Algorithms and Algorithmic Barriers in High-Dimensional Statistics and Random Combinatorial Structures

by

Eren C. Kızıldağ

B.S., Boğaziçi University (2014)

S.M., Massachusetts Institute of Technology (2017)

Submitted to the Department of
Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author .....
Department of Electrical Engineering and Computer Science
May 13, 2022

Certified by .....
David Gamarnik
Nanyang Technological University Professor of Operations Research
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Algorithms and Algorithmic Barriers in High-Dimensional Statistics and Random Combinatorial Structures

by

Eren C. Kızıldağ

B.S., Boğaziçi University (2014)

S.M., Massachusetts Institute of Technology (2017)

Submitted to the Department of Electrical Engineering and Computer Science
on May 13, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

We focus on several algorithmic problems arising from the study of random combinatorial structures and of neural network models, with a particular emphasis on computational aspects. Our main contributions are summarized as follows.

1. Our first focus is on two algorithmic problems arising from the study of random combinatorial structures: the *random number partitioning problem* (`NPP`) and the *symmetric binary perceptron model* (`SBP`). Both of these models exhibit a so-called *statistical-to-computational gap*: a striking gap between the existential and the best known algorithmic guarantees with bounded computational power (such as polynomial-time algorithms). We investigate the nature of this gap for the `NPP` and `SBP` by studying their landscape through the lens of statistical physics and in particular spin glass theory. We establish that both models exhibit the *Overlap Gap Property* (OGP), an intricate geometrical property that is known to be a rigorous barrier for large classes of algorithms. We then leverage the OGP to rule out certain important classes of algorithms, including the class of stable algorithms and the Monte Carlo Markov Chain type algorithms. The former is a rather powerful abstract class that captures the implementation of several important algorithms including the approximate message passing and the low-degree polynomial based methods. Our hardness results for the stable algorithms are based on Ramsey Theory from extremal combinatorics. To the best of our knowledge, this is the first usage of Ramsey Theory to show algorithmic hardness for models with random parameters.

2. Our second focus is on the *Sherrington-Kirkpatrick* (SK) spin glass model, a mean-field model for disordered random media. We establish that the algorithmic problem of exactly computing the partition function of the SK model is average-case hard under the assumption that $P \neq \#P$ (an assumption that is milder than $P \neq NP$ and is widely believed to be true) both for the finite-precision arithmetic model and for the real-valued computational model. Our

result is the first provable hardness result for a statistical physics model with random parameters that is based on standard complexity-theoretical assumptions.

3. Our last focus is on *neural network* (NN) models arising from modern machine learning and high-dimensional statistical inference tasks.

    - Our first set of results to this direction establishes *self-regularity* for two-layer NNs with sigmoid, binary step, rectified linear unit (ReLU) activation functions and non-negative output weights in an algorithm-independent manner. That is, we establish that under very mild distributional assumptions on the training data, any such network has a bounded output norm provided that it attains a small training error on polynomially many data. Our results explain why the overparameterization does not hurt the generalization ability for such architectures. This conundrum has been observed empirically in NNs and defies the classical statistical wisdom.
    - Our final focus is on the problem of learning two-layer NNs with quadratic activation functions under the assumption that the training data are generated by a so-called *teacher* network with planted weights. We first investigate the training aspect, establishing that there exists an energy barrier E_0 below which any stationary point of the empirical risk is necessarily a global optimum. That is, there are no spurious stationary points below E_0. Consequently, we show that the gradient descent algorithm, when initialized below E_0, nearly recovers the planted weights in polynomial-time. We then investigate the question of proper initialization under the assumption that the planted weights are generated randomly. By leveraging a certain semicircle law from random matrix theory, we show that a deterministic initialization suffices, provided that the network is sufficiently overparameterized. Finally, we identify a simple necessary and sufficient geometric condition on the training data under which any minimizer of the empirical risk has good generalization. We lastly show that randomly generated data satisfy this condition almost surely under very mild distributional assumptions.

Thesis Supervisor: David Gamarnik
Title: Nanyang Technological University Professor of Operations Research

Acknowledgments

My (formal) studentship has finally come to an end. As I am writing these final lines that conclude a large chapter of my life in a small Harvard library on a rather cold Cambridge spring day[1], I feel a mixture of emotions. Even though the path was not always smooth, but rather blistered and rugged sometimes, I am proud to say that it was totally worth it. It was not only an intellectual journey, but it also made me much stronger and more confident by giving me the opportunity to rediscover myself. In what follows, I will thank certain individuals who made this journey possible.

The first and most sincere thank you goes to my advisor, *David Gamarnik*. Around this time four years ago, I was in a totally different mental state. I had left my previous research group after my masters to work on something more theoretical, but ended up not finding an advisor for quite some time. It was, by and large, the most challenging period of my life. While I was seriously worried about my future, I started taking David's discrete probability class out of pure luck. Things started improving rapidly and interestingly. Eventually, David took me as his student that summer, and has advised me ever since. It is not an exaggeration to say that I do not remember a single meeting with him after which I was not excited and motivated; there is not a single problem he gave me that I didn't find interesting. He helped me discover my passions and foster my curiosity, while providing the most technical support possible. Had I come to MIT a million times, I would have still worked with him every single one of them.

Secondly, I would like to thank my committee members, *John Tsitsiklis* and *Guy Bresler*. John, thank you for sharing your wisdom during my years at MIT. It was a pleasure and an honor to be your TA, and I only wish I had more chances to interact with you. Guy, thank you for the very inspiring conversations, all the chill times in Berkeley, and all of your help in my postdoctoral applications.

Thirdly, I would like to thank *Will Perkins*. Will, when you gave a talk at MIT in May 2019, I would have never thought we would later end up having such a fantastic collaboration. Thank you very much for hosting me (twice) in Chicago, for all of your help in my postdoctoral applications, and above all, for the beautiful questions and for a very fresh research perspective.

Next, I would like to thank *Kadri Özçaldıran* for his continued support, wisdom and unique style since my undergraduate days (even the day I am writing these lines, I had a call with him earlier). Also many thanks to my teachers at *Ankara Fen Lisesi* for providing me with such a solid academic foundation that still lasts to this day. I am very lucky and proud to have graduated from such an amazing school.

The journey, of course, wouldn't be possible without the support of my friends, here and at the other side of the ocean: *Alex, Harun, Kaya, Gürkan, Zied, Anuran, Melih (HSME), Safa, Thomas, Cemil, Orçun, Nusret, Serkan, Ömer, Hajir, Rabih, Cihan, Brice, Sophie, Dana, Enric, Tanay, Min Jae, Suhas, Paxton, Dheeraj, Chenyang, Amine, Houssam, Manon, Miroslav, Lucas, Muhammmad, Nirav, Ganesh, James, Emre* (and many more). *Alex* and *Harun* deserve a special thanks for being

[1] In fact, I don't even think today deserves to be called a 'spring' day.

the voice of reason when I needed it the most. At this point, I concluded the day and continued writing the next day.

It is now May 8th[2], I am at the Widener Library, getting inspired in a beautiful reading room. A splendid place I guess to thank the most important people in my life: my mom, my dad, and my dearest sister. Thank you all for all the effort you put in raising me and making me who I am today. Thank you for being there in the ups as well as in the most turbulent downs. My doctorate is (almost) as much yours as it is mine. Finally, I am now officially the second 'doc' in family!

Last but not the least, to my dear *Ana* (on a finally very beautiful spring day, May 10th). For giving me some of the most unforgettable moments in my life here in Cambridge, Maine, Chicago, and New Orleans. For the amazing food (and adding the weird spelling of the word 'amazing' into my vocabulary). For making me cheer in the most critical moments. Above all, for your love, every moment in the near past (which feels like many many fulfilling years), for present and for our most beautiful future days to come. Είσαι το άλλο μου μισό και η αδερφή ψυχή μου και κάνεις τη ζωή μου πιο όμορφη από ποτέ.

Cambridge, MA
May 10, 2022

[2]The weather is still not worthy of being called a 'spring' one: sunny, but cold and windy.

Contents

List of Figures

Chapter 1

Introduction

A fundamental commonality across many scientific disciplines is the existence of optimization problems involving uncertainty. Such problems are ubiquitous among many fields ranging from computer science, statistics, machine learning and artificial intelligence to biology and social sciences. Furthermore, the modern era of *big data* added yet another aspect to such problems: *high-dimensionality*. This fueled significant research activity at the intersection of uncertainty and high-dimensionality, creating new metafields dubbed as "high-dimensional probability" and "high-dimensional statistics" [119, 282, 283].

This thesis deals with several algorithmic problems arising from the study of random combinatorial structures, from high-dimensional statistical inference tasks and from the study of neural network models. In our quest, we aim at understanding the underlying model through a rather geometric viewpoint. Our contributions are collected under three (nearly) independent parts.

In what follows, we briefly elaborate on the different parts of the thesis. For each part, we introduce the underlying model/problem that we investigate, fundamental questions pertaining to them that we address; and a brief summary of our main contributions.

1.1 Algorithmic Barriers in Random Combinatorial Structures

Our first focus is on the issues surrounding the algorithmic tractability of optimization in certain combinatorial structures involving randomness. Many algorithmic problems arising from the study of these structures (as well as from the high-dimensional inference tasks and machine learning models) exhibit a rather ubiquitous feature, dubbed as a *statistical-to-computational gap*: there often exists a striking gap between the existential guarantee and the best algorithmic guarantee. The existential guarantees are often established by the so-called first/second moment method [13] or by means of information-theoretical arguments [84]. They can, in principle, be attained with unbounded computational power (e.g. in exponential time). The algorithmic guarantee, on the other hand, is for a restricted class of algorithms requiring bounded compu-

tational power, with the class of polynomial-time algorithms being the benchmark. Throughout, we use the terms "polynomial-time algorithm" and "efficient algorithm" interchangeably.

Perhaps the oldest and simplest example of such a problem exhibiting a *statistical-to-computational gap* is the problem of finding a large *clique* in the so-called "dense" Erdös-Rényi random graph, $\mathbb{G}(n, \frac{1}{2})$. (Given a graph $\mathbb{G} = (V, E)$, a clique C is a fully-connected subset $C \subset V$. That is, $C \subset V$ is a clique iff for any distinct $i, j \in C$, $(i, j) \in E$.) This is a simple random graph model erected on n vertices $\{1, 2, \ldots, n\}$ where for any pair $1 \leq i < j \leq n$ of vertices, (i, j) is an edge with probability $\frac{1}{2}$ independently from any other pair. It is a textbook exercise in probabilistic method, see e.g. [13], that the largest clique of this graph is roughly of size $2 \log_2 n$ with high probability (w.h.p.) as $n \to \infty$. On the other hand, the best known polynomial-time algorithm by Karp [184]—a rather trivial greedy algorithm—returns a clique that is half-optimal, namely of size roughly $\log_2 n$. Noting that such a large clique can be found via a brute force search over all subsets of vertices of cardinality $2 \log_2 n$ in quasi-polynomial time, $2^{\Theta(\log^2 n)}$; this highlights a *statistical-to-computational gap* between the existential and the best algorithmic guarantee. (It is worth noting that, here the existential and the algorithmic guarantees are off by a multiplicative factor of 2. For certain other model, in particular for the models we investigate herein, this gap is often much more dramatic.) A problem that remains open to this day is whether one can improve upon the aforementioned greedy algorithm [184].

The list of problems with a *statistical-to-computational gap* grows very rapidly. A partial list of such problems includes certain "non-planted models" such as random constraint satisfaction problems [217, 7, 189, 234, 23], optimization problems over random graphs [136, 77], spin glass models [221, 73, 120, 124] and the largest submatrix problem [134]; as well as certain "planted" models arising in high-dimensional statistical inference tasks, in particular the principle component analysis (PCA) [47, 199, 168, 167, 21], and perhaps most notably the infamous planted clique problem [177, 214, 38].

Unfortunately, there is as yet no analogue of the standard NP-completeness theory for these *average-case* problems with random inputs; current techniques generally fall short of proving the hardness of such problems even under the assumption that $P \neq NP$. However, a notable exception arises when the problem possesses the so-called *random self-reducibility*. Later in Chapter 4, we establish the average-case hardness of the algorithmic problem of exactly computing the partition function of the Sherrington-Kirkpatrick spin glass model under the assumption $P \neq \#P$ by leveraging the random self-reducibility of the partition function. The assumption that $P \neq \#P$ is much weaker than $P \neq NP$ and is widely believed to be true, see e.g. [178].

While there is no solid hardness theory for such average-case problems per se, a very fruitful and active line of research proposed various approaches that serve as *rigorous evidence* of the algorithmic hardness of such problems. These approaches include (but not limited to):

- failure of the Monte Carlo Markov Chain (MCMC) methods [177, 103];

- failure of low-degree polynomial based algorithms and low-degree methods [169, 193, 122, 285, 63];
- sum-of-squares (SoS) lower bounds [168, 167, 240, 38];
- statistical query (SQ) lower bounds [186, 94, 111];
- failure of the approximate message passing (AMP) algorithm [293, 31] [1];
- reductions from the *planted clique* problem (a canonical problem widely believed to be hard on average) [47, 62, 61].

Our focus in this thesis is on yet another approach called the *Overlap Gap Property* (OGP). This approach aims at studying the *intricate geometry* of the problem by leveraging insights from *statistical physics*. We elaborate on it in the following section.

1.1.1 The Overlap Gap Property (OGP)

Discovered implicitly by Mézard, Mora, and Zecchina [217] and Achlioptas and Ricci-Tersenghi [9] (though coined later by Gamarnik and Li in [134]), the OGP approach leverages insights from statistical physics to form a rigorous link between the intricate geometry of the solution space and the formal algorithmic hardness. The OGP is a topological disconnectivity property. To set the stage, consider a canonical combinatorial optimization problem with random input ξ:

$$\min_{\sigma\in\Theta} \mathcal{L}(\sigma,\xi). \tag{1.1}$$

As an example, consider the random number partitioning problem (NPP), where n i.i.d. standard normal numbers stored in vector $\xi \stackrel{d}{=} \mathcal{N}(0, I_n)$ are to be partitioned into two subsets such that the resulting subset-sums are as close as possible. Encoding any such partition into a binary vector $\sigma \in \mathcal{B}_n \triangleq \{-1,1\}^n$, the NPP is indeed an instance of (1.1) with $\Theta = \mathcal{B}_n$ and

$$\mathcal{L}(\sigma,\xi) \triangleq \big|\langle\sigma,\xi\rangle\big|, \qquad \text{where} \qquad \langle\sigma,\xi\rangle = \sum_{1\le j\le n} \sigma_j\xi_j.$$

We will now briefly describe the OGP. Informally, the OGP holds for the "energy level" $\mathcal{E}$ if there exists $0 < \nu_1 < \nu_2$ such that w.h.p. over the random instance ξ, it is the case that for any $\sigma_1, \sigma_2 \in \Theta$ with $\mathcal{L}(\sigma_1,\xi) \le \mathcal{E}$ and $\mathcal{L}(\sigma_2,\xi) \le \mathcal{E}$,

$$\text{distance}(\sigma_1,\sigma_2) \in [0,\nu_1] \cup [\nu_2,\infty),$$

where the distance function is defined by the ambient space Θ.

Namely, any two near-optimal σ_1 and σ_2 are either "close" or "far" from each other; intermediate distances are not allowed. For our purposes, where we optimize

[1] AMP algorithm achieves the information-theoretically optimal performance for various Bayesian inference problems, see e.g. [91, 90].

over the binary cube $\mathcal{B}_n = \{-1, 1\}^n$, it is more convenient to work with the so-called *normalized overlap* which is scale invariant:

$$\mathcal{O}(\sigma_1, \sigma_2) \triangleq \frac{1}{n} \langle \sigma_1, \sigma_2 \rangle \in [-1, 1]. \tag{1.2}$$

Observe that

$$\mathcal{O}(\sigma_1, \sigma_2) = \frac{n - 2d_H(\sigma_1, \sigma_2)}{n}$$

where $d_H(\sigma_1, \sigma_2)$ is the Hamming distance between σ_1 and σ_2. For this reason, a large value of overlap between σ_1 and σ_2 implies they are similar.

We now illustrate OGP in Figure 1-1. The red curve is plot of an objective function, $\mathcal{L}(\sigma, \xi)$. (It is worth noting that while the space Θ of solutions appears on the horizontal axis in Figure 1-1, it need not be (and is not) one-dimensional. This is merely for display purposes.) The space σ of solutions with $\mathcal{L}(\sigma, \xi) \leq \mathcal{E}$ is partitioned into three disjoint clusters such that the diameter of any cluster is at most ν_1, whereas the distance between any two different clusters is at least ν_2, where $\nu_1 < \nu_2$. Consider now any two near-optimal σ_1, σ_2 where $\mathcal{E}$ quantifies 'near-optimality": $\mathcal{L}(\sigma_1, \xi) \leq \mathcal{E}$ and $\mathcal{L}(\sigma_2, \xi) \leq \mathcal{E}$. If they are contained in the same cluster, they are at most ν_1 apart; whereas if they are contained in different clusters, they are at least ν_2 apart. Namely, there exists no near-optimal pair (σ_1, σ_2) with distance$(\sigma_1, \sigma_2) \in [\nu_1, \nu_2]$. This highlights the presence of OGP in the landscape of $\mathcal{L}$.

Here, one should be careful and not confuse the OGP with clustering: OGP is a stronger property than clustering. That is, while OGP does imply clustering; the converse is not necessarily true: clustering does not necessarily imply OGP as illustrated by Figure 1-2. Notice that (a) $c^* = \min_{\sigma \in \Theta} \mathcal{L}(\sigma, \xi)$ and (b) the set of all solutions σ with $\mathcal{L}(\sigma, \xi) \leq c^* + \mu$ are partitioned into two disjoint clusters, where the maximal diameter ν_1 of a cluster is greater than the separation ν_2 between clusters. In this case, the set of all "distance" values achievable by taking pairs of near-optimal σ_1, σ_2 is the connected interval $[0, \nu_1]$: the OGP is absent in this case.

Origins of OGP. The OGP emerged originally in spin glass theory [273]. A precursory link between the OGP and the formal algorithmic hardness was first made in the context of random constraint satisfaction problems (k-SAT) in a series of papers by Achlioptas and Coja-Oghlan [7]; Achlioptas, Coja-Oghlan, and Ricci-Tersenghi [8]; and by Mézard, Mora, and Zecchina [217]. These papers show an intriguing "clustering" property: they establish that a large portion of the set of satisfying assignments is essentially partitioned into "clusters" that are disconnected with respect to the 'natural topology' of the solution space. As the onset of this clustering property coincides roughly with the regime where the known polynomial-time algorithms fail, this property was conjecturally linked with the formal algorithmic hardness. Strictly speaking, these papers do not establish the OGP. However, an inspection of their arguments reveals that they actually do: the normalized overlap between two satisfying assignments takes values in a set $[0, \nu_1] \cup [\nu_2, 1]$ for some $0 < \nu_1 < \nu_2 < 1$. The aforementioned clustering property is then inferred as a consequence of the OGP.

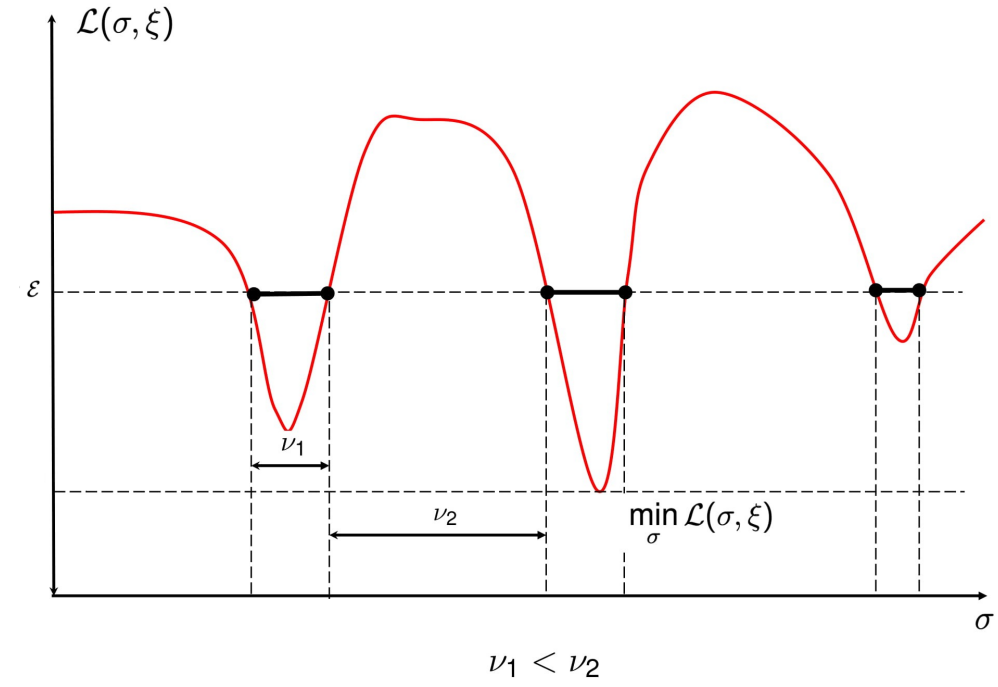

Figure 1-1: OGP for Energy Level $\mathcal{E}$.

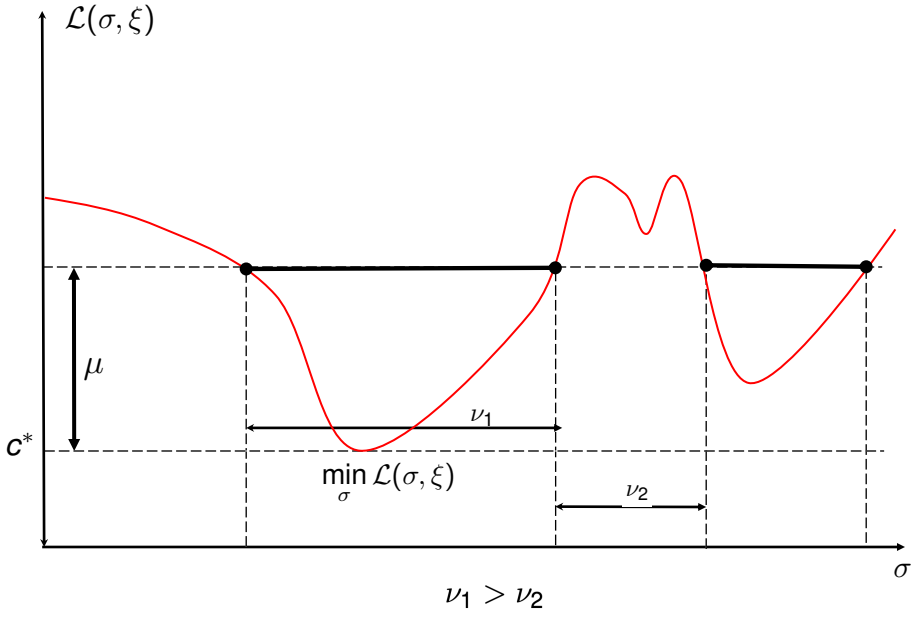

Figure 1-2: Clustering does not imply OGP.

Algorithmic Implications of OGP. The first formal algoritmic implication of the OGP is due to Gamarnik and Sudan [135, 136]. In those papers, theys study the problem of finding a large independent set in sparse random graphs with average degree d. Namely, they study the Erdös-Rényi graph, $\mathbb{G}(n, \frac{d}{n})$ (Their argument also applies to the so-called random $d-$regular graph model, $\mathbb{G}_d(n)$.) It is known, see in particular [117, 118, 45], that the largest independent set of this model is of size $2\frac{\log d}{d}n$ w.h.p., in the double limit as $n \to \infty$ followed by $d \to \infty$. More precisely, let $\mathcal{I}_d$ be the maximal independent set of $\mathbb{G}(n, \frac{d}{n})$ or $\mathbb{G}_d(n)$. Then, w.h.p.

$$\lim_{n\to\infty} \frac{1}{n}|\mathcal{I}_d| = \alpha_d,$$

where the sequence α_d of numbers satisfy $\alpha_d = 2\frac{\log d}{d}(1 + o_d(1))$ as $d \to \infty$. On the other hand, the best known polynomial-time algorithm [184] (a very simple greedy protocol) returns an independent set of cardinality at most $\frac{\log d}{d}n$. In order to reconcile this apparent *statistical-to-computational gap*, Gamarnik and Sudan study the space of all large independent sets. They establish, through a simple first moment argument, that any two independent sets of size greater than $\left(1 + 1/\sqrt{2}\right)\frac{\log d}{d}n$ either have a significant intersection (overlap), or are nearly disjoint. That is, the intermediate values do not occur. Namely they establish that the OGP holds. By leveraging this, they show, through a contradiction argument, that a class of graph algorithms called the *local algorithms/factors of i.i.d.* fails to find an independent set of size greater than $\left(1 + 1/\sqrt{2}\right)\frac{\log d}{d}n$. This refutes an earlier conjecture by Hatami, Lovász, and Szegedy [163]. Subsequent research, again through the lens of OGP, extended this hardness result to the class of *low-degree polynomial* based algorithms [123]. The extra "oversampling" factor, $1/\sqrt{2}$, is an artifact of the overlap analysis. Later research removed this factor by inspecting instead the the overlap pattern of many large independent sets (rather than the pairs) and yielded tight algorithmic hardness guarantees: for any $\epsilon > 0$, there exists an $m \in \mathbb{N}$ such that the set of $m-$tuples of independent sets of size at least $(1 + \epsilon)\frac{\log d}{d}n$ exhibit the multi-OGP; whereas independent sets of cardinality near $\frac{\log d}{d}n$ can be found by means of local algorithms, see e.g. [194]. This was done by Rahman and Virág [244] for *local algorithms*, and by Wein [285] for *low-degree polynomials*. Similar multi-OGP was used by Gamarnik and Sudan in the context of a version of a random constraint satisfaction problem called Not-All-Equal (NAE) random k-SAT problem [137]; and by Bresler and Huang in the context of random k-SAT problem [63]. The approach of looking at the overlap structure between m-tuples of configurations is also at the core of this thesis, and is elaborated further in the next section.

We close this section with a brief list of problems where the OGP is leveraged to rule out certain classes of algorithms. This list includes optimization problems over random graphs and spin glass models [120, 122, 124, 172], random constraint satisfaction problems [137, 63], see also the survey paper by Gamarnik [119] and the references therein.

Multi OGP ($m-$OGP). As we just mentioned, it was previously observed that by considering more intricate overlap patterns, one can potentially lower the (algorithmic) phase transition points further. This idea was employed for the first time by Rahman and Virág [244] in the context of the aforementioned independent set problem. By doing so, they managed to "shave off" the extra $1/\sqrt{2}$ factor present in the earlier result by Gamarnik and Sudan [135, 136]; reaching all the way down to the algorithmic threshold, $\frac{\log d}{d}n$. In a similar vein, Gamarnik and Sudan [137] studied the overlap structure of $m-$tuples $\sigma^{(i)} \in \{0,1\}^n$, $1 \leq i \leq m$ of satisfying assignments in the context of the Not-All-Equal (NAE) $k-$SAT problem. By showing the presence of OGP for $m-$tuples of nearly equidistant points (in $\mathcal{B}_n$), they established near tight hardness guarantees for the class of *sequential local algorithms*: their results match the best computational threshold up to logarithmic-in-k factors.

More recently, $m-$OGP for even more intricate forbidden patterns were considered to establish tight formal algorithmic hardness results in other settings. In particular, by leveraging the $m-$OGP, Wein [285] showed that low-degree polynomials fail to return a large independent set (in sparse random graphs) of size greater than $\frac{\log d}{d}n$, thereby strengthening the earlier result by Gamarnik, Jagannath, and Wein [122]. What is more, Wein's work establishes the *ensemble* variant of OGP (an idea emerged originally in [73]): he considers $m-$tuples of independent sets where each set do not necessarily come from the same random graph, but rather from correlated random graphs. As we elaborate, respectively, in Chapters 2 and 3, the ensemble variant of OGP is used for the random number partitioning problem and the symmetric binary perceptron model considered in this thesis. The ensemble $m-$OGP can be leveraged to rule out *stable algorithms* (appropriately defined); and will also be our focus here. Recently, by leveraging the ensemble $m-$OGP; Bresler and Huang [63] established nearly tight low-degree hardness results for the random $k-$SAT problem: they show that low-degree polynomials fail to return a satisfying assignment when the clause density is only a constant factor off by the computational threshold. In yet another work [172], Huang and Sellke construct a very intricate forbidden structure consisting of an ultrametric tree of solutions, which they refer to as the *branching OGP*. By leveraging this branching OGP, they rule out overlap concentrated algorithms (a class that captures $O(1)$ iteration of gradient descent, approximate message passing; and Langevin Dynamics run for $O(1)$ time) at the algorithmic threshold for the problem of optimizing mixed, even $p-$spin model Hamiltonian.

In this thesis, we carry out the OGP program for two models to give formal evidence of algorithmic hardness and subsequently rule out certain important classes of algorithms.

1.1.2 OGP in Random Number Partitioning Problem

Our first focus is on the random number partitioning problem (`NPP`) mentioned earlier: given n numbers $X = (X_i : 1 \leq i \leq n) \in \mathbb{R}^n$, find a partition $\sigma \in \mathcal{B}_n = \{-1,1\}^n$ such that $\left|\langle \sigma, X \rangle\right|$ is as small as possible. This problem is at the core of a very important application in statistics known as the *design of randomized controlled trials* (which is often considered to be the gold standard for clinical trials [191, 161]) and

has many more other applications, including multiprocessor scheduling, VLSI design, cryptography and so on [76].

The NPP possesses a *statistical-to-computational gap*. When the numbers X_i are i.i.d. standard normal, $X_i \stackrel{d}{=} \mathcal{N}(0,1)$, $1 \le i \le n$;

$$\min_{\sigma \in \mathcal{B}_n} \big|\langle \sigma, X \rangle\big| = \Theta\big(\sqrt{n} 2^{-n}\big),$$

w.h.p. as $n \to \infty$. The best known polynomial-time algorithm, on the other hand, returns a partition $\sigma_{\text{ALG}} \in \mathcal{B}_n$ having an objective of only

$$\big|\langle \sigma_{\text{ALG}}, X \rangle\big| = n^{-\Theta(\log n)}$$

w.h.p. as $n \to \infty$. This highlights a rather striking gap: ignoring the $\sqrt{n}$ factor,

$$2^{-n} \quad \text{vs} \quad 2^{-\Theta(\log^2 n)}.$$

In Chapter 2—which is based on [126]—we initiate the rigorous study of the nature of this gap for NPP. Guided by statistical physics insights, we first conduct a "landscape analysis". Our results are summarized as follows:

- We first study pairs of partitions achieving an objective value of 2^{-E_n}, $E_n = \epsilon n$, and establish the presence of the OGP (for pairs) when $\epsilon > \frac{1}{2}$. We dub this as $2-$OGP.

- Motivated by the fact that the $2-$OGP falls short of explaining the aforementioned *statistical-to-computational gap* all the down to the algorithmic threshold; we next study $m-$tuples of partitions achieving an objective value of 2^{-E_n}, $E_n = \epsilon n$s and establish the presence of multi OGP ($m-$OGP) for any constant $\epsilon > 0$. Importantly, the value of m investigated here remains constant in the natural parameter n of the problem: $m = O(1)$.

- The $m-$OGP approach with $m = O(1)$ falls short of establishing OGP when $E_n = o(n)$. In particular; we show, through a novel and delicate application of the *second moment method*, that the OGP is actually *absent* when $m = O(1)$ and E_n is any arbitrary function having a sub-linear growth in n, $E_n = o(n)$.

- Motivated by the aforementioned investigations demonstrating that the study of $m-$tuples with growing values of m reduces the algorithmic phase transition points further, we finally study $m-$tuples of near-optimal partitions where m itself grows in the natural parameter n of the problem, $m = \omega_n(1)$. In this setting, we establish the presence of $m-$OGP for $m-$tuples of partitions achieving an objective value of 2^{-E_n} for any $E_n = \omega\big(\sqrt{n \log n}\big)$.

The investigation of $m-$tuples where m itself also grows in n is a novel contribution of our work and allows one to establish algorithmic hardness for a much broader range of objective values.

We then turn our attention to the algorithmic front. By leveraging the aforementioned OGP results, we establish following algorithmic hardness results:

- any sufficiently stable algorithm, appropriately defined, fails to find a partition with an objective value of 2^{-E_n}, for any $E_n = \omega(n \log^{-1/5} n)$; and
- a very natural Monte Carlo Markov Chain (MCMC) dynamics fail to return a near-optimal partition.

Our proof of the of the failure of stable algorithm leverages methods from Ramsey Theory from the extremal combinatorics in a crucial way. The application of the Ramsey Theory in this field is very novel.

1.1.3 OGP in Symmetric Binary Perceptron Model

Our second focus is on the symmetric binary perceptron model (SBP). Fix a $\kappa > 0$ and an $\alpha > 0$; set $M = \lfloor n\alpha \rfloor \in \mathbb{N}$; and generate i.i.d. random vectors $X_i \stackrel{d}{=} \mathcal{N}(0, I_n)$, $1 \le i \le M$. ($\lfloor n\alpha \rfloor$ is the largest integer not exceeding $n\alpha$.) Consider the random set

$$S_\alpha(\kappa) \triangleq \bigcap_{1 \le i \le M} \Big\{ \sigma \in \mathcal{B}_n : \big|\langle \sigma, X_i \rangle\big| \le \kappa \sqrt{n} \Big\} = \Big\{ \sigma \in \mathcal{B}_n : \|\mathcal{M}\sigma\|_\infty \le \kappa \sqrt{n} \Big\},$$

where $\mathcal{M} \in \mathbb{R}^{M \times n}$ with rows $X_i \in \mathbb{R}^n$, $1 \le i \le M$. We refer to the matrix $\mathcal{M}$ as the *disorder matrix*. This is a toy neural network model storing random patterns X_i, where the parameter α is called the *storage capacity* [83, 138, 140, 139]. This model also has deep connections to *constraint satisfaction problems*, where (a) the parameter α is akin to the so-called *constraint density*; and (b) each "constraint" $X_i \in \mathbb{R}^n$ rules out certain $\sigma \in \mathcal{B}_n$. For this reason, we often refer to a $\sigma \in S_\alpha(\kappa)$ as a *satisfying assignment*. Our focus is on algorithmic problem of efficiently finding a $\sigma \in S_\alpha(\kappa)$ whenever $S_\alpha(\kappa) \neq \varnothing$ (w.h.p.).

This model exhibits two conundrums detailed bellow.

A *Statistical-to-Computational Gap.* It has been recently established, independently by Perkins and Xu [234] and Abbe, Li, and Sly [6], that for every $\kappa > 0$, there exists a critical value $\alpha_c(\kappa)$ such that for every $\alpha < \alpha_c(\kappa)$, $S_\alpha(\kappa)$ is non-empty w.h.p.; and for every $\alpha > \alpha_c(\kappa)$, it is empty also w.h.p. The value $\alpha_c(\kappa)$ matches the first moment prediction: $\mathbb{E}\big[\big|S_\alpha(\kappa)\big|\big] = o(1)$ if $\alpha > \alpha_c(\kappa)$ and $\mathbb{E}\big[\big|S_\alpha(\kappa)\big|\big] = \omega(1)$ if $\alpha < \alpha_c(\kappa)$. The value of $\alpha_c(\kappa)$ is given by an explicit formula:

$$\alpha_c(\kappa) = -\frac{1}{\log_2 \mathbb{P}\big[|Z| \le \kappa\big]}, \qquad \text{where} \qquad Z \stackrel{d}{=} \mathcal{N}(0,1).$$

Note that in the regime $\kappa \to 0$, $\alpha_c(\kappa)$ behaves roughly like $-\frac{1}{\log_2 \kappa}$. Assuming $\alpha < \alpha_c(\kappa)$ for which $S_\alpha(\kappa) \neq \varnothing$ w.h.p., a natural follow-up algorithmic question is as follows:

"For which values of $\alpha < \alpha_c(\kappa)$, one can find a $\sigma \in S_\alpha(\kappa)$ in polynomial-time?"

This problem is also very related to the much studied *combinatorial discrepancy theory*; certain algorithmic guarantees are available. To the best of our knowledge,

the best polynomial-time algorithmic guarantee is due to Bansal and Spencer [35]. Given any $\kappa > 0$, their algorithm returns, w.h.p., a $\sigma \in S_\alpha(\kappa)$ as long as $\alpha \leq K\kappa^2$ for some explicit constant $K > 0$ independent of κ. Ignoring the absolute constant $K > 0$, the SBP also exhibits a *statistical-to-computational gap* which is much more profound in the regime $\kappa \to 0$:

$$-\frac{1}{\log_2 \kappa} \qquad \text{vs} \qquad \kappa^2.$$

Extreme Clustering. When $S_\alpha(\kappa) \neq \varnothing$ (w.h.p.), yet another natural question is the geometry of the set $S_\alpha(\kappa)$ of satisfying assignments. To that end, the aforementioned papers by Perkins-Xu [234] and Abbe-Li-Sly [6] establish that the SBP exhibits a very intriguing extreme clustering property known as the *Frozen one-step replica symmetry breaking* (Frozen 1−RSB) in statistical physics literature: for any density $\alpha > 0$ below critical value $\alpha_c(\kappa)$, all but an exponentially small fraction of solutions are isolated (w.h.p.) singletons; the Hamming distance to any other solution is linear in the dimension n. In light of [217, 9] linking clustering to algorithmic hardness; this suggests algorithmic intractability. At the same time, however, the model does admit polynomial-time algorithms at sufficiently low densities as was mentioned before. This conundrum challenges the view that extreme clustering leads to algorithmic hardness.

In order to reconcile this apparent conundrum and address the aforementioned *statistical-to-computational gap*, we first study the landscape of this problem through the lens of OGP; and establish following results.

- **High κ regime, $\kappa = O(1)$:** In this regime, we establish the OGP for the pairs and triples of satisfying assignments in $S_\alpha(\kappa)$, dubbed respectively as 2−OGP and 3−OGP. As a running example, let $\kappa = 1$ for which $\alpha_c(\kappa)$ is approximately 1.8159. Our results show that 2−OGP provably takes place when $\alpha \geq 1.71$, whereas 3−OGP provably takes place when $\alpha \geq 1.667$. That is, the OGP threshold is strictly below the existential threshold; and the 3−OGP threshold is strictly below the 2−OGP threshold. Our results indicate an intricate geometry and suggest algorithmic hardness for densities above the OGP threshold.

- **Low κ regime, $\kappa \to 0$.** In this regime, we establish the OGP for m−tuples of satisfying assignments. That is, we study the m−tuples $\sigma_1, \ldots, \sigma_m \in S_\alpha(\kappa)$ of satisfying assignments; establish the presence of m−OGP for densities $\alpha = \Omega(\kappa^2 \log_2 \frac{1}{\kappa})$. Importantly, this guarantee is *nearly tight*: up to the polylogarithmic factor $\log_2 \frac{1}{\kappa}$, it matches the best algorithmic threshold of κ^2.

We then turn our attention to the algorithmic front; and establish the following results.

- Stable algorithms, appropriately defined, fail to find a $\sigma \in S_\alpha(\kappa)$ for densities above the m−OGP threshold, *i.e.* when $\alpha = \Omega(\kappa^2 \log_2 \frac{1}{\kappa})$.

- Online algorithms, appropriately defined, fail to find a $\sigma \in S_\alpha(\kappa)$ for densities sufficiently close to the satisfiability threshold $\alpha_c(\kappa)$.

- The algorithm by Kim and Roche [188] devised for the binary perceptron model is stable in the sense we consider.

The aforementioned results are established in Chapter 3, which is based on [131]. The most technically involved part of this work is establishing the stability of the known algorithms which, unlike in several prior models, do not appear to fall into the class of low-degree polynomials.

Perhaps even more importantly, our work proposes an alternative view on the interplay between the solution space geometry and algorithmic hardness. The existence of polynomial-time algorithms coincides with clustering, therefore the view that clustering implies algorithmic hardness breaks down. On the other hand, the $m-$OGP takes place slightly above the best known algorithmic threshold and the threshold of $m-$OGP nearly matches the algorithmic threshold: the OGP does imply algorithmic hardness. We conjecture that the onset of $m-$OGP coincides with the onset of algorithmic hardness.

1.2 Average-Case Hardness of Sherrington-Kirkpatrick Spin Glass Model

As was mentioned previously, the standard NP-complexity theory tailored for the worst-case hardness is often inadequate for establishing the hardness of average-case problems involving random inputs. A notable exception to this though is when the problem possesses the so-called *random self-reducibility*. Informally, a function f is called *randomly self-reducible* if for any *instance* x, the evaluation of f at x can be reduced, in polynomial-time, to evaluation of f at several random instances $y_1, \ldots, y_k$ [110]. Consequently, if f is randomly self-reducible, then the average-case complexity of f is the same, to within polynomial factors, as its (randomized) worst-case complexity. In this case, one can still leverage the standard NP-complexity theory to establish average-case complexity of f.

Our next focus in this thesis is on the average-case complexity of a certain algorithmic problem surrounding the Sherrington-Kirkpatrick *spin glass* model, SK model in short. This model was introduced in [255] to propose a "solvable" model for the spin glass phase, an unusual magnetic behavior predicted to occur in spatially random physical systems. We now provide more details on the model; see Chapter 4 for a much more elaborate treatment.

Model. Consider $n \in \mathbb{N}$ "sites" $i \in [n] \triangleq \{1, 2, \ldots, n\}$ in space each of which is equipped with a spin $\sigma_i \in \{-1, 1\}$. Intuitively, each site is occupied by a "tiny magnet". Next, let $\boldsymbol{J} = \big(J_{ij} : 1 \leq i < j \leq n\big) \in \mathbb{R}^{n(n-1)/2}$ be a set of parameters called the *couplings*; and let $\boldsymbol{A} = \big(A_i : 1 \leq i \leq n\big)$ be another set of parameters called the *external field*. For any $1 \leq i < j \leq n$, J_{ij} quantifies the strength of the interaction between spins σ_i and σ_j. Moreover, for any $1 \leq i \leq n$, A_i quantifies the strength of the external field at site i. Equipped with these parameters, the "energy" of any *spin configuration* $\boldsymbol{\sigma} = \big(\sigma_i : 1 \leq i \leq n\big) \in \{-1, 1\}^n$ is given by the *Hamiltonian*

$$H\big(\boldsymbol{\sigma}\big) \triangleq \frac{\beta}{\sqrt{n}} \sum_{1 \leq i < j \leq n} J_{ij}\sigma_i\sigma_j + \sum_{1 \leq i \leq n} A_i\sigma_i, \tag{1.3}$$

where the parameter β is called the *inverse temperature*. The SK model corresponds to the case where J_{ij}, $1 \leq i < j \leq n$ and A_i, $1 \leq i \leq n$, are i.i.d. standard normal; and is a mean-field model: the interaction between sites i, j do not depend on their spatial location.

Having defined the Hamiltonian, one can consider the (random) Gibbs measure on $\mathcal{B}_n$ which assigns, to each $\boldsymbol{\sigma} \in \mathcal{B}_n$, a probability mass of $\exp\big(-H(\boldsymbol{\sigma})\big)/Z$ where the parameter Z is the so-called *partition function* ensuring proper normalization:

$$Z\big(\boldsymbol{J}, \boldsymbol{A}, \beta\big) = \sum_{\sigma \in \mathcal{B}_n} \exp\big(-H(\boldsymbol{\sigma})\big). \tag{1.4}$$

The partition function $Z\big(\boldsymbol{J}, \boldsymbol{A}, \beta\big)$ carries an enormous amount of information about the underlying physical system. Moreover, it is also of great relevance in certain Bayesian inference tasks, see e.g. [31, 132] for a more elaborate discussion. That is, the problem of computing the partition function is of natural interest.

In Chapter 4—which is based on [132]—we study the algorithmic problem of *exactly computing* $Z\big(\boldsymbol{J}, \boldsymbol{A}, \beta\big)$. Building upon a novel recursion allowing us to express the partition function of an $n-$spin system as a weighted sum of two $(n-1)$-spin systems with adjusted parameters and leveraging the random self-reducibility of partition function, we establish the average-case hardness of the aforementioned algorithmic problem under the assumption that $P \neq \#P$. The assumption $P \neq \#P$ is milder than $P \neq NP$, and is widely believed to be true. The recursion mentioned above, in some sense, is analogous to the Laplace expansion for the determinant of a matrix.

To the best of our knowledge, this is the first statistical physics model with random parameters for which such an average-case hardness result is established. In the context of statistical inference problems, this result has certain implications: exactly computing the posterior distributions of certain learning models is hard in computational complexity sense.

1.3 Issues in Neural Network Models: Self-Regularity, Benign Landscape, and Provable Learning

Our final focus is on the *neural network* (NN) models which are at the forefront of modern machine learning methods. These models mark a new era in modern machine learning. They have been shown to be extremely powerful in certain tasks that once appeared impossible: natural language processing [79], image recognition [165], image classification [192], speech recognition [219], and even playing the game Go [257], just to name just a few examples. Despite such breakthroughs nearly trespassing the border between fiction and reality, a mathematical understanding of these models is still somewhat lacking. This has fueled significant research efforts.

Setup. Our focus is on two-layer NN architectures, also known as *shallow neural networks*. A two-layer NN $(a, W) \in \mathbb{R}^m \times \mathbb{R}^{m\times d}$ with m hidden units (neurons) computes, for each $X \in \mathbb{R}^d$,

$$\sum_{1\le j\le m} a_j \sigma\big(w_j^T X\big). \tag{1.5}$$

Here, d is the dimension of data $X \in \mathbb{R}^d$; $\sigma(\cdot) : \mathbb{R} \to \mathbb{R}$ is the activation function applied at each neuron; $w_j \in \mathbb{R}^d$ is the jth row of W carrying the weights of jth neuron; and $a = \big(a_j : 1 \le j \le m\big) \in \mathbb{R}^m$ is the vector carrying output weights. The neuron j computes a weighted sum of the coordinates of input X (where the weights are dictated by vector w_j) and passes them to non-linearity $\sigma(\cdot)$. The resulting numbers are then aggregated over all hidden units, through the weights a_j, to obtain a single output.

Next, let $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $1 \le i \le N$, be an i.i.d. sequence of training data drawn from a (unknown) distribution $\mathcal{D}$ on $\mathbb{R}^d \times \mathbb{R}$. Here, $Y_i \in \mathbb{R}$ are often referred to as *labels*. For any $(a, W) \in \mathbb{R}^m \times \mathbb{R}^{m\times d}$, let

$$\widehat{\mathcal{L}}(a, W) \triangleq \frac{1}{N} \sum_{1\le i\le N} \left(Y_i - \sum_{1\le j\le m} a_j \sigma\big(w_j^T X_i\big) \right)^2 \tag{1.6}$$

be the associated *training error*, also known as the *empirical risk*. In what follows, we use the terms 'training error' and 'empirical risk' interchangeably.

We next describe the canonical "learning" problem. Given a set (X_i, Y_i), $1 \le i \le N$, of training data, solve the so-called *empirical risk minimization* (ERM) problem

$$\min_{(a,W)\in\mathbb{R}^m\times\mathbb{R}^{m\times d}} \widehat{\mathcal{L}}(a, W) \tag{1.7}$$

to find a two-layer NN (a, W) with small training error $\widehat{\mathcal{L}}(a, W)$. That is, this "learned" network is desired to explain the unknown relationship between input/label pairs (X_i, Y_i), $1 \le i \le N$, as accurately as possible, where the accuracy is quantified by (1.6). To solve the ERM (1.7) one can run his favourite *training algorithm*: gradient descent (GD), stochastic gradient descent (SGD), mirror descent, etc. One then

uses the learned network, (a, W), to predict unseen data. Here, prediction accuracy is quantified by the so-called *generalization error*, also known as *population risk*:

$$\mathcal{L}(a, W) \triangleq \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\left(Y - \sum_{1\leq j\leq m} a_j \sigma\big(w_j^T X\big)\right)^2\right]. \tag{1.8}$$

Here, the expectation is taken with respect to a *fresh sample* (X, Y) drawn independently from the same distribution $\mathcal{D}$ generating training data (X_i, Y_i), $1 \leq i \leq N$.

Having described the learning setting, we now mention the main issues regarding neural network models.

The Interplay between Overparameterization and Generalization. The conventional wisdom in classical statistics dictates that the overparameterized models, namely the models with more parameters than necessary, tend to overfit to the training data and thus suffer from a poor generalization performance. Nevertheless, the empirical work by Zhang et al. [294] demonstrate the exact opposite effect for the neural networks. They show that neural network models tend to not suffer from this complication: good generalization is retained despite the presence of overparameterization. A predominant explanation for this phenomenon is that while there are many parameter choices near perfectly fitting the training data; the algorithms used in training (such as the GD and the SGD) prefer "simple" solutions regularized according to some additional criteria, such as "small norm". This additional low-complexity is naturally linked to the observed good generalization ability. A drawback of this line of research, however, is that it is algorithm-dependent: one analyzes the end results of the implementation of, say, GD and SGD.

Training and the Optimization Landscape. As we already mentioned, the problem of "learning" (the weights of) aforementioned `NN` models entails solving the ERM (1.7). This is a high-dimensional optimization problem where the underlying landscape is generally highly non-convex. Hence, the underlying learning problem is potentially difficult. Defying this intuition, it has been observed empirically and established rigorously in certain restricted settings (see below) that the GD/SGD, despite being a simple, local, and first-order procedure, is rather successful in training such networks. Understanding why and to what extent this is the case is an ongoing challenge, and another focus of this thesis.

1.3.1 Self-Regularity of Non-Negative Output Weights for Overparameterized Two-Layer Neural Networks

Our next focus in Chapter 5 is on the two-layer `NN` models (1.5) with sigmoid, rectified linear unit (ReLU) and binary step activation functions under the assumption that the output weights are non-negative: $a_j \geq 0$, $1 \leq j \leq m$. This assumption is employed extensively in the theoretical study of such models [143, 93, 204, 98, 250, 295, 150].

Furthermore, it is well-motivated from a practical point of view and allows interpretability. For instance, the audio data and muscular activity data are inherently non-negative, see [1, 259].

Under the aforementioned setting, we consider the problem of finding such a two-layer NN that "fits" a training data set $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $1 \leq i \leq N$, as accurately as possible as quantified by the training error (1.6) and consider the following question:

"To what extent a small training error itself places a restriction on the weights of the learned network?"

We take an algorithm-independent route and investigate this question under the aforementioned non-negativity assumption. Under very mild assumptions on the data distribution $\mathcal{D}$, we establish following results. (All guarantees are w.h.p. with respect to the training data.)

- **Self-Regularity:** Let (a, W) be **any** two-layer NN with small training error $\widehat{\mathcal{L}}(a, W)$ (1.6). Then, $\|a\|_1$ is necessarily well-controlled: $\|a\|_1 = O(1)$.
- **Generalization Guarantees:** **Any** such two-layer NN (a, W) with small $\widehat{\mathcal{L}}(a, W)$ has small generalization error: $\mathcal{L}(a, W)$ appearing in (1.8) is small.

The constants hidden under $O(1)$ depend only on the training error $\widehat{\mathcal{L}}(a, W)$, $\mathbb{E}[|Y|]$ and certain moments of (coordinates of) data X. Several pertinent remarks are now in order.

Notably, our results (a) require a polynomial (in dimension d) sample complexity and are near linear, $\Theta(d \log d)$, for the important cases of ReLU and step activations; (b) are independent of the number of hidden units (which can potentially be very large); (c) are oblivious to the training algorithm; and (d) require very mild assumptions on the data. In particular the input vector $X \in \mathbb{R}^d$ need not have independent coordinates, and the labels Y are only assumed to have bounded first moment. Moreover, our proofs are rather elementary, and based on a covering number argument. Our generalization guarantees are established through the so-called fat-shattering dimension, a scale-sensitive measure of the complexity class that the network architecture being investigated belongs to. Notably, our generalization bounds also have good sample complexity (polynomials in d with a low degree), and are in fact again near-linear for some important cases of interest.

Our work establishes low-complexity for the trained network in an algorithm-independent manner even under the presence of overparameterization; and in particular resonates with the first research theme mentioned above.

1.3.2 Provably Learning Two-Layer Neural Networks with Quadratic Activation Function

Our final focus in Chapter 6—which is based on [127]—is on the problem of learning two-layer NNs with quadratic activation functions $\sigma(x) = x^2$ under the assumption that the labels are generated by a so-called *teacher network* with planted weights

$W^* \in \mathbb{R}^{m\times d}$. More concretely, let $W^* \in \mathbb{R}^{m\times d}$ be a matrix with rows $W_1^*, \dots, W_m^* \in \mathbb{R}^d$ and $X_i \in \mathbb{R}^d$, $1 \le i \le N$ be i.i.d. data. The labels $Y_i \in \mathbb{R}$ are obtained by pushing the inputs $X_i \in \mathbb{R}^d$ into a teacher `NN` with planted weights W^* and quadratic activation function:

$$Y_i = \sum_{1\le j\le m} \big\langle W_j^*, X\big\rangle^2 = \|W^* X_i\|_2^2, \qquad 1 \le i \le N.$$

Here, $\langle \cdot, \cdot \rangle$ is the usual Euclidean inner product on $\mathbb{R}^d$. We refer to such networks as *quadratic networks* for convenience.

This model, admittedly, is rather stylized. Nevertheless, such quadratic networks have been studied extensively in literature [98, 264, 251, 4]. Moreover, blocks of quadratic networks can be stacked together to approximate more practical deeper architectures with sigmoid activation [210]. In addition, quadratic activation function serves as a second order approximation to general non-linearities [279]. Namely, we study quadratic networks to gain insights on deeper and more practical architectures.

We investigate the aforementioned model under the assumption that $\mathrm{rank}(W^*) = d$ (hence in particular $m \ge d$) and that the data $X_i \in \mathbb{R}^d$ consists of i.i.d. sub-Gaussian coordinates[2]: there exists constants $c_1, c_2 > 0$ such that for every $t > 0$, $1 \le i \le N$ and $1 \le j \le d$, $\mathbb{P}[|X_i(j)| > t] \le c_1 \exp\big(-c_2 t^2\big)$.

Our first focus is on the landscape of empirical risk (1.6), where we drop a appearing in (1.6) as $a_j = 1$ for $1 \le j \le m$. We establish following results (Once again, all guarantees are w.h.p. with respect to training data X_i, $1 \le i \le N$.)

- **Benign Landscape:** There is an 'energy barrier' $E_0 > 0$ such that if

  $$\widehat{\mathcal{L}}(W) < E_0 \qquad \text{and} \qquad \nabla_W \widehat{\mathcal{L}}(W) = 0$$

  then $\widehat{\mathcal{L}}(W) = 0$. Namely, below the barrier E_0, the landscape of $\widehat{\mathcal{L}}(\cdot)$ is benign: any stationary point of $\widehat{\mathcal{L}}(\cdot)$ is necessarily a global optimum.

- **Convergence of Gradient Descent:** Initialize the GD with $W_0 \in \mathbb{R}^{m\times d}$ such that $\widehat{\mathcal{L}}(W_0) < E_0$. Then in time $t = \mathrm{poly}(d, \epsilon^{-1})$, it finds a $W_t \in \mathbb{R}^{m\times d}$ such that

  $$\big\|\nabla_W \widehat{\mathcal{L}}(W_t)\big\|_F \le \epsilon \qquad \text{and} \qquad \big\|W_t^T W_t - (W^*)^T W^*\big\|_F \le \epsilon.$$

  Namely, the gradient descent finds, in polynomial-time, an approximate stationary point $W_t \in \mathbb{R}^{m\times d}$ and nearly recovers the planted weights W^*.

The constant E_0 is explicit; and depends only on the smallest singular value $\sigma_{\min}(W^*)$ of W^* and the (conditional) moments of the coordinates of data X_i.

Having established the convergence of GD when initialized below the aforementioned energy barrier E_0; a natural-follow up question is:

"How to initialize properly?"

[2]It is worth noting that this assumption can sometimes be relaxed, as we elaborate later.

We tackle this question under the assumption that the matrix $W^* \in \mathbb{R}^{m\times d}$ of planted weights consists of i.i.d. entries with zero mean and finite fourth moment. Such networks with random weights have been previously considered in literature in the context of random feature methods, see seminal papers by Rahimi and Recht [241, 242].

To that end, we establish the following result.

- **Deterministic Initialization:** Let $C > 0$ be a large enough constant. Then for $W_0 \in \mathbb{R}^{m\times d}$ with $W_0^T W_0 = mI_d$, $\widehat{\mathcal{L}}(W_0) < E_0$ with high probability, provided that $m > Cd^2$.

Here, I_d is the $d \times d$ identity matrix. Namely, provided that the network is sufficiently overparameterized, a deterministic initialization suffices. This result is based on a semicircle law by Bai and Yin [25, 24] which is a novel application of the random matrix theory in the rigorous study of `NN` models.

Our final focus is on the generalization aspect for such networks. Having found a $W \in \mathbb{R}^{m\times d}$ with small training error $\widehat{\mathcal{L}}(W)$; it is by no means clear whether W has a good generalization error $\mathcal{L}(W)$. This naturally prompts the following question:

"What is the smallest number of samples required to claim that a small empirical risk also implies a small generalization error?"

We answer this question by identifying a necessary and sufficient condition on data under which any minimizer of the empirical risk (which, in the case of planted weights, necessarily interpolates the data and achieves zero training error) has zero generalization error. Let $X_i \in \mathbb{R}^d$, $1 \le i \le N$, be the data which is not necessarily random. Define

$$\mathcal{S} \triangleq \big\{A \in \mathbb{R}^{d\times d} : A^T = A\big\};$$

and for $W \in \mathbb{R}^{m\times d}$, let

$$f(W; X) \triangleq \sum_{1\le j\le m} \langle W_j, X\rangle^2 = \|WX\|_2^2.$$

We then establish the following results.

- **Sufficiency:** Suppose that $\mathrm{span}(X_i X_i^T : 1 \le i \le N) = \mathcal{S}$, and $\widehat{m} \in \mathbb{N}$ is arbitrary. Then, for any $W \in \mathbb{R}^{\widehat{m}\times d}$ with $f(W; X_i) = f(W^*; X_i)$, $1 \le i \le N$, $W^T W = (W^*)^T W^*$. Then, $f(W; x) = f(W^*; x)$ for any $x \in \mathbb{R}$. In particular, W generalizes well: $\mathcal{L}(W) = 0$.

- **Necessity:** Suppose $\mathrm{span}(X_i X_i^T : 1 \le i \le N) \subsetneq \mathcal{S}$. Then for any $\widehat{m} \in \mathbb{N}$, there exists a $W \in \mathbb{R}^{\widehat{m}\times d}$ such that while $f(W; X_i) = f(W^*; X_i)$ for every i, $W^T W \ne (W^*)^T W^*$. In particular, $\mathcal{L}(W) > 0$, where $\mathcal{L}$ is defined with respect to any jointly continuous distribution on $\mathbb{R}^d$.

- **Random Data:** Let $X_i \in \mathbb{R}^d$, $1 \le i \le N$ be i.i.d. random vectors drawn from any arbitrary jointly continuous distribution on $\mathbb{R}^d$. Then,

$$\mathbb{P}\big[\mathrm{span}\big(X_i X_i^T : 1 \le i \le N\big) = \mathcal{S}\big] = 1$$

as soon as $N \geq d(d+1)/2$.

Note that the geometric condition, $\text{span}\big(X_i X_i^T : 1 \leq i \leq N\big) = \mathcal{S}$, that we identify is not retrospective in manner: it can be checked before solving the ERM problem (1.7). Moreover, the interpolating NN can be potentially overparameterized. Finally, randomly generated data X_i, $1 \leq i \leq N$, enjoys this condition almost surely as soon as $N \geq d(d+1)/2$ under very mild distributional assumptions.

Our work provides a very complete picture (for the optimization landscape, training, initialization, generalization, as well as the sample complexity) for two-layer NN models with quadratic activation functions and planted weights.

1.4 Structure of the Thesis and Bibliographical Remarks

Each chapter of this thesis is based on a paper by the author and studies one of the aforementioned problems. We briefly highlight the content of each chapter as follows.

- Chapter 2 establishes the Overlap Gap Property (OGP) in the random number partitioning problem and leverages the OGP to establish algorithmic hardness for the class of stable algorithms and Monte Carlo Markov Chain methods. It is based on preprint [126] which is currently under submission.

- Chapter 3 establishes the OGP in symmetric binary perceptron model and leverages the OGP to establish algorithmic hardness for the class of stable algorithms. It is based on the preprint [131] which is currently under submission.

- Chapter 4 establishes the average-case hardness of the algorithmic problem of exactly computing the partition function of the Sherrington-Kirkpatrick spin glass model. It is based on [132] which appeared in *The Annals of Applied Probability* and has also been presented, in part, at the *2020 IEEE International Symposium on Information Theory (ISIT).*

- Chapter 5 establishes self-regularity for two-layer neural networks under non-negativity assumption on its output weights; and leverages the self-regularity to yield good generalization guarantees. It is based on [130] which appeared in *IEEE Transactions on Signal Processing* and has also been presented, in part, at the *2021 IEEE International Symposium on Information Theory (ISIT)* [129].

- Chapter 6 studies the problem of learning two-layer neural networks with quadratic activation function and planted weights; and establishes a fairly complete picture for that problem by addressing many different aspects, including the optimization landscape, convergence of gradient descent, initialization, generalization, as well as the sample complexity. It is based on preprint [127] which is currently under submission.

We commence each chapter with a rather detailed introduction, describe the corresponding model/problem we investigate; and provide a review of prior work and relevant literature. We then present our main contributions. We conclude each chapter with the complete proofs of all of our results.

During the course of his PhD, the author worked on several other problems resulting in following papers that have not been included in this thesis [125, 128, 107].

Chapter 2

Algorithmic Obstructions in the Random Number Partitioning Problem

2.1 Introduction

In this chapter, we study the number partitioning problem (NPP): given n "items" with associated weights, partition them into two "bins", A and B, such that the subset sums corresponding to A and B are as close as possible. More formally, given n numbers $X_i \in \mathbb{R}$, $1 \leq i \leq n$; find a subset $A \subset [n] \triangleq \{1, 2, \ldots, n\}$ such that the discrepancy $\mathcal{D}(A) \triangleq \left|\sum_{i \in A} X_i - \sum_{i \in A^c} X_i\right|$ is minimized. Encoding the membership $X_i \in A$ as a $+1$ and $X_i \in B$ as a -1; NPP can equivalently be posed as a combinatorial optimization problem over the binary cube $\mathcal{B}_n \triangleq \{-1, 1\}^n$:

$$\min_{\sigma \in \mathcal{B}_n} \left| \sum_{1 \leq i \leq n} \sigma_i X_i \right|. \tag{2.1}$$

Our focus is on the algorithmic problem of solving the minimization problem (2.1) "approximately" and "efficiently" (in polynomial time) when the numbers $X_i \in \mathbb{R}$, $1 \leq i \leq n$, are i.i.d. standard normal. We refer to $\mathbf{X} = (X_i : 1 \leq i \leq n) \in \mathbb{R}^n$ as an *instance* of the NPP. Motivated by connections with statistical physics, we refer to $\sigma \in \mathcal{B}_n$ as a *spin configuration*; and to any approximate minimum σ of the problem (2.1) as a *near ground-state*. In the sequel, we slightly abuse the terminology; and use the word "discrepancy" to refer to the optimal value of the combinatorial optimization problem NPP (2.1) and its high-dimensional variant (2.2) (see below); as well as to the value achieved by any partition.

NPP is a special case of the *vector balancing problem* (VBP), where the goal is to minimize the discrepancy

$$\mathcal{D}_n \triangleq \min_{\sigma \in \mathcal{B}_n} \left\| \sum_{1 \leq i \leq n} \sigma_i X_i \right\|_\infty \tag{2.2}$$

of a collection $X_i \in \mathbb{R}^d$, $1 \le i \le n$, of vectors. This problem is at the heart of a very important application in statistics, dubbed as *randomized controlled trials*, which is often considered to be the gold standard for clinical trials [191, 161]. Consider n individuals participating in a randomized study that seeks inference for an additive treatment effect. Each individual i, $1 \le i \le n$, is associated with a set of covariate information $X_i \in \mathbb{R}^d$, a vector carrying the statistics relevant to them such as their age, weight, height, and so on. The individuals are divided into two groups, the treatment group (denoted by a $+$) and the control group (denoted by a $-$). Each group is then subject to a different condition; and a response is evaluated. Based on this response, one seeks to infer the effect of the treatment. To ensure accurate inference based on the response, it is desirable for the groups to have roughly the same covariates. See the very recent work on the design of such randomized controlled experiments by Harshaw, Sävje, Spielman, and Zhang [161] (and the references therein) for a more elaborate discussion on this front.

Besides its significance in statistics, NPP appears in many other practical applications. One such application is the *multiprocessor scheduling*: each item represents the running time of a certain job and each bin represents a group of items that are run on the same processor in a multiprocessor environment [277]. Other practical applications of the NPP include minimizing the size and the delay of VLSI circuits [76, 277], and the so-called Merkle-Hellman cryptosystem [215], one of the earliest public key cryptosystem. For more practical applications of NPP, see the book by Coffman and Lueker [76].

In addition to its important role in statistics and its wide practical applications, NPP is also of great theoretical importance, especially in theoretical computer science, statistical physics, and combinatorial discrepancy theory (see below). NPP is included in the list of *six basic NP-complete problems* by Garey and Johnson [141]; and is the only such problem in this list dealing with numbers. For this reason, it is often used as a basis for establishing the NP-hardness of other problems dealing with numbers, including bin packing, quadratic programming; and the knapsack problem. In statistical physics, NPP is the first system for which the local REM conjecture was established [55, 56]. That is, NPP is the first system which was shown to behave locally like Derrida's random energy model [88, 89], a feature that was conjectured to be universal in random discrete systems [44]. Last but not the least, NPP is one of the first NP-hard problems for which a certain phase transition is established rigorously, which we now discuss. Let X_i, $1 \le i \le n$, be i.i.d. uniform from the set $\{1, 2, \ldots, M\}$ where $M = 2^m$ (namely X_i consists of m-bits). As a function of the control parameter $\kappa \triangleq m/n$ suggested by Gent and Walsh [147], Mertens [216] gave a very elegant yet nonrigorous statistical mechanics argument, for the existence of a phase transition depending on whether $\kappa < 1$ or $\kappa > 1$: the property of finding a perfect partition (that is, a partition with zero discrepancy if $\sum_{1 \le i \le n} X_i$ is even, and that with a discrepancy of one if $\sum_{1 \le i \le n} X_i$ is odd) undergoes as phase transition as κ crosses the value 1. It has been observed empirically that this phase transition is linked with the change of character of typical computational hardness of this problem. Subsequent work by Borgs, Chayes, and Pittel [57] rigorously confirmed the existence of this phase transition. These results further highlight the significance of NPP at the

intersection of computer science, statistical mechanics, and statistics.

As already mentioned, much work has been done on the NPP and its multi-dimensional version, VBP. The prior work visited below can be broadly classified into two categories, namely uncovering the value of the optimal discrepancy; and finding a near ground-state $\sigma \in \mathcal{B}_n$ by means of an efficient algorithm. This was done broadly in for two settings, one where the inputs $X_i \in \mathbb{R}^d$, $1 \leq i \leq n$, are assumed to be *worst-case*; and one where they are assumed to be i.i.d. samples of a distribution, the latter referred to as the *average-case* setting.

We first recap prior results in the *worst-case* setting. A landmark result of discrepancy theory in this setting is due to Spencer [268]. He established that the discrepancy $\mathcal{D}_n$ of VBP per (2.2) is at most $6\sqrt{n}$ if $d = n$ and $\max_{1 \leq i \leq n} \|X_i\|_\infty \leq 1$. Spencer's method, however, is non-constructive. Later research on this front focused on the algorithmic problem of *efficiently* finding a spin configuration $\sigma \in \mathcal{B}_n$ that *approximately* attains a small discrepancy value. These papers are based on techniques including random walks [32, 211], multiplicative weights [201], random weights [246]; and are tight in the regime $d \geq n$: the algorithms return a spin configuration σ with "objective value" $O(\sqrt{n \log(2d/n)})$; and there exist examples whose discrepancy matches this value.

We next visit the *average-case* results. A canonical assumption often considered is that the inputs are i.i.d. standard normal. The first result to this end is due to Karmarkar et al. [183]. They established, using the second moment method, that the objective value of NPP (2.1) is $\Theta\left(\sqrt{n}2^{-n}\right)$ with high probability as $n \to \infty$. Their result remains valid when $X_i \in \mathbb{R}$, $1 \leq i \leq n$ are i.i.d. samples of a distribution that is sufficiently regular. Later research extended this result to the multi-dimensional version, VBP. In particular, in the case where the dimension d is constant, $d = O(1)$, Costello [82] established that the objective value of VBP (2.2) is $\Theta\left(\sqrt{n}2^{-n/d}\right)$ with high probability. When the dimension d is super-linear, in particular $d \geq 2n$, Chandrasekaran and Vempala [71] established that the optimal discrepancy for VBP per (2.2) is essentially $O(\sqrt{n \log(2d/n)})$, ignoring certain polylogarithmic factors. In the regime where $\omega(1) \leq d \leq o(n)$, Turner et al. [278] showed that the optimal discrepancy achieved per (2.2) is $\Theta\left(\sqrt{n}2^{-n/d}\right)$. Moreover, their result transfer also to the case when $X_i \in \mathbb{R}^d$, $1 \leq i \leq n$ consists of i.i.d. coordinates drawn from a density f that is sufficiently regular (in particular, f is square integrable, even; and the coordinates of X_i have a finite fourth moment) and $d = O(n/\log n)$. In addition [278] also studies the regime where $d \leq \delta n$ for a sufficiently small constant δ. For this regime, they establish that the objective value of (2.2) is $O\left(\sqrt{n}2^{-1/\delta}\right)$ with probability at least 99%. This, together with the results of [71] implies that there exists an explicit function $c(\delta)$ such that the discrepancy is $\Theta\left(c(\delta)\sqrt{n}\right)$ with probability at least 99% for $d = \delta n$ and all $\delta > 0$. On the other hand, Aubin, Perkins, and Zdeborová conjectured in [23] that in fact there exists an explicit function $c(\delta)$ such that the discrepancy is $c(\delta)\sqrt{n}$ *with high probability* for the regime $d = \delta n$ and any $\delta > 0$. This conjecture was confirmed very recently, independently by Perkins and Xu [234] and Abbe, Li, and Sly [6].

We now focus on the known algorithmic results. The best known (polynomial-time) algorithm for the NPP is due to Karmarkar and Karp [182] which, for a broad

class of distributions, produces a discrepancy of $O\left(n^{-\alpha \log n}\right)$ with high probability as $n \to \infty$. The original algorithm that they analyzed rigorously is a rather complicated one. Their algorithm, however, is based on a strikingly simple yet a quite elegant, idea; called the *differencing method*, which is based on the following observation. Given a list L of items, placing $x, y \in L$ to the different sides of the partition amounts to removing x and y from L, and adding $|x-y|$ to L instead, an operation that we refer to as *differencing*. Namely, the *differencing* operations applied on $x, y \in L$ returns a new list $L \cup \{|x-y|\} \setminus \{x, y\}$. Using the *differencing*, Karmarkar and Karp proposes two simple (alternative) ways of creating a partition (though they do not rigorously analyze them): the paired differencing method (PDM) and the largest differencing method (LDM). In the former, the items are ordered, and then $\lfloor n/2 \rfloor$ *differencing* operations are performed on the largest and second largest items, on the third and fourth largest items, and so on. The remaining $\lceil n/2 \rceil$ numbers are ordered again, and the aforementioned procedure is repeated until a single item remains, which is the discrepancy achieved by PDM. In LDM, the numbers are again ordered. The *differencing* operation is now applied on the largest and second largest items. The remaining list (now consisting of $n-1$ items) is ordered again, and the procedure is repeated until a single number remains. Recalling that n items can be sorted in near-linear time $O(n \log n)$, the running times of PDM and LDM are indeed polynomial (in n). They conjectured that these two simple natural heuristics also achieve an objective value of $O\left(n^{-\alpha \log n}\right)$ with high probability. For PDM, this conjecture was disproven by Lueker [212] who showed that when the items X_i, $1 \leq i \leq n$, are i.i.d. uniform on $[0,1]$ then the expected discrepancy achieved by the PDM algorithm is rather poor, $\Theta(n^{-1})$. For LDM, however, Yakir [292] confirmed this conjecture, and showed that the expected discrepancy achieved by the LDM is $n^{-\Theta(\log n)}$, when the items X_i are i.i.d. uniform on $[0,1]$. His proof extends to the case when the items X_i follow the exponential distribution, as well. Later, Boettcher and Mertens [52] studied the constant in the exponent, and argued, using non-rigorous calculations, that the expected discrepancy for LDM is $n^{-\alpha \log n}$ for $\alpha = \frac{1}{2 \ln 2} = 0.721\ldots$.

