0.75, 1, 1.3, 1.6, 1.9, 2.25, 2.5, 2.75\}$, we generate the entries of X iid Unif $\left(\{1, 2, 3, \ldots, 2^N\}\right)$ for $N = \frac{p^2}{2\alpha n}$. For each combination of n, α we generate 20 independent instances of inputs. We plot in Figure 1 the fractions of instances where the output of the ELO algorithm outputs exactly β^* and the average termination time of the algorithm.

Comments: First, we observe that importantly the algorithm recovers the vectors correctly on all $\alpha < 1$-instances with $p = 30$ features, even if our theoretical guarantees are only for large enough p. Second, Theorem 5.2.1 implies that if $N > (2n + p)^2 / 2n$ and large p, ELO recovers β^*, with high probability. In the experiments we observe that indeed ELO algorithm works in that regime, as then $\alpha = \frac{p^2}{2nN} < 1$. Also the experiments show that ELO works for larger values of α. Finally, the termination time of the algorithm was on average 1 minute and worst case 5 minutes, granting it reasonable for many applications.

LBR algorithm: We focus on $p = 30$ features, $n = 10$ samples, $Q = 1$ and $R = 100$. We generate each entry of β^* w.p. 0.5 equal to zero and w.p. 0.5, Unif $(\{1, 2, \ldots, R = 100\})$. We generate the entries of X iid $U(0, 1)$ and of W iid $U(-\sigma, \sigma)$ for $\sigma \in \{0, e^{-20}, e^{-12}, e^{-4}\}$. We generate 20 independent instances for any combination of σ and truncation level N. We plot the fraction of instances where the output of LBR algorithm is exactly β^*.

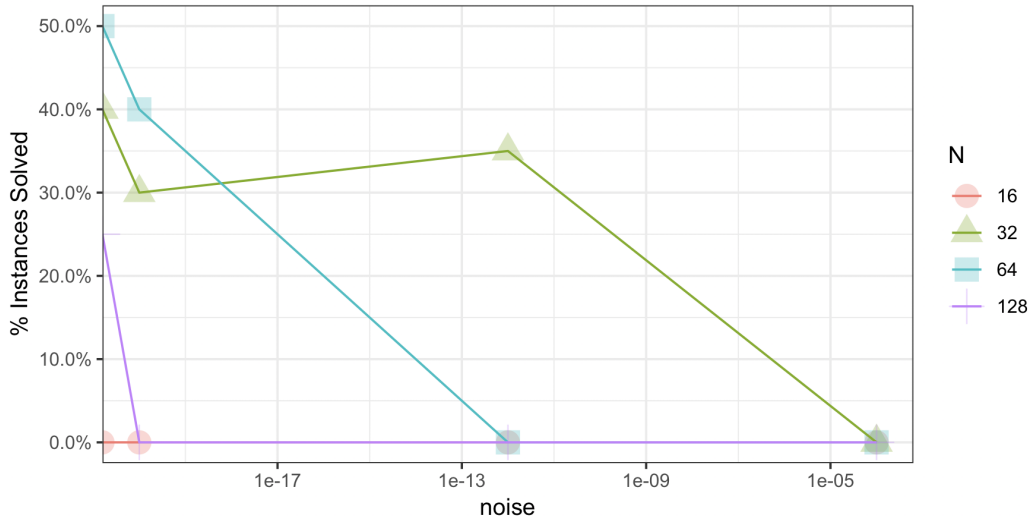


Figure 5-2: Average performance of LBR algorithm for various noise and truncation levels.

Comments: The experiments show that, first LBR works correctly in many cases for the moderate value of $p = 30$ and second that there is indeed an appropriate tuned truncation level $(2n + p)^2/2n < N < \log(1/\sigma)$ for which LBR succeeds. The latter is in exact agreement with Proposition 5.2.6.

5.4 Proof of Theorem 5.2.1

Proof. We first observe that directly from (5.2),

$$\begin{aligned} N &\geq 10 \log\left(\sqrt{p} + \sqrt{n}\left(\|W\|_\infty + 1\right)\right) \\ &\geq 5 \log\left(\sqrt{p}\sqrt{n}\left(\|W\|_\infty + 1\right)\right), \text{ from the elementary } a + b \geq \sqrt{ab} \\ &\geq 2 \log\left(pn\left(\|W\|_\infty + 1\right)\right). \end{aligned}$$

Therefore $2^N \geq \left(pn\left(1 + \|W\|_\infty\right)\right)^2$ which easily implies

$$\frac{\|W\|_\infty}{2^N} \leq \frac{1}{n^2 p^2} = \delta,$$

where we set for convenience $\delta = \delta_p := \frac{1}{n^2 p^2}$.

Lemma 5.4.1. *For all $i \in [n]$, $|(Y_2)_i| \geq \frac{3}{2}\delta 2^N$, w.p. at least $1 - O\left(\frac{1}{np}\right)$.*

Proof. First if $\delta 2^N < 2$, for all $i \in [n]$, $|(Y_2)_i| \geq 3 \geq \frac{3}{2}\delta 2^N$, because of the second step of the algorithm.

Assume now that $\delta 2^N \geq 2$. In that case first observe $Y_1 := Y + XZ = X(\beta^* + Z) + W$ and therefore from the definition of Y_2, $Y_2 = X(\beta^* + Z) + W_1$ for some $W_1 \in \mathbb{Z}^n$ with $\|W_1\|_\infty \leq \|W\|_\infty + 1$. Letting $\beta = \beta^* + Z$ we obtain that for all $i \in [n]$, $Y_i = \langle X^{(i)}, \beta \rangle + (W_1)_i$, where $X^{(i)}$ is the i-th row of X, and therefore

$$(Y_2)_i \geq |\sum_{j=1}^{p} X_{ij}\beta_j| - \|W_1\|_\infty \geq |\sum_{j=1}^{p} X_{ij}\beta_j| - \|W\|_\infty - 1.$$

Furthermore $\hat{R} \geq R$ implies $\beta \in [1, 3\hat{R} + \log p]^p$.

We claim that conditional on $\beta \in [1, 3\hat{R} + p]^p$ for all $i = 1, \ldots, n$, $|\sum_{j=1}^{p} X_{ij}\beta_j| \geq 3\delta 2^N$ w.p. at least $1 - O\left(\frac{1}{np}\right)$ with respect to the randomness of X. Note that this last inequality alongside with $\|W\|_\infty \leq \delta 2^N$ implies for all i, $|(Y_2)_i| \geq 2\delta 2^N - 1$. Hence since $\delta 2^N \geq 2$ we can conclude from the claim that for all i, $|(Y_2)_i| \geq \frac{3}{2}\delta 2^N$ w.p. at least $1 - O\left(\frac{1}{np}\right)$ Therefore it suffices to prove the claim to establish Lemma 5.4.1.

In order to prove the claim, observe that for large enough p,

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{n}\{|\sum_{j=1}^{p} X_{ij}\beta_j| < 3\delta 2^N\}\right) &\leq \sum_{i=1}^{n} \mathbb{P}\left(|\sum_{j=1}^{p} X_{ij}\beta_j| < 3\delta 2^N\right) \\
&= \sum_{i=1}^{n} \sum_{k \in \mathbb{Z} \cap [-3\delta 2^N, 3\delta 2^N]} \mathbb{P}\left(\sum_{j=1}^{p} X_{ij}\beta_j = k\right) \\
&\leq n(6\delta 2^N + 1)\frac{c}{2^N} \\
&\leq 7cn\delta = O\left(\frac{1}{np}\right),
\end{aligned}$$

where we have used that given $\beta_1 \neq 0$ for $i \in [p]$ and $k \in \mathbb{Z}$ the event $\{\sum_{j=1}^{p} X_{ij}\beta_j = k\}$ implies that the random variable X_{i1} takes a specific value, conditional on the realization of the remaining elements $X_{i2}, \ldots, X_{ip}$ involved in the equations. Therefore by our assumption on the iid distribution generating the entries of X, each of these events has probability at most $c/2^N$. Note that the choice of β_1, as opposed to choosing some β_i with $i > 1$, was arbitrary in the previous argument. The last inequality uses the assumption $\delta 2^N \geq 1$ and the final convergence

step is justified from $\delta = O(\frac{1}{n^2 p})$ and that c is a constant. $\square$

Next we use a number-theoretic lemma, which is an extension of a standard result in analytic number theory according to which

$$\lim_{m \to +\infty} \mathbb{P}_{P,Q \sim \text{Unif}\{1,2,\ldots,m\}, P \perp\!\!\!\perp Q} \left[\gcd\left(P, Q\right) = 1\right] = \frac{6}{\pi^2},$$

where $P \perp\!\!\!\perp Q$ refers to P, Q being independent random variables.

Lemma 5.4.2. *Suppose $q_1, q_2, q \in \mathbb{Z}_{>0}$ with $q \to +\infty$ and $\max\{q_1, q_2\} = o(q^2)$. Then*

$$|\{(a,b) \in \mathbb{Z}^2 \cap ([q_1, q_1 + q] \times [q_2, q_2 + q]) : \gcd(a,b) = 1\}| = q^2 \left(\frac{6}{\pi^2} + o_q(1)\right).$$

In other words, if we choose independently one uniform integer in $[q_1, q_1 + q]$ and another uniform integer in $[q_2, q_2 + q]$ the probability that these integers are relatively prime approaches $\frac{6}{\pi^2}$, as $q \to +\infty$.

Proof. We call an integer $n \in \mathbb{Z}_{>0}$ square-free if it is not divisible by the square of a positive integer number other than 1. The **Mobius function** $\mu : \mathbb{Z}_{>0} \to \{-1, 0, 1\}$ is defined to be

$$\mu(n) = \begin{cases} 1, & \text{n is square-free with an even number of prime factors} \\ -1, & \text{n is square-free with an odd number of prime factors} \\ 0, & \text{otherwise} \end{cases}$$

From now on we ease the notation by always referring for this proof to positive integer variables. A standard property for the Mobius function (see Theorem 263 in [HW75]) states that for all $n \in \mathbb{Z}_{>0}$,

$$\sum_{1 \le d \le n, d | n} \mu(d) = \begin{cases} 1, & n = 1 \\ 0, & \text{otherwise} \end{cases}$$

Therefore using the above identity and switching the order of summation we obtain

$$\begin{aligned}
&|(a,b) \in [q_1, q_1+q] \times [q_2, q_2+q], \gcd(a,b) = 1| \\
&= \sum_{(a,b)\in[q_1,q_1+q]\times[q_2,q_2+q]} \left(\sum_{1\le d\le \gcd(a,b), d|\gcd(a,b)} \mu(d) \right) \\
&= \sum_{1\le d\le \max\{q_1,q_2\}+q} \left(\sum_{(a,b)\in[q_1,q_1+q]\times[q_2,q_2+q], d|\gcd(a,b)} \mu(d) \right).
\end{aligned}$$

Now introducing the change of variables $a = kd, b = ld$ for some $k, l \in \mathbb{Z}_{>0}$ and observing that the number of integer numbers in an interval of length $x > 0$ are $x + O(1)$, we obtain

$$\begin{aligned}
&\sum_{1\le d\le \max\{q_1,q_2\}+q} \left(\sum_{\frac{q_1}{d}\le k\le \frac{q_1+q}{d}, \frac{q_2}{d}\le l\le \frac{q_2+q}{d}} \mu(d) \right) \\
&= \sum_{1\le d\le \max\{q_1,q_2\}+q} \left[\left(\frac{q}{d} + O(1) \right)^2 \mu(d) \right] \\
&= \sum_{1\le d\le \max\{q_1,q_2\}+q} \left[\left(\frac{q}{d} \right)^2 \mu(d) + O\left(\frac{q}{d} \right) \mu(d) + O(1)\mu(d) \right]
\end{aligned}$$

Now using $|\mu(d)| \le 1$ for all $d \in \mathbb{Z}_{>0}$, for $n \in \mathbb{Z}_{>0}$,

$$\sum_{d=1}^{n} \frac{1}{d} = O(\log n)$$

and that by Theorem 287 in [HW75] for $n \in \mathbb{Z}_{>0}$,

$$\sum_{d=1}^{n} \frac{\mu(d)}{d^2} = \frac{1}{\zeta(2)} + o_n(1) = \frac{6}{\pi^2} + o_n(1)$$

we conclude that the last quantity equals

$$q^2 \left(\frac{6}{\pi^2} + \frac{1}{q} O(\log(\max\{q_1, q_2\} + q)) + \frac{\max\{q_1, q_2\} + q}{q^2} + o_q(1) \right).$$

Recalling the assumption $q_1, q_2 = o(q^2)$ the proof is complete. □

Claim 5.4.3. *The greatest common divisor of the coordinates of $\beta := \beta^* + Z$ equals to 1, w.p. $1 - \exp(-\Theta(p))$ with respect to the randomness of Z.*

Proof. Each coordinate of β is a uniform and independent choice of a positive integer from an interval of length $2\hat{R}+\log p$ with starting point in $[\hat{R}-R+1, \hat{R}+R+1]$, depending on the value of $\beta_i^* \in [-R, R]$. Note though that Lemma 5.4.2 applies for arbitrary $q_1, q_2 \in [\hat{R}-R+1, \hat{R}+R+1]$ and $q = 2\hat{R} + \log p$ since $q_1, q_2 = o(q^2)$ and $q \to +\infty$. from this we conclude that the probability any two specific coordinates of β have greatest common divisor 1 approaches $\frac{6}{\pi^2}$, as $p \to +\infty$. But the probability the greatest common divisor of all the coordinates is not one implies that the greatest common divisor of the $2i-1$ and $2i$ coordinate is not one, for every $i = 1, 2, \ldots \lfloor \frac{p}{2} \rfloor$. Hence using the independence among the values of the coordinates, we conclude that the greatest common divisor of the coordinates of β is not one with probability at most

$$\left(1 - \frac{6}{\pi^2} + o_p(1)\right)^{\lfloor \frac{p}{2} \rfloor} = \exp(-\Theta(p)).$$

□

Given a vector $z \in \mathbb{R}^{2n+p}$, define $z_{n+1:p} := (z_{n+1}, \ldots, z_{n+p})^t$.

Claim 5.4.4. *The outcome of Step 5 of the algorithm, $\hat{z}$, satisfies*

- $\|\hat{z}\|_2 < m$
- $\hat{z}_{n+1:n+p} = q\beta$, *for some* $q \in \mathbb{Z}^*$, *w.p.* $1 - O\left(\frac{1}{np}\right)$.

Proof. Call $\mathcal{L}_m$the lattice generated by the columns of the $(2n+p) \times (2n+p)$ integer-valued matrix A_m defined in the algorithm; that is $\mathcal{L}_m := \{A_m z | z \in \mathbb{Z}^{2n+p}\}$. Notice that as Y_2 is nonzero at every coordinate, the lattice $\mathcal{L}_m$ is full-dimensional and the columns of A_m define a basis for $\mathcal{L}_m$. Finally, an important vector in $\mathcal{L}_m$ for our proof is $z_0 \in \mathcal{L}_m$ which is defined for $1_n \in \mathbb{Z}^n$ the all-ones vector as

$$z_0 := A_m \begin{bmatrix} \beta \\ 1_n \\ W_1 \end{bmatrix} = \begin{bmatrix} 0_{n\times 1} \\ \beta \\ W_1 \end{bmatrix} \in \mathcal{L}_m. \tag{5.5}$$

Consider the following optimization problem on $\mathcal{L}_m$, known as the shortest vector problem,

$$\begin{aligned} (\mathcal{S}_2) \quad & \min \quad \|z\|_2 \\ & \text{s.t.} \quad z \in \mathcal{L}_m, \end{aligned}$$

If z^* is the optimal solution of $(\mathcal{S}_2)$ we obtain

$$\|z^*\|_2 \leq \|z_0\|_2 = \sqrt{\|\beta\|_2^2 + \|W_1\|_2^2} \leq \|\beta\|_\infty \sqrt{p} + \|W_1\|_\infty \sqrt{n}.$$

and therefore given our assumptions on β, W

$$\|z^*\|_2 \leq \left(3\hat{R} + \log p\right)\sqrt{p} + \left(\|W\|_\infty + 1\right)\sqrt{n}.$$

Using that $\hat{R} \geq 1$ and a crude bound this implies

$$\|z^*\|_2 \leq 4p\left(\hat{R}\sqrt{p} + \left(\|W\|_\infty + 1\right)\sqrt{n}\right).$$

The LLL guarantee and the above observation imply that

$$\|\hat{z}\|_2 \leq 2^{\frac{2n+p}{2}}\|z^*\|_2 \leq 2^{\frac{2n+p}{2}+2}p\left(\hat{R}\sqrt{p} + \left(\|W\|_\infty + 1\right)\sqrt{n}\right) := m_0. \tag{5.6}$$

Now recall that $\hat{W}_\infty \geq \max\{\|W\|_\infty, 1\}$. Since $m \geq 2^{n+\frac{p}{2}+3}p\left(\hat{R}\sqrt{p} + \hat{W}_\infty\sqrt{n}\right)$, we obtain $m > m_0$ and hence $\|\hat{z}\|_2 < m$. This establishes the first part of the Claim.

For the second part, given (5.6) and that $\hat{z}$ is non-zero it suffices to establish that under the conditions of our Theorem there is no non-zero vector in $\mathcal{L}_m \backslash \{z \in \mathcal{L}_m | z_{n+1:n+p} = q\beta, q \in \mathbb{Z}^*\}$ with L_2 norm less than m_0, w.p. $1 - O\left(\frac{1}{np}\right)$. By construction of the lattice for any $z \in \mathcal{L}_m$ there exists an $x \in \mathbb{Z}^{2n+p}$ such that $z = A_m x$. We decompose $x = (x_1, x_2, x_3)^t$ where $x_1 \in \mathbb{Z}^p, x_2, x_3 \in \mathbb{Z}^n$. It must be true

$$z = \begin{bmatrix} m\left(Xx_1 - \text{Diag}_{n\times n}(Y)x_2 + x_3\right) \\ x_1 \\ x_3 \end{bmatrix}.$$

Note that $x_1 = z_{n+1:n+p}$. We use this decomposition of every $z \in \mathcal{L}_m$ to establish our result.

We first establish that for any lattice vector $z \in \mathcal{L}_m$ the condition $\|z\|_2 \leq m_0$ implies necessarily

$$Xx_1 - \mathrm{Diag}_{n\times n}(Y)x_2 + x_3 = 0. \tag{5.7}$$

and in particular $z = (0, x_1, x_3)$. If not, as it is an integer-valued vector, $\|Xx_1 - \mathrm{Diag}_{n\times n}(Y)x_2 + x_3||_2 \geq 1$ and therefore

$$m \leq m\|Xx_1 - \mathrm{Diag}_{n\times n}(Y)x_2 + x_3\|_2 \leq \|z\|_2 \leq m_0,$$

a contradiction as $m > m_0$. Hence, necessarily equation (5.7) and $z = (0, x_1, x_3)$ hold.

Now we claim that it suffices to show that there is no non-zero vector in $\mathcal{L}_m \backslash \{z \in \mathcal{L}_m | z_{n+1:n+p} = q\beta, q \in \mathbb{Z}\}$ with L_2 norm less than m_0, w.p. $1 - O\left(\frac{1}{np}\right)$. Note that in this claim the coefficient q is allowed to take the zero value as well. The reason it suffices to prove this weaker statement is that any non-zero $z \in \mathcal{L}_m$ with $\|z\|_2 \leq m_0$ necessarily satisies that $z_{n+1:n+p} \neq 0$ w.p. $1 - O\left(\frac{1}{np}\right)$ and therefore the case $q = 0$ is not possible w.p. $1 - O\left(\frac{1}{np}\right)$. To see this, we use the decomposition and recall that $x_1 = z_{n+1:n+p}$. Therefore it suffices to establish that there is no triplet $x = (0, x_2, x_3)^t \in \mathbb{Z}^{2n+p}$ with $x_2, x_3 \in \mathbb{Z}^n$ for which the vector $z = A_m x \in \mathcal{L}_m$ is non-zero and $\|z\|_2 \leq m_0$, w.p. $1 - O\left(\frac{1}{np}\right)$. To prove this, we consider such a triplet $x = (0, x_2, x_3)$ and will upper bound the probability of its existence. From equation (5.7) it necessarily holds $\mathrm{Diag}_{n\times n}(Y)x_2 = x_3$, or equivalently

$$\text{for all } i \in [n], Y_i(x_2)_i = (x_3)_i. \tag{5.8}$$

From Lemma 5.4.1 and (5.8) we obtain that

$$\text{for all } i \in [n], \frac{3}{2}\delta 2^N |(x_2)_i| \leq |(x_3)_i| \tag{5.9}$$

w.p. $1 - O\left(\frac{1}{np}\right)$. Since z is assumed to be non-zero and $z = A_m x = (0, 0, x_3)$ there exists $i \in [n]$ with $(x_3)_i \neq 0$. Using (5.8) we obtain $(x_2)_i \neq 0$ as well. Therefore for this value of i it must be simultaneously true that $|(x_2)_i| \geq 1$ and $|(x_3)_i| \leq m_0$. Plugging these inequalities to (5.9) for

this value of i, we conclude that it necessarily holds that

$$\frac{3}{2}\delta 2^N \leq m_0$$

Using the definition of δ, $\delta = \frac{1}{n^2p^2}$, we conclude that it must hold $\frac{1}{n^2p^2}2^N \leq m_0$, or

$$N \leq 2\log(np) + \log m_0.$$

Plugging in the value of m_0 we conclude that for sufficiently large p,

$$N \leq 2\log(np) + \frac{2n+p}{2} + \log p + \log\left(\hat{R}\sqrt{p} + (\|W\|_\infty + 1)\sqrt{n}\right).$$

This can be checked to contradict directly our hypothesis (5.2) and the proof of the claim is complete.

Therefore using the decomposition of every $z \in \mathcal{L}_m$, equation (5.7) and the claim in the last paragraph it suffices to establish that w.p. $1 - O\left(\frac{1}{np}\right)$ there is no triplet (x_1, x_2, x_3) with

(a) $x_1 \in \mathbb{Z}^p, x_2, x_3 \in \mathbb{Z}^n$;

(b) $\|x_1\|_2^2 + \|x_3\|_2^2 \leq m_0$;

(c) $Xx_1 - \mathrm{Diag}_{n\times n}(Y)x_2 - x_3 = 0$;

(d) $\forall q \in \mathbb{Z} : x_1 \neq q\beta$.

We first claim that any such triplet (x_1, x_2, x_3) satifies w.p. $1 - O\left(\frac{1}{np}\right)$

$$\|x_2\|_\infty = O(\frac{m_0 n^2 p^3}{\delta}).$$

To see this let $i = 1, 2, \ldots, n$ and denote by $X^{(i)}$ the i-th row of X. We have because of (c),

$$0 = (Xx_1 - \mathrm{Diag}_{n\times n}(Y)x_2 - x_3)_i = \langle X^{(i)}, x_1\rangle - Y_i(x_2)_i - (x_3)_i,$$

and therefore by triangle inequality

$$|Y_i(x_2)_i| = |\langle X^{(i)}, x_1\rangle - (x_3)_i| \leq |\langle X^{(i)}, x_1\rangle| + |(x_3)_i|. \tag{5.10}$$

But observe that for all $i \in [n], \|X^{(i)}\|_\infty \leq \|X\|_\infty \leq (np)^2 2^N$ w.p. $1 - O\left(\frac{1}{np}\right)$. Indeed using a union bound, Markov's inequality and our assumption on the distribution $\mathcal{D}$ of the entries of X,

$$\mathbb{P}\left(\|X\|_\infty > (np)^2 2^N\right) \leq np\mathbb{P}\left(|X_{11}| > (np)^2 2^N\right) \leq \frac{1}{2^N np}\mathbb{E}[|X_{11}|] \leq \frac{C}{np} = O\left(\frac{1}{np}\right),$$

which establishes the result. Using this, Lemma 5.4.1 and (5.10) we conclude that for all $i \in [n]$ w.p. $1 - O\left(\frac{1}{np}\right)$

$$|(x_2)_i|\frac{3}{2}\delta 2^N \leq (2^N p(np)^2 + 1)m_0$$

which in particular implies

$$|(x_2)_i| \leq O(\frac{m_0 n^2 p^3}{\delta}),$$

w.p. $1 - O\left(\frac{1}{np}\right)$.

Now we claim that for any such triplet (x_1, x_2, x_3) it also holds

$$\mathbb{P}\left(Xx_1 - \mathrm{Diag}_{n\times n}(Y)x_2 - x_3 = 0\right) \leq \frac{c^n}{2^{nN}}. \tag{5.11}$$

To see this note that for any $i \in [n]$ if $X^{(i)}$ is the i-th row of X because $Y = X\beta + W$ it holds $Y_i = \langle X^{(i)}, \beta\rangle + W_i$. In particular, $Xx_1 - \mathrm{Diag}_{n\times n}(Y)x_2 - x_3 = 0$ implies for all $i \in [n]$,

$$\begin{aligned}
&\langle X^{(i)}, x_1\rangle - Y_i(x_2)_i = (x_3)_i \\
\text{or } &\langle X^{(i)}, x_1\rangle - \left(\langle X^{(i)}, \beta\rangle + W_i\right)(x_2)_i = (x_3)_i \\
\text{or } &\langle X^{(i)}, x_1 - (x_2)_i\beta\rangle = (x_3)_i - (x_2)_i W_i
\end{aligned}$$

Hence using independence between rows of X,

$$\mathbb{P}\left(Xx_1 - \mathrm{Diag}_{n\times n}(Y)x_2 - x_3 = 0\right) = \prod_{i=1}^{n}\mathbb{P}\left(\langle X^{(i)}, x_1 - (x_2)_i\beta\rangle = (x_3)_i - (x_2)_i W_i\right) \tag{5.12}$$

But because of (d) for all i, $x_1 - (x_2)_i\beta \neq 0$. In particular, $\langle X^{(i)}, x_1 - (x_2)_i\beta\rangle = (x_3)_i - (x_2)_i W_i$ constraints at least one of the entries of $X^{(i)}$ to get a specific value with respect to the rest of the elements of the row which has probability at most $\frac{c}{2^N}$ by the independence assumption on the entries of X. This observation with (5.12) implies (5.11).

Now, we establish that indeed there are no such triplets, w.p. $1 - O\left(\frac{1}{np}\right)$. Recall the standard fact that for any $r > 0$ there are at most $O(r^n)$ vectors in $\mathbb{Z}^n$ with L_∞-norm at most r. Using this, (5.11) and a union bound over all the integer vectors (x_1, x_2, x_3) with $\|x_1\|_2^2 + \|x_3\|_2^2 \leq m_0$, $\|x_2\|_\infty = O(\frac{m_0 n^2 p^3}{\delta})$ we conclude that the probability that there exist a triplet (x_1, x_2, x_3) satisfying (a), (b), (c), (d) is at most of the order

$$(\frac{m_0 n^2 p^3}{\delta})^n m_0^{n+p} \left[\frac{c^n}{2^{nN}}\right].$$

Plugging in the value of m_0 we conclude that the probability is at most of the order

$$\frac{2^{\frac{1}{2}(2n+p)^2 + n\log(cn^2p^3) + n\log(\frac{1}{\delta}) + (2+\log p)(2n+p)} \left[\hat{R}\sqrt{p} + (\|W\|_\infty + 1)\sqrt{n}\right]^{2n+p}}{2^{nN}}.$$

Now recalling that $\delta = \frac{1}{n^2p^2}$ we obtain $\log(\frac{1}{\delta}) = 2\log(np)$ and therefore the last bound becomes at most of the order

$$\frac{2^{\frac{1}{2}(2n+p)^2 + 5n\log(cnp) + (2+\log p)(2n+p)} \left[\hat{R}\sqrt{p} + (\|W\|_\infty + 1)\sqrt{n}\right]^{2n+p}}{2^{nN}}.$$

We claim that the last quantity is $O\left(\frac{1}{np}\right)$ because of our assumption (5.2). Indeed the logarithm of the above quantity equals

$$\frac{1}{2}(2n+p)\left(2n + p + 4 + 2\log p + 2\log\left(\hat{R}\sqrt{p} + (\|W\|_\infty + 1)\sqrt{n}\right)\right) + 5n\log(cnp) - nN.$$

Using that $\hat{R} \geq 1$ this is upper bounded by

$$\frac{1}{2}(2n+p)\left(2n + p + 10\log\left(R\sqrt{p} + (\|W\|_\infty + 1)\sqrt{n}\right)\right) + 5n\log(cnp) - nN$$

which by our assumption (5.2) is indeed less than $-n\log(np) < -\log(np)$, implying the desired bound. This completes the proof of claim 5.4.4. $\square$

Now we prove Theorem 5.2.1. First with respect to time complexity, it suffices to analyze Step 5 and Step 6. For step 5 we have from [LLL82] that it runs in time polynomial in $n, p, \log \|A_m\|_\infty$ which indeed is polynomial in n, p, N and $\log \hat{R}, \log \hat{W}$. For step 6, recall that the Euclid algorithm to compute the greatest common divisor of p numbers with norm bounded by $\|\hat{z}\|_\infty$ takes time which is polynomial in $p, \log \|\hat{z}\|_\infty$. But from Claim 5.4.4 we have that $\|\hat{z}\|_\infty < m$ and therefore the time complexity is polynomial in $p, \log m$ and therefore again polynomial in n, p, N and $\log \hat{R}, \log \hat{W}$.

Finally we prove that the ELO algorithm outputs exactly β^* w.p. $1 - O\left(\frac{1}{np}\right)$. We obtain from Claim 5.4.4 that $\hat{z}_{n+1:n+p} = q\beta$ for $\beta = \beta^* + Z$ and some $q \in \mathbb{Z}^*$ w.p. $1 - O\left(\frac{1}{np}\right)$. We claim that the g computed in Step 6 is this non-zero integer q w.h.p. To see it notice that from Claim 5.4.3 $\gcd(\beta) = 1$ w.p. $1 - \exp(-\Theta(p)) = 1 - O\left(\frac{1}{np}\right)$ and therefore the g computed in Step 6 satisfies w.p. $1 - O\left(\frac{1}{np}\right)$,

$$g = \gcd(\hat{z}_{n+1:n+p}) = \gcd(q\beta) = q\gcd(\beta) = q.$$

Hence we obtain w.p. $1 - O\left(\frac{1}{np}\right)$.

$$\hat{z}_{n+1:n+p} = g\beta = g\left(\beta^* + Z\right)$$

or w.p. $1 - O\left(\frac{1}{np}\right)$

$$\beta^* = \frac{1}{g}\hat{z}_{n+1:n+p} - Z,$$

which implies based on Step 7 and the fact that $g = q \neq 0$ that indeed the output of the algorithm is β^* w.p. $1 - O\left(\frac{1}{np}\right)$. The proof of Theorem 5.2.1 is complete. $\square$

5.5 Proofs of Theorems 5.2.5.A and 5.2.5.B

Proof of Theorem 5.2.5.A

Proof. We first analyze the algorithm with respect to time complexity. It suffices to analyze step 2 as step 1 runs clearly in polynomial time N, n, p. Step 2 runs the ELO algorithm. From Theorem 5.2.1 we obtain that the ELO algorithm terminates in polynomial time in $n, p, N, \log\left(\hat{Q}\hat{R}\right), \log\left(2\hat{Q}\left(2^N\hat{W} + \hat{R}p\right)\right)$. As the last quantity is indeed polynomial in n, p, N and $\log\hat{R}, \log\hat{Q}, \log\hat{W}$, we are done.

Now we prove that $\hat{\beta}^* = \beta^*$, w.p. $1 - O\left(\frac{1}{np}\right)$. Notice that it suffices to show that the output of Step 3 of the LBR algorithm is exactly $\hat{Q}\beta^*$, as then step 4 gives $\hat{\beta}^* = \frac{Q\beta^*}{Q} = \beta^*$ w.p. $1 - O\left(\frac{1}{np}\right)$.

We first establish that

$$2^N\hat{Q}Y_N = 2^N X_N\hat{Q}\beta^* + W_0 \tag{5.13}$$

for some $W_0 \in \mathbb{Z}^n$ with $\|W_0\|_\infty + 1 \leq 2\hat{Q}\left(2^N\sigma + Rp\right)$. We have $Y = X\beta^* + W$, with $\|W\|_\infty \leq \sigma$. From the way Y_N is defined, $\|Y - Y_N\|_\infty \leq 2^{-N}$. Hence for $W' = W + Y_N - Y$ which satisfies $\|W'\|_\infty \leq 2^{-N} + \sigma$ we obtain

$$Y_N = X\beta^* + W'.$$

Similarly since $\|X - X_N\|_\infty \leq 2^{-N}$ and $\|\beta^*\|_\infty \leq R$ we obtain $\|\left(X - X_N\right)\beta^*\|_\infty \leq 2^{-N}Rp$, and therefore for $W'' = W' + \left(X - X_N\right)\beta^*$ which satisfies $\|W''\|_\infty \leq 2^{-N} + \sigma + 2^{-N}rp$ we obtain,

$$Y_N = X_N\beta^* + W''$$

or equivalently

$$2^N Y_N = 2^N X_N\beta^* + W''',$$

where $W''' := 2^N W''$ which satisfies $\|W'''\|_\infty \leq 1 + 2^N\sigma + Rp$. Multiplying with $\hat{Q}$ we obtain

$$2^N\hat{Q}Y_N = 2^N X_N\left(\hat{Q}\beta^*\right) + W_0,$$

where $W_0 := \hat{Q}W'''$ which satisfies $\|W_0\|_\infty \leq \hat{Q}\left(1 + 2^N\sigma + Rp\right) \leq 2\hat{Q}\left(2^N\sigma + Rp\right) - 1$. This

establishes equation (5.13).

We now apply Theorem 5.2.1 for Y our vector $\hat{Q}2^N Y_N$, X our vector $2^N X_N$, β^* our vector $\hat{Q}\beta^*$, W our vector W_0, R our $\hat{Q}R$, $\hat{R}$ our $\hat{Q}\hat{R}$, $\hat{W}$ our quantity $2\hat{Q}\left(2^N\sigma + Rp\right)$ and finally N our truncation level N.

We fist check the assumption (1), (2), (3) of Theorem 5.2.1. We start with assumption (1). From the definition of X_N we have that $2^N X_N \in \mathbb{Z}^{n\times p}$ and that for all $i \in [n], j \in [p]$,

$$|(2^N X_N)_{ij}| \leq 2^N |X_{ij}|.$$

Therefore for $C = \mathbb{E}[|X_{1,1}|] < \infty$ and arbitrary $i \in [n], j \in [p]$,

$$\mathbb{E}[|(2^N X_N)_{ij}|] \leq 2^N \mathbb{E}[|X_{ij}|] = C2^N,$$

as we wanted. Furthermore, if f is the density function of the distribution $\mathcal{D}$ of the entries of X, recall $\|f\|_\infty \leq c$, by our hypothesis. Now observe for arbitrary $i \in [n], j \in [p]$,

$$\mathbb{P}\left((2^N X_N)_{ij} = k\right) = \mathbb{P}\left(\frac{k}{2^N} \leq X_{ij} \leq \frac{k+1}{2^N}\right) = \int_{\frac{k}{2^N}}^{\frac{k+1}{2^N}} f(u)du \leq \|f\|_\infty \int_{\frac{k}{2^N}}^{\frac{k+1}{2^N}} du \leq \frac{c}{2^N}.$$

This completes the proof that $2^N X_N$ satisfies assumption (1) of Theorem 5.2.1. For assumption (2), notice that $\hat{Q}\beta^*$ is integer valued, as $\hat{Q}$ is assumed to be a mutliple of Q and β^* satisfies Q-rationality. Furthermore clearly

$$\|\hat{Q}\beta^*\|_\infty \leq \hat{Q}R.$$

For the noise level we have by (5.13) $W_0 = 2^N \hat{Q}Y_N - 2^N X_N \hat{Q}\beta^*$ and therefore $W_0 \in \mathbb{Z}^n$ as all the quantities $2^N\hat{Q}Y_N$, $2^N X_N$ and $\hat{Q}\beta^*$ are integer-valued. Finally, Assumption (3) follows exactly from equation (5.13).

Now we check the parameters assumptions of Theorem 5.2.1. We clearly have

$$\hat{Q}R \leq \hat{Q}\hat{R}$$

and

$$\|W\|_\infty \leq 2\hat{Q}\left(2^N\sigma + Rp\right) = \hat{W}.$$

The last step consists of establishing the relation (5.2) of Theorem 5.2.5.A. Plugging in our parameter choice it suffices to prove

$$N > \frac{(2n+p)}{2}\left(2n+p+10\log\left(\hat{Q}\hat{R}\sqrt{p}+2\hat{Q}\left(2^N\sigma+Rp\right)\sqrt{n}\right)\right)+6n\log((1+c)\,np).$$

Using that $\hat{Q}R\sqrt{p} \leq \hat{Q}\left(2^N\sigma+\hat{R}p\right)\sqrt{n}$ and $R \leq \hat{R}$ it suffices to show after elementary algebraic manipulations that

$$N > \frac{(2n+p)}{2}\left(2n+p+10\log 3+10\log\hat{Q}+10\log\left(2^N\sigma+\hat{R}p\right)+5\log n\right)+6n\log((1+c)\,np).$$

Using now that by elementary considerations

$$\frac{(2n+p)}{2}\left(10\log 3+5\log n\right)+4n\log((1+c)\,np) < \frac{(2n+p)}{2}[20\log(3\,(1+c)\,np)] \text{ for all } n \in \mathbb{Z}_{>0},$$

it suffices to show

$$N > \frac{(2n+p)}{2}\left(2n+p+10\log\hat{Q}+10\log\left(2^N\sigma+\hat{R}p\right)+20\log(3\,(1+c)\,np)\right),$$

which is exactly assumption (5.3).

Hence, the proof that we can apply Theorem 5.2.1 is complete. Applying it we conclude that w.p. $1-O\left(\frac{1}{np}\right)$ the output of LBR algorithm at step 3 is $\hat{Q}\beta^*$, as we wanted. □

Proof of Theorem 5.2.5.B

By using a standard union bound and Markov inequality we have

$$\mathbb{P}\left(\|W\|_\infty \leq \sqrt{np}\sigma\right) \geq 1-\sum_{i=1}^{n}\mathbb{P}\left(|W_i| > \sqrt{np}\sigma\right) \geq 1-n\frac{\mathbb{E}\left[W_1^2\right]}{np\sigma^2} \geq 1-\frac{1}{p}.$$

Therefore, conditional on the high probability event $\|W\|_\infty \leq \sqrt{np}\sigma$, we can apply Theorem 5.2.5.A with $\sqrt{np}\sigma$ instead of σ and conclude the result.

5.6 Rest of the Proofs

Proof of Proposition 5.2.6

Proof. If we show that we can apply Theorem 5.2.5.B, the result follows. Since the model assumptions are identical we only need to check the parameter assumptions of Theorem 5.2.5.B. First note that we assume $\hat{R} = R$, we clearly have for the noise $\sigma \leq W_\infty = 1$ and finally $\hat{Q} = Q$. Now for establishing 5.4, we first notice that since $N \leq \log\left(\frac{1}{\sigma}\right)$ is equivalent to $2^N\sigma \leq 1$, we obtain $2^N\sigma\sqrt{np} + Rp \leq 2^{\log(np)+\log(Rp)}$. Therefore it suffices

$$N > \frac{(2n+p)^2}{2n} + 22\frac{2n+p}{n}\log(3(1+c)np) + \frac{2n+p}{n}\log(RQ)$$

Now since $p \geq \frac{300}{\epsilon}\log\left(\frac{300}{c\epsilon}\right)$ it holds

$$22(2n+p)\log(3(1+c)np) < \frac{\epsilon}{2}\frac{(2n+p)^2}{2}, \tag{5.14}$$

for all $n \in \mathbb{Z}_{>0}$.Indeed, this can be equivalently written as

$$22 < \frac{\epsilon}{4}\frac{2n+p}{\log(3(1+c)np)}.$$

But $\frac{2n+p}{\log(3(1+c)np)}$ increases with respect to $n \in \mathbb{Z}_{>0}$ and therefore it is minimized for $n = 1$. In particular it suffices to have

$$22 < \frac{\epsilon}{4}\frac{2+p}{\log(3(1+c)p)},$$

which can be checked to be true for $p \geq \frac{300}{\epsilon}\log\left(\frac{300}{(1+c)\epsilon}\right)$. Therefore using (5.14) it suffices

$$N > (1+\frac{\epsilon}{2})\frac{(2n+p)^2}{2n} + \frac{2n+p}{n}\log(RQ).$$

But observe

$$\begin{aligned} N &\geq (1+\epsilon)\left[\frac{p^2}{2n} + 2n + 2p + (2+\frac{p}{n})\log(RQ)\right] \\ &= (1+\epsilon)\left[\frac{(2n+p)^2}{2n} + (\frac{2n+p}{n})\log(RQ)\right] \\ &> (1+\frac{\epsilon}{2})\frac{(2n+p)^2}{2} + (2n+p)\log(RQ). \end{aligned}$$

The proof of Proposition 5.2.6 is complete. □

Proof of Proposition 5.2.7

Proof. We first establish that $\|X\|_\infty \leq (np)^2$ whp as $p \to +\infty$. By a union bound and Markov inequality

$$\mathbb{P}\left(\max_{i\in[n],j\in[p]} |X_{ij}| > (np)^2\right) \leq np\mathbb{P}\left(|X_{11}| > (np)^2\right) \leq \frac{1}{np}\mathbb{E}[|X_{11}|] = o(1).$$

Therefore with high probability $\|X\|_\infty \leq (np)^2$. Consider the set $T(R,Q)$ of all the vectors $\beta^* \in [-R,R]^p$ satisfying the Q-rationality assumption. The entries of these vectors are of the form $\frac{a}{Q}$ for some $a \in \mathbb{Z}$ with $|a| \leq RQ$. In particular $|T(R,Q)| = (2QR+1)^p$. Now because the entries of X are continuously distributed, all $X\beta^*$ with $\beta^* \in T(R,Q)$ are distinct with probability 1. Furthermore by the above each one of them has L_2 norm satisfies

$$\|X\beta^*\|_2^2 \leq np^2\|X\|_\infty^2\|\beta^*\|_\infty^2 \leq R^2n^5p^6 < R^2(np)^6,$$

w.h.p. as $p \to +\infty$.

Now we establish the proposition by contradiction. Suppose there exist a recovery mechanism that can recover w.h.p. any such vector β^* after observing $Y = X\beta^* + W \in \mathbb{R}^n$, where W has n iid $N(0,\sigma^2)$ entries. In the language of information theory such a recovery guarantee implies that the Gaussian channel with power constraint $R^2(np)^6$ and noise variance σ^2 needs to have capacity at least

$$\frac{\log|T(R,Q)|}{n} = \frac{p\log(2QR+1)}{n}.$$

On the other hand, the capacity of this Gaussian channel with power $\mathcal{R}$ and noise variance Σ^2 is known to be equal to $\frac{1}{2}\log\left(1+\frac{\mathcal{R}}{\Sigma^2}\right)$ (see for example Theorem 10.1.1 in [CT06]). In particular our Gaussian communication channel has capacity $\frac{1}{2}\log\left(1+\frac{R^2(np)^6}{\sigma^2}\right)$.From this we conclude

$$\frac{p\log\left(2QR+1\right)}{n} \leq \frac{1}{2}\log\left(1+\frac{R^2(np)^6}{\sigma^2}\right),$$

which implies

$$\sigma^2 \leq R^2(np)^6 \frac{1}{2^{\frac{2p\log(2QR+1)}{n}}-1},$$

or

$$\sigma \leq R(np)^3\left(2^{\frac{2p\log(2QR+1)}{n}}-1\right)^{-\frac{1}{2}},$$

which completes the proof of the Proposition. □

Proof of Proposition 5.2.8

Proof. Based on Proposition 5.2.6 the amount of noise that can be tolerated is

$$2^{-(1+\epsilon)\left[\frac{p^2}{2n}+2n+2p+(2+\frac{p}{n})\log(RQ)\right]},$$

for an arbitrary $\epsilon > 0$. Since $n = o(p)$ and $RQ = 2^{\omega(p)}$ this simplifies asymptotically to

$$2^{-(1+\epsilon)\left[\frac{p}{n}\log(RQ)\right]},$$

for an arbitrary $\epsilon > 0$. Since $\sigma < \sigma_0^{1+\epsilon}$, we conclude that LBR algorithms is succesfully working in that regime.

For the first part it suffices to establish that under our assumptions for p sufficiently large,

$$\sigma_0^{1-\epsilon} > R(np)^3\left(2^{\frac{2p\log(2QR+1)}{n}}-1\right)^{-\frac{1}{2}}.$$

Since $n = o(\frac{p}{\log p})$ implies $n = o(p)$ we obtain that for p sufficiently large,

$$2^{\frac{2p\log(2QR+1)}{n}} - 1 > 2^{2(1-\frac{1}{2}\epsilon)\frac{p\log(2QR+1)}{n}}$$

which equivalently gives

$$\left(2^{\frac{2p\log(2QR+1)}{n}} - 1\right)^{-\frac{1}{2}} < 2^{-(1-\frac{1}{2}\epsilon)\frac{p\log(2QR+1)}{n}}$$

or

$$R(np)^3\left(2^{\frac{2p\log(2QR+1)}{n}} - 1\right)^{-\frac{1}{2}} < R(np)^3 2^{-(1-\frac{1}{2}\epsilon)\frac{p\log(2QR+1)}{n}}.$$

Therefore it suffices to show

$$R(np)^3 2^{-(1-\frac{1}{2}\epsilon)\frac{p\log(2QR+1)}{n}} \leq \sigma_0^{1-\epsilon} = 2^{-(1-\epsilon)\frac{p\log(QR)}{n}}$$

or equivalently by taking logarithms and performing elementary algebraic manipulations,

$$n\log R + 3n\log(np) \leq \left(1 - \frac{\epsilon}{2}\right)p\log(2 + \frac{1}{RQ}) + \frac{\epsilon}{2}p\log RQ.$$

The condition $n = o(\frac{p}{\log p})$ implies for sufficiently large p, $n\log(np) \leq \frac{\epsilon}{4}p$ and $n\log R \leq \frac{\epsilon}{2}p\log QR$. Using both of these inequalities we conclude that for sufficiently large p,

$$\begin{aligned} n\log R + 3n\log(np) &\leq \frac{\epsilon}{2}p\log QR \\ &\leq \left(1 - \frac{\epsilon}{2}\right)p\log(2 + \frac{1}{RQ}) + \frac{\epsilon}{2}p\log RQ. \end{aligned}$$

This completes the proof.

□

5.7 Conclusion

In this Chapter, we consider the high dimensional linear regression model under exponential-in-p small noise level. We focus on X having iid entries generated from an, almost arbitrary, continuous distribution and β^* being an, almost arbitrary, rational-valued vector of coefficients. We propose a lattice-based method based on the celebrated Lenstra-Lenstra-Lovasz lattice basis reduction algorithm. The algorithms reduces the high dimensional linear regression problem to a shortest vector problem on an appropriately designed lattice. Interestingly, we prove that the algorithm correctly recovers exactly the vector β^* with one sample $n = 1$ and $p \to +\infty$. This is a significant improvement to standard compressed sensing methods, such as LASSO and Basis Pursuit, which are provably requiring diverging number of samples to succeed. Finally, we establish that, under mild assumptions on the range of values for the entries of β^*, our proposed algorithm obtains nearly-optimal noise tolerance.

Chapter 6

The Landscape of the Planted Clique Problem: Dense subgraphs and the Overlap Gap Property

6.1 Introduction

In this Chapter we study the planted clique problem, first introduced in [Jer92]. In this problem one observes an n-vertex undirected graph G sampled in two stages; in the first stage, the graph is sampled according to an Erdős-Rényi graph $G\left(n, \frac{1}{2}\right)$ and in the second stage, k out of the n vertices are chosen uniformly at random and all the edges between these k vertices are deterministically added (if they did not already exist due to the first stage sampling). We call the second stage chosen k-vertex subgraph the *planted clique* $\mathcal{PC}$. The inference task of interest is to recover $\mathcal{PC}$ from observing G. The focus is on the asymptotic setting where both $k = k_n, n \to +\infty$ and the recovery should hold with probability tending to one as $n \to +\infty$ (w.h.p.).

It is a standard result in the literature that as long as $k \geq (2+\epsilon)\log_2 n$, the graph G will have only $\mathcal{PC}$ as a k-clique in G w.h.p. (see e.g. [Bol85]). In particular under this assumption, $\mathcal{PC}$ is recoverable w.h.p. by the brute-force algorithm which checks every k-vertex subset of

whether they induce a k-clique or not. Note that the exhaustive algorithm requires $\binom{n}{k}$ time to terminate, making it in principle not polynomial-time for the values of k of interest. For any $k \geq (2+\epsilon)\log_2 n$, a relatively simple quasipolynomial-time algorithm, that is an algorithm with termination time $n^{O(\log n)}$, can be also proven to recover $\mathcal{PC}$ correctly w.h.p. as $n \to +\infty$ (see e.g. the discussion in [FGR$^+$17] and references therein). Note that a quasipolynomial-time termination time outperforms the termination time of the exhaustive search for $k = \omega(\log n)$.

The first polynomial-time (greedy) recovery algorithm of $\mathcal{PC}$ came out of the observation in [Kuč95] according to which when $k \geq C\sqrt{n \log n}$ for some sufficiently large $C > 0$, the k-highest degree nodes in G are the vertices of $\mathcal{PC}$ w.h.p. A fundamental work [AKS98] proved that a polynomial-time algorithm based on spectral methods recovers $\mathcal{PC}$ when $k \geq c\sqrt{n}$ for any fixed $c > 0$ (see also [FR10], [DM], [DGGP14] and references therein.) Furthermore, in the regime $k/\sqrt{n} \to 0$, various computational barriers have been established for the success of certain classes of polynomial-time algorithms, such as the Sum of Squares Hierarchy [BHK$^+$16], the Metropolis Process [Jer92] and statistical query algorithms [FGR$^+$17]. Nevertheless, no general algorithmic barrier such as NP-hardness has been proven for recovering $\mathcal{PC}$ when $k/\sqrt{n} \to 0$. The absence of polynomial-time algorithms together with the absence of an NP-hardness explanation in the regime where $k \geq (2+\epsilon)\log n$ and $k/\sqrt{n} \to 0$ gives rise to arguably one of the most celebrated and well-studied computational-statistical gaps in the literature, known as the *planted clique problem.*

Computational gaps Computational gaps between what existential or brute-force methods promise and what computationally efficient algorithms achieve is an ubiquitous phenomenon in the analysis of algorithmic tasks in random environments. Such gaps arise for example in the study of several *"non-planted"* models like the maximum-independent-set problem in sparse random graphs [GSa], [COE11], the largest submatrix problem of a random Gaussian matrix [GL16], the diluted 4-spin-model [CGPR17] and the study of random k-SAT [MMZ05], [ACO08]. Recently, such computational gaps started appearing in *"planted"* inference algorithmic tasks in statistics literature such as the high dimensional linear regression problem [GZ17a], [GZ17b], the tensor principal component analysis (PCA) [BAGJ18], [BR13] the stochastic block model (see [Abb17], [BBH18] and references therein) and, of course, the planted clique problem described

above. Towards the fundamental study of such computational gaps the following two methods have been considered.

(1) **Computational gaps: Average-Case Complexity Theory and the central role of Planted Clique**

None of the above gaps have been proven to be an NP-hard algorithmic task. Nevertheless, in correspondence with the well-studied worst-case NP-Completeness complexity theory (see e.g. [Kar72]), some very promising attempts have been made towards building a similar theory for planted inference algorithmic tasks (see e.g. [BR13], [CLR17], [WBP16], [BBH18] and references therein). The goal of this line of research is to show that for two conjecturally computationally hard statistical tasks the existence of a polynomial-time algorithm for one task implies a polynomial-time recovery algorithm for the other. In particular, (computational hardness of) the latter task reduces to (computational hardness of) the former. Notably, the *planted clique problem* seem to play a central role in these developments, similar to the role the boolean-satisfiability problem played in the development of the worst-case NP-completeness theory. Specifically in the context of statistical reduction, multiple statistical tasks in their conjecturally hard regime such as Sparse-PCA [BR13], submatrix localization [CLR17], RIP certification [WBP16], rank-1 Submatrix Detection, Biclustering [BBH18] have been proven to reduce to the planted clique problem in the regime $k/\sqrt{n} \to 0$.

(2) **Computational Gaps: A Spin Glass Perspective (Overlap Gap Property)**

For several of the above-mentioned computational gaps, an inspiring connection have been drawn between the geometry of their solution space, appropriately defined, and their algorithmic difficulty. Specifically it has been repeatedly observed that the appearance of a certain disconnectivity property in the solution space called *Overlap Gap Property (OGP)*, originated in spin glass theory, coincides with the conjectured algorithmic hard phase for the problem. Furthermore, it has also been seen that at the absence of this property even greedy algorithms can exploit the smooth geometry and succeed.

The connection between algorithmic performance and OGP was initially made in the study of the celebrated example of random k-SAT (independently by [MMZ05], [ACORT11]) but

then has been established for other “non-planted" models such as maximum independent set in random graphs [GSa], [RV14] but also “planted models" such as high dimensional linear regression [GZ17a], [GZ17b] and tensor PCA [BAGJ18]. Despite the fundamental nature of the planted clique problem in the development of average-case complexity theory, OGP has not been studied for the planted clique problem. The study of OGP in the context of the planted clique problem is the main focus of this work.

We start with providing some intuition on what OGP is in the context of “non-planted" problems. Motivated by the study of concentration of the associated Gibbs measures [Tal10] for low enough temperature, the OGP concerns the geometry of the near (optimal) solutions. It has been observed that any two “near-optimal" solutions for many such modes exhibit the disconnectivity property stating that that their overlap, measured as a rescaled Hamming distance, is either very large or very small, which we call *the Overlap Gap Property (OGP)* [ACORT11], [ACO08], [MRT11], [COE11], [GSa], [RV14], [CGPR17] [GSb]. For example, the independent sets achieving nearly maximal size in sparse random graph exhibit the OGP [GSa]. An interesting rigorous link also appears between OGP and the power of local algorithms. For example OGP has been used in [GSa] to establish a fundamental barriers on the power of a class of local algorithms called i.i.d. factors for finding nearly largest independent sets in sparse random graphs (see also [RV14] for a tighter later result). Similar negative results have been established in the context of the random NAE-K-SAT problem for the Survey propagation [GSb], of random NAE-K-SAT for the Walksat algorithm [COHH16] and of the max-cut problem in random hypergraphs for the family of i.i.d. factors [CGPR17], As mentioned also above, when OGP disappears the picture changes and, for many of these problems, greedy methods successfully work [ACO08], [AKKT02]. Importantly, because of this connection it is conjectured that the onset of the phase transition point for the presence of OGP corresponds to the onset of algorithmic hardness.

It is worth mentioning that other properties such as the shattering property and the condensation, which have been extensively studied in the context of random constraint satisfaction problems, such as random K-SAT, are topological properties of the solution space which have been linked with algorithmic difficulty (see e.g. [ACO08], [KMRT$^+$07] for appropriate

definitions). We would like to importantly point out that neither of them is identical with OGP. OGP implies for trivial reasons the shattering property but the other implication does not hold. For example, consider the model of random linear equations [ACOGM17], where recovery can be obtained efficiently via the Gaussian elimination when the system is satisfiable. In [ACOGM17] it is established that OGP never appears as the overlaps concentrate on a single point but shattering property does hold in a part of the satisfiability regime. Furthermore, OGP is also not the same with condensation. For example, in the solution space of random K-SAT, OGP appears for multioverlaps around ratio clauses to variables about $2^K \log 2/K$ (up to poly-log K factors) [GSb] which is far below condensation which appears around ratio $2^K \log 2$ [KMRT$^+$07]. It should be noted that in random k-SAT the onset of the apparent algorithmic hardness also occurs around $2^K \log 2/K$ [GSb], [Het16]. The exact connection between each of these properties and algorithmic hardness is an ongoing and fascinating research direction.

Recently the study of OGP has been initiated for "planted" problems as well, for example for the high dimensional linear regression problem [GZ17a], [GZ17b]. For this "planted" problem, the goal is to recover a hidden k-sparse binary vector from noisy linear observations of it. The strategy followed in this Chapter is comprised of two steps. First the task is reduced into an average-case optimization task associated with a natural empirical risk objective. Then, as a second step, a geometric analysis of the region of feasible solutions is performed and the OGP (or the lack of it) is established. Interestingly, in this line of work the "overlaps" considered are between the "near-optimal" solutions of the optimization task and the planted structure itself. In the present paper we follow a similar path to identify the OGP phase transition point for the planted clique problem.

Contribution and Discussion

In this Chapter we analyze the presence of OGP for the planted clique problem. We first turn the inference goal into an average-case optimization problem by adopting an "empirical risk" objective and then perform the OGP analysis on the landscape of near-optimal solutions. The first natural choice for the empirical risk is the log-likelihood of the recovery problem which assigns to any k-subset $C \subseteq V(G)$ the risk value $-\log \mathbb{P}(\mathcal{PC} = C|G)$. A relatively straightforward analysis of

this choice implies that when $k \geq (2+\epsilon)\log_2 n$ the only k-subset obtaining a non-trivial log-likelihood is the planted clique itself, since there are no other cliques of size k in the graph w.h.p. as $n \to +\infty$. In particular, this perspective of studying the near-optimal solutions and OGP fails to provide anything fruitful.

The Dense Subgraphs Landscape and OGP We adopt the "relaxed" *k-Densest-Subgraph* objective of the observed graph G which assigns to any k-subset $C \subseteq V(G)$ the empirical risk $-|\mathrm{E}[C]|$, that is we would like to solve

$$\mathcal{D}(G) : \max_{C \subseteq V(G), |C|=k} |E[C]|,$$

where by $\mathrm{E}[C]$ we refer to the set of edges in the induced subgraph defined by C. Notice that $\mathcal{D}(G)$ is equivalent with maximizing the log-likelihood of a similar recovery problem, the planted k-dense subgraph problem where the edges of $\mathcal{PC}$ are only placed with some specific probability $1 > p > 1/2$ and the rest of the edges are still drawn with probability $\frac{1}{2}$ as before (see e.g. [BBH18] and references therein). Also, notice that, interestingly, $\mathcal{D}(G)$ does not depend on the value of p; that is it is universal for all values of $p \in (\frac{1}{2}, 1)$. Now the planted clique model we are interested in can be seen as the extreme case of the planted k-dense subgraph problem when $p \to 1^-$. In this work we analyze the *overparametrized* version of $\mathcal{D}(G)$, $\bar{k}$-densest-subgraph problem, where for some parameter $\bar{k} \geq k$ the focus is on

$$\mathcal{D}_{\bar{k},k}(G) : \max_{C \subseteq V(G), |C|=\bar{k}} |E[C]|, \tag{6.1}$$

while importantly *the planted clique in G remains of size k.* In this work we study the following question:

How much can a near-optimal solution of $\mathcal{D}_{\bar{k},k}(G)$ intersect the planted clique $\mathcal{PC}$?

The Overlap Gap Property *($\bar{k}$-OGP)* for the $\bar{k}$-Densest subgraph problem would mean that near-optimal solution of $\mathcal{D}_{\bar{k},k}(G)$ (sufficiently dense $\bar{k}$-subgraphs of G) have *either a large or small intersection with the planted clique* (see Definition 6.2.1 below for more details on the notion).

To study the presence of $\bar{k}$-OGP we focus on the monotonicity of the overlap-restricted optimal values for $z = \lfloor \frac{\bar{k}k}{n} \rfloor, \lfloor \frac{\bar{k}k}{n} \rfloor + 1, \ldots, k$;

$$d_{\bar{k},k}(G)(z) = \max_{C \subseteq V(G), |C|=\bar{k}, \text{overlap}(C)=z} |E[C]|,$$

where $\text{overlap}(C) := |C \cap \mathcal{PC}|$. Note that we define the overlaps beginning from $\lfloor \frac{\bar{k}k}{n} \rfloor$ as this level of overlap with $\mathcal{PC}$ is trivially obtained from a uniformly at random chosen $\bar{k}$-vertex subgraph.

Monotonicity and OGP It is not hard to see that the monotonicity (or lack of) of $d_{\bar{k},k}(G)(z)$ might be linked with the presence or absence of $\bar{k}$-OGP. For example, assume that for some realization of G the curve $d_{\bar{k},k}$ satisfies that for some $z^* \in (\lfloor \frac{\bar{k}k}{n} \rfloor, k) \cap \mathbb{Z}$,

$$d_{\bar{k},k}(G)(z^*) < \min\{d_{\bar{k},k}(G)(0), d_{\bar{k},k}(G)(k)\}. \tag{6.2}$$

then $\bar{k}$-OGP holds. Indeed, choosing any $\mathcal{T} > 0$ with

$$d_{\bar{k},k}(G)(z^*) < \mathcal{T} < \min\{d_{\bar{k},k}(G)(0), d_{\bar{k},k}(G)(k)\}$$

we notice that (1) since $\mathcal{T} > d_{\bar{k},k}(G)(z^*)$ any "dense" $\bar{k}$-subgraph with at least $\mathcal{T}$ edges cannot overlap at exactly z^* vertices with $\mathcal{PC}$ and (2) since $\mathcal{T} < \min\{d_{\bar{k},k}(G)(0), d_{\bar{k},k}(G)(k)\}$ there exist both zero and full overlap "dense" $\bar{k}$-subgraphs with that many edges. On the other hand, when the curve is monotonic with respect to overlap z, $\bar{k}$-OGP does not hold for a similar reasoning. Furthermore, note, that when the curve is monotonically increasing the near-optimal solutions of $\mathcal{D}_{\bar{k},k}(G)$ have almost full intersection with $\mathcal{PC}$ (hence, considered *relevant* for recovery), while when it is monotonically decreasing the near-optimal solutions of $\mathcal{D}_{\bar{k},k}(G)$ have almost empty intersection with $\mathcal{PC}$ (hence, considered *irrelevant* for recovery).

Monotonicity of the First Moment Curve Using an optimized union-bound argument (first moment method) we obtain a deterministic upper bound function (we call it *first moment curve*) $\Gamma_{\bar{k},k}(z)$ such that for all overlap values z,

$$d_{\bar{k},k}(G)(z) \leq \Gamma_{\bar{k},k}(z), \tag{6.3}$$

	Low Overparametrization $\bar{k}$	High Overparametrization $\bar{k}$
$k = o(\sqrt{n})$	$\Gamma_{\bar{k},k}$ non-monotonic	$\Gamma_{\bar{k},k}$ monotonically decreasing
$k = \omega(\sqrt{n})$	$\Gamma_{\bar{k},k}$ non-monotonic	$\Gamma_{\bar{k},k}$ monotonically increasing

Table 6.1: The monotonicity phase transitions of $\Gamma_{\bar{k},k}$ at $k = \sqrt{n}$ and varying $\bar{k}$.

which is also provably tight, up-to-lower order terms, at the end-point $z = 0$ (Proposition 6.2.3). For this reason, with the hope that $\Gamma_{\bar{k},k}(z)$ provides a tight upper bound in (6.3), we perform a monotonicity analysis of $\Gamma_{\bar{k},k}(z)$.

We discover that when $k = o(\sqrt{n})$, and relatively small $\bar{k}$ (including $\bar{k} = k$) $\Gamma_{\bar{k},k}$ is *non-monotonic* satisfying a relation similar to (6.2) for some z^*, while for relatively large $\bar{k}$ it is *decreasing*. On the other hand, when $k = \omega(\sqrt{n})$ for relatively small $\bar{k}$ $\Gamma_{\bar{k},k}$ is *non-monotonic* satisfying a relation similar to (6.2) for some z^*, while for relatively large $\bar{k}$ it is *increasing*. In particular, an exciting phase transition is taking place at the critical size $k = \sqrt{n}$ and high overparametrization $\bar{k}$. A summary is produced in Table 1. Theorem 6.2.5 and the discussion that follows provide exact details of the above statements.

Assuming the tightness of $\Gamma_{\bar{k},k}$ in (6.3) we arrive at a *conjecture* regarding the $\bar{k}$-OGP of the landscape. In the apparently algorithmically ihard regime $k = o(\sqrt{n})$ the landscape is either exhibiting $\bar{k}$-OGP or is uniformative. On the other hand, in the algorithmically tractable regime $k = \omega(\sqrt{n})$ for appropriately large $\bar{k}$ there is no $\bar{k}$-OGP and the optimal solutions of $\mathcal{D}_{\bar{k},k}(G)$ have almost full overlap with $\mathcal{PC}$. Of course this is only a prediction for the monotonicity of $d_{\bar{k},k}(G)$, as the function $\Gamma_{\bar{k},k}$ corresponds only to an upper bound. For this reason we establish results proving parts of the picture suggested by the monotonicity of $\Gamma_{k,\bar{k}}$.