Another algorithm is due to Krieger et al. [191] which achieves an objective value of $O\left(n^{-2}\right)$. It is worth noting that albeit having a poor performance, the algorithm of Krieger et al. finds a *balanced* partition: a spin configuration $\sigma \in \mathcal{B}_n$ with $\sum_{1 \leq i \leq n} \sigma_i \in \{0, 1\}$ depending on the parity of n. This is of practical relevance in the design of randomized trials where the treatment and control groups are often desired to have roughly similar size. Moreover, for the multi-dimensional case $d \geq 2$, they also argue that their algorithm achieves a performance of $O\left(n^{-2/d}\right)$. Finally, Turner et al. [278] devised a generalized version of the Karmarkar-Karp algorithm [182], which returns a partition with discrepancy $2^{-\Theta\left(\log^2 n/d\right)}$ provided the dimension $d \geq 2$ satisfies $d = O\left(\sqrt{\log n}\right)$.

The results summarized above highlight a striking gap between what the existential methods guarantee and what the polynomial-time algorithms achieve. To recap, in the case when X_i, $1 \leq i \leq n$, are i.i.d. standard normal, the optimal discrepancy of the NPP per (2.1) is $\Theta\left(\sqrt{n} 2^{-n}\right)$ with high probability; whereas the-state-of-the-art algorithm (by Karmarkar and Karp) only achieves a performance of $2^{-\Theta(\log^2 n)}$, which

is exponentially worse.

Our Contributions

In this part of the thesis, we the study of the nature of the apparent *statistical-to-computational gap* of the NPP and VBP. Our approach is based on establishing and leveraging the aforementioned *Overlap Gap Property* (OGP). For the sake of presentation clarity, it is convenient to interpret the aforementioned gap in terms of the *"exponent"* E_n of the energy level 2^{-E_n}. Thus the statistical guarantee is $E_n = n$; whereas the best (efficient) computational guarantee available is only when $E_n = \Theta(\log^2 n)$. Our main contributions are now described.

The regime $E_n = \Theta(n)$. In this regime, our first result is as follows. Let $X \in \mathbb{R}^n$ be a random vector with i.i.d. standard normal coordinates. Then for any $\epsilon > 0$, there exist $m \in \mathbb{N}$ and $\beta > \eta > 0$, such that with high probability as n diverges, there does not exist an m-tuple $\left(\sigma^{(i)} : 1 \leq i \leq m\right)$ of spin configurations $\sigma^{(i)} \in \mathcal{B}_n$ such that each $\sigma^{(i)}$ is a near ground-state in the sense $\left|\left\langle \sigma^{(i)}, X \right\rangle\right| = O\left(\sqrt{n} 2^{-n\epsilon}\right)$, $1 \leq i \leq m$; and their pairwise overlaps satisfy $\mathcal{O}\left(\sigma^{(i)}, \sigma^{(j)}\right) \in [\beta - \eta, \beta]$, $1 \leq i < j \leq m$, where the overlap $\mathcal{O}(\cdot, \cdot)$ is defined as

$$\mathcal{O}\left(\sigma_1, \sigma_2\right) \triangleq n^{-1} \left|\langle \sigma_1, \sigma_2 \rangle\right| \in [0, 1]. \tag{2.3}$$

This is the m-OGP and it is the subject of Theorem 2.2.3. We establish Theorem 2.2.3 using the so-called *first moment method*; and the smallest m (for fixed $\epsilon > 0$) for which this result holds true is of order $1/\epsilon$. While we state and prove this result for the NPP (2.1) for simplicity, an inspection of our proof reveals that it extends to the VBP (2.2), when $d = o(n)$.

Note that this geometric result pertains the overlap structure of an m-tuple, rather than a pair, of configurations. This is necessary to cover all values of $\epsilon \in (0, 1]$, since as we show in Theorem 2.2.2, the OGP for pairs (that is for $m = 2$) we can only establish when $\epsilon \in (1/2, 1]$. Furthermore, we conjecture that that the pairwise OGP *does not* hold when $\epsilon < 1/2$, but we are not able to prove this yet. The overlap structure we rule out is essentially the same as the one considered in [137] in the context of random constraint satisfaction problem. Moreover, as we establish in the theorem; this result holds also for a family of correlated random vectors $X_i \in \mathbb{R}^n$, $1 \leq i \leq m$ rather than a single instance. This is known as the "ensemble" variant of the OGP, and it is instrumental in proving the failure of any "sufficiently stable" algorithm.

The regime $E_n = o(n)$. To complement our first result, we investigate the overlap structure when the exponent E_n is sublinear, $E_n = o(n)$. Perhaps rather surprisingly, we establish the *absence* of $m-$OGP—for $m = O(1)$—when $E_n = o(n)$. To that end, let $X \in \mathbb{R}^n$ be a random vector with i.i.d. standard normal coordinates. We establish that for every $E_n \in o(n)$, $m \in \mathbb{N}$, $\rho \in (0, 1)$ and $\bar{\rho} \ll \rho$, it is the case that with high probability there exists an m-tuple $\left(\sigma^{(i)} : 1 \leq i \leq m\right)$ of spin configurations $\sigma^{(i)} \in \mathcal{B}_n$

such that they are near ground-states, namely $\left|\left\langle \sigma^{(i)}, X\right\rangle\right| = O\left(\sqrt{n}2^{-E_n}\right)$, $1 \leq i \leq m$, and their pairwise overlaps satisfy $\mathcal{O}\left(\sigma^{(i)}, \sigma^{(j)}\right) \in [\rho - \bar{\rho}, \rho + \bar{\rho}]$, $1 \leq i < j \leq m$. Namely, the overlaps "span" the interval $[0, 1]$. This is our next result; and is the subject of Theorem 2.2.5.

Theorem 2.2.5 is shown by using the so-called *second moment method* together with a careful overcounting idea. While we state and prove this result for a *single* instance $X \in \mathbb{R}^n$ for simplicity, it is conceivable that our technique extends also to correlated instances $X_i \in \mathbb{R}^n$, $1 \leq i \leq m$ albeit perhaps at the cost of more computational details. We stress once again that this result is shown under the assumption that m is a constant $O(1)$ with respect to n.

Despite the results of Theorem 2.2.3 and Theorem 2.2.5, the aforementioned *statistical-to-computational gap* of NPP still persists. That is, the exponents "ruled out" in Theorem 2.2.3, $E_n = \epsilon n$ for $0 < \epsilon < 1$, are still far greater than the current computational limit, $\Theta\left(\log^2 n\right)$, and furthermore, in the case E_n is sub-linear, $E_n = o(n)$; the $m-$OGP (for $m = O(1)$) is actually *absent* as shown in Theorem 2.2.5.

The rationale for studying the multioverlap version of the OGP ($m-$OGP), noted first by Rahman and Virág [244], was the observation that studying the overlap structures of $m-$tuples (of spin configurations), as opposed to pairs, lowers the "threshold" above which the algorithms can be ruled out. The prior work studying $m-$OGP provided such results when m remains constant with respect to n, $m = O(1)$. This was the case for the NAE-K-SAT problem in [137], and for the maximum independent set problem as in [244] and [285].

However, in our context the $m-$OGP with $m = O(1)$ still falls short of going from $\Theta(n)$ all the way down to current computational threshold, $\Theta\left(\log^2 n\right)$. Thus it is quite natural to ask what happens when m is super-constant, $m = \omega_n(1)$. This is what we do next and to the best of our knowledge this is the first example of a problem where considering m-OGP with super-constant m appears necessary. Specifically, when $m = \omega_n(1)$ we establish the presence of the $m-$OGP for $E_n = \omega\left(\sqrt{n \log n}\right)$. This is the subject of Theorem 2.2.6. At the same time, in Section 2.4.1, we give an informal argument which explains that $\omega\left(\sqrt{n \log n}\right)$ is the best exponent eliminated through this technique that one could hope for. Bridging the gap between $E_n = \omega\left(\sqrt{n \log n}\right)$ and $E_n = \Theta(\log^2 n)$ is another problem we leave open. As for Theorem 2.2.3, the result of Theorem 2.2.6 also pertains to the the case of the "ensemble" variant of the OGP. It is Theorem 2.2.6 which we use to rule out any "sufficiently stable" algorithm, appropriately defined.

Failure of "Stable" Algorithms. Next we focus on the algorithmic questions, where we view an algorithm $\mathcal{A}$ (potentially randomized) as a mapping $\mathcal{A} : \mathbb{R}^n \to \mathcal{B}_n$, which takes an $X \in \mathbb{R}^n$ as its input (numbers/items to be partitioned) and returns a spin configuration $\mathcal{A}(X) \in \mathcal{B}_n$ (from which the partition is inferred). Informally, our main algorithmic result is summarized as follows: the "ensemble" $m-$OGP with $m = \omega(1)$ established in Theorem 2.2.6 is an obstruction to any "sufficiently stable" algorithm. In particular, we establish the following result. Let $\epsilon \in \left(0, \frac{1}{5}\right)$ be arbitrary;

and E_n be an energy exponent with

$$\omega\left(n\log^{-\frac{1}{5}+\epsilon}n\right)\leq E_n\leq o(n).$$

Then, there exists no "sufficiently stable" (in an appropriate sense), and potentially randomized, algorithm $\mathcal{A}$ such that with high probability, $n^{-1/2}\left|\langle X,\mathcal{A}(X)\rangle\right|=2^{-E_n}$. Here, the probability is taken with respect to the randomness in $X\stackrel{d}{=}\mathcal{N}(0,I_n)$, as well as the randomization underlying the algorithm itself. This is the subject of Theorem 2.3.2. It is worth noting that the algorithm $\mathcal{A}$ need not be a polynomial-time algorithm: as long as $\mathcal{A}$ is stable in our sense, there is no restriction on its runtime. It was shown in [123] that stable algorithms include algorithms based on low-degree polynomials, which in their own right include the approximate message passing algorithms. Thus Theorem 2.3.2 applies to these classes of algorithms as well.

It is thus natural to inquire whether the stability property holds for the algorithms known the be successful for the NPP, in particular the LDM algorithm which achieves the state of the art $n^{-\Theta(\log n)}$. We were not able to establish the stability of this algorithm, and instead resorted to simulation study which is reported in Subsection 2.3.2. The simulations are conducted by running the LDM on two correlated instances of the NPP and measuring the overlap of the algorithm results as a function of the correlation. The simulation results suggest that indeed the LDM algorithm is stable in the sense we define. Curiously, it reveals additionally an interesting property. Recall the constant $\alpha=\frac{1}{2\ln 2}=0.721\ldots$ which is suggested heuristically as the leading constant in the performance of the algorithm. We discover a phase transition: when the correlation between two instances is of the order at least approximately $1-n^{-\alpha\log_2 n}$ (in other words the level of "perturbation is order $n^{-\alpha\log_2 n}$), the two outputs of the algorithm are identical or nearly identical. Whereas, when the correlation is smaller than this value, there appears to be a linear discrepancy between the two outcomes. The coincidence of this phase transition with the objective value $n^{-\alpha\log_2 n}$ is remarkable and at this stage we do not have an explanation for it.

Failure of an MCMC Family. A consequence of the 2−OGP established in Theorem 2.2.2 (which holds for energy levels $E_n=\epsilon n$, $\epsilon\in\left(\frac{1}{2},1\right]$) is the presence of a certain property, called a *free energy well* (FEW), in the landscape of the NPP. This property is known to be a rigorous barrier for a family of Markov Chain Monte Carlo (MCMC) methods [21] and has been previously employed for other average-case problems (e.g. [121]) to establish slow mixing of the Markov chain associated with the MCMC method and thus the failure of the method. We establish the presence of a FEW in the landscape of NPP in Theorem 2.3.3; and leverage this property in Theorem 2.3.4 to establish the failure of a very natural class of MCMC dynamics tailored for the NPP. More concretely, Theorem 2.3.3 establishes the presence of the FEW of exponentially small "Gibbs mass" in the landscape of NPP. Theorem 2.3.4 then leverages this property, and shows that for a very natural MCMC dynamics with an appropriate initialization, it takes an exponential time for this chain to reach a region of non-trivial Gibbs mass. See the corresponding section for further details.

Study of Local Optima. Our final focus is on the local optima of this model. For any spin configuration $\sigma \in \mathcal{B}_n$, denote by $\sigma^{(i)} \in \mathcal{B}_n$, $1 \leq i \leq n$, the configuration obtained by flipping the $i-$th bit of σ. A spin configuration σ is called a *local optimum* if $\left|\left\langle \sigma^{(i)}, X \right\rangle\right| \geq |\langle \sigma, X \rangle|$ for $1 \leq i \leq n$. Namely, σ is a local optimum if the "swapping" the place (with respect to σ) of any "item" returns a worse partition. To further complement our landscape analysis in the "hard" regime, $E_n = \Theta(n)$, we study the expectation of number N_ϵ of local optima with energy value $O\left(\sqrt{n}2^{-n\epsilon}\right)$, $0 < \epsilon < 1$. We show that this expectation is exponential in n, and we give a precise, linear, trade-off between the "exponent" of $\mathbb{E}[N_\epsilon]$ and ϵ. This is the subject of Theorem 2.2.8. This suggests that a very simple greedy algorithm, starting from an arbitrary $\sigma \in \mathcal{B}_n$ and proceeding by flipping a single spin so as to "reduce energy" as long as there is such a spin, will likely fail to find a ground-state solution for the NPP.

This analysis is inspired by the work of Addario-Berry et al. [10] who carried out an analogous analysis for the local optima of the Hamiltonian of the Sherrington-Kirkpatrick spin glass model.

Connections With the Perceptron Model and Discussion on Frozen Variables

Our result for the pair-wise OGP in the regime $E_n \geq \epsilon n, \epsilon > 1/2$ described in Theorem 2.2.2 will imply in particular that for every two partitions σ, σ' which achieve this energy level, it is the case that either σ and σ' coincide or σ and σ' are at least $\Theta(n)$ apart. That is, partitions achieving value better than $2^{-\frac{n}{2}}$ are essentially isolated points of $\mathcal{B}_n$ separated by linear Hamming distance. This behavior is very related to the so-called "freezing" phenomenon and the "frozen one step Replica Symmetry Breaking (1-RSB)" picture emerging from the statistical physics regarding the solution space geometry of a very much related perceptron model. Given a near-optimal $\sigma \in \mathcal{B}_n$, a coordinate $i \in [n]$ is called *free* if $\sigma^{(i)} \in \mathcal{B}_n$, the configuration obtained from σ by flipping its i^{th} coordinate, is also near-optimal. If a coordinate i is not free, it is called *frozen*. It was recently established independently by Perkins and Xu [234] and Abbe, Li, and Sly [6] that the symmetric Ising perceptron model (a model that is quite similar to the VBP (2.2) with d proportional to n) exhibits an extreme form of freezing, as conjectured by Huang, Wong, and Kabashima [174]: *typical* solutions (solutions sampled uniformly at random) are completely frozen. That is, *all* coordinates of a *typical* solution are frozen, and every other solution is far from it. Interestingly, our Theorem 2.2.2 implies that for the case $d = 1$, in fact even a stronger property holds: *every* near-optimal solution is isolated with no exceptions (and in particular, one does not need to randomly sample a solution).

Overview of Our Techniques

Presence of the OGP. We establish the presence of the overlap gap property (Theorems 2.2.2, 2.2.3, and 2.2.6) using the so-called first moment method. Specifically, we let a certain random variable count the number of tuples (either pairs, or

$m-$tuples) of near ground-state spin configurations with a prescribed overlap pattern. We then show that expectation of these random variables are exponentially small, establishing the presence of the OGP via Markov inequality. At a technical level, this requires a delicate analysis of a certain covariance structure governing the joint probability.

Absence of the OGP. Recall that in the regime $E_n = o(n)$ we establish in Theorem 2.2.5 that the $m-$OGP (for $m = O(1)$) is absent. This is achieved by letting a certain random variable count the number of configurations of interest, and by leveraging the *second moment method* for this counting object. In addition, the proof requires a novel overcounting idea, in order to "decorrelate" pairs of tuples of spin configurations encountered during the second moment computation. Again at a technical level, the proof also requires a delicate analysis of a block covariance matrix; as well as a probabilistic method argument.

Failure of Stable Algorithms. Our theorem 2.3.2 establishing the failure of stable algorithms (appropriately defined) is perhaps the most technically involved proof; and combines many different ideas, including the $m-$OGP result shown in Theorem 2.2.6 and certain concentration inequalities. Furthermore, interestingly, the proof also uses ideas from the extremal combinatorics and Ramsey Theory, see Theorems 2.6.7 and 2.6.6; and Propositions 2.6.9 and 2.6.12. These results from the Ramsey Theory concern the sizes of the largest graph permitting coloring of graphs without the appearance of monochromatic subgraphs of a given size. We use these results so as to generate a forbidden configuration which when coupled with the stability of algorithms contradict the OGP. In order to guide the reader, we provide in Section 2.6.7 a brief outline of the proof.

Failure of an MCMC Family. The failure of the MCMC algorithm is established using the FEW property. The latter is established using our $2-$OGP result, Theorem 2.2.2, and then utilizing a slightly refined property on the energy landscape of `NPP`, borrowed from [183]. The failure of the MCMC method then relies on a fairly routine arguments, but is included nevertheless in full details for completeness.

Chapter Organization. The rest of the chapter is organized as follows. Our main results regarding the geometry of the energy landscape of `NPP` are found in Section 2.2. Specifically, our result establishing the presence of the $m-$OGP for $m = O(1)$ for energy levels $2^{-\Theta(n)}$ is presented in Section 2.2.1; our result showing the absence of $m-$OGP for energy levels $2^{-o(n)}$ is presented in Section 2.2.2; our result showing the presence of $m-$OGP for energy levels 2^{-E_n} with $\omega\left(\sqrt{n\log_2 n}\right) \leq E_n \leq o(n)$, for $m = \omega_n(1)$ is presented in Section 2.2.3. Our result on the the expected number of local optima is found in Section 2.2.4. The failure of stable algorithms is discussed in Section 2.3.1. The same section contains the simulation results. The limitations of our proofs for establishing the $m-$OGP in the case when m is super-constant, $m = \omega_n(1)$, are studied in Section 2.4.1. We briefly recapitulate our conclusions and

outline several interesting open problems and future research directions in Section 2.5. Finally, the proofs of all of our results are presented in Section 2.6.

Notation. The set of real numbers is denoted by $\mathbb{R}$. The sets $\mathbb{N}$ and $\mathbb{Z}_+$ denote the set of positive integers. For any $N \in \mathbb{N}$, the set $\{1, 2, \ldots, N\}$ is denoted by $[N]$. For two sets A, B; their Cartesian product $\{(a, b) : a \in A, b \in B\}$ is denoted by $A \times B$. For any set A, $|A|$ denotes its cardinality. For any $r \in \mathbb{R}$, the largest integer not exceeding r (that is, the floor of r) is denoted by $\lfloor r \rfloor$; and the smallest integer not less than r (that is, the ceiling of r) is denoted by $\lceil r \rceil$. For any $X = (X_i : 1 \leq i \leq n) \in \mathbb{R}^n$, its Euclidean ℓ_2 norm, $\sqrt{\sum_{1 \leq i \leq n} X_i^2}$ and its Euclidean ℓ_∞ norm, $\max_{1 \leq i \leq n} |X_i|$ are denoted respectively by $\|X\|_2$ and $\|X\|_\infty$. For any $X, Y \in \mathbb{R}^n$, their Euclidean inner product, $\sum_{1 \leq i \leq n} X_i Y_i$, is denoted by $\langle X, Y \rangle$. The symbol $\mathbb{1}\{\mathcal{E}\}$ denotes the indicator of $\mathcal{E}$, which is equal to one if $\mathcal{E}$ is true; and equal to zero if $\mathcal{E}$ is false. $\mathcal{B}_n$ denotes the discrete cube $\{-1, 1\}^n$. For any $\sigma, \sigma' \in \mathcal{B}_n$, their Hamming distance $\sum_{1 \leq i \leq n} \mathbb{1}\{\sigma_i \neq \sigma'_i\}$ is denoted by $d_H(\sigma, \sigma')$, their normalized overlap $n^{-1} |\langle \sigma, \sigma' \rangle|$ is denoted by $\mathcal{O}(\sigma, \sigma')$; and their normalized inner product $n^{-1} \langle \sigma, \sigma' \rangle$ is denoted by $\overline{\mathcal{O}}(\sigma, \sigma')$. log and $\log_2$ denote respectively the logarithms with respect to base e and with respect to base 2. For any $r \in \mathbb{R}$, 2^r is denoted by $\exp_2(\alpha)$; and e^r is denoted by $\exp(r)$. Binary entropy function (that is, the entropy of a Bernoulli random variable with parameter p) is denoted by $h(p) = -p \log_2 p - (1-p) \log_2 (1-p)$. $\mathcal{N}(0, 1)$ denotes the standard normal random variable; and $\mathcal{N}(0, I_n)$ denotes the distribution of a random vector $X = (X_i : 1 \leq i \leq n) \in \mathbb{R}^n$ where $X_i \stackrel{d}{=} \mathcal{N}(0, 1)$, i.i.d. For any matrix $\mathcal{M}$, we denote its Frobenius norm, spectral norm, spectrum, smallest singular value, largest singular value, determinant, and trace by $\|\mathcal{M}\|_F$, $\|\mathcal{M}\|_2$, $\sigma(\mathcal{M})$, $\sigma_{\min}(\mathcal{M})$, $\sigma_{\max}(\mathcal{M})$, $|\mathcal{M}|$, and trace$(\mathcal{M})$, respectively. A graph $\mathbb{G} = (V, E)$ is a collection of vertices V with some edges $(v, v') \in E$ between $v, v' \in E$. In the sequel, we consider only *simple* graphs, that is, graphs that are undirected with no loops. A clique is a complete graph, that is a graph $\mathbb{G} = (V, E)$ where for every distinct $v, v' \in V$; $(v, v') \in E$. The clique on $m-$vertices is denoted by K_m. A subset $S \subset V$ of vertices (of $\mathbb{G}$) is called an independent set if for every distinct $v, v' \in S$; $(v, v') \notin E$. The largest cardinality of such an independent set is called the independence number of $\mathbb{G}$; and is denoted by $\alpha(\mathbb{G})$. A $q-$coloring of a graph $\mathbb{G} = (V, E)$ is a function $\varphi : V \to \{1, 2, \ldots, q\}$ assigning to each edge of $\mathbb{G}$ one of q available colors.

We employ the standard Bachmann-Landau asymptotic notation, e.g. $\Theta(\cdot), O(\cdot), o(\cdot)$, $\omega(\cdot)$, and $\Omega(n)$ throughout the chapter. Whenever a function $f(n)$, say, has growth $o(n)$, we either denote by $f(n) = o(n)$ or $f(n) \in o(n)$. Finally, whenever f has a lower and upper bound on its growth, we abuse the notation slightly and use inequalities. For instance, when $f(n) \in \omega_n(1)$ and $f(n) \in o(n)$ (that is, f is super-constant but sub-linear), we often find it convenient to write $\omega_n(1) \leq f(n) \leq o(n)$.

In order to keep our presentation simple, we omit all floor and ceiling operators.

2.2 Main Results. The Landscape of the NPP

In this section, we present our results regarding the geometry of the energy landscape of the number partitioning problem (NPP).

Our results concern the overlap structures of the tuples of near ground-state configurations, formalized next. Recall the definition of the overlap $\mathcal{O}$ from (2.3).

Definition 2.2.1. *Fix an $m \in \mathbb{N}$, and $0 < \eta < \beta < 1$. Let $X_i \stackrel{d}{=} \mathcal{N}(0, I_n)$, $0 \le i \le m$, be i.i.d. random vectors; and let $\mathcal{I}$ be any subset of $[0,1]$. Denote by $\mathcal{S}(\beta, \eta, m, E_n, \mathcal{I})$ the set of all $m-$tuples $\left(\sigma^{(i)} : 1 \le i \le m\right)$ of spin configurations $\sigma^{(i)} \in \mathcal{B}_n$, such that the following holds:*

(a) **(Pairwise Overlap Condition)** *For any $1 \le i < j \le m$,*

$$\beta - \eta \le \mathcal{O}\left(\sigma^{(i)}, \sigma^{(j)}\right) \le \beta.$$

(b) **(Near Ground-State Condition)** *There exists $\tau_i \in \mathcal{I}$, $1 \le i \le m$, such that*

$$\frac{1}{\sqrt{n}} \left|\left\langle \sigma^{(i)}, Y_i\left(\tau_i\right)\right\rangle\right| \le 2^{-E_n}, \quad \text{where} \quad Y_i(\tau_i) = \sqrt{1-\tau_i^2} X_0 + \tau_i X_i, \quad \text{for} \quad 1 \le i \le m.$$

Here, m refers to size of the tuple we investigate; the quantities β and η control the overlap region; E_n controls the "exponent" of the energy level 2^{-E_n} with respect to which $\sigma^{(i)} \in \mathcal{B}_n$ are near ground-state; and $\mathcal{I}$ is a certain index set for describing the correlated instances (more on this later).

The set $\mathcal{S}(\beta, \eta, m, E_n, \mathcal{I})$ is the set of all $m-$tuples of spin configurations $\sigma^{(i)} \in \mathcal{B}_n$, $1 \le i \le m$; where i) the pairwise overlaps between $\sigma^{(i)}$ lie in the interval $[\beta - \eta, \beta]$; and ii) each $\sigma^{(i)}$, $1 \le i \le m$, is a (near) ground-state with respect to an instance of the NPP dictated by the entries of the vector $Y_i(\tau_i) \in \mathbb{R}^n$. Note that the instances $Y_i\left(\tau_i\right)$ with respect to which $\sigma^{(i)}$ are near-optimal need not be the same; each individually distributed as $\mathcal{N}(0, I_n)$; and are correlated. This will later turn out to be useful in ruling out "sufficiently stable" algorithms, appropriately defined.

2.2.1 Overlap Gap Property for the Energy Levels $2^{-\Theta(n)}$

Our first focus is on the pairs of near ground-state configurations with respect to energy levels $E_n = \epsilon n, \epsilon \in \left(\frac{1}{2}, 1\right]$.

Theorem 2.2.2. *Let $X \stackrel{d}{=} \mathcal{N}(0, I_n)$; and $\epsilon \in \left(\frac{1}{2}, 1\right]$ be arbitrary. There exists $\rho \triangleq \rho(\epsilon) \in (0,1)$ such that with probability $1 - \exp\left(-\Theta(n)\right)$, there are no pairs $(\sigma, \sigma') \in \mathcal{B}_n \times \mathcal{B}_n$ for which $\mathcal{O}\left(\sigma, \sigma'\right) \in \left[\rho, \frac{n-2}{n}\right]$; $\frac{1}{\sqrt{n}} \left|\langle \sigma, X\rangle\right| = O\left(2^{-n\epsilon}\right)$, and $\frac{1}{\sqrt{n}} \left|\langle \sigma', X\rangle\right| = O\left(2^{-n\epsilon}\right)$.*

The proof of Theorem 2.2.2 is based on a first moment argument, and is provided in Section 2.6.2. Several remarks are now in order. Theorem 2.2.2 establishes that for the energy levels $E_n = \epsilon n$ with $\epsilon \in \left(\frac{1}{2}, 1\right]$, it is the case that with high probability, the

overlap of any pair of near ground-state spin configurations exhibits a gap. Namely, in the language of Definition 2.2.1, Theorem 2.2.2 establishes that for any $\epsilon \in \left(\frac{1}{2}, 1\right]$ there exists a $\rho \triangleq \rho(\epsilon) \in (0,1)$ such that the set $\mathcal{S}(\beta, \eta, m, E_n, \mathcal{I})$ with parameters $\beta = 1 - \frac{2}{n}$, $\eta = 1 - \frac{2}{n} - \rho$, $m = 2$, $E_n = 2^{-\epsilon n}$ and $\mathcal{I} = \{0\}$ is empty with probability at least $1 - \exp\left(-\Theta(n)\right)$. Note that the rightmost end of this gap is independent of ϵ; and reaches $\frac{n-2}{n}$: this is the largest overlap that can be attained by two spin configurations σ, σ' with $\sigma \neq \pm\sigma'$. That is, if $1 \leq d_H(\sigma, \sigma') \leq n-1$ then $\mathcal{O}(\sigma, \sigma') \leq \frac{n-2}{n}$ with equality if and only if $d_H(\sigma, \sigma') \in \{1, n-1\}$. In particular, w.h.p. for any $\sigma, \sigma' \in \mathcal{B}_n$ with $n^{-\frac{1}{2}} |\langle \sigma, X \rangle| = O(2^{-n\epsilon})$ and $n^{-\frac{1}{2}} |\langle \sigma', X \rangle| = O(2^{-n\epsilon})$, it is the case that either σ and σ' coincide or are at least $\Theta(n)$ apart, $d_H(\sigma, \sigma') = \Theta(n)$. That is, partitions achieving value better than $2^{-\frac{n}{2}}$ are essentially isolated points of $\mathcal{B}_n$ separated by linear Hamming distance, as we have discussed in the introduction.

We will use our pairwise OGP result, Theorem 2.2.2, to establish the existence of the *free energy well* (FEW), which will be used as a barrier for the Markov chain type algorithms.

The energy levels E_n addressed by Theorem 2.2.2 is $E_n \geq \epsilon n, \epsilon > 1/2$. Our next result concerns all linear size energy values E_n. It shows that the NPP exhibits $m-$*Overlap Gap Property* ($m-$OGP)—for constant $m = O(1)$ for such energy levels.

Theorem 2.2.3. *Let $\epsilon > 0$. There exists an $m \triangleq m(\epsilon) \in \mathbb{N}$; β, and η with $0 < \eta < \beta < 1$ such that the following holds. For i.i.d. random vectors $X_i \in \mathbb{R}^n$, $0 \leq i \leq m$ with distribution $\mathcal{N}(0, I_n)$ and any subset $\mathcal{I} \subset [0,1]$ with $|\mathcal{I}| = 2^{o(n)}$,*

$$\mathbb{P}\Big(\mathcal{S}\left(\beta, \eta, m, \epsilon, \mathcal{I}\right) \neq \varnothing\Big) \leq \exp_2\left(-\Theta\left(n\right)\right).$$

Here $\mathcal{S}(\beta, \eta, m, \epsilon, \mathcal{I})$ stands for $\mathcal{S}(\beta, \eta, m, E_n, \mathcal{I})$ with $E_n = n\epsilon$, which was introduced in the Definition 2.2.1.

The proof of Theorem 2.2.3 is based on a first moment argument; and is provided in Section 2.6.3.

Note that Theorem 2.2.3 pertains the "ensemble" variant of the OGP: the spin configurations $\sigma^{(i)}$, $1 \leq i \leq m$, need not be near ground-states for the same instance of the problem; and are instead near ground-states for potentially correlated instances.

Remark 2.2.4. *It is worth noting that while we state and prove Theorem 2.2.3 for the NPP (2.1) for simplicity; our result still remains valid for the high-dimensional version, VBP (2.2). More concretely, recalling that the optimal value of (2.2) for random i.i.d. standard normal inputs $X_i \stackrel{d}{=} \mathcal{N}(0, I_d)$, $1 \leq i \leq n$ is $\Theta\left(\sqrt{n} 2^{-n/d}\right)$ for $\omega(1) \leq d \leq o(n)$; our approach still remains valid and the $m-$OGP still takes place for energy levels $\Theta\left(\sqrt{n} 2^{-\epsilon n/d}\right)$ for any $\epsilon \in (0,1]$.*

2.2.2 Absence of m-Overlap Gap Property for Energy Levels $2^{-o(n)}$

We now focus our attention to the sub-linear exponent regime, $E_n = o(n)$. We establish that the m-OGP is actually *absent* in this regime, when m is constant.

That is, the overlaps of spin configurations achieving sublinear energy levels "span" the entire interval, in an certain sense concretized as follows.

Theorem 2.2.5. *Let $X \stackrel{d}{=} \mathcal{N}(0, I_n)$. Fix any $\eta > 0$ and $m \in \mathbb{N}$. Suppose that $f(n) : \mathbb{N} \to \mathbb{R}^+$ is any arbitrary function with $f(n) = \omega_n(1)$ and $f(n) = o(n)$. Then,*

$$\lim_{n\to\infty} \mathbb{P}\Big(\forall \beta \in [0,1] : \mathcal{S}\left(m, \beta, \eta, f(n), \{0\}\right) \neq \varnothing\Big) = 1;$$

where the set $\mathcal{S}$ is introduced in Definition 2.2.1 with the following modification on the pairwise overlap condition: $\beta - \eta \leq \mathcal{O}\left(\sigma^{(i)}, \sigma^{(j)}\right) \leq \beta + \eta$ for $1 \leq i < j \leq m$.

In particular, since η in the statement of Theorem 2.2.5 is arbitrary, we conclude that the overlaps indeed "span" the entire interval. The $m-$tuples $\left(\sigma^{(i)} : 1 \leq i \leq m\right)$ that we consider in Theorem 2.2.5 consist of spin configurations that are near ground-state with respect to the same instance of the problem: $n^{-1/2}\left|\left\langle \sigma^{(i)}, X\right\rangle\right| \leq \exp_2(-f(n))$ for $1 \leq i \leq m$. Moreover, our proof will demonstrate something stronger: one can find such $m-$tuples $\sigma^{(i)} \in \mathcal{B}_n$, $1 \leq i \leq m$ satisfying not only the constraints on absolute values of inner products but inner products themselves: $\beta - \eta \leq \frac{1}{n}\left\langle \sigma^{(i)}, \sigma^{(j)}\right\rangle \leq \beta + \eta$. The slight modification of Definition 2.2.1 (where the interval $[\beta - \eta, \beta + \eta]$ is considered instead of $[\beta - \eta, \beta]$) is for convenience.

The proof of Theorem 2.2.5 uses the probabilistic method and the second moment method together with a crucial overcounting idea; and is provided in Section 2.6.4.

2.2.3 m-Overlap Gap Property Above $2^{-\Theta(n)}$: Super-Constant m

We now establish the existence of $m-$OGP, where m is super-constant, $m = \omega_n(1)$, for certain energy levels whose exponents are sub-linear.

Theorem 2.2.6. *Let $E_n : \mathbb{N} \to \mathbb{R}^+$ be any arbitrary "energy exponent" with growth condition*

$$E_n \in \omega\left(\sqrt{n \log_2 n}\right) \quad \text{and} \quad E_n \in o(n).$$

Suppose that $X_i \in \mathbb{R}^n$, $1 \leq i \leq m$, are i.i.d. with distribution $\mathcal{N}(0, I_n)$, and $\mathcal{I} \subset [0,1]$ with $|\mathcal{I}| = n^{O(1)}$. Define the sequences $(m_n)_{n\geq 1}$, $(\beta_n)_{n\geq 1}$ and $(\eta_n)_{n\geq 1}$ with $1 > \beta_n > \eta_n > 0$, $n \geq 1$ by

$$m_n \triangleq \frac{2n}{E_n}, \quad \beta_n \triangleq 1 - \frac{2g(n)}{E_n}, \quad \text{and} \quad \eta_n = \frac{g(n)}{2n}, \tag{2.4}$$

where $g(n)$ is any arbitrary function with growth condition

$$\omega(1) \leq g(n) \leq o\left(\frac{E_n^2}{n \log_2 n}\right). \tag{2.5}$$

Then,

$$\mathbb{P}\Big(\mathcal{S}\left(\beta_n, \eta_n, m_n, E_n, \mathcal{I}\right) \neq \varnothing\Big) \leq \exp\left(-\Theta(n)\right). \tag{2.6}$$

Here, $\mathcal{S}(\beta_n, \eta_n, m_n, E_n, \mathcal{I})$ is the set introduced in Definition 2.2.1 with the modification that the pairwise inner products (as opposed to the overlaps) are constrained, that is

$$\beta - \eta \leq \frac{1}{n}\left\langle \sigma^{(i)}, \sigma^{(j)} \right\rangle \leq \beta, \quad \textit{for} \quad 1 \leq i < j \leq m.$$

Moreover, in the special case when

$$\omega\left(n \cdot \log^{-\frac{1}{5}+\epsilon} n\right) \leq E_n \leq o(n)$$

and $\epsilon \in (0, \frac{1}{5})$ is arbitrary, (2.6) still holds with $g(n)$ satisfying

$$g(n) = n \cdot \left(\frac{E_n}{n}\right)^{2+\frac{\epsilon}{8}}. \tag{2.7}$$

The idea of the proof of Theorem 2.2.6 is quite similar to that of Theorem 2.2.3, yet it does not follow directly from Theorem 2.2.3. This is due to the fact that Theorem 2.2.6 uses different asymptotic bounds for certain cardinality terms; and requires a more careful asymptotic analysis. For this reason, we provide a separate and complete proof in Section 2.6.5.

Several remarks are in order. In what follows, we suppress the subscript n from m_n, β_n, η_n; while the reader should keep in his mind that all these quantities are functions of n. We treat the case $E_n = \omega(n \cdot \log^{-1/5+\epsilon} n)$ with a different choice of $g(n)$ (though keeping m, β, η—as functions of $g(n)$—the same). In particular, in this case the $g(n)$ parameter (hence the η parameter) appearing in Theorem 2.2.6 can be taken to be larger. Note that this yields a stronger conclusion: it implies that the length η of the forbidden region is larger. This will be needed later in Theorem 2.3.2 when we leverage Theorem 2.2.6 for the case $E_n = \omega(n \cdot \log^{-1/5+\epsilon} n)$ with $g(n)$ chosen as in (2.7) (and β, η, m prescribed according to (2.4)) to rule out stable algorithms.

Our next remark pertains to the size of the index set, $\mathcal{I}$. While we restricted our attention to sets with $|\mathcal{I}| \triangleq |\mathcal{I}(n)| = n^{O(1)} = \text{poly}(n)$, it appears that our technique still remains valid, so long as $|\mathcal{I}(n)| \leq 2^{CE_n}$, where C is a small enough constant.

We now comment on the energy exponent, E_n. For Theorem 2.2.6 to hold true, E_n should grow faster than $\omega\left(\sqrt{n \log_2 n}\right)$. While we do not rule out the OGP for smaller values of E_n, we will provide an argument which shows that $\omega\left(\sqrt{n \log_2 n}\right)$ is tight for the methods employed in this chapter. See Section 2.4.1 for more details.

An Illustration of Theorem 2.2.6 with a Concrete Choice of Parameters.

We now illustrate Theorem 2.2.6 with a concrete choice of the energy exponent E_n and concrete choices of parameters m, β, and η.

Fix $\delta \in (0, \frac{1}{2})$ and consider "ruling out" the energy levels $E_n = n^{1-\delta}$. That is, our goal is to establish the presence of the m-OGP for $E_n = n^{1-\delta}$ for appropriate m, β, and η parameters. Next, take $m = 2n/E_n = 2n^{\delta}$, per (2.4). Choose a $\delta' > 0$ such

that $\delta' + 2\delta < 1$. Then, set $g(n) = n^{\delta'}$. It is easily verified that

$$\omega_n(1) \le g(n) \le o\left(\frac{E_n^2}{n \log_2 n}\right) = o\left(\frac{n^{1-2\delta}}{\log_2 n}\right).$$

We then take, again per (2.4),

$$\beta_n = 1 - \frac{2g(n)}{E_n} = 1 - 2n^{\delta'+\delta-1} \quad \text{and} \quad \eta_n = \frac{g(n)}{2n} = \frac{1}{2}n^{-1+\delta'}.$$

Note that the overlap region, $[\beta - \eta, \beta]$, has a length η. For the statement of the theorem to be non-vacuous, n times the overlap length must contain some integer values: $|[n\beta - n\eta, n\beta] \cap \mathbb{Z}| = \Omega(1)$ must hold. We verify that indeed $n\eta = \frac{1}{2}n^{\delta'} = \omega_n(1)$. Namely, n times the length of the overlap interval grows polynomially in n.

2.2.4 Expected Number of Local Optima

In this section, we complement our earlier analysis in the "hard" regime, $2^{-\Theta(n)}$. Specifically, we focus on the local optima at these energy levels.

Definition 2.2.7. *Let $\sigma \in \mathcal{B}_n$ be a spin configuration. For every $1 \le i \le n$, denote by $\sigma^{(i)} \in \mathcal{B}_n$ the spin configuration obtained by flipping $i-$th bit of σ. That is, $\sigma^{(i)}(i) = -\sigma(i)$ and $\sigma^{(i)}(j) = \sigma(j)$ for $j \ne i$. Given $X \in \mathbb{R}^n$, a spin configuration $\sigma \in \mathcal{B}_n$ is called a* **local optimum** *if*

$$\left|\left\langle \sigma^{(i)}, X \right\rangle\right| \ge |\langle \sigma, X \rangle|, \quad \text{for} \quad 1 \le i \le n.$$

For energy exponents E_n of form ϵn, $0 < \epsilon < 1$, we now compute the expected number of local optima below the energy level 2^{-E_n}.

Theorem 2.2.8. *Let $X \overset{d}{=} \mathcal{N}(0, I_n)$. Fix any $\epsilon \in (0,1)$, and let N_ϵ be the number of spin configurations $\sigma \in \mathcal{B}_n$ which satisfies the following:*

(a) σ is a local optimum in the sense of Definition 2.2.7.

(b) $\frac{1}{\sqrt{n}} |\langle \sigma, X \rangle| = O(2^{-n\epsilon})$.

Then, $\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}[N_\epsilon] = 1 - \epsilon$.

The proof of Theorem 2.2.8 is provided in Section 2.6.6. The notion of local optimality per Definition 2.2.7 is the same one that Addario-Berry et al. considered in [10]. In particular, they study the local optima of the Hamiltonian of the Sherrington-Kirkpatrick spin glass; and carry out a very similar analysis—namely they show the expected number of local optima is exponentially large, and compute the exponent (though the proofs corresponding to these models are different).

Theorem 2.2.8 gives a precise trade-off between the exponent of the energy value and the exponent of the expectation: the exponent of the (expected) number of

local optima decays linearly in the exponent ϵ of the energy level, as ϵ varies in $(0,1)$. In particular, the expected number of the local optima is exponential with exponent growing linearly as the energy level moves away from the energy of the ground state. Theorem 2.2.8 suggests the likely failure of a very simple, yet natural, greedy algorithm. Consider a greedy algorithm which starts from a spin configuration $\sigma \in \mathcal{B}_n$ and performs a sequence of local, greedy, moves: at each step, flip a spin configuration that decreases the energy, $|\langle \sigma, X\rangle|$. This greedy algorithm continues until one cannot move any further, therefore reaching a local optimum, in the sense of Definition 2.2.7. Theorem 2.2.8 shows that there exists, in expectation, exponentially many such local optima. This suggests that the greedy algorithm will likely fail to find a ground-state solution.

It is important to note that Theorem 2.2.8 should be viewed only as an *evidence* based on the first moment method. Indeed; while it is true, per Theorem 2.2.8, that the expected number of local minima is exponential (in n), it might still be possible that one in fact has $O(1)$ (or even one) local minimizer with high probability: this is the case, e.g., if there exists a rare event on which there are exponentially many local minimizers (due to the so-called *lottery effect*). A typical way to overcome this caveat is to perform a second moment calculation, which unfortunately appears to be quite involved in our case. We leave this for a future work.

2.3 Main Results. Failure of Algorithms

2.3.1 $m-$Overlap Gap Property Implies Failure of Stable Algorithms

The focus in this section understanding the power of stable algorithms in solving the optimization problem (2.1) when the input X_i, $1 \le i \le n$, consists of i.i.d. standard normal weights.

Algorithmic Setting. We interpret an algorithm $\mathcal{A}$ as a mapping from the Euclidean space $\mathbb{R}^n$ to the binary cube $\mathcal{B}_n \triangleq \{-1,1\}^n$. We also allow $\mathcal{A}$ to be potentially randomized. More concretely, we assume that there exists a probability space $(\Omega, \mathbb{P}_\omega)$, such that $\mathcal{A} : \mathbb{R}^n \times \Omega \to \mathcal{B}_n$ and for every $\omega \in \Omega$, $\mathcal{A}(\cdot, \omega) : \mathbb{R}^n \to \mathcal{B}_n$. Here, $X \in \mathbb{R}^n$ denotes the "items" to be partitioned; whereas for a fixed $\omega \in \Omega$, $\mathcal{A}(X,\omega) \in \mathcal{B}_n$ is the spin configuration returned by this potentially randomized algorithm, $\mathcal{A}$; which encodes a partition.

We now formalize the class of "sufficiently stable" algorithms.

Definition 2.3.1. *Let $E > 0$; $f \in \mathbb{N}$, $L \in \mathbb{R}^+$ and $\rho', p_f, p_{\text{st}} \in [0,1]$. A randomized algorithm $\mathcal{A} : \mathbb{R}^n \times \Omega \to \mathcal{B}_n$ for the* `MPP` *(2.1) is called $(E, f, L, \rho', p_f, p_{\text{st}})-$optimal if the following are satisfied.*

- **(Near-Optimality)** *For $(X,\omega) \sim \mathcal{N}(0, I_n) \otimes \mathbb{P}_\omega$,*

$$\mathbb{P}_{(X,\omega)}\left(\frac{1}{\sqrt{n}}\left|\langle X, \mathcal{A}(X,\omega)\rangle\right| \le E\right) \ge 1 - p_f.$$

- **(Stability)** *For every $\rho \in [\rho', 1]$, it holds that*

$$\mathbb{P}_{(X,Y,\omega):X\sim_\rho Y}\Big(d_H\left(\mathcal{A}(X,\omega),\mathcal{A}(Y,\omega)\right) \le f + L\|X-Y\|_2^2\Big) \ge 1 - p_{\mathrm{st}}.$$

Here, the probability is taken with respect to joint randomness $\mathbb{P}_{X,Y}\otimes\mathbb{P}_\omega$ of (X,Y,ω): $X,Y \overset{d}{=} \mathcal{N}(0,I_n)$ with $\mathrm{Cov}(X,Y) = \rho I_n$ (which together uniquely specify the joint distribution $\mathbb{P}_{X,Y}$ denoted by $X \sim_\rho Y$); and $\omega \sim \mathbb{P}_\omega$, which is the "coin flips" of the algorithm.

In what follows, we will abuse the notation, and refer to $\mathcal{A} : \mathbb{R}^n \to \mathcal{B}_n$ (by suppressing ω) as a *randomized algorithm*. The parameter, E, refers to the cost (i.e., objective value) achieved by the partition returned by $\mathcal{A}$. The parameter, p_f, controls the "failure" probability—the probability that algorithm fails to return a partition with cost below E.

An important feature of Definition 2.3.1 is that the stability guarantee is probabilistic; and the parameters, $f, L, \rho', p_{\mathrm{st}}$, control the stability of the algorithm. Specifically, in order to talk about stability in a probabilistic setting, one has to consider two random input vector that are potentially correlated. ρ' controls the region of correlation parameters that the inputs are allowed to take. The parameter, p_{st}, controls the stability probability. L essentially acts like a Lipschitz constant, whereas f is introduced so that when X and Y are "too close", the algorithm is still allowed to make roughly "f flips". This "extra room" of f bits is introduced to allow for greater flexibility of the algorithm and it only makes our negative result stronger.

We now state our main result regarding the failure of stable algorithms for solving the NPP.

Theorem 2.3.2. *Fix any $\epsilon \in \left(0, \frac{1}{5}\right)$ and $L > 0$. Let $E_n : \mathbb{N} \to \mathbb{R}^+$ be an energy exponent satisfying*

$$\omega\left(n\log^{-\frac{1}{5}+\epsilon} n\right) \le E_n \le o(n).$$

For any $c > 0$, define

$$T(c) \triangleq \exp_2\left(2^{8cL\left(\frac{n}{E_n}\right)^{5+\frac{\epsilon}{4}}\log_2\left(cL\left(\frac{n}{E_n}\right)^{4+\frac{\epsilon}{4}}\right)}\right); \tag{2.8}$$

and set

$$\rho_n'(c) \triangleq 1 - \frac{1}{cL}\left(\frac{E_n}{n}\right)^{4+\frac{\epsilon}{4}}, \quad p_{f,n}(c) \triangleq \frac{1}{4T(c)\left(cL\left(\frac{n}{E_n}\right)^{4+\frac{\epsilon}{4}}+1\right)}, \quad \text{and} \quad p_{\mathrm{st},n}(c) \triangleq \frac{(E_n/n)^{8+\frac{\epsilon}{2}}}{9c^2L^2T(c)}. \tag{2.9}$$

Then, there exists constant $c_1, c_2 > 0$ and an $N^* \in \mathbb{N}$ such that the following holds. For every $n \ge N^*$, there exists no randomized algorithm, $\mathcal{A} : \mathbb{R}^n \to \mathcal{B}_n$ such that $\mathcal{A}$ is

$$\left(2^{-E_n}, c_1 n(E_n/n)^{4+\frac{\epsilon}{4}}, L, \rho_n'(c_2), p_{f,n}(c_2), p_{\mathrm{st},n}(c_2)\right) - \textit{optimal}$$

(for the NPP) in the sense of Definition 2.3.1.

The proof of Theorem 2.3.2 is provided in Section 2.6.7.

Several remarks are in order. In what follows, one should keep in their mind that $E_n/n = \log^{-O(1)} n$. Note first that there is no restriction on the runtime of $\mathcal{A}$, provided that it is stable. It is easy to check that $p_{\mathrm{st},n} \to 1$ as $n \to \infty$. Thus the algorithms that are ruled out satisfy with high probability

$$d_H\left(\mathcal{A}(X), \mathcal{A}(Y)\right) \le c_1 n \log^{-O(1)} n + L\|X - Y\|_2^2.$$

In particular, while $\mathcal{A}$ is stable, it is still allowed to make $\Omega\left(n \log^{-O(1)} n\right)$ "flips" even when X and Y are "too close".

Next, since c_2 and L are constant in n,

$$\rho_n'(c_2) = 1 - \frac{1}{c_2 L}\left(\frac{E_n}{n}\right)^{4+\frac{\epsilon}{4}} = 1 - \log^{-O(1)} n.$$

Namely, for our stability assumption, we restrict our attention to $\rho \in [1 - \log^{-O(1)} n, 1]$. It is worth noting that in the case when ρ is constant, $\rho = O(1)$, the stability per Definition 2.3.1 holds (with a sufficiently large constant L) irrespective of the algorithm: by the law of large numbers, $\|X-Y\|_2^2$ is $\Theta(n)$, whereas $d_H(\mathcal{A}(X,\omega), \mathcal{A}(Y,\omega)) \le n$ for any $X, Y \in \mathbb{R}^n$ and $\omega \in \Omega$; hence for L large enough (though constant), $L\|X-Y\|_2^2 > d_H(\mathcal{A}(X,\omega), \mathcal{A}(Y,\omega))$. In particular, in some sense the interesting regime is indeed when $\rho = 1 - o_n(1)$, as we investigate here.

Our next remark pertains to the term $T(c)$ appearing in (2.8). Keeping in mind that c and L are constants (in n); and n/E_n is $\omega(1)$, it follows that

$$8cL\left(\frac{n}{E_n}\right)^{5+\frac{\epsilon}{4}} \log_2\left(cL\left(\frac{n}{E_n}\right)^{4+\frac{\epsilon}{4}}\right) = \Theta\left(\left(\frac{n}{E_n}\right)^{5+\frac{\epsilon}{4}} \log_2\left(\frac{n}{E_n}\right)\right).$$

By assumption on E_n, $E_n/n = \log^{O(1)} n$. This yields the following order of growth for $T(c)$:

$$T(c) = \exp_2\left(2^{o\left(\log^{c'} n\right)}\right) \quad \text{for some} \quad c' \in (0,1).$$

In fact, any $c' > \left(\frac{1}{5} - \epsilon\right)\left(5 + \frac{\epsilon}{2}\right)$ works above. Since $\left(\frac{1}{5} - \epsilon\right)\left(5 + \frac{\epsilon}{2}\right) = 1 - 49\epsilon/10 + \Theta(\epsilon^2)$, the interval for c' is indeed non-vacuous as long as $\epsilon > 0$. Moreover, this interval gets larger as $\epsilon \to 1/5$ (more on this below).

An inspection of the terms $p_{f,n}(c)$ and $p_{\mathrm{st},n}(c)$ appearing in (2.9) reveals that they have the same order of growth as $T(c)^{-1}$. That is,

$$p_{f,n}(c), p_{\mathrm{st},n}(c) = \exp_2\left(-2^{o\left(\log^{c'} n\right)}\right) \quad \text{for any} \quad c' > \left(\frac{1}{5} - \epsilon\right)\left(5 + \frac{\epsilon}{2}\right).$$

In particular, while Theorem 2.3.2 requires high probability guarantees, these guarantees need not be exponential: a sub-exponential choice suffices. Moreover, as

$\epsilon \to \frac{1}{5}$, the restrictions become milder. In the limit (which corresponds essentially to $E_n = \Theta(n)$); it suffices to take a (large) constant probability of success and stability (as we elaborate below).

While the lower bound on the energy exponent E_n can potentially be improved slightly to $E_n = \omega\left(n \cdot \log^{-1/5} n \cdot \log\log n\right)$; it appears that $E_n = \Omega\left(n \cdot \log^{-1/5} n\right)$ is, in fact, necessary; see Section 2.4.2 for an informal argument. For the sake of keeping our presentation simple, we do not pursue this improvement.

While Theorem 2.3.2 rules out algorithms that are sufficiently stable in the sense of Definition 2.3.1, we are unable to prove that the LDM algorithm of Karmarkar and Karp [182] is stable, even though our simulation results, reported in Section 2.3.2, suggest that it is. We leave this as a very interesting, yet we believe an approachable, open problem.

On Energy Levels $2^{-\Theta(n)}$.

Theorem 2.3.2 addresses energy levels 2^{-E_n} with E_n lower bounded by some explicitly given $o(n)$ value. This naturally includes the energy levels $2^{-\Theta(n)}$. It appears, however, that for energy levels $2^{-n\epsilon}$ with $\epsilon > 0$; it is possible to strengthen Theorem 2.3.2 in various aspects, which we comment now.

It appears that a straightforward modification of Theorem 2.3.2—in particular invoking the $m-$OGP result, Theorem 2.2.3, with $m = O(1)$ as opposed to Theorem 2.2.6—yields that f/n can be taken to be constant (in n): the algorithm is then allowed to make $\Theta(n)$ flips even when X and Y are too close. Perhaps more importantly, the probability of success and the stability guarantee can also be boosted: this yields the failure of "stable" algorithms even with a *constant probability* of success/stability (where the constant is sufficiently close to one).

2.3.2 Stability of the LDM Algorithm. Simulation Results

In this section we report simulation results on running the LDM on correlated pairs of n-dimensional gaussian vectors. Thus let $X, X' \stackrel{d}{=} \mathcal{N}(0, I_n)$ be independent, and let $Y_i = \sqrt{1-\tau^2} X_i + \tau X'_i, 1 \le i \le n$ for a fixed value $\tau \in [0,1]$. Then $Y \stackrel{d}{=} \mathcal{N}(0, I_n)$ as well. We run the LDM algorithm on instances X and Y and denote the results by σ and $\sigma(\tau)$ respectively. We measure the overlap as $(1/n)\langle \sigma, \sigma(\tau)\rangle$ and report the results. The simulations were conducted for $n = 50, 100$ and 500 and reported on Figures 2-1,2-2 and 2-3 respectively. The horizontal axis corresponds to the value $\rho \triangleq -\log_2(\tau)$. So as τ decreases to zero and thus the correlation approaches unity, this parameter diverges to infinity. The logarithmic scale is motivated by scaling purposes explained below. As increasing ρ corresponds to higher level of correlation between X and Y, it should reduce the overlap between the corresponding outputs, as indeed this is seen on the figures. For each fixed value of n and ρ we compute the average overlap of 10 runs of the experiment and this is the value reported on the figure. We see that increasing the correlation continuously leads to continuous increase of the average overlap, suggesting that the stability indeed takes place. Curiously though,

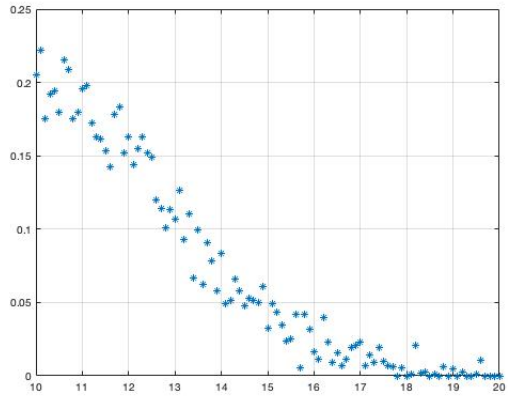

Figure 2-1: Average overlap as a function of correlation parameter ρ for $n = 50$.

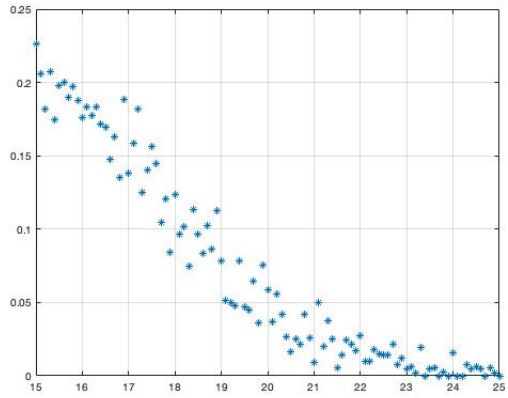

Figure 2-2: Average overlap as a function of correlation parameter ρ for $n = 100$.

in addition to the observed stability, the empirical average of the overlaps drops to a nearly zero level *precisely* at $\tau \approx n^{-\alpha \log n} = \exp(-\alpha \log^2 n)$, corresponding to $\rho = \alpha \log^2 n / \log 2$, at the threshold $\alpha = \frac{1}{2 \ln 2} = 0.721\ldots$ which is the leading constant conjectured for the performance of the LDM, as discussed in the introduction. To check this, note that the values of ρ above for $n = 50, 100$ and 500 are $15.91, 22.05$ and 40.17 respectively, for this choice of α, and this is close to the values where the overlaps touch the zero axis. At this point, we don't have a theoretical explanation for this phase transition. It is conceivable that the algorithm produces the smallest possible discrepancy which is stable under the perturbation above. We leave it as an interesting challenge for further investigation.

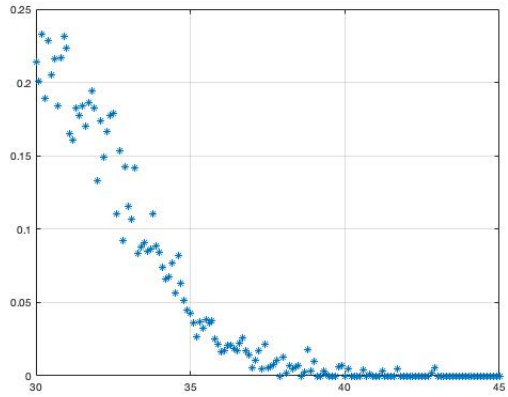

Figure 2-3: Average overlap as a function of correlation parameter ρ for $n = 500$.

2.3.3 2−Overlap Gap Property Implies Failure of an MCMC Family

In this section, we show that the overlap gap property for pairs of spin configurations established in Theorem 2.2.2, is a barrier for a family of Markov Chain Monte Carlo (MCMC) methods for solving the NPP.

Let as before $X \stackrel{d}{=} \mathcal{N}(0, I_n)$, denote the items to be partitioned.

The MCMC dynamics. We begin with specifying the relevant dynamics. Let $\beta_n \geq 0, n \geq 1$ be a sequence of inverse temperatures. For any $\sigma \in \mathcal{B}_n$, define the Hamiltonian $H(\sigma)$ by

$$H(\sigma) = \frac{1}{\sqrt{n}} |\langle \sigma, X \rangle| .$$

The Gibbs measure $\pi_\beta(\cdot)$ at temperature β^{-1} defined on $\mathcal{B}_n$ is specified by the probability mass function

$$\pi_\beta(\sigma) = \frac{1}{Z_\beta} \exp\left(-\beta H(\sigma)\right) \quad \text{where} \quad Z_\beta \triangleq \sum_{\sigma \in \mathcal{B}_n} \exp\left(-\beta H(\sigma)\right). \tag{2.10}$$

Here, Z_β is the "partition function" which ensures proper normalization for π_β. Note that a minus sign is added in front of $H(\sigma)$ in order to ensure that for β sufficiently large, that is for low enough temperatures, the Gibbs measure is concentrated on near ground-state configurations, i.e., on $\sigma \in \mathcal{B}_n$ with a small Hamiltonian value $H(\sigma)$). Indeed, observe that for

$$\sigma^* \triangleq \arg\min_{\sigma \in \mathcal{B}_n} H(\sigma),$$

we have

$$2\exp\left(-\beta H\left(\sigma^*\right)\right) \leq Z_\beta = \sum_{\sigma \in \mathcal{B}_n} \exp\left(-\beta H(\sigma)\right) \leq 2^n \exp\left(-\beta H\left(\sigma^*\right)\right).$$

Taking logarithms and dividing by $\beta > 0$, we arrive at

$$\frac{\ln 2}{\beta} - H(\sigma^*) \leq \frac{\ln Z_\beta}{\beta} \leq n\frac{\ln 2}{\beta} - H(\sigma^*).$$

For β sufficiently large, specifically when $\beta = \Omega\left(n2^{n\epsilon}\right)$ (which will be our eventual choice) we have

$$\frac{\ln Z_\beta}{\beta} + H\left(\sigma^*\right) \leq n\frac{\ln 2}{\beta} = O\left(2^{-n\epsilon}\right).$$

Hence, for $\beta = \Omega\left(n2^{n\epsilon}\right)$, it is the case that the Gibbs distribution $\pi_\beta(\cdot)$ is essentially concentrated on those $\sigma \in \mathcal{B}_n$ with $H(\sigma) = O\left(2^{-n\epsilon}\right)$.

We next construct the undirected graph $\mathbb{G}$ on 2^n vertices with edge set E on which the aforementioned MCMC dynamics is run.

- Each vertex corresponds to a spin configuration $\sigma \in \mathcal{B}_n$.
- For $\sigma, \sigma' \in \mathcal{B}_n$, $(\sigma, \sigma') \in E$ iff $d_H\left(\sigma, \sigma'\right) = 1$.

Let $X_0 \in \mathcal{B}_n$ be a spin configuration at which we initialize the MCMC dynamics. Let $\left(X_t\right)_{t \geq 0}$ be any nearest neighbor discrete time Markov chain on $\mathbb{G}$ initialized at X_0 and reversible with respect to the stationary distribution π_β. For example, X_t is discretized version of the Markov process with rates from σ to σ' defined by $\exp(\beta(H(\sigma') - H(\sigma)))$ when σ' is a neighbor of σ and is zero otherwise. Then the transition matrix $Q(\cdot, \cdot)$ for $\left(X_t\right)_{t \geq 0}$ satisfies the detailed balance equations for π_β: $\pi_\beta(\sigma)Q(\sigma, \sigma') = \pi_\beta(\sigma')Q(\sigma', \sigma)$ for every pair σ, σ' with $d_H\left(\sigma, \sigma'\right) = 1$.

Free energy wells. We now establish that the overlap gap property (shown in Theorem 2.2.2) induces a property called a *free energy well* (FEW) in the landscape of the NPP. This is a provable barrier for the MCMC methods [21], and has been employed to show slow mixing in other settings, see e.g. [21, 121].

Let $\epsilon \in \left(\frac{1}{2}, 1\right)$ and $\rho \in (0, 1)$ be the parameter dictated by Theorem 2.2.2. We define the following sets.

- $I_1 = \left\{\sigma \in \mathcal{B}_n : -\rho \leq \frac{1}{n}\left\langle \sigma, \sigma^* \right\rangle \leq \rho\right\}$.
- $I_2 \triangleq \left\{\sigma \in \mathcal{B}_n : \rho \leq \frac{1}{n}\left\langle \sigma, \sigma^* \right\rangle \leq \frac{n-2}{n}\right\}$, and $\overline{I_2} = \{-\sigma : \sigma \in I_2\}$.
- $I_3 = \{\sigma^*\}$ and $\overline{I_3} = \{-\sigma^*\}$.

We now establish the FEW property.

Theorem 2.3.3. *Let $\epsilon \in \left(\frac{1}{2}, 1\right)$ be arbitrary; and $\beta = \Omega\left(n2^{n\epsilon}\right)$. Then*

$$\min\left\{\pi_\beta\left(I_1\right), \pi_\beta\left(I_3\right)\right\} \geq \exp\left(\Omega\left(\beta 2^{-n\epsilon}\right)\right) \pi_\beta\left(I_2\right)$$

with high probability (with respect to X), as $n \to \infty$.

The proof of Theorem 2.3.3 is provided in Section 2.6.8.

Namely, the FEW property simply states that the set I_2 (of spins $\sigma \in \mathcal{B}_n$ having a "medium" overlap with σ^*) is a "well" of exponentially small (Gibbs) mass separating I_3 and $I_1 \cup \overline{I_2} \cup \overline{I_3}$.

Failure of MCMC. We now establish, as a consequence of the FEW property, Theorem 2.3.3, that the very natural MCMC dynamics introduced earlier provably fails for solving the NPP for "low enough temperatures", specifically when the temperature is exponentially small. This is a slow mixing result. More concretely, we establish that under an appropriate initialization, it requires an exponential amount of time for the aforementioned MCMC dynamics to "hit" a region of "non-trivial Gibbs mass".

To set the stage, let

$$\partial S \triangleq \{\sigma \in \mathcal{B}_n : d_H(\sigma, \sigma^*) = 1\}.$$

Clearly for any $\sigma \in \partial S$, $\mathcal{O}(\sigma, \sigma^*) = \frac{n-2}{n}$. Thus $\partial S \subset I_2$. Now, let us initialize the MCMC via $X_0 \stackrel{d}{=} \pi_\beta(\cdot | I_3 \cup \partial S)$. Define also the "escape time"

$$\tau_\beta \triangleq \inf\{t \in \mathbb{N} : X_t \notin I_3 \cup \partial S \mid X_0 \sim \pi_\beta(\cdot \mid I_3 \cup \partial S)\}. \tag{2.11}$$

We now establish the following "slow mixing" result.

Theorem 2.3.4. *Let $\epsilon \in \left(\frac{1}{2}, 1\right)$, and $\beta = \Omega(n2^{n\epsilon})$. Then, the following holds.*

(a) I_1 and $\overline{I_3}$ collectively contain at least a constant proportion of the Gibbs mass:

$$\pi_\beta\left(I_1 \cup \overline{I_3}\right) \geq \frac{1}{2}(1 + o_n(1)),$$

with high probability as $n \to \infty$.

(b) With high probability (over $X \stackrel{d}{=} \mathcal{N}(0, I_n)$) as $n \to \infty$

$$\tau_\beta = \exp\left(\Omega\left(\beta 2^{-n\epsilon}\right)\right).$$

In particular, for $\beta = \omega(n2^{n\epsilon})$, we obtain $\tau_\beta = \exp(\Omega(n))$ w.h.p. as $n \to \infty$.

The proof of Theorem 2.3.4 is provided in Section 2.6.9.

Per Theorem 2.3.4, when the chain starts in the ground state σ^* or one of its neighboring states, it takes an exponential amount of time for the chain to exit this set of states, and in particular, it takes an exponential amount of time to enter a region of nearly half of Gibbs mass; implying slow mixing.

It is worth noting that Theorem 2.3.4 is shown when the temperature β^{-1} is low enough, more specifically is exponentially small. This ensures the Gibbs measure is well-concentrated on ground states. We leave the analysis of the MCMC dynamics in

the high-temperature regime (i.e., lower values of β) as an interesting open problem for future work. We conjecture that any setting of the temperature will either result in slow mixing or will lead to a Gibbs measure dominated by low values of E_n. We don't have a guess for the exact value of such low energy values. We leave this question for a future exploration.

2.4 Limitations of Our Techniques

Given that our methods fall short of addressing the *statistical-to-computational gap* of the NPP all the way down to $2^{-\Theta(\log^2 n)}$; it is natural to explore the limits of the techniques exploited in this chapter.

2.4.1 Limitation of the $m-$Overlap Gap Property for Growing m

In this section, we give an informal argument suggesting the absence of the $m-$OGP when the energy level E_n is $O\left(\sqrt{n\log_2 n}\right)$. Our informal argument will reveal the following. It appears not possible to establish $m-$OGP (for super-constant m) as we do in Theorem 2.2.6, for energy levels above $\exp_2\left(-\omega\left(\sqrt{n\log_2 n}\right)\right)$. We now detail this.

Step 1: $E_n = \omega(\sqrt{n})$ is Necessary.

We first note, upon studying the proof of Theorem 2.2.6 more carefully, that for the first moment argument to work, one should take $\beta = 1 - o_n(1)$. For convenience, let $\beta = 1 - 2\nu_n$, where $\nu_n = o_n(1)$ is a sequence of positive reals. Furthermore, to ensure the invertibility of a certain covariance matrix arising in the analysis, one should also take $\eta \lesssim \nu_n/m$ (see the proof for further details on this matter).

Now, for the OGP to be meaningful, it should be the case that $n\eta = \Omega(1)$, as noted already previously. Indeed, otherwise the overlap region is void, since no admissible overlap values ρ can be found within the interval $[\beta - \eta, \beta]$. Now, since $\eta \lesssim \nu_n/m$,

$$n\eta = \Omega(1) \implies \frac{n\nu_n}{m} = \Omega(1) \implies n\nu_n = \Omega(m).$$

Next, for an $m-$tuple $\left(\sigma^{(i)} : 1 \le i \le m\right)$; the energy value, 2^{-E_n}, contributes to a $-mE_n$ in the exponent (we again refer the reader to the proof for further details). Finally, a very crude cardinality bound on the number of $m-$tuples $\left(\sigma^{(i)} : 1 \le i \le m\right)$ with pairwise inner products $\frac{1}{n}\left\langle \sigma^{(i)}, \sigma^{(j)} \right\rangle \in [\beta-\eta, \beta]$, $1 \le i < j \le m$, is the following: using the naïve approximation $\log_2 \binom{n}{k} = (1 + o_n(1))k \log_2 \frac{n}{k}$ valid for $k = o(n)$, we arrive at

$$2^n \binom{n}{n\frac{1-\beta}{2}}^{m-1} \sim \exp_2\left(n + mn\nu_n \log \frac{1}{\nu_n}\right),$$

where we have used $m = \omega_n(1)$ and $\beta = 1 - 2\nu_n$; while ignoring the lower order terms

for convenience. Blending these observations together, we arrive at the following formula, for the exponent of the first moment:

$$\xi(n) \triangleq n + mn\nu_n \log \frac{1}{\nu_n} - mE_n.$$

Now, for the first moment argument to work, it should be the case $-\xi(n) = \omega_n(1)$. Hence, $mE_n = \Omega(n)$ must hold. Since $n\nu_n = \Omega(m)$ as shown above, this yields

$$n\nu_n = \Omega(m) = \Omega\left(\frac{n}{E_n}\right).$$

Now, a final constraint is

$$mE_n = \Omega\left(mn\nu_n \log \frac{1}{\nu_n}\right) \Leftrightarrow E_n = \Omega\left(n\nu_n \log \frac{1}{\nu_n}\right).$$

But since $\nu_n = o_n(1)$, we have $\log \frac{1}{\nu_n} = \omega_n(1)$, and consequently, we must have, at the very least,

$$E_n = \omega\left(\frac{n}{E_n}\right) \Leftrightarrow E_n = \omega(\sqrt{n}).$$

Step 2: from $E_n = \omega\left(\sqrt{n}\right)$ to $E_n = \omega\left(\sqrt{n \log_2 n}\right)$.