Overlap Gap Property for $k = n^{0.0917}$ We establish that under the assumption $k \leq \bar{k} = n^C$, for some $0 < C < C^* = \frac{1}{2} - \frac{\sqrt{6}}{6} \sim 0.0917..$ indeed $\bar{k}$-OGP holds for $\mathcal{D}_{\bar{k},k}(G)$ (notice $k = o(\sqrt{n})$) in the regime. The result holds for all values of $\bar{k}$ (up-to-log factors) where the curve $\Gamma_{k,\bar{k}}$ is proven non-monotonic (Theorem 6.2.9). Specifically, we establish that for some constants $0 < D_1 < D_2$ any $\bar{k}$-subgraph of G which is *"sufficiently dense"* will either intersect $\mathcal{PC}$ in *at most* $D_1\sqrt{\frac{\bar{k}}{\log \frac{n}{\bar{k}}}}$ *nodes* or in *at least* $D_2\sqrt{\frac{\bar{k}}{\log \frac{n}{\bar{k}}}}$ *nodes*. Our proof is based on a delicate second moment method argument for dense subgraphs of Erdős-Rényi graphs. We believe that the second moment method argument can be further improved to extend the result to the case $C^* = 0.5 - \epsilon$ for

arbitrary $\epsilon > 0$. We leave this important step as an open question.

The use of Overparametrization The ability to choose $\bar{k} > k$ is paramount in all the results described here. If we have opted for the arguably more natural choice $\bar{k} = k$, and focused solely on k-vertex subgraphs the monotonicity of the curve $\Gamma_{\bar{k},k}$ exhibits a phase transition at the peculiar threshold $k = n^{\frac{2}{3}}$ (see Remark 6.2.6). To make this more precise, no landscape phase transition is suggested around the apparent algorithmic threshold $k = \sqrt{n}$ if we focus on k-vertex dense subgraphs (see for example the identical nature of Figure 1(a) and Figure 2(a) where k is chosen near $\sqrt{n}$ from below and above respectively). For this reason, the use of overparametrization is fundamental.

Significant inspiration from this overparametrization approach is derived from it's recent success on "smoothening" bad local behavior in landscapes arising predominantly in the context of deep learning [SS17], [VBB18], [LMZ18] but also beyond it (e.g. [XHM18] in the context of learning mixtures of Gaussians). We consider this to be a novel conceptual contribution to this line of research on computational-statistical gaps with potentially various extensions.

$n^{0.5-\epsilon}$-Dense Subgraphs of $G\left(n, \frac{1}{2}\right)$ Proposition 6.2.3 and Theorem 6.2.9 are based on a new result on the K-Densest subgraph of a vanilla Erdős-Rényi model G_0 sampled from $G\left(n, \frac{1}{2}\right)$;

$$d_{\mathrm{ER},K}(G_0) = \max_{C \subseteq V(G_0), |C|=K} |E[C]|,$$

for any $K < n^{\frac{1}{2}-\epsilon}$ where $\epsilon > 0$. The study of $d_{\mathrm{ER},K}(G_0)$ is a natural question in random graph theory which, to the best of our knowledge, remains not well-understood even for moderately large values of $K = K_n$. For small enough values of K, specifically $K < 2\log_2 n$, it is well-known $d_{\mathrm{ER},K}(G_0) = \binom{K}{2}$ w.h.p. as $n \to +\infty$ (originally established in [GM75]). On the other hand when $K = n$, trivially $d_{\mathrm{ER},K}(G_0)$ follows Binom $\left(\binom{K}{2}, \frac{1}{2}\right)$ and hence for any $\alpha_K = \omega(1)$, $d_{\mathrm{ER},K}(G_0) = \frac{1}{2}\binom{K}{2} + O(K\alpha_K)$ w.h.p. as $n \to +\infty$. If we choose for the sake of argument $\alpha_K = \log\log K$ the following natural question can be posed;

How $d_{\mathrm{ER},K}(G_0)$ transitions from $\binom{K}{2}$ for $K < 2\log_2 n$ to $\frac{1}{2}\binom{K}{2} + O(K\log\log K)$ for $K = n$?

A recent result in the literature studies the case $K = C\log n$ for $C > 2$ [BBSV18] and

establishes (it is an easy corollary of the main result of the aforementioned paper),

$$d_{\mathrm{ER},K}(G_0) = h^{-1}\left(\log 2 - \frac{2\left(1+o\left(1\right)\right)}{C}\right)\binom{k}{2}, \tag{6.4}$$

w.h.p. as $n \to +\infty$. Here log is natural logarithm and h^{-1} is the inverse of the (rescaled) binary entropy $h : [\frac{1}{2}, 0] \to [0, 1]$ is defined by

$$h(x) = -x\log x - (1-x)\log x. \tag{6.5}$$

Notice that $\lim_{C\to+\infty} h^{-1}\left(\log 2 - \frac{2(1+o(1))}{C}\right) = \frac{1}{2}$ which means that the result from [BBSV18] agrees with the first order behavior of $d_{\mathrm{ER},K}(G_0)$ at " very large" K such as $K = n$. The proof from [BBSV18] is based on a careful and elegant application of the second moment method, where special care is made to control the way "sufficiently dense" subgraphs overlap.

We study the behavior of $d_{\mathrm{ER},K}(G_0)$ for any $K < n^{\frac{1}{2}-\epsilon}$, for $\epsilon > 0$. Specifically, we build and improve on the second moment method technique from [BBSV18] and establish tight results for first and second order behavior of $d_{\mathrm{ER},K}(G_0)$ when K is a power of n strictly less than $\sqrt{n}$. Specifically in Theorem 6.2.10 we show that for any $K = n^C$ for $C \in (0, \frac{1}{2})$ there exists some positive constant $\beta = \beta(C) \in (0, \frac{3}{2})$ such that

$$d_{\mathrm{ER},K}(G_0) = h^{-1}\left(\log 2 - \frac{\log\binom{n}{K}}{\binom{K}{2}}\right)\binom{K}{2} - O\left(K^{\beta}\sqrt{\log n}\right) \tag{6.6}$$

w.h.p. as $n \to +\infty$.

First notice that as our result are established when K is a power n it does not apply in the logarithmic regime. Nevertheless, it is in agreement with the result of of [BBSV18] since for $K = C\log n$,

$$\frac{\log\binom{n}{K}}{\binom{K}{2}} = (1+o\left(1\right))\frac{K\log\left(\frac{n}{K}\right)}{\frac{K^2}{2}} = (1+o\left(1\right))\frac{2}{C},$$

that is the argument in h^{-1} of (6.6) converges to the argument in h^{-1} of (6.4) at this scaling.

Finally, by Taylor expanding h^{-1} around $\log 2$: $h^{-1}\left(\log 2 - t\right) = \frac{1}{2} + \frac{1}{\sqrt{2}}\sqrt{t} + o\left(\sqrt{t}\right)$ for

$t = o(1)$ (Lemma 6.8.3), using our result we can identify the second order behavior of $d_{\mathrm{ER},K}(G_0)$

$$d_{\mathrm{ER},K}(G_0) = \frac{K^2}{4} + \frac{K^{\frac{3}{2}}\sqrt{\log\left(\frac{n}{K}\right)}}{2} + o\left(K^{\frac{3}{2}}\right),$$

w.h.p. as $n \to +\infty$. See Corollary 2 for the exact statement. Note that the second order behavior is of different order in K that in the case $K = n$. We leave the analysis of the behavior of $d_{\mathrm{ER},K}(G_0)$ in the regime for K between $n^{\frac{1}{2}}$ and n as an intruiguing open question.

Connection with $\bar{k}$-OGP Notice that our result (6.6) holds for any $K = \Theta\left(n^C\right)$, $0 < C < \frac{1}{2}$ but in the discussion above we only claimed of using this result to prove $\bar{k}$-OGP for $C < 0.0917$. This happens because to establish $\bar{k}$-OGP using our non-monotonicity arguments and this result (for $K = \bar{k}$) we need to make sure the error term in (6.6) is $o(K)$, which from our result it can only be established if $C < 0.0917$. The reason is that to transfer the non-monotonicity of the first moment curve $\Gamma_{\bar{k},k}(z)$ to the non-monotonicity of the actual curve $d_{\bar{k},k}(G)$ we need the error term in our approximation gap between $d_{\bar{k},k}(G)(z)$ and $\Gamma_{\bar{k},k}(z)$ to do not alter the non-monotonicity behavior of $\Gamma_{\bar{k},k}(z)$. We quantify the non-monotonicity via its "depth", that is via

$$\min\{\Gamma_{\bar{k},k}(0), \Gamma_{\bar{k},k}(k)\} - \min_{z\in[0,k]} \Gamma_{\bar{k},k}(z).$$

The latter "depth" quantity can be proven to grow with order similar to $\Omega(K) = \Omega\left(\bar{k}\right)$ leading to the necessary order for the error term to make the argument go through.

Notation Throughout the paper we use standard big O notations, e.g., for any real-valued sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, $a_n = \Theta(b_n)$ if there exists an absolute constant $c > 0$ such that $\frac{1}{c} \leq |\frac{a_n}{b_n}| \leq c$; $a_n = \Omega(b_n)$ or $b_n = O(a_n)$ if there exists an absolute constant $c > 0$ such that $|\frac{a_n}{b_n}| \geq c$; $a_n = \omega(b_n)$ or $b_n = o(a_n)$ if $\lim_n |\frac{a_n}{b_n}| = 0$.

For an undirected graph G on n vertices we denote by $V(G)$ the sets of its vertices and $E[G]$ the set of its edges. For a subset S of $V(G)$ we refer to the set of all vertices in $V(G)$ which are connected with an edge to every vertex of S, as the common neighborhood of S.

Throughout the paper we denote by h the (rescaled) binary entropy given by (6.5) and for $\gamma \in (\frac{1}{2}, 1)$, we define

$$r(\gamma, \frac{1}{2}) := \log 2 - h(\gamma). \tag{6.7}$$

6.2 Main Results

6.2.1 The Planted Clique Model and Overlap Gap Property

We start with formally defining the Planted Clique Model and the recovery goal of interest.

Parameter Assumptions Let $k, n \in \mathbb{N}$ with $k \leq n$. We assume that both k, n are known. All of our results focus on the regime where $k = k_n$ grows with n as $n \to +\infty$ with $\omega(\log n) = k = o(n)$.

The Generative Process First sample an n vertex undirected graph G_0 according to the Erdős-Rényi $G(n, \frac{1}{2})$ distribution. Then choose k out of n vertices of G_0 uniformly at random and connect all pairs of these vertices with an undirected edge, creating what we call as the planted clique $\mathcal{PC}$ of size k. We denote the resulting n-vertex undirected graph by $G\left(n, k, \frac{1}{2}\right)$ or G for simplicity.

The Recovery Goal Given one sample of G recover the vertices of the planted clique $\mathcal{PC}$.

6.2.2 The $\bar{k}$-Densest Subgraph Problem for $\bar{k} \geq k = |\mathcal{PC}|$

We study the landscape of the sufficiently dense subgraphs in G. Besides n, k we introduce an additional parameter $\bar{k} \in \mathbb{N}$ with $k \leq \bar{k} \leq n$ that will be optimized. The dense subgraphs we consider are of vertex size $\bar{k}$. We study overlaps between the sufficiently dense $\bar{k}$-dense subgraphs and the planted clique $\mathcal{PC}$. Specifically we focus on the $\bar{k}$-densest subgraph problem on G, $\mathcal{D}_{\bar{k},k}(G)$ defined in (6.8).

We define the $\bar{k}$-Overlap Gap Property of $\mathcal{D}_{\bar{k},k}(G)$.

Definition 6.2.1 ($\bar{k}$-OGP)**.** *$\mathcal{D}_{\bar{k}.k}(G)$ exhibits $\bar{k}$-Overlap Gap Property ($\bar{k}$-OGP) if there exists $\zeta_{1,n}, \zeta_{2,n} \in [k]$ with $\zeta_{1,n} < \zeta_{2,n}$ and $0 < r_n < \binom{k}{2}$ such that;*

(1) There exists $\bar{k}$-subsets $A, A' \subseteq V(G)$ with $|A \cap \mathcal{PC}| \leq \zeta_{1,n}$, $|A' \cap \mathcal{PC}| \geq \zeta_{2,n}$ and $\min\{|\mathrm{E}[A]|, |\mathrm{E}[A']|\} \geq r_n$.

(2) For any $\bar{k}$-subset $A \subset V(G)$ with $|\mathrm{E}[A]| \geq r_n$ it holds, either $|A \cap \mathcal{PC}| \leq \zeta_{1,n}$ or $|A \cap \mathcal{PC}| \geq \zeta_{2,n}$.

Here, the first part of the definition ensures that there are sufficiently dense $\bar{k}$-subgraphs of G with both "low" and "high" overlap with $\mathcal{PC}$. The second condition ensures that any sufficiently dense $\bar{k}$-subgraph of G will have either "low" overlap or "high" overlap with $\mathcal{PC}$, implying gaps in the realizable overlap sizes.

To study $\bar{k}$-OGP we study the following curve. For every $z \in \{\lfloor \frac{k\bar{k}}{n} \rfloor, \lfloor \frac{k\bar{k}}{n} \rfloor + 1, \dots, k\}$ let

$$\mathcal{D}_{\bar{k},k}(G)(z) : \max_{C \subseteq V(G), |C|=k, |C \cap \mathcal{PC}|=z} |E[C]|. \tag{6.8}$$

with optimal value denoted by $d_{\bar{k},k}(G)(z)$. In words, $d_{\bar{k},k}(G)(z)$ corresponds to the number of edges of the densest $\bar{k}$-vertex subgraph with vertex-intersection with the planted clique of cardinality z. Notice that, as explained in the previous section, we restrict ourselves to overlap at least $k\bar{k}/n$ since this level of intersection with $\mathcal{PC}$ is achieved simply by sampling uniformly at random a $\bar{k}$-vertex subgraph of G.

6.2.3 Monotonicity Behavior of the First Moment Curve $\Gamma_{\bar{k},k}$

The following deterministic curve will be of distinct importance in what follows.

Definition 6.2.2 (First moment curve)**.** *We define the first moment curve to be the real-valued function* $\Gamma_{\bar{k},k} : \{\lfloor \frac{k\bar{k}}{n} \rfloor, \lfloor \frac{k\bar{k}}{n} \rfloor + 1, \dots, k\} \to \mathbb{R}_{>0}$*, where for* $z = \bar{k} = k$*,*

$$\Gamma_{\bar{k},k}(k) = \binom{k}{2}$$

and otherwise

$$\Gamma_{\bar{k},k}(z) = \binom{z}{2} + h^{-1}\left(\log 2 - \frac{\log\left(\binom{k}{z}\binom{n-k}{\bar{k}-z}\right)}{\binom{\bar{k}}{2} - \binom{z}{2}}\right)\left(\binom{\bar{k}}{2} - \binom{z}{2}\right)$$

for $z \in \{\lfloor \frac{k\bar{k}}{n} \rfloor, \lfloor \frac{k\bar{k}}{n} \rfloor + 1, \dots, k\}$*,*

Here the function h^{-1} is the inverse function of h, which is defined in (6.5). We establish the following proposition relating $d_{k,\bar{k}}(G)(z)$ and $\Gamma_{\bar{k},k}(G)(z)$.

Proposition 6.2.3. *Let* $k, \bar{k}, n \in \mathbb{N}$ *with* $k \leq \bar{k} \leq n$*.*

(1) For any $z \in \{\lfloor \frac{k\bar{k}}{n} \rfloor, \lfloor \frac{k\bar{k}}{n} \rfloor + 1, \ldots, k\}$

$$d_{\bar{k},k}(G)(z) \leq \Gamma_{\bar{k},k}(z),$$

with high probability as $n \to +\infty$.

(2) Suppose $(\log n)^5 \leq k \leq \bar{k} = \Theta\left(n^C\right)$ *for* $C \in (0, \frac{1}{2})$. *For any* $\beta \in (0, \frac{3}{2})$ *with*

$$\beta = \beta(C) > \frac{3}{2} - \left(\frac{5}{2} - \sqrt{6}\right) \frac{1-C}{C},$$

$$\Gamma_{\bar{k},k}(0) - O\left(\left(\bar{k}\right)^{\beta} \sqrt{\log n}\right) \leq d_{\bar{k},k}(G)(0), \tag{6.9}$$

with high probability as $n \to +\infty$.

The bounds stated in Proposition 6.2.3 are based on the first and second moment methods. The proof of Proposition 6.2.3 is in Section 6.4.

Remark 6.2.4. *Under the assumptions of Part (2) of Proposition 6.2.3 we have*

$$\begin{aligned}\Gamma_{\bar{k},k}(0) &= \frac{1}{2}\binom{\bar{k}}{2} + \left(\frac{1}{\sqrt{2}} + o(1)\right)\sqrt{\binom{\bar{k}}{2}\log\left[\binom{n-k}{\bar{k}}\right]} \\ &= \frac{\left(\bar{k}\right)^2}{4} + \frac{\left(\bar{k}\right)^{\frac{3}{2}}\sqrt{\log\left(\frac{(n-k)e}{\bar{k}}\right)}}{2} + o\left(\left(\bar{k}\right)^{\frac{3}{2}}\sqrt{\log n}\right).\end{aligned}$$

Here we have used Taylor expansion for h^{-1} *around* $\log 2$*:* $h^{-1}(\log 2 - t) = \frac{1}{2} + \left(\frac{1}{\sqrt{2}} + o(1)\right)\sqrt{t}$ *(Lemma 6.8.3) for* $t = \frac{\log\left(\binom{n-k}{\bar{k}}\right)}{\binom{\bar{k}}{2}} = O\left(\frac{\log n}{k}\right) = o(1)$ *and Stirling's approximation. The above calculation shows that the additive error term in (6.9) can change the value of* $\Gamma_{\bar{k},k}(0)$ *only at the third higher order term.*

We explain here how Part (1) of Proposition 6.2.3 is established with a goal to provide intuition for the first moment curve definition. Fix some $z \in \{\lfloor \frac{\bar{k}k}{n} \rfloor, \lfloor \frac{\bar{k}k}{n} \rfloor + 1, \ldots, k\}$. For $\gamma \in (0,1)$ we consider the counting random variable for the number of subgraphs with $\bar{k}$ vertices, z vertices

common with the planted clique and at least $\binom{z}{2} + \gamma\left(\binom{\bar{k}}{2} - \binom{z}{2}\right)$ edges;

$$Z_{\gamma,z} := |\{A \subseteq V(G) : |A| = \bar{k}, |A \cap \mathcal{PC}| = z, |E[A]| \geq \binom{z}{2} + \gamma\left(\binom{\bar{k}}{2} - \binom{z}{2}\right)\}|.$$

Notice that first moment method, or simply Markov's inequality, yields

$$\mathbb{P}\left[Z_{\gamma,z} \geq 1\right] \leq \mathbb{E}\left[Z_{\gamma,z}\right].$$

In particular, if for some $\gamma > 0$ it holds $\mathbb{E}\left[Z_{\gamma,z}\right] = o(1)$ we conclude that $Z_{\gamma,z} = 0$ whp and in particular all dense subgraphs have at most $\binom{z}{2} + \gamma\left(\binom{\bar{k}}{2} - \binom{z}{2}\right)$ edges, that is

$$d_{\bar{k},k}(G)(z) \leq \binom{z}{2} + \gamma\left(\binom{\bar{k}}{2} - \binom{z}{2}\right),$$

w.h.p. as $n \to +\infty$. Therefore the pursuit of finding the tightest upper bound using this technique, consists of finding the $\min \gamma : \mathbb{E}\left[Z_{\gamma,z}\right] = o(1)$.

Note that for any subset $A \subset V(G)$ the number of its induced edges follows a shifted Binomial distribution $\binom{z}{2} + \mathrm{Bin}\left(\binom{\bar{k}}{2} - \binom{z}{2}, \frac{1}{2}\right)$. In particular, we have

$$\mathbb{E}\left[Z_{\gamma,z}\right] = \binom{k}{z}\binom{n-k}{\bar{k}-z}\mathbb{P}\left[\mathrm{Bin}\left(\binom{\bar{k}}{2} - \binom{z}{2}, \frac{1}{2}\right) \geq \gamma\left(\binom{\bar{k}}{2} - \binom{z}{2}\right)\right].$$

From this point on, standard identities connecting the tail of the Binomial distribution with the binary entropy function h (see for example Lemma 6.8.2 below) yield the optimal choice to be

$$\gamma := h^{-1}\left(\log 2 - \frac{\log\left(\binom{k}{z}\binom{n-k}{\bar{k}-z}\right)}{\binom{\bar{k}}{2} - \binom{z}{2}}\right),$$

which yields Part (1) if Proposition 6.2.3. More details are in Section 6.4. The part (2) follows from a much more elaborate second moment method, the discussion of which we defer to Subsection 6.2.5 and Section 6.3.

We study the monotonicity property of the first moment curve. We establish the following proposition which proves that for appropriate choice of the overparametrization level of $\bar{k}$, the first moment curve $\Gamma_{\bar{k},k}(G)$ exhibits a monotonicity phase transitions at the predicted algorithmic

threshold $k = \Theta(\sqrt{n})$.

Theorem 6.2.5 (Monotonicity Phase Transition at $k = \sqrt{n}$)**.** *Let $k, \bar{k}, n \in \mathbb{N}$ with $n \to +\infty$ and $\epsilon > 0$ an arbitrarily small constant. Suppose $k \leq \bar{k} \leq n$ and furthermore $(\log n)^5 \leq \bar{k} = o(n)$. There exist a sufficiently large constant $C_0 = C_0(\epsilon) > 0$ such that for the discretized interval $\mathcal{I} = \mathcal{I}_{C_0} = \mathbb{Z} \cap \left[\lfloor C_0 \frac{\bar{k}k}{n} \rfloor, (1-\epsilon)k\right]$ the following are true for sufficiently large n,*

(1) if $k = o(\sqrt{n})$ then

(1i) for any $\bar{k} = o\left(\frac{k^2}{\log\left(\frac{n}{k^2}\right)}\right)$, the function $\Gamma_{\bar{k},k}(z), z \in \mathcal{I}_{C_0}$ is non-monotonic (Figure 1(a)).

(1ii) for any $\bar{k} = \omega\left(\frac{k^2}{\log\left(\frac{n}{k^2}\right)}\right)$, the function $\Gamma_{\bar{k},k}(z), z \in \mathcal{I}_{C_0}$ is decreasing (Figure 1(b)).

(2) if $k = \omega(\sqrt{n})$ then

(2i) for any $\bar{k} = o\left(\frac{n^2}{k^2 \log\left(\frac{k^2}{n}\right)}\right)$, the function $\Gamma_{\bar{k},k}(z), z \in \mathcal{I}_{C_0}$ is non-monotonic (Figure 2(a)).

(2ii) for any $\bar{k} = \omega\left(\frac{n^2}{k^2 \log\left(\frac{k^2}{n}\right)}\right)$, the function $\Gamma_{\bar{k},k}(z), z \in \mathcal{I}_{C_0}$ is increasing (Figure 2(b)).

Furthermore, in the regime that the function is non-monotonic there are constants $0 < D_1 < D_2$ such that for $u_1 := D_1 \lceil \sqrt{\frac{\bar{k}}{\log\left(\frac{n}{k}\right)}} \rceil$ and $u_2 := D_2 \lceil \sqrt{\frac{\bar{k}}{\log\left(\frac{n}{k}\right)}} \rceil$ and large enough n the following are true.

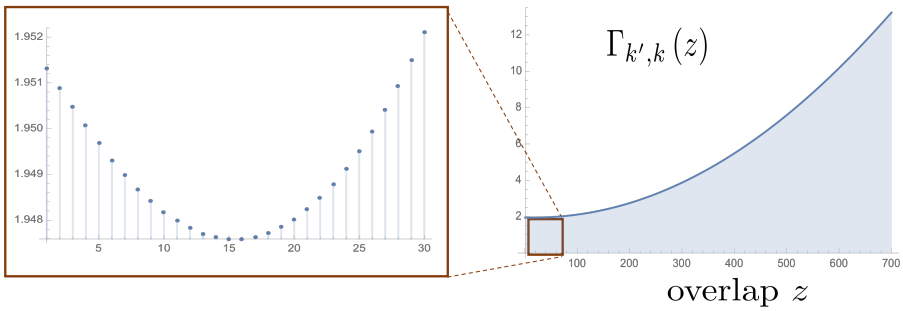
(a) $\lfloor C_0 \frac{\bar{k}k}{n} \rfloor < u_1 < u_2 < (1-\epsilon)k$ and

(b)

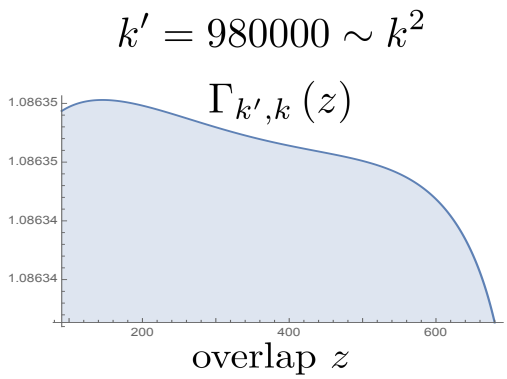
$$\max_{z \in \mathcal{I} \cap [u_1, u_2]} \Gamma_{\bar{k},k}(z) + \Omega\left(\frac{\bar{k}}{\log\left(\frac{n}{k}\right)}\right) \leq \Gamma_{\bar{k},k}(\lfloor C_0 \frac{\bar{k}k}{n} \rfloor) \leq \Gamma_{\bar{k},k}((1-\epsilon)k). \tag{6.10}$$

The proof of the Theorem can be found in Section 6.5.

Remark 6.2.6. *In the special case where $\bar{k} = k$, it can be straightforwardly checked from Theorem 6.2.5 that $\Gamma_{\bar{k},k}$ exhibits a monotonicity phase transition at $k = \Theta\left(n^{\frac{2}{3}}\right)$. In particular, for $\bar{k} = k$, the monotonicity of $\Gamma_{\bar{k},k}$ obtains no phase transition around $k = \Theta(\sqrt{n})$.*



(a) "Low" overparametrization $\bar{k} = k = 700$.



(b) "High" overparametrization $\bar{k} = 2k^2 = 980000$.

Figure 6-1: The behavior $\Gamma_{\bar{k},k}$ for $n = 10^7$ nodes, planted clique of size $k = 700 \ll \lfloor\sqrt{n}\rfloor = 3162$ and "high" and "low" values of $\bar{k}$. We approximate $\Gamma_{\bar{k},k}(z)$ using the Taylor expansion of h^{-1} by $\tilde{\Gamma}_{\bar{k},k}(z) = \frac{1}{2}\left(\binom{k}{2} + \binom{z}{2}\right) + \frac{1}{\sqrt{2}}\sqrt{\left(\binom{k}{2} - \binom{z}{2}\right)\log\left[\binom{k}{z}\binom{n-k}{\bar{k}-z}\right]}$. To capture the monotonicity behavior, we renormalize and plot $\left(\bar{k}\right)^{-\frac{3}{2}}\left(\tilde{\Gamma}_{\bar{k},k}(z) - \frac{1}{2}\binom{\bar{k}}{2}\right)$ versus the overlap sizes $z \in [\lfloor\frac{\bar{k}k}{n}\rfloor, k]$.

Remark 6.2.7. *Note that the monotonicity analysis in Theorem 6.2.5 is performed in the slightly "shrinked" interval* $\mathcal{I}_{C_0} = \mathbb{Z} \cap \left[\lfloor C_0\frac{\bar{k}k}{n}\rfloor, (1-\epsilon)\,k\right]$ *for arbitrarily small* $\epsilon > 0$ *and some constant* $C_0 = C_0(\epsilon) > 0$. *The restriction is made purely for technical reasons as it allows for an easier analysis of the curve's monotonicity behavior. We leave the monotonicity analysis near the endpoints as a topic for future work.*

Theorem 6.2.5 suggests that there are four regimes of interest for the pair $(k, \bar{k})$ and the monotonicity behavior of $\Gamma_{\bar{k}k}(z)$. We explain here the implication of Theorem 6.2.5 under the assumption that $\Gamma_{\bar{k},k}(z)$ *is a tight approximation of* $d_{\bar{k},k}(G)(z)$.

Let us focus first on the regime where the size of the planted clique is $k = o(\sqrt{n})$. Assume first that the level of overparametrization is relatively small, namely $\bar{k} = o\left(\frac{k^2}{\log\left(\frac{n}{k^2}\right)}\right)$, including the case $\bar{k} = k$. In that case the curve is non-monotonic and (6.10) holds (the case of Figure 1(a)). Now this implies that $\bar{k}$-OGP appears for the model. The reason is that under the tightness assumption, (6.10) translates to

$$\max_{z\in\mathcal{I}\cap[u_1,u_2]} d_{\bar{k},k}(G)(z) + \Omega\left(\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}\right) \leq d_{\bar{k},k}(G)(\lfloor C_0\frac{\bar{k}k}{n}\rfloor) \leq d_{\bar{k},k}(G)\left((1-\epsilon)\,k\right).$$

Using that we conclude easily that for sufficiently small constant $c > 0$ any $\bar{k}$-vertex subgraph with number of edges at least $d_{\bar{k},k}(G)(\lfloor C_0\frac{\bar{k}k}{n}\rfloor) - c\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}$ must have either at most u_1 interesection with $\mathcal{PC}$ or at least u_2 intersection with $\mathcal{PC}$ and there exist both empty and full overlap dense subgraphs with at least that many edges.

Now assume that overparametrization is relatively large, that is $\bar{k} = \omega\left(\frac{k^2}{\log\left(\frac{n}{k^2}\right)}\right)$. Then the function $\Gamma_{\bar{k},k}(z)$ is decreasing (the case of Figure 1(b)). This is a regime where $\bar{k}$-OGP disappears but the higher overlap z with $\mathcal{PC}$ implies smaller value of $d_{\bar{k},k}(G)(z)$. In particular, in that case one can efficiently find a sufficiently dense subgraphs but they have almost zero intersection with $\mathcal{PC}$. In conclusion, when $k = o(\sqrt{n})$ (and again under the tightness assumption) either the landscape of the dense subgraphs is *uninformative or it exhibits $\bar{k}$-OGP.*

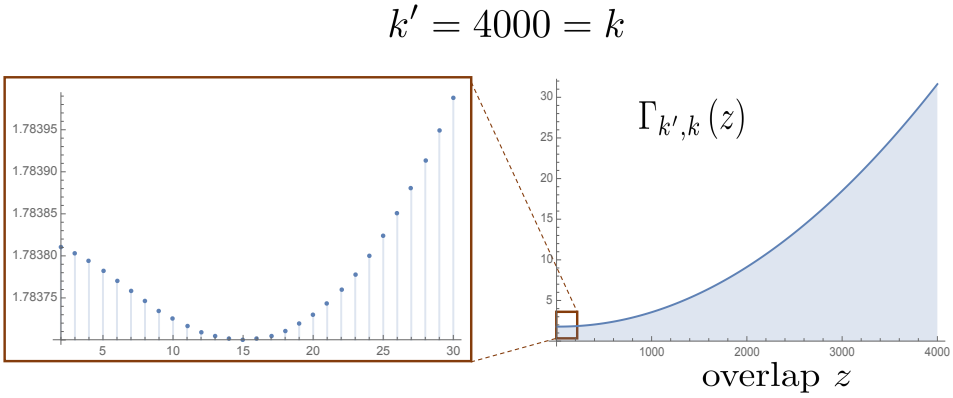
Now suppose $k = \omega(\sqrt{n})$. Assume first that the overparametrization is relatively small, that is $\bar{k} = o\left(\frac{n^2}{k^2\log\left(\frac{k^2}{n}\right)}\right)$. In that regime the curve is non-monotonic (see Figure 2(a)). Then, as in the previous case, $\bar{k}$-OGP appears for the model.

Finally assuming that the overparametrization is relatively large, that is $\bar{k} = \omega\left(\frac{n^2}{k^2\log\left(\frac{k^2}{n}\right)}\right)$, the function $\Gamma_{\bar{k},k}(z)$ becomes increasing (see Figure 2(b)). Under the tightness assumption, it is therefore implied that $\bar{k}$-OGP disappears and higher overlap z with $\mathcal{PC}$ implies higher $d_{\bar{k},k}(G)(z)$. This is an informative case where one can conjecturally find a sufficiently dense subgraphs, using a method of local improvements. Notice that in this regime which now sufficiently dense subgraphs have almost full intersection with $\mathcal{PC}$.