We now let $E_n = \phi(n)\sqrt{n}$, where $\phi(n) = \omega_n(1)$, and plug this in above to study the parameters numerically. Inspecting the lines above, one should take $m = C\frac{\sqrt{n}}{\phi(n)}$, where $C > 1$ is some constant. This, in turn, yields that we require $\nu_n = \frac{g(n)}{\phi(n)\sqrt{n}}$, where $g(n) = \Omega(1)$. In particular, observe that

$$\log \frac{1}{\nu_n} = \frac{1}{2}\log_2 n(1 + o_n(1)) = \Theta(\log_2 n).$$

Now, a final constraint, as one might recall from above, is that the exponent, $-\xi(n)$, should be $\omega_n(1)$ as $n \to \infty$. With this, it should hold

$$n\nu_n \log \frac{1}{\nu_n} = O(E_n) = O(\phi(n)\sqrt{n}).$$

Since

$$n\nu_n \log \frac{1}{\nu_n} = \Theta\left(\frac{g(n)\sqrt{n}\log_2 n}{\phi(n)}\right),$$

it should be the case

$$\frac{g(n)\sqrt{n}\log_2 n}{\phi(n)} \lesssim \phi(n)\sqrt{n},$$

which implies $\phi(n) = \Omega(\sqrt{\log_2 n})$.

Namely, this argument demonstrates the following: if one wants to establish the

overlap gap property for an energy exponent E_n through a first moment technique, E_n should have a growth of at least $\sqrt{n \log_2 n}$; otherwise the moment argument fails.

2.4.2 Limitation of the Ramsey Argument

An important question that remains is whether one can leverage further the $m-$OGP result (Theorem 2.2.6) to establish an analogue of our hardness result (Theorem 2.3.2) for energy levels with an exponent E_n that is at least slightly below $\omega\left(n \log^{-1/5+\epsilon} n\right)$ or all the way to $\omega\left(\sqrt{n \log_2 n}\right)$. We now argue that using our line of argument based on the Ramsey Theory, $E_n = \Omega\left(n \log^{-\frac{1}{5}} n\right)$, no beyond, is essentially the best exponent one would hope to address.

Let E_n be a target exponent for which one wants to establish the hardness; and m be the OGP parameter required per Theorem 2.2.6. Our proof uses, in a crucial way, certain properties regarding Ramsey numbers arising in extremal combinatorics. To that end, let $R_Q(m)$ denotes the smallest $n \in \mathbb{N}$ such that any Q (edge) coloring of K_n contains a monochromatic K_m (see Theorem 2.6.7 for more details). Our argument then contains the following ingredients. We generate a certain number T of "instances" (of the NPP) such that $T \geq R_2(M)$ for $M \geq R_Q(m)$ where Q corresponds to a discretization level we need to address E_n. When then essentially (a) construct a graph $\mathbb{G}$ on T vertices satisfying certain properties, in particular $\alpha(\mathbb{G}) \leq M-1$ (where $\alpha(\mathbb{G})$ is the cardinality of any largest independent set of $\mathbb{G}$) (b) extract a clique K_M of $\mathbb{G}$ whose edges are colored with one of Q available colors; and (c) use $M \geq R_Q(m)$ to conclude that the original graph, $\mathbb{G}$, contains a monochromatic K_m. From here, we then argue that this yields a forbidden configuration, a contradiction with the $m-$OGP.

Using well-known upper and lower bounds on Ramsey numbers (see e.g. [80]) one should then choose $T \geq \exp_2(\Theta(M))$. Moreover, the best lower bound on $R_Q(m)$, due to Lefmann [198], asserts that $R_Q(m) \geq \exp_2(mQ/4)$. Combining these bounds, we then conclude that T should be of order at least

$$T \geq \exp_2\left(2^{\Theta(mQ)}\right). \tag{2.12}$$

Now, an inspection of our proof of Theorem 2.3.2 yields that for certain union bounds, e.g. (2.157), to work; T should be sub-exponential: $T = 2^{o(n)}$. Combining this with (2.12); a necessary condition turns out to be

$$mQ = O\left(\log_2 n\right). \tag{2.13}$$

Now, the discretization Q should be sufficiently fine to ensure that the overlaps are eventually "trapped" within the (forbidden) overlap region of length η dictated by Theorem 2.2.6. In particular, tracing our proof, it appears from (2.141) that

$$Q = \Omega\left(\frac{1}{\eta^2}\right) \tag{2.14}$$

should hold. Furthermore, from the discussion on $m-$OGP; as well as the proof of Theorem 2.2.6, it appears also that mE_n should be $\Omega(n)$, that is

$$m = \Omega\left(\frac{n}{E_n}\right). \tag{2.15}$$

Now, we take the overlap value η to be $g(n)/n$, where $g(n) = \omega(1)$; but it also satisfies other certain, natural, constraints. In particular, using (2.66) and (2.69); for the parameters to make sense, $g(n)$ should be $o(E_n)$. Let

$$E_n = ns(n), \quad g(n) = ns(n)z(n), \quad \text{and} \quad \eta = s(n)z(n) \quad \text{where} \quad s(n), z(n) = o_n(1). \tag{2.16}$$

Combining (2.14), (2.15) and (2.16); we therefore have

$$mQ = \Omega\left(\frac{n}{E_n\eta^2}\right) = \Omega\left(\frac{1}{s(n)^3 z(n)^2}\right). \tag{2.17}$$

Furthermore, to ensure Theorem 2.2.6 applies; the "exponent" of the first moment should not "blow up". For this reason, using (2.85), (2.86), as well as the counting term (2.77), it must at least hold that

$$mE_n = \Omega\left(\frac{mng(n)}{E_n}\right) \iff \frac{E_n}{n} = s(n) = \Omega\left(z(n)\right).$$

This, together with (2.17) as well as the upper bound (2.13), implies that

$$s(n) = \Omega\left(\log^{-\frac{1}{5}} n\right).$$

Hence,

$$E_n = ns(n) = \Omega\left(n\log^{-\frac{1}{5}} n\right),$$

is essentially indeed the best possible. We gave ourselves an ϵ "extra room" in Theorem 2.3.2 so as to avoid complicating relevant quantities any further.

A very interesting question is whether one can by-pass the Ramsey argument altogether. This would help establishing the failure of (presumably more) stable algorithms for even higher energy levels, $E_n = \omega\left(\sqrt{n\log_2 n}\right)$, a regime where Theorem 2.2.6 is applicable.

2.5 Open Problems and Future Work

Our work suggests interesting avenues for future research. While we have focused on the NPP in the present work for simplicity, we believe that many of our results extend to the multi-dimensional case, VBP (2.2), as well; perhaps at the cost of more detailed and computation-heavy proofs. This was noted already in Remark 2.2.4.

Yet another very important direction pertains the *statistical-to-computational gap* of the NPP. The $m-$OGP results that we established hold for energy levels $2^{-\Theta(n)}$ when

$m = O(1)$; and for 2^{-E_n}, $\omega\left(\sqrt{n \log_2 n}\right) \leq E_n \leq o(n)$, when $m = \omega_n(1)$. While we are able to partially explain the aforementioned *statistical-to-computational gap* to some extent, we are unable close it all the way down to the current computational threshold: the best known polynomial-time algorithm to this date achieves an exponent of only $\Theta\left(\log^2 n\right)$. A very interesting open question is whether this gap can be "closed" altogether. That is, either devise a better (efficient) algorithm, improving upon the algorithm by Karmarkar and Karp [182]; or establish the hardness by taking one of the alternative routes (mentioned in the introduction) tailored for proving *average-case hardness*. In light of the fact that not much work has been done in the algorithmic front since the paper [182], it is plausible to hope that better efficient algorithms can indeed be found. In particular, a potential direction appears to be setting up an appropriate Markov Chain dynamics, and establishing rapid mixing. We leave this as an open problem for future work.

While we are able to rule out stable algorithms in the sense of Definition 2.3.1, we are unable to prove that the algorithm by Karmarkar and Karp, in particular the LDM algorithm introduced earlier, is stable with appropriate parameters, although our simulation results suggest that it is. We leave this as yet another open problem.

Another direction pertains the parameters of algorithms that we consider. In particular, one potential direction is to establish Theorem 2.3.2 when the algorithm say has $o_n(1)$ probability of success. That is, p_f in Definition 2.3.1 is $1 - o_n(1)$. We conjecture that the value $p_f = 1 - n^{-O(1)}$ is within the reach.

2.6 Proofs

2.6.1 Auxiliary Results

Below, we record several auxiliary results that will guide our proofs. The first result is the standard asymptotic approximation for the factorial.

$$\log_2 n! = n \log_2 n - n \log_2 e + O\left(\log_2 n\right). \tag{2.18}$$

The second is a very standard approximation for the binomial coefficients, whose proof we include herein for completeness.

Lemma 2.6.1. *Let $n, k \in \mathbb{N}$, where $k = o(n)$. Then,*

$$\log_2 \binom{n}{k} = (1 + o_n(1)) k \log_2 \frac{n}{k}.$$

Proof. Note that for any $0 \leq i \leq k-1$, $\frac{n-i}{k-i} \geq \frac{n}{k}$. Hence,

$$\left(\frac{n}{k}\right)^k \leq \prod_{0 \leq i \leq k-1} \frac{n-i}{k-i} = \binom{n}{k}.$$

Next,

$$\binom{n}{k}\left(\frac{k}{n}\right)^k \le \sum_{0\le t\le n}\binom{n}{t}\left(\frac{k}{n}\right)^t = \left(1+\frac{k}{n}\right)^n.$$

Since $\ln(1+x)\le x$, setting $x=k/n$ yields

$$\left(1+\frac{k}{n}\right)^n \le e^k.$$

Combining these, we obtain

$$\left(\frac{n}{k}\right)^k \le \binom{n}{k} \le \left(\frac{en}{k}\right)^k.$$

Taking now the logarithms both sides, and keeping in mind that $\log_2 \frac{n}{k} = \omega_n(1)$; we arrive at

$$k\log_2\frac{n}{k} \le \log_2\binom{n}{k} \le k\left(\log_2\frac{n}{k}+\log_2 e\right) = k\log_2\frac{n}{k}(1+o_n(1)).$$

Hence,

$$\log_2\binom{n}{k} = (1+o_n(1))k\log_2\frac{n}{k}$$

as claimed. □

The third auxiliary result is a theorem from the matrix theory.

Theorem 2.6.2. (Wielandt-Hoffman)
Let $A, A+E \in \mathbb{R}^{n\times n}$ be two symmetric matrices with respective eigenvalues $\lambda_1(A)\ge\lambda_2(A)\ge\cdots\ge\lambda_n(A)$ and $\lambda_1(A+E)\ge\lambda_2(A+E)\ge\cdots\ge\lambda_n(A+E)$. Then,

$$\sum_{1\le i\le n}\left(\lambda_i(A+E)-\lambda_i(A)\right)^2 \le \|E\|_F^2.$$

For a reference, see e.g. [170, Corollary 6.3.8]; and see [166] for the original paper by Wielandt and Hoffman.

2.6.2 Proof of Theorem 2.2.2

Proof. Let $\epsilon \in \left(\frac{1}{2}, 1\right]$. Let $\rho \in (0,1)$ to be tuned appropriately, and

$$\mathcal{Z}(\rho) \triangleq \left\{(\sigma,\sigma') \in \mathcal{B}_n \times \mathcal{B}_n : \mathcal{O}(\sigma,\sigma') \in \left[\rho, \frac{n-2}{n}\right]\right\}.$$

Set

$$N \triangleq \sum_{(\sigma,\sigma')\in\mathcal{Z}(\rho)} \mathbb{1}\left\{\frac{1}{\sqrt{n}}\left|\langle\sigma, X\rangle\right|, \frac{1}{\sqrt{n}}\left|\langle\sigma', X\rangle\right| = O\left(2^{-n\epsilon}\right)\right\} \tag{2.19}$$

We will establish that $\mathbb{E}[N] = \exp(-\Theta(n))$. This, together with Markov's inequality, will then yield the desired conclusion:

$$\mathbb{P}(N \geq 1) \leq \mathbb{E}[N] = \exp(-\Theta(n)).$$

Step I. Counting. We first upper bound the cardinality $|\mathcal{Z}(\rho)|$. Note that there are 2^n choices for $\sigma \in \mathcal{B}_n$. Having chosen σ; σ' can now be chosen in

$$\sum_{1 \leq k \leq \lceil n\frac{1-\rho}{2} \rceil} \binom{n}{k}$$

different ways. This is due to the fact that if $k = d_H(\sigma, \sigma')$ then $\mathcal{O}(\sigma, \sigma') = \left|1 - 2\frac{k}{n}\right|$. Using Stirling's approximation (2.18), and the fact the sum contains $n^{O(1)} = \exp_2(O(\log_2 n))$ terms, we arrive at the upper bound

$$|\mathcal{Z}(\rho)| \leq \exp_2\left(n + nh\left(\frac{1-\rho}{2}\right) + O(\log_2 n)\right). \tag{2.20}$$

Here $h(x) = -x\log_2 x - (1-x)\log_2(1-x)$ is the binomial entropy function logarithm base two.

Step II. Upper bound on probability. Let $(\sigma, \sigma') \in \mathcal{Z}(\rho)$ with $\mathcal{O}(\sigma, \sigma') = \bar{\rho}$. Set

$$Y_\sigma \triangleq \frac{1}{\sqrt{n}}\langle \sigma, X\rangle \quad \text{and} \quad Y_{\sigma'} \triangleq \frac{1}{\sqrt{n}}\langle \sigma', X\rangle.$$

Note that $Y_\sigma, Y_{\sigma'} \stackrel{d}{=} \mathcal{N}(0,1)$ with correlation $\bar{\rho}$. Now, let $C > 0$ be a constant; and denote by $\mathcal{R}_C$ the region

$$\mathcal{R}_C \triangleq \left[-C2^{-n\epsilon}, C2^{-n\epsilon}\right] \times \left[-C2^{-n\epsilon}, C2^{-n\epsilon}\right].$$

Denote also

$$f(x,y) \triangleq \exp\left(-\frac{1}{2(1-\bar{\rho}^2)}\left(x^2 - 2\bar{\rho}xy + y^2\right)\right)$$

As long as $\bar{\rho} \in (0,1)$, we have $f(x,y) \leq 1$ for every x, y. Furthermore,

$$\sqrt{1 - \left(1 - \frac{2}{n}\right)^2} = \sqrt{\frac{4}{n} - \frac{4}{n^2}} = \frac{2}{\sqrt{n}}(1 + o_n(1)). \tag{2.21}$$

We then have,

$$\mathbb{P}\left((Y_\sigma, Y_{\sigma'}) \in \mathcal{R}_C\right) = \frac{1}{2\pi\sqrt{1-\bar{\rho}^2}} \int_{(x,y)\in\mathcal{R}_C} f\left(x,y\right)\, dx\, dy \tag{2.22}$$

$$\leq \mathcal{C}\frac{1}{\sqrt{1-\left(1-\frac{2}{n}\right)^2}} 2^{-2n\epsilon} \tag{2.23}$$

$$= \mathcal{C}'\left(1+o_n(1)\right) 2^{-2n\epsilon}\sqrt{n}, \tag{2.24}$$

where $\mathcal{C}, \mathcal{C}' > 0$ are some absolute constants. Note that (2.24) is uniform in $\bar{\rho}$: it holds true for **every** $(\sigma, \sigma') \in \mathcal{Z}\left(\rho\right)$.

Step III. Computing the expectation. We now compute $\mathbb{E}[N]$ of N introduced in (2.19). Using linearity of expectation, (2.20), and (2.24), we arrive at

$$\mathbb{E}[N] \leq \exp_2\left(n + nh\left(\frac{1-\rho}{2}\right) - 2n\epsilon + O\left(\log_2 n\right)\right). \tag{2.25}$$

Since $\epsilon > \frac{1}{2}$, there exists a $\rho > 0$ such that

$$1 + h\left(\frac{1-\rho}{2}\right) - 2\epsilon < 0.$$

For this choice of ρ, we indeed have per (2.25) that

$$\mathbb{E}[N] = \exp\left(-\Theta(n)\right),$$

concluding the proof. □

2.6.3 Proof of Theorem 2.2.3

Proof. For any $1 \leq i \leq m$ and $\tau \in \mathcal{I}$; recall $Y_i(\tau) \triangleq \sqrt{1-\tau^2}X_0 + \tau X_i \in \mathbb{R}^n$; and

$$H(\sigma^{(i)}, Y_i(\tau)) \triangleq \frac{1}{\sqrt{n}} |\left\langle \sigma^{(i)}, Y_i(\tau)\right\rangle|.$$

Define,

$$S(\beta, \eta, m) \triangleq \left\{(\sigma^{(1)}, \ldots, \sigma^{(m)}) : \sigma^{(i)} \in \{-1,1\}^n, \mathcal{O}(\sigma^{(i)}, \sigma^{(j)}) \in [\beta-\eta, \beta], 1 \leq i < j \leq m\right\},$$

and

$$N(\beta, \eta, m, \epsilon, \mathcal{I}) = \sum_{(\sigma^{(1)},\ldots,\sigma^{(m)})\in S(\beta,\eta,m)} \mathbb{1}\left\{\exists \tau_1, \ldots, \tau_m \in \mathcal{I} : H(\sigma^{(i)}, Y_i(\tau_i)) = O\left(2^{-n\epsilon}\right), 1 \leq i \leq m\right\}. \tag{2.26}$$

Observe that $N(\beta,\eta,m,\epsilon,\mathcal{I}) = |\mathcal{S}(\beta,\eta,m,\epsilon,\mathcal{I})|$. In what follows, we will establish that for an appropriate choice of parameters β,η, and $m \in \mathbb{Z}_+$,

$$\mathbb{E}[N(\beta,\eta,m,\epsilon,\mathcal{I})] = \exp_2(-\Theta(n)),$$

which will then yield,

$$\mathbb{P}(\mathcal{S}(\beta,\eta,m,\epsilon,\mathcal{I}) \neq \varnothing) = \mathbb{P}(N(\beta,\eta,m,\epsilon,\mathcal{I}) \geq 1) \leq \exp_2(-\Theta(n))$$

through Markov's inequality, and thus the conclusion.

Step I: Counting. We start by upper bounding $|S(\beta,\eta,m)|$. There are 2^n choices for $\sigma^{(1)}$. Now, for any fixed σ, we claim there exists $2\binom{n}{n\frac{1-\rho}{2}}$ sign configurations $\sigma' \in \{-1,1\}^n$ for which $\mathcal{O}(\sigma,\sigma') = \rho$. Indeed, let $k \triangleq \sum_{1\leq i\leq n} \mathbb{1}\{\sigma_i \neq \sigma_i'\}$, the number of coordinates σ and σ' disagree. With this we have $\mathcal{O}(\sigma,\sigma') = \left|\frac{n-2k}{n}\right|$, from which we obtain $k = n\frac{1\pm\rho}{2}$. Equipped with this observation, we now compute the number of choices for $\sigma^{(2)}$ as $2\sum_{\beta-\eta\leq\rho\leq\beta:\rho n\in\mathbb{Z}} \binom{n}{n\frac{1-\rho}{2}}$. We then obtain

$$\begin{aligned}
|S(\beta,\eta,m)| &\leq 2^n \left(2 \sum_{\beta-\eta\leq\rho\leq\beta:\rho n\in\mathbb{Z}} \binom{n}{n\frac{1-\rho}{2}}\right)^{m-1} && (2.27)\\
&= 2^n \left(2 \sum_{\beta-\eta\leq\rho\leq\beta:\rho n\in\mathbb{Z}} \exp_2\left(nh\left(\frac{1-\rho}{2}\right) + O(\log_2 n)\right)\right)^{m-1} && (2.28)\\
&\leq 2^n \left(\exp_2\left(nh\left(\frac{1-\beta+\eta}{2}\right) + O(\log_2 n)\right)\right)^{m-1}. && (2.29)
\end{aligned}$$

We now justify these lines. Recall that $\log_2 n! = n\log_2 n - n\log_2 e + O(\log_2 n)$ by the Stirling's approximation, (2.18). Using this, we obtain $\rho \in (0,1)$, $\binom{n}{\rho n} = \exp_2(nh(\rho) + O(\log_2 n))$ where we recall that $h(x) = -x\log_2 x - (1-x)\log_2(1-x)$ is the binary entropy function logarithm base 2. Thus (2.28) follows. (2.29) is a consequence of the fact that the sum involves $O(n)$ terms. We conclude

$$|S(\beta,\eta,m)| \leq \exp_2\left(n + n(m-1)h\left(\frac{1-\beta+\eta}{2}\right) + (m-1)O(\log_2 n)\right). \quad (2.30)$$

Step II: Probability calculation. Fix $\tau_1,\ldots,\tau_m \in \mathcal{I}$. For any fixed $(\sigma^{(1)},\ldots,\sigma^{(m)}) \in S(\beta,\eta,m)$, we now investigate

$$\mathbb{P}\left(H(\sigma^{(i)}, Y_i(\tau_i)) = O(2^{-n\epsilon}), 1 \leq i \leq m\right).$$

To that end, let $Z_i = \frac{1}{\sqrt{n}}\left\langle \sigma^{(i)}, Y_i(\tau_i)\right\rangle$, and let $\rho_{ij} = \frac{1}{n}\left\langle \sigma^{(i)}, \sigma^{(j)}\right\rangle$. Note that for each $1 \leq i \leq m$, Z_i is standard normal, and moreover, the vector $(Z_1,\ldots,Z_m)$ is a multivariate Gaussian with mean zero and some covariance matrix Σ.

We now investigate this covariance matrix. To that end, let $\gamma_i \triangleq \sqrt{1-\tau_i^2}$, $1 \leq i \leq m$. We first compute $\mathbb{E}[Y_i(\tau_i)Y_j(\tau_j)^T] \in \mathbb{R}^{n\times n}$. We have

$$Y_i(\tau_i)Y_j(\tau_j)^T = \gamma_i\gamma_j X_0X_0^T + \gamma_i\tau_j X_0X_j^T + \gamma_j\tau_i X_iX_0^T + \tau_i\tau_j X_iX_j^T.$$

Since X_0, X_i, X_j are i.i.d., we thus obtain

$$\mathbb{E}[Y_i(\tau_i)Y_j(\tau_j)^T] = \gamma_i\gamma_j I_n \in \mathbb{R}^{n\times n}. \tag{2.31}$$

Equipped with this, we now have for any $1 \leq i < j \leq m$,

$$\begin{aligned}\mathrm{Cov}(Z_i, Z_j) &= \mathbb{E}\left[\frac{1}{\sqrt{n}}\left\langle\sigma^{(i)}, Y_i(\tau_i)\right\rangle \frac{1}{\sqrt{n}}\left\langle\sigma^{(j)}, Y_j(\tau_j)\right\rangle\right] \\ &= \frac{1}{n}(\sigma^{(i)})^T\mathbb{E}[Y_i(\tau_i)Y_j(\tau_j)^T]\sigma^{(j)} \\ &= \rho_{ij}\gamma_i\gamma_j.\end{aligned}$$

Namely, the covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$ of $(Z_1, \ldots, Z_m)$ is given by $\Sigma_{ii} = 1$ for $1 \leq i \leq m$, and $\Sigma_{ij} = \Sigma_{ji} = \rho_{ij}\gamma_i\gamma_j$ for $1 \leq i < j \leq m$.

Now, fix arbitrary constants $C_1, \ldots, C_m > 0$; and let $V \subset \mathbb{R}^m$ be the region defined by

$$V = \left(-C_1 2^{-n\epsilon}, C_1 2^{-n\epsilon}\right) \times \left(-C_2 2^{-n\epsilon}, C_2 2^{-n\epsilon}\right) \times \cdots \times \left(-C_m 2^{-n\epsilon}, C_m 2^{-n\epsilon}\right).$$

Provided Σ is invertible, which we verify independently, the probability of interest evaluates to

$$\mathbb{P}((Z_1, \ldots, Z_m) \in V) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \int_V \exp\left(-\frac{x^T\Sigma^{-1}x}{2}\right) dx.$$

As $\exp\left(-\frac{x^T\Sigma^{-1}x}{2}\right) \leq 1$, we can crudely upper bound this by

$$\frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}}\mathrm{Vol}(V) = \frac{2^{m/2}\prod_{1\leq j\leq m} C_j}{\pi^{m/2}}|\Sigma|^{-1/2}2^{-n\epsilon m}.$$

Observe now that $2^{m/2}, \pi^{m/2}$, and $\prod_{1\leq j\leq m} C_j$ are all constant order $O(1)$ with respect to n. Suppose now that Σ is such that the determinant of Σ is bounded away from zero by an explicit constant controlled solely by m, β, η, regardless of $\mathcal{I}$ and regardless of $\tau_1, \ldots, \tau_m \in \mathcal{I}$. If this is the case, then $|\Sigma|^{-1}$ is $O(1)$ with respect to n. This yields

$$\mathbb{P}\left((Z_1, \ldots, Z_m) \in V\right) \leq \exp_2\left(-n\epsilon m + O(1)\right). \tag{2.32}$$

We now take now a union bound over all $\tau_1, \ldots, \tau_m \in \mathcal{I}$ (note that there are at most

$2^{mo(n)} = 2^{o(n)}$ such terms), and arrive at

$$\mathbb{P}\left(\exists \tau_1, \dots, \tau_m \in \mathcal{I} : H(\sigma^{(i)}, Y_i(\tau_i)) = O\left(2^{-n\epsilon}\right), 1 \le i \le m\right) = \exp_2(-n\epsilon m + o(n)). \tag{2.33}$$

Step III: Calculating the expectation $\mathbb{E}[N(\beta, \eta, m, \epsilon)]$. Provided Σ **is invertible**, we can compute the expectation (2.26) by using (2.30) and (2.33):

$$\mathbb{E}[N(\beta, \eta, m, \epsilon)] \le \exp_2\left(n + n(m-1)h\left(\frac{1-\beta+\eta}{2}\right) + o(n) - n\epsilon m\right).$$

Hence, provided the parameters β, η, m are chosen so that

$$1 + (m-1)h\left(\frac{1-\beta+\eta}{2}\right) - \epsilon m < 0, \tag{2.34}$$

and $|\Sigma|$ is bounded away zero by an explicit constant independent of $\mathcal{I}$, and the choices $\tau_1, \dots, \tau_m$, we indeed obtain $\mathbb{E}[N(\beta, \eta, m, \epsilon)] = \exp_2(-\Theta(n))$, as desired.

We choose $m > \frac{2}{\epsilon}$. With this, $1 - \frac{\epsilon m}{2} < 0$. Observe now that if $0 < \eta < \beta < 1$ are chosen so that $h\left(\frac{1-\beta+\eta}{2}\right) < \frac{\epsilon}{2}$, the condition (2.34) is indeed satisfied. With this, it suffices for $0 < \eta < \beta < 1$ to satisfy

$$\beta - \eta > 1 - 2h^{-1}(\epsilon/2), \tag{2.35}$$

where $h^{-1} : [0,1] \to [0, 1/2]$ is the inverse of the binary entropy function.

Step IV: Invertibility of Σ. We next study the invertibility of covariance matrix Σ, which we recall $\Sigma_{ii} = 1$ for $1 \le i \le m$; and $\Sigma_{ij} = \Sigma_{ji} = \gamma_i \gamma_j \rho_{ij}$, $1 \le i < j \le m$, for some $\tau_1, \dots, \tau_j \in \mathcal{I}$.

Let us now define an auxiliary matrix $\bar{\Sigma} \in \mathbb{R}^{m \times m}$ by $\bar{\Sigma}_{ii} = 1$ for $1 \le i \le m$ and $\bar{\Sigma}_{ij} = \bar{\Sigma}_{ji} = \rho_{ij}$ for any $1 \le i < j \le m$. Namely, $\bar{\Sigma}$ is the covariance matrix Σ when $\gamma_1 = \cdots = \gamma_m = 1$, namely when $\tau_1 = \cdots = \tau_m = 0$, and thus $\sigma^{(1)}, \dots, \sigma^{(m)}$ are near-ground states with respect to the **same** instance $X_0 \in \mathbb{R}^n$ of the problem.

Note that $\rho_{ij} = \mathcal{O}(\sigma^{(i)}, \sigma^{(j)}) = |\bar{\Sigma}_{ij}|$. Thus,

$$\Sigma_{ij} \in [-\beta, -\beta + \eta] \cup [\beta - \eta, \beta],$$

for $1 \le i < j \le m$. In particular, there exists $2^{\binom{m}{2}}$ possible "signs" for the off-diagonal entries for the matrix $\bar{\Sigma}$. With this observation, we now prove an auxiliary lemma.

Lemma 2.6.3. *Let $m \in \mathbb{Z}_+$, $K \triangleq 2^{\binom{m}{2}}$. Construct a family $M_k(x)$, $1 \le k \le K$, of $m \times m$ matrices with unit diagonal entries, where each off-diagonal entry is defined in terms of x, as follows. Fix any "sign-configuration" $\gamma^{(k)} = (\gamma_{ij}^{(k)} : 1 \le i < j \le m) \in \{-1, 1\}^K$, $1 \le k \le K$. let $M_k(x) \in \mathbb{R}^{m \times m}$ be the matrix defined by $(M_k(x))_{ii} = 1$ for $1 \le i \le m$, and $(M_k(x))_{ij} = (M_k(x))_{ji} = \gamma_{ij}^{(k)} x$. Then the following holds:*

(a) *Define* $\varphi_k(x) \triangleq \sigma_{\min}(M_k(x))$, $1 \le k \le K$. *Then, for any* k, *there exists an* $\epsilon_k > 0$ *such that* $\varphi_k(x) > 0$ *for all* $x \in (1-\epsilon_k, 1)$.

(b) *Fix any* $x \in (1 - \min_{k\in[K]} \epsilon_k, 1)$. *Then* $(M_k(x) + E)$ *is invertible for every* $1 \le k \le K$, *provided*

$$\|E\|_2 < \min_{1\le k\le K} \varphi_k(x).$$

Proof. (of Lemma 2.6.3)

(a) Let $D_k(x) = \det(M_k(x))$. Note that $D_k(0) = 1$, thus $D_k \neq 0$ identically. Now observe that D_k is a polynomial in x, of degree m. Thus there indeed exists an $\epsilon_k > 0$ such that $D_k(x) \neq 0$ for $x \in (1-\epsilon_k, 1)$. This yields $\varphi_k(x) > 0$ whenever $x \in (1-\epsilon_k, 1)$ as well.

(b) Fix any $M \in \mathbb{R}^{m\times m}$ with $\operatorname{rank}(M) = m$. Let $E \in \mathbb{R}^{m\times m}$ satisfy $\operatorname{rank}(M+E) < m$. We claim $\|E\|_2 \ge \sigma_{\min}(M)$. To see this, note that if $M + E$ is rank-deficient, then there exists a v with $\|v\|_2 = 1$ such that $(M+E)v = 0$. This yields $Ev = -Mv$, thus

$$\|E\|_2 \ge \|Ev\|_2 = \|Mv\|_2 \ge \sigma_{\min}(M).$$

□

We now return to the proof, where in the remainder we will make use of the quantities defined in Lemma 2.6.3. We express $\bar{\Sigma} = \widehat{\Sigma} + E$. Here, $\widehat{\Sigma} \in \mathbb{R}^{m\times m}$ with unit diagonal entries, and $\widehat{\Sigma}_{ij} = \beta$ if $\left\langle \sigma^{(i)}, \sigma^{(j)} \right\rangle > 0$, and $\widehat{\Sigma}_{ij} = -\beta$ otherwise. The matrix $E \in \mathbb{R}^{m\times m}$ is such that $E_{ii} = 0$ for $1 \le i \le m$; and $|E_{ij}| \le \eta$ for $1 \le i < j \le m$. Note that,

$$\|E\|_F^2 = \sum_{1\le i\le m} \sum_{1\le j\le m} E_{ij}^2 \le m^2\eta^2.$$

Since $\|E\|_2 \le \|E\|_F$, we then conclude $\|E\|_2 \le m\eta$. We now choose $\beta \in (1 - \min_{1\le k\le K} \epsilon_k, 1)$ where the constants ϵ_k are defined in Lemma 2.6.3, and set

$$\eta(\beta) = \frac{\min_{1\le k\le K} \varphi_k(\beta)}{Nm},$$

where N is a (large) positive integer, to be tuned. Using Lemma 2.6.3, we have $\eta(\beta) > 0$ for every $\beta \in (1 - \min_k \epsilon_k, 1)$, and any $N \in \mathbb{Z}_+$. Furthermore, Lemma 2.6.3(b) also yields that under this choice of parameters, $\bar{\Sigma}$ is always invertible. $\bar{\Sigma}$ is, by construction, a covariance matrix thus has non-negative eigenvalues $\lambda_1(\bar{\Sigma}) \ge \cdots \ge \lambda_m(\bar{\Sigma}) > 0$ with

$$m = \operatorname{trace}(\bar{\Sigma}) \ge m\lambda_m(\bar{\Sigma}).$$

Thus we have

$$\lambda_m(\bar{\Sigma}) \le 1. \tag{2.36}$$

Fix now $\tau_1, \ldots, \tau_m \in \mathcal{I} \subset [0,1]$, and recall that $\gamma_i = \sqrt{1-\tau_i^2}$, $1 \le i \le m$. We now express the covariance matrix Σ in terms of $\bar{\Sigma}$, which depends only on m, β, and η.

To that end, let $A = \text{diag}(\gamma_1, \ldots, \gamma_m) \in \mathbb{R}^{m\times m}$ be a diagonal matrix. Observe that,

$$\Sigma = A\bar{\Sigma}A + (I - A^2). \tag{2.37}$$

Observe that as $\bar{\Sigma}$ is positive semidefinite, so do $A\bar{\Sigma}A$. Furthermore, as $1 - \gamma_i^2 \geq 0$ for $1 \leq i \leq m$, the matrix $I - A^2$ is positive semidefinite as well. We now study the smallest eigenvalue $\lambda_m(\Sigma)$.

Lemma 2.6.4. *For any choices of $\tau_1, \ldots, \tau_m \in \mathcal{I}$, it is the case that $\lambda_m(\Sigma) \geq \lambda_m(\bar{\Sigma})$. Hence,*

$$|\Sigma| \geq (\lambda_m(\bar{\Sigma}))^m > 0,$$

which is independent of the indices $\tau_1, \ldots, \tau_m$.

Proof. Recall the Courant-Fischer-Weyl variational characterization of the smallest singular value $\lambda_m(\Sigma)$ of a Hermitian matrix $\Sigma \in \mathbb{R}^{m\times m}$ [170]:

$$\lambda_m(\Sigma) = \inf_{v:\|v\|_2=1} v^T\Sigma v. \tag{2.38}$$

Then for any v with $\|v\|_2 = 1$,

$$\begin{aligned} v^T\Sigma v &= v^T(I - A^2)v + v^T A\bar{\Sigma}Av \\ &\geq \sum_{1\leq i\leq m} (1 - \gamma_i^2)v_i^2 + \lambda_m(\bar{\Sigma})\|Av\|_2^2 \\ &= \sum_{1\leq i\leq m} (1 - \gamma_i^2 + \lambda_m(\bar{\Sigma})\gamma_i^2)v_i^2 \\ &\geq \lambda_m(\bar{\Sigma})\|v\|_2^2 \\ &= \lambda_m(\bar{\Sigma}), \end{aligned}$$

where the first equality uses (2.37), the first inequality uses (2.38), and the last inequality uses $\lambda_m(\bar{\Sigma}) \leq 1$ as established in (2.36). Taking the infimum over all unit norm v, we conclude

$$\lambda_m(\Sigma) \geq \lambda_m(\bar{\Sigma}).$$

Finally

$$|\Sigma| = \prod_{1\leq j\leq m} \lambda_j(\Sigma) \geq \lambda_m(\Sigma)^m \geq \lambda_m(\bar{\Sigma})^m,$$

as desired. □

By the Lemma 2.6.4, we have that $|\Sigma|$ is bounded away from zero by an explicit constant controlled solely by m, β, η, which in particular is independent of $\mathcal{I}$. Thus the union bound leading to (2.33) is indeed valid.

We finally show how (2.35) is fulfilled, which is to ensure

$$\psi_N(\beta) \triangleq \beta - \eta(\beta) = \beta - \frac{\min_{1\leq k\leq K} \varphi_k(\beta)}{Nm} > 1 - h^{-1}\left(\frac{\epsilon}{2}\right).$$

Notice $1 - h^{-1}(\epsilon/2)$ is strictly smaller than 1. Observing now that $\psi_N(\beta) \to 1$ as $N \to +\infty$, and $\beta \to 1$, one can indeed find such β and η. It suffices that β satisfies

$$1 > \beta > \max\left\{1 - \min_{1\le k\le K} \epsilon_k, 1 - \frac{1}{2}h^{-1}\left(\frac{\epsilon}{2}\right)\right\}.$$

Having selected this value of $\beta > 0$, prescribe now $\eta \triangleq \eta(\beta)$, by choosing $N \in \mathbb{Z}_+$ sufficiently large so that

$$\eta = \frac{\min_{1\le k\le K} \varphi_k(\beta)}{Nm} < \frac{1}{2}h^{-1}\left(\frac{\epsilon}{2}\right).$$

This concludes the proof of Theorem 2.2.3. □

2.6.4 Proof of Theorem 2.2.5

The proof of Theorem 2.2.5 is based on the so-called *second moment method*, but in addition uses several other ideas. We provide a short outline below for convenience.

Outline of the Proof of Theorem 2.2.5

- Fix an $m \in \mathbb{N}$, $\rho \in (0,1)$, and a function $f : \mathbb{N} \to \mathbb{R}^+$ with $f(n) \in o(n)$. We first show that with high probability over $X \stackrel{d}{=} \mathcal{N}(0, I_n)$, there exists an $m-$tuple $\left(\sigma^{(i)} : 1 \le i \le m\right)$ of spin configurations $\sigma^{(i)} \in \mathcal{B}_n$ such that i) for $1 \le i \le m$, $n^{-1/2}|\left\langle \sigma^{(i)}, X\right\rangle| \le 2^{-f(n)}$; and ii) for $1 \le i < j \le m$, $\rho - \bar\rho \le n^{-1}\left\langle \sigma^{(i)}, \sigma^{(j)}\right\rangle \le \rho + \bar\rho$, provided $\bar\rho$ is sufficiently small.

- To start with, it is not even clear if for every $\bar\rho$ sufficiently small; there exists—deterministically—$\sigma^{(i)} \in \mathcal{B}_n$, $1 \le i \le m$, such that $\rho - \bar\rho \le n^{-1}\left\langle \sigma^{(i)}, \sigma^{(j)}\right\rangle \le \rho + \bar\rho$ for $1 \le i < j \le m$. We establish this using the so-called *probabilistic method* [13]: we assign the coordinates $\sigma^{(i)}(j) \in \{-1, 1\}$, $1 \le i \le m$ and $1 \le j \le n$ randomly according to the Rademacher distribution with an appropriate parameter[1]; and show that with positive probability, such a configuration exists.

- We then let the random variable M to count the number of $m-$tuples $\left(\sigma^{(i)} : 1 \le i \le m\right)$ of spin configurations which satisfy the desired properties. Our goal is to establish $\mathbb{P}(M \ge 1) = 1 - o_n(1)$. For this goal, we use the so-called *second moment method* which uses the *Paley-Zygmund inequality*: for a non-negative random variable M taking integer values,

$$\mathbb{P}(M \ge 1) \ge \frac{(\mathbb{E}[M]^2)}{\mathbb{E}[M^2]}.$$

  Namely, if the second moment $\mathbb{E}[M^2]$ is asymptotically $\mathbb{E}[M]^2(1 + o_n(1))$, in

[1]Here, we interpret the Rademacher distribution with parameter p as the distribution supported on $\{-1, 1\}$, which takes the value $+1$ with probability p.

other words when $\mathrm{Var}(M) = o\left(\mathbb{E}\left[M\right]^2\right)$, we have that $\mathbb{P}(M \geq 1) = 1 - o_n(1)$, as desired.

- As is rather common with the applications of the second moment method, the computation of the second moment is challenging: it involves an expectation of a sum running over pairs of $m-$tuples of spin configurations. To compute this sum, we employ an overcounting idea.

- To that end, fix an $\epsilon > 0$ small; and let I_ϵ be the set of all integers in the set $[0, n(1-\epsilon)/2] \cup [n(1+\epsilon)/2, 1]$. We now overestimate $\mathbb{E}\left[M^2\right]$ by dividing the sum into two components. Specifically, for two $m-$tuples of spin configurations $\mathcal{T} = \left(\sigma^{(i)} : 1 \leq i \leq m\right)$ and $\overline{\mathcal{T}} = \left(\overline{\sigma}^{(i)} : 1 \leq i \leq m\right)$, we distinguish two cases. The first case pertains the pairs $\left(\mathcal{T}, \overline{\mathcal{T}}\right)$ for which there exists an i, j such that $d_H\left(\sigma^{(i)}, \overline{\sigma}^{(j)}\right) \in I_\epsilon$. The second case pertains the pairs $\left(\mathcal{T}, \overline{\mathcal{T}}\right)$ for which it is the case that for every i, j; $n(1-\epsilon)/2 < d_H\left(\sigma^{(i)}, \overline{\sigma}^{(j)}\right) < n(1+\epsilon)/2$. Namely, the second case essentially corresponds to the pairs of $m-$tuples that are nearly "uncorrelated". The term ϵ introduced above essentially controls the "residual correlation".

- We then find that due to cardinality constraints (via a certain asymptotics pertaining the binomial coefficients), the number of pairs of first kind is small, and the probability term can be neglected. See the proof for details.

- We then observe that the number of pairs of $m-$tuples of second kind dominates the second moment term. For those pairs, however, the computation of their joint probability is tractable due to the fact that they are nearly uncorrelated.

- We then take a union bound over a certain choice of grid, for the goal of obtaining the event which involves a condition over **all** $\beta \in [0, 1]$.

- Finally, sending $n \to \infty$ and $\epsilon \to 0$ carefully; we obtain our desired conclusion.

We now provide the complete proof.

Proof of Theorem 2.2.5

Proof. In what follows, denote by $S(m, \rho, \bar{\rho}, E)$ to be the set of all $m-$tuples $\left(\sigma^{(i)} : 1 \leq i \leq m\right)$ of spin configurations $\sigma^{(i)} \in \mathcal{B}_n$ such that

- For every $1 \leq i < j \leq m$, $\rho - \bar{\rho} \leq \mathcal{O}\left(\sigma^{(i)}, \sigma^{(j)}\right) \leq \rho + \bar{\rho}$.

- For every $1 \leq i \leq m$, $n^{-1/2}\left|\left\langle \sigma^{(i)}, X \right\rangle\right| \leq E$.

Namely, $S(m, \rho, \bar{\rho}, E)$ is a shorthand for the set $\mathcal{S}(m, \rho, \bar{\rho}, \log_2 E, \{0\})$ introduced in Definition 2.2.1 with the modification as in Theorem 2.2.5.

Let $m \in \mathbb{N}$, $\gamma \in (0, \frac{1}{2})$, and $\delta \in (0, \gamma)$. Define the set

$$S(m,\gamma,\delta) \triangleq \left\{ \left(\sigma^{(i)} : 1 \leq i \leq m\right) : \sigma^{(i)} \in \mathcal{B}_n, \frac{1}{n} d_H\left(\sigma^{(i)}, \sigma^{(i')}\right) \in [\gamma - \delta, \gamma + \delta], 1 \leq i < i' \leq m \right\}. \tag{2.39}$$

Call the triple (m, γ, δ) **admisible** if there exists an $N \triangleq N(m, \gamma, \delta) \in \mathbb{N}$ such that for every $n \geq N$, $S(m, \gamma, \delta) \neq \varnothing$. We first prove that for any fixed $m \in \mathbb{N}$, γ, and δ sufficiently small; $S(m, \gamma, \delta) \neq \varnothing$ for all sufficiently large n.

Note that this step is necessary: in order to ensure the the existence of $m-$tuples with desired "energy levels" as required by the Theorem, one needs to ensure first that such $m-$tuples of spin configurations with pairwise constrained overlaps do exist; and this is quite non-trivial for $m > 2$. We will later translate the condition on Hamming distances into a condition on their pairwise (normalized) overlaps.

$S(m, \gamma, \delta) \neq \varnothing$ for n sufficiently large. We choose the spin configurations $\sigma^{(i)} \in \mathcal{B}_n$ randomly. Specifically, let $\sigma^{(i)} = \left(\sigma^{(i)}(j) : 1 \leq j \leq n\right) \in \mathcal{B}_n$ be i.i.d. across $1 \leq i \leq m$ and $1 \leq j \leq n$ with

$$\mathbb{P}\left(\sigma^{(i)}(j) = 1\right) = \eta^*,$$

where $\eta^* \in (0, 1)$ is chosen so that

$$\eta^* \left(1 - \eta^*\right) = \frac{1}{2}\gamma.$$

Define now a sequence $\mathcal{E}_{i,i'}$ of events $1 \leq i < i' \leq m$,

$$\mathcal{E}_{i,i'} \triangleq \left\{ \gamma - \delta \leq \frac{1}{n} d_H\left(\sigma^{(i)}, \sigma^{(i')}\right) \leq \gamma + \delta \right\}.$$

It suffices to establish

$$\mathbb{P}\left(\bigcap_{1 \leq i < i' \leq m} \mathcal{E}_{i,i'} \right) > 0 \Leftrightarrow \mathbb{P}\left(\bigcup_{1 \leq i < i' \leq m} \mathcal{E}^c_{i,i'} \right) < 1.$$

We next study $\mathbb{P}\left(\mathcal{E}^c_{1,2}\right)$. Define $Z_j \triangleq \mathbb{1}\left\{\sigma^{(1)}(j) \neq \sigma^{(2)}(j)\right\}$, $1 \leq j \leq n$. Note that Z_j are i.i.d. Bernoulli variables with mean $\mathbb{E}[Z_j] = 2\eta^*\left(1 - \eta^*\right) = \gamma$. Using now standard concentration results for the sum of i.i.d. Bernoulli variables [281], we have that for any $\delta > 0$

$$\mathbb{P}\left(\mathcal{E}^c_{1,2}\right) = \mathbb{P}\left(\left| \frac{1}{n} \sum_{1 \leq j \leq n} Z_j - \gamma \right| > \delta \right) \leq \exp\left(-Cn\delta^2\right),$$

for an absolute constant $C > 0$. Now, the events $\mathcal{E}_{i,i'}$ are clearly equiprobable across

$1 \le i < i' \le m$. Applying a union bound,

$$\mathbb{P}\left(\bigcup_{1\le i<i'\le m} \mathcal{E}^c_{i,i'}\right) \le \binom{m}{2} \exp\left(-Cn\delta^2\right).$$

Since m is constant, the claim follows.

Fix an arbitrary $\rho \in (0,1)$; a "proxy" for β appearing in the statement of Theorem 2.2.5. Suppose $\bar{\rho}$ is a sufficiently small parameter; a "proxy" for η. Set

$$\gamma \triangleq \frac{1-\rho}{2} \in (0, \frac{1}{2}) \quad \text{and} \quad \delta \triangleq \frac{\bar{\rho}}{2}. \tag{2.40}$$

Observe that

$$\rho - \bar{\rho} \le \frac{1}{n}\left\langle \sigma, \sigma' \right\rangle \le \rho + \bar{\rho} \iff \gamma - \delta \le \frac{1}{n} d_H\left(\sigma, \sigma'\right) \le \gamma + \delta.$$

In what follows, we define certain sets and random variables; which depend on n but is dropped in the notation. Recall the set $S(m,\gamma,\delta)$ from (2.39). As we established, for every $m \in \mathbb{N}$, and γ, $S(m,\gamma,\delta) \neq \varnothing$ for all δ sufficiently small and all n large. Define L_σ to be the cardinality of set

$$S_\sigma = \left\{\left(\sigma^{(i)} : 1 \le i \le m\right) \in S(m,\gamma,\delta) : \sigma^{(1)} = \sigma\right\} \subset S(m,\gamma,\delta).$$

Note that L_σ is independent of σ. So we instead use the notation L for

$$L \triangleq \left|\left\{\left(\sigma^{(i)} : 1 \le i \le m\right) \in S(m,\gamma,\delta) : \sigma^{(1)} = \sigma\right\}\right|. \tag{2.41}$$

We then have

$$|S(m,\gamma,\delta)| = 2^n L. \tag{2.42}$$

Fix $f : \mathbb{N} \to \mathbb{R}^+$, the "energy exponent" with sub-linear growth, $f(n) \in o(n)$; and let $E = 2^{-f(n)}$. Consider

$$M \triangleq M(m,\gamma,\delta,E) = \sum_{\left(\sigma^{(i)}:1\le i\le m\right)\in S\left(m,\gamma,\delta\right)} \mathbb{1}\left\{|Y_i| < 2^{-f(n)}, 1 \le i \le m\right\}, \tag{2.43}$$

where

$$Y_i \triangleq \frac{1}{\sqrt{n}}\left\langle \sigma^{(i)}, X\right\rangle, \quad 1 \le i \le m, \tag{2.44}$$

and, $X \overset{d}{=} \mathcal{N}(0, I_n)$. Then $Y_i \overset{d}{=} \mathcal{N}(0,1)$, $1 \le i \le m$, though not independent.

Namely, for $\gamma = \frac{1-\rho}{2}$, $\delta = \frac{\bar{\rho}}{2}$, and $E = 2^{-f(n)}$, $M(m,\gamma,\delta,E)$ is a lower bound on the cardinality of the set $S(m,\rho,\bar{\rho},E)$ that we study in Theorem 2.2.5. In what follows, we study $\mathbb{P}(M \ge 1)$ and give a lower bound on it for an appropriate choice of parameters.

Second moment method. We recall the Paley-Zygmund inequality: if $M \geq 0$ is a non-negative integer valued random variable, then

$$\mathbb{P}(M \geq 1) = \mathbb{P}(M > 0) \geq \frac{(\mathbb{E}[M])^2}{\mathbb{E}[M^2]}. \tag{2.45}$$

For a short proof, see [58, Exercise 2.4]. In particular, to show $\mathbb{P}(M \geq 1) = 1 - o_n(1)$, it suffices to establish

$$\mathbb{E}\left[M^2\right] = \mathbb{E}[M]^2 (1 + o_n(1)).$$

First moment computation. Fix any $\left(\sigma^{(i)} : 1 \leq i \leq m\right) \in S(m, \gamma, \delta)$ (2.39); and recall Y_i, $1 \leq i \leq m$, from (2.44). To compute the joint probability, we first recover the structure of the covariance matrix $\Sigma \in \mathbb{R}^{m \times m}$. $\Sigma_{ii} = 1$ for $1 \leq i \leq m$; and

$$\mathbb{E}[Y_i Y_j] = \frac{1}{n} \left\langle \sigma^{(i)}, \sigma^{(j)} \right\rangle.$$

Since $\left(\sigma^{(i)} : 1 \leq i \leq m\right) \in S(m, \gamma, \delta)$, it follows that for any $1 \leq i < j \leq m$, $\Sigma_{ij} = \Sigma_{ji} \in [\rho - \bar{\rho}, \rho + \bar{\rho}]$. In particular,

$$\Sigma = (1 - \rho) I_m + \rho 1 1^T + E,$$

where $E \in \mathbb{R}^{m \times m}$ is a perturbation matrix with $E_{ii} = 0$ and $|E_{ij}| = |E_{ji}| \leq \bar{\rho}$ for $1 \leq i < j \leq m$. The spectrum of $(1-\rho) I_m + \rho 1 1^T$ consists of the eigenvalue $1 + \rho(m-1)$ with multiplicity one; and the eigenvalue $1 - \rho$ with multiplicity $m - 1$. Clearly $\|E\|_2 \leq \|E\|_F \leq m\bar{\rho}$. Using now Wielandt-Hoffman Theorem (Theorem 2.6.2), we have that for $\bar{\rho} \ll \frac{1-\rho}{m}$, the matrix Σ is invertible. In what follows, assume $\bar{\rho}$ is in this regime.

With this, we compute that for energy level $E = 2^{-f(n)}$ (with exponent $f(n) \in \omega_n(1) \cap o(n)$),

$$\mathbb{P}\left(|Y_i| < 2^{-f(n)}, 1 \leq i \leq m\right) = \frac{1}{(2\pi)^{\frac{m}{2}} |\Sigma|^{\frac{1}{2}}} \int_{\mathbf{y} \triangleq (y_1, \ldots, y_m) \in [-E, E]^m} \exp\left(-\frac{1}{2} \mathbf{y}^T \Sigma^{-1} \mathbf{y}\right) d\mathbf{y}.$$

Now, observe that for $\mathbf{y} \in [-E, E]^m$, $\exp\left(-\frac{1}{2} \mathbf{y}^T \Sigma^{-1} \mathbf{y}\right) = 1 + o_n(1)$, provided Σ^{-1} is invertible (which we ensured). Under this condition,

$$\mathbb{P}\left(|Y_i| < 2^{-f(n)}, 1 \leq i \leq m\right) = \frac{2^m}{(2\pi)^{\frac{m}{2}} |\Sigma|^{\frac{1}{2}}} E^m (1 + o_n(1)).$$

Equipped with this, we now give two expressions for the first moment. First, using (2.43) and the linearity of expectations, we obtain

$$\mathbb{E}[M] = \sum_{\left(\sigma^{(i)} : 1 \leq i \leq m\right) \in S(m, \gamma, \delta)} \frac{2^m}{(2\pi)^{\frac{m}{2}} |\Sigma|^{\frac{1}{2}}} E^m (1 + o_n(1)). \tag{2.46}$$

Here, the tuple $\left(\sigma^{(i)} : 1 \le i \le m\right)$ induces an "overlap pattern", which, in turn, induces the covariance matrix Σ.

For the second, note that using Wieland-Hoffman inequality, it is the case that for $\bar{\rho} \ll \frac{1-\rho}{m}$, there exists constants $C_1 < C_2$—depending only on $m, \rho, \bar{\rho}$ and are independent of n—such that

$$C_1 < |\Sigma| < C_2.$$

Namely, $|\Sigma| = O_n(1)$ across all $m-$tuples $\left(\sigma^{(i)} : 1 \le i \le m\right) \in S(m, \gamma, \delta)$. With this, we have

$$\mathbb{E}[M] \ge 2^n L \frac{2^m}{(2\pi)^{\frac{m}{2}} C_2} E^m \left(1 + o_n(1)\right). \tag{2.47}$$

Above, we utilized the cardinality bound (2.42).

Second moment computation. The computation for the second moment is more delicate, and involves a sum over pairs of $m-$tuples of spin configurations. For notational purposes, let $\mathcal{T} \triangleq \left(\sigma^{(i)} : 1 \le i \le m\right)$ and $\overline{\mathcal{T}} \triangleq \left(\overline{\sigma}^{(i)} : 1 \le i \le m\right)$. We have

$$\mathbb{E}\left[M^2\right] = \sum_{\mathcal{T}, \overline{\mathcal{T}} \in S(m,\gamma,\delta)} \mathbb{P}\left(|Y_i| \le E, \left|\overline{Y_i}\right| \le E, 1 \le i \le m\right). \tag{2.48}$$

Here the following variables are standard normal

$$Y_i \triangleq \frac{1}{\sqrt{n}} \left\langle \sigma^{(i)}, X \right\rangle \quad \text{and} \quad \overline{Y_i} \triangleq \frac{1}{\sqrt{n}} \left\langle \overline{\sigma}^{(i)}, X \right\rangle, \qquad 1 \le i \le m. \tag{2.49}$$

Now, fix an arbitrary $\epsilon > 0$, and define the set

$$I_\epsilon = \mathbb{Z} \cap \left(\left[0, \frac{n}{2}(1-\epsilon)\right] \cup \left[\frac{n}{2}(1+\epsilon), n\right] \right). \tag{2.50}$$

For the pairs $(\mathcal{T}, \overline{\mathcal{T}}) \in S(m, \gamma, \delta) \times S(m, \gamma, \delta)$ of spin configurations define the following family of m^2 sets

$$S^{(ij)}(\epsilon) \triangleq \left\{ \left(\mathcal{T}, \overline{\mathcal{T}}\right) \in S(m, \gamma, \delta) \times S(m, \gamma, \delta) : d_H\left(\sigma^{(i)}, \overline{\sigma}^{(j)}\right) \in I_\epsilon \right\} \tag{2.51}$$

for $1 \le i, j \le m$. Let

$$\overline{S} \triangleq \left(S(m, \gamma, \delta) \times S(m, \gamma, \delta)\right) \setminus \left(\bigcup_{1 \le i,j \le m} S^{(ij)}(\epsilon) \right). \tag{2.52}$$

Note that the sets $S^{(ij)}(\epsilon)$ potentially intersect for different pairs (i, j). This is the essence of the overcounting we utilize in the remainder, with the key idea being that the overcounting can only increase the second moment.

We next establish an upper bound on the cardinality of $S^{(ij)}(\epsilon)$. Recalling the

quantity L from (2.41), we have

$$\left|S^{(ij)}(\epsilon)\right| = 2^n L^2 \left(\sum_{k\in I_\epsilon} \binom{n}{k}\right).$$

The rationale for this is as follows. The first coordinate of $\mathcal{T}$ is chosen in 2^n different ways; and the remainder are filled in L different ways. Having fixed this $m-$tuple; now using the constraint $d_H\left(\sigma^{(i)}, \bar{\sigma}^{(j)}\right) \in I_\epsilon$, the object $\bar{\sigma}^{(j)}$ can be chosen in $\sum_{k\in I_\epsilon}\binom{n}{k}$ different ways; and finally having fixed $\overline{\sigma}^{(j)}$, the rest of the coordinates of the $m-$tuple $\overline{\mathcal{T}}$ can now be filled in L different ways.

Applying the Stirling's formula (2.18) and using $|I_\epsilon| = n^{O(1)}$

$$\sum_{k\in I_\epsilon}\binom{n}{k} \le n^{O(1)}\binom{n}{n\frac{1-\epsilon}{2}} = \exp_2\left(nh_b\left(\frac{1-\epsilon}{2}\right) + O(\log_2 n)\right).$$

Thus,

$$\left|S^{(ij)}(\epsilon)\right| \le 2^n L^2 \exp_2\left(nh_b\left(\frac{1-\epsilon}{2}\right) + O\left(\log_2 n\right)\right), \quad \text{for} \quad 1 \le i \ne j \le m. \quad (2.53)$$

Overcounting argument. Now, for any pair $(\mathcal{T}, \overline{\mathcal{T}}) \in S(m,\gamma,\delta)\times S(m,\gamma,\delta)$, let

$$p\left(\mathcal{T}, \overline{\mathcal{T}}\right) \triangleq \mathbb{P}\left(|Y_i| \le E, \left|\overline{Y_i}\right| \le E, 1 \le i \le m\right).$$

We now compute the second moment. In terms of the sets introduced in (2.51) (2.52), we have

$$\begin{aligned}\mathbb{E}\left[M^2\right] &= \sum_{(\mathcal{T},\overline{\mathcal{T}})\in S(m,\gamma,\delta)\times S(m,\gamma,\delta)} p\left(\mathcal{T}, \overline{\mathcal{T}}\right) \\ &\le \sum_{1\le i,j\le m}\sum_{(\mathcal{T},\overline{\mathcal{T}})\in S^{(ij)}(\epsilon)} p\left(\mathcal{T}, \overline{\mathcal{T}}\right) + \sum_{(\mathcal{T},\overline{\mathcal{T}})\in \overline{S}} p\left(\mathcal{T}, \overline{\mathcal{T}}\right).\end{aligned}$$

Consequently,

$$\mathbb{E}\left[M^2\right] \le \underbrace{m^2 2^n L^2 \exp_2\left(nh_b\left(\frac{1-\epsilon}{2}\right) + O\left(\log_2 n\right)\right)}_{\triangleq A_\epsilon} + \underbrace{\sum_{(\mathcal{T},\overline{\mathcal{T}})\in\overline{S}} p\left(\mathcal{T}, \overline{\mathcal{T}}\right)}_{\triangleq B_\epsilon}. \quad (2.54)$$

Study of the A_ϵ term. Using the crude lower bound (2.47) on the first moment, we arrive at

$$\begin{aligned}\frac{A_\epsilon}{\mathbb{E}\left[M\right]^2} &\leq \frac{m^2 2^n L^2 \exp_2\left(nh_b\left(\frac{1-\epsilon}{2}\right)+O\left(\log_2 n\right)\right)}{2^{2n}L^2 2^{2m}\left(2\pi\right)^{-m}C_2^{-2}E^{2m}\left(1+o_n(1)\right)} \\ &= \exp_2\left(-n+nh_b\left(\frac{1-\epsilon}{2}\right)-2m\log_2 E+O\left(\log_2 n\right)+O(1)\right) \\ &= \exp_2\left(-n\left(1-h_b\left(\frac{1-\epsilon}{2}\right)\right)-2mf(n)+O(\log_2 n)\right) \\ &= \exp_2\left(-n\left(1-h_b\left(\frac{1-\epsilon}{2}\right)\right)+o(n)\right).\end{aligned}$$

Above, we used the fact that $E = 2^{-f(n)}$ for some exponent $f(n) \in o(n)$; and the fact $\epsilon > 0$ hence $h_b\left(\frac{1-\epsilon}{2}\right) < 1$. Consequently,

$$\frac{A_\epsilon}{\mathbb{E}\left[M\right]^2} \leq \exp_2\left(-\Theta(n)\right). \tag{2.55}$$

Study of the B_ϵ term. This term is more involved, and it is the one that leads to the dominant contribution to the second moment of M.

Fix a pair $(\mathcal{T}, \overline{\mathcal{T}}) \in S(m,\gamma,\delta) \times S(m,\gamma,\delta)$, and recall the associated standard normal variables $Y_i, \overline{Y_i}$, $1 \leq i \leq n$ per (2.49). Our goal is to study the probability

$$p\left(\mathcal{T}, \overline{\mathcal{T}}\right) = \mathbb{P}\left(\left|Y_i\right|, \left|\overline{Y_i}\right| \leq \epsilon, 1 \leq i \leq m\right).$$

To that end, fix $1 \leq i, j \leq m$. We study the covariance between Y_i and $\overline{Y_j}$. Fixing a pair $(\mathcal{T}, \overline{\mathcal{T}}) \in \overline{S}$, we have

$$d_H\left(\sigma^{(i)}, \overline{\sigma}^{(j)}\right) \in \left[\frac{n}{2}(1-\epsilon), \frac{n}{2}(1+\epsilon)\right], \quad \text{for all} \quad i, j.$$

Then,

$$\frac{1}{n}\left\langle \sigma^{(i)}, \overline{\sigma}^{(j)}\right\rangle \in [-\epsilon, \epsilon].$$

Let $\Sigma_\epsilon \in \mathbb{R}^{2m\times 2m}$ be the covariance matrix for the random vector $(Y_1, \ldots, Y_m, \overline{Y_1}, \ldots, \overline{Y_m})$. Observe that it has the following "block" structure:

$$\Sigma_\epsilon = \left(\begin{array}{c|c} \Sigma_{\mathcal{T}} & E \\ \hline E & \Sigma_{\overline{\mathcal{T}}} \end{array}\right) \in \mathbb{R}^{2m\times 2m}.$$

Here, $\Sigma_{\mathcal{T}} \in \mathbb{R}^{m\times m}$ is the covariance matrix corresponding to the random vector $(Y_1, \ldots, Y_m)$; $\Sigma_{\overline{\mathcal{T}}} \in \mathbb{R}^{m\times m}$ is the covariance matrix corresponding to the random vector $(\overline{Y_1}, \ldots, \overline{Y_m})$; and $E \in \mathbb{R}^{m\times m}$ is given by $E_{ij} = \mathbb{E}\left[Y_i \overline{Y_j}\right]$. We have that for $1 \leq i \leq m$,

$$\left(\Sigma_{\mathcal{T}}\right)_{ii} = 1 = \left(\Sigma_{\overline{\mathcal{T}}}\right)_{ii},$$

and for $1 \leq i < j \leq m$,

$$(\Sigma_{\mathcal{T}})_{ij}\,, \quad (\Sigma_{\overline{\mathcal{T}}})_{ij} \quad \in [\rho - \bar{\rho}, \rho + \bar{\rho}].$$

Moreover, for $1 \leq i < j \leq m$,

$$|E_{ij}| = |E_{ji}| \leq \epsilon.$$

We now invert the 2×2 block matrix Σ_ϵ, while keeping in mind that the block, $\Sigma_{\mathcal{T}}$, is invertible. Observe that

$$\left(\begin{array}{c|c} \Sigma_{\mathcal{T}} & E \\ \hline E & \Sigma_{\overline{\mathcal{T}}} \end{array}\right) \left(\begin{array}{c|c} I & -\Sigma_{\mathcal{T}}^{-1}E \\ \hline O & I \end{array}\right) = \left(\begin{array}{c|c} \Sigma_{\mathcal{T}} & O \\ \hline E & \Sigma_{\overline{\mathcal{T}}} - E\Sigma_{\mathcal{T}}^{-1}E \end{array}\right).$$

With this decomposition and the fact that the determinant of a "block triangular" matrix is the product of the determinants of blocks constituting the diagonal, we arrive at

$$\begin{aligned} |\Sigma_\epsilon| &= |\Sigma_{\mathcal{T}}| \cdot \left|\Sigma_{\overline{\mathcal{T}}} - E\Sigma_{\mathcal{T}}^{-1}E\right| \\ &= |\Sigma_{\mathcal{T}}| \cdot |\Sigma_{\overline{\mathcal{T}}}| \cdot \left|I - \Sigma_{\overline{\mathcal{T}}}^{-1}E\Sigma_{\mathcal{T}}^{-1}E\right|, \end{aligned}$$

where we have used the fact that $\Sigma_{\widehat{\mathcal{T}}}$ is invertible as well, to pull the term outside.

Note that for ϵ sufficiently small; the determinant $\left|I - \Sigma_{\overline{\mathcal{T}}}^{-1}E\Sigma_{\mathcal{T}}^{-1}E\right|$ is non-zero over all choices of $\Sigma_{\mathcal{T}}$ and $\Sigma_{\overline{\mathcal{T}}}$. Namely, provided ϵ is small; Σ_ϵ is invertible uniformly across all $\Sigma_{\mathcal{T}}$ and $\Sigma_{\overline{\mathcal{T}}}$. In the remainder, assume $\epsilon > 0$ though sufficiently small. We now define the object

$$\overline{\varphi}(\epsilon) \triangleq \overline{\varphi}\left(\Sigma_{\mathcal{T}}, \Sigma_{\overline{\mathcal{T}}}, E\right) = \left|I - \Sigma_{\overline{\mathcal{T}}}^{-1}E\Sigma_{\mathcal{T}}^{-1}E\right|.$$

Note that $\overline{\varphi}(\cdot)$ is a polynomial in the entries E_{ij}, $1 \leq i, j \leq m$; as well as in the entries of matrices $\Sigma_{\mathcal{T}}$ and $\Sigma_{\overline{\mathcal{T}}}$. Furthermore, $\overline{\varphi} \to 1$ as $\epsilon \to 0$. With this we write

$$|\Sigma_\epsilon| = |\Sigma_{\mathcal{T}}|\,|\Sigma_{\overline{\mathcal{T}}}|\,\overline{\varphi}(\epsilon). \tag{2.56}$$

We now compute

$$\begin{aligned} p\left(\mathcal{T}, \overline{\mathcal{T}}\right) &= (2\pi)^{-m}\,|\Sigma_\epsilon|^{-\frac{1}{2}} \int_{\mathbf{y}=(y_1,\ldots,y_m,\overline{y_1},\ldots,\overline{y_m})\in[-E,E]^{2m}} \exp\left(-\frac{1}{2}\mathbf{y}^T\Sigma_\epsilon^{-1}\mathbf{y}\right)\, d\mathbf{y} \\ &= (2\pi)^{-m}\,|\Sigma_\epsilon|^{-\frac{1}{2}}\, 2^{2m}E^{2m}\,(1 + o_n(1)) \\ &= \left(\overline{\varphi}(\epsilon)\right)^{-\frac{1}{2}}(1 + o_n(1))\left((2\pi)^{-\frac{m}{2}}\,|\Sigma_{\mathcal{T}}|^{-\frac{1}{2}}\,(2E)^m\right)\left((2\pi)^{-\frac{m}{2}}\,|\Sigma_{\overline{\mathcal{T}}}|^{-\frac{1}{2}}\,(2E)^m\right). \end{aligned}$$

Here, the first line uses the definition of $p\left(\mathcal{T}, \overline{\mathcal{T}}\right)$ together with the formulae for the multivariate normal density; the second line uses the fact when $\mathbf{y} \in [-E, E]^{2m}$ then $\exp\left(-\frac{1}{2}\mathbf{y}^T\Sigma_\epsilon^{-1}\mathbf{y}\right) = 1 + o_n(1)$ provided Σ_ϵ is invertible (which we ensured); and the

third line uses (2.56). Thus,

$$B_\epsilon = \sum_{(\mathcal{T},\overline{\mathcal{T}})\in\overline{S}} p\left(\mathcal{T},\overline{\mathcal{T}}\right) = \overline{\varphi}(\epsilon)\left(1+o_n(1)\right) \sum_{(\mathcal{T},\overline{\mathcal{T}})\in\overline{S}} \left((2\pi)^{-\frac{m}{2}} \left|\Sigma_{\mathcal{T}}\right|^{-\frac{1}{2}} (2E)^m\right)\left((2\pi)^{-\frac{m}{2}} \left|\Sigma_{\overline{\mathcal{T}}}\right|^{-\frac{1}{2}} (2E)^m\right). \tag{2.57}$$

We now square the expression (2.46), keep only the terms corresponding to $\overline{S}$; and lower bound the square of the first moment

$$\mathbb{E}\left[M\right]^2 \geq \left(1+o_n(1)\right) \sum_{(\mathcal{T},\overline{\mathcal{T}})\in\overline{S}} \left((2\pi)^{-\frac{m}{2}} \left|\Sigma_{\mathcal{T}}\right|^{-\frac{1}{2}} (2E)^m\right)\left((2\pi)^{-\frac{m}{2}} \left|\Sigma_{\overline{\mathcal{T}}}\right|^{-\frac{1}{2}} (2E)^m\right). \tag{2.58}$$

Combining (2.57) and (2.58), we arrive at

$$\frac{B_\epsilon}{\mathbb{E}\left[M\right]^2} \leq \left(1+o_n(1)\right)\left(\overline{\varphi}(\epsilon)\right)^{-\frac{1}{2}}. \tag{2.59}$$

Applying Paley-Zygmund Inequality. Applying the Paley-Zygmund inequality (2.45),

$$\mathbb{P}\left(M\geq 1\right) \geq \frac{\mathbb{E}\left[M\right]^2}{\mathbb{E}\left[M^2\right]} \geq \frac{\mathbb{E}\left[M\right]^2}{A_\epsilon + B_\epsilon} = \frac{1}{\frac{A_\epsilon}{\mathbb{E}[M]^2} + \frac{B_\epsilon}{\mathbb{E}[M]^2}} \geq \frac{1}{\exp\left(-\Theta(n)\right) + \left(1+o_n(1)\right)\overline{\varphi}(\epsilon)^{-1/2}}. \tag{2.60}$$

Here, the second inequality uses the overcounting upper bound (2.54); and the third inequality uses the upper bounds (2.55) and (2.59).

Combining everything. The reasoning above remains valid if (a) $\rho > \bar{\rho}$ and (b) $\bar{\rho} \ll \frac{1-\rho}{m}$. Now, choose

$$\bar{\rho} = \frac{\eta}{1000m}.$$

Here, the choice of the constant 1000 is arbitrary. Next, let ℓ be the largest positive integer such that $2\ell\eta < 1-\eta$. Consider the "grid" $\rho_k = 2k\eta$, $1\leq k\leq \ell$, and intervals $I_k = [(2k-1)\eta, (2k+1)\eta] = [\rho_k - \eta, \rho_k + \eta]$ centered at ρ_k. Since

$$[\rho_k - \bar{\rho}, \rho_k + \bar{\rho}] \subset [\rho_k - \eta, \rho_k + \eta]$$

it follows by using (2.60) that

$$\mathbb{P}\left(S\left(m,\rho_k,\eta,E\right)\neq\varnothing\right) \geq \mathbb{P}\left(S\left(m,\rho_k,\bar{\rho},E\right)\neq\varnothing\right) \geq \frac{1}{\exp\left(-\Theta(n)\right) + \left(1+o_n(1)\right)\overline{\varphi}_k(\epsilon)^{-1/2}}, \tag{2.61}$$

where $\overline{\varphi}_k(\cdot)$ is a continuous function with the property that $\overline{\varphi}_k(\epsilon) \to 1$ as $\epsilon \to 0$. Taking a union bound over $1 \leq k \leq \ell$, we arrive at

$$\mathbb{P}\left(\underbrace{\bigcap_{1\leq k\leq \ell} \{S(m,\rho_k,\eta,E) \neq \varnothing\}}_{\triangleq \mathcal{E}_{\text{aux}}}\right) \geq 1 - \ell \frac{\exp(-\Theta(n)) + (1+o_n(1))\,\overline{\varphi}(\epsilon)^{-1/2} - 1}{\exp(-\Theta(n)) + (1+o_n(1))\,\overline{\varphi}(\epsilon)^{-1/2}}, \tag{2.62}$$

where $\overline{\varphi}(\cdot) = \min_{1\leq k\leq \ell} \overline{\varphi}_k(\cdot)$. In particular, since ℓ is finite, it follows $\overline{\varphi}(\epsilon) \to 1$ as $\epsilon \to 0$.

We now carefully send n and ϵ to their corresponding limits. Note that the asymptotic expressions (in n) given above are valid so long as $\epsilon > 0$—see, e.g. (2.55). Thus, we must send $n \to \infty$ first, while keeping $\epsilon > 0$ fixed. We clearly have

$$1 \geq \limsup_{n\to\infty} \mathbb{P}(\mathcal{E}_{\text{aux}}).$$

Furthermore, while keeping $\epsilon > 0$ and sending $n \to \infty$ in (2.62), we obtain

$$\liminf_{n\to\infty} \mathbb{P}(\mathcal{E}_{\text{aux}}) \geq 1 - \ell \cdot \frac{\overline{\varphi}(\epsilon)^{-1/2} - 1}{\overline{\varphi}(\epsilon)^{-1/2}}.$$

Note that the sequence $\{\mathbb{P}(\mathcal{E}_{\text{aux}})\}_{n\geq 1}$ (note that $\mathcal{E}_{\text{aux}}$ implicitly depends on n) is not a function of ϵ—and the lower bound holds true for every ϵ sufficiently close to zero. Moreover, ℓ is a constant. For this reason, we can now safely send $\epsilon \to 0$ to obtain

$$\liminf_{n\to\infty} \mathbb{P}(\mathcal{E}_{\text{aux}}) \geq 1.$$

Hence

$$1 \geq \limsup_{n\to\infty} \mathbb{P}(\mathcal{E}_{\text{aux}}) \geq \liminf_{n\to\infty} \mathbb{P}(\mathcal{E}_{\text{aux}}) \geq 1$$

implying

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{E}_{\text{aux}}) = 1.$$

Finally, observe that on high probability event $\mathcal{E}_{\text{aux}}$, it is the case that for each of $[\eta, 3\eta], [3\eta, 5\eta], \ldots$, there exists an $m-$tuple of spin configurations (with appropriate energy) whose pairwise overlaps are contained in the chosen interval. Since for each $\beta \in [0,1]$; $[\beta - 3\eta, \beta + 3\eta]$ contains a full interval $[(2k-1)\eta, (2k+1)\eta]$; we conclude that

$$\mathbb{P}\left(\forall \beta \in [0,1] : \mathcal{S}\left(m, \beta, 3\eta, 2^{-f(n)}\right) \neq \varnothing\right) \geq \mathbb{P}(\mathcal{E}_{\text{aux}}) = 1 - o_n(1).$$

The above reasoning remains true for every $\eta > 0$. Taking $\frac{\eta}{3}$ in place of η yields the desired conclusion.

□

2.6.5 Proof of Theorem 2.2.6

Case 1: $\omega\left(\sqrt{n\log_2 n}\right) \le E_n \le o(n)$**.**

Proof. Let $g(n)$ be an arbitrary function with growth

$$\omega(1) \le g(n) \le o\left(\frac{E_n^2}{n\log_2 n}\right).$$

We take m, β, and η per (2.4), that is

$$m = \frac{2n}{E_n}, \quad \beta = 1 - \frac{2g(n)}{E_n}, \quad \text{and} \quad \eta = \frac{g(n)}{2n}.$$

Define next several auxiliary parameters. First, set $\phi(n)$ by the expression

$$E_n = \phi(n)\sqrt{n\log n}. \tag{2.63}$$

Since $\omega\left(\sqrt{n\log_2 n}\right) \le E_n \le o(n)$, it holds that

$$\omega_n(1) \le \phi(n) \le o\left(\sqrt{\frac{n}{\log n}}\right). \tag{2.64}$$

Moreover, in terms of $\phi(\cdot)$, the growth condition on g translates as

$$\omega_n(1) \le g(n) \le o\left(\phi(n)^2\right). \tag{2.65}$$

Introduce another parameter ν_n via

$$\nu_n = \frac{g(n)}{E_n} = \frac{g(n)}{\phi(n)\sqrt{n\log n}}. \tag{2.66}$$

Thus, in terms of $g(n), \phi(n)$, and ν_n; the parameters m, β, η chosen as above satisfy the following relations:

$$m = 2\frac{n}{E_n} = \frac{2\sqrt{n}}{\phi(n)\sqrt{\log n}}, \tag{2.67}$$

$$\eta = \frac{g(n)}{2n} = \frac{g(n)}{\phi(n)\sqrt{n\log n}} \cdot \frac{\phi(n)\sqrt{\log n}}{2\sqrt{n}} = \frac{\nu_n}{m}; \tag{2.68}$$

and

$$\beta = 1 - \frac{2g(n)}{E_n} = 1 - 2\nu_n = 1 - 2\frac{g(n)}{\phi(n)\sqrt{n\log n}}. \tag{2.69}$$

In particular, it holds that

$$\eta = \frac{1-\beta}{2m}. \tag{2.70}$$

The expressions (2.67)-(2.70) will be convenient for handling certain expressions appearing below.

We will establish $m-$OGP for the interval $[\beta - \eta, \beta]$, where m, β, η are chosen as

above. As a sanity check, note that the interval $[\beta-\eta,\beta]$ has length η, and for our result to be non-vacuous, it should be the case that the overlap region is not void, that is

$$|(n\beta - n\eta, n\beta)\cap\mathbb{Z}| \geq 1.$$

Indeed

$$n\eta = \frac{n\nu_n}{m} = \frac{1}{2}g(n) = \omega_n(1),$$

thus the region is not void.

Recall now

$$Y_i(\tau) = \sqrt{1-\tau^2}X_0 + \tau X_i \in \mathbb{R}^n, \quad \text{for} \quad 1\leq i\leq m \quad \text{and} \quad \tau\in\mathcal{I}.$$

In order to apply first moment method and Markov's inequality, we essentially need two bounds: 1) a bound on the cardinality of the $m-$tuples $(\sigma^{(i)}: 1\leq i\leq m)$ of spin configurations whose pairwise overlaps are constrained to $[\beta-\eta,\beta]$, and 2) a bound on a certain (joint) probability.

To that end, define

$$S(\beta,\eta,m) \triangleq \left\{(\sigma^{(1)},\dots,\sigma^{(m)}) : \sigma^{(i)}\in\{-1,1\}^n, \frac{1}{n}\left\langle\sigma^{(i)},\sigma^{(j)}\right\rangle \in [\beta-\eta,\beta], 1\leq i<j\leq m\right\}$$

and

$$N(\beta,\eta,m,E_n,\mathcal{I}) = \sum_{(\sigma^{(i)}:1\leq i\leq m)\in S(\beta,\eta,m)} \mathbb{1}\left\{\exists\tau_1,\dots,\tau_m\in\mathcal{I} : \frac{1}{\sqrt{n}}|\left\langle\sigma^{(i)},Y_i(\tau_i)\right\rangle| \leq 2^{-E_n}, 1\leq i\leq m\right\}.$$

Observe that, with these notation, we have

$$N(\beta,\eta,m,E_n,\mathcal{I}) = |\mathcal{S}(\beta,\eta,m,E_n,\mathcal{I})|.$$

Thus, by Markov's inequality

$$\mathbb{P}\left(\mathcal{S}(\beta,\eta,m,E_n,\mathcal{I})\neq\varnothing\right) = \mathbb{P}\left(N(\beta,\eta,m,E_n,\mathcal{I})\geq 1\right) \leq \mathbb{E}[N(\beta,\eta,m,E_n,\mathcal{I})].$$

We will establish that with the parameters chosen as above, $\mathbb{E}[N(\beta,\eta,m,E_n,\mathcal{I})] = \exp(-\Theta(n))$, which will conclude the proof.

Step 1. Cardinality upper bound. We now upper bound the number of $m-$tuples $(\sigma^{(i)} : 1\leq i\leq m)$ of spin configurations with (pairwise) overlaps constrained to $[\beta-\eta,\beta]$, that is, we upper bound the cardinality $|S(\beta,\eta,m)|$. For this we rely on Lemma 2.6.1.

Now, for $\sigma^{(1)}$, there are 2^n choices. Furthermore, for any fixed $\rho\in[\beta-\eta,\beta]$, there exists

$$\binom{n}{n\frac{1-\rho}{2}}$$

spin configurations σ' for which $\frac{1}{n}\langle\sigma,\sigma'\rangle = \rho$. With this, the number of choices for

$\sigma^{(2)}$ evaluates

$$\sum_{\rho:\beta-\eta\le\rho\le\beta, n\rho\in\mathbb{Z}} \binom{n}{n\frac{1-\rho}{2}}.$$

With this, the number of all such $m-$tuples $(\sigma^{(i)} : 1 \le i \le m)$ with $n^{-1}\left\langle \sigma^{(i)}, \sigma^{(j)} \right\rangle \in [\beta-\eta, \beta]$, $1 \le i < j \le m$, is at most

$$2^n \left(\sum_{\rho:\beta-\eta\le\rho\le\beta, n\rho\in\mathbb{Z}} \binom{n}{n\frac{1-\rho}{2}} \right)^{m-1}. \tag{2.71}$$

Observe that with our choice of parameters where $1-\beta = o_n(1)$ and $\eta = o_n(1)$,

$$\max_{\substack{\rho:\rho\in[\beta-\eta,\beta] \\ \rho n\in\mathbb{Z}}} \binom{n}{n\frac{1-\rho}{2}} = \binom{n}{n\frac{1-\beta+\eta}{2}}. \tag{2.72}$$

Recalling (2.70) and the fact $m = \omega_n(1)$, and therefore $m^{-1} = o_n(1)$, we have

$$1-\beta+\eta = (1-\beta)\left(1+\frac{1}{2m}\right) = (1-\beta)(1+o_n(1)).$$

Consequently, using (2.69), we conclude

$$\begin{aligned} n\frac{1-\beta+\eta}{2} &= \frac{n}{2}(1-\beta)(1+o_n(1)) && (2.73)\\ &= \frac{n}{2}\frac{2g(n)}{\phi(n)\sqrt{n\log_2 n}}(1+o_n(1)) && (2.74)\\ &= \frac{g(n)\sqrt{n}}{\phi(n)\sqrt{\log_2 n}}(1+o_n(1)) && (2.75)\\ &= \frac{ng(n)}{E_n}(1+o_n(1)). && (2.76) \end{aligned}$$

Next, observe that

$$g(n) = o\left(\phi(n)^2\right) = o\left(\frac{E_n^2}{n\log_2 n}\right) = o\left(E_n\right),$$

as $E_n = o(n)$ which is trivially $o(n\log_2 n)$. Thus, it follows from (2.76) that $n\frac{1-\beta+\eta}{2} = o(n)$. Thus, Lemma 2.6.1 applies. As a sanity check, we also verify $n\frac{1-\beta+\eta}{2} = \omega(1)$, so that the counting bound is not vacuous: using (2.64) and (2.75), we have $\phi(n)^{-1} = \omega\left(\sqrt{\frac{\log_2 n}{n}}\right)$. Thus

$$n\frac{1-\beta+\eta}{2} = \omega(g(n)) = \omega(1).$$

We now proceed to control the term

$$\binom{n}{n\frac{1-\beta+\eta}{2}}.$$

As we have verified, $n\frac{1-\beta+\eta}{2} = o(n)$. Thus we are indeed in the setting of Lemma 2.6.1.

Observe that using (2.76),

$$\log_2 \frac{n}{n\frac{1-\beta+\eta}{2}} = \log_2 \frac{E_n}{g(n)} = O\left(\log_2 n\right),$$

since $E_n = o(n)$. We now apply Lemma 2.6.1 to conclude that

$$\binom{n}{n\frac{1-\beta+\eta}{2}} = \exp_2\left((1+o_n(1))\frac{ng(n)}{E_n}O\left(\log_2 n\right)\right) = \exp_2\left(O\left(\frac{ng(n)}{E_n}\log_2 n\right)\right). \tag{2.77}$$

Note also that by (2.68)

$$\begin{aligned} |[n\beta - n\eta, n\beta] \cap \mathbb{Z}| &= O(n\eta) \\ &= O(g(n)). \end{aligned}$$

Consequently, using (2.72), (2.77), the fact $E_n = \phi(n)\sqrt{n\log_2 n}$ and the cardinality bound above in this order

$$\sum_{\rho:\beta-\eta\leq\rho\leq\beta,\rho n\in\mathbb{Z}} \binom{n}{n\frac{1-\rho}{2}} \leq \exp_2\left(O\left(\frac{g(n)\sqrt{n\log_2 n}}{\phi(n)}\right) + O(\log_2 g(n))\right).$$

Now, since $\frac{1}{\phi} = \omega\left(\sqrt{\frac{\log_2 n}{n}}\right)$, we have

$$\frac{g(n)\sqrt{n\log_2 n}}{\phi(n)} = \omega(g(n)\log_2 n);$$

and therefore the term, $O(\log_2 g(n))$ appearing in the bound above, is lower order. Thus we conclude

$$\sum_{\rho:\beta-\eta\leq\rho\leq\beta:\rho n\in\mathbb{Z}} \binom{n}{n\frac{1-\rho}{2}} \leq \exp_2\left(O\left(\frac{g(n)\sqrt{n\log_2 n}}{\phi(n)}\right)\right). \tag{2.78}$$

We now return back to the earlier bound on the cardinality of the $m-$tuples with overlaps in $[\beta-\eta,\beta]$ as per (2.71). Since $m = \omega_n(1)$, it holds $m-1 = m(1+o_n(1))$.

With this, we obtain

$$\begin{aligned}
2^n \left(\sum_{\rho:\beta-\eta\le\rho\le\beta:\rho n\in\mathbb{Z}} \binom{n}{n\frac{1-\rho}{2}} \right)^{m-1} &\le \exp_2\left(n + (m-1)O\left(\frac{g(n)\sqrt{n\log_2 n}}{\phi(n)} \right) \right) \\
&\le \exp_2\left(n + O\left(\frac{mg(n)\sqrt{n\log_2 n}}{\phi(n)} \right) \right) \\
&= \exp_2\left(n + O\left(\frac{ng(n)}{\phi(n)^2} \right) \right).
\end{aligned}$$

Consequently,

$$|S(\beta,\eta,m)| \le \exp_2\left(n + O\left(\frac{ng(n)}{\phi(n)^2} \right) \right) \le \exp_2(n+o(n)), \tag{2.79}$$

since $g(n) = o\left(\phi(n)^2\right)$.

Step 2. Upper bounding the probability. For the energy exponent E_n defined earlier, suppose that $\mathcal{R}$ is the region

$$\mathcal{R} = \left[-2^{-E_n}, 2^{-E_n}\right] \times \left[-2^{-E_n}, 2^{-E_n}\right] \times \cdots \times \left[-2^{-E_n}, 2^{-E_n}\right] \subset \mathbb{R}^m.$$

Fix $\tau_1, \ldots, \tau_m \in \mathcal{I}$, and fix any $m-$tuple, $(\sigma^{(i)} : 1 \le i \le m) \in S(\beta,\eta,m)$. Recall $Y_i(\tau_i)$, $1 \le i \le m$ from Definition 2.2.1 and $Z_i = \frac{1}{\sqrt{n}} \left\langle \sigma^{(i)}, Y_i(\tau_i) \right\rangle \stackrel{d}{=} \mathcal{N}(0,1)$, $1 \le i \le m$.

Let Σ denotes the covariance matrix of the (centered) vector $(Z_i : 1 \le i \le m) \in \mathbb{R}^m$.

The probability we want to upper bound is the following:

$$\mathbb{P}\left((Z_i : 1 \le i \le m) \in \mathcal{R} \right) = (2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \int_{\mathcal{R}\subset\mathbb{R}^m} \exp\left(-\frac{1}{2} x^T \Sigma^{-1} x \right) \, dx.$$

Since provided $|\Sigma| \neq 0$, $\exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right) \le 1$ for any $x \in \mathbb{R}^m$; we upper bound the probability with

$$\mathbb{P}\left((Z_i : 1 \le i \le m) \in \mathcal{R} \right) \le (2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \mathrm{Vol}(\mathcal{R}) = 2^{\frac{m}{2}} \pi^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} 2^{-mE_n}. \tag{2.80}$$

Studying the covariance matrix, Σ. The lines below are almost identical to Step II in the proof of Theorem 2.2.3; and kept for convenience.

To control the probability in (2.80), we study the covariance matrix Σ. In particular, our goal is to lower bound $|\Sigma|$ away from zero, uniformly for all choices of $(\sigma^{(i)} : 1 \le i \le m) \in S(\beta,\eta,m)$, and for every $\tau_1, \ldots, \tau_m \in \mathcal{I}$.

Using the exact same route as in Step II of the proof of Theorem 2.2.3, we arrive at the conclusion that $\Sigma \in \mathbb{R}^{m\times m}$ has the following structure:

$$\Sigma_{ii} = 1, \quad \text{for} \quad 1 \le i \le m; \quad \text{and} \quad \Sigma_{ij} = \Sigma_{ji} = \gamma_i\gamma_j\bar{\Sigma}_{ij}, \quad \text{for} \quad 1 \le i < j \le m,$$

where $\bar{\Sigma}_{ij} = \bar{\Sigma}_{ji} = \rho_{ij} = \frac{1}{n}\left\langle \sigma^{(i)}, \sigma^{(j)} \right\rangle$, $1 \leq i < j \leq m$. Here, $\gamma_i = \sqrt{1-\tau_i^2}$, $1 \leq i \leq m$.

Namely, $\bar{\Sigma} \in \mathbb{R}^{m\times m}$ is an auxiliary matrix introduced for studying $\Sigma \in \mathbb{R}^{m\times m}$, and has the structure:

$$\bar{\Sigma}_{ii} = 1, \quad \text{for} \quad 1 \leq i \leq m; \quad \text{and} \quad \bar{\Sigma}_{ij} = \bar{\Sigma}_{ji} = \rho_{ij}, \quad \text{for} \quad 1 \leq i < j \leq m.$$

Now, let $A = \text{diag}(\gamma_1, \ldots, \gamma_m) \in \mathbb{R}^{m\times m}$ be a diagonal matrix. It follows that

$$\Sigma = A\bar{\Sigma}A + (I - A^2). \tag{2.81}$$

We next study $\bar{\Sigma}$. To that end, define the matrix $\widehat{\Sigma} \in \mathbb{R}^{m\times m}$, where $\widehat{\Sigma}_{ii} = 1$ for $1 \leq i < j \leq m$; and $\widehat{\Sigma}_{ij} = \widehat{\Sigma}_{ji} = \beta$ for $1 \leq i < j \leq m$. Observe that

$$\widehat{\Sigma} = (1-\beta)I_m + \beta 11^T.$$

Now, the spectrum of the matrix $11^T \in \mathbb{R}^{m\times m}$ consists of the eigenvalue m with multiplicity one; and eigenvalue 0 with multiplicity $m-1$. Furthermore, since $\widehat{\Sigma}$ is obtained by applying a rank-1 perturbation to a multiple of identity matrix, its spectrum consists of the eigenvalue $1-\beta+\beta m$, that is, $1+\beta(m-1)$ with multiplicity one; and $1-\beta$ with multiplicity $m-1$. Since $1+\beta(m-1)$ and $1-\beta$ are both positive, this (symmetric) matrix is also positive definite.

With this notation, we now express $\bar{\Sigma}$ of interest as

$$\bar{\Sigma} = \widehat{\Sigma} + E,$$

where the (symmetric) perturbation matrix $E \in \mathbb{R}^{m\times m}$ satisfies $E_{ii} = 0$ for $1 \leq i \leq m$, and $|E_{ij}| = |E_{ji}| \leq \eta$ for $1 \leq i < j \leq m$. We will bound the spectrum of $\bar{\Sigma}$ away from zero, using Wielandt-Hoffman inequality, Theorem 2.6.2. To that end, let $\lambda_1 = 1 + (m-1)\beta \geq \lambda_2 = \cdots = \lambda_m = 1 - \beta > 0$ denotes the eigenvalues of $\widehat{\Sigma}$; and let $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m$ denotes the eigenvalues of $\bar{\Sigma} = \widehat{\Sigma} + E$. Then, Theorem 2.6.2 yields

$$\sum_{1\leq j\leq m} (\mu_j - \lambda_j)^2 \leq \|E\|_F^2 \leq m(m-1)\eta^2 < (m\eta)^2 = \left(\frac{1-\beta}{2}\right)^2,$$

where we use the facts $E_{ii} = 0$, $|E_{ij}| \leq \eta$ for the second inequality; and (2.70) for the last equality. With this,

$$\frac{1-\beta}{2} > |\mu_m - \lambda_m| \Rightarrow \mu_m > \frac{1-\beta}{2} = \nu_n,$$

using (2.69). Note that, this bound is uniform across all $(\sigma^{(i)} : 1 \leq i \leq m) \in S(\beta, \eta, m)$: no matter which $m-$tuple $(\sigma^{(i)} : 1 \leq i \leq m) \in S(\beta, \eta, m)$ is chosen, the determinant of the (induced) covariance matrix Σ satisfies $|\bar{\Sigma}| > \nu_n^m$. Consequently, $|\bar{\Sigma}|^{-\frac{1}{2}} < \nu_n^{-\frac{m}{2}}$. Having controlled the determinant of $\bar{\Sigma}$, we now return back to the original covariance matrix Σ as per (2.81). Note that, under the aforementioned choice

of parameters, $\bar{\Sigma}$ is invertible. Furthermore, $\bar{\Sigma}$ is, by construction, a covariance matrix thus has non-negative eigenvalues $\lambda_1(\bar{\Sigma}) \geq \cdots \geq \lambda_m(\bar{\Sigma}) > 0$ with

$$m = \text{trace}(\bar{\Sigma}) \geq m\lambda_m(\bar{\Sigma}).$$

Thus we have

$$\lambda_m(\bar{\Sigma}) \leq 1. \tag{2.82}$$

Now, observe that as $\bar{\Sigma}$ is positive semidefinite, so is $A\bar{\Sigma}A$, appearing in the equation (2.81). Furthermore, as $1 - \gamma_i^2 \geq 0$ for $1 \leq i \leq m$, the matrix $I - A^2$ is positive semidefinite as well. We now study the smallest eigenvalue $\lambda_m(\Sigma)$.

Lemma 2.6.5. *For any choices of $\tau_1, \ldots, \tau_m \in \mathcal{I}$, it is the case that $\lambda_m(\Sigma) \geq \lambda_m(\bar{\Sigma})$. Hence,*

$$|\Sigma| \geq (\lambda_m(\bar{\Sigma}))^m > \nu_n^m > 0,$$

which is independent of the indices $\tau_1, \ldots, \tau_m$.

Proof. Recall the Courant-Fischer-Weyl variational characterization of the smallest singular value $\lambda_m(\Sigma)$ of a Hermitian matrix $\Sigma \in \mathbb{R}^{m \times m}$ [170]:

$$\lambda_m(\Sigma) = \inf_{v : \|v\|_2 = 1} v^T \Sigma v.$$

Notice furthermore that for the matrix $\bar{\Sigma} \in \mathbb{R}^{m \times m}$, and any $v' \in \mathbb{R}^m$, we also have

$$(v')^T \bar{\Sigma} v' \geq \lambda_m(\bar{\Sigma}) \|v'\|_2^2. \tag{2.83}$$

Now let $v \in \mathbb{R}^m$ have unit ℓ_2 norm ($\|v\|_2 = 1$). We have,

$$\begin{aligned}
v^T \Sigma v &= v^T (I - A^2) v + v^T A \bar{\Sigma} A v \\
&\geq \sum_{1 \leq i \leq m} (1 - \gamma_i^2) v_i^2 + \lambda_m(\bar{\Sigma}) \|Av\|_2^2 \\
&= \sum_{1 \leq i \leq m} (1 - \gamma_i^2 + \lambda_m(\bar{\Sigma}) \gamma_i^2) v_i^2 \\
&\geq \lambda_m(\bar{\Sigma}) \|v\|_2^2 \\
&= \lambda_m(\bar{\Sigma}),
\end{aligned}$$

where the first equality uses (2.81), the first inequality uses (2.83), and the last inequality uses $\lambda_m(\bar{\Sigma}) \leq 1$ as established in (2.82). Since for any $v \in \mathbb{R}^m$ with unit norm we have $v^T \Sigma v \geq \lambda_m(\bar{\Sigma})$, we thus take the infimum over all unit norm v and conclude

$$\lambda_m(\Sigma) \geq \lambda_m(\bar{\Sigma}).$$

Finally

$$|\Sigma| = \prod_{1 \leq j \leq m} \lambda_j(\Sigma) \geq \lambda_m(\Sigma)^m \geq \lambda_m(\bar{\Sigma})^m > \nu_n^m,$$

as desired. □

As a consequence of this lemma, we obtain that

$$|\Sigma|^{-\frac{1}{2}} < \nu_n^{-\frac{m}{2}}$$

uniformly for all $(\sigma^{(i)} : 1 \leq i \leq m) \in S(\beta, \eta, m)$, and every $\tau_1, \ldots, \tau_m \in \mathcal{I}$; meaning that the upper bound depends only on the choice of m, and ν_n induced by the overlap value β $(= 1 - 2\nu_n)$.

Equipped with this, we now return to the probability upper bound (2.80):

$$\begin{aligned}\mathbb{P}\left((Z_i : 1 \leq i \leq m) \in \mathcal{R}\right) &\leq 2^{\frac{m}{2}} \pi^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} 2^{-mE_n} \\ &\leq \exp_2\left(\frac{m}{2} - \frac{m}{2}\log_2 \pi + \frac{m}{2}\log_2 \frac{1}{\nu_n} - mE_n\right).\end{aligned}$$

In particular, taking union bound over all choices of $\tau_1, \ldots, \tau_m \in \mathcal{I}$, and recalling there are $n^{O(m)} = \exp_2(O(m \log_2 n))$ such choices, we have

$$\mathbb{E}\left[\mathbb{1}\left\{\exists \tau_1, \ldots, \tau_m \in \mathcal{I} : \frac{1}{\sqrt{n}} |\left\langle \sigma^{(i)}, Y_i(\tau_i)\right\rangle| \leq 2^{-E_n}, 1 \leq i \leq m\right\}\right] \tag{2.84}$$

$$= \exp_2\left(O(m \log_2 n) + \frac{m}{2} - \frac{m}{2}\log_2 \pi + \frac{m}{2}\log_2 \frac{1}{\nu_n} - mE_n\right). \tag{2.85}$$

We are now ready to upper bound the expectation.

Step 3. Upper bounding the expectation. Using the linearity of expectation, and the fact $O\left(ng(n)/\phi(n)^2\right) = o(n)$ following from (2.65) we have

$$\mathbb{E}[N(\beta, \eta, m, E_n, \mathcal{I})] \leq \exp_2\left(n + o(n) + O(m \log_2 n) + \frac{m}{2} - \frac{m}{2}\log_2 \pi + \frac{m}{2}\log_2 \frac{1}{\nu_n} - mE_n\right). \tag{2.86}$$

where we used the probability/expectation bound above, and the cardinality bound per (2.79). Keep in mind that $mE_n = 2n$ per (2.67). Thus $n - mE_n = -n$. Now, since $E_n = \omega_n(1)$, we simultaneously have

$$\frac{m}{2}, \frac{m}{2}\log_2 \pi = o(mE_n) = o(n).$$

Since $E_n = \omega(\sqrt{n \log_2 n})$, we also have

$$O(m \log_2 n) = o(mE_n) = o(n).$$

Finally, we study the $\frac{m}{2}\log_2 \frac{1}{\nu_n}$ term. Recalling ν_n from (2.66), we obtain

$$\begin{aligned}\log_2 \frac{1}{\nu_n} &= \left(\frac{1}{2}\log_2 n + \frac{1}{2}\log_2 \log_2 n + \log_2 \phi(n) - \log_2 g(n)\right) \\ &= O(\log_2 n),\end{aligned}$$

using (2.64) and (2.65). Applying now (2.67) and the fact $\phi(n) = \omega_n(1)$, we obtain

$$m \log_2 \frac{1}{\nu_n} = O\left(\frac{\sqrt{n \log_2 n}}{\phi(n)}\right) = o\left(\sqrt{n \log_2 n}\right) = o(n).$$

Consequently,

$$\mathbb{E}[N(\beta, \eta, m, E_n, \mathcal{I})] \leq \exp_2\left(-n + o(n)\right) = \exp_2(-\Theta(n)).$$

Finally, applying Markov's inequality, we conclude

$$\mathbb{P}\left(\mathcal{S}(\beta, \eta, m, E_n) \neq \varnothing\right) \leq \exp\left(-\Theta(n)\right),$$

as claimed. This concludes the proof when

$$\omega\left(\sqrt{n \log_2 n}\right) \leq E_n \leq o(n).$$

□

Case 2: The Special Case, $E_n = \omega\left(n \cdot \log^{-1/5+\epsilon} n\right)$.

Proof. Let $E_n = \omega\left(n \cdot \log^{-1/5+\epsilon} n\right)$ for an $\epsilon \in (0, \frac{1}{5})$. We set

$$g(n) \triangleq n \cdot \left(\frac{E_n}{n}\right)^{2+\frac{\epsilon}{8}}.$$

Take m, β, and η per (2.4), that is

$$m = \frac{2n}{E_n}, \quad \beta = 1 - \frac{2g(n)}{E_n}, \quad \text{and} \quad \eta = \frac{g(n)}{2n}.$$

We next introduce several auxiliary quantities. Set $s(n) \triangleq E_n/n$, and $z(n) = s(n)^{1+\frac{\epsilon}{8}}$. In terms of $s(n)$ and $z(n)$; the parameters $g(n), m, \beta, \eta$ chosen as above satisfy now the relations:

$$g(n) = n \cdot s(n) \cdot z(n) \quad \text{where} \quad z(n) = s(n)^{1+\frac{\epsilon}{8}}, \tag{2.87}$$

$$m = \frac{2n}{E_n} = \frac{2}{s(n)}, \tag{2.88}$$

and

$$\beta = 1 - 2\frac{g(n)}{E_n} = 1 - 2z(n) \quad \text{and} \quad \eta = \frac{g(n)}{2n} = \frac{s(n)z(n)}{2}. \tag{2.89}$$

We also have

$$\omega\left(\log^{-1/5+\epsilon} n\right) \leq s(n) \leq o(1). \tag{2.90}$$

Moreover, analogous to previous case, define $\nu_n = g(n)/E_n$, which now becomes

$$\nu_n = \frac{g(n)}{E_n} = z(n) = s(n)^{1+\frac{\epsilon}{8}}. \tag{2.91}$$

This is clearly $o_n(1)$ due to (2.90). Furthermore, the length of the interval $[n\beta - n\eta, n\beta]$ is $\Theta(n\eta)$ which is $\Theta(g(n)) = \omega_n(1)$. Hence, the interval is not vacuous. Moreover, in terms of this parameter, the expressions $\beta = 1 - 2\nu_n$, $\eta = \nu_n/m$ and $\eta = \frac{1-\beta}{2m}$ are still valid.

Most of the steps of the proof remains the same as the previous case. Below, we only point out the necessary changes.

Step 1. Cardinality upper Bound. The expressions (2.71) and (2.72) for the counting term remain the same. We now analyze the $n^{\frac{1-\beta+\eta}{2}}$ term. Using (2.89), we have

$$\begin{aligned} n\frac{1-\beta+\eta}{2} &= \frac{n}{2}\left(1 - (1-2z(n)) + \frac{s(n)z(n)}{2}\right) && (2.92)\\ &= \frac{n}{2}\left(2z(n) + \frac{s(n)z(n)}{2}\right) && (2.93)\\ &= nz(n)\left(1 + \frac{s(n)}{4}\right) && (2.94)\\ &= nz(n)\left(1 + o_n(1)\right), && (2.95) \end{aligned}$$

where the last step uses (2.90). Since $z(n) = s(n)^{1+\frac{\epsilon}{8}} = o(1)$ as well, we obtain $n^{\frac{1-\beta+\eta}{2}}$ to be $o(n)$. Hence, Lemma 2.6.1 applies. Applying it with,

$$k = n\frac{1-\beta+\eta}{2} = nz(n)(1+o_n(1)) = ns(n)^{1+\frac{\epsilon}{8}}(1+o_n(1)),$$

the expression (2.77) modifies to

$$\begin{aligned} \binom{n}{n\frac{1-\beta+\eta}{2}} &= \exp_2\left(\left(1+o_n(1)\right) ns(n)^{1+\frac{\epsilon}{8}} \log\frac{1}{s(n)^{1+\frac{\epsilon}{8}}}\right) && (2.96)\\ &= \exp_2\left(O\left(ns(n)^{1+\frac{\epsilon}{8}}\log\frac{1}{s(n)}\right)\right). && (2.97) \end{aligned}$$

Next, $\left|[n\beta - n\eta, n\beta] \cap \mathbb{Z}\right| = O(n\eta) = O(g(n))$, and thus this term contributes to $O\left(\log_2 g(n)\right)$ in the exponent. Using $g(n) = ns(n)^{2+\frac{\epsilon}{8}}$ per (2.87) as well as $s(n) \geq \omega\left(\log^{-\frac{1}{5}+\epsilon} n\right)$ per (2.90), we find

$$n\log^{-O(1)} n \leq g(n) \leq n \implies \log_2 g(n) = \Theta\left(\log_2 n\right).$$

Using $1/s(n) = \omega_n(1)$, we have

$$ns(n)^{1+\frac{\epsilon}{8}} \log \frac{1}{s(n)} = \omega\left(ns(n)^{1+\frac{\epsilon}{8}}\right) = \omega\Big(\log_2 n\Big) = \omega\Big(\log_2 g(n)\Big). \tag{2.98}$$

Consequently,

$$\sum_{\rho:\beta-\eta\le\rho\le\beta,\rho n\in\mathbb{Z}} \binom{n}{n\frac{1-\rho}{2}} \le O(g(n)) \cdot \binom{n}{n\frac{1-\beta+\eta}{2}} \tag{2.99}$$

$$\le \exp_2\left(O\left(ns(n)^{1+\frac{\epsilon}{8}} \log \frac{1}{s(n)}\right)\right), \tag{2.100}$$

where the second step uses (2.97) and (2.98). Next,

$$2^n \left(\sum_{\rho:\beta-\eta\le\rho\le\beta:\rho n\in\mathbb{Z}} \binom{n}{n\frac{1-\rho}{2}}\right)^{m-1} = \exp_2\left(n + (m-1)\log_2\left(\sum_{\rho:\beta-\eta\le\rho\le\beta,\rho n\in\mathbb{Z}} \binom{n}{n\frac{1-\rho}{2}}\right)\right). \tag{2.101}$$

Applying (2.100) in (2.101), we obtain the following modification for the cardinality bound appearing in (2.79) :

$$|S(\beta,\eta,m)| \le \exp_2\left(n + O\left(mns(n)^{1+\frac{\epsilon}{8}} \log \frac{1}{s(n)}\right)\right). \tag{2.102}$$

Step 2. Upper bounding the probability. The entire analysis of the probabilistic term remains intact. In particular, (2.85) remains valid. (The ν_n term appearing in (2.85) is now given by (2.91).)

We are now ready to upper bound the first moment.

Step 3. Upper bounding the expectation. Recalling (2.90) and (2.91), we obtain

$$\log \frac{1}{\nu_n} = \Theta\left(\log \frac{1}{s(n)}\right) = O\left(\log\log n\right) = o\left(E_n\right).$$

Consequently, the term $\frac{m}{2}\log_2 \frac{1}{\nu_n}$ appearing in (2.85) is $o(mE_n)$. The remaining terms, $O(m\log_2 n)$, $\frac{m}{2}$, and $\frac{m}{2}\log_2 \pi$, are still $o(mE_n)$ as in the first case, since $E_n = \omega(1)$. Hence, (2.85) becomes

$$\exp_2\left(O(m\log_2 n) + \frac{m}{2} - \frac{m}{2}\log_2 \pi + \frac{m}{2}\log_2 \frac{1}{\nu_n} - mE_n\right) = \exp_2\left(-mE_n + o\left(mE_n\right)\right). \tag{2.103}$$

After incorporating the modified cardinality bound (2.102) into the probability bound (2.103), the expression (2.86) for the first moment now becomes

$$\mathbb{E}[N(\beta,\eta,m,E_n,\mathcal{I})] \le \exp_2\left(n + O\left(mns(n)^{1+\frac{\epsilon}{8}} \log \frac{1}{s(n)}\right) - mE_n + o(mE_n)\right). \tag{2.104}$$

Recalling $m = 2n/E_n$, this bound is

$$\exp_2\left(-n + O\left(mns(n)^{1+\frac{\epsilon}{8}} \log \frac{1}{s(n)}\right) + o(n)\right). \tag{2.105}$$

To finish the proof, that is to establish $\mathbb{E}[N(\beta, \eta, m, E_n, \mathcal{I})] \leq \exp(-\Theta(n))$, it suffices to verify

$$O\left(mns(n)^{1+\frac{\epsilon}{8}} \log \frac{1}{s(n)}\right) = o(n).$$

Recall by (2.88) that $m = 2s(n)^{-1}$. Hence

$$mns(n)^{1+\frac{\epsilon}{8}} \log \frac{1}{s(n)} = n\left(2s(n)^{\frac{\epsilon}{8}} \log \frac{1}{s(n)}\right) = o(n),$$

where we used the fact that $s(n) = o_n(1)$ per (2.89) and thus

$$2s(n)^{\frac{\epsilon}{8}} \log \frac{1}{s(n)} = o_n(1), \qquad \forall \epsilon > 0.$$

Hence, the expression in (2.105) is indeed $\exp_2(-\Theta(n))$. This concludes the proof when

$$\omega\left(n \cdot \log^{-\frac{1}{5}+\epsilon} n\right) \leq E_n \leq o(n).$$

□

2.6.6 Proof of Theorem 2.2.8

We first have

$$N_\epsilon = \sum_{\sigma \in \{\pm 1\}^n} \mathbb{1}\left\{n^{-\frac{1}{2}} |\langle \sigma, X\rangle| = O(2^{-n\epsilon}), \sigma \text{ is locally optimal}\right\}.$$

Hence,

$$\mathbb{E}[N_\epsilon] = 2^n \mathbb{P}\left(n^{-\frac{1}{2}} |\langle \sigma, X\rangle| = O(2^{-n\epsilon}), \sigma \text{ is locally optimal}\right).$$

To start with, notice $n^{-\frac{1}{2}} \langle \sigma, X\rangle \stackrel{d}{=} \mathcal{N}(0,1)$, thus

$$\begin{aligned}\mathbb{P}\left(n^{-\frac{1}{2}} |\langle \sigma, X\rangle| = O(2^{-n\epsilon}), \sigma \text{ is locally optimal}\right) &\leq \mathbb{P}\left(n^{-\frac{1}{2}} |\langle \sigma, X\rangle| = O(2^{-n\epsilon})\right) \\ &\leq C2^{-n\epsilon},\end{aligned}$$

where $C > 0$ is some absolute constant. Hence,

$$\mathbb{E}[N_\epsilon] \leq C2^{n(1-\epsilon)} \Rightarrow \limsup_{n\to\infty} \frac{1}{n} \log_2 \mathbb{E}[N_\epsilon] \leq 1 - \epsilon. \tag{2.106}$$

We now investigate the lower bound. To that end, let $Y_i \triangleq \sigma_i X_i$. Note that Y_i, $1 \le i \le n$, is a collection of i.i.d. standard normal random variables. Now, local optimality of σ per Definition 2.2.7, namely, $\left\langle \sigma^{(i)}, X \right\rangle^2 \ge \langle \sigma, X \rangle^2$ for $1 \le i \le n$ is equivalent to

$$\sum_{j:1\le j\le n, j\neq i} Y_i Y_j \le 0$$

for $1 \le i \le n$. With this, we arrive at

$$\left\{ n^{-\frac{1}{2}} |\langle \sigma, X\rangle| = O(2^{-n\epsilon}), \sigma \text{ is locally optimal} \right\} = \bigcap_{1\le i\le n} \left\{ \sum_{1\le j\le n: j\neq i} Y_i Y_j \le 0 \right\} \cap \left\{ \left| \frac{1}{\sqrt{n}} \sum_{1\le j\le n} Y_j \right| \le 2^{-n\epsilon} \right\}$$

where we ignored the constant hidden under $O(2^{-n\epsilon})$ for convenience.

Observe now the following union of events, that are disjoint up to a measure zero set:

$$\left\{ \left| \frac{1}{\sqrt{n}} \sum_{1\le j\le n} Y_j \right| \le 2^{-n\epsilon} \right\} = \left\{ -2^{-n\epsilon} \le \frac{1}{\sqrt{n}} \sum_{1\le j\le n} Y_j \le 0 \right\} \cup \left\{ 0 \le \frac{1}{\sqrt{n}} \sum_{1\le j\le n} Y_j \le 2^{-n\epsilon} \right\}.$$

This brings us

$$\begin{aligned}
&\left\{ n^{-\frac{1}{2}} |\langle \sigma, X\rangle| = O(2^{-n\epsilon}), \sigma \text{ is locally optimal} \right\} \\
&= \underbrace{\left(\bigcap_{1\le i\le n} \left\{ \sum_{1\le j\le n: j\neq i} Y_i Y_j \le 0 \right\} \cap \left\{ \frac{1}{\sqrt{n}} \sum_{1\le j\le n} Y_j \in [-2^{-n\epsilon}, 0] \right\} \right)}_{\triangleq \mathcal{E}_1} \\
&\bigcup \underbrace{\left(\bigcap_{1\le i\le n} \left\{ \sum_{1\le j\le n: j\neq i} Y_i Y_j \le 0 \right\} \cap \left\{ \frac{1}{\sqrt{n}} \sum_{1\le j\le n} Y_j \in [0, 2^{-n\epsilon}] \right\} \right)}_{\triangleq \mathcal{E}_2}.
\end{aligned}$$

Note that $(Y_1, \ldots, Y_n) \stackrel{d}{=} (-Y_1, \ldots, -Y_n)$. From here $\mathbb{P}(\mathcal{E}_1) = \mathbb{P}(\mathcal{E}_2)$. Furthermore, the events $\mathcal{E}_1$ and $\mathcal{E}_2$ are disjoint, up to a set of measure zero; thus $\mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_2) = 2\mathbb{P}(\mathcal{E}_2)$.

We now compute $\mathbb{P}(\mathcal{E}_2)$. For convenience set $S \triangleq \sum_{1\le j\le n} Y_j$. Then the condition $\sum_{1\le j\le n, j\neq i} Y_i Y_j \le 0$ is equivalent to $Y_i(Y_i - S) \ge 0$. Namely,

$$\bigcap_{1\le i\le n} \left\{ \sum_{1\le j\le n: j\neq i} Y_i Y_j \le 0 \right\} \cap \left\{ n^{-\frac{1}{2}} S \in [0, 2^{-n\epsilon}] \right\} = \bigcap_{1\le i\le n} \{ Y_i \notin [0, S] \} \cap \left\{ n^{-\frac{1}{2}} S \in [0, 2^{-n\epsilon}] \right\}.$$

Define now the auxiliary variables

$$\bar{Y}_i \triangleq Y_i - \frac{1}{n} S, \quad 1 \le i \le n.$$

Clearly, $(\bar{Y}_1, \ldots, \bar{Y}_n)$ and S are jointly normal. Notice, furthermore, that for any

fixed $1 \le i \le n$, $\mathbb{E}\left[\bar{Y}_i S\right] = 0 = \mathbb{E}\left[\bar{Y}_i\right] = \mathbb{E}\left[S\right]$. This yields $(\bar{Y}_1, \ldots, \bar{Y}_n)$ and S are independent. Furthermore, $n^{-\frac{1}{2}} S \stackrel{d}{=} \mathcal{N}(0,1)$.

With these, we obtain

$$\mathbb{P}(\mathcal{E}_2) = \mathbb{P}\left(\bigcap_{1 \le i \le n} \left\{ \bar{Y}_i \notin \left[-\frac{1}{n} S, \frac{n-1}{n} S \right] \right\} \cap \left\{ n^{-\frac{1}{2}} S \in [0, 2^{-n\epsilon}] \right\} \right) \tag{2.107}$$

$$= \int_{z \in [0, 2^{-n\epsilon}]} \mathbb{P}\left(\bigcap_{1 \le i \le n} \left\{ \bar{Y}_i \notin \left[-\frac{1}{n} S, \frac{n-1}{n} S \right] \right\} \;\middle|\; n^{-\frac{1}{2}} S = z \right) \varphi(z)\, dz \tag{2.108}$$

$$= \int_{z \in [0, 2^{-n\epsilon}]} \mathbb{P}\left(\bigcap_{1 \le i \le n} \left\{ \bar{Y}_i \notin \left[-\frac{1}{\sqrt{n}} z, \frac{n-1}{\sqrt{n}} z \right] \right\} \right) \varphi(z)\, dz \tag{2.109}$$

$$= \int_{z \in [0, 2^{-n\epsilon}]} \left(1 - \mathbb{P}\left(\bigcup_{1 \le i \le n} \left\{ \bar{Y}_i \in \left[-\frac{1}{\sqrt{n}} z, \frac{n-1}{\sqrt{n}} z \right] \right\} \right) \right) \varphi(z)\, dz \tag{2.110}$$

$$\ge \int_{z \in [0, 2^{-n\epsilon}]} \left(1 - n\mathbb{P}\left(\bar{Y}_1 \in \left[-\frac{1}{\sqrt{n}} z, \frac{n-1}{\sqrt{n}} z \right] \right) \right) \varphi(z)\, dz \tag{2.111}$$

$$\ge \int_{z \in [0, 2^{-n\epsilon}]} \left(1 - n\mathbb{P}\left(\mathcal{N}(0,1) \in \left[-\frac{1}{\sqrt{n-1}} 2^{-n\epsilon}, \sqrt{n-1} \cdot 2^{-n\epsilon} \right] \right) \right) \varphi(z)\, dz \tag{2.112}$$

$$\ge \int_{z \in [0, 2^{-n\epsilon}]} \left(1 - \frac{n^2}{\sqrt{n-1}} 2^{-n\epsilon} \right) \varphi(z)\, dz \tag{2.113}$$

$$= \left(1 - \frac{n^2}{\sqrt{n-1}} 2^{-n\epsilon} \right) (2\pi)^{-\frac{1}{2}} 2^{-n\epsilon} (1 + o_n(1)) \tag{2.114}$$

$$= (2\pi)^{-\frac{1}{2}} (1 + o_n(1)) 2^{-n\epsilon}. \tag{2.115}$$

We now justify each of these lines, where $\varphi(z) \triangleq (2\pi)^{-\frac{1}{2}} \exp(-z^2/2)$ being the standard normal density. (2.107) is the definition of $\mathcal{E}_2$; (2.108) follows from the law of total probability; (2.109) uses the fact that the random vector $(\bar{Y}_i : 1 \le i \le n)$ and S are independent; (2.110) uses De Morgan's law; (2.111) uses union bound; (2.112) uses the fact that $\bar{Y}_1 \stackrel{d}{=} \mathcal{N}(0, \frac{n-1}{n})$ where $\mathcal{N}(0,1)$ is standard normal; (2.113) uses the trivial upper bound $\mathbb{P}(\mathcal{N}(0,1) \in I) \le |I|$ for any interval I; (2.114) uses the fact that in the interval $[0, 2^{-n\epsilon}]$, $\varphi(z) = (2\pi)^{-\frac{1}{2}}(1 + o_n(1))$; and finally (2.115) uses the fact $1 - n^2(n-1)^{-\frac{1}{2}} 2^{-n\epsilon} = 1 + o_n(1)$.

Therefore, using (2.115) we arrive at

$$\mathbb{P}\left(n^{-\frac{1}{2}} |\, \langle \sigma, X \rangle \,| = O(2^{-n\epsilon}), \sigma \text{ is locally optimal} \right) = \mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_2) = 2\mathbb{P}(\mathcal{E}_2) \ge 2(2\pi)^{-\frac{1}{2}} (1 + o_n(1)) 2^{-n\epsilon}.$$

With this, we conclude,

$$\mathbb{E}[N_\epsilon] = 2^n \mathbb{P}\left(n^{-\frac{1}{2}} |\, \langle \sigma, X \rangle \,| = O(2^{-n\epsilon}), \sigma \text{ is locally optimal} \right) \ge 2(2\pi)^{-\frac{1}{2}} (1 + o_n(1)) 2^{n(1-\epsilon)}.$$

Thus,

$$\liminf_{n\to\infty} \frac{1}{n} \log_2 \mathbb{E}\left[N_\epsilon\right] \geq 1-\epsilon. \tag{2.116}$$

Finally, we arrive at the desired conclusion by combining (2.106) and (2.116):

$$\lim_{n\to\infty} \frac{1}{n} \log_2 \mathbb{E}\left[N_\epsilon\right] = 1-\epsilon.$$

2.6.7 Proof of Theorem 2.3.2

The proof of Theorem 2.6.7 uses several interesting ideas. In order to present them in a coherent way, we first provide an informal outline sketching the proof.

Outline of the Proof of Theorem 2.3.2

Fix E_n (with prescribed growth condition) corresponding to the exponent of energy level 2^{-E_n} we want to rule out. We use the $m-$OGP property established in Theorem 2.2.6. Specifically, let $m \in \mathbb{N}$ and $1 > \beta > \eta > 0$ be the parameters prescribed by Theorem 2.2.6 for this choice of E_n.

- We first reduce the proof to the case of deterministic algorithms. That is, instead of considering $\mathcal{A} : \mathbb{R}^n \times \Omega \to \mathcal{B}_n$, we find a $\omega^* \in \Omega$, set $\mathcal{A}(\cdot) \triangleq \mathcal{A}(\cdot, \omega^*)$; and consider instead this deterministic choice $\mathcal{A} : \mathbb{R}^n \to \mathcal{B}_n$ in the remainder.

- We then study a certain high-probability event. This event will establish that for any $m-$tuple $\left(\sigma^{(i)} : 1 \leq i \leq m\right)$ of spin configurations that are near-optimal with respect to *independent instances* $X_i \stackrel{d}{=} \mathcal{N}(0, I_n)$, $1 \leq i \leq m$, there is a pair $1 \leq i < j \leq m$ such that $\overline{\mathcal{O}}\left(\sigma^{(i)}, \sigma^{(j)}\right)$ is contained in an interval of form $[0, 1-\eta']$, which is below $[\beta-\eta, \beta]$ interval prescribed by the OGP result, Theorem 2.2.6.

- We then set $X_0 \stackrel{d}{=} \mathcal{N}(0, I_n)$ and generate T "replicas" $X_i \in \mathbb{R}^n$, i.i.d. random vectors each with distribution $\mathcal{N}(0, I_n)$. We then divide $[0,1]$ into Q equal pieces via $0 = \tau_0 < \tau_1 < \cdots < \tau_Q = 1$; and interpolate, for each $1 \leq i \leq T$, between X_0 and X_i in the following way:

  $$Y_i(\tau_k) = \sqrt{1-\tau_k^2} X_0 + \tau_k X_i, \quad 1 \leq i \leq T,\, 0 \leq k \leq Q.$$

  The numbers T and Q will be tuned appropriately.

- We next establish that the pairwise "overlaps" are "stable" along each interpolation trajectory: for $1 \leq i < j \leq T$ and $0 \leq k \leq Q-1$, we show

  $$\left|\overline{\mathcal{O}}^{(ij)}\left(\tau_k\right) - \overline{\mathcal{O}}^{(ij)}\left(\tau_{k+1}\right)\right| \quad \text{is small.}$$

- We then use the guarantee on the probability of the success of the algorithm to arrive at a guarantee that the algorithm will produce, for each interpolation

trajectory $1 \leq i \leq T$ and time instance $0 \leq k \leq Q$ (that is for $Y_i(\tau_k) \in \mathbb{R}^n$), a solution that is near ground-state: the solution $\mathcal{A}\left(Y_i(\tau_k)\right) \in \mathcal{B}_n$ generated achieves an objective value 2^{-E_n} for (2.1). That is,

$$\frac{1}{\sqrt{n}} \left|\langle Y_i(\tau_k), \mathcal{A}\left(Y_i(\tau_k)\right)\rangle\right| = 2^{-E_n}.$$

- We then take a union bound over all subsets $S = \{i_1, \ldots, i_m\}$ of $[T]$ of cardinality $|S| = m$ to extend the previous high-probability event (the event pertaining the spin configurations that are near ground with respect to independent instances) when the indices come from the set S. Now, $\tau = 0$ in the beginning of the interpolation. Thus, it is the case that for every $1 \leq i < j \leq T$,

$$Y_i(\tau) = Y_j(\tau) \implies \mathcal{A}(Y_i(\tau)) = \mathcal{A}(Y_i(\tau)),$$

  when $\tau = 0$. On the other hand, due to the previous property applied to this subset $\mathcal{S}$, there exists indices $1 \leq k < \ell \leq m$ such that the overlap between i_k and i_ℓ is eventually below $\beta - \eta$. Since the overlaps are stable, that is, they do not change *abruptly*, this implies that there is a time τ such that the overlap between $\mathcal{A}(Y_{i_k}(\tau))$ and $\mathcal{A}(Y_{i_\ell}(\tau))$ is contained in $(\beta - \eta, \beta)$.

- Equipped with this, we then construct a certain graph $\mathbb{G} = (V, E)$. Specifically, we let $|V| = T$ where each vertex corresponds to a replica (i.e., an interpolation trajectory); and for $1 \leq i < j \leq m$, we let $(i, j) \in E$ if there is a time τ such that the overlap between $Y_i(\tau)$ and $Y_j(\tau)$ is contained in $(\beta - \eta, \beta)$. Moreover, each edge (i, j) is *colored* with one of Q different colors: color the edge $(i, j) \in E$ with color $1 \leq t \leq Q$ if τ_t is the first time such that the overlap between $\mathcal{A}\left(Y_i(\tau_t)\right)$ and $\mathcal{A}\left(Y_j(\tau_t)\right)$ is contained in $(\beta - \eta, \beta)$. With this, the graph has following properties. For every subset $S \subset V$ of $|S| = m$ vertices, there exists $1 \leq i_S < j_S \leq T$ such that $(i_S, j_S) \in E$. Moreover, each edge of this graph is colored with one of Q colors. We then establish, using tools from the extremal graph theory and the Ramsey theory, that $\mathbb{G}$ contains a monochromatic $m-$clique provided that T is sufficiently large.

- Finally, if $\mathbb{G}$ contains a monochromatic $m-$clique, this means there exists indices $1 \leq i_1 < i_2 < \cdots < i_m \leq T$ and a time $\tau' \in \{\tau_1, \ldots, \tau_Q\}$ such that the overlap between $\mathcal{A}\left(Y_{i_k}(\tau')\right)$ and $\mathcal{A}\left(Y_{i_\ell}(\tau')\right)$ is contained in $(\beta - \eta, \beta)$ for any $1 \leq k < \ell \leq m$. Setting

$$\sigma^{(k)} \triangleq \mathcal{A}\left(Y_{i_k}(\tau')\right), \quad 1 \leq k \leq m,$$

  we then deduce the $m-$tuple $\left(\sigma^{(k)} : 1 \leq k \leq m\right)$ of near ground-state spin configurations $\sigma^{(k)} \in \mathcal{B}_n$ violates the $m-$OGP established in Theorem 2.2.6. This will conclude the proof.

Before we formally start proving Theorem 2.3.2, we state several auxiliary results.

Auxiliary Results from Ramsey Theory and Extremal Graph Theory

Our first auxiliary result pertains the so-called two-color Ramsey numbers.

Theorem 2.6.6. *Let $k, \ell \geq 2$ be integers; and $R(k,\ell)$ denotes the smallest $n \in \mathbb{N}$ such that any red/blue (edge) coloring of K_n contains either a red K_k or a blue K_ℓ. Then*

$$R(k,\ell) \leq \binom{k+\ell-2}{k-1} = \binom{k+\ell-2}{\ell-1}. \tag{2.117}$$

Proof. To that end, we show $R(k,\ell)$ exists for any $k, \ell \in \mathbb{N}$; and moreover for $k, \ell \geq 2$, it holds that

$$R(k,\ell) \leq R(k,\ell-1) + R(k-1,\ell). \tag{2.118}$$

The elegant argument below is due to Erdös and Szekeres [108] and is reproduced herein for completeness. The argument is via induction on $k+\ell$. The base case is clear. Suppose for every i, j with $i + j \leq n-1$ the numbers $R(i,j)$ exist. Now, we consider $R(k,\ell)$ for $k+\ell = n$, $k, \ell \geq 2$. By inductive hypothesis, both $R(k-1,\ell)$ and $R(k,\ell-1)$ exists. Now, let $m \triangleq R(k-1,\ell) + R(k,\ell-1)$, and consider any red/blue (edge) coloring of K_m. For any vertex $v \in K_m$, either (a) v is adjacent to at least $R(k-1,\ell)$ vertices through a red edge; or (b) v is adjacent to at least $R(k,\ell-1)$ vertices through a blue edge. Assume case (a). By inductive hypothesis, any $R(k-1,\ell)$ such neighbors of v contains either a red K_{k-1} or a blue K_ℓ. Adding v, the resulting graph indeed has either a red K_ℓ or a blue K_ℓ. The case (b) is handled similarly. This establishes (2.118).

(2.117) now follows from (2.118) again by induction on $k+\ell$. The base cases are verified easily. Assume $k, \ell \geq 3$. Then by inductive hypothesis

$$R(k-1,\ell) \leq \binom{k+\ell-3}{k-2} \quad \text{and} \quad R(k,\ell-1) \leq \binom{k+\ell-3}{k-1}.$$

Thus,

$$R(k,\ell) \leq R(k-1,\ell) + R(k,\ell-1) \leq \binom{k+\ell-3}{k-2} + \binom{k+\ell-3}{k-1} = \binom{k+\ell-2}{k-1}.$$

□

The second result pertains the so-called multicolor Ramsey numbers.

Theorem 2.6.7. *Let $q, m \in \mathbb{N}$. Denote by $R_q(m)$ the smallest $n \in \mathbb{N}$ for which any $q-$coloring of the edges of K_n necessarily contains a monochromatic K_m. Then*

$$R_q(m) \leq q^{qm}. \tag{2.119}$$

Theorem 2.6.7 can be shown using a minor modification of the neighborhood-chasing argument given by Erdös and Szekeres [108]. See [81, Page 6] for more information.

We next define a certain graph property.

Definition 2.6.8. *Fix a positive integer $M \in \mathbb{N}$. A graph $\mathbb{G} = (V, E)$ is called* $M-$admissible *if for any $S \subset V$, $|S| = M$; there exists distinct $i, j \in S$ such that $(i, j) \in E$.*

Namely, $\mathbb{G}$ is $M-$admissible if $\alpha(\mathbb{G}) \leq M - 1$, where $\alpha(\mathbb{G})$ is the independence number of $\mathbb{G}$.

We now state and prove our second auxiliary result, an extremal graph theory result.

Proposition 2.6.9. *Let $M \in \mathbb{N}$. Any $M-$admissible graph $\mathbb{G} = (V, E)$ with*

$$|V| \geq \binom{2M-2}{M-1}$$

contains an $M-$clique.

Proof of Proposition 2.6.9

Proof. Let $\mathbb{G}$ be an $M-$admissible graph on $|V| \geq \binom{2M-2}{M-1}$ vertices. Theorem 2.6.6 then yields that $|V| \geq R(M, M)$. Now, for any $i, j \in V$; we say (i, j) is colored "red" if $(i, j) \in E$; and (i, j) is colored "blue" otherwise. Due to the Ramsey property, $\mathbb{G}$ contains either a red K_M or a blue K_M; that is, $\mathbb{G}$ contains either a clique of size M or an independent set of size M. But since $\mathbb{G}$ is $M-$admissible, $\alpha(\mathbb{G}) \leq M - 1$. Thus the latter is not the case. Hence $\mathbb{G}$ contains a K_M. □

We are now ready to start formally proving Theorem 2.3.2.

Proof of Theorem 2.3.2

Proof. In what follows, recall the notation that for $\sigma, \sigma' \in \mathcal{B}_n$;

$$\overline{\mathcal{O}}(\sigma, \sigma') = \frac{1}{n} \langle \sigma, \sigma' \rangle = \frac{1}{n} \sum_{1 \leq i \leq n} \sigma_i \sigma'_i.$$

Recall also that all floor/ceiling signs are omitted for the sake of a clear presentation.

Let $L > 0$ be fixed (which is constant in n); and $\exp_2(-E_n)$ be the target energy level whose "exponent" E_n satisfies, for some $\epsilon \in (0, \frac{1}{5})$,

$$\omega\left(n \cdot \log^{-\frac{1}{5}+\epsilon} n\right) \leq E_n \leq o(n). \tag{2.120}$$

In what follows, we choose

$$c_1 = \frac{1}{6400} \quad \text{and} \quad c_2 = 8 \cdot 480^2, \tag{2.121}$$

and establish that there exists no randomized algorithm $\mathcal{A} : \mathbb{R}^n \times \Omega \to \mathcal{B}_n$ that is $\left(2^{-E_n}, f, L, \rho', p_f, p_{\text{st}}\right)$-optimal for every sufficiently large n, where the parameter f is

specified in Theorem 2.3.2, as

$$f = \frac{1}{6400} \cdot n \cdot \left(\frac{E_n}{n}\right)^{4+\frac{\epsilon}{4}}; \tag{2.122}$$

and the parameters $\rho', p_f, p_{\text{st}}$ are given by (2.9) with c_2 chosen as in (2.121) (we supress the dependence of ρ', p_f and p_{st} on n and c_2 for convenience). The proof is by a contradiction argument.

Choice of Auxiliary Parameters. We choose parameters m, β, η (all functions of n) as in the second part of Theorem 2.2.6, which we recall for convenience:

$$m \triangleq m(n) = \frac{2n}{E_n}, \tag{2.123}$$

$$g(n) = n \cdot \left(\frac{E_n}{n}\right)^{2+\frac{\epsilon}{8}}; \tag{2.124}$$

and

$$\beta \triangleq \beta(n) = 1 - 2\frac{g(n)}{E_n} \quad \text{and} \quad \eta \triangleq \eta(n) = \frac{g(n)}{2n}. \tag{2.125}$$

We now establish certain convenient expression for the parameters $f, \rho', p_f, p_{\text{st}}$ in terms of the quantities m, β, η above. For f chosen per (2.122), define

$$C_1 \triangleq \frac{f}{n} = \frac{1}{6400}\left(\frac{E_n}{n}\right)^{4+\frac{\epsilon}{4}}.$$

Using (2.124) and (2.125), it follows that

$$C_1 = \frac{g(n)^2}{6400 \cdot n^2} = \frac{\eta^2}{1600}. \tag{2.126}$$

Define next

$$Q = \frac{480^2 \cdot 2L}{\eta^2}. \tag{2.127}$$

Using (2.124), (2.125), and (2.127), it follows that

$$Q = \frac{2 \cdot 480^2 \cdot L}{\eta^2} = \frac{2 \cdot 480^2 \cdot L}{(g(n)/2n)^2} = \left(8 \cdot 480^2 \cdot L\right) \cdot \left(\frac{n}{E_n}\right)^{4+\frac{\epsilon}{4}}. \tag{2.128}$$

In particular, with $c_2 = 8 \cdot 480^2$ as above, the parameter $T(c_2)$ defined per (2.8) becomes

$$T(c_2) \triangleq T = \exp_2\left(2^{4mQ\log_2 Q}\right) \tag{2.129}$$

for m chosen in (2.123) and Q chosen in (2.127).

Moreover, for p_f and p_{st} chosen per (2.9), it holds that

$$p_f = \frac{1}{4T(Q+1)} \tag{2.130}$$

and

$$p_{\mathrm{st}} = \frac{1}{9TQ^2}. \tag{2.131}$$

Define next the function

$$\Psi(x) \triangleq \sqrt{\left(1-\frac{x^2}{Q^2}\right)\left(1-\frac{(x+1)^2}{Q^2}\right)} + \frac{x(x+1)}{Q^2}, \quad 0 \le x \le Q-1.$$

We show $\Psi(\cdot)$ is decreasing on $[0, Q-1]$. For this, it suffices to verify $\Psi'(x) \le 0$ for $0 \le x \le Q-1$. We have

$$\Psi'(x) = \frac{1}{Q^2}\left(2x+1-\frac{x\left(Q^2-(x+1)^2\right)+(x+1)\left(Q^2-x^2\right)}{\sqrt{\left(Q^2-x^2\right)\left(Q^2-(x+1)^2\right)}}\right)$$

Now set $u \triangleq Q^2-(x+1)^2$, $\lambda \triangleq \frac{x}{2x+1}$; and $\bar{\lambda} = 1-\lambda = \frac{x+1}{2x+1}$ (while suppressing x dependence). Clearly $u \ge 0$ as $0 \le x \le Q-1$ and $\bar{\lambda} > 1/2$. It then boils down verifying

$$\Psi'(x) \le 0 \Leftrightarrow \sqrt{u(u+2x+1)} \le \lambda u + \bar{\lambda}(u+2x+1).$$

Applying the weighted AM-GM inequality, we find

$$\lambda u + \bar{\lambda}(u+2x+1) \ge u^{\lambda} \cdot (u+2x+1)^{\bar{\lambda}}.$$

Hence, it suffices to verify

$$u^{\lambda} \cdot (u+2x+1)^{\bar{\lambda}} \ge \sqrt{u(u+2x+1)} \Leftrightarrow (u+2x+1)^{\frac{1}{4x+2}} > u^{\frac{1}{4x+2}},$$

which is immediate as $u+2x+1 > u$. Having established that $\Psi(\cdot)$ is decreasing on $[0, Q-1]$, thus $\min_{0\le k\le Q-1}\Psi(k) = \Psi(Q-1)$; it holds that

$$\min_{0\le k\le Q-1}\Psi(k) = \min_{0\le k\le Q-1}\sqrt{\left(1-\frac{k^2}{Q^2}\right)\left(1-\frac{(k+1)^2}{Q^2}\right)} + \frac{k(k+1)}{Q^2} = 1-\frac{1}{Q}.$$

In particular, ρ' chosen as in (2.9) admits

$$\rho' = 1-\frac{1}{8\cdot 480^2\cdot L}\left(\frac{E_n}{n}\right)^{4+\frac{\epsilon}{4}} = 1-\frac{1}{Q} = \min_{0\le k\le Q-1}\Psi(k), \tag{2.132}$$

where Q is the parameter studied in (2.127), (2.128).

To prove. In what follows, our goal is to establish that there exists no randomized algorithm $\mathcal{A}: \mathbb{R}^n \times \Omega \to \mathcal{B}_n$ that is $\left(2^{-E_n}, C_1 n, L, \rho', p_f, p_{\mathrm{st}}\right)$-optimal for every suffi-

ciently large n, where $C_1, \rho', p_f, p_{\rm st}$ admit the convenient expressions (2.126), (2.132), (2.130), and (2.131); respectively.

Reduction to deterministic algorithms. We first reduce the proof to the case $\mathcal{A}$ is deterministic. Let $\mathbb{P}_X \otimes \mathbb{P}_\omega$ denotes the joint law of (X, ω). Here, ω is the randomness of $\mathcal{A}$. Define now the event

$$\mathcal{E}_s(\omega) \triangleq \left\{ \frac{1}{\sqrt{n}} \Big| \langle X, \mathcal{A}(X;\omega) \rangle \Big| \leq 2^{-E_n} \right\}.$$

Observe that

$$\mathbb{P}_{X,\omega} \left(\frac{1}{\sqrt{n}} \Big| \langle X, \mathcal{A}(X,\omega) \rangle \Big| > 2^{-E_n} \right) = \mathbb{E}_\omega \Big[\mathbb{P}_X \Big(\mathcal{E}_s \left(\omega\right)^c \Big) \Big].$$

We now perceive $\mathbb{P}_X \left(\mathcal{E}_s \left(\omega\right)^c \right)$ as a random variable whose source of randomness is ω (as the randomness over X is "integrated" over $\mathbb{P}_X$). Using Markov's inequality

$$\mathbb{P}_\omega \Big(\mathbb{P}_X \Big(\mathcal{E}_s \left(\omega\right)^c \Big) \geq 2p_f \Big) \leq \frac{\mathbb{E}_\omega \left[\mathbb{P}_X \left(\mathcal{E}_s \left(\omega\right)^c \right) \right]}{2p_f} \leq \frac{1}{2}.$$

Set

$$\Omega_1 \triangleq \Big\{ \omega \in \Omega : \mathbb{P}_X \left(\mathcal{E}_s \left(\omega\right)^c \right) < 2p_f \Big\} \implies \mathbb{P}_\omega \left(\Omega_1 \right) \geq \frac{1}{2}.$$

Now, divide the interval $[0,1]$ into Q subintervals $0 = \tau_0 < \tau_1 < \cdots < \tau_Q = 1$, each of size Q^{-1} for Q introduced in (2.127). Next, set

$$\rho_k \triangleq \sqrt{\left(1 - \tau_k^2\right) \left(1 - \tau_{k+1}^2\right)} + \tau_k \tau_{k+1}, \quad 0 \leq k \leq Q-1. \tag{2.133}$$

For ρ' introduced in (2.132), we have that $\rho_k \in [\rho', 1]$, $0 \leq k \leq Q-1$. Define next

$$\mathcal{E}_2 \left(\omega\right) \triangleq \Big\{ d_H \Big(\mathcal{A}(X,\omega), \mathcal{A}(Y,\omega) \Big) \leq C_1 n + L \|X - Y\|_2^2 \Big\}.$$

Define also the sequence $A_{k,\omega}$ of random variables

$$A_{k,\omega} \triangleq \mathbb{P}_{(X,Y):X \sim_{\rho_k} Y} \Big(\mathcal{E}_2(\omega)^c \Big), \quad 0 \leq k \leq Q-1, \quad \text{and} \quad \omega \in \Omega.$$

The source of randomness in each $A_{k,\omega}$ is due to $\mathbb{P}_\omega$. Observe now that for any fixed $0 \leq k \leq Q-1$, using Markov's inequality similar to above,

$$\begin{aligned}
\mathbb{P}_\omega \left(A_{k,\omega} \geq 3Qp_{\rm st} \right) &\leq \frac{1}{3Qp_{\rm st}} \mathbb{E}_\omega \Big[A_{k,\omega} \Big] \\
&= \frac{1}{3Qp_{\rm st}} \mathbb{P}_{(X,Y,\omega):X \sim_{\rho_k} Y} \Big(d_H \Big(\mathcal{A}(X,\omega), \mathcal{A}(Y,\omega) \Big) > C_1 n + L \|X - Y\|_2^2 \Big) \\
&\leq \frac{1}{3Q}.
\end{aligned}$$

Taking now a union bound over $0 \leq k \leq Q-1$,

$$\mathbb{P}_\omega\left(\bigcup_{0\leq k\leq Q-1}\left\{A_{k,\omega} \geq 3Qp_{\mathrm{st}}\right\}\right) \leq \frac{1}{3}.$$

Hence,

$$\Omega_2 \triangleq \left\{\omega\in\Omega : A_{k,\omega} < 3Qp_{\mathrm{st}}, 0\leq k\leq Q-1\right\} \implies \mathbb{P}_\omega(\Omega_2)\geq \frac{2}{3}.$$

Since $\mathbb{P}_\omega(\Omega_1)+\mathbb{P}_\omega(\Omega_2)\geq \frac{1}{2}+\frac{2}{3}>1$, it follows that $\Omega_1\cap\Omega_2\neq\varnothing$. Consequently, there exists an $\omega^*\in\Omega$, such that

$$\mathbb{P}_X\left(\frac{1}{\sqrt{n}}\left|\langle X, \mathcal{A}(X,\omega^*)\rangle\right| \leq 2^{-E_n}\right) \geq 1-2p_f. \tag{2.134}$$

and

$$\mathbb{P}_{(X,Y):X\sim_{\rho_k}Y}\left(d_H\left(\mathcal{A}(X,\omega^*),\mathcal{A}(Y,\omega^*)\right)\leq C_1 n + L\|X-Y\|_2^2\right)\geq 1-3Qp_{\mathrm{st}}, \quad \text{for} \quad 0\leq k\leq Q-1. \tag{2.135}$$

In the remainder, we fix this choice of $\omega^*\in\Omega$, and interpret $\mathcal{A}(\cdot)\triangleq\mathcal{A}(\cdot,\omega^*)$ as a deterministic (that is, no "coin flip" ω) map acting between $\mathbb{R}^n$ and $\mathcal{B}_n$.