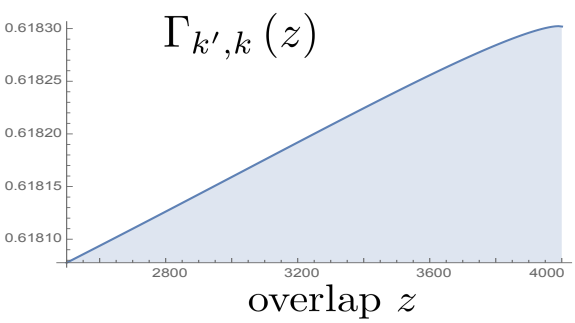
Summing this up we arrive at the following conjecture based on Theorem 6.2.5.

Conjecture 6.2.8. *Suppose* $(\log n)^5 \leq k = o(n)$.

(1) If $k = o(\sqrt{n})$ *then*



(a) "Low" overparametrization $\bar{k} = k = 4000$.



(b) "High" overparametrization $\bar{k} = n^2/k^2 = 6250000$.

Figure 6-2: The behavior $\Gamma_{\bar{k},k}$ for $n = 10^7$ nodes, planted clique of size $k = 4000 \gg \lfloor\sqrt{n}\rfloor = 3162$ and "high" and "low" values of $\bar{k}$. The rest of the plotting details are identical with that of Figure 1.

(1i) *for any* $\bar{k} = o\left(k^2 \log\left(\frac{n}{k^2}\right)\right)$ *there is* $\bar{k}$*-Overlap Gap Property w.h.p. as* $n \to +\infty$.

(1ii) *for any* $\bar{k} = \omega\left(k^2 \log\left(\frac{n}{k^2}\right)\right)$ *there is no* $\bar{k}$*-Overlap Gap Property, but* $d_{\bar{k},k}(G)(z)$ *is decreasing as a function of* z *w.h.p. as* $n \to +\infty$. *In particular, the near-optimal solutions of* $\mathcal{D}_{\bar{k},k}(G)$ *are uninformative about recovering* $\mathcal{PC}$.

(2) *if* $k = \omega(\sqrt{n})$,

(2i) *for any* $\bar{k} = o\left(\frac{n^2}{k^2 \log\left(\frac{k^2}{n}\right)}\right)$ *there is* $\bar{k}$*-Overlap Gap Property w.h.p. as* $n \to +\infty$.

(2ii) *for any* $\omega\left(\frac{n^2}{k^2 \log\left(\frac{k^2}{n}\right)}\right) = \bar{k} = o(n)$, *there is no* $\bar{k}$*-Overlap Gap Property and* $d_{\bar{k},k}(G)(z)$ *is increasing as a function of* z *w.h.p. as* $n \to +\infty$. *In particular, the near-optimal*

solutions of $\mathcal{D}_{\bar{k},k}(G)$ are informative about recovering $\mathcal{PC}$.

In the following two subsections we establish rigorously parts of Conjecture 6.2.8.

6.2.4 $\bar{k}$-Overlap Gap Property for $k = n^{0.0917}$

We now turn to the regime $k = o(\sqrt{n})$. In this regime Theorem 6.2.5 and Conjecture 6.2.8 suggests the presence of $\bar{k}$-OGP when $\bar{k} = o\left(k^2 \log\left(\frac{n}{k^2}\right)\right)$, which includes $\bar{k} = k$. We establish here the result that $\bar{k}$-OGP indeed holds as long as both $k, \bar{k}$ are less than n^C for $C \sim 0.0917...$

Theorem 6.2.9. *[$\bar{k}$-Overlap Gap Property]*
Suppose $(\log n)^5 \leq k \leq \bar{k} = \Theta\left(n^C\right)$ for some $C > 0$ with $0 < C < \frac{1}{2} - \frac{\sqrt{6}}{6} \sim 0.0917..$ and furthermore $\bar{k} = o\left(k^2 \log\left(\frac{n}{k^2}\right)\right)$.
Then there are constants $C_0 > 0$ and $0 < D_1 < D_2$ such that for $u_1 := D_1 \lceil \sqrt{\frac{\bar{k}}{\log\left(\frac{n}{k}\right)}} \rceil$ and $u_2 := D_2 \lceil \sqrt{\frac{\bar{k}}{\log\left(\frac{n}{k}\right)}} \rceil$ and large enough n the following holds.

(a) $\lceil C_0 \frac{\bar{k}k}{n} \rceil < u_1 < u_2 < \frac{k}{2}$ and

(b) $d_{\bar{k},k}(G)(z)$ is non-monotonic with

$$\min\{d_{\bar{k},k}(G)(0), d_{\bar{k},k}(G)\left(\frac{k}{2}\right)\} - \max_{z \in \mathcal{I} \cap [u_1, u_2]} d_{\bar{k},k}(G)(z) = \Omega\left(\frac{\bar{k}}{\log\left(\frac{n}{k}\right)}\right) \tag{6.11}$$

with high probability as $n \to +\infty$.

In particular, $\bar{k}$-Overlap Gap Property holds for the choice $\zeta_1 = u_1, \zeta_2 = u_2$ and $r_n = \Gamma_{\bar{k},k}(0) - \Theta\left(\frac{\bar{k}}{\log\left(\frac{n}{k}\right)}\right)$, with high probability as $n \to +\infty$.

The proof of the Theorem 6.2.9 is in Section 6.6.

6.2.5 K-Densest Subgraph Problem for $G\left(n, \frac{1}{2}\right)$

Of instrumental importance towards Theorem 6.2.3 and Theorem 6.2.9 is a new result on the value of the densest K-subgraph of a vanilla Erdős-Rényi graph $G_0\left(n, \frac{1}{2}\right)$. In this section we present this result. To the best of our knowledge it is the first such result for super-logarithmic-in-n values of K (see [BBSV18] and the Introduction of the present paper for details).

Let $1 \leq K \leq n$. We study the maximum number of edges of a subgraph of $G_0 \sim G(n, \frac{1}{2})$ with K vertices, that is

$$d_{\mathrm{ER},K}(G_0) := \max_{A \subseteq V(G), |A|=K} |\mathrm{E}[A]|. \tag{6.12}$$

We establish the following result.

Theorem 6.2.10. *Suppose $K = \Theta\left(n^C\right)$ for any constant $C \in (0, \frac{1}{2})$. For any fixed $\beta \in (0, \frac{3}{2})$ with*

$$\beta = \beta(C) > \max\{\frac{3}{2} - \left(\frac{5}{2} - \sqrt{6}\right)\frac{1-C}{C}, 0\}$$

it holds,

$$h^{-1}\left(\log 2 - \frac{\log\binom{n}{K}}{\binom{K}{2}}\right)\binom{K}{2} - O\left(K^{\beta}\sqrt{\log n}\right) \leq d_{\mathrm{ER},K}(G_0) \leq h^{-1}\left(\log 2 - \frac{\log\binom{n}{K}}{\binom{K}{2}}\right)\binom{K}{2}, \tag{6.13}$$

with high probability as $n \to +\infty$.

The proof of the theorem is given in Section 6.3.

Remark 6.2.11. *Let $C_{\mathrm{crit}} := 5/8 - \sqrt{6}/4$ be the unique positive solution to $\frac{3}{2} - \left(\frac{5}{2} - \sqrt{6}\right)\frac{1-C}{C} = 0$. Notice that Theorem 6.2.10 provides a qualitative different concentration result in the regime where $C \leq C_{\mathrm{crit}}$ and when $C > C_{\mathrm{crit}}$. In the former case it implies that for any arbitrarily small constant $\beta > 0$ (6.13) holds, while in the latter case for (6.13) to hold the exponent β needs to be assumed to be larger than $\frac{3}{2} - \left(\frac{5}{2} - \sqrt{6}\right)\frac{1-C}{C} > 0$.*

For any value of $C \in (0, 1/2)$ we can choose some $0 < \beta = \beta\,(C) < \frac{3}{2}$ so that Theorem 6.2.10, and in particular (6.13), holds for this value of β. Combining (6.13) with a direct applications of the Taylor expansion of h^{-1} (Lemma 6.8.3) and the Stirling's approximation for $\binom{n}{K}$ we obtain the following asymptotic behavior of $d_{\mathrm{ER},K}(G)$, for any $K = \Theta\left(n^C\right), C \in \left(0, \frac{1}{2}\right)$.

Corollary 2. *Suppose $K = \Theta\left(n^C\right)$ for any fixed $C \in (0, \frac{1}{2})$. Then,*

$$d_{\mathrm{ER},K}(G_0) = \frac{K^2}{4} + \frac{K^{\frac{3}{2}}\sqrt{\log\left(\frac{n}{K}\right)}}{2} + o\left(K^{\frac{3}{2}}\right), \tag{6.14}$$

with high probability as $n \to +\infty$.

6.3 Proof of Theorem 6.2.10

In this section we establish Theorem 6.2.10. We first provide a proof techniques section and then establish in separate subsections the lower and upper bounds of (6.13). Finally an intermediate subsection is devoted to certain key lemmas for the proof.

6.3.1 Roadmap

For $\gamma \in (\frac{1}{2}, 1)$ let $Z_{K,\gamma}$ the random variable that counts the number of K-vertex subgraphs of $G \sim G(n, \frac{1}{2})$ with edge density at least γ (equivalently with number of edges at least $\gamma\binom{K}{2}$), that is

$$Z_{K,\gamma} := \sum_{A \subset V(G): |A|=K} 1\left(|E[A]| \geq \gamma\binom{K}{2}\right). \tag{6.15}$$

Markov's inequality (on the left) and Paley-Zygmund inequality (on the right) give

$$\mathbb{E}\left[Z_{K,\gamma}\right] \geq \mathbb{P}\left[Z_{K,\gamma} \geq 1\right] \geq \frac{\mathbb{E}\left[Z_{K,\gamma}\right]^2}{\mathbb{E}\left[Z_{K,\gamma}^2\right]}. \tag{6.16}$$

(6.16) has two important implications.

First if for some $\gamma > 0$,

$$\lim_n \mathbb{E}\left[Z_{K,\gamma}\right] = 0$$

then (6.16) gives $Z_{K,\gamma} = 0$ w.h.p. as $n \to +\infty$ and therefore the densest K-subgraph has at most $\gamma\binom{K}{2}$ edges w.h.p. as $n \to +\infty$. This is called *the first moment method* for the random variable $Z_{k,\gamma}$.

Second if for some $\gamma > 0$,

$$\lim_n \frac{\mathbb{E}\left[Z_{k,\gamma}\right]^2}{\mathbb{E}\left[Z_{K,\gamma}^2\right]} = 1$$

then $Z_{K,\gamma} \geq 1$ w.h.p. as $n \to +\infty$ and therefore the densest-K subgraph has at least $\gamma\binom{K}{2}$ edges w.h.p. as $n \to +\infty$. This is called *the second moment method* for the random variable $Z_{k,\gamma}$.

Combining the two observations and a Taylor Expansion result described in Lemma 6.3.3, to

establish Theorem 6.2.10 it suffices to establish for some $\alpha \leq \beta(C) - \frac{1}{2}$ and

$$\gamma = h^{-1}\left(\log 2 - \frac{\log\binom{n}{K} - O\left(K^{\alpha}\log n\right)}{\binom{K}{2}}\right),$$

that it holds

$$\lim_n \mathbb{E}\left[Z_{K,\gamma}\right] = 0, \lim_n \frac{\mathbb{E}\left[Z_{K,\gamma}\right]^2}{\mathbb{E}\left[Z_{K,\gamma}^2\right]} = 1.$$

We establish the upper bound provided in Theorem 6.2.10 exactly in this way, by showing that for $\alpha = 0$ and $\gamma = h^{-1}\left(\log 2 - \frac{\log\binom{n}{K}}{\binom{K}{2}}\right)$ it holds $\lim_n \mathbb{E}\left[Z_{K,\gamma}\right] = 0$. We present this argument in Subsection 6.3.2.

The lower bound appears much more challenging to obtain. A crucial difficulty is that by writing $Z_{K,\gamma}$ as a sum of indicators as in (6.15) and expanding $\mathbb{E}\left[Z_{K,\gamma}^2\right]$ we need to control various complicated ways that two dense K-subgraphs overlap. This is not an uncommon difficulty in the literature of second moment method applications where certain conditioning is usually necessary for the second moment method to provide tight results (see e.g. [BMR$^+$18], [GZ17a], [WX18], [BPW18], [RXZ19] and references therein).

To control the ways dense subgraphs overlap we follow a similar, but not identical, path to [BBSV18] which analyzed the K-densest subgraph problem for $K = \Theta\left(\log n\right)$ and also used a conditioning technique. We do not analyze directly the second moment of $Z_{K,\gamma}$ but instead we focus on the second moment for another counting random variable that counts sufficiently dense subgraphs satisfying also an additional *flatness condition.* The condition is established to hold with high probability under the Erdős-Rényi structure (Lemma 6.3.4) and under this condition the dense subgraphs overlap in more "regular" ways leading to an easier control of the second moment. More details and the analysis of the second moment method under the flatness condition are in Subsections 6.3.3 and 6.3.4.

6.3.2 Proof of the Upper Bound

Using (6.16) it suffices to show that for $\gamma := h^{-1}\left(\log 2 - \frac{\log\binom{n}{K}}{\binom{K}{2}}\right)$, $\mathbb{E}[Z_{K,\gamma}] = o(1)$.

We have by linearity of expectation and (6.15)

$$\begin{aligned}\mathbb{E}\left[Z_{K,\gamma}\right] &= \binom{n}{K}\mathbb{P}\left[|\mathrm{E}[A]| \geq \gamma\binom{K}{2}\right], \text{for some } A \subseteq V(G), |A| = K \\ &= \binom{n}{K}\mathbb{P}\left[\mathrm{Bin}\left(\binom{K}{2}, \frac{1}{2}\right) \geq \gamma\binom{K}{2}\right] \end{aligned} \tag{6.17}$$

Using the elementary inequality $\binom{n}{K} \leq n^K$ we have

$$\frac{\log\binom{n}{K}}{\binom{K}{2}} = O\left(\frac{\log n}{K}\right) = o(1) \tag{6.18}$$

since by our assumption $\omega(\log n) = K$.

By Lemma 6.8.3 and (6.18) we have,

$$\gamma = \frac{1}{2} + \Omega\left(\sqrt{\frac{\log\binom{n}{K}}{\binom{K}{2}}}\right) = \frac{1}{2} + o(1).$$

Therefore $\lim_n \gamma = \frac{1}{2}$ and by Stirling's approximation,

$$\left(\gamma - \frac{1}{2}\right)\sqrt{\binom{K}{2}} = \Omega\left(\sqrt{\log\binom{n}{K}}\right) = \Omega\left(\sqrt{K\log\frac{n}{K}}\right) = \omega(1).$$

Hence both assumptions of Lemma 6.8.2 are satisfied and hence (6.17) implies

$$\mathbb{E}\left[Z_{K,\gamma}\right] \leq \binom{n}{K} O\left(\exp\left(-\binom{K}{2} r(\gamma, \frac{1}{2}) - \Omega\left(\sqrt{K\log\frac{n}{K}}\right)\right)\right), \tag{6.19}$$

where recall that $r(\gamma, \frac{1}{2})$ is defined in (6.7). Now notice that for our choice of γ,

$$r(\gamma, \frac{1}{2}) = \log 2 - h(\gamma) = \frac{\log\binom{n}{K}}{\binom{K}{2}}.$$

In particular using (6.19) we conclude that

$$\mathbb{E}\left[Z_{K,\gamma}\right] = \exp\left(-\Omega\left(\sqrt{K\log\frac{n}{K}}\right)\right) = o(1). \tag{6.20}$$

The completes the proof of the upper bound.

6.3.3 (γ, δ)-flatness and auxiliary lemmas

We start with appropriately defining the flatness condition mention in Subsection 6.3.1. Specifically, for $K \in \mathbb{N}$ we introduce a notion of a (γ, δ)-flat K-vertex graph G, where $\gamma, \delta \in (0,1)$. This generalizes the corresponding definition from [BBSV18, Section 3].

For $0 \leq \ell \leq K$ let

$$D_K(\ell, \delta) := \begin{cases} \sqrt{2\gamma(2+\delta)\min\left(\binom{K}{2} - \binom{\ell}{2}, \binom{\ell}{2}\right)\left(\log\binom{K}{\ell} + 2\log K\right)} & 0 \leq \ell < \frac{2K}{3} \\ \sqrt{2\gamma(1+\delta)\min\left(\binom{K}{2} - \binom{\ell}{2}, \binom{\ell}{2}\right)\left(\log\binom{K}{\ell} + 2\log K\right)} & \frac{2K}{3} \leq \ell \leq K \end{cases} \tag{6.21}$$

Definition 6.3.1 ((γ, δ)-flat graph)**.** *Call a K-vertex graph G, (γ, δ)-flat if*

- $|E[G]| = \left\lceil \gamma\binom{K}{2} \right\rceil$ *and*
- *for all $A \subset V(G)$ with $\ell = |A| \in \{2, 3, \ldots, K-1\}$ we have $|E[A]| \leq \left\lceil \gamma\binom{\ell}{2} \right\rceil + D_K(\ell, \delta)$.*

Notice that a (γ, δ)-flat subgraph of $G \sim G(n, \frac{1}{2})$ has edge density approximately γ and is constrained to do not have arbitrarily dense subgraphs. In particular, two (γ, δ)-flat subgraphs of G cannot overlap in "extremely" dense subgraphs. This property leads to an easier control of the second moment of the random variable which counts the number of (γ, δ)-flat subgraphs compared to the second moment of $Z_{K,\gamma}$ defined in Definition 6.15. Using the second moment method we establish the existence of an appropriate (γ, δ)-flat subgraph leading to the desired lower bound stated in Theorem 6.2.10. Even under the flatness restriction, the control of the second moment remains far from trivial and requires a lot of careful and technical computations. For this reason we devote the rest of this subsection on stating and proving four auxiliary lemmas. In the following subsection we provide the proof of the lower bound.

Lemma 6.3.2. *Let $\alpha \in (0,1)$. Suppose $K = \Theta(n^C)$ for $C \in (0,1)$.*
For any γ satisfying $\gamma = h^{-1}\left(\log 2 - \frac{\log\binom{n}{K} - O(K^\alpha \log n)}{\binom{K}{2}}\right)$ it holds

$$\gamma = \frac{1}{2} + (1 + o(1))\sqrt{\frac{\log\frac{n}{K}}{K}} = \frac{1}{2} + \Theta\left(\sqrt{\frac{\log n}{K}}\right).$$

Furthermore,

$$r\left(\gamma, \frac{1}{2}\right) = \log 2 - h(\gamma) = (1 + o(1))\frac{2\log\frac{n}{K}}{K} = \Theta\left(\frac{\log n}{K}\right).$$

Proof. We first observe that since $K = \Theta(n^C)$ for $C \in (0,1)$ by Stirling approximation we have $\log \binom{n}{K} = (1+o(1))\, K \log \frac{n}{K}$. Therefore, since $C < 1$ and $\alpha < 1$, it also holds

$$\frac{\log \binom{n}{K} - O\left(K^{\alpha} \log n\right)}{\binom{K}{2}} = (1+o(1))\, \frac{K \log \frac{n}{K}}{\frac{K^2}{2}} = (1+o(1))\, \frac{2 \log \frac{n}{K}}{K}.$$

Hence γ satisfies

$$\gamma = h^{-1}\left(\log 2 - (1+o(1))\, \frac{2 \log \frac{n}{K}}{K}\right). \tag{6.22}$$

By Lemma 6.8.3 we have $h^{-1}(\log 2 - \epsilon) = \frac{1}{2} + \left(\frac{1}{\sqrt{2}} + o(1)\right)\sqrt{\epsilon}$. Since $\frac{2 \log \frac{n}{K}}{K} = o(1)$ we have that

$$\gamma = \frac{1}{2} + (1+o(1))\sqrt{\frac{\log \frac{n}{K}}{K}} = \frac{1}{2} + \Theta\left(\sqrt{\frac{\log n}{K}}\right).$$

Furthermore by (6.22) we directly have

$$r\left(\gamma, \frac{1}{2}\right) = \log 2 - h(\gamma) = (1+o(1))\, \frac{2 \log \frac{n}{K}}{K} = \Theta\left(\frac{\log n}{K}\right).$$

□

Lemma 6.3.3. *Suppose $\omega(\log n) = K = o(\sqrt{n})$. Then for any fixed $\alpha \in (0,1)$,*

$$h^{-1}\left(\log 2 - \frac{\log \binom{n}{K} - O\left(K^{\alpha} \log n\right)}{\binom{K}{2}}\right)\binom{K}{2} = h^{-1}\left(\log 2 - \frac{\log \binom{n}{K}}{\binom{K}{2}}\right)\binom{K}{2} - O\left(K^{\alpha+\frac{1}{2}}\sqrt{\log n}\right).$$

Proof. Equivalently we need to show that

$$h^{-1}\left(\log 2 - \frac{\log \binom{n}{K} - O\left(K^{\alpha} \log n\right)}{\binom{K}{2}}\right) = h^{-1}\left(\log 2 - \frac{\log \binom{n}{K}}{\binom{K}{2}}\right) - O\left(K^{\alpha-\frac{3}{2}}\sqrt{\log n}\right).$$

Now from Lemma 6.8.3 we know that for $\epsilon = o(1)$, $h^{-1}(\log 2 - \epsilon) = \frac{1}{2} + \Theta(\sqrt{\epsilon})$. By Stirling approximation since $K = o(\sqrt{n})$ we have $\binom{n}{K} = \Theta\left(\left(\frac{ne}{K}\right)^K\right)$. Using $\alpha \in (0,1)$,

$$\log \binom{n}{K} = \Theta\left(K \log\left(\frac{ne}{K}\right)\right) = \omega\left(K^{\alpha} \log n\right).$$

Hence,

$$\left|\frac{\log\binom{n}{K} - O\left(K^{\alpha}\log n\right)}{\binom{K}{2}}\right| = O\left(\frac{\log n}{K}\right) = o\left(1\right).$$

Therefore by Lemma 6.8.3

$$\begin{aligned}
&h^{-1}\left(\log 2 - \frac{\log\binom{n}{K}}{\binom{K}{2}}\right) - h^{-1}\left(\log 2 - \frac{\log\binom{n}{K} - O\left(K^{\alpha-2}\log n\right)}{\binom{K}{2}}\right)\\
&= \Theta\left(\sqrt{\frac{\log\binom{n}{K}}{\binom{K}{2}}} - \sqrt{\frac{\log\binom{n}{K} - O(K^{\alpha}\log n)}{\binom{K}{2}}}\right)\\
&= O\left(\frac{K^{\alpha-2}\log n}{\sqrt{\frac{\log\binom{n}{K}}{\binom{K}{2}}}}\right), \text{ using } \sqrt{a} - \sqrt{b} = (a-b)\,/\left(\sqrt{a} + \sqrt{b}\right)\\
&= O\left(K^{\alpha-\frac{3}{2}}\sqrt{\log n}\right).
\end{aligned}$$

The proof of the Lemma is complete.

□

The lemma below generalizes Lemma 4 from [BBSV18].

Lemma 6.3.4. *Let $\gamma, \delta \in (0,1)$. Suppose G' is an Erdős-Rényi $G\left(K, \frac{1}{2}\right)$ conditioned on having $\left\lceil \gamma\binom{K}{2} \right\rceil$ edges. Then G' is (γ,δ)-flat (defined in Definition 6.3.1) w.h.p. as $K \to +\infty$.*

Proof. For any $C \subset V(G)$, let $e\left(C\right) := |E[C]|/\binom{|C|}{2}$.

Consider any $2 \leq \ell \leq K-1$ and any $C \subset V(G)$ with $|C| = \ell$. By identical reasoning we have from equation (4), page 6 in [BBSV18] that for any $r > 0$,

$$\mathbb{P}\left(\left|E\left[C\right]\right| \geq \gamma\binom{\ell}{2} + r\right) \leq \exp\left(-\frac{r(r-1)}{2\gamma\min(\binom{K}{2} - \binom{\ell}{2}, \binom{\ell}{2})}\right).$$

Therefore by union bound,

$$\begin{aligned}
&\binom{K}{\ell}\mathbb{P}\left(|E\left[C\right]| \geq \gamma\binom{\ell}{2} + D_K\left(\ell,\delta\right)\right) \\
&\leq \binom{K}{\ell}\sum_{r=D_K(\ell,\delta)}^{\binom{\ell}{2}} \exp\left(-\frac{r(r-1)}{2\gamma\min(\binom{K}{2}-\binom{\ell}{2},\binom{\ell}{2})}\right) \\
&\leq \exp\left(\log\binom{K}{\ell} + \log\binom{\ell}{2} - \frac{\left(D_K\left(\ell,\delta\right)-1\right)^2}{2\gamma\min(\binom{K}{2}-\binom{\ell}{2},\binom{\ell}{2})}\right) \\
&\leq \exp\left(\log\binom{K}{\ell} + 2\log K - \frac{\left(D_K\left(\ell,\delta\right)-1\right)^2}{2\gamma\min(\binom{K}{2}-\binom{\ell}{2},\binom{\ell}{2})}\right).
\end{aligned}$$

Therefore plugging in the value for $D_K\left(\ell,\delta\right)$ we conclude that for $\ell < \frac{2K}{3}$,

$$\binom{K}{\ell}\mathbb{P}\left(|E\left[C\right]| \geq \gamma\binom{\ell}{2} + D_K\left(\ell,\delta\right)\right) \leq \exp(-(1+\delta)\log\binom{K}{\ell})$$

and for $\ell \geq \frac{2K}{3}$,

$$\binom{K}{\ell}\mathbb{P}\left(|E\left[C\right]| \geq \gamma\binom{\ell}{2} + D_K\left(\ell,\delta\right)\right) \leq \exp(-\delta\log\binom{K}{\ell}).$$

Using union bound and the above two inequalities we have that G' is not (γ,δ)-flat with probability at most

$$\begin{aligned}
&\sum_{\ell=2}^{K-1}\binom{K}{\ell}\mathbb{P}\left(|E\left[C\right]| \geq \gamma\binom{\ell}{2} + D_K\left(\ell,\delta\right)\right) \\
&\leq \sum_{\ell=1}^{\lfloor\frac{2K}{3}\rfloor}\binom{K}{\ell}^{-1-\delta} + \sum_{\ell=\lceil\frac{2K}{3}\rceil}^{K-1}\binom{K}{\ell}^{-\delta}
\end{aligned} \tag{6.23}$$

Using now that for ℓ satisfying $\frac{2k}{3} \leq \ell \leq K - K^{\frac{\delta}{2}}$ we have

$$\binom{K}{\ell} = \binom{K}{K-\ell} \geq \left(\frac{K}{K-\ell}\right)^{K-\ell} \geq 3^{K-\ell} \geq 3^{K^{\frac{\delta}{2}}}$$

and otherwise if $\ell \leq \frac{2K}{3}$, $\binom{K}{\ell} \geq \binom{K}{1} = K$ the right hand side of (6.23) is at most

$$KK^{-1-\delta} + K3^{-K^{\frac{\delta}{2}}} + K^{\frac{\delta}{2}}K^{-\delta} \leq K^{-\delta} + K3^{-K^{\frac{\delta}{2}}} + K^{-\frac{\delta}{2}}$$

which is $o(1)$. The proof of the Lemma is complete.

□

Assume $G \sim G\left(n, \frac{1}{2}\right)$ and $K \leq n$. For $2 \leq \ell \leq K-1, 0 \leq L \leq \binom{\ell}{2}$ and $A, B \subset V(G)$ with $|A| = K, |B| = K$ and $|A \cap B| = \ell$ let

$$g_\ell(L) := \mathbb{P}\left(|\mathrm{E}[A]| = |\mathrm{E}[B]| = \lceil \gamma \binom{K}{2} \rceil, |\mathrm{E}[|A \cap B|]| = L\right). \tag{6.24}$$

Lemma 6.3.5. *For $2 \leq \ell \leq K-1$ and $\gamma \in (\frac{1}{2}, 1)$ let $\lambda := \exp\left(\frac{2\gamma-1}{1-\gamma} + \frac{1}{\gamma\left[\binom{K}{2} - \binom{\ell}{2}\right]}\right)$. Then*

(1) for any $r \geq 0$,

$$\frac{g_\ell(\lceil \gamma\binom{\ell}{2} \rceil + r)}{\mathbb{P}\left(|\mathrm{E}[A]| = \lceil \gamma\binom{K}{2} \rceil\right)^2} \leq \lambda^r \exp\left(\binom{\ell}{2} r(\gamma, \frac{1}{2}) + O(1)\right).$$

(2) for any $r \leq 0$,

$$\frac{g_\ell(\lceil \gamma\binom{\ell}{2} \rceil + r)}{\mathbb{P}\left(|\mathrm{E}[A]| = \lceil \gamma\binom{K}{2} \rceil\right)^2} \leq \exp\left(\binom{\ell}{2} r(\gamma, \frac{1}{2}) + O(1)\right).$$

Proof. This follows from the proof of [BBSV18, Lemma 6] for $p = \frac{1}{2}$ and minor adjustment in the choice of λ. The minor adjustment is justified by the second displayed equation on Page 9 in the aforementioned paper. In that equation if we apply the elementary inequality $1 + x \leq e^x$ once for $x = \frac{2\gamma-1}{1-\gamma}$ and once for $x = \frac{1}{\gamma\left[\binom{K}{2} - \binom{\ell}{2}\right]}$ we obtain the new choice of λ. With this modification, following the proof of [BBSV18, Lemma 6], *mutatis mutandis*, gives the Lemma. □

6.3.4 Proof of the Lower Bound

We turn now to the lower bound of (6.13).

For $\gamma \in (\frac{1}{2}, 1)$ we again define $Z_{K,\gamma}$ as in (6.15). Furthermore for any $\delta > 0$, let $\hat{Z}_{K,\gamma,\delta}$ the

random variable that counts the number of (γ, δ)-flat K-vertex subgraphs of G;

$$\hat{Z}_{K,\gamma,\delta} := \sum_{A \subset V(G): |A|=K} 1\left(A \text{ is } (\gamma,\delta)\text{-flat}\right). \tag{6.25}$$

Notice that clearly by definition of (γ, δ)-flatness we have that for any choice of K, γ and any $\delta > 0$ almost surely

$$Z_{K,\gamma} \geq \hat{Z}_{K,\gamma,\delta}. \tag{6.26}$$

We establish the following proposition.

Proposition 6.3.6. *Suppose that $K = \Theta(n^C)$ for some constant $C \in (0, \frac{1}{2})$. Let any $\alpha \in (0,1)$ satisfying*

$$\alpha > 1 - \left(\frac{5}{2} - \sqrt{6}\right)\frac{1-C}{C} \tag{6.27}$$

and set

$$\gamma = h^{-1}\left(\log 2 - \frac{\log\binom{n}{K} - K^{\alpha}\log n}{\binom{K}{2}}\right).$$

Then there exists $\delta > 0$ small enough such that

$$\frac{\mathbb{E}\left[\left(\hat{Z}_{K,\gamma,\delta}\right)^2\right]}{\mathbb{E}\left[\hat{Z}_{K,\gamma,\delta}\right]^2} = 1 + o(1). \tag{6.28}$$

In particular, $Z_{K,\gamma} \geq \hat{Z}_{K,\gamma,\delta} \geq 1$ with high probability as $n \to +\infty$.

Using this proposition for $\alpha := \beta(C) + \frac{1}{2}$ and the Taylor expansion argument from Lemma 6.3.3 we conclude the desired lower bound of Theorem 6.2.10.

Proof of Proposition 6.3.6. Notice that $\hat{Z}_{K,\gamma,\delta} \geq 1$ with high probability as $n \to +\infty$ follows by (6.28) using Paley-Zigmund inequality. Thus we focus on establishing (6.28).

We begin by choosing $\delta > 0$ to satisfy

$$1 - C(2\alpha - 1) + 4(\sqrt{(1-\alpha)} + \delta)\sqrt{C(1-C)} - 2(1-C) < 0. \tag{6.29}$$

To establish the existence of such δ notice that (6.27) by elementary algebra is equivalent with

$$C(1-\alpha) < (\sqrt{\frac{3}{2}} - 1)^2(1-C)$$

or

$$\sqrt{C(1-\alpha)} + \sqrt{1-C} < \sqrt{\frac{3}{2}(1-C)}$$

which by squaring both sides yields

$$C(1-\alpha) + 1 - C + 2\sqrt{(1-\alpha)}\sqrt{C(1-C)} < \frac{3}{2}(1-C)$$

or equivalently by multiplying both sides by 2 and rearranging

$$1 - C(2\alpha - 1) + 4\sqrt{(1-\alpha)}\sqrt{C(1-C)} - 2(1-C) < 0.$$

Now, since $C \in (0,1)$, the last inequality implies the existence of some sufficiently small $\delta > 0$ such that (6.29) holds.