An auxiliary high-probability event. We now study a certain auxiliary high-probability event. This event pertains to the spin configurations that are near-optimal with respect to *independent* instances.

Let $\mathcal{M}$ be an index set with cardinality m, $X_i \overset{d}{=} \mathcal{N}(0,I_n)$, $i\in\mathcal{M}$, be i.i.d. Let $S_{\mathcal{M}}$ be a shorthand for the set

$$\mathcal{S}_{\mathcal{M}} \triangleq \mathcal{S}\left(1, \frac{3g(n)}{E_n}, m, E_n, \{1\}\right)$$

(in the sense of Definition 2.2.1) of $m-$tuples $\left(\sigma^{(i)} : i\in\mathcal{M}\right)$ of spin configurations $\sigma^{(i)}\in\mathcal{B}_n$ with a modification that the $\mathcal{O}(\cdot,\cdot)$ in Definition 2.2.1 is replaced with $\overline{\mathcal{O}}(\cdot,\cdot)$, the normalized inner product, as studied in Theorem 2.2.6. Here, we keep the m parameter as in (2.88); but modify the η parameter into $3g(n)/E_n$. Note that with these choices, $\eta=\frac{1-\beta}{2m}$ no longer holds, but as we expand below this does not cause any problems.

Namely, $S_{\mathcal{M}}$ is the set of spin configurations that i) have a large inner product and ii) are near ground-state with respect to *independent* instances $X_i\in\mathbb{R}^n$. We claim

Lemma 2.6.10.

$$\mathbb{P}(\mathcal{E}_{\mathcal{M}}) \leq \exp(-\Theta(n)), \tag{2.136}$$

where

$$\mathcal{E}_{\mathcal{M}} \triangleq \{S_{\mathcal{M}}\neq\varnothing\} = \{|S_{\mathcal{M}}|\geq 1\}. \tag{2.137}$$

Namely on $\mathcal{E}_{\mathcal{M}}$, it is the case that for any $\left(\sigma^{(i)} : i \in \mathcal{M}\right)$ that are near-optimal, there exists $i < j$, $i, j \in \mathcal{M}$, such that

$$\overline{\mathcal{O}}\left(\sigma^{(i)}, \sigma^{(j)}\right) \in \left[0, 1 - \frac{3g(n)}{E_n}\right]. \tag{2.138}$$

Proof of Lemma 2.6.10. The proof of this claim is nearly identical to (and in fact easier than) that of Theorem 2.2.6, Case 2. Thus we only point out the necessary modifications.

The term $g(n)$ and E_n, as functions of $s(n)$ and $z(n)$, remain the same as (2.87). That is,

$$g(n) = n \cdot s(n) \cdot z(n), \quad E_n = n \cdot s(n), \quad \text{where} \quad z(n) = s(n)^{1+\frac{\epsilon}{8}}.$$

The expression (2.89) regarding parameters β, η now modifies to

$$\beta = 1 \quad \text{and} \quad \eta = \frac{3g(n)}{E_n}.$$

Note that, in this case the covariance matrix Σ is always identity due to the independence of X_i, $1 \leq i \leq m$. Moreover, with $\beta = 1$; the counting term $n^{\frac{1-\beta+\eta}{2}}$ studied in (2.95) is now

$$n\frac{1-\beta+\eta}{2} = \frac{3ng(n)}{2E_n} = \frac{3}{2}nz(n).$$

This is clearly $o(n)$ since $z(n) = o(1)$. Thus, Lemma 2.6.1 is applicable, and the counting bound, (2.97), now becomes

$$\begin{aligned}\binom{n}{n\frac{1-\beta+\eta}{2}} &= \exp_2\left((1+o_n(1))\,\frac{3nz(n)}{2}\log_2\frac{2n}{3nz(n)}\right)\\ &= \exp_2\left(O\left(ns(n)^{1+\frac{\epsilon}{8}}\log_2\frac{1}{s(n)}\right)\right),\end{aligned}$$

where we used the fact $z(n) = s(n)^{1+\frac{\epsilon}{8}}$. Note now that

$$\left|[n\beta - n\eta, n\beta] \cap \mathbb{Z}\right| = O\left(n\eta\right) = O\left(\frac{ng(n)}{E_n}\right) = O(nz(n)) = o(n),$$

using $z(n) = o(1)$. Hence,

$$\log\left|[n\beta - n\eta, n\beta] \cap \mathbb{Z}\right| = O(\log n) = o\left(ns(n)^{1+\frac{\epsilon}{8}}\log_2\frac{1}{s(n)}\right)$$

Thus, (2.100) remains the same. Hence, the cardinality upper bound (2.102) is still

of form

$$\exp_2\left(n + O\left(mns(n)^{1+\frac{\epsilon}{8}}\log_2\frac{1}{s(n)}\right)\right).$$

Now, since the covariance matrix Σ is identity, there is no contribution of a term of form $\frac{m}{2}\log_2\frac{1}{\nu_n}$ (the determinant contribution) to (2.103); and the "dominant" contribution of the probability term (2.103) to the exponent of the first moment is $-mE_n$.

Hence, (2.103), (2.104); and (2.105) all remain the same. Thus, Theorem 2.2.6 indeed still remains valid.

□

Construction of interpolation paths. Our proof will use the so-called "interpolation method". To that end, let $X_i \in \mathbb{R}^n$, $0 \le i \le T$, be i.i.d. random vectors (dubbed as *replicas*), each having distribution $\mathcal{N}(0, I_n)$, where T is specified in (2.129).

Recall now $Y_i(\tau)$, $\tau \in [0,1]$ and $1 \le i \le T$, from Definition 2.2.1. Notice that for any $\tau \in [0,1]$ and any $1 \le i \le T$, $Y_i(\tau) \stackrel{d}{=} \mathcal{N}(0, I_n)$. At $\tau = 0$, it is the case that $Y_i(\tau) = Y_j(\tau) = X_0$ for $1 \le i < j \le T$. Thus, for $\tau = 0$,

$$\mathcal{A}\left(Y_i(\tau)\right) = \mathcal{A}\left(Y_j(\tau)\right), \quad 1 \le i < j \le T.$$

At $\tau = 1$, on the other hand, $Y_i(\tau)$, $1 \le i \le T$, is a collection of T i.i.d. random vectors, each with distribution $\mathcal{N}(0, I_n)$.

Divide the interval $[0,1]$ into Q subintervals $0 = \tau_0 < \tau_1 < \cdots < \tau_Q = 1$, each of size $1/Q$, where Q is specified in (2.127). Define next the pairwise overlaps

$$\overline{\mathcal{O}}^{(ij)}\left(\tau_k\right) \triangleq \frac{1}{n}\left\langle \mathcal{A}\left(Y_i\left(\tau_k\right)\right), \mathcal{A}\left(Y_j\left(\tau_k\right)\right)\right\rangle \tag{2.139}$$

for $1 \le i < j \le T$ and $0 \le k \le Q$.

Stability of successive steps. We now establish, using the stability of $\mathcal{A}$, that for $1 \le i < j \le T$ and $0 \le k \le Q-1$,

$$\left|\overline{\mathcal{O}}^{(ij)}\left(\tau_k\right) - \overline{\mathcal{O}}^{(ij)}\left(\tau_{k+1}\right)\right|$$

is small. More concretely we establish

Lemma 2.6.11.

$$\mathbb{P}(\mathcal{E}_3) \ge 1 - (T+1)\exp\Big(-\Theta(n)\Big) - 3TQ^2 p_{\rm st} \tag{2.140}$$

where

$$\mathcal{E}_3 \triangleq \bigcap_{1\le i<j\le T}\bigcap_{0\le k\le Q}\left\{\left|\overline{\mathcal{O}}^{(ij)}(\tau_k) - \overline{\mathcal{O}}^{(ij)}(\tau_{k+1})\right| \le 4\sqrt{C_1} + \frac{48\sqrt{2L}}{\sqrt{Q}}\right\}. \tag{2.141}$$

Later, we study the asymptotics of T and show that the bound in (2.140) is not vacuous.

Proof. We first establish that for every $1 \leq i \leq T$, $\|X_i\|_2 \leq 6\sqrt{n}$ w.h.p. Let $X_i = (X_i(j) : 1 \leq j \leq n)$, where $X_i(j)$, $1 \leq j \leq n$ are i.i.d. standard normal. Appealing to Bernstein's inequality as in the proof of [282, Theorem 3.1.1], we have that for every $t \geq 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{1\leq j\leq n} X_i(j)^2 - 1\right| \geq t\right) \leq \exp(-cn\min\{t^2, t\}).$$

Using now a union bound over $1 \leq i \leq T$, we conclude

$$\mathbb{P}\left(\|X_i\| \leq 6\sqrt{n}, 0 \leq i \leq T\right) \geq 1 - (T+1)\exp(-\Theta(n)).$$

Here the choice of 6 is arbitrary, any constant larger than 1 works.

Fix any $1 \leq i \leq T$. We now upper bound $\|Y_i(\tau_k) - Y_i(\tau_{k+1})\|$, where $Y_i(\cdot)$ is defined in Definition 2.2.1. Note that

$$\begin{aligned}\|Y_i(\tau_k) - Y_i(\tau_{k+1})\| &\leq \left|\sqrt{1-\tau_k^2} - \sqrt{1-\tau_{k+1}^2}\right| \|X_0\| + |\tau_k - \tau_{k+1}|\|X_i\| \\ &\leq \left|\sqrt{1-\tau_k^2} - \sqrt{1-\tau_{k+1}^2}\right| \|X_0\| + Q^{-1}\|X_i\|\end{aligned}$$

using triangle inequality, and the fact $|\tau_k - \tau_{k+1}| \leq Q^{-1}$. Next, observe that using $\tau_k, \tau_{k+1} \in [0,1]$,

$$\sqrt{\left|\tau_k^2 - \tau_{k+1}^2\right|} = \sqrt{|\tau_k - \tau_{k+1}|}\sqrt{\tau_k + \tau_{k+1}} \leq \sqrt{2\,|\tau_k - \tau_{k+1}|}.$$

We now show

$$\frac{\sqrt{\left|\tau_k^2 - \tau_{k+1}^2\right|}}{\sqrt{1-\tau_k^2} + \sqrt{1-\tau_{k+1}^2}} \leq 1.$$

Squaring, and using $\tau_{k+1} > \tau_k$, this is equivalent to having

$$\tau_{k+1}^2 - \tau_k^2 \leq 1 - \tau_k^2 + 1 - \tau_{k+1}^2 + 2\sqrt{(1-\tau_k^2)\left(1-\tau_{k+1}^2\right)} \iff \tau_{k+1}^2 \leq 1 + \sqrt{(1-\tau_k^2)\left(1-\tau_{k+1}^2\right)},$$

which holds as $0 \leq \tau_i \leq 1$ for all i. Equipped with the previous bounds, we thus have

$$\begin{aligned}\left|\sqrt{1-\tau_k^2}-\sqrt{1-\tau_{k+1}^2}\right| &= \frac{|\tau_k^2-\tau_{k+1}^2|}{\sqrt{1-\tau_k^2}+\sqrt{1-\tau_{k+1}^2}} \\ &\leq \sqrt{2|\tau_k-\tau_{k+1}|}\underbrace{\frac{\sqrt{|\tau_k^2-\tau_{k+1}^2|}}{\sqrt{1-\tau_k^2}+\sqrt{1-\tau_{k+1}^2}}}_{\leq 1} \\ &\leq \sqrt{2}|\tau_k-\tau_{k+1}|^{\frac{1}{2}} \\ &\leq \sqrt{2}Q^{-\frac{1}{2}},\end{aligned}$$

where the last line uses $|\tau_k - \tau_{k+1}| \leq Q^{-1}$. In particular, on the high probability event[2]

$$\mathcal{E}_{\text{norm}} \triangleq \left\{\|X_i\|_2 \leq 6\sqrt{n}, 0 \leq i \leq T\right\}$$

it holds that

$$\left\|Y_i(\tau_k) - Y_i(\tau_{k+1})\right\| \leq 6\sqrt{2}n^{\frac{1}{2}}Q^{-\frac{1}{2}} + 6Q^{-1}n^{\frac{1}{2}} \leq 12\sqrt{2}n^{\frac{1}{2}}Q^{-\frac{1}{2}}. \tag{2.142}$$

Next, we observe that $Y_i(\tau_k), Y_i(\tau_{k+1})$ are both distributed $\mathcal{N}(0, I_n)$ with correlation

$$\mathbb{E}\left[Y_i(\tau_k)\, Y_i(\tau_{k+1})^T\right] = \rho_k I$$

where ρ_k, $0 \leq k \leq Q-1$, is per (2.133). Define now the event

$$\mathcal{E}_{\text{stability}} = \bigcap_{0\leq i\leq T}\bigcap_{0\leq k\leq Q-1}\left\{d_H\Big(\mathcal{A}(Y_i(\tau_k)), \mathcal{A}(Y_i(\tau_{k+1}))\Big) \leq C_1 n + L\|Y_i(\tau_k) - Y_i(\tau_{k+1})\|_2^2\right\}$$

which is the event that the algorithm is stable for each interpolation trajectory $1 \leq i \leq T$ along time indices $0 \leq k \leq Q-1$. Using (2.135) together with a union bound over $1 \leq i \leq T$ and $0 \leq k \leq Q-1$, it holds that with probability at least $1 - 3TQ^2 p_{\text{st}}$,

$$\mathbb{P}(\mathcal{E}_{\text{stability}}) \geq 1 - 3TQ^2 p_{\text{st}}.$$

Consequently, taking a union bound, this time for $\mathcal{E}_{\text{norm}} \cap \mathcal{E}_{\text{stability}}$, we arrive at

$$\mathbb{P}(\mathcal{E}_{\text{norm}} \cap \mathcal{E}_{\text{stability}}) = 1 - (T+1)\exp(-\Theta(n)) - 3TQ^2 p_{\text{st}}.$$

We now compute $\left|\overline{\mathcal{O}}^{(ij)}(\tau_k) - \overline{\mathcal{O}}^{(ij)}(\tau_{k+1})\right|$ on the event $\mathcal{E}_{\text{norm}} \cap \mathcal{E}_{\text{stability}}$, while treating the stability condition deterministic due to the conditioning.

For notational convenience, let $\mathcal{A}_i(k) \triangleq \mathcal{A}(Y_i(\tau_k))$, $1 \leq i \leq T$ and $0 \leq k \leq Q$.

[2] As we verify soon, $T = 2^{o(n)}$, hence this is indeed a high probability event.

We first observe that

$$\|\mathcal{A}_i(k) - \mathcal{A}_j(k)\| = 2\sqrt{d_H\left(\mathcal{A}_i(k), \mathcal{A}_j(k)\right)}.$$

Using the stability condition,

$$d_H\left(\mathcal{A}_i(k), \mathcal{A}_i(k+1)\right) \le C_1 n + L\|Y_i(k) - Y_i(k+1)\|_2^2, \quad 1 \le i \le T, \quad 0 \le k \le Q-1$$

together with the trivial inequality $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$, valid for all $u, v \ge 0$, we have

$$\|\mathcal{A}_i(k) - \mathcal{A}_i(k+1)\|_2 \le 2\sqrt{C_1}\sqrt{n} + 2\sqrt{L}\,\|Y_i(k) - Y_i(k+1)\|_2\,. \tag{2.143}$$

Next,

$$\begin{aligned}
\left|\overline{\mathcal{O}}^{(ij)}(\tau_k) - \overline{\mathcal{O}}^{(ij)}(\tau_{k+1})\right| &= \frac{1}{n}\left|\langle \mathcal{A}_i(k), \mathcal{A}_j(k)\rangle - \langle \mathcal{A}_i(k+1), \mathcal{A}_j(k+1)\rangle\right| && (2.144)\\
&\le \frac{1}{n}\left|\langle \mathcal{A}_i(k) - \mathcal{A}_i(k+1), \mathcal{A}_j(k)\rangle\right| + \frac{1}{n}\left|\langle \mathcal{A}_i(k+1), \mathcal{A}_j(k) - \mathcal{A}_j(k+1)\rangle\right| && (2.145)\\
&\le \frac{1}{\sqrt{n}}\Big(\|\mathcal{A}_i(k) - \mathcal{A}_i(k+1)\|_2 + \|\mathcal{A}_j(k) - \mathcal{A}_j(k+1)\|_2\Big) && (2.146)\\
&\le \frac{1}{\sqrt{n}}\left(4\sqrt{C_1}\sqrt{n} + 2\sqrt{L}\|Y_i(\tau_k) - Y_i(\tau_{k+1})\| + 2\sqrt{L}\|Y_j(\tau_k) - Y_j(\tau_{k+1})\|\right) && (2.147)\\
&\le 4\sqrt{C_1} + \frac{4\sqrt{L}\cdot 12\sqrt{2}n^{\frac{1}{2}}Q^{-\frac{1}{2}}}{\sqrt{n}} && (2.148)\\
&= 4\sqrt{C_1} + \frac{48\sqrt{2L}}{\sqrt{Q}}. && (2.149)
\end{aligned}$$

Above, (2.144) uses the definition; (2.145) uses the triangle inequality; (2.146) uses the Cauchy-Schwarz inequality, and the fact $\|\mathcal{A}_i(k+1)\|_2 = \|\mathcal{A}_j(k)\| = \sqrt{n}$; (2.147) uses (2.143); and finally (2.148) uses (2.142). □

Success along the trajectory. We now study the event that the algorithm $\mathcal{A}$ is "successful" along each interpolation trajectory. We claim that we have

$$\mathbb{P}\left(\mathcal{E}_4\right) \ge 1 - 2T(Q+1)p_f \tag{2.150}$$

where the event $\mathcal{E}_4$ is defined as

$$\mathcal{E}_4 \triangleq \bigcap_{1\le i\le T}\bigcap_{0\le k\le Q}\left\{\frac{1}{\sqrt{n}}\left|\langle \mathcal{A}_i(k), Y_i(\tau_k)\rangle\right| \le 2^{-E_n}\right\}. \tag{2.151}$$

Namely, the event $\mathcal{E}_4$ says that the algorithm $\mathcal{A}$ creates a near ground-state at each "discrete time instance" $0 \leq k \leq Q$ along each interpolation trajectory $1 \leq i \leq T$.

We now prove this claim. Note that as $X_i \in \mathbb{R}^n$, $1 \leq i \leq T$ are i.i.d. $\mathcal{N}(0, I_n)$, it follows that for each $1 \leq i \leq T$ and $0 \leq k \leq Q$, $Y_i(\tau_k) \stackrel{d}{=} \mathcal{N}(0, I_n)$. Using now (2.134) together with a union bound over $1 \leq i \leq T$ and $0 \leq k \leq Q$ settles the result.

The order of growth of parameters. Before we put everything together, we now study the order of growth of relevant parameters. This is necessary for applying union bound arguments that will follow.

First, combining (2.120) and (2.123), we obtain

$$\omega(1) \leq m \leq o\left(\log^{\frac{1}{5}-\epsilon} n\right). \tag{2.152}$$

This is not vacuous since $\epsilon < \frac{1}{5}$.

Next, since L is constant in n, the asymptotics of Q given in (2.128) becomes

$$Q = \left(8 \cdot 480^2 \cdot L\right) \cdot \left(\frac{n}{E_n}\right)^{4+\frac{\epsilon}{4}} = \Theta\left(\left(\frac{n}{E_n}\right)^{4+\frac{\epsilon}{4}}\right).$$

Recalling now the condition (2.120) on E_n, we obtain

$$Q = o\left(\log^{\left(\frac{1}{5}-\epsilon\right)\left(4+\frac{\epsilon}{4}\right)} n\right). \tag{2.153}$$

Moreover, (2.153) yields also that

$$\log Q = \log\left(o\left(\log^{\left(\frac{1}{5}-\epsilon\right)\left(4+\frac{\epsilon}{4}\right)} n\right)\right) = O\left(\log\log n\right). \tag{2.154}$$

Combining bounds (2.152), (2.153), and (2.154), we arrive at

$$mQ\log Q = o\left(\log^{\left(\frac{1}{5}-\epsilon\right)\left(5+\frac{\epsilon}{4}\right)} n\right) O\left(\log\log n\right) = o\left(\log^{\left(\frac{1}{5}-\epsilon\right)\left(5+\frac{\epsilon}{2}\right)} n\right), \tag{2.155}$$

where we used

$$O\left(\log\log n\right) = o\left(\log^{w} n\right)$$

valid for any constant $w > 0$. Next, observe that for $\epsilon > 0$,

$$\left(\frac{1}{5}-\epsilon\right)\left(5+\frac{\epsilon}{2}\right) = 1 - \left(\frac{49}{10}\epsilon + \frac{\epsilon^2}{2}\right) < 1.$$

Thus, by combining (2.129), (2.155), and the fact $mQ\log Q + \log m = \Theta(mQ\log Q)$, we arrive at

$$\binom{T}{m} \leq T^m = \exp_2\left(m 2^{4mQ\log_2 Q}\right) = 2^{o(n)}. \tag{2.156}$$

Putting everything together. We now put everything together. For any $\mathcal{M} \subset [T]$ with $|\mathcal{M}| = m$, recall the event $\mathcal{E}_{\mathcal{M}}$ as in (2.137), and define the event

$$\mathcal{E}_1 \triangleq \bigcap_{\mathcal{M} \subset [T] : |\mathcal{M}| = m} \mathcal{E}_{\mathcal{M}}. \tag{2.157}$$

Using (2.136), together with a union bound, we obtain

$$\mathbb{P}\left(\mathcal{E}_1^c\right) = \mathbb{P}\left(\bigcup_{\mathcal{M} \subset [T] : |\mathcal{M}| = m} \mathcal{E}_{\mathcal{M}}^c\right) \leq \binom{T}{m} \exp\left(-\Theta\left(n\right)\right).$$

Since $\binom{T}{m} = 2^{o(n)}$ per (2.156), we deduce

$$\mathbb{P}\left(\mathcal{E}_1\right) \geq 1 - \exp\left(-\Theta(n)\right). \tag{2.158}$$

For the events $\mathcal{E}_1$ defined in (2.157) (refer to (2.158) for its probability), $\mathcal{E}_3$ defined in (2.141) (refer to (2.140) for its probability), and $\mathcal{E}_4$ defined in (2.151) (refer to (2.150) for its probability), define their intersection by

$$\mathcal{F} = \mathcal{E}_1 \cap \mathcal{E}_3 \cap \mathcal{E}_4.$$

Check that using (2.140), the fact $T = 2^{o(n)}$ per (2.156) as well as the choice of p_{st} per (2.131), we have

$$\begin{aligned} \mathbb{P}\left(\mathcal{E}_3^c\right) &\leq (T+1) \exp\left(-\Theta(n)\right) + 3TQ^2 p_{\mathrm{st}} \\ &\leq \exp\left(-\Theta(n)\right) + \frac{1}{3}. \end{aligned}$$

Moreover, using (2.150) as well as the choice of p_f per (2.130), we arrive at

$$\mathbb{P}(\mathcal{E}_4^c) \leq \frac{1}{2}.$$

A union bound over $\mathcal{E}_1^c, \mathcal{E}_2^c$ and $\mathcal{E}_4^c$ then yields

$$\mathbb{P}\left(\mathcal{F}\right) = \mathbb{P}\left(\mathcal{E}_1 \cap \mathcal{E}_3 \cap \mathcal{E}_4\right) = 1 - \mathbb{P}\left(\mathcal{E}_1^c \cup \mathcal{E}_3^c \cup \mathcal{E}_4^c\right) \geq \frac{1}{6} - \exp\left(-\Theta(n)\right). \tag{2.159}$$

In the remainder, assume that we are on the event $\mathcal{F}$.

Note that, from the choice of C_1 per (2.126) and Q per (2.127), we obtain that on the event $\mathcal{F}$, it holds that

$$\left|\overline{\mathcal{O}}^{(ij)}\left(\tau_k\right) - \overline{\mathcal{O}}^{(ij)}\left(\tau_{k+1}\right)\right| \leq \frac{\eta}{5}, \tag{2.160}$$

for $1 \leq i < j \leq T$ and $0 \leq k \leq Q$.

Now, fix any subset $S \subset [T]$ with $|S| = m$. A consequence of the event $\mathcal{E}_1$,

through (2.138), is that there exists distinct $i_S, j_S \in S$ such that

$$\overline{\mathcal{O}}^{(i_S,j_S)}(\tau_Q) \in \left[0, 1 - \frac{3g(n)}{E_n}\right].$$

We verify that this interval is "below" the forbidden region, $(\beta - \eta, \beta)$: per (2.125) it suffices to ensure

$$\beta - \eta = \underbrace{1 - \frac{2g(n)}{E_n}}_{=\beta} - \underbrace{\frac{g(n)}{2n}}_{=\eta} > 1 - \frac{3g(n)}{E_n} \Leftrightarrow \frac{g(n)}{E_n} > \frac{g(n)}{2n} \Leftrightarrow 2n > E_n.$$

Since $E_n = o(n)$, this indeed holds for all sufficiently large n.

Take now $\delta = \frac{\eta}{100}$. We next show there exists a $k' \in [1, Q] \cap \mathbb{Z}$ such that

$$\overline{\mathcal{O}}^{(i_S,j_S)}(\tau_{k'}) \in (\beta - \eta + 3\delta, \beta - 3\delta),$$

where $(\beta - \eta, \beta)$ is the forbidden overlap region as per Theorem 2.2.6. Take indeed K_0 to be the last index (in $[1, Q] \cap \mathbb{Z}$) where $\overline{\mathcal{O}}^{(i_S j_S)}(\tau_{K_0}) \geq \beta - 3\delta$. Note that such a K_0 must exist since $\overline{\mathcal{O}}^{(ij)}(0) = 1$ for every $1 \leq i < j \leq T$. Then if $\overline{\mathcal{O}}^{(i_S j_S)}(\tau_{K_0+1}) \leq \beta - \eta + 3\delta$, we obtain

$$\left|\overline{\mathcal{O}}^{(i_S,j_S)}(\tau_{K_0}) - \overline{\mathcal{O}}^{(i_S j_S)}(\tau_{K_0+1})\right| \geq \eta - 6\delta > 0,$$

which contradicts with the event $\mathcal{E}_3$ and in particular with (2.160) for sufficiently large n. Namely,

$$\overline{\mathcal{O}}^{(i_S,j_S)}(\tau_{K_0+1}) \in (\beta - \eta + 3\delta, \beta - 3\delta).$$

In particular, keeping in mind that S was arbitrary, we conclude that for every subset $S \subset [T]$ of cardinality $|S| = m$, there exists $1 \leq i_S < j_S \leq m$ such that for some $\tau_S \in \{\tau_1, \ldots, \tau_Q\}$ it is the case that

$$\overline{\mathcal{O}}^{(i_S,j_S)}(\tau_S) \in (\beta - \eta + 3\delta, \beta - 3\delta) \subsetneq (\beta - \eta, \beta).$$

Equipped with this, we now construct a certain graph $\mathbb{G} = (V, E)$ such that the following holds.

- Its vertex set V coincides with $[T]$. That is, $V = \{1, 2, \ldots, T\}$, where each vertex corresponds to an interpolation trajectory $1 \leq i \leq T$.

- For any $1 \leq i < j \leq T$, $(i, j) \in E$ if and only if there exists a time $\tau \in [0, 1]$ such that
$$\overline{\mathcal{O}}^{(ij)}(\tau) \in (\beta - \eta, \beta).$$

Next, we "color" each edge of $\mathbb{G}$ with one of Q colors. Specifically, for any $1 \leq i < j \leq T$ with $(i, j) \in E$, this edge is colored with color t, $1 \leq k \leq Q$ where t is the first

time instance $\{\tau_1, \tau_2, \ldots, \tau_Q\}$ such that

$$\overline{\mathcal{O}}^{(ij)}(\tau_t) \in (\beta - \eta, \beta).$$

In particular, $\mathbb{G}$ enjoys the following properties.

- $\mathbb{G} = (V, E)$ has $|V| = T$ vertices; with the property that for any subset $S \subset V$ of cardinality $|S| = m$, there exists a distinct pair $i, j \in S$ of vertices such that $(i, j) \in E$.
- Any edge $(i, j) \in E$ of $\mathbb{G}$ is colored with one of Q colors.

Proposition 2.6.12. *The graph $\mathbb{G}$ contains a monochromatic $m-$clique K_m.*

Proof of Proposition 2.6.12. Recall from (2.129) that $\mathbb{G}$ has

$$T = \exp_2\left(2^{4mQ\log_2 Q}\right)$$

vertices. Define now

$$M \triangleq Q^{mQ} = 2^{mQ\log_2 Q} \tag{2.161}$$

Note that $\mathbb{G}$ is $m-$admissible, in the sense of Definition 2.6.8. Since $M > m$ for $Q > 1$, it is also $M-$admissible. Observe that

$$T = \exp_2\left(2^{4mQ\log_2 Q}\right) \geq \exp_2\left(2 \cdot \underbrace{2^{mQ\log_2 Q}}_{=M}\right) = 4^M \geq \binom{2M-2}{M-1}.$$

Applying Proposition 2.6.9, we find that $\mathbb{G}$ contains an M, that is a Q^{Qm}, clique, K_M. Finally, since each edge of K_M is colored with one of Q colors and $R_Q(m) \leq Q^{Qm}$ per Theorem 2.6.7, we obtain that K_M contains a monochromatic $m-$clique. Namely, $\mathbb{G}$ contains a monochromatic m-clique K_m since all graphs above we worked with are subgraphs of $\mathbb{G}$. This concludes the proof of Proposition 2.6.12. □

We now complete the proof of Theorem 2.3.2. Observe what it means for $\mathbb{G}$ to contain a monochromatic $m-$clique: there exists an $m-$tuple $1 \leq i_1 < i_2 < \cdots < i_m \leq T$ of vertices (i.e. replicas) and a color (i.e. a time $\tau' \in \{\tau_1, \ldots, \tau_Q\}$) such that

$$\overline{\mathcal{O}}^{(i_k, i_\ell)}(\tau') \in (\beta - \eta, \beta), \quad 1 \leq k < \ell \leq m.$$

Now, define

$$\sigma^{(k)} \triangleq \mathcal{A}\left(Y_{i_k}(\tau')\right), \quad 1 \leq k \leq m.$$

It follows that $\left(\sigma^{(k)} : 1 \leq k \leq m\right)$ enjoys the following conditions:

- Since we are on the event $\mathcal{F}$ which is a subset of the success event $\mathcal{E}_4$ (2.151), it holds that

$$\frac{1}{\sqrt{n}}\left|\left\langle \sigma^{(k)}, Y_{i_k}(\tau')\right\rangle\right| \leq 2^{-E_n}$$

- For $1 \leq k < \ell \leq m$,
$$\overline{\mathcal{O}}\left(\sigma^{(k)}, \sigma^{(\ell)}\right) \in (\beta - \eta, \beta).$$

Namely, for the choice $\zeta \triangleq \{i_1, i_2, \ldots, i_m\}$ of the $m-$tuple of distinct indices, the set $\mathcal{S}_\zeta \triangleq \mathcal{S}(\beta, \eta, m, E_n, \mathcal{I})$ introduced in Definition 2.2.1—with modification that inner products are considered—(where the indices $1, 2, \ldots, m$ there is replaced with $i_1, \ldots, i_m$) with $\mathcal{I} = \{\tau_0, \tau_1, \ldots, \tau_Q\}$ is non-empty. Namely,

$$\mathbb{P}\Big(\exists \zeta \subset [T], |\zeta| = m : S_\zeta \neq \varnothing\Big) \geq \mathbb{P}(\mathcal{F}) \geq \frac{1}{2} - \exp\left(-\Theta(n)\right).$$

We now use the $m-$OGP result, Theorem 2.2.6. Taking a union bound over $\zeta \subset [T]$ with $|\zeta| = m$ in Theorem 2.2.6, we obtain

$$\mathbb{P}\Big(\exists \zeta \subset [T], |\zeta| = m : S_\zeta \neq \varnothing\Big) \leq \binom{T}{m} \exp\left(-\Theta(n)\right) = \exp\left(-\Theta(n)\right),$$

since $\binom{T}{m} = 2^{o(n)}$. But this yields

$$\exp\left(-\Theta(n)\right) \geq \mathbb{P}\Big(\exists \zeta \subset [T], |\zeta| = m : S_\zeta \neq \varnothing\Big) \geq \frac{1}{2} - \exp\left(-\Theta(n)\right),$$

that is

$$\exp\left(-\Theta(n)\right) \geq \frac{1}{6} - \exp\left(-\Theta(n)\right).$$

This is a contradiction for sufficiently large n. Therefore, the proof is complete. □

2.6.8 Proof of Theorem 2.3.3

Proof. We start by recalling that

$$H(\sigma^*) = H(-\sigma^*) = \Theta\left(2^{-n}\right),$$

with high probability, as noted in the introduction. Now, using Theorem 2.2.2, it follows that

$$\min_{\sigma \in I_2} H(\sigma) = \Omega\left(2^{-n\epsilon}\right)$$

with high probability. Indeed, for ρ chosen as above, with high probability no two spin configurations with overlap $\left[\rho, \frac{n-2}{n}\right]$ can achieve simultaneously an energy of $O\left(2^{-n\epsilon}\right)$.

In what follows next, the constants hidden under $\Theta(\cdot)$ and $\Omega(\cdot)$ are absorbed into the inverse temperature $\beta > 0$.

We have the following trivial lower bound:

$$\pi_\beta(I_3) = \pi_\beta\left(\overline{I_3}\right) = \pi_\beta(\sigma^*) = \frac{1}{Z_\beta} \exp\left(-\beta H(\sigma^*)\right) = \frac{1}{Z_\beta} \exp\left(-\beta 2^{-n}\right).$$

Notice, on the other hand, that for any $\sigma \in I_2$,

$$\pi_\beta(\sigma) \leq \frac{1}{Z_\beta} \exp\left(-\beta 2^{-n\epsilon}\right).$$

Next, we upper bound

$$|I_2| \leq \sum_{1 \leq k \leq \left\lceil \frac{n(1-\rho)}{2} \right\rceil} \binom{n}{k} = \exp_2\left(nh\left(\frac{1-\rho}{2}\right) + O\left(\log_2 n\right)\right),$$

where $h(\cdot)$ is the binary entropy function. Consequently

$$\pi_\beta(I_2) = \sum_{\sigma \in I_2} \pi_\beta(\sigma) \leq \frac{|I_2|}{Z_\beta}\left(-\beta 2^{-n\epsilon}\right) \leq \frac{1}{Z_\beta} \exp\left(nh\left(\frac{1-\rho}{2}\right) + O\left(\log_2 n\right) - \beta 2^{-n\epsilon}\right).$$

Hence

$$\pi_\beta\left(I_3\right) \geq \exp\left(-\beta 2^{-n} + \beta 2^{-n\epsilon} - nh\left(\frac{1-\rho}{2}\right) + O\left(\log_2 n\right)\right) \pi_\beta(I_2).$$

Finally, in the regime $\beta = \Omega\left(n 2^{n\epsilon}\right)$, it is the case that

$$-\beta 2^{-n} + \beta 2^{-n\epsilon} - nh\left(\frac{1-\rho}{2}\right) + O\left(\log_2 n\right) = \Omega\left(\beta 2^{-n\epsilon}\right) = \Omega(n).$$

Hence,

$$\pi_\beta\left(I_3\right) \geq e^{\Omega(n)} \pi_\beta(I_2).$$

We next apply this reasoning for the set I_1, which is slightly more delicate.

To that end, fix an $\epsilon' \in (\epsilon, 1)$ (recall that $\epsilon < 1$). We will show that with probability $1 - O(1/n)$, there exists a $\sigma' \in \mathcal{B}_n$ such that $H(\sigma') = \Theta\left(2^{-n\epsilon'}\right)$. For this, it suffices to use [183, Theorem 3.1] (with parameters $\beta = \sqrt{n} 2^{-n\epsilon'}$ and $\epsilon = \frac{\beta}{2}$, in terms of their notation).

It is evident, due to the OGP as well as the fact I_3 and $\overline{I_3}$ contains only ground states $\pm\sigma^*$, that $\sigma' \notin \left(\overline{I_2} \cup I_2\right) \cup \left(\overline{I_3} \cup I_3\right)$. Consequently, $\sigma' \in I_1$. With this, we have the trivial lower bound

$$\pi_\beta\left(I_1\right) \geq \pi_\beta\left(\sigma'\right) = \frac{1}{Z_\beta} \exp\left(-\beta 2^{-n\epsilon'}\right).$$

Repeating the exact same reasoning while keeping in mind $\epsilon' > \epsilon$, we conclude

$$\pi_\beta\left(I_1\right) \geq \exp\left(\Omega\left(\beta 2^{-n\epsilon}\right)\right) \pi_\beta(I_2).$$

This concludes the proof. □

2.6.9 Proof of Theorem 2.3.4

Proof. In what follows, we have $\beta = \Omega\left(n2^{n\epsilon}\right)$.

Part (a)

Using the FEW property established in Theorem 2.3.3, we have that

$$\min\left\{\pi_\beta(I_1), \pi_\beta(I_3)\right\} \geq \exp\left(\Omega(n)\right)\pi_\beta(I_2).$$

We now use the facts $\pi_\beta(I_2) = \pi_\beta\left(\overline{I_2}\right)$, $\pi_\beta(I_3) = \pi_\beta\left(\overline{I_3}\right)$; and $\pi_\beta(I_1) + \pi_\beta(I_2) + \pi_\beta\left(\overline{I_2}\right) + \pi_\beta(I_3) + \pi_\beta\left(\overline{I_3}\right) \geq 1$ to arrive at $\left(\pi_\beta(I_1) + 2\pi_\beta\left(I_3\right)\right)\left(1 + \exp\left(-\Omega(n)\right)\right) \geq 1$. Consequently, we have $\pi_\beta(I_1) + 2\pi_\beta\left(I_3\right) \geq 1 + o_n(1)$. With this we conclude that

$$\pi_\beta(I_1) + \pi_\beta\left(I_3\right) \geq \frac{1}{2}(1 + o_n(1)),$$

as claimed.

Part (b)

Theorem 2.3.4(b) is a consequence of following proposition.

Proposition 2.6.13. *Let $\beta = \Omega\left(n2^{n\epsilon}\right)$. Then, for any $T > 0$, the "escape time" τ_β introduced in* (2.11) *satisfies*

$$\mathbb{P}\left(\tau_\beta \leq T\right) \leq T\exp\left(-\Omega\left(\beta 2^{-n\epsilon}\right)\right),$$

with high probability (over the randomness of $X \stackrel{d}{=} \mathcal{N}(0, I_n)$) as $n \to \infty$.

Proof of Proposition 2.6.13. The proof uses standard arguments and in particular is a straightforward adaptation of [21, Theorem 7.4] and [121, Theorem 3.2]. We reproduce it herein for completeness.

Let $\overline{X}_t$ be the Markov chain defined on $I_3 \cup \partial S$ which is X_t reflected at the boundary $\mathscr{A} \triangleq \partial\left(I_3 \cup \partial S\right)$ of $I_3 \cup \partial S$. Observe that

$$\sigma \in \mathscr{A} \iff d_H(\sigma, \sigma^*) = 1 \iff \frac{1}{n}\left\langle \sigma, \sigma^* \right\rangle = \frac{n-2}{n}.$$

We now specify the transition kernel $\overline{Q}(x, y)$ of $\overline{X}_t$. If $x \in \left(I_3 \cup \partial S\right) \setminus \mathscr{A}$, then $\overline{Q}(x, y) = Q(x, y)$ for any $y \in I_3 \cup \partial S$. If $x \in \mathscr{A}$, then $\overline{Q}(x, y) = Q(x, y)$ for $y \in I_3 \cup \partial S$; and $\overline{Q}(x, y) = 0$ otherwise. Note that by detailed balance, $\overline{X}_t$ is reversible with respect to $\pi_\beta\left(\cdot \mid I_3 \cup \partial S\right)$, namely the invariant measure of X_t conditioned on $I_3 \cup \partial S$.

We now couple the initialization of the chains; $\overline{X}_0 = X_0 \sim \pi_\beta\left(\cdot \mid I_3 \cup \partial S\right)$, to arrive at the conclusion that so long as $t \leq \tau_\beta$, almost surely X_t and $\overline{X}_t$ follow the same trajectory: $\overline{X}_t = X_t$ and $\overline{X}_t \sim \pi_\beta\left(\cdot \mid I_3 \cup \partial S\right)$.

We next proceed in the exact same manner as in [121, Theorem 3.2]. Note that

$$\mathbb{P}(\tau_\beta \leq T) \leq \mathbb{P}\big(\exists i \leq T : \tau_\beta = i, X_{i-1} \in \mathscr{A}\big) = \mathbb{P}\big(\exists i \leq T : \tau_\beta = i, \overline{X}_{i-1} \in \mathscr{A}\big)$$

from the definition of τ_β and the fact that the Markov Chains $\overline{X}_t$ and X_t, started from a common state in $I_3 \cup \partial S$, follow the same trajectory for $t \leq \tau_\beta$. Using the stationarity, namely the fact that $\overline{X}_t \sim \pi_\beta(\cdot \mid I_3 \cup \partial S)$, via a union bound; we further conclude

$$\mathbb{P}\big(\exists i \leq T : \tau_\beta = i, \overline{X}_{i-1} \in \mathscr{A}\big) \leq \mathbb{P}\big(\exists i \leq T : \overline{X}_{i-1} \in \mathscr{A}\big) \leq T\pi_\beta\big(\mathscr{A} \mid I_3 \cup \partial S\big).$$

Thus

$$\mathbb{P}(\tau_\beta \leq T) \leq T\pi_\beta\big(\mathscr{A} \mid I_3 \cup \partial S\big). \tag{2.162}$$

We now employ the FEW property to conclude the proof. Observe that $\mathscr{A} = \partial\left(I_3 \cup \partial S\right) \subset I_3 \cup \partial S$. Moreover, observe that $\mathscr{A} = \partial S \subset I_2$, hence $\pi_\beta(\mathscr{A}) \leq \pi_\beta(I_2)$; and $\pi_\beta(I_3 \cup \partial S) \geq \pi_\beta(I_3)$. Combining these, we obtain

$$\pi_\beta\left(\mathscr{A} \mid I_3 \cup \partial S\right) = \frac{\pi_\beta\left(\mathscr{A}\right)}{\pi_\beta\left(I_3 \cup \partial S\right)} \leq \frac{\pi_\beta\left(I_2\right)}{\pi_\beta\left(I_3\right)} \leq \exp\left(-\Omega\left(\beta 2^{-n\epsilon}\right)\right). \tag{2.163}$$

The last inequality uses the FEW property established in Theorem 2.3.3. Combining (2.162) and (2.163) we conclude the proof:

$$\mathbb{P}\left(\tau_\beta \leq T\right) \leq T \exp\left(-\Omega\left(\beta 2^{-n\epsilon}\right)\right).$$

□

With this, the proof of Theorem 2.3.4 is complete. □

Chapter 3

Algorithms and Barriers in the Symmetric Binary Perceptron Model

3.1 Introduction

In this chapter, we study the *perceptron model*. Proposed initially in the 1960's [180, 289, 287, 83], this is a toy model of one-layer neural network storing random patterns as well as a very natural model in high-dimensional probability. Let $X_i \in \mathbb{R}^n$, $1 \leq i \leq M$, be i.i.d. random patterns to be stored. *Storage* of these patterns is achieved if one finds a vector of synaptic weights $\sigma \in \mathbb{R}^n$ consistent with all X_i: that is, $\langle X_i, \sigma \rangle \geq 0$ for $1 \leq i \leq M$. There are two main variants of the perceptron: when the vector σ lies on on sphere in $\mathbb{R}^n$ (the spherical perceptron) and when $\sigma \in \mathcal{B}_n \triangleq \{-1, 1\}^n$ (the binary or Ising perceptron). For more on the spherical perceptron see [139, 254, 269, 274, 11]; in this chapter we will focus only on the binary perceptron.

A key quantity associated to the perceptron is the *storage capacity*: the maximum number M^* of such patterns for which there exists a vector of weights $\sigma \in \mathcal{B}_n$ that is consistent with all X_i, $1 \leq i \leq M^*$. Investigations beginning with Gardner [138, 139] and Gardner-Derrida [140] in the statistical physics literature provided a detailed, yet non-rigorous, picture for the storage capacity in the case of patterns distributed as n-dimensional Gaussian vectors.

More general perceptron models are defined by an activation function $U : \mathbb{R} \to \{0, 1\}$[1]. We say a pattern X_i is stored by σ with respect to U if $U(\langle X_i, \sigma \rangle) = 1$. Much recent work on these models have focused on two classes of activity functions: $U(x) = \mathbf{1}_{x \geq \kappa\sqrt{n}}$ and $U(x) = \mathbf{1}_{|x| \leq \kappa\sqrt{n}}$. The first defines the *asymmetric binary perceptron*, the second the *symmetric binary perceptron*. We now detail some of the previous work on these models.

[1]For an even more general setting see [54].

3.1.1 Perceptron models

Asymmetric Binary Perceptron

We now define the classic binary perceptron, which we call the *asymmetric binary perceptron* (ABP) throughout. Fix $\kappa \in \mathbb{R}$, $\alpha > 0$; and set $M = \lfloor n\alpha \rfloor \in \mathbb{N}$. Let $X_i \overset{d}{=} \mathcal{N}(0, I_n)$, $1 \leq i \leq M$, be i.i.d. random vectors, where $\mathcal{N}(0, I_n)$ denotes the $n-$dimensional multivariate normal distribution with zero mean and identity covariance. Consider the (random) set

$$S_\alpha^{\mathrm{A}}(\kappa) \triangleq \bigcap_{1 \leq i \leq M} \Big\{ \sigma \in \mathcal{B}_n : \langle \sigma, X_i \rangle \geq \kappa \sqrt{n} \Big\}. \tag{3.1}$$

The vectors $X_i \in \mathbb{R}^n$, $1 \leq i \leq M$, are collectively referred to as the *disorder*. In what follows, we slightly abuse the terminology and use "disorder" to refer to both the vectors X_i, $1 \leq i \leq M$; as well as the matrix $\mathcal{M} \in \mathbb{R}^{M \times n}$ whose rows are X_i. The set $S_\alpha^{\mathrm{A}}(\kappa)$ is the *solution space*, a random subset of $\mathcal{B}_n$.

The computer science take on the perceptron model is to view it as an instance of a *random constraint satisfaction problem*. Indeed, observe that $S_\alpha^{\mathrm{A}}(\kappa)$ is an intersection of M random halfspaces, each defined by the *constraint vector* X_i (and threshold κ). Each constraint rules out certain solutions in the space $\mathcal{B}_n$ of all possible solutions; and the parameter α plays a role akin to the constraint density in the literature on random $k-$SAT, see e.g. [23, 234, 6] for more discussion. For these reasons, we refer to α as the *constraint density* in the sequel.

Perhaps the most important *structural* question is whether $S_\alpha^{\mathrm{A}}(\kappa)$ is empty/non-empty (w.h.p., as $n \to \infty$). Krauth and Mézard conjectured in [190] that the event, $\big\{ S_\alpha^{\mathrm{A}}(\kappa) \neq \varnothing \big\}$, exhibits what is known as a *sharp threshold*: there is an explicit threshold $\alpha_{\mathrm{KM}}(\kappa)$ such that

$$\lim_{n \to \infty} \mathbb{P}\big[S_\alpha^{\mathrm{A}}(\kappa) \neq \varnothing \big] = \begin{cases} 0, & \text{if } \alpha > \alpha_{\mathrm{KM}}(\kappa) \\ 1, & \text{if } \alpha < \alpha_{\mathrm{KM}}(\kappa) \end{cases}. \tag{3.2}$$

Using non-rigorous calculations based on the so-called *replica method*, Krauth and Mézard [190] conjecture a precise value of $\alpha_{\mathrm{KM}}(0)$ around 0.833. It is worth noting that this value deviates significantly from the *first moment threshold*: note that for $\kappa = 0$, $\mathbb{E}\big[\big| S_\alpha(\kappa) \big| \big] = \exp_2\big(n - n\alpha \big)$, which is exponentially small (in n) only for $\alpha > 1$.

The structure of $S_\alpha^{\mathrm{A}}(\kappa)$ and the aforementioned phase transition still (largely) remain as open problems. Even the very existence of such a sharp phase transition point remains open, though Xu[2] [291] has shown sharpness of the threshold around a possibly n-dependent value $\alpha_c^{(n)}(\kappa)$, as in [116] in the setting of random CSP's. With that in mind, we can define

$$\alpha_c^*(\kappa) = \inf \Big\{ \alpha : \lim_{n \to \infty} \mathbb{P}\Big[S_\alpha^{\mathrm{A}}(0) = \varnothing \Big] = 1 \Big\}.$$

[2]Xu establishes this in a slightly different setting, where the disorder X_i consists of i.i.d. Rademacher entries.

The work by Ding and Sun [95] establishes, using an elegant second-moment argument, that for every $\alpha \leq \alpha_{\rm KM}(0)$,

$$\liminf_{n\to\infty} \mathbb{P}\Big[S_\alpha^{\rm A}(0) \neq \varnothing\Big] > 0.$$

Hence, $\alpha_c^*(0) \geq \alpha_{\rm KM}(0)$. However, a matching upper bound is still missing: the best known bound is due to Kim and Roche [188, Theorem 1.2], which show $\alpha_c^*(0) \leq 0.9963$. More precisely, they show for any $\epsilon < 0.0037$, $\mathbb{P}\Big[S_{1-\epsilon}^{A}(0) \neq \varnothing\Big] = o(1)$. For a similar negative result with a stronger convergence guarantee; that is a guarantee of form

$$\mathbb{P}\Big[S_{1-\delta}^{A}(0) \neq \varnothing\Big] \leq \exp\big(-\delta n\big)$$

for some small $\delta > 0$ (though potentially worse than 0.0037), see Talagrand [271].

When $S_\alpha^{\rm A}(\kappa) \neq \varnothing$ (w.h.p.), a follow-up algorithmic question is whether such a satisfying $\sigma \in \mathcal{B}_n$ can be found algorithmically (in polynomial time). Regarding such positive results, the best known guarantee is again due to Kim and Roche. They devise in [188] an (multi-stage majority) algorithm that w.h.p. returns a solution $\sigma \in S_\alpha^{\rm A}(0)$ as long as $\alpha < 0.005$. (In particular, their algorithm is a constructive proof that $S_\alpha^{\rm A}(0) \neq \varnothing$ w.h.p. for $\alpha < 0.005$.) Later in Section 3.3.4, we informally describe the implementation of their algorithm and establish that it is stable in an appropriate sense.

Symmetric Binary Perceptron

Proposed initially by Aubin, Perkins, and Zdeborová in [23]; the symmetric binary perceptron (SBP) model is our main focus in the present chapter. Similar to the asymmetric case, fix a $\kappa > 0$, $\alpha > 0$; and set $M = \lfloor n\alpha \rfloor$. Let $X_i \overset{d}{=} \mathcal{N}(0, I_n)$, $1 \leq i \leq M$, be i.i.d. random vectors, and consider

$$S_\alpha(\kappa) \triangleq \bigcap_{1\leq i\leq M} \Big\{\sigma \in \mathcal{B}_n : \big|\langle \sigma, X_i\rangle\big| \leq \kappa\sqrt{n}\Big\} = \Big\{\sigma \in \mathcal{B}_n : \big\|\mathcal{M}\sigma\big\|_\infty \leq \kappa\sqrt{n}\Big\}, \quad (3.3)$$

where $\mathcal{M} \in \mathbb{R}^{M\times n}$ with rows $X_1, \ldots, X_M$. This model is called *symmetric* since $\sigma \in S_\alpha(\kappa)$ iff $-\sigma \in S_\alpha(\kappa)$. It turns out that the symmetry makes the SBP more amenable to analysis compared to its asymmetric counterpart, while retaining the relevant conjectural structural properties nearly intact, see [29]. Though not our focus here, it is worth mentioning that this is analogous to the random $k-$SAT model. Its symmetric variant, NAE $k-$SAT, is mathematically more tractable, yet at the same time exhibits similar structural properties.

As its asymmetric counterpart, it was conjectured that the SBP also undergoes a *sharp phase transition*. More concretely, it was conjectured that there exists a $\alpha_c(\kappa)$ such that the event, $\{S_\alpha(\kappa) \neq \varnothing\}$, undergoes a sharp phase transition as α crosses

$\alpha_c(\kappa)$. Notably, $\alpha_c(\kappa)$ matches with the first moment prediction:

$$\alpha_c(\kappa) \triangleq -\frac{1}{\log_2 \mathbb{P}\big[|Z| \le \kappa\big]}, \quad \text{where} \quad Z \sim \mathcal{N}(0,1). \tag{3.4}$$

It was established in [23] that (a) $\lim_{n\to\infty} \mathbb{P}\big[S_\alpha(\kappa) \neq \varnothing\big] = 0$ for $\alpha > \alpha_c(\kappa)$; and (b) $\liminf_{n\to\infty} \mathbb{P}\big[S_\alpha(\kappa) \neq \varnothing\big] > 0$ for $\alpha < \alpha_c(\kappa)$. The latter guarantee uses the so-called *second moment method*, though falling short of establishing the high probability guarantee. Subsequent works by Perkins and Xu [234]; and Abbe, Li, and Sly [6] establish that $\mathbb{P}\big[S_\alpha(\kappa) \neq \varnothing\big] = 1 - o(1)$ for all $\alpha < \alpha_c(\kappa)$. Namely, $\alpha_c(\kappa)$ is indeed a sharp threshold for the SBP. Having established the existence and the location of such a sharp phase transition; the next question, once again, is whether such a $\sigma \in S_\alpha(\kappa)$ can be found *efficiently*; that is, by means of polynomial-time algorithms. This is our main focus in the present chapter.

The SBP is closely related to *combinatorial discrepancy theory* [268, 213]. Given a matrix $\mathcal{M} \in \mathbb{R}^{M\times n}$, a central problem in discrepancy theory is to compute, approximate, or bound its *discrepancy* $\mathcal{D}(\mathcal{M})$:

$$\mathcal{D}(\mathcal{M}) \triangleq \min_{\sigma\in\mathcal{B}_n} \big\|\mathcal{M}\sigma\big\|_\infty.$$

Several different settings are considered in the discrepancy literature: *worst-case* $\mathcal{M}$ and *average-case* $\mathcal{M}$ (where the entries of $\mathcal{M}$ either i.i.d. Rademacher or i.i.d. Gaussian); and both existential and algorithmic results are sought. In the *proportional regime*, the discrepancy perspective is to fix the aspect ratio $\alpha = M/n$ and find a solution σ with small $\|\mathcal{M}\sigma\|_\infty$. This is the inverse of the perceptron perspective: fixing $\kappa > 0$ and finding the largest α for which a solution σ exists. In particular, the sharp threshold result for the SBP described above settles the question of discrepancy in the random proportional regime: for $\mathcal{M} \in \mathbb{R}^{M\times n}$ with i.i.d. $\mathcal{N}(0,1)$ entries, $\mathcal{D}(\mathcal{M}) = (1+o(1))f(\alpha)\sqrt{n}$ w.h.p. where $f(\cdot)$ is the inverse function of α_c. The first and second moment methods can also be employed to establish the value of discrepancy in the random setting in other regimes, e.g. [238, 278, 14]. Moreover, as we describe below, discrepancy algorithms (e.g. [32, 211, 71, 36, 239]) can be employed for the SBP.

3.1.2 Main Results

From an algorithmic point of view, the most striking fact about the SBP is the existence of a large *statistical-to-computational gap*. Explanations for both the algorithmic hardness of the model and for the success of efficient algorithms at low densities have been put forth recently.

A Statistical-to-Computational Gap. A random constraint satisfaction problem like the SBP is said to exhibit a *statistical-to-computational gap* if the density below which solutions are known to exist w.h.p. is higher than the densities at which known efficient algorithms can find a solution. As we now demonstrate, the SBP ex-

hibits a statistical-to-computational gap for all $\kappa > 0$, but this gap is most pronounced in the regime of small κ. In this regime, the best known algorithmic guarantee for finding a solution in the SBP is due to Bansal and Spencer [36] from the literature on combinatorial discrepancy. As we detail in Section 3.3.3 and show in Corollary 3.3.6, their algorithm works for $\alpha = O(\kappa^2)$ as $\kappa \to 0$. This stands in stark contrast to the threshold for the existence of solutions. From (3.4), $\alpha_c(\kappa)$ behaves like $\frac{1}{\log_2(1/\kappa)}$:

$$\alpha_c(\kappa) = -\frac{1}{\log_2 \mathbb{P}\big[|Z| \le \kappa\big]} = -\frac{1}{\frac{1}{2}\log_2 \frac{2}{\pi} + \log_2\big(1 + o_\kappa(1)\big) + \log_2 \kappa} = \frac{1}{\log_2(1/\kappa)}\big(1+o_\kappa(1)\big).$$

Namely, $\alpha_c(\kappa)$ is asymptotically much larger than the algorithmic κ^2 threshold. The main motivation of the present chapter is to inquire into the origins of this gap in the SBP by leveraging insights from statistical physics. In particular, we will establish the presence of a geometric property known as the *Overlap Gap Property* (OGP), and use it to rule out classes of *stable algorithms*, appropriately defined.

Freezing, rare clusters, and algorithms. The SBP exhibits striking structural properties which are thought to contribute to both the success of polynomial-time algorithms at low densities and the failure of efficient algorithms at higher densities.

On one hand, the model exhibits the "frozen one-step Replica Symmetry Breaking (1-RSB)" scenario at all positive densities $\alpha < \alpha_c$. This states that whp over the instance, almost every solution σ is *totally frozen* and isolated: the nearest other solution is at linear Hamming distance to σ. This extreme form of clustering was conjectured to hold for the ABP and SBP in [190, 174, 23, 29], and subsequently established for the SBP in [234, 6]. In light of the earlier works by Mézard, Mora, and Zecchina [217] and Achlioptas and Ricci-Tersenghi [9] positing a link between clustering, freezing, and algorithmic hardness, it is tempting to postulate that finding a solution σ for the SBP is hard for every $\alpha \in (0, \alpha_c(\kappa))$, but this is contradicted by the existence of efficient algorithms at low densities such as that of [36, 60, 28, 26, 27] including the algorithm by Bansal and Spencer discussed above. Combining these facts, we arrive at the conclusion that the SBP exhibits an intriguing phenomenon: the existence of polynomial-time algorithms can coexist with the frozen 1-RSB phenomenon. This conundrum challenges the view that clustering and freezing necessarily lead to algorithmic hardness.

In an attempt to explain this apparent conundrum, it was conjectured in [30] that while a $1 - o(1)$ fraction of all solutions are totally frozen, an exponentially small fraction of solutions appear in clusters of exponential (in n) size; and the efficient learning algorithms that manage to find solutions find solutions belonging to such rare clusters, see [234] for further discussion. In this direction, Abbe, Li, and Sly [5] established very recently that whp a connected cluster of solutions of linear diameter does indeed exist at *all* densities $\alpha < \alpha_c$. Furthermore, they show that an efficient multi-stage majority algorithm (based on that of [188]) can find such a large cluster at densities $\alpha = O(\kappa^{10})$ in the $\kappa \to 0$ regime[3].

[3]See in particular α_0 appearing in [5, Page 6].

These results and conjectures prompt several questions regarding the statistical-to-computational gap exhibited by the SBP. If large connected clusters exist at all subcritical densities, what is the reason for the apparent algorithmic hardness? Do the efficient algorithms for densities $\alpha = O(\kappa^2)$ also find solutions lying in one of these large connected clusters? At what densities are these large clusters algorithmically accessible? In particular, while we now know detailed structural information about the SBP, its statistical-to-computational gap remains a mystery.

Our results on the Overlap Gap Property and failure of stable algorithms. We investigate the statistical-to-computational gap in the SBP via the *Overlap Gap Property (OGP)*, an intricate geometrical property of the solution space that has been used to rigorously rule out large classes of search algorithms for many important random computational problems including random $k-$SAT [137, 78, 63] and independent sets in sparse random graphs [135, 245, 285], see also the survey paper by Gamarnik [119]. We will describe the OGP in more detail below. At a high level, it asserts the non-existence of tuples of solutions at prescribed distances in the solution space.

Our first main result establishes the OGP for $m-$tuples of solutions (dubbed as $m-$OGP) at densities $\Omega(\kappa^2 \log_2 \frac{1}{\kappa})$:

Theorem 3.1.1 (Informal, see Theorem 3.2.4)**.** *For densities $\alpha = \Omega(\kappa^2 \log_2 \frac{1}{\kappa})$, the SBP exhibits the $m-$OGP for appropriately chosen parameters.*

We also establish the presence of $2-$OGP and the $3-$OGP for the SBP in the high κ regime, i.e. when $\kappa = 1$, respectively in Theorem 3.2.2 and Theorem 3.2.3. As we show in Theorem 3.5.2 through the multi-dimensional version of Berry-Esseen Theorem, our OGP results enjoy *universality*: they remain valid under milder distributional assumptions on the entries of $\mathcal{M}$.

Our next main result shows that the $m-$OGP rules out the class of *stable algorithms* formalized in Definition 3.3.1. At a high level, an algorithm is stable if a small perturbation of its input results in a small perturbation of the solution σ it outputs. In the literature on other random computational problems, it has been shown that the class of stable algorithms captures powerful classes of algorithms including Approximate Message Passing algorithms [120], low-degree polynomials [122, 63], and low-depth circuits [124].

Theorem 3.1.2 (Informal, see Theorem 3.3.2)**.** *The $m-$OGP implies the failure of stable algorithms for the SBP.*

Thus, we obtain the following corollary:

Corollary 3.1.3 (Informal, see Theorem 3.3.2)**.** *Stable algorithms (with appropriate parameters) fail to find a solution for the SBP for densities $\alpha = \Omega(\kappa^2 \log_2 \frac{1}{\kappa})$.*

In particular, this hardness result matches the algorithmic κ^2 threshold up to a logarithmic factor. Hence, while the view that freezing implies algorithmic hardness

for the SBP breaks down, the rigorous link between the OGP and algorithmic hardness remains intact.

In addition to stable algorithms; we also consider the class of *online algorithms* which includes the Bansal-Spencer algorithm [36]. Informally, an algorithm $\mathcal{A}$ is online if the t^{th} coordinate of the solution it outputs depends only on the first t columns of $\mathcal{M}$.

Theorem 3.1.4 (Informal, see Theorem 3.3.4)**.** *Online algorithms fail to find a solution for the SBP for sufficiently high densities.*

Having established the hardness of stable algorithms for the SBP at the $m-$OGP threshold; a natural follow-up question is whether the known efficient algorithms for perceptron models are stable and whether the ABP also exhibits the $m-$OGP. To that end, we investigate the stability property of the Kim-Roche algorithm [188] for the ABP.

Theorem 3.1.5 (Informal, see Theorem 3.3.8)**.** *The Kim-Roche algorithm [188] for the ABP is stable in the sense of Definition 3.3.1.*

Investigating the stability of the Bansal-Spencer algorithm [36] and whether the ABP also exhibits the OGP are among several open questions we discuss in Section 3.1.4.

3.1.3 Background and Related Work

Statistical-to-Computational Gaps. As we noted, the SBP model exhibits a *statistical-to-computational gap* (SCG): a gap between what the existential results guarantee (and thus what can be found with unbounded computational power), and what algorithms with bounded computational power (such as polynomial-time algorithms) can promise. Such SCGs are a ubiquitous feature in many algorithmic problems (with *random inputs*) appearing in high-dimensional statistical inference tasks and in the study of random combinatorial structures. A partial (and certainly incomplete) list of problems with an SCG includes constraint satisfaction problems [217, 9, 7], optimization problems over random graphs [135, 77, 136] and spin glass models [73, 122, 120, 124], number partitioning problem [126], principal component analysis [47, 199, 200], and the "infamous" planted clique problem [177, 92, 38]; see also the introduction of [126], the recent survey [119]; and the references therein.

Unfortunately, due to the so-called *average-case* nature of these problems, the standard NP-completeness theory often fails to establish hardness for those problems even under the assumption $P \neq NP$. (It is worth noting though that a notable exception to this is when the problem exhibits *random self-reducibility*, see e.g. [133] for such a hardness result regarding a spin glass model, conditional on a weaker assumption $P \neq \#P$.) Nevertheless, a very fruitful (and still active) line of research proposed certain forms of *rigorous evidences* of algorithmic hardness for such average-case problems. These approaches include the failure of Monte Carlo Markov Chain methods [177, 103], low-degree methods and failure of low-degree polynomials [169,

193, 122, 285, 63], Sum-of-Squares [168, 167, 240, 38] and Statistical Query [186, 94, 111, 112] lower bounds, failure of the approximate message passing algorithm (an algorithm that is information-theoretically optimal for certain important problems, see e.g. [91, 90]) [293, 31]; and the reductions from the planted clique problem [47, 62, 61], just to name a few. Yet another very promising such approach is through the *intricate geometry* of the problem, via the so-called *Overlap Gap Property* (OGP).

Overlap Gap Property (OGP). Implicitly discovered by Mézard, Mora, and Zecchina [217] and Achlioptas and Ricci-Tersenghi [9] (though coined later in [134]), the OGP approach leverages insights from the statistical physics to form a rigorous link between the intricate geometry of the solution space and formal algorithmic hardness. Informally, the OGP is a topological disconnectivity property, and states (in the context of a random combinatorial optimization problem, say over $\mathcal{B}_n$) that (w.h.p. over the randomness) any two near-optimal $\sigma_1, \sigma_2 \in \mathcal{B}_n$ are either "close" or "far" from each other: there exists $0 < \nu_1 < \nu_2 < 1$ such that $n^{-1}\langle \sigma_1, \sigma_2 \rangle \in [0, \nu_1] \cup [\nu_2, 1]$. That is, their (normalized) overlaps do not admit *intermediate values*; and no two near-optimal solutions of intermediate distance can be found. It has been shown (see below) that the OGP is a rigorous barrier for large classes of algorithms. See [119] for a survey on OGP.

Algorithmic Implications of OGP. The line of research relating the OGP to algorithmic hardness was initiated by Gamarnik and Sudan [135, 136]. They consider the problem of finding a large independent set in the sparse random graphs with average degree d. It is known, see e.g. [117, 118, 45], that in the double limit (first sending $n \to \infty$, then letting $d \to \infty$), the largest independent set of this graph is of size $2\frac{\log d}{d}n$. On the other hand, the best known polynomial-time algorithm [184] (a very simple greedy protocol) returns an independent set that is half optimal, namely of size $\frac{\log d}{d}n$. In order to reconcile this apparent SCG, Gamarnik and Sudan study the space of all large independent sets. They establish that any two independent sets of size greater than $(1+1/\sqrt{2})\frac{\log d}{d}n$ exhibit OGP. By leveraging this, they show, through a contradiction argument, that *local algorithms* (known as the *factors of i.i.d.*) fail to find an independent set of size greater than $(1+1/\sqrt{2})\frac{\log d}{d}n$. Subsequent research (again via the lens of OGP) extended this hardness result to the class of low-degree polynomials [122]. The extra "oversampling" factor, $1/\sqrt{2}$, was removed by inspecting instead the the overlap pattern of many large independent sets (rather than the pairs), therefore establishing hardness all the way down to the algorithmic threshold. This was done by Rahman and Virág [245] for *local algorithms*, and by Wein [285] for *low-degree polynomials*; and is also our focus here (see below). A list of problems where the OGP is leveraged to rule out certain classes of algorithms includes optimization over random graphs and spin glass models [120, 122, 124, 172], number partitioning problem [126], random constraint satisfaction problems [137, 63].

Multi OGP ($m-$OGP). As we just mentioned, it was previously observed that by considering more intricate overlap patterns, one can potentially lower the (algo-

rithmic) phase transition points further. This idea was employed for the first time by Rahman and Virág [245] in the context of the aforementioned independent set problem. They managed to "shave off" the extra $1/\sqrt{2}$ factor present in the earlier result by Gamarnik and Sudan [135, 136], and reached all the way down to the algorithmic threshold, $\frac{\log d}{d}n$. In a similar vein, Gamarnik and Sudan [137] studied the overlap structure of $m-$tuples $\sigma^{(i)} \in \mathcal{B}_n$, $1 \leq i \leq m$ of satisfying assignments in the context of the Not-All-Equal (NAE) $k-$SAT problem. By showing the presence of OGP for $m-$tuples of nearly equidistant points (in $\mathcal{B}_n$), they established nearly tight hardness for *sequential local algorithms*: their results match the computational threshold modulo factors that are polylogarithmic (in k). A similar overlap pattern (for $m-$tuples consisting of nearly equidistant points) was also considered by Gamarnik and Kızıldağ [126] in the context of random number partitioning problem (NPP), where they established hardness well below the existential threshold. (It is worth noting that [126] considers $m-$tuples where m itself also grows in n, $m = \omega_n(1)$.)

More recently, $m-$OGP for more intricate forbidden patterns were considered to establish formal hardness in other settings. In particular, by leveraging $m-$OGP, Wein [285] showed that low-degree polynomials fail to return a large independent set (in sparse random graphs) of size greater than $\frac{\log d}{d}n$, thereby strengthening the earlier result by Gamarnik, Jagannath, and Wein [122]. Wein's work establishes the *ensemble* variant of OGP (an idea emerged originally in [73]): he considers $m-$tuples of independent sets where each set do not necessarily come from the same random graph, but rather from correlated random graphs. The ensemble variant of OGP was also considered in [126] for the NPP. While technically more involved to establish, it appears that the ensemble $m-$OGP can be leveraged to rule out virtually any *stable algorithm* (appropriately defined); and will also be our focus here. More recently, by leveraging the ensemble $m-$OGP; Bresler and Huang [63] established nearly tight low-degree hardness results for the random $k-$SAT problem: they show that low-degree polynomials fail to return a satisfying assignment when the clause density is only a constant factor off by the computational threshold. In yet another work, Huang and Sellke [172] construct a very intricate forbidden structure consisting of an ultrametric tree of solutions, which they refer to as the *branching OGP*. By leveraging this branching OGP, they rule out overlap concentrated algorithms[4] at the algorithmic threshold for the problem of optimizing mixed, even $p-$spin model Hamiltonian.

3.1.4 Open Problems

Location of the Algorithmic Threshold. We establish in Theorem 3.2.4 that the SBP exhibits $m-$OGP if $\alpha = \Omega\big(\kappa^2 \log_2 \frac{1}{\kappa}\big)$. On the other hand, we have per Corollary 3.3.6 that the Bansal-Spencer algorithm [36] works when $\alpha = O(\kappa^2)$. In light of these, we make the following conjecture:

Conjecture 3.1.6. *As $\kappa \to 0$, the algorithmic threshold for the SBP is at $\widetilde{\Theta}(\kappa^2)$.*

[4] A class that captures $O(1)$ iterations of gradient descent, approximate message passing; and Langevin Dynamics run for $O(1)$ time.

In particular, we conjecture that up to factors that are polylogarithmic in $\frac{1}{\kappa}$, the Bansal-Spencer algorithm is the best possible within the class of efficient algorithms. That is, up to polylogarithmic factors no polynomial-time algorithm succeeds above the $m-$OGP threshold. An interesting question is whether the $\log_2 \frac{1}{\kappa}$ factor is necessary or it can be 'shaved off'. We believe it might be possible to remove this factor by considering a more intricate overlap pattern, e.g. similar to those considered in [285, 63, 172].

We now make Conjecture 3.1.6 more precise. Given an $m \in \mathbb{N}$ and $\kappa > 0$, let $\alpha_m^*(\kappa)$ be the smallest subcritical density such that the SBP exhibits $m-$OGP with appropriate parameters when $\alpha \geq \alpha_m^*$. We conjecture that the $m \to \infty$ limit of the $m-$OGP threshold marks the true algorithmic threshold: for every $\epsilon > 0$ and κ small enough, there do not exist polynomial-time algorithms for the SBP when $\alpha \geq (1+\epsilon) \lim_{m\to\infty} \alpha_m^*(\kappa)$. See Conjecture 3.3.7 for details. This conjecture is backed up by the evidence that for many random computational problems including random $k-$SAT [63], independent sets in sparse random graphs [245, 285], and mixed even $p-$spin model [172], the $m-$OGP matches or nearly matches the best known algorithmic threshold.

Abbe, Li and Sly ask in [5, Question 1] whether the algorithmic threshold for the SBP coincides with the threshold for the existence of a 'wide web': a cluster of solutions with maximum possible diameter n. One one hand, the existence of a wide web rules out the 2-OGP: pairs of solutions of every possible overlap exist. It would be very interesting to determine whether the threshold for existence of the wide web coincides with the conjectured algorithmic threshold of $\tilde{\Theta}(\kappa^2)$ above, or even more precisely the limiting m-OGP threshold $\lim_{m\to\infty} \alpha_m^*$ (at least asymptotically as $\kappa \to 0$).

The Asymmetric Model. As we noted earlier, the ABP is more challenging from a mathematical perspective, and some of its basic properties are still far from being rigorously understood. In particular, even the very existence of a sharp phase transition and the frozen 1-RSB picture—both rigorously known to hold for the SBP—remain open.

The ABP also exhibits a statistical-to-computational gap. On one hand, Kim-Roche algorithm [188] finds solutions at low enough densities, specifically when $\alpha < 0.005$. On the other hand, the result of Ding and Sun [95] shows that solutions do exist (with probability bounded away from 0) when $\alpha < \alpha_{\mathrm{KM}}(0) \approx 0.83$. It would be interesting to show that the ABP exhibits $m-$OGP for some densities $\alpha < \alpha_{\mathrm{KM}}(0)$. To understand the statistical-to-computational gap of ABP further, it would be interesting to explore the model in the regime $\kappa \to \infty$ and investigate the $m-$OGP threshold and threshold for the existence of efficient algorithms. Further, there are other perceptron models one could explore in this regard, e.g. the *$U-$function binary perceptron* introduced in [23].

Stability of Other Algorithms. We established in Theorem 3.3.8 that the Kim-Roche algorithm for ABP is stable. In light of this, we make the following conjecture regarding the SBP:

Conjecture 3.1.7. *There exists a stable algorithm that finds a solution for the* SBP *w.h.p. when* $\alpha = O(\kappa^2)$.

In particular, proving stability of the Bansal-Spencer algorithm would resolve Conjecture 3.1.7, but this seems challenging: the presence of a certain non-linear potential function (see [36, Equation 2.5]) renders the stability analysis difficult.

The algorithm of [5] is a variant of the Kim-Roche algorithm that works for the SBP for $\alpha = O\big(\kappa^{10}\big)$. Proving the stability of this algorithm would be an interesting first step towards resolving Conjecture 3.1.7.

Broader Research Agendas on the OGP. As mentioned above, the OGP is a provable barrier for a broad class of algorithms for many random computational problems. A list of such algorithms includes local/sequential local algorithms, Monte Carlo Markov Chain (MCMC) methods, low-degree polynomials, Langevin dynamics, approximate message passing type algorithms, low-depth circuits, and stable algorithms in general. In many random computational problems (like k-SAT and independent sets) the OGP coincides with the threshold for the existence of known efficient search algorithms. One might then conjecture (as we do here) that the OGP marks the true algorithmic threshold. It would thus be very surprising and very interesting to find a case where efficient algorithms succeed in the face of the OGP[5]. While random k-SAT, independent sets in random graphs, and other random CSP's have been studied for decades without finding such algorithms, algorithms for perceptron models have not been studied as extensively, especially not in the limiting regime $\kappa \to 0$ we focus on here, and thus this might be fruitful direction to pursue.

3.1.5 Organization and Notation

Chapter Organization. The rest of the chapter is organized as follows. Our OGP results are stated in Section 3.2. In particular, we establish $2-$OGP and $3-$OGP for the high κ case ($\kappa = 1$) in Section 3.2.2; and the $m-$OGP for the regime $\kappa \to 0$ in Section 3.2.3. We then take an algorithmic route, and establish our main hardness result in Section 3.3.1; and formulate a conjecture pertaining the true algorithmic threshold in Section 3.3.3. In Section 3.3.4 we describe the Kim-Roche algorithm and show that it is stable. We record certain limitations of our approach in Section 3.4. We show in Section 3.5 that our OGP results enjoy *universality* and extend beyond the Gaussian disorder. We provide complete proofs in Section 3.6. Finally in Appendix A, we provide a MATLAB code for verifying Lemma 3.6.1 using which we establish $2-$OGP and $3-$OGP for $\kappa = 1$.

Notation. For any $n \in \mathbb{N}$, $[n] \triangleq \{1, 2, \ldots, n\}$. The binary cube $\{-1, 1\}^n$ is denoted by $\mathcal{B}_n$. For any set A, $|A|$ denotes its cardinality. For any $r > 0$ and $x \in \mathbb{R}$; $\exp_r(x)$ and $\log_r(x)$ denote respectively the exponential and logarithm functions base r. For

[5]Beyond those cases where algebraic techniques like Gaussian elimination can find solutions to 'noiseless' problems like solving random linear equations.

any $v = (v_i : 1 \le i \le n) \in \mathbb{R}^n$ and $p > 0$, $\|v\|_p \triangleq \left(\sum_{1\le i\le n} |v_i|^p\right)^{1/p}$, and $\|v\|_\infty = \max_{1\le i\le n} |v_i|$. For any $v, v' \in \mathbb{R}^n$, $\langle v, v'\rangle \triangleq \sum_{1\le i\le n} v_i v_i'$ and $\mathcal{O}vv' \triangleq n^{-1}\langle v, v'\rangle$. For any $\sigma, \sigma' \in \mathcal{B}_n$, $d_H(\sigma, \sigma')$ denotes their Hamming distance. For $k \in \mathbb{N}$, $\mathbf{e} \in \mathbb{R}^k$ denotes the vector of all ones (where the dimension will be clear from the context); and I_k denotes the $k \times k$ identity matrix. For any $x \in \mathbb{R}$, $\lfloor x \rfloor$ and $\lceil x \rceil$ respectively denote its floor and ceil. For $p \in [0,1]$, $h(p) \triangleq -p\log_2 p - (1-p)\log_2(1-p)$ is the binary entropy function (logarithm base two). For any $n \in \mathbb{N}$, $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$; $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the $n-$dimensional random vector having multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. For any event $\mathcal{E}$, $\mathbb{1}\{\mathcal{E}\}$ denotes its indicator. Given a matrix $\mathcal{M}$; $\|\mathcal{M}\|_F$, $\|\mathcal{M}\|_2$, $\sigma(\mathcal{M})$, $\sigma_{\min}(\mathcal{M})$, $\sigma_{\max}(\mathcal{M})$, $|\mathcal{M}|$ and trace$(\mathcal{M})$ denote, respectively, its Frobenius norm, spectral norm, spectrum (that is, the set of its eigenvalues), smallest and largest singular values, determinant, and trace. A graph $\mathbb{G} = (V, E)$ is a collection of vertices V together with some edges $(v, v') \in E$ between $v, v' \in V$. We consider herein only the simple graphs, namely those that are undirected with no loops. A graph $\mathbb{G} = (V, E)$ is called a *clique* if for every distinct $v, v' \in V$, $(v, v') \in E$. We denote the clique on $m-$vertices ($m \in \mathbb{N}$) by K_m. A subset $S \subset V$ of vertices (of a $\mathbb{G} = (V, E)$) is called an *independent set* if for every distinct $v, v' \in V$, $(v, v') \notin E$. The largest cardinality of such an independent set is called the *independence number* of $\mathbb{G}$, denoted $\alpha(\mathbb{G})$. A $q-$coloring of a graph $\mathbb{G} = (V, E)$ is a function $\varphi : E \to \{1, 2, \ldots, q\}$ assigning to each $e \in E$ one of q available colors.

Throughout the chapter, we employ the standard Bachmann-Landau asymptotic notation, e.g. $\Theta(\cdot), O(\cdot), o(\cdot)$, and $\Omega(\cdot)$. If there is no subscript, the asymptotic is with respect to $n \to \infty$. In the case where we consider asymptotics other than $n \to \infty$, we reflect this by a subscript: for instance, if f is a function such that $f(\kappa) \to \infty$ as $\kappa \to 0$, we denote this by $f = \omega_\kappa(1)$. To keep our exposition clean, we omit floor/ceiling signs whenever appropriate.

3.2 OGP in the Symmetric Binary Perceptron

In this section, we establish landscape results, dubbed as *ensemble $m-$OGP*, concerning the overlap structures of $m-$tuples $\left(\sigma^{(i)} : 1 \le i \le m\right)$, $\sigma^{(i)} \in \mathcal{B}_n$, that satisfy "box constraints" with respect to potentially correlated instances of Gaussian disorder.

3.2.1 Technical Preliminaries

We next formalize the notion of correlated instances through an appropriate interpolation scheme.

Definition 3.2.1. *Fix a $\kappa > 0$, and recall*

$$\alpha_c(\kappa) = -\frac{1}{\log_2 \mathbb{P}(|\mathcal{N}(0,1)| \le \kappa)}.$$

Let $0 < \alpha < \alpha_c(\kappa)$, $m \in \mathbb{N}$, $0 < \eta < \beta < 1$, and $\mathcal{I} \subset [0, \pi/2]$. Set $M = \lfloor n\alpha \rfloor$ and

suppose that $\mathcal{M}_i \in \mathbb{R}^{M\times n}$, $0 \le i \le m$, *is a sequence of i.i.d. random matrices, each having i.i.d.* $\mathcal{N}(0,1)$ *coordinates. Denote by* $\mathcal{S}_\kappa(\beta,\eta,m,\alpha,\mathcal{I})$ *the set of all* $m-$*tuples* $\big(\sigma^{(i)} : 1 \le i \le m\big)$, $\sigma^{(i)} \in \mathcal{B}_n$, *satisfying the following conditions.*

(a) **(Pairwise Overlap Condition)** *For any* $1 \le i < j \le m$,

$$\beta - \eta \le \mathcal{O}\sigma^{(i)}\sigma^{(j)} \le \beta,$$

where $\mathcal{O}\sigma^{(i)}\sigma^{(j)} \triangleq n^{-1}\left\langle \sigma^{(i)}, \sigma^{(j)} \right\rangle$.

(b) **(Rectangular Constraints)** *There exists* $\tau_i \in \mathcal{I}$, $1 \le i \le m$, *such that*

$$\left\|\mathcal{M}_i(\tau_i)\sigma^{(i)}\right\|_\infty \le \kappa\sqrt{n}, \quad 1 \le i \le m$$

where

$$\mathcal{M}_i(\tau_i) = \cos(\tau_i)\mathcal{M}_0 + \sin(\tau_i)\mathcal{M}_i \in \mathbb{R}^{M\times n}, \quad 1 \le i \le m. \tag{3.5}$$

The interpretations of the parameters appearing in Definition 3.2.1 are as follows. The parameter m is the size of the tuples we inspect; κ is the constraint threshold; and α is the constraint density. That is, we consider $M = \lfloor n\alpha \rfloor$ random constraints. Parameters β and η control the (forbidden) region of pairwise overlaps. Finally, the index set, $\mathcal{I}$, is used for generating correlated instances of random constraints via interpolation $\mathcal{M}_i(\tau_i)$ in (3.5), $\tau_i \in \mathcal{I}$. This is necessary to study the ensemble OGP, see below.

As a concrete example to Definition 3.2.1, consider the toy setting $m = 2$ and $\mathcal{I} = \{0\}$. In this case, $S_\kappa\big(\beta,\eta,2,\alpha,\{0\}\big)$ is simply the set of all pairs $(\sigma_1,\sigma_2) \in \mathcal{B}_n \times \mathcal{B}_n$ such that (a) $\beta - \eta \le n^{-1}\langle \sigma_1,\sigma_2\rangle \le \beta$ and (b) $\left\|\mathcal{M}\sigma_i\right\|_\infty \le \kappa\sqrt{n}$ for $i = 1,2$; where $\mathcal{M} \in \mathbb{R}^{\lfloor \alpha n\rfloor \times n}$ is a random matrix with i.i.d. standard normal entries.

3.2.2 Landscape Results: High κ Regime

Our first focus is on the regime where κ is *large*. While we set $\kappa = 1$ (thus $\alpha_c(\kappa)$ is approximately 1.8159) for simplicity; our results extend easily to any fixed $\kappa > 0$. In this case, we also drop the subscript κ appearing in Definition 3.2.1, and simply use the notation $\mathcal{S}(\beta,\eta,m,\alpha,\mathcal{I})$ to denote $\mathcal{S}_1(\beta,\eta,m,\alpha,\mathcal{I})$.

Our first result establishes $2-$OGP above $\alpha \ge 1.71$.

Theorem 3.2.2. *Let* $1.71 \le \alpha \le \alpha_c(1) \approx 1.8159$. *Then, there exists* $0 < \eta_2^* < \beta_2^* < 1$ *and a constant* $c^* > 0$ *such that the following holds. Fix any* $\mathcal{I} \subset [0,\pi/2]$ *with* $|\mathcal{I}| \le \exp_2\big(c^* n\big)$. *Then,*

$$\mathbb{P}\Big[\mathcal{S}\big(\beta_2^*,\eta_2^*,2,\alpha,\mathcal{I}\big) \ne \varnothing\Big] \le \exp_2\big(-\Theta(n)\big).$$

By considering the overlap structure of triples, one can further reduce the threshold (on α) to approximately 1.667 above which the overlap gap property takes place.

Theorem 3.2.3. *Let* $1.667 \leq \alpha \leq \alpha_c(1) \approx 1.8159$. *Then, there exists* $0 < \eta_3^* < \beta_3^* < 1$ *and a constant* $c^* > 0$ *such that the following holds. Fix any* $\mathcal{I} \subset [0, \pi/2]$ *with* $|\mathcal{I}| \leq \exp_2\big(c^* n\big)$. *Then,*

$$\mathbb{P}\Big[\mathcal{S}\big(\beta_3^*, \eta_3^*, 3, \alpha, \mathcal{I}\big) \neq \varnothing\Big] \leq \exp_2\big(-\Theta(n)\big).$$

The proof of Theorem 3.2.3 is provided in Section 3.6.2. The proof of Theorem 3.2.2 is quite similar to that of Theorem 3.2.3 (and in fact much simpler in terms of technical details); and is omitted.

Theorem 3.2.3 implies that 3−OGP (with appropriate parameters) takes place for $\alpha \geq 1.667$, which is indeed strictly smaller than the corresponding threshold of $\alpha \geq 1.71$ for 2−OGP established in Theorem 3.2.2. An inspection of the proof reveals that our choice of η^* satisfies $\eta^* \ll \beta^*$. That is, the structure that Theorem 3.2.3 rules out corresponds essentially to (nearly) equilateral triangles in Hamming space.

Theorem 3.2.3 is established using the *first-moment method*. More specifically, we let a certain random variable count the number of such triples. We then leverage Lemma 3.6.1 to ensure that the exponent of the first moment of that random variable is negative under appropriate choices of parameters. That is, the expectation is exponentially small (in n). Markov's inequality then yields Theorem 3.2.3. At a technical level, this amounts, in particular, to (a) counting the number of nearly equilateral triangles in the Hamming space; and (b) applying a Gaussian comparison inequality by Sidák [256] (reproduced herein as Theorem 3.6.5 for completeness). It is worth noting though that unlike [126], our counting bound is exact (up to lower-order terms). This appears necessary. Indeed, it appears not possible to improve upon Theorem 3.2.2 if one considers instead the relaxation to the "star-shaped" forbidden structures (where the overlap constraint is relaxed to $\mathcal{O}\sigma^{(1)}\sigma^{(j)} \in [\beta - \eta, \beta]$, $j \geq 2$) as in the counting step of [126, Theorems 2.3 and 2.6]. The aforementioned counting term appears more involved for $m \geq 4$.

As we noted earlier in the introduction, we do not pursue the m−OGP improvement for $m \geq 4$ in the high κ regime. This is due to the following reason: the first moment method employed for establishing m−OGP actually fails as m gets larger. That is, one can in fact (a) lower bound the first moment of the number N of m−tuples corresponding to the forbidden structure that m−OGP deals with, and (b) show that for m large, the value of α above which $\mathbb{E}[N]$ is $o(1)$ is actually strictly larger than 1.71. This, of course, is only a failure of the first moment method, and does not necessarily imply that the m−OGP itself yields a worse threshold. In fact, given the previously mentioned prior work employing m−OGP as well as the fact that m−OGP deals with a more nested structure, it indeed makes sense that m−OGP (for $m \geq 4$) should hold for a much broader range of α.[6] For this reason, it is plausible to conjecture that considering m−OGP beyond $m \in \{2, 3\}$ lowers the threshold on α. We leave the formal verification of this for future investigation.

Before we close this section, we remark that Baldassi, Della Vecchia, Lucibello,

[6]Here, it is worth noting that such a strict monotonicity in m has also been conjectured by Ben Arous and Jagannath in the context of spherical spin glass models [22].

and Zecchina established in [29] similar OGP results for the high κ case. To that end, fix any $x \in [0,1]$ and $K > 0$. Using a first moment argument, they show the existence of a critical threshold $\alpha_{\rm UB}^{(m)}(x,K)$ such that the following holds: fix any $\alpha > \alpha_{\rm UB}^{(m)}(x,K)$; then w.h.p. there exists no $m-$tuple $\sigma_i \in S_\alpha(K)$ with fixed pairwise Hamming distances of $\lfloor nx \rfloor$. Namely, their results correspond to the case $\eta_2^* = \eta_3^* = 0$. Furthermore, their results are rigorous for $m \in \{2,3,4\}$. However, they also suffer from technical difficulties similar to ours arising from the combinatorial terms for $m > 4$. For this reason, they resort to non-rigorous calculations and a replica symmetric ansatz to study $m-$tuples beyond $m = 4$.

3.2.3 Landscape Results: The Regime $\kappa \to 0$.

We now turn to our results in the regime $\kappa \to 0$. Observe that for any fixed $\kappa > 0$, the volume of the "rectangular box" $[-\kappa,\kappa]^m$ (which eventually controls the probabilistic term) appearing in Definition 3.2.1 is $(2\kappa)^m$. When $\kappa \to 0$, this term actually shrinks further by increasing m. Thus, one can hope to pursue the $m-$OGP improvement. This is the subject of the present subsection. Our main result to that end is as follows.

Theorem 3.2.4. *Let*

$$\alpha_{\rm OGP}(\kappa) \triangleq 10\kappa^2 \log \frac{1}{\kappa}. \tag{3.6}$$

Then, for every sufficiently small $\kappa > 0$ and $\alpha \geq \alpha_{\rm OGP}(\kappa)$, there exist $0 < \eta < \beta < 1$, $c > 0$, and an $m \in \mathbb{N}$ such that the following holds. Fix any $\mathcal{I} \subset [0,\pi/2]$ with $|\mathcal{I}| \leq \exp_2\big(cn\big)$. Then,

$$\mathbb{P}\Big[\mathcal{S}_\kappa\big(\beta,\eta,m,\alpha,\mathcal{I}\big) \neq \varnothing\Big] \leq \exp_2\big(-\Theta(n)\big).$$

The proof of Theorem 3.2.4 is in Section 3.6.3.

Recall from our earlier discussion (also see Section 3.3.3 and Corollary 3.3.6 therein) that the algorithm by Bansal and Spencer [36] works for $\alpha = O(\kappa^2)$. On the other hand, no (efficient) algorithm is known for $\alpha \geq C\kappa^2$, where $C > 0$ is a large absolute constant. Namely, the current known algorithmic threshold for the symmetric binary perceptron model is $\Theta(\kappa^2)$. In light of these facts, Theorem 3.2.4 shows that the OGP threshold $\alpha_{\rm OGP}(\kappa)$ is *nearly matching*: the onset of OGP coincides up to polylogarithmic (in κ) factors with the threshold (on α) above which no polynomial-time algorithms are known to work. The choice of the constant 10 appearing in (3.6) is for convenience and can potentially be improved.

We now comment on the extra $\log_2 \frac{1}{\kappa}$ factor appearing in (3.6). As we detail in Section 3.4, the exponent of the first moment of the cardinality term, $\big|S_\kappa\big(\beta,\eta,m,\alpha,\mathcal{I}\big)\big|$, appears to be strictly positive (for every β,η,m) if $\alpha = O\big(\kappa^2 \log_2(1/\kappa)\big)$. That is, Theorem 3.2.4 is in a sense the best possible using our techniques. However, it is plausible that by considering a more delicate forbidden structure (akin to the ones studied in [285, 63, 172]), one may in fact be able to remove this logarithmic factor. This suggests two conjectures: (a) in the regime $\kappa \to 0$, the algorithm by Bansal and

Spencer [36] is best possible (up to constant factors); and that (b) the OGP marks the onset of algorithmic hardness.

3.3 Algorithmic Barriers for the Perceptron Model

3.3.1 $m-$Overlap Gap Property Implies Failure of Stable Algorithms

We commence this section by recalling our setup. We fix a $\kappa > 0$, and an $\alpha < \alpha_c(\kappa)$ so that w.h.p. as $n \to \infty$, there exists a $\sigma \in S_\alpha(\kappa)$, where $S_\alpha(\kappa)$ is the (random) set introduced in (3.3). Having ensured that $S_\alpha(\kappa)$ is (w.h.p.) non-empty; our focus in this section is the problem of finding such a σ by using *stable algorithms*, formalized below.

Algorithmic Setting. We interpret an algorithm $\mathcal{A}$ as a mapping from $\mathbb{R}^{M\times n}$ to $\mathcal{B}_n$. We allow $\mathcal{A}$ to be potentially randomized: we assume there exists an underlying probability space $(\Omega, \mathbb{P}_\omega)$ such that $\mathcal{A} : \mathbb{R}^{M\times n} \times \Omega \to \mathcal{B}_n$. That is, for any $\omega \in \Omega$ and disorder matrix $\mathcal{M} \in \mathbb{R}^{M\times n}$; $\mathcal{A}(\cdot, \omega)$ returns a $\sigma_{\rm ALG} \triangleq \mathcal{A}(\mathcal{M}, \omega) \in \mathcal{B}_n$; and we want $\sigma_{\rm ALG}$ to satisfy $\|\mathcal{M}\sigma_{\rm ALG}\|_\infty \le \kappa\sqrt{n}$.

We now formalize the class of stable algorithms that we investigate in the present chapter.

Definition 3.3.1. *Fix a $\kappa > 0$, an $\alpha < \alpha_c(\kappa)$; and set $M = \lfloor n\alpha \rfloor$. An algorithm $\mathcal{A} : \mathbb{R}^{M\times n} \times \Omega \to \mathcal{B}_n$ is called $(\rho, p_f, p_{\rm st}, f, L)-$stable for the* `SBP` *model, if it satisfies the following for all sufficiently large n.*

- **(Success)** *Let $\mathcal{M} \in \mathbb{R}^{M\times n}$ be a random matrix with i.i.d $\mathcal{N}(0,1)$ coordinates. Then,*
$$\mathbb{P}_{(\mathcal{M},\omega)}\Big[\big\|\mathcal{M}\mathcal{A}(\mathcal{M},\omega)\big\|_\infty \le \kappa\sqrt{n}\Big] \ge 1 - p_f.$$
- **(Stability)** *Let $\mathcal{M}, \overline{\mathcal{M}} \in \mathbb{R}^{M\times n}$ be random matrices, each with i.i.d. $\mathcal{N}(0,1)$ coordinates such that $\mathbb{E}\big[\mathcal{M}_{ij}\overline{\mathcal{M}}_{ij}\big] = \rho$ for $1 \le i \le M$ and $1 \le j \le n$. Then,*
$$\mathbb{P}_{(\mathcal{M},\overline{\mathcal{M}},\omega)}\Big[d_H\big(\mathcal{A}(\mathcal{M},\omega), \mathcal{A}(\overline{\mathcal{M}},\omega)\big) \le f + L\|\mathcal{M} - \overline{\mathcal{M}}\|_F\Big] \ge 1 - p_{\rm st}.$$

Definition 3.3.1 is similar to the notion of stability considered in [126, Definition 3.1]. It is also worth noting that Definition 3.3.1 applies also to deterministic algorithms $\mathcal{A}$. In this case, we simply modify the probability statements to reflect the fact that the only source of randomness is the input $\mathcal{M}$ (and $\overline{\mathcal{M}}$) to the algorithm. In the remainder of the chapter, we often abuse the notation by dropping ω and simply referring to $\mathcal{A} : \mathbb{R}^{M\times n} \to \mathcal{B}_n$ as a randomized algorithm.

We next highlight the operational parameters appearing in Definition 3.3.1. κ is the "width" of the "rectangles" defined by the constraints. α is the constraint density (also known as the aspect ratio). That is, $M = \lfloor n\alpha \rfloor$ is the number of constraints.

The parameter p_f controls the success guarantee. The parameters ρ, p_{st}, f and L collectively control the stability guarantee. The parameter ρ essentially controls the amount of correlation. Stability parameters p_{st}, f and L describe the amount of sensitivity of the algorithm's output to the correlation values. Our stability guarantee is probabilistic, where the probability is taken with respect to the joint randomness in $\mathcal{M}, \overline{\mathcal{M}}$ as well as to the coin flips ω of $\mathcal{A}$. The "extra room" of f bits makes our negative result only stronger: even when $\mathcal{M}$ and $\overline{\mathcal{M}}$ are very close, the algorithm is still allowed to make roughly f flips.

We now state our next main result.

Theorem 3.3.2. *Fix any sufficiently small $\kappa > 0$, $\alpha \geq \alpha_{\text{OGP}}(\kappa) = 10\kappa^2 \log_2 \frac{1}{\kappa}$, and $L > 0$. Let $m \in \mathbb{N}$ and $0 < \eta < \beta < 1$ be the $m-$OGP parameters prescribed by Theorem 3.2.4. Set*

$$C = \frac{\eta^2}{1600}, \qquad Q \triangleq \frac{4800 L \pi}{\eta^2}\sqrt{\alpha}, \qquad \text{and} \qquad T = \exp_2\Big(2^{4mQ \log_2 Q}\Big). \tag{3.7}$$

Then, there exists an $n_0 \in \mathbb{N}$ such that the following holds. For every $n \geq n_0$, there exists no randomized algorithm $\mathcal{A} : \mathbb{R}^{M \times n} \to \mathcal{B}_n$ that is

$$\left(\cos\left(\frac{\pi}{2Q}\right), \frac{1}{9(Q+1)T}, \frac{1}{9Q(T+1)}, Cn, L\right) - \textit{stable}$$

for the `SBP`*, in the sense of Definition 3.3.1.*

The proof of Theorem 3.3.2 is provided in Section 3.6.4. Several remarks are now in order. First, observe that there is no restriction on the running time of $\mathcal{A}$: as long as it is stable in the sense of Definition 3.3.1 with appropriate parameters, Theorem 3.3.2 applies.

Our second remark pertains to the scaling of parameters in the regime $n \to \infty$. Observe that the parameters α, L, m and η are all $O(1)$ (in n) as $n \to \infty$; hence the parameters C, Q, and T appearing in (3.7) are all constants. In particular, p_f and p_{st} are of constant order. This is an important feature of our result: the algorithms that we rule out have a *constant probability* of success/stability. Namely, $\mathcal{A}$ need not have a high-probability guarantee. This is a notable departure from the main hardness result in [126, Theorem 3.2], as well as from those appeared in prior works: unlike our case, the algorithms ruled out via OGP in those papers are required to succeed with high probability.

Our next remark pertains to the stability guarantee. Note that the algorithms that we rule out satisfy

$$d_H\Big(\mathcal{A}(\mathcal{M}), \mathcal{A}(\overline{\mathcal{M}})\Big) \leq Cn + L\big\|\mathcal{M} - \overline{\mathcal{M}}\big\|_F.$$

Namely, under our notation of stability the algorithm is still allowed to make $\Theta(n)$ flips when $\mathcal{M}$ and $\overline{\mathcal{M}}$ are "nearly identical".

Our final remark pertains to the parameter L. We establish Theorem 3.3.2 for the case when L is constant in order to keep our exposition clean. However, an inspection

of our argument reveals L can be pushed to $O\left(\frac{\log n}{\log\log n}\right)$.

3.3.2 Failure of Online Algorithms for SBP

Our next focus is on the class of *online algorithms*, formalized below.

Definition 3.3.3. *Fix a $\kappa > 0$, an $\alpha < \alpha_c(\kappa)$; and set $M = \lfloor n\alpha \rfloor \in \mathbb{N}$. Let $\mathcal{M} \in \mathbb{R}^{M\times n}$ be a disorder matrix with columns $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n \in \mathbb{R}^M$, and $\mathcal{A} : \mathbb{R}^{M\times n} \to \mathcal{B}_n$ be an algorithm where*

$$\mathcal{A}\big(\mathcal{M}\big) = \sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n) \in \mathcal{B}_n.$$

We call $\mathcal{A}$ p_f-online if the following hold.

- **(Success)** *For $\mathcal{M}$ consisting of i.i.d. $\mathcal{N}(0,1)$ entries,*

$$\mathbb{P}\Big[\big\|\mathcal{M}\mathcal{A}(\mathcal{M})\big\|_\infty \le \kappa\sqrt{n}\Big] \ge 1 - p_f.$$

- **(Online)** *There exists deterministic functions f_t, $1 \le t \le n$ such that*

$$\sigma_t = f_t\big(\mathcal{C}_i : 1 \le i \le t\big) \in \{-1, 1\} \qquad \text{for} \qquad 1 \le t \le n.$$

Several remarks are now in order. The parameter p_f is the failure probability of $\mathcal{A}$: $\mathcal{A}(\mathcal{M}) \in S_\alpha(\kappa)$ w.p. at least $1 - p_f$. The second condition states that for all $1 \le t \le n$, σ_t is a function of $\mathcal{C}_1, \ldots, \mathcal{C}_t$ only. More precisely, the signs $\sigma_i \in \{-1, 1\}$, $1 \le i \le t-1$, have been assigned at the end of round $t-1$. A new column $\mathcal{C}_t \in \mathbb{R}^M$ arrives in the beginning of round t, and $\mathcal{A}$ assigns a $\sigma_t \in \{-1, 1\}$ depending only on the previous decisions. This highlights the online nature of $\mathcal{A}$.

Definition 3.3.3 is an abstraction that captures, in particular, the algorithm by Bansal and Spencer [36]. Our next result establishes that online algorithms fail to return a $\sigma \in S_\alpha(\kappa)$ for densities α close to the critical threshold $\alpha_c(\kappa)$. Similar to our treatment in Section 3.2.2, we stick to the case $\kappa = 1$ for simplicity, even though our argument easily extends to arbitrary $\kappa > 0$.

Theorem 3.3.4. *Let $1.77 \le \alpha \le \alpha_c(1) \approx 1.8159$. Then, there exists a constant $c_f > 0$ such that the following holds. For any $p_f < \frac{1}{2} - \exp\big(-c_f n\big)$, there exists no $\mathcal{A}$ for SBP which is p_f-online in the sense of Definition 3.3.3.*

The proof of Theorem 3.3.4 is provided in Section 3.6.5. The proof is based on a contradiction argument, which we informally describe. Given $\Delta \in (0,1)$, $\mathcal{M} \in \mathbb{R}^{M\times n}$, let $\mathcal{M}_\Delta \in \mathbb{R}^{M\times n}$ be obtained from $\mathcal{M}$ by independently resampling the last $\Delta \cdot n$ columns on $\mathcal{M}$. Fix an online algorithm $\mathcal{A}$, and let $\sigma \triangleq \mathcal{A}\big(\mathcal{M}\big)$, $\sigma_\Delta \triangleq \big(\mathcal{M}_\Delta\big)$. Then w.p. at least $1 - 2p_f$, $\|\mathcal{M}\sigma\|_\infty \le \sqrt{n}$ and $\|\mathcal{M}_\Delta\sigma_\Delta\|_\infty \le \sqrt{n}$. Furthermore, σ and σ_Δ agree on first $n - \Delta n$ coordinates due to the online nature of $\mathcal{A}$. Namely, assuming such an $\mathcal{A}$ exists, we have $\mathbb{P}\big[\Xi(\Delta) \ne \varnothing\big] \ge 1 - 2p_f$, where $\Xi(\Delta)$ is the set of all pairs $(\sigma, \sigma_\Delta) \in \mathcal{B}_n \times \mathcal{B}_n$ such that $\|\mathcal{M}\sigma\|_\infty \le \sqrt{n}$, $\|\mathcal{M}_\Delta\sigma_\Delta\|_\infty \le \sqrt{n}$

and $n^{-1}\langle \sigma, \sigma_\Delta \rangle \geq 1 - 2\Delta$. On the other hand, a first moment argument (see in particular Proposition 3.6.17) reveals that for the same choice of Δ, we actually have $\mathbb{P}\big[\Xi(\Delta) \neq \varnothing\big] \leq \exp(-\Theta(n))$. This yields a contradiction and proves Theorem 3.3.4.

The contradiction argument described above is slightly different than 2−OGP, yielding a lower bound $\alpha \geq 1.77$. Notice that this is strictly larger than the corresponding 2−OGP threshold, i.e. $\alpha \geq 1.71$, for the same setting ($\kappa = 1$) per Theorem 3.2.2. Lastly, the online algorithms that we rule out need not have a high probability guarantee: a success probability slightly above $\frac{1}{2}$ suffices.

3.3.3 Algorithmic Threshold in SBP: A Lower Bound and a Conjecture

Algorithmic Lower Bound in SBP. Heretofore, we used $\Theta\big(\kappa^2\big)$ as our baseline for the current computational threshold for the SBP. Namely, against this threshold; we (a) formulated the aforementioned *statistical-to-computational gap* and (b) compared our hardness result, Theorem 3.3.2, for the stable algorithms established via the m−OGP approach. In this section, we justify this choice for the algorithmic threshold, from the lower bound perspective.

As we mentioned in the introduction, the SBP is closely related to the well-known problem of minimizing the discrepancy of a matrix (or set system). The discrepancy minimization problem received much attention in the field of combinatorics and theoretical computer science; several efficient algorithms have been devised for it, see e.g. [246, 201, 106, 36]. In what follows, we use the recent work by Bansal and Spencer [36] as our baseline for postulating a computational threshold on α as one varies κ; though several of the algorithms cited above essentially yield the same $\Theta\big(\kappa^2\big)$ guarantee modulo different absolute constants. Before we proceed with the result of Bansal and Spencer [36]; it is worth noting that there is yet another complementary line of research focusing on the so-called online guarantees, see e.g. [34, 33, 15, 209]. However, all of these algorithms suffer from extra polylogarithmic factors; and therefore their implied guarantees on α are poorer. That is they provably work only for α asymptotically much smaller than κ^2.

The work by Bansal and Spencer (see in particular [36, Section 3.3]) establishes the following.

Theorem 3.3.5. *[36, Theorem 3.4] Let $T \in \mathbb{N}$ be an arbitrary time horizon, and $v_i \sim \mathrm{Unif}(\mathcal{B}_M)$, $1 \leq i \leq T$, be i.i.d. random vectors. Then there exists a value $K > 0$ and an algorithm that returns signs $s_1, \ldots, s_T \in \{-1, 1\}$ in $\mathrm{Poly}(M, T)$ time such that*

$$\mathbb{P}\left[\left\|\sum_{i \leq T} s_i v_i\right\|_\infty \leq K\sqrt{M}\right] \geq 1 - \exp\big(-cM\big).$$

Here, $c, K > 0$ are absolute constants independent of M and T.

Corollary 3.3.6. *There exists an absolute constant $K > 0$ such that the following holds. Fix any $\kappa > 0$, $\alpha < (\kappa/K)^2$; and consider the matrix $\mathcal{M} \in \mathbb{R}^{\alpha n \times n}$ with*

i.i.d. entries subject to the condition

$$\mathbb{P}\big[\mathcal{M}_{ij} = +1\big] = \frac{1}{2} = \mathbb{P}\big[\mathcal{M}_{ij} = -1\big], \qquad \textit{for all} \qquad 1 \le i \le \alpha n, 1 \le j \le n.$$

Then, there exists an algorithm $\mathcal{A}$, running in $\mathrm{poly}(n)$ *time, such that w.h.p.*

$$\Big\|\mathcal{M} \cdot \mathcal{A}\big(\mathcal{M}\big)\Big\|_\infty \le \kappa\sqrt{n}.$$

Corollary 3.3.6 is a direct consequence of Theorem 3.3.5. Indeed, consider $\mathcal{M} \in \{\pm 1\}^{\alpha T \times T}$ with $\alpha = n/T$, whose columns are v_i, $1 \le i \le T$. Then one can find, in polynomial (in n, T) time, a $\sigma \in \mathcal{B}_T$ such that

$$\big\|\mathcal{M}\sigma\big\|_\infty \le K\sqrt{n} = K\sqrt{\alpha T}.$$

Since $\alpha < (\kappa/K)^2$, the claim follows.

Admittedly, their result is established for the case of i.i.d. Rademacher disorder. Nevertheless, due to the aforementioned universality guarantees encountered in perceptron-like models, it is expected that the exact same guarantee (perhaps with a modified constant K) remains true for the case of i.i.d. standard normal disorder.

A Conjecture on the Algorithmic Threshold. Recall from our prior discussion that for many random computational problems, the $m-$OGP threshold coincides (or nearly coincides) with conjectured algorithmic threshold. Examples include the problem of finding the largest independent set in random sparse graphs [245, 285], NAE-k-SAT [137], random $k-$SAT [63], mixed even $p-$spin model [172], and so on. In light of the preceding discussion, this is also the case for the `SBP` model: the limit of known algorithms is at $\Theta(\kappa^2)$, whereas, as we establish in Theorem 3.2.4, the ensemble $m-$OGP holds for densities $\Omega\big(\kappa^2 \log_2 \frac{1}{\kappa}\big)$ in the regime $\kappa \to 0$.

On the other hand, unlike models such as the independent set problem, $k-$SAT, or the planted clique; prior to this work no conjectures were proposed regarding the threshold for algorithmic hardness in `SBP` model in the $\kappa \to 0$ regime. Here, we do put forward such a conjecture. To that end, let

$$\alpha^*_m(\kappa) \triangleq \inf\Big\{\alpha \in \big[0, \alpha_c(\kappa)\big] : \exists 1 > \beta > \eta > 0, \liminf_{n\to\infty} \mathbb{P}\Big[\mathcal{S}_\kappa\big(\beta, \eta, m, \alpha, \{0\}\big) = \varnothing\Big] = 1\Big\}. \tag{3.8}$$

That is, $\alpha^*_m(\kappa)$ is the threshold for the $m-$OGP (with appropriate β, η). Let

$$\alpha^*_\infty(\kappa) \triangleq \lim_{m\to\infty} \alpha^*_m(\kappa), \tag{3.9}$$

where the limit is well-defined since $\big(\alpha^*_m\big)_{m\ge 1}$ is a non-increasing sequence of non-negative real numbers. Then we conjecture $\alpha^*_\infty(\kappa)$ marks the true algorithmic threshold for this problem.

Conjecture 3.3.7. *For any $\epsilon > 0$, there exists a $\kappa^*(\epsilon) > 0$ such that the following hold for every $\kappa \le \kappa^*(\epsilon)$:*

- *There exists no polynomial-time search algorithms for the* SBP *if* $\alpha > (1+\epsilon)\alpha^*_\infty(\kappa)$.

- *There exists a polynomial-time search algorithm for the* SBP *if* $\alpha < (1-\epsilon)\alpha^*_\infty(\kappa)$.

Recall that per Theorem 3.2.4, $\alpha^*_\infty(\kappa) = O\big(\kappa^2 \log_2 \frac{1}{\kappa}\big)$. Notice that the $\alpha^*_m(\kappa)$ (hence the $\alpha^*_\infty(\kappa)$) are defined for the non-ensemble variant of $m-$OGP, $\mathcal{I} = \{0\}$. That is, $\sigma^{(i)}$, $1 \le i \le m$, satisfy constraints dictated by the rows of the **same** disorder matrix $\mathcal{M} \in \mathbb{R}^{M\times n}$ with i.i.d. $\mathcal{N}(0,1)$ (or Rademacher) entries, where $M = \lfloor \alpha n \rfloor$. This is merely for simplicity: the ensemble $m-$OGP and the non-ensemble $m-$OGP often take place at the **exact same** threshold. The former, on the other hand, is just technically more involved; and is necessary to rule out certain classes of algorithms via an interpolation/contradiction argument as we do in this work. The structural property implied by the non-ensemble OGP already suffices to predict the desired algorithmic threshold.

3.3.4 Stability of the Kim-Roche Algorithm

Having established that the $m-$OGP is a provable barrier for the class of stable algorithms, it is then natural to inquire whether the class of stable algorithms captures the implementations of known algorithms for perceptron models. In this section, we investigate this question for a certain algorithm devised for the *asymmetric* model, which we recall from (3.1).

Kim and Roche devised in [188] an algorithm which admits, as its input, a disorder matrix $\mathcal{M} \in \mathbb{R}^{k\times n}$ with i.i.d. entries; and returns a $\sigma \in \mathcal{B}_n$ such that $\mathcal{M}\sigma \in \mathbb{R}^k$ is entry-wise non-negative as long as $k < 0.005n$. That is, their algorithm provably returns a $\sigma \in S^A_\alpha(0)$ as long as $\alpha < 0.005$. (We use k in place of M for the number of constraints so as to be consistent with their notation.) We denote their algorithm by $\mathcal{A}_{\text{KR}} : \mathbb{R}^{k\times n} \to \mathcal{B}_n$ as a shorthand notation. It is worth noting that while their results are established for the case where $\mathcal{M}$ consists of i.i.d. Rademacher entries, they easily extend to the case of Gaussian $\mathcal{N}(0,1)$ entries, which will be our focus here. $\mathcal{A}_{\text{KR}}$ takes $O\big(\log_{10}\log n_{10}\big)$ steps, each requiring poly(n) time.[7] Namely, $\mathcal{A}_{\text{KR}}$ is an *efficient* algorithm that provably works in the so-called *linear regime*, $k = \Theta(n)$. Admittedly, $\mathcal{A}_{\text{KR}}$ is tailored for the asymmetric model. Nevertheless, there are only a few known algorithms with rigorous guarantees for perceptron models; thus it is indeed natural to explore the stability of $\mathcal{A}_{\text{KR}}$.

Operational Parameters. We next provide details of the Kim-Roche algorithm from [188]. Let

$$f_0 = 1, \quad f_1 = \frac{1}{200}, \qquad \text{and} \quad f_j = 10^{-2^j}, \qquad \text{for} \qquad 2 \le j \le N, \tag{3.10}$$

[7] Throughout this section, we consider all logarithms in base 10 in order to be consistent with the notation of [188].

as in [188, Equation 5.42], where $N = \lceil C \log_{10} \log_{10} n \rceil$ is the total number of rounds. Next, let

$$k_0 = k, \quad k_1 = 2\lfloor (1/2)(n/10^8) \rfloor + 1, \quad \text{and} \quad k_s = 2\lfloor (1/2)(n \cdot f_s^3) \rfloor + 1 \quad \text{for} \quad 2 \le s \le N \tag{3.11}$$

as in [188, Equation 5.46]. Set

$$A \triangleq \sum_{0 \le j \le N} f_j; \tag{3.12}$$

and let n_j be defined by

$$n_0 = \lfloor n/A \rfloor, \quad \text{and} \quad n_j = \left\lfloor \frac{n}{A} \sum_{0 \le i \le j} f_i \right\rfloor - \left\lfloor \frac{n}{A} \sum_{0 \le i \le j-1} f_i \right\rfloor, \quad 1 \le j \le N \tag{3.13}$$

per [188, Equation 5.44].

Informal Description of the Algorithm. We now describe the $\mathcal{A}_{\mathrm{KR}}$ algorithm. To that end, denote by $R_1, \ldots, R_k \in \mathbb{R}^n$ the rows of $\mathcal{M}$. Given a set P of rows and a set Q of columns, let $\mathcal{M}(P, Q)$ denote the $|P| \times |Q|$ submatrix M_{ij} obtained by retaining rows $i \in P$ and columns $j \in Q$.

- In the first round, $\mathcal{A}_{\mathrm{KR}}$ assigns n_0 coordinates of σ by taking the majority vote in submatrix $\mathcal{M}([1, k] : [1, n_0])$. That is,

$$\sigma_j = \operatorname{sgn}\left(\sum_{1 \le i \le k} \mathcal{M}_{ij} \right) \qquad \text{for} \qquad 1 \le j \le n_0.$$

- For each row R_i of $\mathcal{M}$, it then computes the partial inner products $\langle R_i, \sigma \rangle$ (restricted to $\mathbb{R}^{n_0}$), finds an index set $\mathcal{I}_1$ corresponding to k_1 smallest indices; and takes a majority vote in the submatrix $\mathcal{M}(\mathcal{I}_1 : [n_0 + 1, n_0 + n_1])$ and repeats this procedure.

- In particular, at the beginning of round $j \ge 1$, one has a vector $\sigma \in \{\pm 1\}^{\sum_{0 \le s \le j-1} n_s}$. One then computes the partial inner products $\langle R_i, \sigma \rangle$, $1 \le i \le k$, and computes an index set $\mathcal{I}_j$ with $|\mathcal{I}_j| = k_j$ such that $i \in \mathcal{I}_j$ iff $\langle R_i, \sigma \rangle$ is among the k_j smallest (partial) inner products. Taking then the majority vote in the submatrix

$$\mathcal{M}\left(\mathcal{I}_j : \left[1 + \sum_{0 \le s \le j-1} n_s, \sum_{0 \le s \le j} n_s \right] \right)$$

  settles next n_j entries of σ, that is the entries σ_j, $1 + \sum_{0 \le s \le j-1} n_s \le j \le \sum_{0 \le s \le j} n_s$. Namely, a $k_j \times n_j$ submatrix is used for determining next n_j components of σ and $\overline{\sigma}$.

Numerically, $n_0 \approx 0.995n$. Thus even at the beginning, $\mathcal{A}_{\mathrm{KR}}$ already settles *most* of the entries of $\sigma \in \mathcal{B}_n$.

Having described the $\mathcal{A}_{\mathrm{KR}}$ informally, we are now in a position to state our main result. We show that $\mathcal{A}_{\mathrm{KR}}$ is stable in the sense of Definition 3.3.1.

Theorem 3.3.8. *Let $\mathcal{M} \in \mathbb{R}^{k\times n}$ and $\mathcal{M}' \in \mathbb{R}^{k\times n}$ be two i.i.d. random matrices each with i.i.d.$\mathcal{N}(0,1)$ entries; and let*

$$\overline{\mathcal{M}}(\tau) \triangleq \cos(\tau)\mathcal{M} + \sin(\tau)\mathcal{M}' \in \mathbb{R}^{k\times n}, \qquad \tau \in \left[0, \frac{\pi}{2}\right]. \tag{3.14}$$

Set $\tau = n^{-0.02}$. Then,

$$\mathbb{P}\Big[d_H\Big(\mathcal{A}_{\mathrm{KR}}(\mathcal{M}), \mathcal{A}_{\mathrm{KR}}\big(\overline{\mathcal{M}}(\tau)\big)\Big) = o(n)\Big] \geq 1 - O\left(n^{-\frac{1}{41}}\right).$$

As a result, the Kim-Roche algorithm is

$$\Big(\cos\big(n^{-0.02}\big), o\big(1/n\big), O\left(n^{-1/41}\right), Cn, L\Big) - \text{stable}$$

in the sense of Definition 3.3.1 for any $C > 0$ and $L > 0$ (see below for further details).

In order to establish Theorem 3.3.8, we first establish in Section 3.6.6 an auxiliary result, Proposition 3.6.18, which pertains to the partial implementation of $\mathcal{A}_{\mathrm{KR}}$. That is, we analyze $\mathcal{A}_{\mathrm{KR}}$ run for $c\log_{10}\log_{10} n$ rounds (where $c > 0$ is a small enough constant) as opposed to its full $N = \lceil C\log_{10}\log_{10} n\rceil$ round implementation; and show that it is stable. We then show in Section 3.6.6 that the number of unassigned coordinates, $\sum_{c\log_{10}\log_{10} n+1\leq j\leq N} n_j$, is $o(n)$. This, together with Proposition 3.6.18, establishes Theorem 3.3.8.

Several pertinent remarks are now in order. We first highlight that $\mathcal{A}_{\mathrm{KR}}$ is indeed stable in the sense of Definition 3.3.1 with the parameters noted above. (Here, we suppress the randomness, and the success guarantee is now for the event $\big\{\mathcal{M}\mathcal{A}(\mathcal{M}) \geq 0\big\}$ entry-wise.) To that end, an inspection of [188, Theorem 1.4] reveals that for $\mathcal{M} \in \mathbb{R}^{k\times n}$ with $k = \alpha n$ having i.i.d. $\mathcal{N}(0,1)$ entries,

$$\mathbb{P}_{\mathcal{M}}\Big[\mathcal{M}\mathcal{A}(\mathcal{M}) \geq 0\Big] \geq 1 - o\big(n^{-1}\big),$$

as long as $\alpha < 0.005$. With this, we obtain that $\mathcal{A}_{\mathrm{KR}}$ is

$$\Big(\cos\big(n^{-0.02}\big), o\big(1/n\big), O\left(n^{-1/41}\right), Cn, L\Big) - \text{stable}$$

for any $\alpha < 0.005$, $C > 0$ and $L > 0$ for the asymmetric binary perceptron in the sense of Definition 3.3.1. (In fact, one can take $L = 0$ as $C > 0$.) Note though that this parameter scaling is not comparable with Theorem 3.2.4 since Theorem 3.2.4 pertains to the symmetric model, see below for more details.

Recalling

$$\mathcal{O}\big(\sigma, \overline{\sigma}\big) = n^{-1}\left\langle \sigma, \overline{\sigma}\right\rangle = 1 - 2d_H\big(\sigma, \overline{\sigma}\big)/n,$$

it follows that in the setting of Theorem 3.3.8, $\mathcal{O}\big(\sigma, \overline{\sigma}\big) = 1 - o(1)$. That is, σ and $\overline{\sigma}$ agree on all but a vanishing fraction of coordinates. Informally, this suggests that $\mathcal{A}_{\mathrm{KR}}$

cannot overcome the overlap barrier of η appearing in Theorems 3.2.2, 3.2.3, and 3.2.4 as $\eta = O(1)$. However, we established the OGP results for the symmetric case as opposed to the asymmetric model for which $\mathcal{A}_{\text{KR}}$ is devised. Thus Theorem 3.3.8 is not exactly compatible with the hardness result, Theorem 3.3.2. A more compelling picture would be to show that the OGP takes place also for the asymmetric model, with an η that is of order $O(1)$; and then couple such result with Theorem 3.3.8. We leave this as a very interesting direction for future work.

Lastly, it would also be very interesting to prove that the algorithm by Abbe, Li and Sly [5] devised for the `SBP` is also stable in the relevant sense. An inspection of [5] reveals that several of the key steps are similar to [188], but there are a few differences which prevent immediate verification of stability. We now elaborate on this by highlighting fundamental differences between their algorithm and $\mathcal{A}_{\text{KR}}$. Inspecting [5, Page 7], it appears that a major change is the incorporation of extra sign parameters inside the summation: using the exact same notation as in [5], this is

$$\operatorname{sgn}\left(\sum_{r\in\mathcal{R}_i} -\operatorname{sgn}\big(S^{(r)}(0:i-1)\big)G_{r,j}\right).$$

This extra term is independent of the summands, and therefore, is benign. As a result, it appears that our Lemmas 3.6.22 and 3.6.23 apply almost verbatimly. Their algorithm has two additional steps, one in the beginning and one at the end; these steps appear to be stable, as well. A main technical challenge, however, is that their algorithm requires $O(\sqrt{\log n})$ rounds, as opposed to the Kim-Roche algorithm that requires only $O(\log\log n)$ rounds. The choice of $\log\log n$ is crucial for our argument, see in particular Proposition 3.6.18 below. It is not clear though if they need $O(\sqrt{\log n})$ steps to find a large cluster and whether one can achieve the much more modest goal of finding a solution σ in $O(\log\log n)$ rounds. We leave the formal investigation of this as a very interesting direction for future research.

3.4 Natural Limitations of Our Techniques

Recall from our earlier discussion that the algorithmic threshold for the SBP model appears to be $\Theta(\kappa^2)$, whereas we established $m-$OGP for densities above $\Omega\big(\kappa^2\log_2\frac{1}{\kappa}\big)$. That is, the OGP threshold is off by a polylogaritmic (in $1/\kappa$) factor.

In this section, we investigate whether one can shave off this extra $\log_2\frac{1}{\kappa}$ factor. In a nutshell, we provide an informal argument suggesting that to establish $m-$OGP for the structure that we consider, $\alpha = \Omega\big(\kappa^2\log_2\frac{1}{\kappa}\big)$ appears necessary.

To that end, fix first a $\kappa > 0$, where one should think of κ to be sufficiently small. An inspection of the proof of Theorem 3.2.4 reveals that the first moment is controlled by a certain $\Upsilon(\beta,\alpha)$ appearing in (3.55), which we repeat below for convenience:

$$\Upsilon(\beta,\alpha) = h\left(\frac{1-\beta}{2}\right) - \frac{\alpha}{2}\log_2(2\pi) + \alpha\log_2(2\kappa) - \frac{\alpha}{2}\log_2(1-\beta). \tag{3.15}$$

In particular, for the first moment argument to work, it should be the case that $\Upsilon(\beta, \alpha)$ is negative for an appropriate choice of parameters β, α.

Now set $\delta \triangleq \frac{1-\beta}{2}$. Yet another inspection of the proof of Theorem 3.2.4 shows that for the first moment method to be applicable, β should be close to one. For this reason, the regime of interest below is therefore $\delta \to 0$ and $\kappa \to 0$.

Step 1: $\delta > \kappa^2$ is necessary.

Note that as $\delta \to 0$,

$$h(\delta) = -\delta \log_2 \delta - (1-\delta)\log_2(1-\delta) = -\delta \log_2 \delta + \Theta_\delta\big(\delta\big),$$

using the Taylor expansion,

$$\log_2(1-\delta) = -\frac{1}{\ln 2}\delta + o(\delta).$$

With this, we manipulate (3.15) to arrive at

$$\Upsilon(\beta, \alpha) = \delta \log_2 \frac{1}{\delta} + \frac{\alpha}{2}\log_2 \frac{1}{\delta} - \alpha \log_2 \frac{1}{\kappa} + \Theta_\delta\big(\delta\big) + o_{\delta,\kappa}\big(\alpha \log_2 \kappa\big).$$

Now, for $\Upsilon(\beta, \alpha)$ to be negative, we must have $\alpha \log_2 \frac{1}{\kappa} > \frac{\alpha}{2}\log_2 \frac{1}{\delta}$. This immediately yields $\delta > \kappa^2$ to be a *necessary* condition.

Step 2: $\alpha = \Omega\big(\kappa^2 \log_2 \frac{1}{\kappa}\big)$ is necessary.

Set $\delta = C\kappa^2$ for $C \triangleq C(\kappa) > 1$. The expression for $\Upsilon(\beta, \alpha)$ then becomes

$$\begin{aligned}
\Upsilon(\beta, \alpha) &= \underbrace{h\left(\frac{1-\beta}{2}\right)}_{-\delta \log_2 \delta + \Theta_\delta(\delta)} - \frac{\alpha}{2}\log_2(2\pi) + \alpha \log_2(2\kappa) - \frac{\alpha}{2}\log_2(1-\beta) \\
&= -\delta \log_2 \delta + \Theta_\delta(\delta) + \alpha \log_2 \kappa - \frac{\alpha}{2}\log_2 \delta - \frac{\alpha}{2}\log_2 \pi \\
&= 2C\kappa^2 \log_2 \frac{1}{\kappa} + C\kappa^2 \log_2 \frac{1}{C} - \frac{\alpha}{2}\log_2 C - \frac{\alpha}{2}\log_2 \pi. \qquad (3.16)
\end{aligned}$$

Note that $\delta < 1$, and thus $C < \frac{1}{\kappa^2}$. Thus $\log_2 C < 2\log_2 \frac{1}{\kappa}$. Equipped with this observation, we investigate two separate cases for the growth of C.

Case 1: $\log_2 C = o_\kappa(\log_2 \frac{1}{\kappa})$**.** Then $C\kappa^2 \log_2 \frac{1}{\kappa}$ dominates the term, $C\kappa^2 \log_2 \frac{1}{C}$ apearing in (3.16). In this case for $\Upsilon(\beta, \alpha)$ to be negative, one must indeed ensure

$$\frac{\alpha}{2}\log_2 C > 2C\kappa^2 \log_2 \frac{1}{\kappa}.$$

Rearranging this, we find

$$\alpha > \frac{4C}{\log_2 C}\kappa^2 \log_2 \frac{1}{\kappa} \implies \alpha = \Omega\left(\kappa^2 \log_2 \frac{1}{\kappa}\right),$$

which is precisely our claim. In fact, the parameter β for which Theorem 3.2.4 is established is of form $\beta = 1 - \Theta(\kappa^2)$, see (3.56). Hence, one has $\delta = \Theta(\kappa^2)$ and $C = \Theta_\kappa(1)$, thus $\log_2 C$ is indeed $o_\kappa\big(\log_2 \frac{1}{\kappa}\big)$.

Case 2: $\log_2 C = \Theta_\kappa\big(\log_2 \frac{1}{\kappa}\big)$**.** In this case, we now show that the threshold on α is worse than the one appearing in the previous case.

To that end, set $C \sim \kappa^{-\gamma}$, $\gamma < 2$: that is, we assume

$$\lim_{\kappa \to 0} \frac{\log_2 C}{\log_2 \frac{1}{\kappa}} = \gamma.$$

We focus on certain terms appearing in (3.16). Note that,

$$C\kappa^2 \log_2 \frac{1}{\kappa} \sim 2\kappa^{2-\gamma} \log_2 \frac{1}{\kappa}, \quad C\kappa^2 \log_2 \frac{1}{C} = -\gamma\kappa^{2-\gamma} \log_2 \frac{1}{\kappa}, \quad \text{and} \quad \frac{\alpha}{2} \log_2 C \sim \frac{\alpha}{2}\gamma \log_2 \frac{1}{\kappa}.$$

Combining these findings, we immediately observe that for $\Upsilon(\beta, \alpha)$ to be negative, one must have $\alpha \sim \kappa^{2-\gamma}$, where any $\gamma < 2$ works. Notice that this threshold is strictly worse than $\kappa^2 \log_2 \frac{1}{\kappa}$.

Hence, $\alpha = \Omega\big(\kappa^2 \log_2 \frac{1}{\kappa}\big)$ is indeed necessary for $m-$OGP (for the configuration we consider with a sufficiently large $m \in \mathbb{N}$ and $0 < \eta < \beta < 1$) to take place. It is though conceivable that by establishing the OGP for a potentially more intricate structure, like the ones considered in [285, 63, 172], one may in fact reach all the way down to $\Theta(\kappa^2)$. We leave this extension as an interesting future research direction.

3.5 Universality in OGP: Beyond Gaussian Disorder

Our OGP results, Theorems 3.2.2, 3.2.3 and 3.2.4, are established for the case where the disorder matrix $\mathcal{M} \in \mathbb{R}^{M \times n}$ consists of i.i.d. $\mathcal{N}(0, 1)$ entries. However, much like many other properties regarding the perceptron model, the OGP also enjoys the *universality*. In other words, the exact details of the distribution (of disorder) are immaterial; and provided that certain (rather mild) conditions on the distribution are satisfied, the OGP results still remain valid.

We now (somewhat informally) elaborate on the mechanics of this extension. The main technical tool that we employ is the multi-dimensional version of the Berry-Esseen Theorem, which is reproduced herein for convenience.

Theorem 3.5.1. *Let $Y_1, Y_2, \ldots, Y_n \in \mathbb{R}^m$ be independent centered random vectors. Suppose $S = \sum_{1 \le i \le n} Y_i$ and $\Sigma = \mathrm{Cov}(S) \in \mathbb{R}^{m \times m}$ is invertible. Let $Z \sim \mathcal{N}(0, \Sigma)$ be an $m-$dimensional multivariate normal random vector, whose covariance is Σ. Then,*

there exists a universal constant C such that for all convex $U \subseteq \mathbb{R}^d$,

$$\Big|\mathbb{P}[S \in U] - \mathbb{P}[Z \in U]\Big| \leq C m^{\frac{1}{4}} \sum_{1 \leq j \leq n} \mathbb{E}\Big[\big\|\Sigma^{-\frac{1}{2}} Y_j\big\|_2^3\Big].$$

We will apply Theorem 3.5.1 for $U = [-\kappa, \kappa]^m$. While our results still transfer to the ensemble OGP, we restrict our attention to the non-ensemble variant for simplicity. That is, we focus on the case where the set $\mathcal{I}$ appearing in Definition 3.2.1 is $\{0\}$.

Theorem 3.5.2. *Let $\mathcal{D}$ be a distribution on $\mathbb{R}$ with the property that for $T \sim \mathcal{D}$,*

$$\mathbb{E}\big[T\big] = 0, \quad \mathbb{E}\big[T^2\big] = 1; \qquad \textit{and} \qquad \mathbb{E}\big[T^3\big] < \infty.$$

Fix $\kappa > 0$, $\alpha < \alpha_c(\kappa)$, $m \in \mathbb{N}$, $0 < \eta < \beta < 1$. Then,

$$\mathbb{E}_{\mathcal{M} \in \mathbb{R}^{M \times n} : \mathcal{M}_{ij} \sim \mathcal{D}, i.i.d.} \Big[\mathcal{S}_\kappa\big(\beta, \eta, m, \alpha, \{0\}\big)\Big] \leq \mathbb{E}_{\mathcal{M} \in \mathbb{R}^{M \times n} : \mathcal{M}_{ij} \sim \mathcal{N}(0,1), i.i.d.} \Big[\mathcal{S}_\kappa\big(\beta, \eta, m, \alpha, \{0\}\big)\Big] e^{O(\sqrt{n})}.$$

The proof of Theorem 3.5.2 is provided in Section 3.6.7. Hence, if $0 < \eta < \beta < 1$ and $m \in \mathbb{N}$ are such that

$$\mathbb{E}_{\mathcal{M} \in \mathbb{R}^{M \times n} : \mathcal{M}_{ij} \sim \mathcal{N}(0,1), \text{i.i.d.}} \Big[\mathcal{S}_\kappa\big(\beta, \eta, m, \alpha, \{0\}\big)\Big] = \exp\big(-\Theta(n)\big),$$

then

$$\mathbb{E}_{\mathcal{M} \in \mathbb{R}^{M \times n} : \mathcal{M}_{ij} \sim \mathcal{D}, \text{i.i.d.}} \Big[\mathcal{S}_\kappa\big(\beta, \eta, m, \alpha, \{0\}\big)\Big] = \exp\big(-\Theta(n)\big).$$

A particular case of interest is when the (i.i.d.) entries of the disorder matrix is Rademacher. That is, $\mathbb{P}[\mathcal{M}_{ij} = 1] = 1/2 = \mathbb{P}[\mathcal{M}_{ij} = -1]$, i.i.d. across $1 \leq i \leq M$ and $1 \leq j \leq n$. In this case, Theorem 3.5.2 asserts that the $m-$OGP holds with the exact same parameters appearing in Theorem 3.2.4.

3.6 Proofs

3.6.1 Some Auxiliary Results

Our $2-$OGP and $3-$OGP results for the high κ case (namely Theorem 3.2.2 and Theorem 3.2.3) require the following auxiliary result. We remind the reader that $h(\cdot)$ is the binary entropy function logarithm base two.

Lemma 3.6.1. *Let*

$$f_1(\Delta, \alpha) = 1 + h(\Delta) + \alpha \log_2 \mathbb{P}\big[|Z_1| \leq 1, |Z_2| \leq 1\big], \tag{3.17}$$

where $(Z_1, Z_2) \sim \mathcal{N}\big(0, \Delta I + (1-\Delta)\mathbf{e}\mathbf{e}^T\big)$. Let

$$f_2(\beta, \alpha) \triangleq 1 + h\left(\frac{1-\beta}{2}\right) + \alpha \log_2 \mathbb{P}\big[|Z_1| \leq 1, |Z_2| \leq 1\big], \tag{3.18}$$

where $(Z_1, Z_2) \sim \mathcal{N}\big(0, (1-\beta)I_2 + \beta \mathbf{e}\mathbf{e}^T\big)$*; and let*

$$f_3(\beta, \alpha) \triangleq 1 + \frac{1-\beta}{2} + h\left(\frac{1-\beta}{2}\right) + \frac{1+\beta}{2} h\left(\frac{1-\beta}{2(1+\beta)}\right) + \alpha \log_2 \mathbb{P}\big[|Z_i| \leq 1, 1 \leq i \leq 3\big], \tag{3.19}$$

where $(Z_1, Z_2, Z_3) \sim \mathcal{N}\big(0, (1-\beta)I_3 + \beta \mathbf{e}\mathbf{e}^T\big)$*. Then, the following holds.*

(a) Let $S_1(\alpha) = \big\{\Delta \in [0.00001, 0.1] : f_1(\Delta, \alpha) < 0\big\}$*. Then,* $S_1(1.77) \neq \varnothing$*. Hence,* $S_1(\alpha) \neq \varnothing$*,* $\forall \alpha \geq 1.77$*.*

(b) Let $S_2(\alpha) = \big\{\beta \in (0,1) : f_2(\beta, \alpha) < 0\big\}$*. Then,* $S_2(1.71) \neq \varnothing$*. Hence,* $S_2(\alpha) \neq \varnothing$*,* $\forall \alpha \geq 1.710$*.*

(c) Let $S_3(\alpha) = \big\{\beta \in (0,1) : f_3(\beta, \alpha) < 0\big\}$*. Then,* $S_3(1.667) \neq \varnothing$*. Hence,* $S_3(\alpha) \neq \varnothing$*,* $\forall \alpha \geq 1.667$*.*

Lemma 3.6.1 is established numerically using MATLAB's `mvncdf` function to evaluate the probability term. The accompanying code is provided in Appendix A.

We next record two useful auxiliary results regarding bivariate Gaussian random variables. These will later be useful in Section 3.6.6 to prove the stability of Kim-Roche algorithm. Our first lemma to that end pertains to the quadrant probabilities for the bivariate normal distribution.

Lemma 3.6.2. *Let* (X, Y) *be a bivariate normal random variable with*

$$(X, Y) \overset{d}{=} \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right).$$

Then,

$$\mathbb{P}\big(X \geq 0, Y \geq 0\big) = \frac{1}{4} + \frac{1}{2\pi} \sin^{-1}(\rho).$$

Lemma 3.6.2 is quite well-known; a proof is provided below for completeness.

Proof of Lemma 3.6.2. Note that the pair (X, Y) has bivariate normal distribution with parameter ρ. Next, define

$$Z \triangleq \frac{Y - \rho X}{\sqrt{1-\rho^2}}.$$

Clearly X and Z are i.i.d. standard normals. Let $\zeta \triangleq -\frac{\rho}{\sqrt{1-\rho^2}}$. Observe that

$$\begin{aligned}
\mathbb{P}\big(X \geq 0, Y \geq 0\big) &= \mathbb{P}\left(X \geq 0, Z \geq -\frac{\rho}{\sqrt{1-\rho^2}} X\right) \\
&= \int_{x=0}^{\infty} \int_{z=\zeta x}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{x^2+z^2}{2}\right)\, dx\, dz \\
&= \int_{\theta=\tan^{-1}(\zeta)}^{\frac{\pi}{2}} \int_{r=0}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{r^2}{2}\right) r\, dr\, d\theta \\
&= \frac{1}{2\pi} \int_{\theta=\tan^{-1}(\zeta)}^{\frac{\pi}{2}} d\theta \\
&= \frac{1}{4} - \frac{1}{2\pi} \tan^{-1}(\zeta) \\
&= \frac{1}{4} + \frac{1}{2\pi} \sin^{-1} \rho.
\end{aligned}$$

Here, the second line uses the independence of X and Z; the third line is obtained upon passing to polar coordinates; and the last line follows from the fact $\tan^{-1}$ is an odd function, and that if $\tan\theta = \frac{\rho}{\sqrt{1-\rho^2}}$ then $\sin\theta = \rho$. □

Our next lemma is as follows.

Lemma 3.6.3. *Let $Z_1, Z_2 \stackrel{d}{=} \mathcal{N}(0,1)$ where (Z_1, Z_2) is a bivariate normal with parameter ρ: $\mathbb{E}[Z_1 Z_2] = \rho$. Then*

$$\mathbb{E}\big[Z_1 | Z_2 \geq 0\big] = \rho\sqrt{\frac{2}{\pi}}.$$

Proof of Lemma 3.6.3. Note that $Z_3 := \frac{Z_1 - \rho Z_2}{\sqrt{1-\rho^2}}$ is a standard normal, independent of Z_2 (as $\mathbb{E}[Z_2 Z_3] = 0$ and (Z_2, Z_3) is also bivariate normal). Hence

$$\begin{aligned}
\mathbb{E}\big[Z_1 | Z_2 \geq 0\big] &= \mathbb{E}\big[\rho Z_2 + \sqrt{1-\rho^2} Z_3 | Z_2 \geq 0\big] \\
&= \rho \mathbb{E}\big[Z_2 | Z_2 \geq 0\big] \\
&= \rho \frac{\mathbb{E}\big[Z_2 \mathbb{1}\{Z_2 \geq 0\}\big]}{\mathbb{P}(Z_2 \geq 0)} \\
&= 2\rho \int_0^{\infty} \frac{1}{\sqrt{2\pi}} x \exp\left(-\frac{x^2}{2}\right)\, dx \\
&= \rho\sqrt{\frac{2}{\pi}},
\end{aligned}$$

yielding Lemma 3.6.3. □

3.6.2 Proof of Theorem 3.2.3

Our proof is based on the *first moment method*. We begin by observing the following monotonicity: if

$$\alpha \le \alpha' \le \alpha_c(1) = -\frac{1}{\log_2 \mathbb{P}\big(|\mathcal{N}(0,1)| \le 1\big)} \approx 1.8159,$$

then

$$\mathbb{P}\Big[\mathcal{S}(\beta,\eta,3,\alpha',\mathcal{I}) \neq \varnothing\Big] \le \mathbb{P}\Big[\mathcal{S}(\beta,\eta,3,\alpha,\mathcal{I}) \neq \varnothing\Big].$$

For this reason, it suffices to consider $\alpha = 1.667$.

Counting term. Let $0 < \eta < \beta < 1$. We first count the number of triples $(\sigma^{(i)} : 1 \le i \le 3)$ in $\mathcal{B}_n$ subject to the overlap condition. (In what follows, we omit floor/ceiling operations to keep our exposition clean.)

Lemma 3.6.4. *Let $0 < \eta < \beta < 1$ be fixed. Denote by $M(\beta,\eta)$ the number of triples $(\sigma^{(i)} : 1 \le i \le 3)$ with $\sigma^{(i)} \in \mathcal{B}_n$ subject to the condition*

$$\beta - \eta \le \mathcal{O}\sigma^{(i)}\sigma^{(j)} = \frac{1}{n}\left\langle \sigma^{(i)}, \sigma^{(j)} \right\rangle \le \beta, \qquad 1 \le i < j \le 3.$$

Then,

$$M(\beta,\eta) \le \exp_2\Big(n\varphi_{\mathrm{Count}}(\beta,\eta) + O(\log_2 n)\Big), \tag{3.20}$$

where

$$\varphi_{\mathrm{Count}}(\beta,\eta) = 1 + h\left(\frac{1-\beta+\eta}{2}\right) + \frac{1-\beta+\eta}{2} + \frac{1+\beta}{2}h\left(\frac{1-\beta+2\eta}{2(1+\beta)}\right). \tag{3.21}$$

In particular, for any fixed β, the map $\eta \mapsto \varphi_{\mathrm{Count}}(\beta,\eta)$ is continuous at $\eta = 0$.