For an arbitrary K-vertex subset $A \subseteq V(G)$ and linearity of expectation, (6.25) gives

$$\begin{aligned}
\mathbb{E}[\hat{Z}_{K,\gamma,\delta}] &= \binom{n}{K}\mathbb{P}\left(A \text{ is } (\gamma,\delta)\text{-flat}\right) \\
&= (1-o(1))\binom{n}{K}\mathbb{P}\left(|\mathrm{E}[A]| = \left\lceil \gamma\binom{K}{2}\right\rceil\right), \text{ using Lemma 6.3.4} \\
&= \binom{n}{K}\exp\left(-\binom{K}{2}r(\gamma,\frac{1}{2}) - \frac{1}{2}\log\binom{K}{2} + O(1)\right), \text{ using Lemma 6.8.2} \\
&= \exp\left(\log\binom{n}{K} - \binom{K}{2}r(\gamma,\frac{1}{2}) - \frac{1}{2}\log K + O(1)\right). \qquad (6.30)
\end{aligned}$$

Using that for our choice of γ,

$$r(\gamma,\frac{1}{2}) = \frac{\log\binom{n}{K} - K^\alpha \log n}{\binom{K}{2}}$$

we conclude that,

$$\mathbb{E}[\hat{Z}_{K,\gamma,\delta}] = \exp\left(K^\alpha \log n - \frac{1}{2}\log K + O(1)\right) = \exp\left(\Omega\left(K^\alpha \log n\right)\right), \tag{6.31}$$

since $K^\alpha = \Theta(n^{C\alpha}) = \omega(1)$.

We now proceed to the second moment calculation. For $A \subset V(G)$ with $|A| = K$ define the events $E_A := \{A \text{ is } (\gamma,\delta)\text{-flat}\}$ and $E'_A := \{|E[A]| = \lceil \gamma \binom{K}{2} \rceil\}$. Note

$$\hat{Z}_{K,\gamma,\delta} = \sum_{A \subset V(G), |A|=K} 1(E_A).$$

For $\ell = |A \cap B|$ we have via standard expansion,

$$\begin{aligned}
&\frac{\mathbb{E}[(\hat{Z}_{K,\gamma,\delta})^2]}{\mathbb{E}[\hat{Z}_{K,\gamma,\delta}]^2} - 1 \\
&= \frac{\mathbb{E}[(\hat{Z}_{K,\gamma,\delta})^2] - \mathbb{E}[\hat{Z}_{K,\gamma,\delta}]^2}{\mathbb{E}[\hat{Z}_{K,\gamma,\delta}]^2} \\
&= \sum_{\ell=2}^{K} \binom{K}{\ell}\binom{n-K}{K-\ell}\binom{n}{K}^{-1} \frac{\mathbb{P}\left(E_A \cap E_B\right) - \mathbb{P}\left(E_A\right)^2}{\mathbb{P}\left(E_A\right)^2} \\
&= \sum_{\ell=2}^{K-1} \binom{K}{\ell}\binom{n-K}{K-\ell}\binom{n}{K}^{-1} \frac{\mathbb{P}\left(E_A \cap E_B\right) - \mathbb{P}\left(E_A\right)^2}{\mathbb{P}\left(E_A\right)^2} + \frac{1 - \mathbb{P}\left(E_A\right)}{\mathbb{E}[\hat{Z}_{K,\gamma,\delta}]} \\
&\leq \sum_{\ell=2}^{K-1} \binom{K}{\ell}\binom{n-K}{K-\ell}\binom{n}{K}^{-1} \frac{\mathbb{P}\left(E_A \cap E_B\right)}{\mathbb{P}\left(E_A\right)^2} + o(1), \text{ since by (6.31) } \mathbb{E}[\hat{Z}_{k,\gamma,\delta}] = \omega(1) \\
&\leq (1 + o\left(1\right)) \sum_{\ell=2}^{K-1} \binom{K}{\ell}\binom{n-K}{K-\ell}\binom{n}{K}^{-1} \frac{\mathbb{P}\left(E_A \cap E_B\right)}{\mathbb{P}\left(E'_A\right)^2} + o(1), \text{ from Lemma 6.3.4.}
\end{aligned}$$

Now for fixed $\ell \in \{2, 3, \ldots, K-1\}$ and (γ,δ)-flat K-subgraphs A, B with $\ell = |A \cap B|$ we have from the definition of (γ,δ)-flatness that the graph induced by $A \cap B$ contains at most

$\lceil \gamma\binom{k}{2} \rceil + D_K(\ell, \delta)$ edges. In particular,

$$\begin{aligned}
\frac{\mathbb{P}\left(E_A \cap E_B\right)}{\mathbb{P}\left(E'_A\right)^2} &= \sum_{L=0}^{\lceil \gamma\binom{\ell}{2} \rceil + D_K(\ell,\delta)} \frac{\mathbb{P}\left(E_A \cap E_B, E[A \cap B] = L\right)}{\mathbb{P}\left(E'_A\right)^2} \\
&\leq \sum_{L=0}^{\lceil \gamma\binom{\ell}{2} \rceil + D_K(\ell,\delta)} \frac{\mathbb{P}\left(E'_A \cap E'_B, E[A \cap B] = L\right)}{\mathbb{P}\left(E'_A\right)^2}, \text{ using that } E_A \subseteq E'_A, E_B \subseteq E'_B \\
&= \sum_{L=0}^{\lceil \gamma\binom{\ell}{2} \rceil + D_K(\ell,\delta)} \frac{g_\ell(L)}{\mathbb{P}\left(E'_A\right)^2}, \text{ using notation (6.24)} \\
&\leq \sum_{L=0}^{\lceil \gamma\binom{\ell}{2} \rceil + D_K(\ell,\delta)} \lambda^{D_K(\ell,\delta)} \exp\left(\binom{\ell}{2} r(\gamma, \frac{1}{2}) + O(1)\right), \text{ using Lemma 6.3.5 and } \lambda \geq 1 \\
&\leq \binom{\ell}{2} \lambda^{D_K(\ell,\delta)} \exp\left(\binom{\ell}{2} r(\gamma, \frac{1}{2}) + O(1)\right) \\
&= \exp\left(D_K(\ell, \delta) \log \lambda + \binom{\ell}{2} r(\gamma, \frac{1}{2}) + O(\log \ell)\right).
\end{aligned}$$

Therefore we conclude

$$\begin{aligned}
\frac{\mathbb{E}[(\hat{Z}_{K,\gamma,\delta})^2]}{\mathbb{E}[\hat{Z}_{K,\gamma,\delta}]^2} \leq 1 + \sum_{\ell=2}^{K-1} \binom{K}{\ell}\binom{n-K}{K-\ell}\binom{n}{K}^{-1} \exp\left(D_K(\ell, \delta) \log \lambda + \binom{\ell}{2} r(\gamma, \frac{1}{2}) + O(\log \ell)\right) \\
+ o(1).
\end{aligned}$$

We proceed from now on in two steps to complete the proof. First we show that for some sufficiently small constant $\delta_1 > 0$,

$$\sum_{\ell=2}^{\lfloor \delta_1 K \rfloor} \binom{K}{\ell}\binom{n-K}{K-\ell}\binom{n}{K}^{-1} \exp\left(D_K(\ell, \delta) \log \lambda + \binom{\ell}{2} r(\gamma, \frac{1}{2}) + O(\log \ell)\right) = o(1). \quad (6.32)$$

In the second step we show that for the constant $\delta_1 > 0$ chosen in the first step,

$$\sum_{\ell=\lceil \delta_1 K \rceil}^{K-1} \binom{K}{\ell}\binom{n-K}{K-\ell}\binom{n}{K}^{-1} \exp\left(D_K(\ell, \delta) \log \lambda + \binom{\ell}{2} r(\gamma, \frac{1}{2}) + O\left(\log n\right)\right) = o(1). \quad (6.33)$$

Note here that for these values of ℓ for the second step we have replaced $O(\log \ell)$ with the equivalent bound $O\left(\log K\right) = O\left(\log n\right)$ since $K = \Theta(n^C)$ for $C > 0$.

First Step, proof of (6.32): For the combinatorial term we use a simple inequality derived from Stirling's approximation (see e.g. page 11 in [BBSV18]),

$$\binom{K}{\ell}\binom{n-K}{K-\ell}\binom{n}{K}^{-1} \leq (1+o(1))\left(\frac{K^2}{n}\right)^{\ell}. \tag{6.34}$$

We now bound the terms in the exponent. Plugging in the value of λ from Lemma 6.3.5 we have

$$D_K(\ell,\delta)\log\lambda = \frac{D_K(\ell,\delta)}{\gamma\left[\binom{K}{2}-\binom{\ell}{2}\right]} + \frac{2\gamma-1}{1-\gamma}D_K(\ell,\delta).$$

By the definition of $D_K(\ell,\delta)$ (6.21) we have

$$\frac{D_K(\ell,\delta)}{\gamma\left[\binom{K}{2}-\binom{\ell}{2}\right]} = O\left(\sqrt{\frac{\log\binom{K}{\ell}+\log K}{\binom{K}{2}-\binom{\ell}{2}}}\right) \leq O\left(\sqrt{\frac{K}{\binom{K}{2}-\binom{K-1}{2}}}\right) = O(1),$$

since $\ell \leq \delta_1 K \leq K-1$, assuming $\delta_1 < 1$. From Lemma 6.3.2 we have $\gamma = \frac{1}{2} + \Theta\left(\sqrt{\frac{\log n}{K}}\right)$. Furthermore, by (6.21), $K = \Theta\left(n^C\right)$ and $\binom{K}{\ell} \leq K^{\ell}$ we have

$$D_K(\ell,\delta) = O\left(\sqrt{\ell^2\left(\log\binom{K}{\ell}+\log K\right)}\right) = O\left(\sqrt{\ell^3\log n}\right).$$

Combining the two last equalities we conclude

$$\frac{2\gamma-1}{1-\gamma}D_K(\ell,\delta) = O\left(\frac{\ell^{3/2}\log n}{\sqrt{K}}\right). \tag{6.35}$$

Finally, again by Lemma 6.3.2, $r(\gamma,\frac{1}{2}) = \Theta(\frac{\log n}{K})$ and therefore

$$\binom{\ell}{2}r(\gamma,\frac{1}{2}) = O\left(\frac{\ell^2\log n}{K}\right). \tag{6.36}$$

Combining (6.34), (6.35) and (6.36) we conclude that for any $\delta_1 > 0$, supposing $\ell < \delta_1 K$ we get

$$\binom{K}{\ell}\binom{n-K}{K-\ell}\binom{n}{K}^{-1}\exp\left(D_K(\ell,\delta)\log\lambda+\binom{\ell}{2}r(\gamma,\frac{1}{2})+O\left(\log\ell\right)\right) \tag{6.37}$$

$$=\exp\left[-\ell\log\left(\frac{n}{K^2}\right)+O\left(\frac{\ell^{3/2}\log n}{\sqrt{K}}\right)+O\left(\frac{\ell^2\log n}{K}\right)+O\left(\log\ell\right)\right]$$

$$=\exp\left[-\ell\log n\left(1-2C-O\left(\sqrt{\frac{\ell}{K}}\right)-O\left(\frac{\ell}{K}\right)-O\left(\frac{\log\ell}{\ell\log n}\right)\right)\right],\ \text{using } K=\Theta(n^C)$$

$$\leq\exp\left[-\ell\log n\left(1-2C-O\left(\sqrt{\delta_1}\right)-O\left(\delta_1\right)-O\left(\frac{\log\ell}{\ell\log n}\right)\right)\right],\ \text{using } \ell\leq\delta_1 K$$

$$\leq\exp\left[-\ell\log n\left(1-2C-O\left(\sqrt{\delta_1}\right)-O\left(\delta_1\right)-O\left(\frac{1}{\log n}\right)\right)\right], \tag{6.38}$$

where we have used $\log\ell\leq\ell$ for all $\ell\geq 1$. Since $C<\frac{1}{2}$ we choose $\delta_1>0$ small enough but constant such that for some $\delta_2>0$ and large enough n,

$$1-2C-O\left(\sqrt{\delta_1}\right)-O\left(\delta_1\right)-O\left(\frac{1}{\log n}\right)>\delta_2. \tag{6.39}$$

Hence for this choice of constants $\delta_1,\delta_2>0$ if $\ell\leq\delta_1 K$ using (6.38) and (6.39) we conclude that the expression (6.37) is at upper bounded by

$$\exp\left(-\delta_2\ell\log n\right)=n^{-\delta_2\ell}.$$

Therefore we have,

$$\sum_{\ell=2}^{\lfloor\delta_1 K\rfloor}\binom{K}{\ell}\binom{n-K}{K-\ell}\binom{n}{K}^{-1}\exp\left(D_K(\ell,\delta)\log\lambda+\binom{\ell}{2}r(\gamma,\frac{1}{2})+O\left(\log\ell\right)\right)\leq\sum_{\ell\geq 1}n^{-\delta_2\ell}=O\left(n^{-\delta_2}\right).$$

This completes the proof of (6.32).

Second step, proof of (6.33): For the second step we start by multiplying both numerator and denominator of the left hand side of (6.33) with the two sides of (6.30); $\mathbb{E}\left[\hat{Z}_{K,\gamma,\delta}\right]=$

$\binom{n}{K}\exp\left(-\binom{K}{2}r(\gamma,\frac{1}{2})+O\left(\log n\right)\right)$, to get that it suffices to show

$$\frac{1}{\mathbb{E}[\hat{Z}_{K,\gamma,\delta}]}\sum_{\ell=\lceil\delta_1 K\rceil}^{K-1}\binom{K}{\ell}\binom{n-K}{K-\ell}\exp\left(D_K(\ell,\delta)\log\lambda-\left(\binom{K}{2}-\binom{\ell}{2}\right)r(\gamma,\frac{1}{2})+O\left(\log n\right)\right)=o(1).$$

Since by equation (6.31) we have $\mathbb{E}[\hat{Z}_{k,\gamma,\delta}]\geq\exp\left(D_0K^\alpha\log n\right)$ for some universal constant $D_0>0$ and $K=\omega(1)$ it suffices that

$$\sum_{\ell=\lceil\delta_1 K\rceil}^{K-1}\binom{K}{\ell}\binom{n-K}{K-\ell}\exp\left(D_K(\ell,\delta)\log\lambda-\left(\binom{K}{2}-\binom{\ell}{2}\right)r(\gamma,\frac{1}{2})-D_0K^\alpha\log n\right)=o(1).$$

Plugging in the value of λ we have

$$\sum_{\ell=\lceil\delta_1 K\rceil}^{K-1}\binom{K}{\ell}\binom{n-K}{K-\ell}\exp\left(D_K(\ell,\delta)\log\lambda-\left(\binom{K}{2}-\binom{\ell}{2}\right)r(\gamma,\frac{1}{2})-D_0K^\alpha\log n\right)$$

which is of the order

$$\sum_{\ell=\lceil\delta_1 K\rceil}^{K-1}\binom{K}{\ell}\binom{n-K}{K-\ell}\times\exp\left[\frac{D_K(\ell,\delta)}{\gamma\left[\binom{K}{2}-\binom{\ell}{2}\right]}+\frac{2\gamma-1}{1-\gamma}D_K(\ell,\delta)-\left(\binom{K}{2}-\binom{\ell}{2}\right)r(\gamma,\frac{1}{2})-D_0K^\alpha\sqrt{\log n}\right]$$

By (6.21) we have

$$\frac{D_K(\ell,\delta)}{\gamma\left[\binom{K}{2}-\binom{\ell}{2}\right]}=O\left(\sqrt{\frac{\log\binom{K}{\ell}+\log K}{\binom{K}{2}-\binom{\ell}{2}}}\right)\leq O\left(\sqrt{\frac{K}{\binom{K}{2}-\binom{K-1}{2}}}\right)=O\left(1\right),\tag{6.40}$$

since $\ell\leq K-1$. Furthermore by Lemma 6.3.2,

$$\begin{aligned}\left(\binom{K}{2}-\binom{\ell}{2}\right)r(\gamma,\frac{1}{2})&\geq\left(\binom{K}{2}-\binom{\ell}{2}\right)(1-o(1))\frac{2\log(\frac{n}{K})}{K}\\&=\left(\binom{K}{2}-\binom{\ell}{2}\right)\frac{2\log(\frac{n}{K})}{K}-o\left((K-\ell)\log n\right)\end{aligned}\tag{6.41}$$

Hence, combining (6.40) and (6.41) we conclude that it suffices to show

$$\sum_{\ell=\lceil \delta_1 K \rceil}^{K-1} \exp\left[F(\ell)\right] = o(1) \tag{6.42}$$

where $F(\ell)$ equals

$$\begin{aligned} &\log(\binom{K}{\ell}\binom{n-K}{K-\ell}) + \frac{2\gamma-1}{1-\gamma} D_K(\ell,\delta) \\ &- \left(\binom{K}{2} - \binom{\ell}{2}\right) \frac{2\log(\frac{n}{K})}{K} - D_0 K^\alpha \log n + o\left((K-\ell)\log n\right). \end{aligned} \tag{6.43}$$

Now we separate three cases to study $F(\ell)$.

Case 1 (large values of ℓ) : We assume $K-1 \geq \ell \geq K - c_1 K^\alpha \log n$, where $c_1 > 0$ is a universal constant defined below.

In this case we bound the combinatorial term using $\binom{K}{\ell} \leq K^{K-\ell}$ and $\binom{n-K}{K-\ell} \leq n^{K-\ell}$ to conclude

$$\binom{K}{\ell}\binom{n-K}{K-\ell} \leq K^{K-\ell} n^{K-\ell} = \exp\left(O\left[(K-\ell)\log n\right]\right). \tag{6.44}$$

Furthermore,

$$\begin{aligned} \frac{2\gamma-1}{1-\gamma} D_K(\ell,\delta) &= O\left((2\gamma-1)\sqrt{\left(\binom{K}{2} - \binom{\ell}{2}\right)\left(\log\binom{K}{\ell} + \log K\right)}\right), \text{using (6.21)} \\ &= O\left((2\gamma-1)\sqrt{(K-\ell)\left(K+\ell-1\right)\left(\log\binom{K}{\ell} + \log K\right)}\right) \\ &\leq O\left(\sqrt{\frac{\log n}{K}}\sqrt{(K-\ell)K \cdot (K-\ell)\log K}\right), \text{from Lemma 6.3.2, } \binom{K}{\ell} \leq K^{K-\ell} \\ &\leq O\left[(K-\ell)\log n\right], \end{aligned} \tag{6.45}$$

Therefore using (6.45) and (6.46) for ℓ with $K-1 \geq \ell \geq K - c_1 K^\alpha \log n$ we have

$$\begin{aligned} F(\ell) &\leq O\left((K-\ell)\log n\right) - D_0 K^\alpha \log n \\ &\leq C\,(K-\ell)\log n - D_0 K^\alpha \log n, \text{ for some universal constant } C > 0 \\ &\leq (Cc_1 - D_0) K^\alpha \log n, \text{ by the assumption on } \ell \\ &\leq -\frac{D_0}{2} K^\alpha \log n, \text{ by choosing } c_1 := D_0/2C, \end{aligned}$$

which gives

$$\sum_{\ell=\lceil K - c_1 K^\alpha \log n\rceil}^{K-1} \exp\left[F(\ell)\right] = O\left(\exp\left(\log K - \frac{D_0}{2} K^\alpha \log n\right)\right) = o\left(1\right) \tag{6.46}$$

where the last equality is because $K = \omega(1)$.

Case 2 (moderate values of ℓ) : $(1-\delta')K \leq \ell \leq K - c_1 K^\alpha \log n$, where $c_1 > 0$ is defined in Case 1 and $\frac{1}{3} > \delta' > 0$ is a sufficiently small but constant positive number such that

$$1 - C(2\alpha - 1) + 4(\sqrt{(1-\alpha)} + \delta)\sqrt{C(1-C)} - 2(1-\delta')(1-C) < -\delta'. \tag{6.47}$$

Note that such a $\frac{1}{3} > \delta' > 0$ exists because of our choice of δ satisfying (6.29) and because $C < 1$.

We start with the standard $\binom{K}{\ell} \leq \left(\frac{Ke}{K-\ell}\right)^{K-\ell}$ and $\binom{n-K}{K-\ell} \leq \left(\frac{(n-K)e}{K-\ell}\right)^{K-\ell}$ to conclude

$$\log\left(\binom{K}{\ell}\binom{n-K}{K-\ell}\right) \leq (K-\ell)\log\left(\frac{nKe^2}{(K-\ell)^2}\right) \tag{6.48}$$

$$\leq \left(1 - C(2\alpha-1) + o(1)\right)(K-\ell)\log n, \tag{6.49}$$

where for the last step we used $K - \ell \geq \Omega\left(K^\alpha\right)$ and that $K = \Theta(n^C)$. Furthermore for this

values of ℓ we have $\ell > \frac{2K}{3}$. Therefore from (6.21),

$$\begin{aligned}
D_K(\ell,\delta) &\le (1+\delta)\sqrt{2\gamma\left(\binom{K}{2}-\binom{\ell}{2}\right)\log\left(2K\binom{k}{\ell}\right)} \\
&\le (1+\delta+o(1))\sqrt{\left(\binom{K}{2}-\binom{\ell}{2}\right)\log\left(2K\binom{K}{\ell}\right)}, \text{ using Lemma 6.3.2} \\
&\le (1+\delta+o(1)))\sqrt{\left(\binom{K}{2}-\binom{\ell}{2}\right)\left((K-\ell)\log(\frac{Ke}{K-\ell})+2\log K\right)} \\
&\le (1+\delta+o(1))\,(K-\ell)\,\sqrt{K\log\left(O\left(K^{1-\alpha}\right)\right)}, \qquad (6.50) \\
&\text{(where we used } \binom{K}{2}-\binom{\ell}{2} \le K(K-\ell), K-\ell \ge \Omega(K^{\alpha})) \\
&\le \left(\sqrt{1-\alpha}+\delta+o(1)\right)(K-\ell)\,\sqrt{K\log K} \qquad (6.51)
\end{aligned}$$

From Lemma 6.3.2 we have

$$\frac{2\gamma-1}{1-\gamma} = (4+o(1))\,\sqrt{\frac{\log\frac{n}{K}}{K}}.$$

Hence combining it with (6.52),

$$\begin{aligned}
\frac{2\gamma-1}{1-\gamma}D_K(\ell,\delta) &\le \left(\sqrt{1-\alpha}+\delta+o(1)\right)4(K-\ell)\sqrt{\frac{\log\frac{n}{K}}{K}}\sqrt{K\log K} \\
&= 4\left(\sqrt{(1-\alpha)}+\delta+o(1)\right)(K-\ell)\sqrt{\log(\frac{n}{K})\log K} \qquad (6.52)
\end{aligned}$$

Now by dropping the term $-D_0K^{\alpha}\log n < 0$, $F(\ell)$ is at most

$$\log(\binom{K}{\ell}\binom{n-K}{K-\ell}) + \frac{2\gamma-1}{1-\gamma}D_K(\ell,\delta) - \left(\binom{K}{2}-\binom{\ell}{2}\right)\frac{2\log(\frac{n}{K})}{K} + o\left((K-\ell)\log n\right).$$

which using (6.50), (6.53) is at most $1+o\left(1\right)$ times

$$
\begin{aligned}
&(K-\ell)\\
&\times\left[(1-C(2\alpha-1))\log n+4\left(\sqrt{(1-\alpha)}+\delta\right)\sqrt{\log(\frac{n}{K})\log K}-\frac{2\left(\binom{K}{2}-\binom{\ell}{2}\right)\log\frac{n}{K}}{K(K-\ell)}+o(\log n)\right]\\
&\leq(K-\ell)\\
&\times\left[(1-C(2\alpha-1))\log n+4\left(\sqrt{(1-\alpha)}+\delta\right)\sqrt{\log(\frac{n}{K})\log K}-2\left(1-\delta'\right)\log\frac{n}{K}+o(\log n)\right]\\
&=(K-\ell)\log n\left[1-C(2\alpha-1)+4\left(\sqrt{(1-\alpha)}+\delta\right)\frac{\sqrt{\log(\frac{n}{K})\log K}}{\log n}-2\left(1-\delta'\right)\frac{\log\frac{n}{K}}{\log n}+o(1)\right],
\end{aligned}
$$

where for the last inequality we used that for $\ell\geq(1-\delta')k$, $\binom{k}{2}-\binom{\ell}{2}\geq(1-\delta'-o\left(1\right))k(k-\ell)$. Using that $K=\Theta\left(n^{C}\right)$ we conclude,

$$
\begin{aligned}
&F(\ell)\leq\\
&\left[(1-C(2\alpha-1))+4\left(\sqrt{(1-\alpha)}+\delta\right)\sqrt{C(1-C)}-2\left(1-\delta'\right)\left(1-C\right)+o(1)\right](K-\ell)\log n.
\end{aligned}
$$

From (6.48) we know that for large n

$$
(1-C(2\alpha-1))+4\left(\sqrt{(1-\alpha)}+\delta\right)\sqrt{C(1-C)}-2\left(1-\delta'\right)\left(1-C\right)+o(1)<-\delta'.
$$

Therefore we conclude for all ℓ with $(1-\delta')K\leq\ell\leq K-c_{1}K^{\alpha}\log n$

$$
F(\ell)\leq-\delta'\left(K-\ell\right)\log n\leq-\Omega\left(K^{\alpha}\left(\log n\right)^{2}\right).
$$

Hence,

$$
\sum_{\ell=\lceil(1-\delta')K\rceil}^{\lfloor K-c_{1}K^{\alpha}\log n\rfloor}\exp\left[F(\ell)\right]=O\left(K\exp\left(-\Omega(K^{\alpha}(\log n)^{2}\right)\right)=O\left(\exp\left(\log n-\Omega(K^{\alpha}(\log n)^{2}\right)\right)=o(1),
\tag{6.53}
$$

where the last equality is because $K=\Theta(n^{C})$ for $C>0$.

Case 3 (small values of ℓ) : $\delta_1 K \le \ell \le (1-\delta')K$ where δ' is defined in Case 2 and δ_1 in Part 1.

Similar to (6.49) we have

$$\log\left(\binom{K}{\ell}\binom{n-K}{K-\ell}\right) \le (K-\ell)\log\left(\frac{nKe^2}{(K-\ell)^2}\right)$$
$$\le (1+o(1))\,(K-\ell)\log\frac{n}{K}, \tag{6.54}$$

where we have used for the last inequality that $\ell = \Theta(K)$.

Furthermore using (6.21) and Lemma 6.3.2 we have

$$\begin{aligned}
\frac{2\gamma-1}{1-\gamma}D_K(\ell,\delta) &\le O\left(\sqrt{\frac{\log n}{K}}\sqrt{\left(\binom{K}{2}-\binom{\ell}{2}\right)\left(\log\binom{K}{\ell}+\log K\right)}\right)\\
&\le O\left(\sqrt{\frac{\log n}{K}}\sqrt{\left(\binom{K}{2}-\binom{\ell}{2}\right)\left((K-\ell)\log(\frac{Ke}{K-\ell})+\log K\right)}\right)\\
&\le O\left(\sqrt{\frac{\log n}{K}}(K-\ell)\sqrt{K}\right), \text{using } K-\ell=\Theta(K), \binom{K}{2}-\binom{\ell}{2}\le K(K-\ell)\\
&= o\left((K-\ell)\log n\right).
\end{aligned}$$

Combining it with (6.55) we have that $F(\ell)$ is at most

$$\begin{aligned}
&(1+o(1))\left[(K-\ell)\log(\frac{n}{K})+o((K-\ell)\log n)-\frac{2\left(\binom{K}{2}-\binom{\ell}{2}\right)}{K}(\log\frac{n}{K})\right]\\
&\le^{(*)} (1+o(1))\,(K-\ell)\left[\log(\frac{n}{K})+o(\log n)-2\left(\frac{1+\delta_1}{2}\right)(\log\frac{n}{K})\right]\\
&\le (K-\ell)\log n\left(1-C-2\left(\frac{1+\delta_1}{2}\right)(1-C)+o(1)\right), \text{ using } K=\Theta(n^c)\\
&= (K-\ell)\log n\left(-\delta_1(1-C)+o(1)\right),
\end{aligned} \tag{6.55}$$

where to derive (*) we use that for $\ell \ge \delta_1 K$,

$$\binom{K}{2}-\binom{\ell}{2} \le (\frac{1+\delta_1+o(1)}{2})K(K-\ell).$$

Since $\delta_1(1-C) > 0$ we conclude from (6.56) that for all ℓ with $\delta_1 K \le \ell \le (1-\delta')K$ and large

enough n,

$$F(\ell) \leq -\Theta\left(K \log n\right)$$

Hence,

$$\sum_{\ell=\lceil \delta_1 K \rceil}^{\lfloor (1-\delta')K \rfloor} \exp\left[F(\ell)\right] \leq O\left(K \exp\left[(\log n - \Theta(K \log n))\right)\right) \leq O\left(\exp\left(\log n - \Theta(K \log n)\right)\right) = o\left(1\right). \tag{6.56}$$

Combining (6.47), (6.54) and (6.57) we conclude the proof of (6.33). This completes the proof of the Proposition and of the Theorem. □

6.4 Proofs for First Moment Curve Bounds

6.4.1 Proof of first part of Proposition 6.2.3

Proof of first part of Proposition 6.2.3. If $z = \bar{k} = k$ then trivially

$$d_{k,k}(G)(k) = |E[\mathcal{PC}]| = \binom{k}{2} = \Gamma_{k,k}(k)$$

almost surely.

Otherwise, we fix some $z \in \{\lfloor \frac{k\bar{k}}{n} \rfloor, \lfloor \frac{k\bar{k}}{n} \rfloor + 1, \ldots, k\}$. Since $z = \bar{k} = k$ does not hold, we have $z < \bar{k}$. For $\gamma \in (0,1)$ we consider the counting random variable

$$Z_{\gamma,z} := |\{A \subseteq V(G) : |A| = \bar{k}, |A \cap \mathcal{PC}| = z, |E[A]| \geq \binom{z}{2} + \gamma_z \left(\binom{\bar{k}}{2} - \binom{z}{2}\right)\}|.$$

By Markov's inequality, $\mathbb{P}\left[Z_{\gamma,z} \geq 1\right] \leq \mathbb{E}\left[Z_{\gamma,z}\right]$. In particular, if for some $\gamma_z > 0$ it holds

$$\sum_{z=\lfloor \frac{\bar{k}k}{n} \rfloor}^{k} \mathbb{E}\left[Z_{\gamma_z,z}\right] = o(1) \tag{6.57}$$

we conclude using a union bound that for all z, $Z_{\gamma_z,z} = 0$ w.h.p. as $n \to +\infty$ and in particular for all z, $d_{\bar{k},k}(G)(z) \leq \binom{z}{2} + \gamma_z \left(\binom{\bar{k}}{2} - \binom{z}{2}\right)$, w.h.p. as $n \to +\infty$. Therefore it suffices to show

that for

$$\gamma_z := h^{-1}\left(\log 2 - \frac{\log\left(\binom{k}{z}\binom{n-k}{\bar{k}-z}\right)}{\binom{\bar{k}}{2}-\binom{z}{2}}\right),$$

(6.58) holds. Notice that γ_z is well-defined exactly because $z = \bar{k} = k$ does not hold. We fix this choice of γ_z from now on.

Let us fix z. We start with bounding the expectation for arbitrary $\gamma > 0$. By linearity of expectation we have

$$\begin{aligned}
\mathbb{E}\left[Z_{\gamma_z,z}\right] &= \binom{k}{z}\binom{n-k}{\bar{k}-z}\mathbb{P}\left[|\mathrm{E}[A]| \geq \binom{z}{2} + \gamma_z\left(\binom{\bar{k}}{2}-\binom{z}{2}\right)\right], \text{where } |A| = \bar{k}, |A \cap \mathcal{PC}| = z \\
&= \binom{k}{z}\binom{n-k}{\bar{k}-z}\mathbb{P}\left[\binom{z}{2} + \mathrm{Bin}\left(\binom{\bar{k}}{2}-\binom{z}{2}\right) \geq \binom{z}{2} + \gamma_z\left(\binom{\bar{k}}{2}-\binom{z}{2}\right)\right] \\
&= \binom{k}{z}\binom{n-k}{\bar{k}-z}\mathbb{P}\left[\mathrm{Bin}\left(\binom{\bar{k}}{2}-\binom{z}{2}\right) \geq \gamma_z\left(\binom{\bar{k}}{2}-\binom{z}{2}\right)\right] \qquad (6.58)
\end{aligned}$$

Using the elementary inequalities

$$\binom{k}{z} \leq k^{k-z} \leq n^{\bar{k}-z}$$

and

$$\binom{n-k}{\bar{k}-z} \leq n^{\bar{k}-z},$$

we conclude

$$\frac{\log\left(\binom{k}{z}\binom{n-k}{\bar{k}-z}\right)}{\binom{\bar{k}}{2}-\binom{z}{2}} = O\left(\frac{\log n}{\bar{k}+z}\right) = o(1) \qquad (6.59)$$

by our assumption $\omega(\log n) = \bar{k}$.