Proof of Lemma 3.6.4. For $\sigma^{(k)} \in \mathcal{B}_n$, denote its i^{th} coordinate $(1 \le i \le n)$ by $\sigma_i^{(k)}$.

Clearly, there are 2^n ways of choosing $\sigma^{(1)}$. Having fixed $\sigma^{(1)}$, there are $\binom{n}{n\frac{1-\rho}{2}}$ ways of choosing $\sigma^{(2)}$ with $\mathcal{O}\sigma^{(1)}\sigma^{(2)} = \rho \in [\beta-\eta, \beta]$. Assume now that both $\sigma^{(1)}$ and $\sigma^{(2)}$ are fixed, and define $I \subset [n]$ with $|I| = n\frac{1-\rho}{2}$ as

$$I \triangleq \Big\{1 \le i \le n : \sigma_i^{(1)} \neq \sigma_i^{(2)}\Big\},$$

and let $I^c \triangleq [n] \setminus I$ with $|I^c| = n\frac{1+\rho}{2}$. In particular, $\sigma^{(1)}$ and $\sigma^{(2)}$ agree on coordinates in I^c and disagree on coordinates in I. Having fixed $\sigma^{(1)}$ and $\sigma^{(2)}$, now let $N_3(\rho)$ denote the number of all admissible $\sigma^{(3)}$ satisfying the inner product condition (with $\sigma^{(1)}$ and $\sigma^{(2)}$). Then, it is evident that

$$M(\beta,\eta) = 2^n \sum_{\rho:\beta-\eta\le\rho\le\beta,\rho n\in\mathbb{N}} \binom{n}{n\frac{1-\rho}{2}} N_3(\rho). \tag{3.22}$$

Now, suppose that

$$t_1 \triangleq \left|\left\{i \in I : \sigma_i^{(2)} = \sigma_i^{(3)}\right\}\right| \qquad \text{and} \qquad t_2 \triangleq \left|\left\{i \in I^c : \sigma_i^{(2)} \neq \sigma_i^{(3)}\right\}\right|.$$

Then

$$d_H\big(\sigma^{(1)}, \sigma^{(3)}\big) = t_1 + t_2 \quad \text{and} \quad d_H\big(\sigma^{(2)}, \sigma^{(3)}\big) = n\frac{1-\rho}{2} - t_1 + t_2.$$

Next, observe that using the condition on $\mathcal{O}\sigma^{(1)}\sigma^{(3)}$ and $\mathcal{O}\sigma^{(2)}\sigma^{(3)}$, we arrive at

$$n\frac{1-\beta}{2} \leq d_H\big(\sigma^{(1)}, \sigma^{(3)}\big) = t_1 + t_2 \leq n\frac{1-\beta+\eta}{2},$$

and

$$n\frac{1-\beta}{2} \leq d_H\big(\sigma^{(2)}, \sigma^{(3)}\big) = t_2 - t_1 + n\frac{1-\rho}{2} \leq n\frac{1-\beta+\eta}{2}.$$

We thus arrive at

$$n\frac{1-\beta}{2} \leq t_1 + t_2 \leq n\frac{1-\beta+\eta}{2} \quad \text{and} \quad n\frac{\rho-\beta}{2} \leq t_2 - t_1 \leq n\frac{\rho-\beta+\eta}{2}.$$

This yields the following lower and upper bounds on t_1, t_2:

$$n\frac{1-\rho-\eta}{4} \leq t_1 \leq n\frac{1-\rho+\eta}{4} \tag{3.23}$$

$$n\frac{1+\rho-2\beta}{4} \leq t_2 \leq n\frac{1+\rho-2\beta+2\eta}{4}. \tag{3.24}$$

Define now the rectangle

$$\mathcal{T} \triangleq \left[\frac{1-\rho-\eta}{4}, \frac{1-\rho+\eta}{4}\right] \times \left[\frac{1+\rho-2\beta}{4}, \frac{1+\rho-2\beta+2\eta}{4}\right].$$

Note that (a) for η small enough, $\mathcal{T} \subset (0,\infty)^2$; and (b) the set of all admissible $(t_1, t_2) \in \mathbb{N}^2$ pairs are precisely the set of all lattice points in the box $n\mathcal{T}$. Having fixed $\sigma^{(1)}$ and $\sigma^{(2)}$, the number $N_3(\rho)$ of admissible $\sigma^{(3)}$ then computes as

$$N_3(\rho) = \sum_{(t_1,t_2)\in\mathbb{N}^2\cap n\mathcal{T}} \binom{n\frac{1-\rho}{2}}{t_1}\binom{n\frac{1+\rho}{2}}{t_2}. \tag{3.25}$$

Note that using the fact $\binom{n}{\alpha}$ is maximized for $\alpha = \lfloor n/2 \rfloor$, we obtain

$$\binom{n\frac{1-\rho}{2}}{t_1}\binom{n\frac{1+\rho}{2}}{t_2} \leq \binom{n\frac{1-\rho}{2}}{n\frac{1-\rho}{4}}\binom{n\frac{1+\rho}{2}}{n\frac{1+\rho-2\beta+2\eta}{4}}, \quad \forall (t_1,t_2) \in \mathbb{N}^2 \cap n\mathcal{T}. \tag{3.26}$$

Now, we use the well-known asymptotic on binomial coefficients: for any $r \in (0,1)$, $\binom{n}{nr} = \exp_2(nh(r) + O(\log_2 n))$. Combining this fact together with (3.25) and (3.26),

we obtain the following upper bound on number of such $\sigma^{(3)}$:

$$N_3(\rho) \le \exp_2\left(n\frac{1-\rho}{2} + n\frac{1+\rho}{2}h\left(\frac{1+\rho-2\beta+2\eta}{2(1+\rho)}\right) + O(\log_2 n)\right). \tag{3.27}$$

We next study the argument $(1+\rho-2\beta+2\eta)/2(1+\rho)$ of the entropy term appearing in (3.27). Clearly, as $\eta < \beta$, the argument is less than $1/2$. Now, observe that

$$\begin{aligned}
&\frac{1+\rho_1-2\beta+2\eta}{2(1+\rho_1)} < \frac{1+\rho_2-2\beta+2\eta}{2(1+\rho_2)} \\
&\Leftrightarrow \frac{1}{2} - \frac{\beta-\eta}{1+\rho_1} < \frac{1}{2} - \frac{\beta-\eta}{1+\rho_2} \\
&\Leftrightarrow \frac{1}{1+\rho_2} < \frac{1}{1+\rho_1} \Leftrightarrow \rho_1 < \rho_2.
\end{aligned}$$

Consequently, using the monotonicity of h in $[0, \frac{1}{2}]$,

$$h\left(\frac{1+\rho-2\beta+2\eta}{2(1+\rho)}\right) < h\left(\frac{1-\beta+2\eta}{2(1+\beta)}\right).$$

Using this and (3.27), $N_3(\rho)$ is further upper bounded by

$$N_3(\rho) \le \exp_2\left(n\frac{1-\beta+\eta}{2} + n\frac{1+\beta}{2}h\left(\frac{1-\beta+2\eta}{2(1+\beta)}\right) + O(\log_2 n)\right). \tag{3.28}$$

Finally, we have

$$\binom{n}{n\frac{1-\rho}{2}} \le \binom{n}{n\frac{1-\beta+\eta}{2}} = \exp_2\left(nh\left(\frac{1-\beta+\eta}{2}\right) + O(\log_2 n)\right). \tag{3.29}$$

Combining (3.22), (3.28) and (3.29), we obtain

$$\begin{aligned}
M(\beta,\eta) &\le \exp_2\left(n + nh\left(\frac{1-\beta+\eta}{2}\right) + n\frac{1-\beta+\eta}{2} + n\frac{1+\beta}{2}h\left(\frac{1-\beta+2\eta}{2(1+\beta)}\right) + O(\log_2 n)\right) \\
&= \exp_2\Big(n\varphi_{\text{count}}(\beta,\eta) + O(\log_2 n)\Big),
\end{aligned}$$

yielding (3.20). Since the continuity follows immediately from the continuity of the entropy, the proof of Lemma 3.6.4 is complete. □

Probability term. Now, fix any $\big(\sigma^{(i)} : 1 \le i \le 3\big)$ with $\beta - \eta \le \mathcal{O}\sigma^{(i)}\sigma^{(j)} \le \beta$. More concretely, let

$$\mathcal{O}\sigma^{(i)}\sigma^{(j)} \triangleq \beta - \eta_{ij}, \qquad \text{where} \qquad 0 \le \eta_{ij} \le \eta, \quad 1 \le i < j \le 3. \tag{3.30}$$

We control the probability term

$$\mathbb{P}\left[\exists \tau_1, \tau_2, \tau_3 \in \mathcal{I} : n^{-\frac{1}{2}} \left|\mathcal{M}_i(\tau_i)\sigma^{(i)}\right| \leq \mathbf{e}, 1 \leq i \leq 3\right], \tag{3.31}$$

where $\mathbf{e} \in \mathbb{R}^{M\times 1}$ is the vector of all ones, and the inequality is coordinate-wise. As a first step, we take a union bound over $\mathcal{I}$ to obtain

$$\begin{aligned}&\mathbb{P}\left[\exists \tau_1, \tau_2, \tau_3 \in \mathcal{I} : n^{-\frac{1}{2}} \left|\mathcal{M}_i(\tau_i)\sigma^{(i)}\right| \leq \mathbf{e}, 1 \leq i \leq 3\right] \\ &\leq |\mathcal{I}|^3 \max_{\tau_i \in \mathcal{I}, 1\leq i\leq 3} \mathbb{P}\left[n^{-\frac{1}{2}} \left|\mathcal{M}_i(\tau_i)\sigma^{(i)}\right| \leq \mathbf{e}, 1 \leq i \leq 3\right]. \end{aligned} \tag{3.32}$$

Next, let the first row of $\mathcal{M}_i(\tau_i)$ be $\mathtt{R}_i \stackrel{d}{=} \mathcal{N}(0, I_n) \in \mathbb{R}^n$, $1 \leq i \leq 3$. Using the independence across rows, we have

$$\mathbb{P}\left[n^{-\frac{1}{2}} \left|\mathcal{M}_i(\tau_i)\sigma^{(i)}\right| \leq \mathbf{e}, 1 \leq i \leq 3\right] = \mathbb{P}\left[n^{-\frac{1}{2}} \left|\left\langle \mathtt{R}_i, \sigma^{(i)}\right\rangle\right| \leq 1, 1 \leq i \leq 3\right]^{\alpha n}. \tag{3.33}$$

To upper bound the probability appearing in (3.33), observe that $\left(n^{-1/2}\left\langle \mathtt{R}_i, \sigma^{(i)}\right\rangle : 1 \leq i \leq 3\right)$ is a multivariate normal with each component having zero mean and unit variance. We now compute its covariance matrix $\overline{\Sigma} \in \mathbb{R}^{3\times 3}$. Observe that for $i \neq j$,

$$\begin{aligned}\overline{\Sigma}_{ij} &= \mathbb{E}\left[n^{-\frac{1}{2}}\left\langle \mathtt{R}_i, \sigma^{(i)}\right\rangle \cdot n^{-\frac{1}{2}}\left\langle \mathtt{R}_j, \sigma^{(j)}\right\rangle\right] \\ &= \frac{1}{n}\left(\sigma^{(i)}\right)^T \mathbb{E}\left[\mathtt{R}_i \mathtt{R}_j^T\right]\sigma^{(j)} \\ &= \cos(\tau_i)\cos(\tau_j)\mathcal{O}\sigma^{(i)}\sigma^{(j)} \\ &= \cos(\tau_i)\cos(\tau_j)\left(\beta - \eta_{ij}\right),\end{aligned}$$

where the last line uses (3.30). In order to remove the dependence of $\overline{\Sigma}$ on τ_i, we now employ the following Gaussian comparison inequality established by Sidák [256, Corollary 1].

Theorem 3.6.5. *Let $(X_1, \ldots, X_k) \in \mathbb{R}^k$ be a multivariate normal random vector each of whose coordinates have zero mean and unit variance. Suppose that its covariance matrix $\Sigma \in \mathbb{R}^{k\times k}$ has the following form: there exists $\lambda_1, \ldots, \lambda_k$ ($0 \leq \lambda_i \leq 1$, $1 \leq i \leq k$) such that for every $1 \leq i \neq j \leq k$, $\Sigma_{ij} = \lambda_i\lambda_j\rho_{ij}$ where $(\rho_{ij} : 1 \leq i \neq j \leq k)$ is some fixed covariance matrix. Fix any $c_1, \ldots, c_k > 0$, and denote*

$$P(\lambda_1, \ldots, \lambda_k) = \mathbb{P}\left[|X_1| < c_1, |X_2| < c_2, \ldots, |X_k| < c_k\right].$$

Then, $P(\lambda_1, \ldots, \lambda_k)$ is a non-decreasing function of each λ_i, $i = 1, 2, \ldots, k$, $0 \leq \lambda_i \leq 1$. That is,

$$P(\lambda_1, \lambda_2, \ldots, \lambda_k) \leq P(1, 1, \ldots, 1).$$

Applying Theorem 3.6.5, we find that

$$\max_{\tau_i \in \mathcal{I}, 1\leq i\leq 3} \mathbb{P}\left[n^{-\frac{1}{2}} \left|\left\langle \mathtt{R}_i, \sigma^{(i)}\right\rangle\right| \leq 1, 1 \leq i \leq 3\right] \leq \mathbb{P}\left[|Z_1| \leq 1, |Z_2| \leq 1, |Z_3| \leq 1\right] \tag{3.34}$$

where

$$(Z_1, Z_2, Z_3) \stackrel{d}{=} \mathcal{N}\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}1 & \beta-\eta_{12} & \beta-\eta_{13}\\ \beta-\eta_{12} & 1 & \beta-\eta_{23}\\ \beta-\eta_{13} & \beta-\eta_{23} & 1\end{bmatrix}\right). \tag{3.35}$$

Now, note that

$$\Sigma \triangleq \begin{bmatrix}1 & \beta-\eta_{12} & \beta-\eta_{13}\\ \beta-\eta_{12} & 1 & \beta-\eta_{23}\\ \beta-\eta_{13} & \beta-\eta_{23} & 1\end{bmatrix} = (1-\beta)I_3 + \beta\mathbf{e}\mathbf{e}^T + E, \tag{3.36}$$

where $E \in \mathbb{R}^{3\times 3}$ has zero diagonal entries, and $E_{ij} = -\eta_{ij}$ for $1 \le i \ne j \le 3$. In particular, $\|E\|_F \le \eta\sqrt{6}$. Since the eigenvalues of $(1-\beta)I_3 + \beta\mathbf{e}\mathbf{e}^T$ are $1+2\beta$ (with multiplicity one) and $1-\beta$ (with multiplicity two), it follows that provided $\eta < (1-\beta)/\sqrt{6}$, the covariance matrix Σ appearing in (3.36) is invertible. We assume that this is indeed the case from this point on.

Define

$$\varphi_{\mathrm{Prob}}\big(\beta, \eta_{12}, \eta_{13}, \eta_{23}\big) \triangleq \mathbb{P}\big(|Z_1| \le 1, |Z_2| \le 1, |Z_3| \le 1\big) \tag{3.37}$$

where (Z_1, Z_2, Z_3) has distribution in (3.35). We now combine (3.31), (3.32), (3.33), (3.34), and (3.37) to arrive at

$$\mathbb{P}\left[\exists \tau_1, \tau_2, \tau_3 \in \mathcal{I} : n^{-\frac{1}{2}}\left|\mathcal{M}_i(\tau_i)\sigma^{(i)}\right| \le \mathbf{e}, 1 \le i \le 3\right] \le |\mathcal{I}|^3 \varphi_{\mathrm{Prob}}\big(\beta, \eta_{12}, \eta_{13}, \eta_{23}\big)^{\alpha n}. \tag{3.38}$$

Our next technical result pertains to φ_{Prob}.

Lemma 3.6.6. *Fix any $\beta \in (0,1)$. Then the map*

$$\big(\eta_{12}, \eta_{13}, \eta_{23}\big) \mapsto \log_2 \varphi_{\mathrm{Prob}}\big(\beta, \eta_{12}, \eta_{13}, \eta_{23}\big)$$

(from $\mathbb{R}^3$ to $\mathbb{R}$) is continuous at $(0,0,0)$.

Proof of Lemma 3.6.6. Define a sequence $(\boldsymbol{\zeta_k})_{k\ge 1}$ such that

$$\boldsymbol{\zeta_k} = \big(\eta_{12}(k), \eta_{13}(k), \eta_{23}(k)\big) \in \mathbb{R}^3 \quad \text{and} \quad \lim_{k\to\infty} \boldsymbol{\zeta_k} = (0,0,0).$$

Let $\Sigma_k \in \mathbb{R}^{3\times 3}$ be the matrix Σ appearing in (3.36) with parameters $\boldsymbol{\zeta_k}$. Let $\mathbf{v} = (x,y,z)$, and define functions

$$f_k(\mathbf{v}) \triangleq \exp\left(-\frac{1}{2}\mathbf{v}^T\Sigma_k^{-1}\mathbf{v}\right).$$

Moreover, let

$$\Sigma_\infty \triangleq (1-\beta)I + \beta\mathbf{e}\mathbf{e}^T \quad \text{and} \quad f_\infty(\mathbf{v}) \triangleq \exp\left(-\frac{1}{2}\mathbf{v}^T\Sigma_\infty^{-1}\mathbf{v}\right).$$

Note that, $|f_k(\mathbf{v})| \le 1$ for every $\mathbf{v} \in \mathbb{R}^3$ (as long as Σ_k is positive definite). Moreover,

we have the pointwise convergence:

$$\lim_{k\to\infty} f_k(\mathbf{v}) = f_\infty(\mathbf{v}) \qquad \text{for all} \qquad \mathbf{v}\in\mathbb{R}^3.$$

Therefore, by the dominated convergence theorem

$$\lim_{k\to\infty}\int_{[-1,1]^3} f_k(\mathbf{v})\,d\mathbf{v} = \int_{[-1,1]^3} f_\infty(\mathbf{v})\,d\mathbf{v}. \tag{3.39}$$

Moreover,

$$\lim_{k\to\infty}\big|\Sigma_k\big| = \big|\Sigma_\infty\big|. \tag{3.40}$$

Finally,

$$\log_2\varphi_{\mathrm{Prob}}\big(\beta,\boldsymbol{\zeta_k}\big) = -\frac{3}{2}\log_2(2\pi) - \frac{1}{2}\log_2\big|\Sigma_k\big| + \log_2\left(\int_{[-1,1]^3} f_k(\mathbf{v})\,d\mathbf{v}\right).$$

Since $x\mapsto\log_2 x$ is continuous, we obtain by combining (3.39) and (3.40) that

$$\lim_{k\to\infty}\log_2\varphi_{\mathrm{Prob}}\big(\beta,\boldsymbol{\zeta_k}\big) = \log_2\varphi_{\mathrm{Prob}}(\beta,0,0,0).$$

Since the sequence $(\boldsymbol{\zeta_k})_{k\geq 1}$ is arbitrary, the proof of Lemma 3.6.6 is complete. □

Choice of β,η. In the remainder, let $\alpha^* = 1.667$. Notice next that

$$f_3(\beta,\alpha) = \varphi_{\mathrm{Count}}\big(\beta,0\big) + \alpha\cdot\log_2\varphi_{\mathrm{Prob}}\big(\beta,0,0,0\big),$$

where f_3 is defined in (3.19), φ_{Count} is defined in (3.21); and φ_{Prob} is defined in (3.37). Let β^* be such that

$$f^* \triangleq f_3\big(\beta^*,\alpha^*\big) = \inf_{\beta\in[0,1]} f_3(\beta,\alpha^*). \tag{3.41}$$

Since $S_3(\alpha^*)\neq\varnothing$ by Lemma 3.6.1, it follows $f^*<0$. Having fixed β^*, let

$$\epsilon^* \triangleq -f^*/8\alpha^* > 0 \qquad \text{and} \qquad c^* \triangleq -f^*/24. \tag{3.42}$$

Using Lemma 3.6.6, it follows that there exists a $\delta_1^* \triangleq \delta_1^*\big(\beta^*,\epsilon^*\big) > 0$, such that

$$\sup_{\substack{(\eta_{12},\eta_{13},\eta_{23})\\ |\eta_{ij}|<\delta_1^*, 1\leq i<j\leq 3}} \Big|\log_2\varphi_{\mathrm{Prob}}(\beta^*,\eta_{12},\eta_{13},\eta_{23}) - \log_2\varphi_{\mathrm{Prob}}(\beta^*,0,0,0)\Big| < \epsilon^*. \tag{3.43}$$

Take $\eta<\delta_1^*$, ensuring $0\leq\eta_{ij}\leq\eta<\delta_1^*$, $1\leq i<j\leq 3$. Using Markov's inequality, we have

$$\mathbb{P}\big(\mathcal{S}(\beta^*,\eta,3,\alpha^*,\mathcal{I})\neq\varnothing\big) = \mathbb{P}\big(\big|\mathcal{S}(\beta^*,\eta,3,\alpha^*,\mathcal{I})\big|\geq 1\big) \leq \mathbb{E}\big[\big|\mathcal{S}(\beta^*,\eta,3,\alpha^*,\mathcal{I})\big|\big].$$

We now combine the counting bound (3.20), the probability bound (3.38) and the bound (3.43) to upper bound $\mathbb{E}\big[\big|\mathcal{S}(\beta^*,\eta,3,\alpha^*,\mathcal{I})\big|\big]$:

$$\begin{aligned}\mathbb{E}\big[\big|\mathcal{S}(\beta^*,\eta,3,\alpha^*,\mathcal{I})\big|\big] &\le \exp_2\Big(n\varphi_{\rm Count}(\beta^*,\eta)+n\alpha^*\log_2\varphi_{\rm Prob}(\beta^*,0,0,0)+n\alpha^*\epsilon^*+3\log_2|\mathcal{I}|+O(\log_2 n)\\ &\le \exp_2\left(n\left(\varphi_{\rm Count}(\beta^*,\eta)+\alpha^*\log_2\varphi_{\rm Prob}(\beta^*,0,0,0)-\frac{f^*}{4}+O\left(\frac{\log_2 n}{n}\right)\right)\right),\end{aligned} \tag{3.44}$$

where (3.44) uses $\max\{\alpha^*\epsilon^*, 3\log_2|\mathcal{I}|\} \le -f^*/8$ which follows from (3.42). Now, using the continuity of $\eta \mapsto \varphi_{\rm Count}(\beta^*,\eta)$ at $\eta=0$, it follows that there is a $\delta_2^* \triangleq \delta_2^*(\beta^*) > 0$ such that

$$|\eta| < \delta_2^* \implies \varphi_{\rm Count}(\beta^*,\eta) < \varphi_{\rm Count}(\beta^*,0) - \frac{f^*}{4}. \tag{3.45}$$

Finally, we let

$$\eta^* = \frac{1}{2}\min\big\{\delta_1^*,\delta_2^*\big\}. \tag{3.46}$$

With this choice of η^*, we have by using (3.44) and (3.45) that

$$\begin{aligned}\mathbb{E}\big[\big|\mathcal{S}(\beta^*,\eta^*,3,\alpha^*,\mathcal{I})\big|\big] &\le \exp_2\left(n\left(\varphi_{\rm Count}(\beta^*,\eta^*)+\alpha^*\log_2\varphi_{\rm Prob}(\beta^*,0,0,0)-\frac{f^*}{4}+O\left(\frac{\log_2 n}{n}\right)\right)\right)\\ &\le \exp_2\left(n\left(\varphi_{\rm Count}(\beta^*,0)+\alpha^*\log_2\varphi_{\rm Prob}(\beta^*,0,0,0)-\frac{f^*}{2}+O\left(\frac{\log_2 n}{n}\right)\right)\right)\\ &\le \exp_2\left(n\left(\frac{f^*}{2}+O\left(\frac{\log_2 n}{n}\right)\right)\right)\\ &= \exp_2\Big(-\Theta(n)\Big),\end{aligned}$$

using the fact per (3.41) that $f^* < 0$. This completes the proof of Theorem 3.2.3.

3.6.3 Proof of Theorem 3.2.4

Fix $\kappa > 0$. We start by observing the following obvious monotonicity property: for any fixed $0 < \eta < \beta < 1$, $m \in \mathbb{N}$, $\mathcal{I} \subset [0,\pi/2]$, and $\alpha \le \alpha'$; we have

$$\mathbb{P}\Big[\mathcal{S}_\kappa(\beta,\eta,m,\alpha',\mathcal{I}) \neq \varnothing\Big] \le \mathbb{P}\Big[\mathcal{S}_\kappa(\beta,\eta,m,\alpha,\mathcal{I}) \neq \varnothing\Big].$$

For this reason, it suffices to establish the result for $\alpha = \alpha_{\rm OGP}(\kappa) = 10\kappa^2\log\frac{1}{\kappa}$. We will do so by using the *first moment method*: note that by Markov's inequality,

$$\mathbb{P}\Big[\mathcal{S}_\kappa(\beta,\eta,m,\alpha,\mathcal{I}) \neq \varnothing\Big] = \mathbb{P}\Big[\Big|\mathcal{S}_\kappa(\beta,\eta,m,\alpha,\mathcal{I})\Big| \ge 1\Big] \le \mathbb{E}\Big[\Big|\mathcal{S}_\kappa(\beta,\eta,m,\alpha,\mathcal{I})\Big|\Big]. \tag{3.47}$$

We now study $\mathbb{E}\Big[\Big|\mathcal{S}_\kappa(\beta,\eta,m,\alpha,\mathcal{I})\Big|\Big]$.

Counting term. Fix any $m \in \mathbb{N}$, and $0 < \eta < \beta < 1$. We upper bound the number $M(m, \beta, \eta)$ of the $m-$tuples $\big(\sigma^{(i)} : 1 \le i \le m\big)$, $\sigma^{(i)} \in \mathcal{B}_n$, subject to the constraint $\beta - \eta \le n^{-1} \left\langle \sigma^{(i)}, \sigma^{(j)} \right\rangle \le \beta$ for $1 \le i < j \le m$.

Lemma 3.6.7.

$$M(m, \beta, \eta) \le \exp_2 \left(n + n(m-1) h \left(\frac{1 - \beta + \eta}{2} \right) + O(\log_2 n) \right). \tag{3.48}$$

Proof of Lemma 3.6.7. Note that for any $\sigma, \sigma' \in \mathcal{B}_n$, $\langle \sigma, \sigma' \rangle = n - 2 d_H(\sigma, \sigma')$. There are 2^n choices for $\sigma^{(1)}$. Having chosen a $\sigma^{(1)}$; any $\sigma^{(i)}$, $2 \le i \le m$, can be chosen in

$$\sum_{\substack{\rho : \frac{1-\beta}{2} \le \rho \le \frac{1-\beta+\eta}{2} \\ \rho n \in \mathbb{N}}} \binom{n}{n\rho} \le \binom{n}{n\frac{1-\beta+\eta}{2}} n^{O(1)},$$

different ways, subject to the constraint that $\beta - \eta \le n^{-1} \left\langle \sigma^{(1)}, \sigma^{(i)} \right\rangle \le \beta$. For any $\rho \in (0,1)$, $\binom{n}{n\rho} = \exp_2\big(nh(\rho) + O\big(\log_2 n\big)\big)$ by Stirling's approximation. Combining these, and recalling $m = O(1)$ (as $n \to \infty$), we obtain (3.48). □

Probability term. Fix any $\big(\sigma^{(i)} : 1 \le i \le m\big)$ with the pairwise overlaps

$$n^{-1} \left\langle \sigma^{(i)}, \sigma^{(j)} \right\rangle = \beta - \eta_{ij}, \qquad 1 \le i < j \le m.$$

Evidently, $\eta_{ij} \ge 0$ for all $i < j$. Moreover, if we set

$$\boldsymbol{\eta} = \big(\eta_{ij} : 1 \le i < j \le m\big) \in \mathbb{R}^{m(m-1)/2},$$

then,

$$\|\boldsymbol{\eta}\|_\infty \le \eta.$$

Lemma 3.6.8. *Let $\Sigma(\boldsymbol{\eta}) \in \mathbb{R}^{m \times m}$ be a matrix with the property that*

(a) $\big(\Sigma(\boldsymbol{\eta})\big)_{ii} = 1$ *for* $1 \le i \le m$.

(b) $\big(\Sigma(\boldsymbol{\eta})\big)_{ij} = \big(\Sigma(\boldsymbol{\eta})\big)_{ji} = \beta - \eta_{ij}$ *for every* $1 \le i < j \le m$.

Then,

(i) $\Sigma(\boldsymbol{\eta})$ *is positive definite (PD) if* $\eta < \frac{1-\beta}{m}$.

(ii) Assume $\Sigma(\boldsymbol{\eta})$ *is PD. Let* $(Z_1, Z_2, \ldots, Z_m) \sim \mathcal{N}\big(0, \Sigma(\boldsymbol{\eta})\big)$ *be a multivariate normal random vector, and define*

$$\varphi_{\mathrm{Prob}}\big(\beta, \boldsymbol{\eta}, \kappa\big) = \mathbb{P}\Big[|Z_i| \le \kappa, 1 \le i \le m \Big]. \tag{3.49}$$

Then,

$$\mathbb{P}\Big[\exists \tau_i \in \mathcal{I}, 1 \le i \le m : \big| \mathcal{M}_i(\tau_i) \sigma^{(i)} \big| \le (\kappa\sqrt{n}) \mathbf{e}, 1 \le i \le m \Big] \le |\mathcal{I}|^m \varphi_{\mathrm{Prob}}\big(\beta, \boldsymbol{\eta}, \kappa\big)^{\alpha n}. \tag{3.50}$$

(iii) We have

$$\varphi_{\text{Prob}}\big(\beta, \boldsymbol{\eta}, \kappa\big) \le \big(2\pi\big)^{-\frac{m}{2}} \big|\Sigma(\boldsymbol{\eta})\big|^{-\frac{1}{2}} \big(2\kappa\big)^m. \tag{3.51}$$

Proof of Lemma 3.6.8. (i) Note that $\Sigma(\boldsymbol{\eta}) = (1-\beta)I + \beta \mathbf{e}\mathbf{e}^T + E$, where $E \in \mathbb{R}^{m\times m}$ with $0 \le E_{ij} \le \eta$ for $1 \le i < j \le m$. In particular, $\|E\|_2 \le \|E\|_F \le \eta m$. Noting that the smallest eigenvalue of $(1-\beta)I + \beta\mathbf{e}\mathbf{e}^T$ is $1-\beta$, the result follows.

(ii) The result follows by taking a union bound over $\mathcal{I}$; and then applying the Gaussian comparison inequality, Theorem 3.6.5, in the exact same way as in the proof of Theorem 3.2.3.

(iii) Recall that the multivariate normal density for $(Z_i : 1 \le i \le m)$ is given by

$$f(z_1, \ldots, z_m) = \big(2\pi\big)^{-\frac{m}{2}} \big|\Sigma(\boldsymbol{\eta})\big|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{z}^T\Sigma(\boldsymbol{\eta})\boldsymbol{z}\right),$$

where $\boldsymbol{z} = (z_i : 1 \le i \le m)$, and thus

$$\mathbb{P}\Big[(Z_i : 1 \le i \le m) \in [-\kappa, \kappa]^m\Big] \le \big(2\pi\big)^{-\frac{m}{2}} \big|\Sigma(\boldsymbol{\eta})\big|^{-\frac{1}{2}} \big(2\kappa\big)^m;$$

using the fact

$$\exp\left(-\frac{1}{2}\boldsymbol{z}^T\Sigma(\boldsymbol{\eta})\boldsymbol{z}\right) \le 1,$$

for every $\boldsymbol{z} \in \mathbb{R}^m$ as $\Sigma(\boldsymbol{\eta})$ is PD.

□

Upper bounding the expectation. Assume $0 < \eta < \beta < 1$ and $m \in \mathbb{N}$ are fixed as $n \to \infty$; and that η is small enough, so that $\Sigma(\boldsymbol{\eta})$ is positive definite. (We will tune η eventually.) Combining the counting bound (3.48) arising from Lemma 3.6.7; the probability bounds (3.50) and (3.51) arising from Lemma 3.6.8; and $|\mathcal{I}| \le 2^{cn}$, we upper bound the expectation by

$$\begin{aligned}\mathbb{E}\Big[\Big|\mathcal{S}_\kappa(\beta, \eta, m, \alpha, \mathcal{I})\Big|\Big] &\le \exp_2\left(n + n(m-1)h\left(\frac{1-\beta+\eta}{2}\right) + cmn - \frac{m\alpha n}{2}\log_2(2\pi)\right.\\ &\qquad \left. + m\alpha n \log_2(2\kappa) - \frac{\alpha n}{2}\inf_{\boldsymbol{\eta}:\|\boldsymbol{\eta}\|_\infty \le \eta} \log_2\big|\Sigma(\boldsymbol{\eta})\big| + O(\log_2 n)\right)\\ &\le \exp_2\Big(n\Psi(c, m, \beta, \eta, \kappa) + O\big(\log_2 n\big)\Big),\end{aligned} \tag{3.52}$$

where

$$\begin{aligned}\Psi(c, \beta, \eta, m, \alpha) \triangleq 1 + cm + mh\left(\frac{1-\beta+\eta}{2}\right) - \frac{\alpha m}{2}\log_2(2\pi)\\ + \alpha m \log_2(2\kappa) - \frac{\alpha}{2}\inf_{\boldsymbol{\eta}:\|\boldsymbol{\eta}\|_\infty \le \eta} \log_2\big|\Sigma(\boldsymbol{\eta})\big|,\end{aligned} \tag{3.53}$$

will be called the *free energy* term.

Making the free energy negative. Recall $\alpha = 10\kappa^2 \log_2 \frac{1}{\kappa}$. We claim that for κ small, there exist $0 < \eta < \beta < 1$, $c > 0$ and $m \in \mathbb{N}$ such that $\Psi(c, \beta, \eta, m, \alpha) < 0$.

To that end, we first establish $\Psi(c, \beta, 0, m, \alpha) < 0$ for appropriately chosen $0 < \beta < 1$, $m \in \mathbb{N}$ and $c > 0$. Once this is ensured, observe that the result follows immediately: both the binary entropy and $\log_2 |\Sigma(\boldsymbol{\eta})|$ are continuous, and the domain, $\boldsymbol{\eta} \in [0, \eta]^{m(m-1)/2}$ is compact; and thus $\Psi(c, \beta, \eta, m, \alpha) < 0$ for any sufficiently small $\eta > 0$.

Let $\Sigma \triangleq \Sigma(0) = (1-\beta)I + \beta \mathbf{e}\mathbf{e}^T$. Then, the spectrum of Σ consists of the eigenvalue $1 - \beta + \beta m$ with multiplicity one; and the eigenvalue $1 - \beta$ with multiplicity $m - 1$. Consequently,

$$\begin{aligned}\Psi(c, \beta, 0, m, \alpha) &= 1 + mh\left(\frac{1-\beta}{2}\right) - \frac{\alpha m}{2}\log_2(2\pi) + \alpha m \log_2(2\kappa) + cm \\ &\quad - \frac{\alpha}{2}(m-1)\log_2(1-\beta) - \frac{\alpha}{2}\log_2(1 - \beta + \beta m) \\ &\leq m\left(\frac{1}{m} - \frac{\alpha}{2m}\log_2(1 - \beta + \beta m) + c + \Upsilon(\beta, \alpha)\right), \end{aligned} \tag{3.54}$$

where

$$\Upsilon(\beta, \alpha) \triangleq h\left(\frac{1-\beta}{2}\right) - \frac{\alpha}{2}\log_2(2\pi) + \alpha \log_2(2\kappa) - \frac{\alpha}{2}\log_2(1-\beta). \tag{3.55}$$

Set

$$\beta = 1 - 4\kappa^2, \tag{3.56}$$

and recall $\alpha = 10\kappa^2 \log_2 \frac{1}{\kappa}$. With these β and α and $\kappa > 0$ sufficiently small,

$$\frac{1}{m} - \frac{\alpha}{2m}\log_2\big(1 - \beta + \beta m\big) = o_m(1), \quad \text{as} \quad m \to \infty.$$

For this reason, it suffices to verify that for every $\kappa > 0$ sufficiently small, $\Upsilon(\beta, \alpha) < 0$.

Analyzing $\Upsilon(\beta, \alpha)$. Note that $1 - \beta = 4\kappa^2 = (2\kappa)^2$, and thus

$$\alpha \log_2(2\kappa) - \frac{\alpha}{2}\log_2(1-\beta) = 0. \tag{3.57}$$

Using the Taylor expansion, $\log_2(1-x) = -\frac{x}{\ln 2} + o(x)$ as $x \to 0$, we have

$$\log_2\big(1 - 2\kappa^2\big) = -\frac{2}{\ln 2}\kappa^2 + o_\kappa\big(\kappa^2\big),$$

as $\kappa \to 0$. Consequently,

$$h\left(\frac{1-\beta}{2}\right) = h\big(2\kappa^2\big) = -2\kappa^2 \log_2\big(2\kappa^2\big) - \big(1-2\kappa^2\big)\log_2\big(1-2\kappa^2\big)$$
$$= 4\kappa^2 \log_2 \frac{1}{\kappa} + \Theta_\kappa\big(\kappa^2\big). \tag{3.58}$$

Combining (3.57) and (3.58), we thus obtain

$$\Upsilon\big(\beta,\alpha\big) = \Big(-5\log_2(2\pi)+4\Big)\kappa^2 \log_2 \frac{1}{\kappa} + \Theta\big(\kappa^2\big), \tag{3.59}$$

which is indeed negative for every κ small.

Combining everything. We now complete the argument. For our choice of β and α, (3.59) implies that $\Upsilon(\beta,\alpha) < 0$ for every κ small. Having ensured this (for a fixed κ), we then simultaneously set $m \in \mathbb{N}$ to be sufficiently large and $c > 0$ to be sufficiently small, so that

$$\frac{1}{m} - \frac{\alpha}{2m}\log_2\big(1-\beta+\beta m\big) + c + \Upsilon(\beta,\alpha) < 0.$$

This, via (3.54), ensures $\Psi(c,\beta,0,m,\alpha) < 0$. Finally, (uniform) continuity in η ensures that for every small enough $\eta > 0$, $\Psi(c,\beta,\eta,m,\alpha) < 0$; hence

$$\mathbb{E}\Big[\Big|\mathcal{S}_\kappa(\beta,\eta,m,\alpha,\mathcal{I})\Big|\Big] = \exp\big(-\Theta(n)\big)$$

by (3.52). Finally, inserting this into (3.47), we complete the proof.

3.6.4 Proof of Theorem 3.3.2

Our proof is quite similar to that of [126, Theorem 3.2], including the aforementioned Ramsey argument. In order to guide the reader, we commence this section with an outline of the proof.

Proof Outline for Theorem 3.3.2

Fix a $\kappa > 0$, an $\alpha \geq 10\kappa^2 \log_2 \frac{1}{\kappa}$; and recall the parameters, $m \in \mathbb{N}$ and $0 < \eta < \beta < 1$, prescribed by our $m-$OGP result, Theorem 3.2.4. In a nutshell, our proof is based on a contradiction argument. To that end, assume such a stable $\mathcal{A}$ exists. Using $\mathcal{A}$; we create, with positive probability, an instance of the forbidden structure ruled out by the $m-$OGP. A brief roadmap is as follows.

- As customary, we first reduce to the case of deterministic algorithms. That is, we find an $\omega^* \in \Omega$, set $\mathcal{A}^*(\cdot) = \mathcal{A}(\cdot,\omega^*) : \mathbb{R}^{M\times n} \to \mathcal{B}_n$, and operate with this deterministic $\mathcal{A}^*$. This is the subject of Lemma 3.6.11.

- We then study a certain high-probability event, dubbed as *chaos* event. This event pertains to $m-$tuples $\sigma^{(i)} \in \mathcal{B}_n$, $1 \le i \le m$, where $\|\mathcal{M}_i \sigma^{(i)}\|_\infty \le \kappa\sqrt{n}$ for i.i.d. random matrices $\mathcal{M}_i \in \mathbb{R}^{M\times n}$, each with i.i.d. $\mathcal{N}(0,1)$ coordinates. Namely, $\sigma^{(i)}$ satisfy constraints dictated by independent instances of disorder. We show the existence of a β', such that w.h.p. it is the case that for any such $m-$tuple, there exists $1 \le i < j \le m$ such that $\mathcal{O}(\sigma^{(i)}, \sigma^{(j)}) \le \beta'$. Notably, $\beta' < \beta - \eta$. This is the subject of Lemma 3.6.12.

- We then (a) generate $T+1$ i.i.d. random matrices $\mathcal{M}_i \in \mathbb{R}^{M\times n}$, $0 \le i \le T$ (dubbed as *replicas*); (b) divide the interval $[0, \pi/2]$ into Q equal pieces, $0 = \tau_0 < \tau_1 < \cdots < \tau_Q = 1$; (c) construct *interpolation trajectories*

$$\mathcal{M}_i(\tau_k) = \cos(\tau_k)\mathcal{M}_0 + \sin(\tau_k)\mathcal{M}_i \in \mathbb{R}^{M\times n}, \quad 1 \le i \le T, \quad 0 \le k \le Q;$$

  and (d) evaluate $\mathcal{A}^*$ along each trajectory and time step, by setting

$$\sigma_i(\tau_k) \triangleq \mathcal{A}^*\big(\mathcal{M}_i(\tau_k)\big) \in \mathcal{B}_n.$$

  We tune $T, Q \in \mathbb{N}$ appropriately.

- We next show in Proposition 3.6.13 that since $\mathcal{A}^*$ is stable; the overlaps evolve smoothly along each trajectory. That is, we show that for every $1 \le i < j \le T$ and $0 \le k \le Q-1$,

$$\Big|\mathcal{O}^{(ij)}(\tau_k) - \mathcal{O}^{(ij)}(\tau_{k+1})\Big|, \qquad \text{where} \qquad \mathcal{O}^{(ij)}(\tau) \triangleq \frac{1}{n}\left\langle \sigma_i(\tau), \sigma_j(\tau)\right\rangle,$$

  is small.

- We then show, by taking a union bound over $1 \le i \le T$ and $0 \le k \le Q$; that with positive probability, the algorithm is successful along each trajectory and time step:

$$\Big\|\mathcal{M}_i(\tau_k)\sigma_i(\tau_k)\Big\|_\infty \le \kappa\sqrt{n}, \quad 1 \le i \le T, \quad 0 \le k \le Q.$$

  This is the subject of Lemma 3.6.14.

- We next take a union bound, over all subsets $A \subset [T]$ with $|A| = m$, to extend the aforementioned *chaos* event to all such subsets.

- We then let τ evolve from $\tau_0 = 0$ to $\tau_Q = \pi/2$. Notice that in the beginning, $\sigma_i(\tau_0)$ are all equal; whereas at the end, $\sigma_i(\tau_Q)$ are obtained by applying $\mathcal{A}^*$ to i.i.d. matrices, $\mathcal{M}_i(\tau_Q)$. Using this observation, the chaos property, as well as the stability of the overlaps (in the sense of above), we establish in Proposition 3.6.15 the following. For every $A \subset [T]$ with $|A| = m$, there exists $1 \le i_A < j_A \le T$ and a time $\tau_A \in \{\tau_1, \ldots, \tau_Q\}$ such that

$$\mathcal{O}^{(i_A, j_A)}\big(\tau_A\big) \in (\beta - \eta, \beta).$$

- We then construct a graph $\mathbb{G} = (V, E)$ on $|V| = T$ vertices. More specifically, (a) vertex $i \in V$ of $\mathbb{G}$ corresponds to ith interpolation trajectory; and (b) for any $1 \leq i < j \leq T$, $(i, j) \in E$ iff there is a time $t \in \{1, 2, \ldots, Q\}$ such that $\mathcal{O}^{(ij)}(\tau_t) \in (\beta - \eta, \beta)$. Note, from the previous bullet point, that the largest independent set of $\mathbb{G}$ is of size at most $m - 1$. That is $\alpha(\mathbb{G}) \leq m - 1$. We next color each edge $(i, j) \in E$ of $\mathbb{G}$ with the first time $t \in \{1, 2, \ldots, Q\}$ such that $\mathcal{O}^{(ij)}(\tau_t) \in (\beta - \eta, \beta)$.

- We next apply the Ramsey argument twice. We first use the so-called *two-color* version of Ramsey Theory, Theorem 3.6.9. Using the fact $\alpha(\mathbb{G}) \leq m - 1$; it follows that $\mathbb{G}$ contains a large clique, provided that the number T of vertices is sufficiently large. Call this large clique K_M, and observe that each edge of K_M is colored with one of Q potential colors. We then apply the so-called *multicolor* version of Ramsey Theory, Theorem 3.6.10. Provided M is large (which is ensured by our eventual choice of parameters), we deduce the original graph $\mathbb{G}$ contains a monochromatic $m-$clique K_m. These are done in Proposition 3.6.16.

- We now interpret the monochromatic K_m extracted above. There exists $1 \leq i_1 < i_2 < \cdots < i_m \leq T$ and a time $t \in \{1, 2, \ldots, Q\}$ such that $\mathcal{O}^{(i_k, i_\ell)}(\tau_t) \in (\beta - \eta, \beta)$ for $1 \leq k < \ell \leq m$. Setting $\sigma^{(i)} \triangleq \sigma_{i_k}(\tau_t) \in \mathcal{B}_n$, this $m-$tuple $\sigma^{(i)} \in \mathcal{B}_n$ is precisely the forbidden configuration ruled out by our $m-$OGP result.

- Finally, under the assumption that such an $\mathcal{A}^*$ exists; the whole process outlined above happens with positive probability. That is, we generate such an $m-$tuple with positive probability; contradicting with the $m-$OGP result, where the guarantee is exponentially small in n. This settles Theorem 3.3.2.

Before we provide the complete proof, we record the following auxiliary results.

Auxiliary Results from Ramsey Theory in Extremal Combinatorics

As was already noted, our proof uses Ramsey Theory in extremal combinatorics in a crucial way. To that end, we provide two auxiliary results. The first result pertains to the so-called two-color Ramsey numbers.

Theorem 3.6.9. *Let $k, \ell \geq 2$ be integers; and $R(k, \ell)$ denotes the smallest $n \in \mathbb{N}$ such that any red/blue (edge) coloring of K_n necessarily contains either a red K_k or a blue K_ℓ. Then,*

$$R(k, \ell) \leq \binom{k + \ell - 2}{k - 1} = \binom{k + \ell - 2}{\ell - 1}.$$

In the special case where $k = \ell = M \in \mathbb{N}$, we thus have

$$R(M, M) \leq \binom{2M - 2}{M - 1}.$$

Theorem 3.6.9 is folklore. For a proof, see e.g. [126, Theorem 6.6].

The second auxiliary result pertains to the so-called multicolor Ramsey numbers.

Theorem 3.6.10. *Let $q, m \in \mathbb{N}$. Denote by $R_q(m)$ the smallest $n \in \mathbb{N}$ such that any $q-$coloring of the edges of K_n necessarily contains a monochromatic K_m. Then,*

$$R_q(m) \leq q^{qm}.$$

Theorem 3.6.10 can be established by using a minor modification of the so-called *neighborhood chasing* argument due to Erdös and Szekeres [108], see [81, Page 6] for a more elaborate discussion.

Proof of Theorem 3.3.2

Let $\kappa > 0$ be a sufficiently small (fixed) constant, $\alpha \geq \alpha_{\text{OGP}}(\kappa) = 10\kappa^2 \log_2 \frac{1}{\kappa}$ (with $\alpha < \alpha_c(\kappa)$); and $M = \lfloor n\alpha \rfloor \in \mathbb{N}$.

We establish the hardness result for stable algorithms. That is, we show that there exists no randomized algorithm $\mathcal{A} : \mathbb{R}^{M \times n} \times \Omega \to \mathcal{B}_n$ which is

$$\big(\rho, p_f, p_{\text{st}}, f, L\big) - \text{stable}$$

(for every sufficiently large n) for the `SBP`, in the sense of Definition 3.3.1. We argue by contradiction: suppose such an $\mathcal{A}$ exists.

Parameter Choice. For the above choice of α and κ, let $m \in \mathbb{N}$, and $0 < \eta < \beta = 1 - 4\kappa^2 < 1$ be $m-$OGP parameters prescribed by Theorem 3.2.4. Observe that the $m-$OGP statement still holds with parameters $0 < \eta' < \beta < 1$ if $\eta' < \eta$. For this reason, we assume

$$\eta < \kappa^2, \qquad \beta - \eta > 1 - 5\kappa^2. \tag{3.60}$$

We first set

$$f = Cn \quad \text{where} \quad C = \frac{\eta^2}{1600}; \tag{3.61}$$

then define auxiliary parameters Q and T, where

$$Q = \frac{4800 L \pi}{\eta^2}\sqrt{\alpha} \qquad \text{and} \qquad T = \exp_2\left(2^{4mQ \log_2 Q}\right); \tag{3.62}$$

and finally prescribe p_f, p_{st}, and ρ where

$$p_f = \frac{1}{9(Q+1)T}, \qquad p_{\text{st}} = \frac{1}{9Q(T+1)}, \qquad \text{and} \qquad \rho = \cos\left(\frac{\pi}{2Q}\right). \tag{3.63}$$

Reduction to Deterministic Algorithms. We next establish that randomness do not improve the performance of a stable algorithm by much.

Lemma 3.6.11. *Let $\kappa > 0$, $\alpha < \alpha_c(\kappa)$, and $M = \lfloor n\alpha \rfloor$. Suppose that $\mathcal{A} : \mathbb{R}^{M \times n} \times \Omega \to \mathcal{B}_n$ is a randomized algorithm that is $\big(\rho, p_f, p_{\text{st}}, f, L\big)-$stable (for the `SBP`). Then, there exists a deterministic algorithm $\mathcal{A}^* : \mathbb{R}^{M \times n} \to \mathcal{B}_n$ that is $\big(\rho, 3p_f, 3p_{\text{st}}, f, L\big)-$stable*[8].

[8]Lemma 3.6.11 applies also to the deterministic algorithms, see remarks following Definition 3.3.1.

Proof of Lemma 3.6.11. For any $\omega \in \Omega$, define the event

$$\mathcal{E}_1(\omega) \triangleq \Big\{ \big| \mathcal{M}\mathcal{A}\big(\mathcal{M}, \omega\big) \big| \leq (\kappa\sqrt{n})\mathbf{e} \Big\},$$

where $\mathcal{M} \in \mathbb{R}^{M \times n}$, $\mathcal{A}(\mathcal{M}, \omega) \in \mathcal{B}_n$, and the inequality is coordinate-wise. Observe that

$$\mathbb{P}_{\mathcal{M},\omega}\big[\big| \mathcal{M}\mathcal{A}\big(\mathcal{M}, \omega\big) \big| > (\kappa\sqrt{n})\mathbf{e} \big] = \mathbb{E}_{\omega}\big[\mathbb{P}_{\mathcal{M}}\big(\mathcal{E}_1^c(\omega) \big) \big].$$

Perceiving $\mathbb{P}_{\mathcal{M}}(\mathcal{E}_1^c)$ as a random variable with source of randomness ω (note that the randomness over $\mathcal{M}$ is "integrated" over $\mathbb{P}_{\mathcal{M}}$), we have by Markov's inequality

$$\mathbb{P}_{\omega}\big[\mathbb{P}_{\mathcal{M}}\big[\mathcal{E}_1^c(\omega) \big] \geq 3p_f \big] \leq \frac{\mathbb{E}_{\omega}\big[\mathbb{P}_{\mathcal{M}}\big[\mathcal{E}_1^c(\omega) \big] \big]}{3p_f} \leq \frac{1}{3}.$$

Hence, $\mathbb{P}[\Omega_1] \geq 2/3$, where

$$\Omega_1 \triangleq \Big\{ \omega \in \Omega : \mathbb{P}_{\mathcal{M}}\big[\mathcal{E}_1^c(\omega) \big] < 3p_f \Big\}.$$

Defining next

$$\mathcal{E}_2(\omega) \triangleq \Big\{ d_H\big(\mathcal{A}(\mathcal{M}, \omega), \mathcal{A}(\overline{\mathcal{M}}, \omega) \big) \leq f + L\|\mathcal{M} - \overline{\mathcal{M}}\|_F \Big\},$$

where $\mathcal{M}, \overline{\mathcal{M}} \in \mathbb{R}^{M \times n}$ with i.i.d. standard normal coordinates subject to the constraint $\mathbb{E}\big[\mathcal{M}_{11}\overline{\mathcal{M}}_{11} \big] = \rho$. Applying the exact same logic, we find $\mathbb{P}[\Omega_2] \geq 2/3$, where

$$\Omega_2 \triangleq \Big\{ \omega \in \Omega : \mathbb{P}_{\mathcal{M},\overline{\mathcal{M}}}\big[\mathcal{E}_2^c(\omega) \big] < 3p_{\mathrm{st}} \Big\}.$$

Noting $\mathbb{P}[\Omega_1] + \mathbb{P}[\Omega_2] = 4/3 > 1$, it follows $\Omega_1 \cap \Omega_2 \neq \varnothing$. Now take any $\omega^* \in \Omega_1 \cap \Omega_2$, and set $\mathcal{A}^*(\cdot) \triangleq \mathcal{A}(\cdot, \omega^*)$. Clearly, $\mathcal{A}^*$ is $\big(\rho, 3p_f, 3p_{\mathrm{st}}, f, L\big)-$stable, establishing Lemma 3.6.11. □

In the remainder, we restrict our attention to deterministic $\mathcal{A}^* : \mathbb{R}^{M \times n} \to \mathcal{B}_n$ appearing in Lemma 3.6.11 which is $\big(\rho, 3p_f, 3p_{\mathrm{st}}, f, L\big)-$stable.

Chaos event. We now focus on the so-called *chaos* event, which pertains to $m-$tuples $\sigma^{(i)} \in \mathcal{B}_n$, $1 \leq i \leq m$, where $\big\|\mathcal{M}_i\sigma^{(i)}\big\|_{\infty} \leq \kappa\sqrt{n}$ for i.i.d. random matrices $\mathcal{M}_i \in \mathbb{R}^{M \times n}$. Namely, we investigate $m-$tuples satisfying constraints dictated by independent instances of disorder.

Lemma 3.6.12. *For every sufficiently small $\kappa > 0$ and $\alpha \geq \alpha_{\mathrm{OGP}}(\kappa) = 10\kappa^2 \log_2 \frac{1}{\kappa}$, and sufficiently large $m \in \mathbb{N}$,*

$$\mathbb{P}\Big[S_{\kappa}\big(1, 5\kappa^2, m, \alpha, \{\pi/2\}\big) \neq \varnothing \Big] \leq \exp_2\big(-\Theta(n)\big).$$

Proof of Lemma 3.6.12. The proof is quite similar to (and in fact, much simpler than) that of Theorem 3.2.4. Hence, we only provide a brief sketch. Check that for any

$\sigma \in \mathcal{B}_n$ and $X \overset{d}{=} \mathcal{N}(0, I_n)$,

$$\mathbb{P}\left[-\kappa \le n^{-\frac{1}{2}} \langle \sigma, X \rangle \le \kappa\right] \le \frac{2}{\sqrt{2\pi}}\kappa.$$

Endowed with this, a straightforward first moment argument yields

$$\begin{aligned}\mathbb{E}\Big[\big|S_\kappa\big(5\kappa^2, 1, m, \alpha, \{\pi/2\}\big)\big|\Big] &\le \exp_2\left(n\left(1 + mh\left(\frac{5\kappa^2}{2}\right) + \alpha m \log_2\left(\frac{2\kappa}{\sqrt{2\pi}}\right)\right) + O\big(\log_2 n\big)\right) \\ &= \exp_2\left(nm\left(\frac{1}{m} + h\left(\frac{5\kappa^2}{2}\right) + \alpha \log_2\left(\frac{2\kappa}{\sqrt{2\pi}}\right)\right) + O\big(\log_2 n\big)\right).\end{aligned}$$

We now apply the Taylor expansion, $\log_2(1-x) = -\frac{x}{\ln 2} + o(x)$ as $x \to 0$ to obtain

$$\begin{aligned}h\left(\frac{5\kappa^2}{2}\right) &= \frac{5\kappa^2}{2}\log_2\left(\frac{5\kappa^2}{2}\right) + \left(1 - \frac{5\kappa^2}{2}\right)\underbrace{\log_2\left(1 - \frac{5\kappa^2}{2}\right)}_{=-\Theta_\kappa\left(\kappa^2\right)} \\ &= -5\kappa^2 \log_2 \kappa + \Theta_\kappa\big(\kappa^2\big).\end{aligned}$$

Consequently,

$$h\left(\frac{5\kappa^2}{2}\right) + \alpha \log_2\left(\frac{2\kappa}{\sqrt{2\pi}}\right) = -10\kappa^2\left(\log_2\frac{1}{\kappa}\right)^2 + \Theta_\kappa\big(\kappa^2 \log_2 \kappa\big),$$

which is indeed negative for all sufficiently small $\kappa > 0$. Having ensured $\kappa > 0$ is sufficiently small, note that if $h(5\kappa^2/2) + \alpha \log_2(2\kappa/\sqrt{2\pi}) < 0$ then for every sufficiently large $m \in \mathbb{N}$,

$$\frac{1}{m} + h\left(\frac{5\kappa^2}{2}\right) + \alpha \log_2\left(\frac{2\kappa}{\sqrt{2\pi}}\right) < 0.$$

This yields the conclusion by Markov's inequality. □

In particular, for every $\kappa > 0$, $\alpha \ge 10\kappa^2 \log_2 \frac{1}{\kappa}$ and $m \in \mathbb{N}$ large enough, w.h.p. it is the case that for every $m-$tuple $\sigma^{(i)} \in \mathcal{B}_n$, $1 \le i \le m$ with $\|\mathcal{M}_i \sigma^{(i)}\|_\infty \le \kappa\sqrt{n}$ (where $\mathcal{M}_i$ are i.i.d. random matrices with i.i.d. $\mathcal{N}(0,1)$ coordinates), there exists $1 \le i < j \le m$ such that

$$\frac{1}{n}\left\langle \sigma^{(i)}, \sigma^{(j)} \right\rangle \le 1 - 5\kappa^2. \tag{3.64}$$

Construction of Interpolation Paths. Our proof uses *interpolation* ideas. To that end, let $\mathcal{M}_i \in \mathbb{R}^{M \times n}$, $0 \le i \le T$ (recall T from (3.62)), be a sequence of i.i.d. random matrices, each with i.i.d. $\mathcal{N}(0,1)$ coordinates. Recall the interpolation appearing in (3.5), repeated below for convenience:

$$\mathcal{M}_i(\tau) \triangleq \cos(\tau)\mathcal{M}_0 + \sin(\tau)\mathcal{M}_i, \quad 1 \le i \le T, \quad \tau \in [0, \pi/2]. \tag{3.65}$$

Observe that for any fixed $\tau \in [0, \pi/2]$, $\mathcal{M}_i(\tau)$ consists of i.i.d. standard normal entries.

Next for Q appearing in (3.62); we discretize $[0, \pi/2]$ into Q sub intervals—each of size $\Theta\big(Q^{-1}\big)$—where the endpoints are given by

$$0 = \tau_0 < \tau_1 < \cdots < \tau_Q = \frac{\pi}{2}. \tag{3.66}$$

We apply $\mathcal{A}^*$ to each $\mathcal{M}_i(\tau_k)$:

$$\sigma_i(\tau_k) \triangleq \mathcal{A}^*\big(\mathcal{M}_i(\tau_k)\big) \in \mathcal{B}_n, \quad 1 \le i \le T, \quad 0 \le k \le Q. \tag{3.67}$$

For every $1 \le i < j \le T$ and $0 \le k \le Q$, define pairwise overlaps

$$\mathcal{O}^{(ij)}\big(\tau_k\big) \triangleq \frac{1}{n} \left\langle \sigma_i\big(\tau_k\big), \sigma_j\big(\tau_k\big) \right\rangle. \tag{3.68}$$

A useful observation is for $k = 0$, $\sigma_1(\tau_0) = \cdots = \sigma_T(\tau_0)$; and therefore the overlaps are all unity.

Successive Steps are Stable. We now show the stability of overlaps, that is,

$$\big|\mathcal{O}^{(ij)}(\tau_k) - \mathcal{O}^{(ij)}(\tau_{k+1})\big|$$

is small for all $1 \le i < j \le T$ and $0 \le k \le Q - 1$. More concretely, we establish the following proposition.

Proposition 3.6.13.

$$\mathbb{P}\big[\mathcal{E}_{\mathrm{St}}\big] \ge 1 - 3(T+1)Qp_{\mathrm{st}} - (T+1)\exp\big(-\Theta(n^2)\big),$$

where

$$\mathcal{E}_{\mathrm{St}} \triangleq \bigcap_{1 \le i < j \le T} \bigcap_{0 \le k \le Q-1} \left\{ \left| \mathcal{O}^{(ij)}(\tau_k) - \mathcal{O}^{(ij)}(\tau_{k+1}) \right| \le 4\sqrt{C} + 4\sqrt{3L\pi Q^{-1}}\alpha^{\frac{1}{4}} \right\} \tag{3.69}$$

for C defined in (3.61).

Proof of Proposition 3.6.13. We first establish an auxiliary concentration result. Let $\mathcal{M} \in \mathbb{R}^{M \times n}$ be a random matrix with i.i.d. $\mathcal{N}(0,1)$ coordinates. Then by applying Bernstein's inequality like in the proof of [281, Theorem 3.1.1], we obtain that for some absolute constant $c > 0$ and any $t \ge 0$,

$$\mathbb{P}\left[\left| \frac{1}{Mn} \sum_{1 \le i \le M} \sum_{1 \le j \le n} \mathcal{M}_{ij}^2 - 1 \right| \ge t \right] \le \exp\Big(-cMn \min\{t, t^2\} \Big).$$

Taking a union bound and recalling $M = \lfloor n\alpha \rfloor = \Theta(n)$, we thus have

$$\mathbb{P}\Big[\|\mathcal{M}_i\|_F \le 6\sqrt{Mn}, 0 \le i \le T \Big] \ge 1 - (T+1)\exp\Big(-\Theta(n^2)\Big). \tag{3.70}$$

The constant 6 is chosen arbitrarily, and any constant greater than 1 works.

Next, we show a simple Lipschitzness property for $\cos(\cdot)$ and $\sin(\cdot)$: we claim that for every $x, y \in \mathbb{R}$,

$$\big|\cos(x) - \cos(y)\big| \le |x - y| \qquad \text{and} \qquad \big|\sin(x) - \sin(y)\big| \le |x - y|.$$

Indeed, by the mean value theorem, for $x < y$, it holds that for some $c \in (x, y)$; $|\cos(x) - \cos(y)| = |x - y| \cdot |\sin(c)| \le |x - y|$. The result for the $\sin(\cdot)$ is analogous. Consequently,

$$\max\Big\{\big|\cos(\tau_k) - \cos(\tau_{k+1})\big|, \big|\sin(\tau_k) - \sin(\tau_{k+1})\big|\Big\} \le |\tau_k - \tau_{k+1}| = \frac{\pi}{2Q}, \tag{3.71}$$

where we used (3.66) at the last step.

We now employ (3.71) to upper bound $\big\|\mathcal{M}_i(\tau_k) - \mathcal{M}_i(\tau_{k+1})\big\|_F$ on the high probability event appearing in (3.70). Assuming (3.70) takes place, we have that for any fixed $1 \le i \le T$ and $0 \le k \le Q - 1$,

$$\begin{aligned}
\Big\|\mathcal{M}_i(\tau_k) - \mathcal{M}_i(\tau_{k+1})\Big\|_F &= \Big\|\cos(\tau_k)\mathcal{M}_0 + \sin(\tau_k)\mathcal{M}_i - \cos(\tau_{k+1})\mathcal{M}_0 - \sin(\tau_{k+1})\mathcal{M}_i\Big\|_F \\
&\le \big|\cos(\tau_k) - \cos(\tau_{k+1})\big|\|\mathcal{M}_0\|_F + \big|\sin(\tau_k) - \sin(\tau_{k+1})\big|\|\mathcal{M}_i\|_F \qquad (3.72) \\
&\le \frac{\pi}{2Q}\big(\|M_0\|_F + \|M_i\|_F\big) \qquad (3.73) \\
&\le \frac{3\pi}{Q}\sqrt{Mn}; \qquad (3.74)
\end{aligned}$$

where (3.72) uses triangle inequality for the Frobenius norm; (3.73) uses (3.71); and finally (3.74) uses the fact that on the event (3.70), $\|\mathcal{M}_i\|_F \le 6\sqrt{Mn}$ for $0 \le i \le T$.

We next observe that for any fixed $1 \le i \le T$ and $0 \le k \le Q - 1$; each of $\mathcal{M}_i(\tau_k)$ and $\mathcal{M}_i(\tau_{k+1})$ has i.i.d. $\mathcal{N}(0, 1)$ entries subject to

$$\mathbb{E}\Big[\big(\mathcal{M}_i(\tau_k)\big)_{\ell,j}\big(\mathcal{M}_i(\tau_{k+1})\big)_{\ell,j}\Big] = \cos\big(\tau_{k+1} - \tau_k\big) = \cos\left(\frac{\pi}{2Q}\right). \tag{3.75}$$

for $1 \le \ell \le M$ and $1 \le j \le n$. Recall now by Lemma 3.6.11 that $\mathcal{A}^*$ is stable with stability probability $1 - 3p_{\rm st}$. Taking thus a union bound, we find

$$\begin{aligned}
&\mathbb{P}\Big[d_H\Big(\sigma_i(\tau_k), \sigma_i(\tau_{k+1})\Big) \le Cn + L\big\|\mathcal{M}_i(\tau_k) - \mathcal{M}_i(\tau_{k+1})\big\|_F, 1 \le i \le T, 0 \le k \le Q - 1\Big] \\
&\ge 1 - 3(T + 1)Qp_{\rm st}. \qquad (3.76)
\end{aligned}$$

We now combine (3.74) (valid on event (3.70)) and the event (3.76) by a union bound. We find that

$$\mathbb{P}[\mathcal{E}] \ge 1 - 3(T + 1)Qp_{\rm st} - (T + 1)\exp\big(-\Theta(n^2)\big), \tag{3.77}$$

where

$$\mathcal{E} \triangleq \bigcap_{1\le i\le T}\bigcap_{0\le k\le Q-1}\left\{d_H\big(\sigma_i(\tau_k),\sigma_i(\tau_{k+1})\big) \le Cn + \frac{3L\pi}{Q}\sqrt{Mn}\right\}. \tag{3.78}$$

In the remainder of the proof, assume we operate on the event $\mathcal{E}$ (3.78).

Observe that for any $\sigma,\sigma' \in \mathcal{B}_n$, $\|\sigma-\sigma'\|_2 = 2\sqrt{d_H(\sigma,\sigma')}$; and recall from (3.67) the notation, $\sigma_i(\tau_k)$. We have that for any $1 \le i \le T$ and $0 \le k \le Q-1$,

$$\begin{aligned}\Big\|\sigma_i(\tau_k)-\sigma_i(\tau_{k+1})\Big\|_2 &= 2\sqrt{d_H\big(\sigma_i(\tau_k),\sigma_k(\tau_{k+1})\big)}\\ &\le 2\sqrt{Cn+3L\pi Q^{-1}\sqrt{Mn}} &(3.79)\\ &\le \sqrt{n}\Big(2\sqrt{C}+2\sqrt{3L\pi Q^{-1}}\alpha^{\frac14}\Big); &(3.80)\end{aligned}$$

where (3.79) follows from the fact we are on event $\mathcal{E}$ (3.78); and (3.80) uses the fact $M \le n\alpha$ and the trivial inequality $\sqrt{u+v} \le \sqrt{u}+\sqrt{v}$ valid for all $u,v \ge 0$.

Equipped with (3.80), we are now in a position to conclude. Fix any $1 \le i < j \le T$ and $0 \le k \le Q-1$. We have the following chain of inequalities:

$$\begin{aligned}\Big|\mathcal{O}^{(ij)}(\tau_k)-\mathcal{O}^{(ij)}(\tau_{k+1})\Big| &= \frac1n\Big|\langle\sigma_i(\tau_k),\sigma_j(\tau_k)\rangle-\langle\sigma_i(\tau_{k+1}),\sigma_j(\tau_{k+1})\rangle\Big|\\ &\le \frac1n\Big(\Big|\langle\sigma_i(\tau_k)-\sigma_i(\tau_{k+1}),\sigma_j(\tau_k)\rangle\Big|+\Big|\langle\sigma_i(\tau_{k+1}),\sigma_j(\tau_k)-\sigma_j(\tau_{k+1})\rangle\Big|\Big) &(3.81)\\ &\le \frac{1}{\sqrt{n}}\Big(\Big\|\sigma_i(\tau_k)-\sigma_i(\tau_{k+1})\Big\|_2+\Big\|\sigma_j(\tau_k)-\sigma_j(\tau_{k+1})\Big\|_2\Big) &(3.82)\\ &\le 4\sqrt{C}+4\sqrt{3L\pi Q^{-1}}\alpha^{\frac14}. &(3.83)\end{aligned}$$

Indeed, (3.81) follows from the triangle inequality; (3.82) uses Cauchy-Schwarz inequality with the fact $\|\sigma\|_2 = \sqrt{n}$ for any $\sigma \in \mathcal{B}_n$; and (3.83) uses (3.80). Recalling the probability bound (3.77) on the event $\mathcal{E}$ that we operated under, the proof of Proposition 3.6.13 is complete. □

$\mathcal{A}^*$ is Successful along Each Trajectory. We next study the event that $\mathcal{A}^*$ is *successful* along each interpolation trajectory and across times. We have

Lemma 3.6.14.

$$\mathbb{P}\big[\mathcal{E}_{\text{Suc}}\big] \ge 1-3T(Q+1)p_f,$$

where

$$\mathcal{E}_{\text{Suc}} \triangleq \bigcap_{1\le i\le T}\bigcap_{0\le k\le Q}\Big\{\big\|\mathcal{M}_i(\tau_k)\sigma_i(\tau_k)\big\|_\infty \le \kappa\sqrt{n}\Big\}. \tag{3.84}$$

Proof of Lemma 3.6.14. The results follows immediately by (a) recalling, from Lemma 3.6.11, that

$$\mathbb{P}_{\mathcal{M}}\Big[\big\|\mathcal{M}\mathcal{A}^*(\mathcal{M})\big\|_\infty \le \kappa\sqrt{n}\Big] \ge 3p_f,$$

(b) observing $\mathcal{M}_i(\tau_k) \stackrel{d}{=} \mathcal{M}_0$ for all i and k; and (c) taking a union bound over $1 \le i \le T$ and $0 \le k \le Q$.

Combining Everything. Fix any subset $A \subset [T]$ with $|A| = m$, and let $\mathcal{E}_A$ be

$$\mathcal{E}_A \triangleq \left\{ \exists\big(\sigma^{(i)} \in \mathcal{B}_n, i \in A\big) : \max_{i \in A} \big\|\mathcal{M}_i(1)\sigma^{(i)}\big\|_\infty \le \kappa\sqrt{n},\ \beta - \eta \le n^{-1}\left\langle \sigma^{(i)}, \sigma^{(j)} \right\rangle \le \beta, i, j \in A, i \ne j \right\}.$$

Namely, $\mathcal{E}_A$ is nothing but the chaos event in the sense of Lemma 3.6.12, where the indices are restricted to $A \subset [T]$. In particular, $\mathbb{P}[\mathcal{E}_A] \ge \exp(-\Theta(n))$ due to Lemma 3.6.12. Taking a union bound over $A \subset [T]$, we obtain

$$\mathbb{P}[\mathcal{E}_{\rm Ch}] \triangleq \mathbb{P}\left[\bigcap_{A \subset [T], |A| = m} \mathcal{E}_A^c\right] \ge 1 - \binom{T}{m} e^{-\Theta(n)} = 1 - \exp\big(-\Theta(n)\big), \tag{3.85}$$

where we used the fact $\binom{T}{m} = O(1)$ (as $n \to \infty$) since $T = O(1)$ per (3.62) and $m = O(1)$. Let

$$\mathcal{F} \triangleq \mathcal{E}_{\rm St} \cap \mathcal{E}_{\rm Suc} \cap \mathcal{E}_{\rm Ch}, \tag{3.86}$$

where $\mathcal{E}_{\rm St}$, $\mathcal{E}_{\rm Suc}$, and $\mathcal{E}_{\rm Ch}$ are defined, respectively, in (3.69), (3.84), and (3.85). We then have

$$\begin{aligned}
\mathbb{P}[\mathcal{F}] &\ge 1 - \mathbb{P}\big[\mathcal{E}_{\rm St}^c\big] - \mathbb{P}\big[\mathcal{E}_{\rm Suc}^c\big] - \mathbb{P}\big[\mathcal{E}_{\rm Ch}^c\big] && (3.87)\\
&\ge 1 - 3(T+1)Qp_{\rm st} - (T+1)e^{-\Theta(n^2)} - 3T(Q+1)p_f - e^{-\Theta(n)} && (3.88)\\
&\ge \frac{1}{3} - \exp\big(-\Theta(n)\big), && (3.89)
\end{aligned}$$

where (3.87) follows from a union bound; (3.88) uses Proposition 3.6.13, Lemma 3.6.14 and (3.85); and (3.89) recalls (3.63) for p_f and $p_{\rm st}$. We operate on the event $\mathcal{F}$ in the remainder of the proof.