By Lemma 6.8.3 and (6.60) we have,

$$\frac{1}{2} + \Omega\left(\sqrt{\frac{\log\left(\binom{k}{z}\binom{n-k}{\bar{k}-z}\right)}{\binom{\bar{k}}{2}-\binom{z}{2}}}\right) \leq \gamma_z \leq \frac{1}{2} + o(1).$$

Therefore $\lim_n \gamma_z = \frac{1}{2}$ and the elementary $\binom{n-k}{\bar{k}-z} \geq \left((n-k)/(\bar{k}-z)\right)^{\bar{k}-z}$ since $z < \bar{k}$,

$$\left(\gamma_z - \frac{1}{2}\right)\sqrt{\binom{\bar{k}}{2}-\binom{z}{2}} = \Omega\left(\sqrt{\log\left(\binom{k}{z}\binom{n-k}{\bar{k}-z}\right)}\right) = \Omega\left(\sqrt{(\bar{k}-z)\log\frac{n}{\bar{k}-z}}\right).$$

Hence both assumption of Lemma 6.8.2 are satisfied (notice that the Binomial distribution of interest is defined on population size $\binom{\bar{k}}{2} - \binom{z}{2}$) and hence (6.59) implies

$$\mathbb{E}\left[Z_{\gamma_z,z}\right] \leq \binom{k}{z}\binom{n-k}{\bar{k}-z} O\left(\exp\left(-\left(\binom{\bar{k}}{2} - \binom{z}{2}\right) r(\gamma, \frac{1}{2}) - \Omega\left(\sqrt{(\bar{k}-z)\log\frac{n}{\bar{k}-z}}\right)\right)\right) \tag{6.60}$$

Now notice that for our choice of γ_z,

$$r(\gamma, \frac{1}{2}) = \log 2 - h(\gamma) = \frac{\log\left(\binom{k}{z}\binom{n-k}{\bar{k}-z}\right)}{\binom{\bar{k}}{2} - \binom{z}{2}}.$$

In particular using (6.61) we conclude that for any z

$$\mathbb{E}\left[Z_{\gamma_z,z}\right] = \exp\left(-\Omega\left(\sqrt{(\bar{k}-z)\log\frac{n}{\bar{k}-z}}\right)\right). \tag{6.61}$$

Hence,

$$\begin{aligned}
&\sum_{z=\lfloor\frac{\bar{k}k}{n}\rfloor}^{k} \mathbb{E}\left[Z_{\gamma_z,z}\right] \\
&= \sum_{z=\lfloor\frac{\bar{k}k}{n}\rfloor}^{k} \exp\left(-\Omega\left(\sqrt{(\bar{k}-z)\log\frac{n}{\bar{k}-z}}\right)\right) \\
&= \sum_{z=\min\{k,\bar{k}-\left((\log n)^2\right)\}}^{k} \exp\left(-\Omega\left(\sqrt{(\bar{k}-z)\log\frac{n}{\bar{k}-z}}\right)\right) \\
&+ \sum_{z=\lfloor\frac{\bar{k}k}{n}\rfloor}^{\min\{k,\bar{k}-\left((\log n)^2\right)\}} \exp\left(-\Omega\left(\sqrt{(\bar{k}-z)\log\frac{n}{\bar{k}-z}}\right)\right) \\
&\leq (\log n)^2 \exp\left(-\Omega\left(\sqrt{\log n}\right)\right) + k\exp\left(-\Omega\left(\sqrt{(\log n)^{\frac{3}{2}}}\right)\right) \\
&\leq \exp\left(-\Omega\left(\sqrt{\log n}\right)\right) + k\exp\left(-\Omega\left(\sqrt{(\log n)^{\frac{3}{2}}}\right)\right) \\
&= o\left(1\right),
\end{aligned}$$

which is (6.58) as we wanted. □

6.4.2 Proof of second part of Proposition 6.2.3

Proof of second part of Proposition 6.2.3. The result follows from Theorem 6.2.10 by observing that $d_{\bar{k},k}(G)(0)$ corresponds to the number of edges of the $\bar{k}$-densest subgraph of a vanilla $G(n-k,\frac{1}{2})$ random graph. □

6.5 Proofs for First Moment Curve Monotonicity results

6.5.1 Key lemmas

Lemma 6.5.1. *Suppose $1 \leq k \leq \bar{k} \leq n$ with $n \to +\infty$, $\bar{k} = o(n)$ and $\epsilon \in (0,1)$ arbitrarily small constant. For $z \in [0,(1-\epsilon)k] \cap \mathbb{Z}$ let*

$$A(z) := \log\left(\binom{k}{z}\binom{n-k}{\bar{k}-z}\right). \tag{6.62}$$

Then for any $z \in [0,(1-\epsilon)k] \cap \mathbb{Z}$,

$$A(z) = \Theta\left(\bar{k}\log\left(\frac{n}{\bar{k}}\right)\right). \tag{6.63}$$

and

$$A(z+1) - A(z) = \log\left(\frac{k\bar{k}}{(z+1)n}\right) - O(1). \tag{6.64}$$

Proof. First

$$A(z) = \log\left(\binom{k}{z}\right) + \log\left(\binom{n-k}{\bar{k}-z}\right).$$

Since, $\binom{k}{z} \leq 2^k$ we have $\log\left(\binom{k}{z}\right) = O(k)$. Hence,

$$A(z) = \log\left(\binom{n-k}{\bar{k}-z}\right) + O(k). \tag{6.65}$$

For any $z \in [0,(1-\epsilon)k]$ since $k \leq \bar{k}$ we have

$$\epsilon\bar{k} \leq \bar{k} - z \leq \bar{k}.$$

Hence, since for large n we have $\bar{k} < \frac{n}{2}$, by standard monotonicity arguments on the binomial

coefficients we have

$$\log\left(\binom{n-k}{\epsilon\bar{k}}\right) \leq \log\left(\binom{n-k}{\bar{k}-z}\right) \leq \log\left(\binom{n-k}{\bar{k}}\right)$$

which using Stirling's approximation since $k \leq \bar{k} = o(n)$ and ϵ is a positive constant yields

$$\log\left(\binom{n-k}{\bar{k}-z}\right) = \Theta\left(\bar{k}\log\left(\frac{n}{\bar{k}}\right)\right).$$

Combining this with (6.66) and $k \leq \bar{k} = o(n)$ we conclude (6.64).

For the final part, simple algebra and that $\frac{k-z}{k} = \Omega(1)$, $\frac{\bar{k}-z}{\bar{k}} = \Omega(1)$ for the z of interest yields,

$$\begin{aligned}
A(z+1) - A(z) &= \log\left(\frac{(k-z)\left(\bar{k}-z\right)}{(z+1)\left(n-k-\bar{k}+z+1\right)}\right) \\
&= \log\left(\frac{k\bar{k}}{(z+1)n}\right) + \log\left(\frac{k-z}{k}\right) + \log\left(\frac{\bar{k}-z}{\bar{k}}\right) + \log\left(\frac{n}{n-k-\bar{k}+z+1}\right) \\
&= \log\left(\frac{k\bar{k}}{(z+1)n}\right) - O(1) - O\left(\frac{\bar{k}}{n}\right) \\
&= \log\left(\frac{k\bar{k}}{(z+1)n}\right) - O(1),
\end{aligned}$$

which is (6.65).

□

Lemma 6.5.2. *Suppose $k \leq \bar{k} \leq n$ with $(\log n)^5 \leq \bar{k}$. Then for any $z \in \mathbb{Z}_{>0}$ for which it holds $\lfloor \frac{\bar{k}k}{n} \rfloor \leq z \leq k$ we have,*

$$|\Gamma_{\bar{k},k}(z) - \Phi_{\bar{k}}(z)| = O(1), \tag{6.66}$$

for

$$\Phi_{\bar{k}}(z) := \frac{1}{2}\left(\binom{\bar{k}}{2} + \binom{z}{2}\right) + \frac{1}{\sqrt{2}}\sqrt{A(z)\left(\binom{\bar{k}}{2} - \binom{z}{2}\right)} - \frac{1}{6\sqrt{2}}\sqrt{\frac{A(z)^3}{\binom{\bar{k}}{2} - \binom{z}{2}}}. \tag{6.67}$$

and $A(z)$ is defined in (6.63).

Proof. Let

$$\epsilon := \frac{\log\left(\binom{k}{z}\binom{n-k}{\bar{k}-z}\right)}{\binom{\bar{k}}{2}-\binom{z}{2}}.$$

Combining the elementary inequalities

$$\binom{k}{z} = \binom{k}{k-z} \leq k^{k-z} \leq n^{\bar{k}-z}$$

and

$$\binom{n-k}{\bar{k}-z} \leq n^{\bar{k}-z}$$

with

$$\bar{k} \geq (\log n)^5 = \omega(\log n)$$

we conclude

$$\epsilon = O\left(\frac{\left(\bar{k}-z\right)\log n}{\left(\bar{k}-z\right)\left(\bar{k}+z-1\right)}\right) = O\left(\frac{\log n}{\bar{k}}\right) = o(1).$$

For our choice of ϵ, $\Gamma_{\bar{k},k}$ can be simply expressed as

$$\Gamma_{\bar{k},k}(z) = \binom{z}{2} + h^{-1}(\log 2 - \epsilon)\left(\binom{\bar{k}}{2} - \binom{z}{2}\right).$$

Since $\epsilon = o(1)$, Lemma 6.8.3 implies

$$|\Gamma_{\bar{k},k}(z) - \Phi_{\bar{k}}(z)| = O\left(\sqrt{\frac{A(z)^5}{\left(\binom{\bar{k}}{2}-\binom{z}{2}\right)^3}}\right), \tag{6.68}$$

where $\Phi_{\bar{k}}(z)$ and $A(z)$ are defined in (6.68) and (6.63) respectively.

Using $\binom{\bar{k}}{z}\binom{n-k}{\bar{k}-z} \leq n^{2\left(\bar{k}-z\right)}$ we have

$$A(z) \leq 2\left(\bar{k}-z\right)\log n.$$

Furthermore $\binom{\bar{k}}{2} - \binom{z}{2} \geq \frac{\bar{k}\left(\bar{k}-z\right)}{2}$. Hence, combining the last two inequalities, (6.69) can be

simplified to

$$|\Gamma_{\bar{k},k}(z) - \Phi_{\bar{k}}(z)| = O\left(\sqrt{\frac{(\log n)^5}{\bar{k}}}\right) = O(1), \tag{6.69}$$

where the last step is due to $(\log n)^5 \leq \bar{k}$. This concludes the proof of the Lemma.

□

Lemma 6.5.3. *Suppose $k \leq \bar{k} \leq n$ with $(\log n)^5 \leq \bar{k}$ and $\epsilon > 0$. Then for some sufficiently large constant $C_0 = C_0(\epsilon) > 0$, if $\lfloor C_0 \frac{\bar{k}k}{n} \rfloor \leq z \leq (1-\epsilon)k$,*

$$\Gamma_{\bar{k},k}(z+1) - \Gamma_{\bar{k},k}(z) = z\left(\frac{1}{2} - o(1)\right) - \Theta\left[\sqrt{\frac{\bar{k}}{\log(\frac{n}{\bar{k}})}} \log\left(\frac{(z+1)n}{k\bar{k}}\right)\right] + O(1). \tag{6.70}$$

Proof. First we choose $C_0 > 0$ large enough so that so that $\log\left(\frac{(z+1)n}{k\bar{k}}\right)$ dominates the constant additional factor in the right hand side of (6.65) and therefore for all z of interest

$$A(z+1) - A(z) = \Theta\left(\log\left(\frac{k\bar{k}}{(z+1)n}\right)\right) = -\Theta\left(\log\left(\frac{(z+1)n}{k\bar{k}}\right)\right). \tag{6.71}$$

In light of Lemma 6.5.2 we can prove (6.71) with $\Phi_{\bar{k}}(z+1) - \Phi_{\bar{k}}(z)$ (defined in (6.68)) instead of $\Gamma_{\bar{k},k}(z+1) - \Gamma_{\bar{k},k}(z)$ at the expense only of $O(1)$ terms on the right hand sides. We write the difference $\Phi_{\bar{k}}(z+1) - \Phi_{\bar{k}}(z)$ as a summation of three parts.

$$\begin{aligned}\Phi_{\bar{k}}(z+1) - \Phi_{\bar{k}}(z) = &\underbrace{\frac{1}{2}\left(\binom{\bar{k}}{2} + \binom{z+1}{2}\right) - \frac{1}{2}\left(\binom{\bar{k}}{2} + \binom{z}{2}\right)}_{\text{First Part}} \\ &+ \frac{1}{\sqrt{2}}\left(\underbrace{\sqrt{A(z+1)\left(\binom{\bar{k}}{2} - \binom{z+1}{2}\right)} - \sqrt{A(z)\left(\binom{\bar{k}}{2} - \binom{z}{2}\right)}}_{\text{Second Part}}\right) \\ &- \frac{1}{6\sqrt{2}}\left(\underbrace{\sqrt{\frac{A(z+1)^3}{\left(\binom{\bar{k}}{2} - \binom{z+1}{2}\right)}} - \sqrt{\frac{A(z)^3}{\left(\binom{\bar{k}}{2} - \binom{z}{2}\right)}}}_{\text{Third Part}}\right).\end{aligned}$$

The first part can be straightforwardly simplified to $\frac{z}{2}$.

We write the second part as follows,

$$\begin{aligned}
&\sqrt{A(z+1)\left(\binom{\bar{k}}{2}-\binom{z+1}{2}\right)}-\sqrt{A(z)\left(\binom{\bar{k}}{2}-\binom{z}{2}\right)}\\
=&\left(\sqrt{A(z+1)}-\sqrt{A(z)}\right)\sqrt{\binom{\bar{k}}{2}-\binom{z+1}{2}}+\sqrt{A(z)}\left(\sqrt{\binom{\bar{k}}{2}-\binom{z+1}{2}}-\sqrt{\binom{\bar{k}}{2}-\binom{z}{2}}\right)\\
=&\left(\frac{A(z+1)-A(z)}{\sqrt{A(z+1)}+\sqrt{A(z)}}\right)\sqrt{\binom{\bar{k}}{2}-\binom{z+1}{2}}-\sqrt{A(z)}\frac{\binom{z+1}{2}-\binom{z}{2}}{\sqrt{\binom{\bar{k}}{2}-\binom{z+1}{2}}+\sqrt{\binom{\bar{k}}{2}-\binom{z}{2}}}\\
=&\left(\frac{A(z+1)-A(z)}{\sqrt{A(z+1)}+\sqrt{A(z)}}\right)\sqrt{\binom{\bar{k}}{2}-\binom{z+1}{2}}-\sqrt{A(z)}\frac{z}{\sqrt{\binom{\bar{k}}{2}-\binom{z+1}{2}}+\sqrt{\binom{\bar{k}}{2}-\binom{z}{2}}}.
\end{aligned} \tag{6.72}$$

Since $z\leq(1-\epsilon)k\leq(1-\epsilon)\bar{k}$ applying (6.64) from Lemma 6.5.1, the last quantity is of the order

$$\begin{aligned}
&\Theta\left[\left(\frac{A(z+1)-A(z)}{\sqrt{\bar{k}\log(\frac{n}{\bar{k}})}}\right)\bar{k}\right]-\Theta\left[\sqrt{\bar{k}\log(\frac{n}{\bar{k}})}\frac{z}{\bar{k}}\right],\text{using }\binom{\bar{k}}{2}-\binom{z}{2}=\Theta\left(\left(\bar{k}\right)^2\right)\\
=&\Theta\left[\left(\frac{\left(A(z+1)-A(z)\right)\sqrt{\bar{k}}}{\sqrt{\log(\frac{n}{\bar{k}})}}\right)\right]-\Theta\left[\frac{\sqrt{\log(\frac{n}{\bar{k}})}z}{\sqrt{\bar{k}}}\right]\\
=&-\Theta\left[\sqrt{\frac{\bar{k}}{\log(\frac{n}{\bar{k}})}}\log\left(\frac{(z+1)\,n}{k\bar{k}}\right)\right]-o\left(z\right).
\end{aligned} \tag{6.73}$$

where for the last equality we used (6.72) and $\bar{k}=\omega\left(\log n\right)$.

For the third part we write,

$$\begin{aligned}
&\sqrt{\frac{A(z+1)^3}{\left(\binom{\bar{k}}{2}-\binom{z+1}{2}\right)}}-\sqrt{\frac{A(z)^3}{\binom{\bar{k}}{2}-\binom{z}{2}}}\\
=&\frac{A(z+1)^{\frac{3}{2}}-A(z)^{\frac{3}{2}}}{\sqrt{\binom{\bar{k}}{2}-\binom{z+1}{2}}}+A(z)^{\frac{3}{2}}\left(\frac{1}{\sqrt{\binom{\bar{k}}{2}-\binom{z+1}{2}}}-\frac{1}{\sqrt{\binom{\bar{k}}{2}-\binom{z}{2}}}\right)
\end{aligned}$$

Using $a^{\frac{3}{2}}-b^{\frac{3}{2}}=\left(a^3-b^3\right)/\left(a^{\frac{3}{2}}+b^{\frac{3}{2}}\right)$ and $a^3-b^3=(a-b)\left(a^2+b^2+ab\right)=O\left((a-b)\left(a^2+b^2\right)\right)$

for $a, b \in \mathbb{R}$, we have that the quantity is at most

$$O\left[\frac{\left(A(z+1)-A(z)\right)\left(A(z+1)^2+A(z)^2\right)}{\left(A(z+1)^{\frac{3}{2}}+A(z)^{\frac{3}{2}}\right)\sqrt{\binom{\bar{k}}{2}-\binom{z+1}{2}}}\right]+O\left[A(z)^{\frac{3}{2}}\left(\frac{\sqrt{\binom{\bar{k}}{2}-\binom{z}{2}}-\sqrt{\binom{\bar{k}}{2}-\binom{z+1}{2}}}{\sqrt{\left(\binom{\bar{k}}{2}-\binom{z+1}{2}\right)\left(\binom{\bar{k}}{2}-\binom{z}{2}\right)}}\right)\right]$$

which by Lemma 6.5.1 and (6.72) is at most

$$\begin{aligned}
&O\left[\frac{\log\left(\frac{(z+1)n}{k\bar{k}}\right)\sqrt{\bar{k}\log\left(\frac{n}{\bar{k}}\right)}}{\sqrt{\binom{\bar{k}}{2}-\binom{z+1}{2}}}\right]+O\left[\left(\bar{k}\log\left(\frac{n}{\bar{k}}\right)\right)^{\frac{3}{2}}\left(\frac{\sqrt{\binom{\bar{k}}{2}-\binom{z}{2}}-\sqrt{\binom{\bar{k}}{2}-\binom{z+1}{2}}}{\sqrt{\left(\binom{\bar{k}}{2}-\binom{z+1}{2}\right)\left(\binom{\bar{k}}{2}-\binom{z}{2}\right)}}\right)\right]\\
=&O\left[\frac{\left(\log\left(\frac{n}{\bar{k}}\right)\right)^{\frac{3}{2}}}{\sqrt{\bar{k}}}\right]+O\left[\frac{\left(\binom{z+1}{2}-\binom{z}{2}\right)\left(\log\left(\frac{n}{\bar{k}}\right)\right)^{\frac{3}{2}}}{\left(\bar{k}\right)^{\frac{3}{2}}}\right],
\end{aligned}$$

where for the last equality we used the elementary $\sqrt{a}-\sqrt{b}=(a-b)/\left(\sqrt{a}+\sqrt{b}\right)$ and that $z \leq (1-\epsilon)\bar{k}$. Finally, the last displayed quantity is at most

$$\begin{aligned}
&O\left[\frac{\left(\log\left(\frac{n}{\bar{k}}\right)\right)^{\frac{3}{2}}}{\sqrt{\bar{k}}}\right]+O\left[\frac{z\left(\log\left(\frac{n}{\bar{k}}\right)\right)^{\frac{3}{2}}}{\left(\bar{k}\right)^{\frac{3}{2}}}\right]\\
=&O\left[\frac{\left(\log\left(\frac{n}{\bar{k}}\right)\right)^{\frac{3}{2}}}{\sqrt{\bar{k}}}\right], \text{using } z \leq \bar{k}\\
=&o\left(1\right),
\end{aligned}$$

since $\bar{k}=\omega\left(\log^3 n\right)$ by our assumptions.

Combining the three parts gives

$$\Phi_{\bar{k}}(z+1)-\Phi_{\bar{k}}(z)=z\left(\frac{1}{2}-o(1)\right)-\Theta\left[\sqrt{\frac{\bar{k}}{\log(\frac{n}{\bar{k}})}}\log\left(\frac{(z+1)\,n}{k\bar{k}}\right)\right]+o\left(1\right). \tag{6.74}$$

which based on Lemma 6.5.2 implies (6.71).

The proof of the Lemma is complete. □

Lemma 6.5.4. *Suppose $k \leq \bar{k} \leq n$ with $(\log n)^5 \leq \bar{k}$ and $\epsilon > 0$. Let*

$$T_n := \sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}} \log\left(\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}} \left(\frac{\bar{k}k}{n}\right)^{-1}\right). \tag{6.75}$$

For some sufficiently large constant $C_0 = C_0(\epsilon) > 0$ and sufficiently large enough values of n the following monotonicity properties hold in the discretized interval

$$\mathcal{I} = \mathcal{I}_{C_0} = [\lfloor C_0 \frac{\bar{k}k}{n} \rfloor, (1-\epsilon)k] \cap \mathbb{Z}.$$

(1) If $T_n = o\left(\frac{\bar{k}k}{n}\right)$ then $\Gamma_{\bar{k},k}$ is monotonically increasing on $\mathcal{I}$.

(2) If $T_n = \omega(k)$ then $\Gamma_{\bar{k},k}$ is monotonically decreasing on $\mathcal{I}$.

(3) If $\omega\left(\frac{\bar{k}k}{n}\right) = T_n = o(k)$ then $\Gamma_{\bar{k},k}$ is non-monotonous on $\mathcal{I}$ with the property that for some constants $0 < D_1 < D_2$, $u_1 := D_1 \lceil \sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}} \rceil$ and $u_2 := D_2 \lceil \sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}} \rceil$ the following are true.

(a) $\lfloor C_0 \frac{\bar{k}k}{n} \rfloor < u_1 < u_2 < (1-\epsilon)k$ and

(b)

$$\max_{z \in \mathcal{I} \cap [u_1, u_2]} \Gamma_{\bar{k},k}(z) + \Omega\left(\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}\right) \leq \Gamma_{\bar{k},k}(\lfloor C \frac{\bar{k}k}{n} \rfloor) \leq \Gamma_{\bar{k},k}\left((1-\epsilon)k\right). \tag{6.76}$$

Proof. We start with the case $T_n = o\left(\frac{\bar{k}k}{n}\right)$ which can be equivalently written as

$$\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}} \left(\frac{\bar{k}k}{n}\right)^{-1} \log\left(\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}} \left(\frac{\bar{k}k}{n}\right)^{-1}\right) = o(1)$$

or using part (a) of Lemma 6.8.4,

$$\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}} \left(\frac{\bar{k}k}{n}\right)^{-1} = o(1). \tag{6.77}$$

Using (6.71) from Lemma 6.5.3 we have that for some universal constant $C_1 > 0$ and large

enough n,

$$\Gamma_{\bar{k},k}(z+1)-\Gamma_{\bar{k},k}(z) \geq \frac{z}{4}-C_1\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\log\left(\frac{n\left(z+1\right)}{\bar{k}k}\right)-O\left(1\right)$$
$$=\frac{k\bar{k}}{4n}\log\left(\frac{n\left(z+1\right)}{\bar{k}k}\right)\left(\frac{\frac{nz}{\bar{k}k}}{\log\left(\frac{n(z+1)}{\bar{k}k}\right)}-4C_1\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\left(\frac{\bar{k}k}{n}\right)^{-1}\right)-O\left(1\right).$$

The second term in the parenthesis in the last displayed quantity is $o\left(1\right)$ from (6.78). Now notice that since $\bar{k}=\omega\left(\log n\right)$, from (6.78) we have

$$\frac{\bar{k}k}{n}=\omega\left(1\right). \tag{6.78}$$

Therefore choosing $C_0>0$ large enough we have that $z\geq\lfloor C_0\frac{\bar{k}k}{n}\rfloor$ implies that the first term in the parenthesis can be made to be at least 1. Finally, the multiplicative term outside the parenthesis satisfies

$$\frac{k\bar{k}}{4n}\log\left(\frac{n\left(z+1\right)}{\bar{k}k}\right)\geq\frac{k\bar{k}}{4n}\log e^4=\frac{k\bar{k}}{n}$$

by choosing $z+1\geq\lfloor\frac{C_0k\bar{k}}{n}\rfloor$ for say $C_0>e^4$. Hence, indeed for some sufficiently large $C_0>0$ if $z\in\mathcal{I}_{C_0}$,

$$\Gamma_{\bar{k},k}(z+1)-\Gamma_{\bar{k},k}(z)\geq\frac{k\bar{k}}{n}\left(1-o\left(1\right)\right)-O\left(1\right)$$

which according to (6.79) implies that for some sufficiently large $C_0>0$ for n large enough if $z\in\mathcal{I}_{C_0}$,

$$\Gamma_{\bar{k},k}(z+1)\geq\Gamma_{\bar{k},k}(z),$$

that is the curve is increasing.

We now turn to Part (2) where $T_n=\omega\left(k\right)$ which can be equivalently written as

$$\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\left(\frac{\bar{k}k}{n}\right)^{-1}\log\left(\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\left(\frac{\bar{k}k}{n}\right)^{-1}\right)=\omega\left(\frac{n}{\bar{k}}\right)$$

or using that $\bar{k} = o(n)$ and part (c) of Lemma 6.8.4,

$$\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\left(\frac{\bar{k}k}{n}\right)^{-1} = \omega\left(\frac{\frac{n}{\bar{k}}}{\log\left(\frac{n}{\bar{k}}\right)}\right). \tag{6.79}$$

which simplifies to

$$\sqrt{\bar{k}\log\left(\frac{n}{\bar{k}}\right)} = \omega(k). \tag{6.80}$$

Now using (6.71) from Lemma 6.5.3 we have that for some universal constants $U_1 > 0$ and large enough n,

$$\begin{aligned}\Gamma_{\bar{k},k}(z+1) - \Gamma_{\bar{k},k}(z) &\leq \frac{3(z+1)}{4} - U_1\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\log\left(\frac{n(z+1)}{\bar{k}k}\right) + O(1) \\ &= \frac{3k\bar{k}}{4n}\log\left(\frac{n(z+1)}{\bar{k}k}\right)\left(\frac{\frac{n(z+1)}{\bar{k}k}}{\log\left(\frac{n(z+1)}{\bar{k}k}\right)} - \frac{4}{3}U_1\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\left(\frac{\bar{k}k}{n}\right)^{-1}\right) + O(1).\end{aligned} \tag{6.81}$$

Recall that for $x > e$, $x/\log x$ is increasing from elementary reasoning. Therefore if $C_0 > e$ using

$$\frac{n(z+1)}{\bar{k}k} \geq e \tag{6.82}$$

and the trivial $\frac{n(z+1)}{\bar{k}k} \leq \frac{n}{\bar{k}}$, we have

$$\frac{\frac{n(z+1)}{\bar{k}k}}{\log\left(\frac{n(z+1)}{\bar{k}k}\right)} \leq \frac{\frac{n}{\bar{k}}}{\log\left(\frac{n}{\bar{k}}\right)}.$$

Hence, by (6.80) we conclude

$$\frac{\frac{n(z+1)}{\bar{k}k}}{\log\left(\frac{n(z+1)}{\bar{k}k}\right)} = o\left(\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\left(\frac{\bar{k}k}{n}\right)^{-1}\right).$$

Therefore indeed the term inside the parenthesis in (6.82) is at most $-U_1\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\left(\frac{\bar{k}k}{n}\right)^{-1}$ for

large enough n, which allows to conclude that for large enough n (6.82) implies for all $z \in \mathcal{I}_C$,

$$\begin{aligned}\Gamma_{\bar{k},k}(z+1) - \Gamma_{\bar{k},k}(z) &\leq -\frac{3}{4}U_1\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\log\left(\frac{n\,(z+1)}{\bar{k}k}\right) + O\,(1) \\ &\leq -\frac{3}{4}U_1\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}} + O\,(1) \end{aligned} \tag{6.83}$$

where we have used $\log\left(\frac{n(z+1)}{\bar{k}k}\right) \geq 1$ according to (6.83). Using now that $\bar{k} = \omega\,(\log n)$ we conclude based on (6.84), that indeed for some sufficiently large $C_0 > 0$ and large enough n, if $z \in \mathcal{I}_{C_0}$, $\Gamma_{\bar{k},k}(z+1) \leq \Gamma_{\bar{k},k}(z)$, that is the curve is decreasing.

We finally turn to Part (3) where $\omega\left(\frac{\bar{k}k}{n}\right) = T_n = o\,(k)$. By similar arguments as for (6.78) and (6.81) we conclude that in this case

$$\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}} = \omega\left(\frac{\bar{k}k}{n}\right) \tag{6.84}$$

and

$$\sqrt{\bar{k}\log\left(\frac{n}{\bar{k}}\right)} = o\,(k)\,. \tag{6.85}$$

Notice that because of (6.85) and (6.86) we have that for any choice of $D_1, D_2 > 0$ and sufficiently large n,

$$\lfloor C_0\frac{\bar{k}k}{n}\rfloor < u_1 = D_1\lceil\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\rceil < u_2 = D_2\lceil\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\rceil < (1-\epsilon)k.$$

It suffices now to establish (6.77) as non-monotonicity is directly implied by it.

By definition of $\Gamma_{\bar{k},k}$ to establish for large n

$$\Gamma_{\bar{k},k}\left((1-\epsilon)k\right) \geq \Gamma_{\bar{k},k}(\lfloor C_0\frac{\bar{k}k}{n}\rfloor) \tag{6.86}$$

it suffices to establish that for large n,

$$\binom{k}{2} + h^{-1}\left(\log 2 - \frac{\log\left(\binom{k}{(1-\epsilon)k}\binom{n-k}{\bar{k}-(1-\epsilon)k}\right)}{\binom{\bar{k}}{2} - \binom{(1-\epsilon)k}{2}}\right)\left(\binom{\bar{k}}{2} - \binom{(1-\epsilon)k}{2}\right)$$

is bigger than

$$\binom{\lfloor C_0 \frac{\bar{k}k}{n} \rfloor}{2} + h^{-1}\left(\log 2 - \frac{\log \binom{k}{k-\lfloor C_0 \frac{\bar{k}k}{n} \rfloor}\binom{n-k}{\bar{k}}}{\binom{\bar{k}}{2} - \binom{\lfloor C \frac{\bar{k}k}{n} \rfloor}{2}}\right)\left(\binom{\bar{k}}{2} - \binom{\lfloor C \frac{\bar{k}k}{n} \rfloor}{2}\right).$$

Since $\bar{k} = \omega(\log n)$ both the arguments of h^{-1} in the displayed equations are $\log 2 - o(1)$. Hence by Lemma 6.8.3 it suffices for large n to prove

$$\frac{1}{2}\left(\binom{\bar{k}}{2} + \binom{(1-\epsilon)k}{2}\right) + \Theta\left(\sqrt{\log\left(\binom{k}{(1-\epsilon)\,k}\binom{n-k}{\bar{k}-(1-\epsilon)k}\right)\left(\binom{\bar{k}}{2} - \binom{(1-\epsilon)k}{2}\right)}\right)$$

is bigger than

$$\frac{1}{2}\left(\binom{\bar{k}}{2} + \binom{\lfloor C_0 \frac{\bar{k}k}{n} \rfloor}{2}\right) + \Theta\left(\sqrt{\log\left[\binom{k}{\lfloor C_0 \frac{\bar{k}k}{n} \rfloor}\binom{n-k}{\bar{k}-\lfloor C_0 \frac{\bar{k}k}{n} \rfloor}\right]\left(\binom{\bar{k}}{2} - \binom{\lfloor C_0 \frac{\bar{k}k}{n} \rfloor}{2}\right)}\right).$$

Since by (6.85) and (6.86) we have $k = \omega\left(\frac{\bar{k}k}{n}\right)$ it suffices that k^2 is

$$\omega\left(\sqrt{\log\left[\binom{k}{\lfloor C_0 \frac{\bar{k}k}{n} \rfloor}\binom{n-k}{\bar{k}-\lfloor C_0 \frac{\bar{k}k}{n} \rfloor}\right]\binom{\bar{k}}{2}} - \sqrt{\log\left[\binom{k}{(1-\epsilon)\,k}\binom{n-k}{\bar{k}-(1-\epsilon)\,k}\right]\left(\binom{\bar{k}}{2} - \binom{k}{2}\right)}\right).$$

Using that $(1-\epsilon)\,k \leq k$ and that for large n, $\bar{k} \leq \frac{n}{2}$ by standard monotonicity arguments we have

$$\binom{k}{(1-\epsilon)\,k}\binom{n-k}{\bar{k}-(1-\epsilon)\,k} \geq \binom{n-k}{\bar{k}-k}.$$

Hence it suffices to show

$$k^2 = \omega\left(\sqrt{\log\left[\binom{k}{\lfloor C_0 \frac{\bar{k}k}{n} \rfloor}\binom{n-k}{\bar{k}-\lfloor C_0 \frac{\bar{k}k}{n} \rfloor}\right]\binom{\bar{k}}{2}} - \sqrt{\log\left[\binom{n-k}{\bar{k}-k}\right]\left(\binom{\bar{k}}{2} - \binom{k}{2}\right)}\right).$$

Now since for large n, $\bar{k} < \frac{n-k}{2}$, using the elementary

$$\binom{k}{\lfloor C_0 \frac{\bar{k}k}{n} \rfloor}\binom{n-k}{\bar{k}-\lfloor C_0 \frac{\bar{k}k}{n} \rfloor} \leq 2^k\binom{n-k}{\bar{k}} \leq e^{O\left(\bar{k}\log\left(\frac{n-k}{\bar{k}}\right)\right)}$$

and

$$\binom{n-k}{\bar{k}-k} \geq \left(\frac{\left(n-\bar{k}\right)}{\bar{k}}\right)^{\bar{k}-k}$$

we conclude that it suffices to have

$$k^2 = \omega\left(\left(\left(\bar{k}\right)^{\frac{3}{2}} - \sqrt{\bar{k}}\left(\bar{k}-k\right)\right)\log\left(\frac{n-k}{\bar{k}}\right)\right) = \omega\left(\sqrt{\bar{k}}k\log\left(\frac{n-k}{\bar{k}}\right)\right),$$

which follows directly from (6.86). This establishes (6.87) for large enough n.

Now using (6.71) from Lemma 6.5.3 to conclude that for some universal constants $U_1 > 0$, large enough n and such z,

$$\Gamma_{\bar{k},k}(z+1) - \Gamma_{\bar{k},k}(z) \leq z - U_1\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\log\left(\frac{n\left(z+1\right)}{\bar{k}k}\right) + O\left(1\right). \tag{6.87}$$

Using $z+1 \geq \lfloor C_0\frac{\bar{k}k}{n}\rfloor$ for $C_0 > e$ and focusing only on $\lfloor C_0\frac{\bar{k}k}{n}\rfloor \leq z \leq \frac{U_1}{2}\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}$ (existence of such z follows by (6.85)) we conclude for any such z and large enough n,

$$\Gamma_{\bar{k},k}(z+1) - \Gamma_{\bar{k},k}(z) \leq z - U_1\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}} + O\left(1\right) \leq -\frac{U_1}{2}\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}} \tag{6.88}$$

where we used the fact that $\bar{k} = \omega\left(\log n\right)$. Now set

$$D_1 := \frac{U_1}{4}, D_2 := \frac{U_1}{2}.$$

Fix any $Z \in \mathcal{I}$ with $D_1\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}} \leq Z \leq D_2\lceil\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\rceil$. Focus on $z \in \mathcal{I}$ with $\lfloor C_0\frac{\bar{k}k}{n}\rfloor \leq z \leq Z-1$. (6.85) yields that the the number of such z's for large n is at least $\frac{D_1}{2}\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}$. By telescopic summation of (6.89) over these z we have

$$\Gamma_{\bar{k},k}\left(Z\right) + D_1^2\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)} \leq \Gamma_{\bar{k},k}(\lfloor C_0\frac{\bar{k}k}{n}\rfloor). \tag{6.89}$$

Since Z was arbitrary we conclude that,

$$\max_{z\in\mathcal{I}\cap\left[D_1\lceil\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\rceil,D_2\lceil\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\rceil\right]}\Gamma_{\bar{k},k}(z)+\Omega\left(\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}\right)\leq\Gamma_{\bar{k},k}(\lfloor C_0\frac{\bar{k}k}{n}\rfloor). \tag{6.90}$$

Equations (6.91) and (6.87) imply (6.77). The proof of the Lemma is complete.

□

6.5.2 Proof of Theorem 6.2.5

Proof of Theorem 6.2.5. We start with the case where $k=o\left(\sqrt{n}\right)$. Notice that $k=o\left(\sqrt{n}\right)$ together with $\bar{k}=o\left(n\right)$ trivially imply

$$\frac{\frac{n}{\bar{k}}}{\log\frac{n}{\bar{k}}}=\omega\left(\frac{k^2}{n}\right)$$

which can be written equivalently as

$$\frac{n}{k^2}=\omega\left(\frac{\bar{k}}{n}\log\left(\frac{n}{\bar{k}}\right)\right)$$

or

$$\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}=\omega\left(\frac{\bar{k}k}{n}\right)$$

or

$$\left(\frac{\bar{k}k}{n}\right)^{-1}\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}=\omega\left(1\right).$$

Using part (b) of Lemma 6.8.4 we have

$$\left(\frac{\bar{k}k}{n}\right)^{-1}\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\log\left(\left(\frac{\bar{k}k}{n}\right)^{-1}\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\right)=\omega\left(1\right)$$

or

$$T_n=\omega\left(\frac{\bar{k}k}{n}\right), \tag{6.91}$$

where T_n is defined in equation (6.76).

First, we consider the subcase where $\bar{k} = o\left(\frac{k^2}{\log\left(\frac{n}{k^2}\right)}\right)$. In that case, we have

$$\frac{n}{\bar{k}} = \omega\left(\frac{n}{k^2}\log\left(\frac{n}{k^2}\right)\right)$$

which since $k^2 = o(n)$ which according to part (d) of Lemma 6.8.4 implies

$$\frac{\frac{n}{\bar{k}}}{\log\left(\frac{n}{\bar{k}}\right)} = \omega\left(\frac{n}{k^2}\right) \tag{6.92}$$

which is equivalent with

$$k = \omega\left(\sqrt{\bar{k}\log\left(\frac{n}{\bar{k}}\right)}\right)$$

or

$$\frac{\frac{n}{\bar{k}}}{\log\left(\frac{n}{\bar{k}}\right)} = \omega\left(\left(\frac{\bar{k}k}{n}\right)^{-1}\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\right)$$

or

$$\frac{n}{\bar{k}} = \omega\left(\left(\frac{\bar{k}k}{n}\right)^{-1}\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\log\left(\left(\frac{\bar{k}k}{n}\right)^{-1}\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\right)\right).$$

The last equality can be rewritten

$$\frac{n}{\bar{k}} = \omega\left(\left(\frac{\bar{k}k}{n}\right)^{-1}T_n\right),$$

where T_n is defined in equation (6.76), which now simplifies to

$$T_n = o(k). \tag{6.93}$$

Combining (6.92) with (6.94), according to Part (3) of Lemma 6.5.4 we conclude the desired non-monotonicity result in that subcase.

Second, we consider the subcase $\bar{k} = \omega\left(\frac{k^2}{\log\left(\frac{n}{k^2}\right)}\right)$. In tha case, following similar to the derivation of (6.94) by simply the order of comparison (in more detail, reversing the o-notation with the ω-notation and applying the other direction of part (d) of Lemma 6.8.4), we conclude that in this case $T_n = \omega(k)$. In particular, according to Part (2) of Lemma 6.5.4 we can conclude that the curve is decreasing in that subcase.

We turn now to the case $k = \omega(\sqrt{n})$. In that case, together with $\bar{k} = o(n)$ we have

$$\frac{\frac{n}{\bar{k}}}{\log\left(\frac{n}{k}\right)} = \omega\left(\frac{n}{k^2}\right)$$

which is exactly (6.93). Following the identical derivation following (6.93) we conclude that (6.94) holds in this case.

First, we consider the subcase $\bar{k} = o\left(\frac{\frac{n^2}{k^2}}{\log\left(\frac{n}{k^2}\right)}\right)$ which can be written equivalently as

$$\frac{n}{\bar{k}} = \omega\left(\frac{k^2}{n}\log\left(\frac{k^2}{n}\right)\right).$$

Since $k^2 = \omega(n)$ according to part (d) of Lemma 6.8.4 we have

$$\frac{k^2}{n} = o\left(\frac{\frac{n}{\bar{k}}}{\log\left(\frac{n}{\bar{k}}\right)}\right)$$

or

$$\frac{\bar{k}k}{n} = o\left(\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\right)$$

or

$$\left(\frac{\bar{k}k}{n}\right)^{-1}\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}} = \omega(1)$$

which according to part (b) of Lemma 6.8.4 gives

$$\left(\frac{\bar{k}k}{n}\right)^{-1}\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\log\left(\left(\frac{\bar{k}k}{n}\right)^{-1}\sqrt{\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}}\right) = \omega(1).$$

The last equality simplifies to (6.92). In particular, in this regime both (6.92) and (6.94) are hence according to Part (3) of Lemma 6.5.4 we can conclude the desired non-monotonicity result in this subcase.

Second, we consider the subcase $\bar{k} = \omega\left(\frac{\frac{n^2}{k^2}}{\log\left(\frac{n}{k^2}\right)}\right)$. Following similar reasoning to the derivation of (6.92) under the assumption $\bar{k} = o\left(\frac{\frac{n^2}{k^2}}{\log\left(\frac{n}{k^2}\right)}\right)$ one can establish $T_n = o\left(\frac{\bar{k}k}{n}\right)$ from $\bar{k} = \omega\left(\frac{\frac{n^2}{k^2}}{\log\left(\frac{n}{k^2}\right)}\right)$ imply (in more detail, we need to reverse the o-notation with the ω-notation at all places and apply the other direction of part (b) of Lemma 6.8.4)). Hence, using Part (1) of

Lemma 6.5.4 allows us to conclude that the curve is increasing in this subcase.

This completes the proof of Theorem 6.2.5.

□

6.6 Proof of the Presence of the Overlap Gap Property

Proof of Theorem 6.2.9. First, we apply Theorem 6.2.5 for $\epsilon = \frac{1}{2}$ and we denote by $C_0 = C_0\left(\frac{1}{2}\right) > 0$ the constant implied by Theorem 6.2.5. Notice that since under our assumptions both $\bar{k}, k$ are $o(\sqrt{n})$, $\bar{k}k = o(n)$, and therefore for large n, $\lfloor C_0 \frac{\bar{k}k}{n} \rfloor = 0$. In particular, the interval containg the overlap sizes of interest simplifies to

$$\mathcal{I} = \left[0, \frac{k}{2}\right] \cap \mathbb{Z}.$$

Furthermore according to our parameter assumptions on $\bar{k}, k, n$ we are in the case (1i) of Theorem 6.2.5 where $\Gamma_{\bar{k},k}(z), z \in \mathcal{I}$ is non-monotonic and satisfies (6.10). Specifically, let D_1, D_2, u_1, u_2 as in Theorem 6.2.5 so that for large enough n,

$$\lfloor C_0 \frac{\bar{k}k}{n} \rfloor = 0 < u_1 < u_2 < \frac{k}{2} \tag{6.94}$$

and (6.10) holds.

We first establish (6.11) for D_1, D_2, u_1, u_2 as chosen above. By Proposition 6.2.3 we know that for all $z \in \mathcal{I}$, $d_{\bar{k},k}(G)(z) \leq \Gamma_{\bar{k},k}(z)$, w.h.p. as $n \to +\infty$. Combining with (6.10) we have that for some constant $c_0 > 0$:

$$\Gamma_{\bar{k},k}(0) \geq \max_{z \in \mathcal{I} \cap [u_1, u_2]} d_{\bar{k},k}(G)(z) + c_0 \frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}, \tag{6.95}$$

w.h.p. as $n \to +\infty$. Hence, to establish (6.11) from (6.96) it suffices to establish that

$$\min\{d_{\bar{k},k}(G)(0), d_{\bar{k},k}(G)(\frac{k}{2})\} \geq \Gamma_{\bar{k},k}(0) - o\left(\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}\right), \tag{6.96}$$

w.h.p. as $n \to +\infty$. Indeed, combining (6.97) with (6.96) implies

$$\min\{d_{\bar{k},k}(G)(0), d_{\bar{k},k}(G)(\frac{k}{2})\} \geq \max_{z \in \mathcal{I} \cap [u_1,u_2]} d_{\bar{k},k}(G)(z) + \frac{c_0}{2} \frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}, \tag{6.97}$$

w.h.p. as $n \to +\infty$ which implies (6.11).

We first prove

$$d_{\bar{k},k}(G)(0) \geq \Gamma_{\bar{k},k}(0) - o\left(\frac{\bar{k}}{\log\left(\frac{n}{\bar{k}}\right)}\right), \tag{6.98}$$

w.h.p. as $n \to +\infty$. Notice that the exponent $C = \log k / \log n$ satisfies

$$C < \frac{1}{2} - \frac{\sqrt{6}}{6} = 1 - \frac{1}{6 - 2\sqrt{6}}$$

and as it can straightforwardly checked it also satisfies $\frac{3}{2} - \left(\frac{5}{2} - \sqrt{6}\right) \frac{1-C}{C} < 1$. Therefore some $\beta(C) > 0$ satisfies

$$\frac{3}{2} - \left(\frac{5}{2} - \sqrt{6}\right) \frac{1-C}{C} < \beta(C) < 1.$$

Part (2) of Theorem 6.2.3 gives for this value of $\beta = \beta(C)$

$$d_{k,\bar{k}}(G)(0) \geq \Gamma_{\bar{k},k}(0) - O\left(\left(\bar{k}\right)^{\beta} \sqrt{\log n}\right) \tag{6.99}$$

w.h.p. as $n \to +\infty$. Since $\beta < 1$, we have $\left(\bar{k}\right)^{\beta} \sqrt{\log n} = o\left(\frac{\bar{k}}{\log(n/\bar{k})}\right)$. Hence, using (6.100), we can directly conclude (6.99).

We now proceed with proving

$$d_{\bar{k},k}(G)\left(\frac{k}{2}\right) \geq \Gamma_{\bar{k},k}(0), \tag{6.100}$$

w.h.p. as $n \to +\infty$. Note that (6.101) combined with (6.99) imply (6.97). First, denote by $G_0 := G \setminus \mathcal{PC}$ the graph obtained by deleting from G the vertices of $\mathcal{PC}$ and notice that G_0 is simply distributed as an Erdős-Rényi model $G_0 \sim G\left(n-k, \frac{1}{2}\right)$. Second, we fix an arbitrary $\frac{k}{2}$-vertex subgraph S_1 of $\mathcal{PC}$ and then optimize over the $N := \binom{n-k}{\bar{k}-\frac{k}{2}}$ different $(\bar{k} - \frac{k}{2})$-vertex

subgraphs S_2 of G_0 to get

$$d_{\bar{k},k}(G)\binom{k}{2} \geq \max_{S_2} |\mathrm{E}\left(S_1 \cup S_2\right)| = \binom{\frac{k}{2}}{2} + \max_{S_2}\{|\mathrm{E}(S_1,S_2)| + |\mathrm{E}(S_2)|\}, \tag{6.101}$$

where we used the fact that S_1 is a subset of the planted clique and by $\mathrm{E}(S_1,S_2)$ we denote to the set of edges with one endpoint in S_1 and one endpoint in S_2. We now index the subsets S_2 by $S^i, i = 1,2,\ldots,N$ and set $X_i = |\mathrm{E}(S_1,S^i)|$ and $Y_i = |\mathrm{E}(S^i)|$. It is straightforward to see because of the distribution of G_0 that

(1) for each $i \in [N]$, $X_i \sim \mathrm{Bin}\left(\binom{\bar{k}-k/2}{2}, \frac{1}{2}\right)$

(2) $Y_i, i \in [N]$ are i.i.d. $\mathrm{Bin}\left((\bar{k}-\frac{k}{2})\frac{k}{2}, \frac{1}{2}\right)$

(2) the sequence $X_i, i \in [N]$ is independent from the sequence $Y_j, j \in [N]$ and

(4) $\max_{i\in[N]}\{X_i\} = d_{\mathrm{ER},\bar{k}-\frac{k}{2}}(G_0)$, where $d_{\mathrm{ER},K}(\cdot)$ is defined for any $K \in [|V(G_0)|]$ in (6.12).

Hence, combining (6.102) and the above four observations with Lemma 6.8.1 we have

$$\begin{aligned} d_{\bar{k},k}(G)\binom{k}{2} &\geq \binom{\frac{k}{2}}{2} + \max_{i=1,2,\ldots,N}\{X_i\} + \max\{\frac{(\bar{k}-\frac{k}{2})\frac{k}{2}}{2} - \sqrt{(\bar{k}-\frac{k}{2})\frac{k}{2}\log\log N}, 0\} \\ &\geq \binom{\frac{k}{2}}{2} + d_{\mathrm{ER},\bar{k}-\frac{k}{2}}(G_0) + \max\{\frac{(\bar{k}-\frac{k}{2})\frac{k}{2}}{2} - \sqrt{\bar{k}k\log\log N}, 0\} \\ &= \binom{\frac{k}{2}}{2} + d_{\mathrm{ER},\bar{k}-\frac{k}{2}}(G_0) + \max\{\frac{(\bar{k}-\frac{k}{2})\frac{k}{2}}{2} - O\left(\sqrt{\bar{k}k\log n}\right), 0\}, \end{aligned} \tag{6.102}$$

where for the last equality we have used that $N = \binom{n-\bar{k}}{\bar{k}-k} \leq 2^{n-\bar{k}} \leq 2^n$ and therefore $\log\log N = O\left(\log n\right)$.

Since by our assumption $\bar{k} = \Theta\left(n^C\right)$ for $C < \frac{1}{2}$ and $k \leq \bar{k}$ we have $\frac{\bar{k}}{2} \leq \bar{k} - \frac{k}{2} \leq \bar{k}$ and therefore $\bar{k} - \frac{k}{2} = \Theta\left(n^C\right)$. Hence Theorem 6.2.10 can be applied, according to which for any $\beta > 0$ with $\frac{3}{2} - \left(\frac{5}{2} - \sqrt{6}\right)\frac{1-C}{C} < \beta < 1$ it holds,

$$d_{\mathrm{ER},\bar{k}-\frac{k}{2}}(G_0) \geq h^{-1}\left(\log 2 - \frac{\log\binom{n-k}{\bar{k}-\frac{k}{2}}}{\binom{\bar{k}-\frac{k}{2}}{2}}\right)\binom{\bar{k}-\frac{k}{2}}{2} - O\left(\left(\bar{k}\right)^{\beta}\sqrt{\log n}\right).$$

Hence, using (6.103),

$$d_{\bar{k},k}(G)\binom{k}{2} \geq \binom{\frac{k}{2}}{2} + h^{-1}\left(\log 2 - \frac{\log\binom{n-k}{\bar{k}-\frac{k}{2}}}{\binom{\bar{k}-\frac{k}{2}}{2}}\right)\binom{\bar{k}-\frac{k}{2}}{2} + \frac{\left(\bar{k}-\frac{k}{2}\right)k}{2} - O\left(\sqrt{\bar{k}k\log n} + \left(\bar{k}\right)^{\beta}\sqrt{\log n}\right), \tag{6.103}$$

w.h.p. as $n \to +\infty$.

Using Lemma 6.8.3 for Taylor expanding h^{-1} the lower bound of (6.104) simplifies and yield that $d_{\bar{k},k}(G)\left(\frac{k}{2}\right)$ is at least

$$\frac{1}{2}\binom{\bar{k}-\frac{k}{2}}{2} + \binom{\frac{k}{2}}{2} + \frac{\left(\bar{k}-\frac{k}{2}\right)k}{2} + \Theta\left(\sqrt{\log\left[\binom{n-k}{\bar{k}-\frac{k}{2}}\right]\binom{\bar{k}-\frac{k}{2}}{2}}\right) - O\left(\sqrt{\bar{k}k\log n} + \left(\bar{k}\right)^{\beta}\sqrt{\log n}\right)$$

which since $\beta < 1$ and $k \leq \bar{k}$ is at least

$$\frac{1}{2}\binom{\bar{k}-\frac{k}{2}}{2} + \binom{\frac{k}{2}}{2} + \frac{\left(\bar{k}-\frac{k}{2}\right)k}{2} + \Theta\left(\sqrt{\log\left[\binom{n-k}{\bar{k}-\frac{k}{2}}\right]\binom{\bar{k}-\frac{k}{2}}{2}}\right) - O\left(\bar{k}\sqrt{\log n}\right). \tag{6.104}$$

Furthermore, Lemma 6.8.3 implies

$$\Gamma_{\bar{k},k}(0) = \frac{1}{2}\binom{\bar{k}}{2} + \Theta\left(\sqrt{\log\left[\binom{n-k}{\bar{k}}\right]\binom{\bar{k}}{2}}\right). \tag{6.105}$$

Hence, to establish (6.101) using (6.105), (6.106) it suffices to show that

$$\frac{1}{2}\binom{\bar{k}-\frac{k}{2}}{2} + \binom{\frac{k}{2}}{2} + \frac{\left(\bar{k}-\frac{k}{2}\right)\frac{k}{2}}{2} + \Theta\left(\sqrt{\log\left[\binom{n-k}{\bar{k}-\frac{k}{2}}\right]\binom{\bar{k}-\frac{k}{2}}{2}}\right)$$

is bigger than

$$\frac{1}{2}\binom{\bar{k}}{2} + \Theta\left(\sqrt{\log\left[\binom{n-k}{\bar{k}}\right]\binom{\bar{k}}{2}}\right) + \omega\left(\bar{k}\sqrt{\log n}\right).$$

By direct computation we have

$$\frac{1}{2}\binom{\bar{k}-\frac{k}{2}}{2} + \binom{k/2}{2} + \frac{\left(\bar{k}-\frac{k}{2}\right)\frac{k}{2}}{2} - \frac{1}{2}\binom{\bar{k}}{2} = \frac{k^2}{16} - O\left(\bar{k}\right).$$

Hence, it suffices to have

$$k^2 = \omega\left(\sqrt{\log\left[\binom{n-k}{\bar{k}}\right]\binom{\bar{k}}{2}} - \sqrt{\log\left[\binom{n-k}{\bar{k}-k}\right]\binom{\bar{k}-k}{2}}\right) + \omega\left(\bar{k}\sqrt{\log n}\right). \tag{6.106}$$

Now using the elementary $\binom{n-k}{\bar{k}} \leq (\frac{ne}{\bar{k}})^{\bar{k}}$ and $\binom{n-k}{\bar{k}-k} \geq (\frac{(n-\bar{k})}{\bar{k}})^{\bar{k}-k}$ and the fact that $\bar{k} = \Theta\left(n^C\right)$ for $C < 1/2$ we conclude for (6.107) to hold, it suffices to have

$$k^2 = \omega(\left((\bar{k})^{\frac{3}{2}} - (\bar{k}-k)^{\frac{3}{2}}\right)\sqrt{\log n}) + \omega\left(\bar{k}\sqrt{\log n}\right).$$

Using the elementary inequality, implied by mean value theorem, that for $0 < a \leq b$, $a^{\frac{3}{2}} - b^{\frac{3}{2}} \leq \frac{3}{2}\sqrt{a}\,(a-b)$ it suffices

$$k^2 = \omega\left(\sqrt{\bar{k}}k\sqrt{\log n}\right) + \omega\left(\bar{k}\sqrt{\log n}\right)$$

which now follows directly from our assumptions $k^2 = \omega\left(\bar{k}\log\frac{n}{k^2}\right)$ and $k \leq \bar{k} = n^C$ for $C < 1/2$. The proof of (6.101) and therefore of (6.98) and (6.11) is complete.

We now show how (6.96), (6.99) and (6.101) established above imply the presence of OGP w.h.p. as $n \to +\infty$. We set

$$\zeta_1 := u_1, \zeta_2 := u_2 \text{ and } r := \Gamma_{\bar{k},k}(0) - \frac{c_0}{2}\frac{\bar{k}}{\log\left(\frac{n}{k}\right)}.$$

We start with the second property of $\bar{k}$-OGP. By the definition of ζ_1, ζ_2, r and (6.96) we have

$$\max_{z \in \mathcal{I} \cap [\zeta_1, \zeta_2]} d_{\bar{k},k}(G)(z) < r,$$

w.h.p. as $n \to +\infty$. Using now the definition of $d_{\bar{k},k}(G)(z)$ the last displayed equality directly implies that there is no $\bar{k}$-vertex subset A with $|\mathrm{E}[A]| \geq r$ with $|A \cap \mathcal{PC}| \in [\zeta_1, \zeta_2]$, as we wanted.

For the first part, notice that (6.99), (6.101) and the definition of r imply

$$\min\{d_{\bar{k},k}(G)(0), d_{\bar{k},k}(G)(\frac{k}{2})\} > r,$$

w.h.p. as $n \to +\infty$. Using the definitions of $d_{\bar{k},k}(G)(0), d_{\bar{k},k}(G)(\frac{k}{2})$ respectively we conclude the existence of a $\bar{k}$-vertex subset A with $|\mathrm{E}[A]| \geq r$ and $|A \cap \mathcal{PC}| = 0$ and of a $\bar{k}$-vertex subset B

with $|\mathrm{E}[B]| \geq r$ and $|B \cap \mathcal{PC}| = \frac{k}{2}$, w.h.p. as $n \to +\infty$. Since (6.95) holds, we conclude the first property of $\bar{k}$-OGP. This completes the proof of the presence of $\bar{k}$-OGP and of Theorem 6.2.9.

□

6.7 Conclusion and future directions

The work presented in this Chapter studies the OGP for the planted clique problem. We focus on the way dense subgraphs of the observed graph $G\left(n, \frac{1}{2}, k\right)$ overlap with the planted clique and offer first moment evidence of a fundamental OGP phase transition at $k = \Theta\left(\sqrt{n}\right)$. We establish part of the conjectured OGP phase transition by showing that for any $k, \bar{k}$ satisfying for large n, $k \leq \bar{k} = O\left(n^{0.0917}\right)$ OGP does hold. All of our results are for overparametrized $\bar{k}$-vertex dense subgraphs, where $\bar{k} \geq k$. Introducing this additional free parameter is essential for establishing our results.

Our work prompts to multiple future research directions.

(1) The first and most relevant future direction is establishing the rest parts of Conjecture 6.2.8. We pose this as the main open problem out of this work.

(2) Our result on the value of the K-densest subgraph of an Erdős-Rényi model $G\left(n, \frac{1}{2}\right)$ shows tight concentration of the first and second order behavior of the quantity $d_{\mathrm{ER},K}(G_0)$ defined in (6.12), and applies for any $K \leq n^{0.5-\epsilon}$, for $\epsilon > 0$.

Improving on the third order bounds established in Corollary 2 is of high interest. If the third order term can be proven to be $o\left(K\right)$ for any $K \leq n^{0.5-\epsilon}$ (currently established only for $K \leq n^{0.0917}$) then the first part 1(a) of Conjecture 6.2.8 can be established by following the arguments of this Chapter.

Studying the K-densest subgraph problem for higher values of K appears also an interesting mathematical quest. According to Corollary 2 the second order term behaves as $\Theta\left(K^{\frac{3}{2}}\right)$ (up-to-log n terms) when $K \leq n^{\frac{1}{2}-\epsilon}$. The identification of the exact constant in front of $K^{\frac{3}{2}}$ is interesting. When $K = \Theta\left(n\right)$ the constant is naturally expected to be related to the celebrated Parisi formula (see e.g. [JS18] for similar results for the random MAX-CUT problem and [Sen] for a general technique).

(3) In this Chapter, we use the overparametrization technique as a way to study the landscape of the planted clique problem. Overparametrization has been used extensively in the literature for smoothening "bad" landscapes, but predominantly in the context of deep learning. To the best of our knowledge this is the first time it is used to study computational-

statistical gaps. Without overparametrization the first moment curve obtains a phase transition at the peculiar threshold $k = n^{\frac{2}{3}}$, far above the conjectured onset of algorithmic hardness threshold $k = \sqrt{n}$. *Can the technique of overparametrization be used to study the OGP phase transition of other computational-statistical gaps?* An interesting candidate would be the 3-tensor PCA problem. In this problem, a landscape property called free-energy wells, which is similar to OGP, seems to be appearing in a different place from the conjectured algorithmic threshold (see [BAGJ18] and the discussion therein). It would be very interesting if the free energy wells-algorithmic gap could close using the overparametrization technique.

(4) Last but not least, our work suggests an algorithmic direction. As explained in the introduction, the presence of OGP is rigorously linked with the failure of local methods in multiple problems in the literature. We consider an interesting direction to rigorously show the failure of various fundamental local search methods for finding the planted clique, for example MCMC methods such as Metropolis-Hastings algorithm or the Glauber dynamics, using the presence of $\bar{k}$-OGP as defined in this Chapter.

6.8 Auxilary lemmas

Lemma 6.8.1. *Let $M, N \in \mathbb{N}$ with $N \to +\infty$. Let $X_1, X_2, \ldots, X_N$ abritrary correlated random variables and $Y_1, Y_2, \ldots, Y_N$ i.i.d.* $\mathrm{Bin}\left(M, \frac{1}{2}\right)$*, all living in the same probability space. We also assume $(Y_i)_{i=1,2,\ldots,N}$ are independent of $(X_i)_{i=1,2,\ldots,N}$. Then*

$$\max_{i=1,2,\ldots,N}\{X_i + Y_i\} \geq \max_{i=1,2,\ldots,N}\{X_i\} + \max\{\frac{M}{2} - \sqrt{M \log\log N}, 0\},$$

w.h.p. as $N \to +\infty$.

Proof. It suffices to show that for $i^* := \arg\max_{i=1,2,\ldots,N} X_i$,

$$Y_{i^*} \geq \frac{M}{2} - \sqrt{M \log\log N}$$

w.h.p. as $N \to +\infty$. The result now easily follows from standard Chernoff bound and independence between $(Y_i)_{i=1,2,\ldots,N}$ and i^*. □

For the following two lemmas recall that h is defined in (6.5) and for $\gamma \in (\frac{1}{2}, 1)$, $r(\gamma, \frac{1}{2})$ is defined in (6.7).

Lemma 6.8.2. *Let $N \in \mathbb{N}$ growing to infinity and $\gamma = \gamma_N > \frac{1}{2}$ with $\lim_N \gamma_N = \frac{1}{2}$ and $\lim_N \left(\gamma_N - \frac{1}{2}\right)\sqrt{N} = +\infty$.*

Then for X following $\mathrm{Bin}(N, \frac{1}{2})$ and $N \to +\infty$ it holds

$$\mathbb{P}\left(X = \lceil \gamma N \rceil\right) = \exp\left(-Nr(\gamma, \frac{1}{2}) - \frac{\log N}{2} + O\left(1\right)\right)$$

and

$$\mathbb{P}\left(X \geq \lceil \gamma N \rceil\right) = \exp\left(-Nr(\gamma, \frac{1}{2}) - \Omega\left(\log\left(\left(\gamma - \frac{1}{2}\right)\sqrt{N}\right)\right)\right),$$

where $r\left(\gamma, \frac{1}{2}\right)$ is defined in (6.7).

Proof. We have by Stirling approximation

$$\binom{N}{\lceil \gamma N \rceil} = \exp\left(Nh\left(\gamma\right) - \frac{1}{2}\log\left(N\gamma\left(1-\gamma\right)\right) + O\left(1\right)\right) \tag{6.107}$$

In particular, using $r\left(\gamma,\frac{1}{2}\right)=h\left(\frac{1}{2}\right)-h\left(\gamma\right)=\log 2-h\left(\gamma\right)$ and that $\gamma=\frac{1}{2}+o_N\left(1\right)$ we conclude

$$\mathbb{P}\left(X=\lceil\gamma N\rceil\right)=\binom{N}{\lceil\gamma N\rceil}\frac{1}{2^N}=\exp(-Nr(\gamma,\frac{1}{2})-\frac{1}{2}\log N+O\left(1\right))$$

Now using standard binomial coefficient inequalities (see e.g. Proposition 1(c) in [Kla00]) we have that for any $1\leq k\leq N/2$,

$$\mathbb{P}\left(X\geq\left\lceil\frac{N}{2}+k\right\rceil\right)\leq\frac{\frac{N}{2}+k}{2k+1}\mathbb{P}\left(X=\left\lceil\frac{N}{2}+k\right\rceil\right).$$

Hence for large enough N if we set $k=\left(\gamma-\frac{1}{2}\right)N$ we have,

$$\begin{aligned}
\mathbb{P}\left(X\geq\lceil\gamma N\rceil\right)&\leq\left(\frac{\gamma}{2\gamma-1}+o\left(1\right)\right)\mathbb{P}\left(X=\lceil\gamma N\rceil\right)\\
&=\left(\frac{\gamma}{2\gamma-1}+o\left(1\right)\right)\exp(-Nr(\gamma,\frac{1}{2})-\frac{1}{2}\log N+O\left(1\right))\\
&=\left(1+o\left(1\right)\right)\left(\frac{\gamma}{2\gamma-1}\right)\exp(-Nr(\gamma,\frac{1}{2})-\frac{1}{2}\log N+O\left(1\right)),\text{ since }\lim_N\gamma_N=\frac{1}{2}>0\\
&=\exp\left(-Nr(\gamma,\frac{1}{2})+\log\left(\frac{2\gamma}{(2\gamma-1)\sqrt{N}}\right)+O(1)\right)\\
&=\exp\left(-Nr(\gamma,\frac{1}{2})-\Omega\left(\log\left(\left(\gamma-\frac{1}{2}\right)\sqrt{N}\right)\right)\right).
\end{aligned}$$

The proof of the Lemma 6.8.2 is complete. □

Lemma 6.8.3. *For $\epsilon=\epsilon_n\to 0$, it holds*

$$h^{-1}\left(\log 2-\epsilon\right)=\frac{1}{2}+\frac{1}{\sqrt{2}}\sqrt{\epsilon}-\frac{1}{6\sqrt{2}}\epsilon^{\frac{3}{2}}+O\left(\epsilon^{\frac{5}{2}}\right).$$

Proof. Let $\Phi(x):=\sqrt{\log 2-h\left(\frac{1}{2}+x\right)},x\in[0,\frac{1}{2}]$. We straightforwardly calculate that for the sequence of derivatives at 0, $a_i:=\Phi^{(i)}(0)$, $i\in\mathbb{Z}_{\geq 0}$ it holds $a_0=0,a_1=\sqrt{2},a_2=0,a_3=2\sqrt{2}$ and $a_4=0$.

Notice that for all $\epsilon\in(0,\log 2)$ and Φ^{-1} the inverse of Φ,

$$h^{-1}\left(\log 2-\epsilon\right)=\frac{1}{2}+\Phi^{-1}(\sqrt{\epsilon}).$$

Lemma follows if we establish that Taylor expansion of Φ^{-1} around $y = 0$ is given by

$$\Phi^{-1}(y) = \frac{1}{\sqrt{2}}y - \frac{1}{6\sqrt{2}}y^3 + O\left(y^5\right). \tag{6.108}$$

Clearly $\Phi^{-1}(0) = 0$. For $b_i := \left(\Phi^{-1}\right)^{(i)}(0)$, $i \in \mathbb{Z}_{\geq 0}$ by standard calculations using the Lagrange inversion theorem we have $b_0 = 0$,

$$b_1 = \frac{1}{a_1} = \frac{1}{\sqrt{2}},$$

$$b_2 = -\frac{a_2}{2a_1} = 0,$$

$$b_3 = \frac{1}{2\sqrt{2}}\left[-\frac{a_3}{a_1} + 3\left(\frac{a_2}{a_1}\right)^2\right] = -\frac{1}{\sqrt{2}}$$

and

$$b_4 = \frac{1}{4}\left[-\frac{a_4}{a_1} + \frac{10}{3}\frac{a_2 a_3}{a_1^2} - 60\frac{a_2}{a_1}\right] = 0.$$

From this point, Taylor expansion yields that for small y

$$\Phi^{-1}(y) = b_0 + b_1 y + \frac{b_2}{2}y^2 + \frac{b_3}{6}y^3 + \frac{b_4}{24}y^4 + O\left(y^5\right)$$

which given the values of $b_i, i = 0, 1, 2, 3, 4$ yields (6.109). The proof of the Lemma is complete.

□

The following elementary calculus properties are used throughout the proof sections.

Lemma 6.8.4. *Suppose $(a_n)_{n\in\mathbb{N}}, (b_n)_{n\in\mathbb{N}}$ are two sequences of positive real numbers. The following are true.*

(a) The sequence $a_n \log a_n$ converges to zero if and only if a_n converges to zero.

(b) The sequence $a_n \log a_n$ diverges to infinity if and only if a_n diverges to infinity.

(c) Suppose b_n diverges to infinity. Then $a_n \log a_n = \omega(b_n)$ if and only if $a_n = \omega\left(\frac{b_n}{\log b_n}\right)$.

(d) Suppose b_n diverges to infinity. Then $a_n = \omega(b_n \log b_n)$ if and only if $\frac{a_n}{\log a_n} = \omega(b_n)$.

Proof. Both properties (a), (b) follow in a straightforward way from the continuity of the mapping

$$x \in (0, \infty) \to x \log x \in \mathbb{R},$$

and the limiting behaviors

$$\lim_{x \to 0} x \log x = 0, \lim_{x \to +\infty} x \log x = +\infty.$$

Regarding property (c): For the one direction, assume

$$\lim_n \frac{a_n \log a_n}{b_n} = +\infty \tag{6.109}$$

and c_n is defined by $a_n = \frac{c_n b_n}{\log b_n}$. It suffices to show c_n diverges to infinity. By (6.110) we know

$$\lim_n c_n \frac{\log\left(\frac{c_n b_n}{\log b_n}\right)}{\log b_n} = +\infty. \tag{6.110}$$

Assuming $\liminf_n c_n < +\infty$ it follows since $\lim_n b_n = +\infty$ that

$$\liminf_n c_n \frac{\log\left(\frac{c_n b_n}{\log b_n}\right)}{\log b_n} \leq \liminf_n c_n \frac{\log c_n + \log b_n}{\log b_n} \leq \liminf_n c_n < \infty,$$

a direct contradiction with (6.111). This completes the proof of this direction.

For the other direction, assume

$$\lim_n \frac{a_n \log b_n}{b_n} = +\infty \tag{6.111}$$

and c_n is defined by $a_n \log a_n = c_n b_n$. It suffices to show c_n diverges to infinity. By (6.112) we know

$$\lim_n c_n \frac{\log\left(\frac{a_n \log a_n}{c_n}\right)}{\log a_n} = +\infty. \tag{6.112}$$

Note that since b_n diverges to infinity (6.112) implies that a_n diverges to infinity as well. Assuming

$\liminf_n c_n < +\infty$ it follows since $\lim_n a_n = +\infty$ that

$$\liminf_n c_n \frac{\log\left(\frac{a_n \log a_n}{c_n}\right)}{\log a_n} \leq \liminf_n c_n \frac{\log a_n + \log\log a_n - \log c_n}{\log a_n} \leq \liminf_n c_n < \infty,$$

a direct contradiction with (6.113). This completes the proof of this direction.

Property (d) follows by similar reasoning as in the case of property (c). □

Chapter 7

Conclusion

In this thesis, we study the computational and statistical challenges of two well-established high dimensional statistical models; the high dimensional linear regression model and the planted clique model. We establish multiple results regarding their statistical and computational limits. From a statistical perspective, in Chapter 2 we identify, under certain assumptions, sharply the statistical limit of high dimensional linear regression revealing an all-to-nothing phase transition. From a computational perspective, in Chapter 5 we propose a new polynomial-time algorithm for noiseless high dimensional linear regression using lattice basis reduction, which can provably recover the vector of coefficients with access to only one sample, $n = 1$.

A large focus of this thesis is dedicated to studying the property that for both these models, statistical inference using unbounded computational power becomes possible in regimes where no computationally efficient method is known to succeed, a property known as a computational-statistical gap. In Chapters 3 and 4 we study the computational statistical gap of high dimensional linear regression, and in Chapter 6, we study the computational statistical gap of the planted clique model. In both cases we offer a possible explanation to this phenomenon by providing a rigorous link between the computational statistical gap and the presence of a certain Overlap Gap Property. The Overlap Gap Property find its origin in spin glass theory and is known to be linked with algorithmic hardness. We conjecture that this connection is of fundamental nature and the study of the Overlap Gap Property can provide a rigorous generic explanation for the appearance of computational-statistical gaps in high dimensional statistical models.

Bibliography

[Abb17] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *arXiv*, 2017.

[ACO08] Dimitris Achlioptas and Amin Coja-Oghlan. Algorithmic barriers from phase transitions. In *FOCS'08 IEEE 49th Annual Symposium on Foundations of Computer Science, 2008*, pages 793–802, 2008.

[ACOGM17] Peter Ayre, Amin Coja-Oghlan, Pu Gao, and Noela Muller. The satisfiability threshold for random linear equations. *arXiv*, 2017.

[ACORT11] 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.

[AKJ17] Ahmed El Alaoui, Florent Krzakala, and Michael I Jordan. Finite size corrections and likelihood ratio fluctuations in the spiked Wigner model. *arXiv preprint arXiv:1710.02903*, 2017.

[AKKT02] Dimitris Achlioptas, Jeong Han Kim, Michael Krivelevich, and Prasad Tetali. Two-coloring random hypergraphs. *Random Structures & Algorithms*, 20(2):249–259, 2002.

[AKS98] Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. *Random Structures and Algorithms*, 13(3-4):457–466, 1998.

[ASZ10] Shuchin Aeron, Venkatesh Saligrama, and Manqi Zhao. Information theoretic bounds for compressed sensing. *IEEE Transactions on Information Theory*, 56(10):5111–5130, October 2010.

[AT10] Mehmet Akcakaya and Vahid Tarokh. Shannon-theoretic limits on noisy compressive sampling. *IEEE Transactions on Information Theory*, 56(1):492–504, December 2010.

[BAGJ18] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Algorithmic thresholds for tensor pca. *arXiv*, 2018.

[Bal67] Bruno Baldessari. The distribution of a quadratic form of normal random variables. *The Annals of Mathematical Statistics*, 38(6):1700–1704, 1967.

[BB99] L. Brunel and J. Boutros. Euclidean space lattice decoding for joint detection in cdma systems. In *Proceedings of the 1999 IEEE Information Theory and Communications Workshop (Cat. No. 99EX253)*, 1999.

[BBH18] Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. *Conference on Learning Theory (COLT)*, 2018.

[BBHL09] Peter J. Bickel, James B. Brown, Haiyan Huang, and Qunhua Li. An overview of recent developments in genomics and associated statistical methods. *Phil. Trans. R. Soc. A*, 2009.

[BBSV18] Paul Balister, Bela Bollobas, Julian Sahasrabudhe, and Alexander Veremyev. Dense subgraphs in random graphs. *arXiv*, 2018.

[BC12] A. R. Barron and S. Cho. High-rate sparse superposition codes with iteratively optimal estimates. *Proc. IEEE Int. Symp. Inf. Theory*, 2012.

[BCSZ18] Christian Borgs, Jennifer T. Chayes, Adam D. Smith, and Ilias Zadik. Revealing network structure confidentially: Improved rates for node-private graphon estimation. In *Symposium of the Foundations of Computer Science (FOCS)*, 2018.

[BD09] Thomas Blumensath and Mike E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265 – 274, 2009.

[BDDW08] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, Dec 2008.

[BDMK16] Jean Barbier, Mohamad Dia, Nicolas Macris, and Florent Krzakala. The mutual information in random linear estimation. In *Proceedings of the Allerton Conference on Communication, Control, and Computing*, Monticello, IL, 2016.

[BHK+16] Boaz Barak, Samuel Hopkins, Jonathan Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. *57th Annual Symposium on Foundations of Computer Science (FOCS)*, 2016.

[BJPD17] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G. Dimakis. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, pages 537–546, 2017.

[BLH+14] Zhao Bo, Wenmiao Lu, T. Kevin Hitchens, Fan Lam, Chien Ho, and Zhi-Pei Liang. Accelerated mr parameter mapping with low-rank and sparsity constraints. *Magnetic Resonance in Medicine*, 2014.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[BMNN16] Jess Banks, Cristopher Moore, Joe Neeman, and Praneeth Netrapalli. Information-theoretic thresholds for community detection in sparse networks. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, NY, June 23-26 2016*, pages 383–416, 2016.

[BMR+18] J. Banks, C. Moore, Vershynin R., N. Verzelen, and J. Xu. Information-theoretic bounds and phase transitions in clustering, space pca, and submatrix localization. *IEEE Transactions on Information Theory*, 2018.

[BMV+18] J. Banks, C. Moore, R. Vershynin, N. Verzelen, and J. Xu. Information-theoretic bounds and phase transitions in clustering, sparse pca, and submatrix localization. *IEEE Transactions on Information Theory*, 64(7):4872–4894, 2018.

[Bol85] B. Bollobas. *Random Graphs.* Academic Press, Inc., 1985.

[Bor11] Mazen Al Borno. Reduction in solving some integer least squares problems. *arXiv Preprint*, 2011.

[BPW18] Afonso S Bandeira, Amelia Perry, and Alexander S. Wein. Notes on computational-statistical gaps: predictions using statistical physics. *Arxiv preprint arXiv:1803.11132.pdf*, 2018.

[BR13] Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*, pages 1046–1066, 2013.

[BRT09a] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 08 2009.

[BRT09b] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 08 2009.

[BTK+17] I. Ben Atitallah, C. Thrampoulidis, A. Kammoun, T. Y. Al-Naffouri, M. Alouini, and B. Hassibi. The box-lasso with application to gssk modulation in massive mimo systems. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1082–1086, June 2017.

[CCL+08] Carlos M. Carvalho, Jeffrey Chang, Joseph E. Lucas, Joseph R. Nevins, Quanli Wang, and Mike West. High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 2008.

[CDS01] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, January 2001.

[CESV15] Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.

[CGPR17] Wei-Kuo Chen, David Gamarnik, Dmitry Panchenko, and Mustazee Rahman. Suboptimality of local algorithms for a class of max-cut problems, 2017.

[CH90] Alan Miller. Chapman and Hall. Subset selection in regression. *Chapman and Hall*, 1990.

[Cho14] S. Cho. High-dimensional regression with random design, including sparse superposition codes. *Ph.D. dissertation, Dept. Statist., Yale Univ., New Haven, CT, USA*, 2014.

[CL99] V.C. Chen and Hao Ling. Joint time-frequency analysis for radar signal and image processing. *IEEE Transactions on Signal Processing*, 1999.

[CLR17] Tony Cai, Tengyuan Liang, and Alexander Rakhlin. Computational and statistical boundaries for submatrix localization in a large noisy matrix. *The Annals of Statistics*, 2017.

[COE11] 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.

[COHH16] Amin Coja-Oghlan, Amir Haqshenas, and Samuel Hetterich. Walksat stalls well below the satisfiability threshold. *arXiv preprint arXiv:1608.00346*, 2016.

[CRT06] Emmanuel J. Candes, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[CT05] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[CT07] Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 12 2007.

[CW11] T. T. Cai and L. Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory*, 57(7):4680–4688, July 2011.

[CWD16] B. P. Chapman, A. Weiss, and P. R. Duberstein. Statistical learning theory for high dimensional prediction: Application to criterion-keyed scale development. *Psychological Methods*, 21(4):603–620, 2016.

[DGGP14] Yael Dekel, Ori Gurel-Gurevich, and Yuval Peres. Finding hidden cliques in linear time with high probability. *Combinatorics, Probabability and Computing*, 2014.

[DJM13] D. L. Donoho, A. Javanmard, and A. Montanari. Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Transactions on Information Theory*, 59(11):7434–7464, Nov 2013.

[DM] Y. Deshpande and A. Montanari. Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. *arxiv.org/abs/1304.7047*.

[DMM09] David L. Donoho, Arian Maleki, and Andrea Montanari. Message passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences of the United States of America*, 106 45:18914–9, 2009.

[Don06] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[DT05] David L Donoho and Jared Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9452–9457, 2005.

[DT09] David Donoho and Jared Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293, 2009.

[DT10] David L. Donoho and Jared Tanner. Counting the faces of randomly-projected hypercubes and orthants, with applications. *Discrete & Computational Geometry*, 43(3):522–541, Apr 2010.

[Dud17] J. Dudczyk. A method of feature selection in the aspect of specific identification of radar signals. *Bulletin of the Polish Academic of Sciences, Technical Sciences*, 2017.

[EACP11] Emmanuel J. Cand'es Ery Arias-Castro and Yaniv Plan. Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, 2011.

[EL85] P. Erdos and G.G. Lorentz. On the probability that n and g(n) are relatively prime. *Acta Arith.*, 5:524–531, 1985.

[FGR+17] Vitaly Fedman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *Journal of the ACM (JACM)*, 2017.

[FHY07] Jianqing Fan, Peter Hall, and Qiwei Yao. To how many simultaneous hypothesis tests can normal, student's t or bootstrap calibration be applied? *Journal of the American Statistical Association*, 102(480):1282–1288, 2007.

[Fis22] R. A. Fisher. The goodness of fit of regression formula, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, 85(4):597–612, 1922.

[FR10] Uriel Feige and Dorit Ron. Finding hidden cliques in linear time. In *AofA*, 2010.

[FR13] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Springer, 2013.

[FRG09] Alyson K. Fletcher, Sundeep Rangan, and Vivek K Goyal. Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Transactions on Information Theory*, 55(12):5758–5772, November 2009.

[Fri86] Alan M. Frieze. On the lagarias-odlyzko algorithm for the subset sum problem. *SIAM J. Comput.*, 15:536–539, 1986.

[Gal85] Francis Galton. Presidential address. *Section H, Anthropology*, 1885.

[GAM+18] J. O. Garcia, A. Ashourvan, S. Muldoon, J. M. Vettel, and D. S. Bassett. Applications of community detection techniques to brain graphs: Algorithmic considerations and implications for neural function. *Proceedings of the IEEE*, 106(5):846–867, May 2018.

[Geo12] Edward I. George. The variable selection problem. *Journal of the American Statistical Association*, 2012.

[GL16] David Gamarnik and Quan Li. Finding a large submatrix of a gaussian random matrix. *arXiv preprint arXiv:1602.08529*, 2016.

[GM75] G. Grimmett and C. McDiarmid. On colouring random graphs. *Mathematical Proceedings of the Cambridge Philosophical Society*, 1975.

[GSa] David Gamarnik and Madhu Sudan. Limits of local algorithms over sparse random graphs. *Annals of Probability.* To appear.

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

[GV05] Dongning Guo and Sergio Verdú. Randomly spread CDMA: Asymptotics via statistical physics. *IEEE Transactions on Information Theory*, 51(6):1983–2010, June 2005.

[GZ17a] David Gamarnik and Ilias Zadik. High dimensional linear regression with binary coefficients: Mean squared error and a phase transition. *Conference on Learning Theory (COLT)*, 2017.

[GZ17b] David Gamarnik and Ilias Zadik. Sparse high dimensional linear regression: Algorithmic barrier and a local search algorithm. 2017.

[GZ18] David Gamarnik and Ilias Zadik. High dimensional linear regression using lattice basis reduction. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[GZ19] David Gamarnik and Ilias Zadik. The landscape of the planted clique problem: Dense subgraphs and the overlap gap property. *arXiv Preprint arXiv:1904.07174*, 2019.

[HB98] A. Hassibi and S. Boyd. Integer parameter estimation in linear models with applications to gps. *IEEE Transactions on Signal Processing*, 1998.

[HC08] Stefani M. dos Remedios C. G. Ho, J. W. and M. A. Charleston. Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics*, 2008.

[Het16] Samuel Hetterich. Analysing survey propagation guided decimation on random formulas. *arXiv preprint arXiv:1602.08519*, 2016.

[HG10] Qiu X. Hu, R. and G. Glazko. A new gene selection procedure based on the covariance distance. *Bioinformatics*, 2010.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2009.

[HTW15] Trevor Hastie, Robert Tibshirani, and Martin J. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability, 2015.

[HV02] B. Hassibi and H. Vikalo. On the expected complexity of integer least-squares problems. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[HW71] D. L. Hanson and F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Statist.*, 42(3):1079–1083, 06 1971.

[HW75] G.H. Hardy and E.M. Wright. *An Introduction to the Theory of Numbers*. Oxford Science Publications, fifth edition edition, 1975.

[HY09] Qiu X. Glazko G. Klevanov L. Hu, R. and A. Yakovlev. Detecting intergene correlation changes in microarray analysis: A new approach to gene selection. *BMC Bioinformatics*, 2009.

[JB12] Antony Joseph and Andrew R. Barron. Least sqaures superposition codes of moderate dictionarysize are reliable at rates up to capacity. *IEEE Transactions on Information Theory*, 2012.

[JB14] A. Joseph and A. R. Barron. Fast sparse superposition codes have near exponential error probability for r $<$ c,. *IEEE Trans. Inf. Theory, vol. 60, no. 2, pp. 919–942*, 2014.

[JBC17] Lucas Janson, Rina Foygel Barber, and Emmanuel Candès. Eigenprism: inference for high dimensional signal-to-noise ratios. *Journal of the Royal Statistical Society. Series B*, 2017.

[Jer92] Mark Jerrum. Large cliques elude the metropolis process. *Random Structures & Algorithms*, 1992.

[JKR11] Yuzhe Jin, Young-Han Kim, and Bhaskar D Rao. Limits on support recovery of sparse signals via multiple-access communication techniques. *IEEE Transactions on Information Theory*, 57(12):7877–7892, 2011.

[JS18] Aukosh Jagannath and Subhabrata Sen. On the unbalanced cut problem and the generalized sherrington-kickpatrick model. *arXiv Preprint*, 2018.

[Kar72] Richard M Karp. Reducibility among combinatorial problems. *Complexity of Computer Computations*, 1972.

[Kea98] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6):983–1006, November 1998.

[KKM+17] Shrinivas Kudekar, Santhosh Kumar, Marco Mondelli, Henry D Pfister, Eren Şaşoğlu, and Rüdiger L Urbanke. Reed–muller codes achieve capacity on erasure channels. *IEEE Transactions on Information Theory*, 63(7):4298–4316, 2017.

[Kla00] Bernhard Klar. Bounds on tail probabilities of discrete distributions. *Probability in the Engineering and Informational Sciences*, 2000.

[KMRT+07] F. Krzakała, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborová. Gibbs states and the set of solutions of random constraint satisfaction problems. *Proceedings of the National Academy of Sciences*, 104(25):10318–10323, 2007.

[KMS+12] F. Krzakala, M. Mézard, F. Sausset, Y. F. Sun, and L. Zdeborová. Statistical-physics-based reconstruction in compressed sensing. *Phys. Rev. X*, 2:021005, May 2012.

[Kuč95] Luděk Kučera. Expected complexity of graph partitioning problems. *Discrete Applied Mathematics*, 1995.

[Kum10] Nirman Kumar. Bounding the volume of hamming balls. `https://cstheory.wordpress.com/2010/08/13/bounding-the-volume-of-hamming-balls/`, Aug. 2010.

[Las01] Jean B. Lasserre. An explicit exact sdp relaxation for nonlinear 0-1 programs. In Karen Aardal and Bert Gerards, editors, *Integer Programming and Combinatorial Optimization*, pages 293–303, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[LDSP08] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly. Compressed sensing mri. *IEEE Signal Processing Magazine*, 25(2):72–82, March 2008.

[Lem79] Abraham Lempel. Cryptology in transition. *ACM Comput. Surv.*, 11(4):285–303, December 1979.

[LLL82] Arjen Klaas Lenstra, Hendrik Willem Lenstra, and László Lovász. Factoring polynomials with rational coefficients. *Mathematische Annalen*, 261(4):515–534, 1982.

[LM00] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[LMZ18] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. *Conference on Learning Theory (COLT)*, 2018.

[LO85] Jeffrey C Lagarias and Andrew M Odlyzko. Solving low-density subset sum problems. *Journal of the ACM (JACM)*, 32(1):229–246, 1985.

[LPW06] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2006.

[LVZ17] Miles Lubin, Juan Pablo Vielma, and Ilias Zadik. Mixed integet convex representability. In *Integer Programming and Combinatorial Optimization (IPCO) Conference*, 2017.

[LVZ18] Miles Lubin, Juan Pablo Vielma, and Ilias Zadik. Regularity in mixed integer convex representability. *arxiv*, 2018.

[Mar77] Brian G. Marsden. Carl friedrich gauss, astronomer. *Journal of the Royal Astronomical Society of Canada*, 71, 08 1977.

[MB06a] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 2006.

[MB06b] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 06 2006.

[MH78] R. Merkle and M. Hellman. Hiding information and signatures in trapdoor knapsacks. *IEEE Transactions on Information Theory*, 24(5):525–530, Sep 1978.

[MMU08] Cyril Méasson, Andrea Montanari, and Rüdiger Urbanke. Maxwell construction: The hidden bridge between iterative and maximum a posteriori decoding. *IEEE Transactions on Information Theory*, 54(12):5277–5307, 2008.

[MMZ05] M. Mézard, T. Mora, and R. Zecchina. Clustering of solutions in the random satisfiability problem. *Physical Review Letters*, 94(19):197205, 2005.

[MNS15] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015.

[MRT11] 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.

[MSZ18] L. W. Mackey, V. Syrgkanis, and I. Zadik. Orthogonal machine learning: Power and limitations. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[Nag76] S. V. Nagaev. An estimate of the remainder term in the multidimensional central limit theorem. pages 419–438. Lecture Notes in Math., Vol. 550, 1976.

[NP33] Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 1933.

[NT18] Mohamed Ndaoud and Alexandre B Tsybakov. Optimal variable selection and adaptive noisy compressed sensing. *arXiv preprint arXiv:1809.03145*, 2018.

[OTH13] S. Oymak, C. Thrampoulidis, and B. Hassibi. The squared-error of generalized lasso: A precise analysis. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1002–1009, Oct 2013.

[OWJ11] Guillaumem Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 2011.

[Par00] Pablo A. Parrilo. Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization. Technical report, 2000.

[PDFV05] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[PW15] Yury Polyanskiy and Yihong Wu. Lecture Notes on Information Theory. Feb 2015. `http://people.lids.mit.edu/yp/homepage/data/itlectures_v4.pdf`.

[PWB16] Amelia Perry, Alexander S. Wein, and Afonso S. Bandeira. Statistical limits of spiked tensor models. arXiv:1612.07728, Dec. 2016.

[PZHS16] Bao Peng, Zhi Zhao, Guangjie Han, and Jian Shen. Consensus-based sparse signal reconstruction algorithm for wireless sensor networks. *International Journal of Distributed Sensor Networks*, 2016.

[QMP+12] Giorgio Quer, Riccardo Masiero, Gianluigi Pillonetto, Michele Rossi, and Michele Zorzi. Sensing, compression, and recovery for wsns: Sparse signal modeling and monitoring framework. *IEEE Transactions on Wireless Communications*, 2012.

[Rad11] K. Rahnama Rad. Nearly sharp sufficient conditions on exact sparsity pattern recovery. *IEEE Transactions on Information Theory*, 57(7):4672–4679, July 2011.

[Ree17] Galen Reeves. Conditional central limit theorems for Gaussian projections. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 3055–3059, Aachen, Germany, June 2017.

[RG12] G. Reeves and M. Gastpar. The sampling rate-distortion tradeoff for sparsity pattern recovery in compressed sensing. *IEEE Transactions on Information Theory*, 58(10):3065–3092, May 2012.

[RG13] Galen Reeves and Michael Gapstar. Approximate sparsity pattern recovery: Information-theoretic lower bounds. *IEEE Trans. Information Theory*, 2013.

[RGV17] C. Rush, A. Greig, and R. Venkataramanan. Capacity-achieving sparse superposition codes via approximate message passing decoding. *IEEE Trans. Inf. Theory, vol. 63, pp. 1476–1500*, 2017.

[RP16] Galen Reeves and Henry D. Pfister. The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 665 – 669, Barcelona, Spain, July 2016. arXiv. Available: `https://arxiv.org/abs/1607.02524`.

[RV14] Mustazee Rahman and Balint Virag. Local algorithms for independent sets are half-optimal. *arXiv preprint arXiv:1402.0485*, 2014.

[RXZ19] Galen Reeves, Jiaming Xu, and Ilias Zadik. The all-or-nothing phenomenon in sparse linear regression. *Conference on Learning Theory (COLT)*, 2019.

[SC15] Jonathan Scarlett and Volkan Cevher. Limits on support recovery with probabilistic models: An information-theoretic framework. *IEEE International Symposium on Information Theory (ISIT)*, 2015.

[SC17] Jonathan Scarlett and Volkan Cevher. Limits on support recovery with probabilistic models: An information-theoretic framework. *IEEE Transactions on Information Theory*, 63(1):593–620, September 2017.

[Sen] Subhabrata Sen. Optimization on sparse random hypergraphs and spin glasses. *Random Structures and Algorithms,* to appear.

[Sha82] A. Shamir. A polynomial time algorithm for breaking the basic merkle-hellman cryptosystem. In *23rd Annual Symposium on Foundations of Computer Science (sfcs 1982)*, pages 145–152, Nov 1982.

[SS17] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. *arXiv*, 2017.

[Sti81] Stephen M. Stigler. Gauss and the invention of least squares. *Ann. Statist.*, 9(3):465–474, 05 1981.

[SW95] A. Shwartz and A. Weiss. *Large deviations for performance analysis*. Chapman and Hall, 1995.

[Tal10] M. Talagrand. *Mean Field Models for Spin Glasses: Volume I: Basic Examples*. Springer, 2010.

[Tan02] T. Tanaka. A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors. *IEEE Transactions on Information Theory*, 48(11):2888–2910, November 2002.

[TWY12] Cai Tony, Liu Weidong, and Xia Yin. Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association*, 2012.

[TZP19] Christos Thrampoulidis, Ilias Zadik, and Yury Polyanskyi. A simple bound on the ber of the map decoder for massive mimo systems. *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, 2019.

[VBB18] Luca Venturi, Afonso S Bandeira, and Joan Bruna. Spurious valleys in two-layer neural network optimization landscapes. *arXiv Preprint arxiv:1802.06384*, 2018.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[Wai09a] Martin J Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741, 2009.

[Wai09b] 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, 2009.

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[WBP16] Tengyao Wang, Quentin Berthet, and Yaniv Plan. Average-case hardness of rip certification. *Neural Information Processing Systems (NeurIPS)*, 2016.

[WWR10] Wei Wang, Martin J Wainwright, and Kannan Ramchandran. Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *IEEE Transactions on Information Theory*, 56(6):2967–2979, 2010.

[WX18] Yihong Wu and Jiaming Xu. Statistical problems with planted structures: Information theoretical and computational limits. *arXiv Preprint*, 2018.

[XHM18] Ji Xu, Daniel J Hsu, and Arian Maleki. Benefits of over-parameterization with em. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10662–10672. Curran Associates, Inc., 2018.

[XZB01] Y. Shi X. Zhang and Z. Bao. A new feature vector using selected bispectra for signal classification with application in radar target recognition. *IEEE Transactions on Signal Processing*, 2001.

[Yul97] Udny Yule. On the theory of correlation. *Journal of the Royal Statistical Society*, 60:812–54, 1897.

[ZCZ+09] Z. Zhu, G. Cao, S. Zhu, S. Ranjan, and A. Nucci. A social network based patching scheme for worm containment in cellular networks. In *IEEE INFOCOM 2009*, pages 1476–1484, April 2009.

[Zha93] Ping Zhang. Model selection via multifold cross validation. *The Annals of Statistics*, 1993.

[ZTP19] Ilias Zadik, Christos Thrampoulidis, and Yury Polyanskyi. Improved bounds on gaussian mac and sparse regression via gaussian inequalities. *IEEE International Symposium on Information Theory (ISIT)*, 2019.

[ZY06] Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, December 2006.