Now, inserting into (3.69) the choice of C per (3.61) and Q per (3.62); it is the case that on $\mathcal{F}$,

$$\left|\mathcal{O}^{(ij)}(\tau_k) - \mathcal{O}^{(ij)}(\tau_{k+1})\right| \le \frac{\eta}{5} \tag{3.90}$$

for every $1 \le i < j \le T$ and $0 \le k \le Q-1$. Fix next any $A \subset [T]$ with $|A| = m$. We establish the following proposition.

Proposition 3.6.15. *For every $A \subset [T]$ with $|A| = m$, there exists $1 \le i_A < j_A \le m$ and $\tau_A \in \{\tau_1, \ldots, \tau_Q\}$ such that for $\delta = \frac{\eta}{100}$,*

$$\mathcal{O}^{(i_A, j_A)}(\tau_A) \in \big(\beta - \eta + 3\delta, \beta - 3\delta\big) \subsetneq (\beta - \eta, \beta).$$

Proof of Proposition 3.6.15. A consequence of $\mathcal{E}_{\rm Ch}$ (part of $\mathcal{F}$) is that there exists distinct $i_A, j_A \in A$ such that

$$\mathcal{O}^{(i_A, j_A)}(\tau_Q) \le 1 - 5\kappa^2,$$

where we utilized (3.64). Recall now the choice of $\beta = 1 - 4\kappa^2$ and η such that $\beta - \eta > 1 - 5\kappa^2$. In particular, $\mathcal{O}^{(i_A,j_A)}(\tau_Q) < \beta - \eta$. We now claim for $\delta = \eta/100$, there exists a $k' \in \{1, 2, \ldots, Q\}$ such that

$$\mathcal{O}^{(i_A,j_A)}(\tau_{k'}) \in \big(\beta - \eta + 3\delta, \beta - 3\delta\big).$$

To that end, take $K_0 \in \{1, 2, \ldots, Q\}$ to be the *last time* such that $\mathcal{O}^{(i_A,j_A)}(\tau_{K_0}) \geq \beta - 3\delta$. Note that such a K_0 must exist as $\mathcal{O}^{(ij)}(0) = 1$ for every $1 \leq i < j \leq T$. Then if $\mathcal{O}^{(i_A,j_A)}(\tau_{K_0+1}) \leq \beta - \eta + 3\delta$, we obtain

$$\Big|\mathcal{O}^{(i_A,j_A)}(\tau_{K_0}) - \mathcal{O}^{(i_A,j_A)}(\tau_{K_0+1})\Big| \geq \eta - 6\delta,$$

contradicting (3.90) for sufficiently large n. That is,

$$\mathcal{O}^{(i_A,j_A)}(\tau_{K_0+1}) \in (\beta - \eta + 3\delta, \beta - 3\delta).$$

Since $A \subset [T]$ was arbitrary, Proposition 3.6.15 is established. □

Constructing an Appropriate Graph, and Applying Ramsey Theory. We now construct a certain graph $\mathbb{G} = (V, E)$ satisfying the following properties.

- The vertex set V coincides with $[T]$. That is, $V = \{1, 2, \ldots, T\}$, where each vertex i corresponds to the interpolation trajectory i, $1 \leq i \leq T$.
- For any $1 \leq i < j \leq T$, we add $(i, j) \in E$ iff there exists a time $\tau \in [0, 1]$ such that $\mathcal{O}^{(ij)}(\tau) \in (\beta - \eta, \beta)$.

Namely, $\mathbb{G}$ is a graph with a potentially large number of vertices, and a certain number of edges.

We next *color* each $(i, j) \in E$ with one of Q colors. Specifically, for any $1 \leq i < j \leq T$ with $(i, j) \in E$; we color the edge $(i, j) \in E$ with color t, $1 \leq t \leq Q$, where $\tau_t \in \{\tau_1, \ldots, \tau_Q\}$ is the first time such that

$$\mathcal{O}^{(ij)}(\tau_t) \in (\beta - \eta, \beta).$$

Having done this coloring, $\mathbb{G} = (V, E)$ satisfies the following properties:e

(a) We have $|V| = T$; and for every $A \subset V$ with $|A| = m$, there exists distinct $i_A, j_A \in A$ such that $(i_A, j_A) \in E$. Namely, $\mathbb{G}$ contains no independent sets of size larger than $m - 1$.

(b) Any $(i, j) \in E$ is colored with one of colors $\{1, 2, \ldots, Q\}$.

We claim

Proposition 3.6.16. *$\mathbb{G} = (V, E)$ defined above contains a monochromatic $m-$clique, K_m.*

Proof of Proposition 3.6.16. Recall from (3.62) that $\mathbb{G}$ contains $T = \exp_2\big(\exp_2\big(4mQ\log_2 Q\big)\big)$ vertices. Set

$$M \triangleq Q^{mQ} = 2^{mQ\log_2 Q}. \tag{3.91}$$

Extracting a Large Clique K_M. Recall from Theorem 3.6.9 that

$$R_2(M,M) \le \binom{2M-2}{M-1}.$$

As a result, any graph with at least $\binom{2M-2}{M-1}$ vertices contains either an independent set of cardinality M, or an $M-$clique, K_M. Now, from property (a) above, the largest independent set of $\mathbb{G}$ is of size at most $m-1$, which is less than M. Since

$$T = \exp_2\left(2^{4mQ\log_2 Q}\right) \ge 2^{2M} = 4^M \ge \binom{2M-2}{M-1}$$

for M defined in (3.91), it follows that $\mathbb{G}$ contains a K_M, where $M = Q^{Qm}$, each of whose edges is colored with one of Q colors.

Further Extracting a Monochromatic K_m. Now that we extracted a K_M with $M = Q^{Qm}$. Since $R_Q(m) \le Q^{Qm}$ per Theorem 3.6.10; we obtain, by applying the multicolor version of Ramsey Theory, that K_M contains a monochromatic K_m. Since K_M is a subgraph of $\mathbb{G}$, this establishes Proposition 3.6.16. □

We now finalize the proof of Theorem 3.3.2. We interpret K_m of $\mathbb{G}$ extracted in Proposition 3.6.16: there exists an $m-$tuple, $1 \le i_1 < i_2 < \cdots < i_m \le T$ and a color $t \in \{1,2,\ldots,Q\}$ such that

$$\mathcal{O}^{(i_k,i_\ell)}(\tau_t) \in (\beta-\eta,\beta), 1 \le k < \ell \le m.$$

Now, set $\sigma^{(k)} \triangleq \mathcal{A}^*\big(\mathcal{M}_{i_k}(\tau_t)\big) \in \mathcal{B}_n$, $1 \le k \le m$. Observe the following for this $m-$tuple:

- Noting we are on $\mathcal{F}$ (3.86), and in particular $\mathcal{F} \subset \mathcal{E}_{\text{Suc}}$ defined in (3.84); we have
$$\left\|\mathcal{M}_{i_k}(\tau_t)\sigma^{(k)}\right\|_\infty \le \kappa\sqrt{n}, \quad 1 \le k \le m.$$
- For $1 \le k < \ell \le m$,
$$\beta - \eta < \frac{1}{n}\left\langle \sigma^{(k)}, \sigma^{(\ell)}\right\rangle < \beta.$$

In particular, for the choice $\zeta = \{i_1, i_2, \ldots, i_m\}$ of the $m-$tuple of distinct indices, the set $\mathcal{S}_\zeta \triangleq \mathcal{S}_\kappa\big(\beta,\eta,m,\alpha,\mathcal{I}\big)$ with $\mathcal{I} = \{\tau_i : 0 \le i \le Q\}$ is non-empty. That is,

$$\mathbb{P}\Big[\exists \zeta \in [T] : |\zeta| = m, \mathcal{S}_\zeta \ne \varnothing\Big] \ge \mathbb{P}[\mathcal{F}] \ge \frac{1}{3} - \exp\big(-\Theta(n)\big).$$

Notice, on the other hand, that using the $m-$OGP result, Theorem 3.2.4, we have

$$\mathbb{P}\Big[\exists \zeta \in [T] : |\zeta| = m, \mathcal{S}_\zeta \neq \varnothing\Big] \leq \binom{T}{m} e^{-\Theta(n)} = \exp\big(-\Theta(n)\big),$$

by taking a union bound and recalling $\binom{T}{m} = O(1)$. Combining these, we therefore obtain

$$\exp\big(-\Theta(n)\big) \geq \frac{1}{3} - \exp\big(-\Theta(n)\big),$$

which is clearly a contradiction for all n large enough, establishing the result. □

3.6.5 Proof of Theorem 3.3.4

We first provide an auxiliary result.

Proposition 3.6.17. *Fix $\Delta \in (0, \frac{1}{2})$. Let $\mathcal{M} \in \mathbb{R}^{M \times n}$ be a matrix with i.i.d. $\mathcal{N}(0,1)$ coordinates; and let $\mathcal{M}_\Delta \in \mathbb{R}^{M \times n}$ be the matrix obtained from $\mathcal{M}$ by resampling its last $\Delta \cdot n$ columns independently from $\mathcal{N}(0,1)$. Let $\Xi(\Delta) \subset \mathcal{B}_n \times \mathcal{B}_n$ be the set of all $(\sigma, \sigma_\Delta) \in \mathcal{B}_n \times \mathcal{B}_n$ satisfying the following conditions.*

- $\big\|\mathcal{M}\sigma\big\|_\infty \leq \sqrt{n}$ *and* $\big\|\mathcal{M}_\Delta \sigma_\Delta\big\|_\infty \leq \sqrt{n}$.
- $n^{-1} \langle \sigma, \sigma_\Delta \rangle \in [1 - 2\Delta, 1]$.

Then, there is a $\Delta > 0$ such that

$$\mathbb{P}\big[\Xi(\Delta) = \varnothing\big] \geq 1 - \exp\big(-\Theta(n)\big).$$

Assuming Proposition 3.6.17, we now show how to establish Theorem 3.3.4. Suppose such an $\mathcal{A}$ that is p_f-online for $p_f < \frac{1}{2} - \exp(-c_f n)$ exists. Let $\mathcal{M} \in \mathbb{R}^{M \times n}$ with i.i.d. $\mathcal{N}(0,1)$ entries, $\mathcal{M}_\Delta \in \mathbb{R}^{M \times n}$ be the matrix obtained from $\mathcal{M}$ by independently resampling its last Δn columns; and set

$$\sigma \triangleq \mathcal{A}\big(\mathcal{M}\big) \in \mathcal{B}_n \qquad \text{and} \qquad \sigma_\Delta \triangleq \mathcal{A}\big(\mathcal{M}_\Delta\big) \in \mathcal{B}_n.$$

By a union bound, it is the case that w.p. at least $1 - 2p_f$, $\big\|\mathcal{M}\sigma\big\|_\infty \leq \sqrt{n}$ and $\big\|\mathcal{M}_\Delta \sigma_\Delta\big\|_\infty \leq \sqrt{n}$. Moreover, since the algorithm is online per Definition 3.3.3, it follows that $\sigma(i) = \sigma_\Delta(i)$ for $1 \leq i \leq n - \Delta n$. Hence,

$$\frac{1}{n} \langle \sigma, \sigma_\Delta \rangle = \frac{1}{n} \sum_{1 \leq i \leq n - \Delta n} \sigma(i)\sigma_\Delta(i) + \frac{1}{n} \sum_{n - \Delta n + 1 \leq i \leq n} \sigma(i)\sigma_\Delta(i) \geq 1 - 2\Delta.$$

As $(\sigma, \sigma_\Delta) \in \Xi(\Delta)$, we have $\mathbb{P}\big[\Xi(\Delta) \neq \varnothing\big] \geq 1 - 2p_f \geq 2\exp(-c_f n)$. On the other hand, $\mathbb{P}\big[\Xi(\Delta) \neq \varnothing\big] \leq \exp\big(-\Theta(n)\big)$. This is a clear contradiction if $c_f > 0$ is small enough. Therefore, it suffices to prove Proposition 3.6.17.

Proof of Proposition 3.6.17. The proof is similar to that of 2−OGP result, Theorem 3.2.2; and is based, in particular, on the first moment method. Let

$$N = \sum_{\sigma,\sigma_\Delta : n^{-1}\langle \sigma,\sigma_\Delta\rangle \in [1-2\Delta,1]} \mathbb{1}\left\{ \left\|\mathcal{M}\sigma\right\|_\infty \le \sqrt{n}, \left\|\mathcal{M}_\Delta\sigma_\Delta\right\|_\infty \le \sqrt{n}\right\}$$

Clearly, $N = |\Xi(\Delta)|$. By Markov's inequality,

$$\mathbb{P}\big[\Xi(\Delta) \neq \varnothing\big] = \mathbb{P}\big[N \ge 1\big] \le \mathbb{E}[N]. \tag{3.92}$$

Thus, it suffices to show $\mathbb{E}[N] = \exp(-\Theta(n))$ for $\Delta > 0$ small.

Counting term. There are 2^n choices for $\sigma \in \mathcal{B}_n$. Note that $n^{-1}\langle \sigma, \sigma_\Delta\rangle \in [1-2\Delta, 1] \iff d_H\big(\sigma,\sigma_\Delta\big) \le \Delta n$. Thus, having fixed a σ, there are

$$\sum_{k\in\mathbb{N}\cap[0,\Delta n]} \binom{n}{k} \le (1+\Delta n)\cdot \binom{n}{\Delta n} = \exp_2\Big(nh(\Delta) + O(\log_2 n)\Big)$$

choices for $\sigma_\Delta \in \mathcal{B}_n$, where we used the fact $\binom{n}{k} \le \binom{n}{\Delta n}$ for any $k \le \Delta n$ (as $\Delta < 1/2$) and Stirling's approximation. Thus,

$$\left|\left\{(\sigma,\sigma_\Delta) \in \mathcal{B}_n\times\mathcal{B}_n : n^{-1}\langle\sigma,\sigma_\Delta\rangle \ge 1-2\Delta\right\}\right| \le \exp_2\Big(n+nh(\Delta)+O(\log_2 n)\Big). \tag{3.93}$$

Probability term. Now, fix σ, σ_Δ with $n^{-1}\langle\sigma,\sigma_\Delta\rangle \ge 1-2\Delta$. Let $R = (Z_1, Z_2, \ldots, Z_n) \in \mathbb{R}^n$ and $R_\Delta = (Z_1, Z_2, \ldots, Z_{n-\Delta n}, Z'_{n-\Delta n+1}, \cdots, Z'_n) \in \mathbb{R}^n$ respectively be the first rows of $\mathcal{M}$ and $\mathcal{M}_\Delta$, where $Z_1, \ldots, Z_n, Z'_{n-\Delta n+1}, \ldots, Z'_n$ are i.i.d. standard normal. Using the independence of rows of $\mathcal{M}$ and $\mathcal{M}_\Delta$, we have

$$\mathbb{P}\Big[\left\|\mathcal{M}\sigma\right\|_\infty \le \sqrt{n}, \left\|\mathcal{M}_\Delta\sigma_\Delta\right\|_\infty \le \sqrt{n}\Big] = \mathbb{P}\Big[n^{-\frac12}\big|\langle R,\sigma\rangle\big| \le 1, n^{-\frac12}\big|\langle R_\Delta,\sigma_\Delta\rangle\big| \le 1\Big]^{\alpha n}.$$

We next study bivariate normal variables $n^{-\frac12}\langle R,\sigma\rangle \overset{d}{=} \mathcal{N}(0,1)$ and $n^{-\frac12}\langle R_\Delta,\sigma_\Delta\rangle \overset{d}{=} \mathcal{N}(0,1)$. Note that

$$\begin{aligned}\frac1n\mathbb{E}\Big[\langle R,\sigma\rangle\langle R_\Delta,\sigma_\Delta\rangle\Big] &= \frac1n\sum_{1\le i\le n-\Delta n}\mathbb{E}\big[Z_i^2\sigma(i)\sigma_\Delta(i)\big] + \frac1n\sum_{n-\Delta n+1\le i\le n}\underbrace{\mathbb{E}\big[Z_iZ'_i\sigma(i)\sigma_\Delta(i)\big]}_{=0}\\ &= \frac1n\sum_{1\le i\le n-\Delta n}\sigma(i)\sigma_\Delta(i) \in [1-2\Delta, 1-\Delta]\end{aligned}$$

since $d_H\big(\sigma,\sigma_\Delta\big) \le \Delta n$. Letting $\lambda \triangleq n^{-1}\mathbb{E}\Big[\langle R,\sigma\rangle\langle R_\Delta,\sigma_\Delta\rangle\Big]/(1-\Delta) \in \big[\frac{1-2\Delta}{1-\Delta}, 1\big]$, we therefore obtain that $n^{-\frac12}\langle R,\sigma\rangle, n^{-\frac12}\langle R_\Delta,\sigma_\Delta\rangle$ is bivariate normal with parameter $\lambda(1-\Delta)$. Let $p(\rho) \triangleq \mathbb{P}\big[(Z_1,Z_2)\in[-1,1]^2\big]$ where $\rho\in[0,1]$ and (Z_1,Z_2) bivariate normal with parameter ρ. Using Theorem 3.6.5 with $k=2$ and $\lambda_1=\lambda_2=\sqrt{\lambda}$, we

thus obtain $\max_{0\le\lambda\le 1} p\big(\lambda(1-\Delta)\big) = p\big(1-\Delta\big)$. Hence,

$$\mathbb{P}\Big[\big\|\mathcal{M}\sigma\big\|_\infty \le \sqrt{n}, \big\|\mathcal{M}_\Delta\sigma_\Delta\big\|_\infty \le \sqrt{n}\Big] \le p\big(1-\Delta\big)^{\alpha n}. \tag{3.94}$$

Upper bounding $\mathbb{E}[N]$. Combining (3.92), (3.93) and (3.94), we obtain

$$\begin{aligned}\mathbb{P}\big[\Xi(\Delta)\neq\varnothing\big] \le \mathbb{E}[N] &\le \exp_2\Big(n\Big(1+h(\Delta)+\alpha\log_2 p\big(1-\Delta\big)\Big)+O(\log_2 n)\Big)\\ &\le \exp_2\big(nf_1(\Delta,\alpha)+O(\log_2 n)\big),\end{aligned}$$

where $f_1(\Delta,\alpha) = 1+h(\Delta)+\alpha\log_2 p(1-\Delta)$, per Lemma 3.6.1. Since $\log_2 p(1-\Delta) < 0$, it suffices to consider $\alpha = 1.77$. Setting Δ such that

$$f_1(\Delta, 1.77) = \inf_{x\in[10^{-5},10^{-1}]} f_1(x, 1.77),$$

Lemma 3.6.1 (a) implies $f_1(\Delta, 1.77) < 0$. With this choice of Δ, we complete the proof of Proposition 3.6.17. □

3.6.6 Proof of Theorem 3.3.8

In this section, we establish Theorem 3.3.8. That is, we show that $\mathcal{A}_{\mathrm{KR}}$ is stable in the probabilistic sense. We first set the stage. Recall from (3.14) the interpolation

$$\overline{\mathcal{M}}(\tau) \triangleq \cos(\tau)\mathcal{M} + \sin(\tau)\mathcal{M}' \in \mathbb{R}^{k\times n}, \qquad \tau\in\left[0,\frac{\pi}{2}\right],$$

where $\mathcal{M},\mathcal{M}'\in\mathbb{R}^{k\times n}$ are two i.i.d. random matrices each with i.i.d. $\mathcal{N}(0,1)$ entries. In particular, $\overline{\mathcal{M}}(\tau)$ has i.i.d. $\mathcal{N}(0,1)$ coordinates for each $\tau\in[0,\pi/2]$.

Next, denote by $R_1,\dots,R_k\in\mathbb{R}^n$ the rows of $\mathcal{M}$; and by $C_1,\dots,C_n\in\mathbb{R}^k$ the columns of $\mathcal{M}$. Likewise, let $\overline{R}_1,\dots,\overline{R}_k\in\mathbb{R}^n$ and $\overline{C}_1,\dots,\overline{C}_n\in\mathbb{R}^k$ be the rows and columns of $\overline{\mathcal{M}}(\tau)$, respectively. (Whenever appropriate, we drop τ for convenience.) As in Theorem 3.3.8, set

$$\sigma = \mathcal{A}_{\mathrm{KR}}\big(\mathcal{M}\big)\in\mathcal{B}_n \qquad \text{and} \qquad \overline{\sigma} = \mathcal{A}_{\mathrm{KR}}\Big(\overline{\mathcal{M}}(\tau)\Big)\in\mathcal{B}_n.$$

We first establish the following proposition which pertains the L round implementation of Kim-Roche algorithm, where $L\le c\log_{10}\log_{10} n$ for $c>0$ sufficiently small (as opposed to its full implementation).

Proposition 3.6.18. *Let $c>0$ be a sufficiently small constant, and $L\le c\log_{10}\log_{10} n$ be an arbitrary non-negative integer. Define*

$$\alpha_\ell = \alpha_0\cdot 10^{-\ell}, \qquad 1\le\ell\le L \quad \textit{with} \quad \alpha_0 = 0.01; \tag{3.95}$$

and set $\tau = n^{-2\alpha_0}$. Let $\sigma\in\{-1,1\}^{\sum_{0\le\ell\le L} n_\ell}$ and $\overline{\sigma}\in\{-1,1\}^{\sum_{0\le\ell\le L} n_\ell}$ respectively be the outputs generated by running L rounds of Kim-Roche algorithm on $\mathcal{M}$ and $\overline{\mathcal{M}}(\tau)$

defined in (3.14). *Define*

$$J_\ell \triangleq \Big\{ i \in [n_0 + n_1 + \cdots + n_\ell] : \sigma_i \neq \overline{\sigma}_i \Big\}, \quad 0 \leq \ell \leq L. \tag{3.96}$$

Then, for any $0 \leq \ell \leq L$,

$$\mathbb{P}\Big[\big| J_\ell \big| \leq n^{1-\alpha_\ell} \Big] \geq 1 - O\left(n^{-\frac{1}{40}+\epsilon} \right),$$

where $\epsilon > 0$ *is arbitrary.*

Proof of Proposition 3.6.18

This section is devoted to the proof of Proposition 3.6.18. We proceed by establishing several auxiliary results.

Majority is stable. As a first step, we establish the stability of the majority algorithm. This algorithm assigns, to each column $C_j \in \mathbb{R}^k$ of $\mathcal{M}$, the sign of entries in C_j. That is,

$$\sigma_j = \operatorname{sgn}\left(\sum_{1 \leq i \leq k} \mathcal{M}_{ij} \right) \in \{-1, 1\}.$$

Namely, this algorithm is simply the very first round of $\mathcal{A}_{\mathrm{KR}}$ assigning $n_0 \leq n$ entries of $\sigma \in \mathcal{B}_n$, where $n_0 \approx n$.

Lemma 3.6.19. *Let* $\mathcal{A}_{\mathrm{maj}} : \mathbb{R}^{k \times n} \to \mathcal{B}_n$ *be the "majority" algorithm defined above. Recall* $\overline{\mathcal{M}}(\tau)$ *from* (3.14). *Then,*

$$d_H\Big(\mathcal{A}_{\mathrm{maj}}(\mathcal{M}), \mathcal{A}_{\mathrm{maj}}\big(\overline{\mathcal{M}}(\tau)\big) \Big) \stackrel{d}{=} \mathrm{Bin}\Big(n, \frac{\tau}{\pi} \Big).$$

Proof of Lemma 3.6.19. Define I_j, $1 \leq j \leq n$ by

$$I_j = \mathbb{1}\Big\{ \mathcal{A}_{\mathrm{maj}}(\mathcal{M})_j \neq \mathcal{A}_{\mathrm{maj}}\big(\overline{\mathcal{M}}(\tau)\big)_j \Big\}.$$

Then, I_j are i.i.d. Bernoulli. In particular, it suffices to show $I_j \sim \mathrm{Ber}(\tau/\pi)$. To that end, we study I_1. Let $(X_1, \ldots, X_k)$ be the first column of $\mathcal{M}$ and $(Y_1, \ldots, Y_k)$ be the first column of $\mathcal{M}'$. Furthermore, set

$$Z_i \triangleq \cos(\tau) X_i + \sin(\tau) Y_i, \quad 1 \leq i \leq k.$$

Note that, $I_i = 1$ if and only if

$$\operatorname{sgn}\left(\sum_{1 \leq i \leq k} X_i \right) \neq \operatorname{sgn}\left(\sum_{1 \leq i \leq k} Z_i \right).$$

From symmetry,

$$\mathbb{P}\left(\operatorname{sgn}\left(\sum_{1\le i\le k} X_i\right) \neq \operatorname{sgn}\left(\sum_{1\le i\le k} Z_i\right)\right) = 2\mathbb{P}\left(k^{-\frac{1}{2}}\sum_{1\le i\le k} X_i > 0, -k^{-\frac{1}{2}}\sum_{1\le i\le k} Z_i > 0\right). \tag{3.97}$$

Observe that $\mathbb{E}[X_i Z_j] = \cos(\tau)\mathbb{1}\{i=j\}$. Hence,

$$\left(k^{-\frac{1}{2}}\sum_{1\le i\le k} X_i, -k^{-\frac{1}{2}}\sum_{1\le i\le k} Z_i\right) \overset{d}{=} \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & -\cos(\tau)\\ -\cos(\tau) & 1\end{bmatrix}\right).$$

Next, applying Lemma 3.6.2, the probability in (3.97) evaluates to

$$2\cdot\left(\frac{1}{4}+\frac{1}{2\pi}\sin^{-1}\left(-\cos(\tau)\right)\right) = \frac{1}{2}+\frac{1}{\pi}\sin^{-1}\left(-\sin\left(\frac{\pi}{2}-\tau\right)\right) = \frac{\tau}{\pi}.$$

Hence $I_j \overset{d}{=} \mathrm{Ber}(\tau/\pi)$, $1\le j\le n$ i.i.d. Finally, since $d_H\Big(\mathcal{A}_{\mathrm{maj}}(\mathcal{M}), \mathcal{A}_{\mathrm{maj}}\big(\overline{\mathcal{M}}(\tau)\big)\Big) = \sum_{1\le j\le n} I_j$, the proof of Lemma 3.6.19 is complete. □

Correlated ensemble is close to the original. Next, assume that for some $T\in\mathbb{N}$, T rounds (of the algorithm) are completed so far. In particular, the algorithm produced $\sigma,\overline{\sigma}\in\{\pm 1\}^{\sum_{0\le j\le T} n_j}$. Recall the variables from Section 3.3.4:

$$\langle R_i,\sigma\rangle, \quad 1\le i\le k \qquad \text{and} \qquad \left\langle \overline{R_i},\overline{\sigma}\right\rangle, \quad 1\le i\le k,$$

where the inner products are defined in $\mathbb{R}^{\sum_{0\le j\le T} n_j}$ and $\overline{R_i}$, $1\le i\le k$ are the rows of $\overline{\mathcal{M}}(\tau)$ appearing in (3.14). We show that these ensembles are "close" to each other in the following sense.

Lemma 3.6.20. *Let $\alpha>0$ satisfy*

$$\alpha \ge 10^{-c\log_{10}(\log_{10} n)} \tag{3.98}$$

for a sufficiently small constant $c>0$ and $k=\Theta(n)$. Then with probability at least $1-\exp(-k/3)$,

$$\sup \sum_{1\le i\le k}\left|\langle R_i,\sigma\rangle - \left\langle \overline{R_i},\overline{\sigma}\right\rangle\right| \le Ck\sqrt{n}\Big(\tau\log_{10} n + n^{-\alpha/2}\Big) \tag{3.99}$$

for all large enough n, where the supremum is over all pairs (σ, J), $\sigma\in\mathcal{B}_n$ and $J=\{i\in[n]:\sigma_i\neq\overline{\sigma}_i\}\subset[n]$ with $|J|\le n^{1-\alpha}$. Here, $C>0$ is an absolute constant.

It is worth noting that due to sup term, Lemma 3.6.20 provides a uniform control for **any** pair $(\sigma,\overline{\sigma})\in\mathcal{B}_n\times\mathcal{B}_n$ with $d_H(\sigma,\overline{\sigma})\le n^{1-\alpha}$. We will show that this, in particular, captures the outputs of Kim-Roche algorithm.

Proof of Lemma 3.6.20. We start by observing that for any $\sigma \in \mathcal{B}_n$, if $J = \{i : \sigma_i \neq \overline{\sigma}_i\}$, then

$$\langle R_i, \sigma \rangle - \langle \overline{R}_i, \overline{\sigma} \rangle = \sum_{j \in J} \big(R_{ij} + \overline{R}_{ij}\big)\sigma_j + \sum_{j \in J^c} \big(R_{ij} - \overline{R}_{ij}\big)\sigma_j. \tag{3.100}$$

Next, for any **fixed** σ and J, we have by (3.14) that

$$\sum_{j \in J} \big(R_{ij} + \overline{R}_{ij}\big)\sigma_j \stackrel{d}{=} \mathcal{N}\Big(0, \Big(\sin^2\tau + \big(1 + \cos(\tau)\big)^2\Big)|J|\Big) \tag{3.101}$$

$$\sum_{j \in J^c} \big(R_{ij} - \overline{R}_{ij}\big)\sigma_j \stackrel{d}{=} \mathcal{N}\Big(0, \Big(\sin^2\tau + \big(1 - \cos(\tau)\big)^2\Big)|J^c|\Big). \tag{3.102}$$

Using simple bounds, $1 - \frac{\tau^2}{2} \leq \cos(\tau) \leq 1$ and $\sin(\tau) \leq \min\{1, \tau\}$, we obtain the following upper bounds

$$\begin{aligned}\left(\sin^2\tau + (1 + \cos\tau)^2\right)|J| &\leq 5|J| \\ \left(\sin^2\tau + (1 - \cos\tau)^2\right)|J^c| &\leq 100n\tau^2\end{aligned}$$

on the variances of variables appearing in (3.101) and (3.102).

We next set the stage to apply Bernstein's inequality [281, Proposition 5.16]: for i.i.d. $Z_i \sim \mathcal{N}(0,1)$, $1 \leq i \leq n$, there exists an absolute constant $A > 0$ such that for all $t > 0$,

$$\mathbb{P}\left(\sum_{1 \leq i \leq n} |Z_i| \geq n\mathbb{E}[|Z_1|] + t\right) \leq \exp\left(-\min\left(\frac{t^2}{4nA^2}, \frac{t}{2A}\right)\right). \tag{3.103}$$

Fix any absolute constant $C > 0$. Note that using the variance upper bound above

$$\mathbb{P}\left(\sum_{1 \leq i \leq k} \left|\sum_{j \in J} \left(R_{ij} + \overline{R}_{ij}\right)\sigma_j\right| \geq Ck\sqrt{|J|}\right) \leq \mathbb{P}\left(\sum_{1 \leq i \leq k} |Z_i| \geq \frac{Ck\sqrt{|J|}}{\sqrt{5|J|}}\right), \tag{3.104}$$

where $Z_i \stackrel{d}{=} \mathcal{N}(0,1)$, $1 \leq i \leq k$ i.i.d. Recall that $\mathbb{E}[|Z_i|] = \sqrt{2/\pi}$. Furthermore, for

$$t = \left(\frac{C}{\sqrt{5}} - \sqrt{\frac{2}{\pi}}\right)k,$$

$k = \Theta(n)$ implies that

$$\min\left(\frac{t^2}{4nA^2}, \frac{t}{2A}\right) \geq k$$

for $C > 0$ large enough. Likewise,

$$\mathbb{P}\left(\sum_{1\le i\le k}\left|\sum_{j\in J^c}\left(R_{ij}-\overline{R}_{ij}\right)\sigma_j\right|\ge Ck\sqrt{n\tau^2}\log_{10}n\right)\le\mathbb{P}\left(\sum_{1\le i\le k}|Z_i|\ge\frac{Ck\sqrt{n\tau^2}\log_{10}n}{10\sqrt{n\tau^2}}\right). \tag{3.105}$$

This time, choosing

$$t=\frac{C}{10}k\log_{10}n-k\sqrt{\frac{2}{\pi}},$$

and recalling $k=\Theta(n)$, we have

$$\min\left(\frac{t^2}{4nA^2},\frac{t}{2A}\right)\ge k\log_{10}n$$

provided $C > 0$ is large. Consequently, applying Bernstein's inequality (3.103) to (3.104) and (3.105), we obtain

$$\mathbb{P}\left(\sum_{1\le i\le k}\left|\sum_{j\in J}\left(R_{ij}+\overline{R}_{ij}\right)\sigma_j\right|\ge Ck\sqrt{|J|}\right)\le\exp(-k) \tag{3.106}$$

$$\mathbb{P}\left(\sum_{1\le i\le k}\left|\sum_{j\in J^c}\left(R_{ij}-\overline{R}_{ij}\right)\sigma_j\right|\ge Ck\sqrt{n\tau^2}\log_{10}n\right)\le\exp(-k\log_{10}n) \tag{3.107}$$

for any sufficiently large constant $C > 0$.

The bounds above are valid for any such (σ, J). We next upper bound the number of all such pairs. Note that,

$$\begin{aligned}\left|(\sigma,J):\sigma\in\{-1,1\}^J,|J|\le n^{1-\alpha}\right|&=\sum_{m\le n^{1-\alpha}}2^m\binom{n}{m}\\&\le\sum_{m\le n^{1-\alpha}}(2n)^m\\&\le n^{1-\alpha}(2n)^{n^{1-\alpha}}\\&\le\exp\bigl(C'n^{1-\alpha}\log_{10}n\bigr),\end{aligned} \tag{3.108}$$

for some absolute $C' > 0$. Likewise,

$$\begin{aligned}\left|(\sigma,J):\sigma\in\{-1,1\}^{J^c},|J|\le n^{1-\alpha}\right|&=\sum_{m\le n^{1-\alpha}}2^{n-m}\binom{n}{m}\\&\le 2^n\sum_{0\le m\le n}2^{-m}\binom{n}{m}\\&=\exp\bigl(n\ln 3\bigr),\end{aligned} \tag{3.109}$$

where we used the binomial theorem, $\sum_{0\le m\le n}2^{-m}\binom{n}{m}=3^n/2^n$.

We now prepare the stage to take union bounds. Note that since $k = \Theta(n)$, $k \log_{10} n = \Theta\big(n \log n\big) = \omega\big(n \ln 3\big)$. In particular, the cardinality term appearing in (3.109) is dominated by the corresponding probability term (3.107). Next, we compare the (order of) cardinality term (3.108) with the corresponding probability term (3.106). Note that

$$10^{-c \log_{10}(\log_{10} n)} = \big(\log_{10} n\big)^{-c}.$$

Employing this and the lower bound (3.98) on α, we obtain

$$\begin{aligned} n^{1-\alpha} \log_{10} n &\leq n^{1-10^{-c \log_{10}(\log_{10} n)}} \log_{10} n \\ &= n \cdot \underbrace{\exp\left(\frac{1}{\log_{10} e} \log_{10}(\log_{10} n) - \frac{1}{\log_{10} e} \big(\log_{10} n\big)^{1-c} \right)}_{= o(1), \text{ provided } c < 1} \\ &= o(n). \end{aligned}$$

Since $k = \Theta(n)$ and $c > 0$ is sufficiently small, it follows that the probability term appearing in (3.106) dominates the cardinality term (3.108).

Taking union bounds, we obtain

$$\mathbb{P}\left(\sup_{\sigma, J} \sum_{1 \leq i \leq k} \left| \sum_{j \in J} \big(R_{ij} + \overline{R}_{ij}\big) \sigma_j \right| \geq Ck \sqrt{n^{1-\alpha}} \right) \leq \exp(-k/2) \tag{3.110}$$

and

$$\mathbb{P}\left(\sup_{\sigma, J} \sum_{1 \leq i \leq k} \left| \sum_{j \in J^c} \big(R_{ij} - \overline{R}_{ij}\big) \sigma_j \right| \geq Ck \sqrt{n\tau^2} \log_{10} n \right) \leq \exp\big(-k \log_{10} n/2\big). \tag{3.111}$$

Finally, combining (3.110) and (3.111) via a union bound, we conclude that (3.99) holds with probability at least $1 - \exp(-k/3)$, completing the proof of Lemma 3.6.20. □

Distribution of inner products. Next, as an auxiliary step, we study the parameters of distribution of $\langle R_i, \sigma \rangle$, where σ is generated by the application of majority protocol: $\sigma_j = \operatorname{sgn}\big(\langle C_j, e \rangle\big)$, where e is the vector of all ones. Note that

$$\langle R_i, \sigma \rangle = \sum_{1 \leq j \leq n} X_{ij} \sigma_j = \sum_{1 \leq j \leq n} X_{ij} \operatorname{sgn}\left(\sum_{1 \leq i \leq k} X_{ij} \right) = \sum_{1 \leq j \leq n} X_{ij} \operatorname{sgn}\left(\frac{1}{\sqrt{k}} \sum_{1 \leq i \leq k} X_{ij} \right).$$

Notice that for any fixed row index i, the collection $X_{ij}\sigma_j$, $1 \leq j \leq n$ is i.i.d. We now compute the relevant statistics.

Lemma 3.6.21. *For any $1 \le i \le k$ and $1 \le j \le n$,*

$$\mathbb{E}\big[X_{ij}\sigma_j\big] = \sqrt{\frac{2}{\pi k}}.$$

Consequently, for any distinct $(i,j),(i',j') \in [k] \times [n]$,

$$\mathbb{E}\big[X_{ij}\sigma_j X_{i'j'}\sigma_{j'}\big] = \mathbb{1}\{j \neq j'\}\frac{2}{\pi k}$$

Proof of Lemma 3.6.21. For simplicity, we drop the index i below whenever convenient. Observe that $\mathbb{P}(\sigma_1 = +1) = \mathbb{P}(\sigma_1 = -1) = \frac{1}{2}$. We then have

$$\begin{aligned}
\mathbb{E}\big[X_1\sigma_1\big] &= \frac{1}{2}\Big(\mathbb{E}\big[X_1|\sigma_1 = +1\big] - \mathbb{E}\big[X_1|\sigma_1 = -1\big]\Big) \\
&= \frac{1}{2}\left(\mathbb{E}\left[X_1\middle|\frac{1}{\sqrt{k}}\sum_{1\le j\le k} X_j \ge 0\right] + \mathbb{E}\left[-X_1\middle|-\frac{1}{\sqrt{k}}\sum_{1\le j\le k} X_j \ge 0\right]\right) \\
&= \mathbb{E}\left[X_1\middle|\frac{1}{\sqrt{k}}\sum_{1\le j\le k} X_j \ge 0\right] = \sqrt{\frac{2}{\pi k}},
\end{aligned}$$

where we applied Lemma 3.6.3 for the bivariate normal

$$\left(X_1, \frac{1}{\sqrt{k}}\sum_{1\le j\le k} X_j\right) \overset{d}{=} \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \frac{1}{\sqrt{k}}\\ \frac{1}{\sqrt{k}} & 1\end{bmatrix}\right).$$

Having established the claim for $\mathbb{E}[X_{ij}\sigma_j]$, the rest is straightforward. Take $(i,j) \neq (i',j')$. Note that if $j = j'$, we are done since X_{ij} and $X_{i'j}$ are independent with mean zero. Assume $j \neq j'$. Then, $X_{ij}\sigma_j$ and $X_{i'j'}\sigma_{j'}$ are i.i.d. This completes the proof of Lemma 3.6.21. □

Thresholding suffices to find k_j indices. We now establish that for finding the k_j (row) indices to be used in round j of the algorithm, it suffices to threshold the inner products. This is a consequence of the following concentration result.

Lemma 3.6.22. *Suppose that $0 < c < \log_{10} 2$ is an arbitrary constant, and $1 \le T \le c\log_{10}\log_{10} n$ is an arbitrary integer. Let $\sigma \in \mathbb{R}^{S_T}$ for $S_T = \sum_{0\le s\le T} n_s$ be the output of $\mathcal{A}_{\mathrm{KR}}$ at the end of Tth round. Then, for any $x \in \mathbb{R}$, and $\epsilon > 0$,*

$$\mathbb{E}\left[\left(\left|\big\{1 \le i \le k : \langle R_i, \sigma\rangle < x\big\}\right| - k\Phi\left(\frac{1}{\sqrt{S_T}}\left(x - \sum_{0\le s\le T} n_s\sqrt{\frac{2k_s}{\pi k^2}}\right)\right)\right)^2\right] \le O\left(n^{\frac{39}{20}+\epsilon}\right),$$

where $\Phi(t) = \mathbb{P}(Z \le t)$ for $Z \sim \mathcal{N}(0,1)$, k_s is defined in (3.11) and n_s is defined in (3.13).

Proof of Lemma 3.6.22. We consider

$$\mathbb{P}\Big(\langle R_i, \sigma\rangle < x\Big) \quad \text{and} \quad \mathbb{P}\Big(\langle R_i, \sigma\rangle < x, \langle R_{i'}, \sigma\rangle < x\Big) \tag{3.112}$$

for $1 \le i, i' \le k$ and $i \neq i'$. Define the running sums

$$S_t = \sum_{0 \le j \le t} n_j, \quad 0 \le t \le T. \tag{3.113}$$

That is, S_t is the number of entries of σ assigned at the end of round t. Suppose that the algorithm run for T rounds, where $1 \le T \le c \log_{10} \log_{10} n$ for $c > 0$ sufficiently small. With this notation,

$$\langle R_i, \sigma\rangle = \sum_{1 \le j \le S_T} X_{ij}\sigma_j = \sum_{1 \le j \le n_0} X_{ij}\sigma_j + \sum_{1 \le t \le T} \sum_{S_{t-1}+1 \le j \le S_t} X_{ij}\sigma_j.$$

To analyze the distribution of this value, let us define

$$U(i, i') \triangleq \sum_{0 \le t \le T} \sum_{S_{t-1}+1 \le j \le S_t} X_{ij}\widetilde{\sigma}_j + \sum_{0 \le t \le T} n_t \mu_t, \tag{3.114}$$

where for $S_{t-1} + 1 \le j \le S_t$, and $0 \le t \le T$ (with the convention $S_{-1} \triangleq 0$, $\mathcal{I}_0 \triangleq [k]$ and $k_0 \triangleq k$)

$$\widetilde{\sigma}_j \triangleq \operatorname{sgn}\left(\sum_{\ell \in \mathcal{I}_t \setminus \{i,i'\}} X_{\ell j}\right) \qquad \text{and} \qquad \mu_t = \mathbb{E}\big[X_{ij}\sigma_j\big].$$

We suppress the dependence of $\widetilde{\sigma}$ on i, i' for convenience. Observe that $\widetilde{\sigma}_j$ is independent of all X_{ij} and $X_{i'j}$.

We now compute μ_t appearing above. To that end, we remind the reader the (index) set $\mathcal{I}_t$ for convenience: for any $1 \le i \le k$, $i \in \mathcal{I}_t$ iff the (partial) inner product, $\langle R_i, \sigma\rangle$, is among the smallest k_t (partial) inner products $\langle R_j, \sigma\rangle$, $1 \le j \le k$.

Next, note that σ_j (the sign assigned to column j) is obtained by taking a majority vote in $k_t \times n_t$ submatrix with row indices prescribed by $\mathcal{I}_t$. Note that if $i \notin \mathcal{I}_t$, X_{ij} and σ_j are independent. Furthermore, given any i,

$$\mathbb{P}(i \in \mathcal{I}_t) = \binom{k}{k_t}^{-1} \binom{k-1}{k_t - 1} = \frac{k_t}{k}.$$

from symmetry. Consequently,

$$\begin{aligned}\mu_t &= \mathbb{E}\big[X_{ij}\sigma_j \mid i \in \mathcal{I}_t\big]\mathbb{P}(i \in \mathcal{I}_t) + \underbrace{\mathbb{E}\big[X_{ij}\sigma_j \mid i \notin \mathcal{I}_t\big]}_{=0}\mathbb{P}(i \notin \mathcal{I}_t) \\ &= \sqrt{\frac{2}{\pi k_t}} \cdot \frac{k_t}{k} = \sqrt{\frac{2k_t}{\pi k^2}},\end{aligned} \tag{3.115}$$

where we used Lemma 3.6.21 to invoke $\mathbb{E}\big[X_{ij}\sigma_j \mid i \in \mathcal{I}_t\big] = \sqrt{\frac{2}{\pi k_t}}$.

Define now

$$\Delta_{i,i'} \triangleq \langle R_i, \sigma\rangle - U(i,i').$$

Since $\mathbb{E}\big[X_{ij}\widetilde{\sigma}_j\big] = 0$, we obtain $\mathbb{E}[\Delta_{i,i'}] = 0$ from the choice of μ_t, $0 \le t \le T$. We now claim

Lemma 3.6.23.

$$\mathrm{Var}\Big(\Delta_{i,i'}\Big) = O\left(\sum_{0\le t\le T} n_t \cdot k_t^{-\frac{1}{4}}\right). \tag{3.116}$$

Moreover, if $T \le c\log_{10}\log_{10} n$ with $c < \log_{10} 2$, then

$$\mathrm{Var}\Big(\Delta_{i,i'}\Big) = O\left(n^{\frac{3}{4}+\epsilon}\right) \tag{3.117}$$

for any $\epsilon > 0$.

Proof of Lemma 3.6.23. Note that for $S_{t-1}+1 \le j \le S_t$, σ_j is a function of a $k_t \times n_t$ submatrix with i.i.d. $\mathcal{N}(0,1)$ entries that has not been inspected yet. Hence

$$\begin{aligned}
\mathrm{Var}\Big(\Delta_{i,i'}\Big) &= \mathrm{Var}\left(\sum_{0\le t\le T}\sum_{S_{t-1}+1\le j\le S_t} X_{ij}\left(\sigma_j - \widetilde{\sigma}_j\right)\right)\\
&= \sum_{0\le t\le T} n_t \mathrm{Var}\Big(X_{ij}\left(\sigma_j - \widetilde{\sigma}_j\right)\Big)\\
&\le \sum_{0\le t\le T} n_t \mathbb{E}\Big[X_{ij}^2\left(\sigma_j - \widetilde{\sigma}_j\right)^2\Big]\\
&\le \sum_{0\le t\le T} n_t \sqrt{\mathbb{E}\left[X_{ij}^4\right]\mathbb{E}\left[\left(\sigma_j - \widetilde{\sigma}_j\right)^4\right]}\\
&= O\left(\sum_{0\le t\le T} n_t \sqrt{\mathbb{P}\left(\sigma_j \ne \widetilde{\sigma}_j\right)}\right),
\end{aligned}$$

where the second line uses the fact that for any fixed t and $S_{t-1}+1 \le j \le S_t$ the distributions of $X_{ij}\big(\sigma_j - \widetilde{\sigma}_j\big)$ are identical; the third line uses $\mathrm{Var}(U) \le \mathbb{E}[U^2]$; and the fourth line uses Cauchy-Schwarz inequality.

We now show that for $S_{t-1}+1 \le j \le S_t$, $0 \le t \le T$,

$$\mathbb{P}\Big(\sigma_j \ne \widetilde{\sigma}_j\Big) = O\left(\frac{1}{\sqrt{k_t}}\right)$$

which will establish Lemma 3.6.23. We have that

$$\mathbb{P}\Big(\sigma_j \ne \widetilde{\sigma}_j\Big) \le \underbrace{\mathbb{P}\Big(\sigma_j \ne \widetilde{\sigma}_j \Big| i, i' \notin \mathcal{I}_t\Big)}_{=0} + \mathbb{P}\Big(\sigma_j \ne \widetilde{\sigma}_j \Big| |\mathcal{I}_t \cap \{i,i'\}| = 1\Big) + \mathbb{P}\Big(\sigma_j \ne \widetilde{\sigma}_j \Big| i, i' \in \mathcal{I}_t\Big).$$

Recall now from Lemma 3.6.2 that for a pair (S_1, S_2) of bivariate normal random variables $S_1, S_2 \stackrel{d}{=} \mathcal{N}(0,1)$ with parameter ρ,

$$\mathbb{P}\Big(\mathrm{sgn}(S_1) \neq \mathrm{sgn}(S_2)\Big) = \frac{1}{2} - \frac{1}{\pi}\sin^{-1}(\rho),$$

which, in particular, is a decreasing function of ρ. Now,

$$\mathbb{P}\Big(\sigma_j \neq \widetilde{\sigma}_j \Big| |\mathcal{I}_t \cap \{i,i'\}| = 1\Big) = \mathbb{P}\left(\mathrm{sgn}\left(\frac{1}{\sqrt{k_t - 1}} \sum_{1 \leq i \leq k_t - 1} Z_i\right) \neq \mathrm{sgn}\left(\frac{1}{\sqrt{k_t}} \sum_{1 \leq i \leq k_t} Z_i\right)\right),$$

where Z_i, $1 \leq i \leq k_t$ are i.i.d. $\mathcal{N}(0,1)$. Setting $S_1 = (k_t - 1)^{-\frac{1}{2}} \sum_{1 \leq i \leq k_t - 1} Z_i$ and $S_2 = k_t^{-\frac{1}{2}} \sum_{1 \leq i \leq k_t} Z_i$, we find that (S_1, S_2) is a bivariate normal with parameter $\sqrt{1 - \frac{1}{k_t}}$. Likewise, a similar argument yields that $\mathbb{P}\Big(\sigma_j \neq \widetilde{\sigma}_j | i, i' \in \mathcal{I}_t\Big) = \mathbb{P}\big(\mathrm{sgn}(S_1') \neq \mathrm{sgn}(S_2')\big)$, where $S_1', S_2' \stackrel{d}{=} \mathcal{N}(0,1)$ is a bivariate normal with parameter $\sqrt{1 - \frac{2}{k_t}}$. Consequently,

$$\mathbb{P}\Big(\sigma_j \neq \widetilde{\sigma}_j\Big) \leq \mathbb{P}\Big(\sigma_j \neq \widetilde{\sigma}_j \Big| |\mathcal{I}_t \cap \{i,i'\}| = 1\Big) + \mathbb{P}\Big(\sigma_j \neq \widetilde{\sigma}_j | i, i' \in \mathcal{I}_t\Big) \leq 2\mathbb{P}\Big(\sigma_j \neq \widetilde{\sigma}_j | i, i' \in \mathcal{I}_t\Big).$$

Next, for fixed $i \neq i'$; set $S \triangleq \sum_{\ell \in \mathcal{I}_t \setminus \{i,i'\}} X_{\ell j}$. We then have,

$$\begin{aligned}
\mathbb{P}\Big(\sigma_j \neq \widetilde{\sigma}_j | i, i' \in \mathcal{I}_t\Big) &= 2\mathbb{P}\big(S + X_i + X_{i'} \geq 0, S \leq 0\big) \\
&= 2\mathbb{P}\left(\frac{1}{\sqrt{k_j}}(S + X_i + X_{i'}) \geq 0, -\frac{1}{\sqrt{k_j - 2}} S \geq 0\right) \\
&= \frac{1}{2} - \frac{1}{\pi}\sin^{-1}\left(\sqrt{1 - \frac{2}{k_j}}\right) \\
&= \frac{1}{2} - \frac{1}{\pi}\sin^{-1}\left(1 - \frac{1}{k_j} + O\left(\frac{1}{k_j^2}\right)\right) \\
&= \frac{1}{2} - \frac{1}{\pi}\left(\frac{\pi}{2} - O\left(\frac{1}{\sqrt{k_j}}\right)\right) = O\left(\frac{1}{\sqrt{k_j}}\right),
\end{aligned}$$

where the first line uses symmetry; the third line uses Lemma 3.6.2; and the fourth line uses $\sqrt{1-x} = 1 - \frac{x}{2} + O(x^2)$ and $\sin^{-1}(1-x) = \frac{\pi}{2} - \sqrt{2x} + O(x^{3/2})$. Hence, we established

$$\mathrm{Var}\Big(\Delta_{i,i'}\Big) = O\left(\sum_{0 \leq t \leq T} n_t \cdot k_t^{-\frac{1}{4}}\right),$$

where $O(\cdot)$ only hides absolute constants. Namely, (3.116) holds. Next, let $c < \log_{10} 2$.

Then we claim

$$\sum_{0\le j\le c\log_{10}\log_{10} n} \frac{n_j}{\sqrt{k_j}} = O\left(n^{\frac{3}{4}+\epsilon}\right)$$

for any $\epsilon > 0$, which will yield (3.117).

In the remainder, we omit floor/ceiling operators whenever convenient. Note that for k_j defined in (3.11),

$$k_j = 2\lfloor (1/2) f_j^3 \cdot n \rfloor + 1 \ge f_j^3 n - 1.$$

Moreover, for n_j appearing in (3.13),

$$n_j = \left\lfloor \frac{n}{A} \sum_{0\le i\le j} f_i \right\rfloor - \left\lfloor \frac{n}{A} \sum_{0\le i\le j-1} f_i \right\rfloor \le \frac{n}{A} f_j + 1 \le n f_j + 1,$$

using the fact that $A \ge 1$ per (3.12). Next, using $f_j \le 1$, we have

$$n f_j \ge n f_j^3 \ge n \cdot \exp_{10}\left(-3 \cdot 2^{c\log_{10}\log_{10} n}\right) = \exp_{10}\left(\log_{10} n - 3 \cdot (\log_{10} n)^{c'}\right) = \omega(1)$$

where $c' = c\log_{10} 2 < (\log_{10} 2)^2 < 1$. Here, we used the fact that $j \le L$, where $L = c\log_{10}\log_{10} n$ appears in Proposition 3.6.18 with $c > 0$ small enough.

Namely, $n f_j, n f_j^3 = \omega(1)$. Employing this, together with $n_j \le n f_j$ and $k_j \ge n f_j^3 - 1$ established above, we have

$$\begin{aligned}
O\left(\sum_{0\le j\le c\log_{10}\log_{10} n} n_j \cdot k_j^{-\frac{1}{4}}\right) &= O\left(\sum_{0\le j\le c\log_{10}\log_{10} n} n f_j \cdot (n f_j^3)^{-\frac{1}{4}}\right) \\
&= O\left(\sum_{0\le j\le c\log_{10}\log_{10} n} n^{\frac{3}{4}} f_j^{\frac{1}{4}}\right) \\
&= O\left(\log_{10}\log_{10} n \cdot n^{\frac{3}{4}}\right) \\
&= O\left(n^{\frac{3}{4}+\epsilon}\right), \quad \forall \epsilon > 0,
\end{aligned}$$

where the first line uses the bounds $n_j \le n f_j + 1$ and $k_j \ge n f_j^3 - 1$ and the third line uses the fact $f_j = o(1)$. This concludes the proof of Lemma 3.6.23. □

Lemma 3.6.23 yields

$$\mathbb{E}\left[\Delta_{i,i'}^2\right] = \mathrm{Var}\left(\Delta_{i,i'}\right) = O\left(n^{\frac{3}{4}+\epsilon}\right) \tag{3.118}$$

for any $\epsilon > 0$. Now, recall that

$$\langle R_i, \sigma\rangle = U(i,i') + \Delta_{i,i'} \qquad \text{and} \qquad \langle R_{i'}, \sigma\rangle = U(i',i) + \Delta_{i',i}.$$

In particular, using Chebyshev's inequality and (3.118), we obtain

$$\mathbb{P}\Big(\big|\Delta_{i',i}\big| > k^{\frac{2}{5}}\Big) = \mathbb{P}\Big(\big|\Delta_{i,i'}\big| > k^{\frac{2}{5}}\Big) \leq k^{-\frac{4}{5}}\mathbb{E}\Big[\Delta_{i,i'}^2\Big] = O\left(n^{-\frac{1}{20}+\epsilon}\right), \tag{3.119}$$

for any $\epsilon > 0$, as $k = \Theta(n)$. Equipped with these,

$$\begin{aligned}\mathbb{P}\Big(\langle R_i, \sigma\rangle < x, \langle R_{i'}, \sigma\rangle < x\Big) &= \mathbb{P}\Big(U(i,i') + \Delta_{i,i'} < x, U(i',i) + \Delta_{i',i} < x\Big) \\ &\leq \mathbb{P}\Big(U(i,i') < x + k^{\frac{2}{5}}, U(i',i) < x + k^{\frac{2}{5}}\Big) + \mathbb{P}\Big(\big|\Delta_{i,i'}\big| > k^{\frac{2}{5}}\Big) + \mathbb{P}\Big(\big|\Delta_{i',i}\big| > \\ &= \mathbb{P}\Big(U(i,i') < x + k^{\frac{2}{5}}, U(i',i) < x + k^{\frac{2}{5}}\Big) + O\left(n^{-\frac{1}{20}+\epsilon}\right),\end{aligned}$$

where the last line uses (3.119). Next, we establish an auxiliary lemma.

Lemma 3.6.24. *The random variables*

$$\frac{1}{\sqrt{S_T}}\sum_{1\leq j\leq S_T} X_{ij}\widetilde{\sigma}_j \qquad \textit{and} \qquad \frac{1}{\sqrt{S_T}}\sum_{1\leq j\leq S_T} X_{i'j}\widetilde{\sigma}_j$$

are i.i.d. standard normal.

Proof of Lemma 3.6.24. Note that X_{ij}, $1 \leq j \leq S_T$ and $X_{i'j}$, $1 \leq j \leq S_T$ are i.i.d. $\mathcal{N}(0,1)$. Moreover, $\widetilde{\sigma}_j$, $1 \leq j \leq S_T$ is an i.i.d. collection with

$$\mathbb{P}\big[\widetilde{\sigma}_j = 1\big] = \frac{1}{2} = \mathbb{P}\big[\widetilde{\sigma}_j = -1\big] \tag{3.120}$$

and that $\widetilde{\sigma}_j$ are independent of X_{ij} and $X_{i'j}$. Next, we show if $X \stackrel{d}{=} \mathcal{N}(0,1)$ and $\widetilde{\sigma}$ has the distribution (3.120) and is independent of X, then $\widetilde{\sigma}X \stackrel{d}{=} \mathcal{N}(0,1)$. To see this, we rely on characteristic functions: for any $t \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\big[e^{it\widetilde{\sigma}X}\big] &= \mathbb{E}\big[e^{it\widetilde{\sigma}X}\big|\widetilde{\sigma} = 1\big]\mathbb{P}\big[\widetilde{\sigma} = 1\big] + \mathbb{E}\big[e^{it\widetilde{\sigma}X}\big|\widetilde{\sigma} = -1\big]\mathbb{P}\big[\widetilde{\sigma} = -1\big] \\ &= \frac{1}{2}\left(\mathbb{E}\big[e^{itX}\big] + \mathbb{E}\big[e^{-itX}\big]\right) \\ &= \exp\big(-\frac{t^2}{2}\big).\end{aligned}$$

Using Lévy's inversion theorem [288, Section 16.6], it follows that $\widetilde{\sigma}X \stackrel{d}{=} \mathcal{N}(0,1)$. Applying this fact, since $X_{ij}\widetilde{\sigma}_j$, $1 \leq j \leq S_T$ is an i.i.d. $\mathcal{N}(0,1)$ collection, we deduce

$$\frac{1}{\sqrt{S_T}}\sum_{1\leq j\leq S_T} X_{ij}\widetilde{\sigma}_j \stackrel{d}{=} \mathcal{N}(0,1) \qquad \text{and} \qquad \frac{1}{\sqrt{S_T}}\sum_{1\leq j\leq S_T} X_{i'j}\widetilde{\sigma}_j \stackrel{d}{=} \mathcal{N}(0,1).$$

Finally, we show $X_{ij}\widetilde{\sigma}_j \perp X_{i'j}\widetilde{\sigma}_j$, which will yield Lemma 3.6.24. Once again, we rely on Lévy's inversion theorem. Let

$$\boldsymbol{Z} = (X_{ij}\widetilde{\sigma}_j, X_{i'j}\widetilde{\sigma}_j) \qquad \text{and} \qquad \boldsymbol{t} = (t_1, t_2).$$

We have

$$\begin{aligned}\mathbb{E}\big[e^{i\boldsymbol{t}^T\boldsymbol{Z}}\big] &= \mathbb{E}\left[e^{i\widetilde{\sigma}_j\left(t_1X_{ij}+t_2X_{i'j}\right)}\right]\\ &= \mathbb{E}\left[e^{i\widetilde{\sigma}_j\left(t_1X_{ij}+t_2X_{i'j}\right)}\Big|\widetilde{\sigma}_j=1\right]\mathbb{P}\big[\widetilde{\sigma}_j=1\big]+\mathbb{E}\left[e^{i\widetilde{\sigma}_j\left(t_1X_{ij}+t_2X_{i'j}\right)}\Big|\widetilde{\sigma}_j=-1\right]\mathbb{P}\big[\widetilde{\sigma}_j=-1\big]\\ &= \exp\left(-\frac{t_1^2+t_2^2}{2}\right),\end{aligned}$$

where we used the fact $t_1X_{ij}+t_2X_{i'j}\stackrel{d}{=}\mathcal{N}\big(0,t_1^2+t_2^2\big)$. Clearly,

$$\mathbb{E}\big[e^{i\boldsymbol{t}^T\boldsymbol{Z}}\big]=\mathbb{E}\big[e^{it_1X_{ij}\widetilde{\sigma}_j}\big]\mathbb{E}\big[e^{it_2X_{i'j}\widetilde{\sigma}_j}\big],$$

since $t_1X_{ij}\widetilde{\sigma}_j\stackrel{d}{=}\mathcal{N}(0,t_1^2)$ and $t2_2X_{i'j}\widetilde{\sigma}_j\stackrel{d}{=}\mathcal{N}(0,t_2^2)$. Since $t_1,t_2\in\mathbb{R}$ are arbitrary, we complete the proof of Lemma 3.6.24 by appealing once again to Lévy's inversion theorem. □

Denote $\Phi(t)=\mathbb{P}[\mathcal{N}(0,1)\le t]$. For $S_T=\sum_{0\le t\le T}n_t$, we have

$$\begin{aligned}&\mathbb{P}\Big(U(i,i')<x+k^{\frac{2}{5}},U(i',i)<x+k^{\frac{2}{5}}\Big)\\ &=\mathbb{P}\left(\frac{1}{\sqrt{S_T}}\sum_{1\le j\le S_T}X_{ij}\widetilde{\sigma}_j,\frac{1}{\sqrt{S_T}}\sum_{1\le j\le S_T}X_{i'j}\widetilde{\sigma}_j<\frac{x}{\sqrt{S_T}}-\frac{1}{\sqrt{S_T}}\sum_{0\le t\le T}n_t\sqrt{\frac{2k_t}{\pi k^2}}+\frac{k^{\frac{2}{5}}}{\sqrt{S_T}}\right)\\ &=\Phi\left(\frac{x}{\sqrt{S_T}}-\frac{1}{\sqrt{S_T}}\sum_{0\le t\le T}n_t\sqrt{\frac{2k_t}{\pi k^2}}+\frac{k^{\frac{2}{5}}}{\sqrt{S_T}}\right)^2,\end{aligned}$$

where the second line uses the expressions for $U(i,i')$ and $U(i',i)$ per (3.114) and for μ_t per (3.115); and the last line uses Lemma 3.6.24. Observe next that $\Phi(\cdot)$ is trivially $1-$Lipschitz:

$$\big|\Phi(t_1)-\Phi(t_2)\big|=\int_{\min\{t_1,t_2\}}^{\max\{t_1,t_2\}}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\,dt\le\big|t_1-t_2\big|.$$

With this, we have

$$\Phi\left(\frac{x}{\sqrt{S_T}}-\frac{1}{\sqrt{S_T}}\sum_{0\le t\le T}n_t\sqrt{\frac{2k_t}{\pi k^2}}+\frac{k^{\frac{2}{5}}}{\sqrt{S_T}}\right)-\Phi\left(\frac{x}{\sqrt{S_T}}-\frac{1}{\sqrt{S_T}}\sum_{0\le t\le T}n_t\sqrt{\frac{2k_t}{\pi k^2}}\right)\le\frac{k^{\frac{2}{5}}}{\sqrt{S_T}}.$$

Moreover, note that $\Theta(n)=n_0\le S_T\le n$ yields $S_T=\Theta(n)$, hence in particular $k^{2/5}/\sqrt{S_T}=\Theta(n^{-1/10})$ as $k=\Theta(n)$ too. Consequently,

$$\Phi\left(\frac{x}{\sqrt{S_T}}-\frac{1}{\sqrt{S_T}}\sum_{0\le t\le T}n_t\sqrt{\frac{2k_t}{\pi k^2}}+\frac{k^{\frac{1}{3}}}{\sqrt{S_T}}\right)^2\le\Phi\left(\frac{x}{\sqrt{S_T}}-\frac{1}{\sqrt{S_T}}\sum_{0\le t\le T}n_t\sqrt{\frac{2k_t}{\pi k^2}}\right)^2+O\left(n^{-\frac{1}{10}}\right).$$

Likewise, using inequality $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq \mathbb{P}(\mathcal{E}_1) - \mathbb{P}(\mathcal{E}_2^c)$ valid for all events $\mathcal{E}_1, \mathcal{E}_2$, we have

$$\begin{aligned}
\mathbb{P}\Big(\langle R_i, \sigma\rangle < x, \langle R_{i'}, \sigma\rangle < x\Big) &\geq \mathbb{P}\Big(U(i,i') < x - k^{\frac{2}{5}}, U(i',i) < x - k^{\frac{2}{5}}, \Delta_{i,i'} < k^{\frac{1}{3}}, \Delta_{i',i} < k^{\frac{2}{5}}\Big) \\
&\geq \mathbb{P}\Big(U(i,i') < x - k^{\frac{2}{5}}, U(i',i) < x - k^{\frac{2}{5}}\Big) - \mathbb{P}\Big(\Delta_{i,i'} > k^{\frac{2}{5}} \quad \text{or} \quad \Delta_{i',i} > k^{\frac{2}{5}}\Big) \\
&\geq \Phi\left(\frac{x}{\sqrt{S_T}} - \frac{1}{\sqrt{S_T}} \sum_{0 \leq t \leq T} n_t \sqrt{\frac{2k_t}{\pi k^2}}\right)^2 - O\left(n^{-\frac{1}{20}+\epsilon}\right),
\end{aligned}$$

where in the last line we have once again used (3.119). Combining these, we arrive at

$$\left|\mathbb{P}\Big(\langle R_i, \sigma\rangle < x, \langle R_{i'}, \sigma\rangle < x\Big) - \Phi\left(\frac{x}{\sqrt{S_T}} - \frac{1}{\sqrt{S_T}} \sum_{0 \leq t \leq T} n_t \sqrt{\frac{2k_t}{\pi k^2}}\right)^2\right| = O\left(n^{-\frac{1}{20}+\epsilon}\right). \tag{3.121}$$

Similarly, we have

$$\left|\mathbb{P}\Big(\langle R_i, \sigma\rangle < x\Big) - \Phi\left(\frac{x}{\sqrt{S_T}} - \frac{1}{\sqrt{S_T}} \sum_{0 \leq t \leq T} n_t \sqrt{\frac{2k_t}{\pi k^2}}\right)\right| = O\left(n^{-\frac{1}{20}+\epsilon}\right). \tag{3.122}$$

Next, let

$$\xi \triangleq \Phi\left(\frac{x}{\sqrt{S_T}} - \frac{1}{\sqrt{S_T}} \sum_{0 \leq t \leq T} n_t \sqrt{\frac{2k_t}{\pi k^2}}\right).$$

Moreover, set

$$\begin{aligned}
\Delta_1 &\triangleq \mathbb{P}\Big(\langle R_i, \sigma\rangle < x, \langle R_{i'}, \sigma\rangle < x\Big) - \Phi\left(\frac{x}{\sqrt{S_T}} - \frac{1}{\sqrt{S_T}} \sum_{0 \leq t \leq T} n_t \sqrt{\frac{2k_t}{\pi k^2}}\right)^2 \\
\Delta_2 &\triangleq \mathbb{P}\Big(\langle R_i, \sigma\rangle < x\Big) - \Phi\left(\frac{x}{\sqrt{S_T}} - \frac{1}{\sqrt{S_T}} \sum_{0 \leq t \leq T} n_t \sqrt{\frac{2k_t}{\pi k^2}}\right).
\end{aligned}$$

Then, $|\Delta_1|, |\Delta_2| = O(n^{-\frac{1}{20}+\epsilon})$. Using (3.121) and (3.122), we obtain

$$\begin{aligned}
&\left|\mathbb{E}\left[\Big(\mathbb{1}\big\{\langle R_i,\sigma\rangle < x\big\} - \xi\Big)\Big(\mathbb{1}\big\{\langle R_{i'},\sigma\rangle < x\big\} - \xi\Big)\right]\right| \\
&= \left|\mathbb{P}\Big(\langle R_i,\sigma\rangle < x, \langle R_{i'},\sigma\rangle < x\Big) - 2\mathbb{P}\big(\langle R_i,\sigma\rangle < x\big)\xi + \xi^2\right| \\
&= \left|\Delta_1 + \xi^2 - 2\xi\big(\Delta_2 + \xi\big) + \xi^2\right| \\
&= \left|\Delta_1 - 2\xi\Delta_2\right| \\
&= O\Big(n^{-\frac{1}{20}+\epsilon}\Big)
\end{aligned}$$

as $\xi < 1$. Hence, we arrive at

$$\begin{aligned}
&\mathbb{E}\left[\left(\left|\{1 \le i \le k : \langle R_i,\sigma\rangle < x\}\right| - k\Phi\left(\frac{x}{\sqrt{S_T}} - \frac{1}{\sqrt{S_T}}\sum_{0\le t\le T} n_t\sqrt{\frac{2k_t}{\pi k^2}}\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{1\le i\le k}\left(\mathbb{1}\big\{\langle R_i,\sigma\rangle < x\big\} - \Phi\left(\frac{x}{\sqrt{S_T}} - \frac{1}{\sqrt{S_T}}\sum_{0\le t\le T} n_t\sqrt{\frac{2k_t}{\pi k^2}}\right)\right)\right)^2\right] \\
&= O(k) + k^2 O\left(n^{-\frac{1}{20}+\epsilon}\right) = O\left(n^{\frac{39}{20}+\epsilon}\right),
\end{aligned}$$

since $k \le n$. This concludes the proof of Lemma 3.6.22. □

Index sets are nearly identical. Next, assume that the algorithm completed $\ell-1$ rounds and generated $\sigma, \overline{\sigma} \in \{\pm 1\}^{\sum_{0\le t\le \ell-1} n_t}$. Recall that $\mathcal{I}_\ell$ is the set of (row) indices $1 \le i \le k$ corresponding to smallest k_ℓ elements among $\left(\sum_{0\le t\le \ell-1} n_t\right)^{-\frac{1}{2}} \langle R_i,\sigma\rangle$, $1 \le i \le k$. Likewise, $\overline{\mathcal{I}}_\ell$ denotes the corresponding set of indices for $\left(\sum_{0\le t\le \ell-1} n_t\right)^{-\frac{1}{2}} \langle \overline{R}_i,\overline{\sigma}\rangle$, $1 \le i \le k$. In particular, Lemma 3.6.22 yields that there is an x_ℓ such that w.h.p.,

$$\mathcal{I}_\ell \approx \Big\{1 \le i \le k : \langle R_i,\sigma\rangle < x_\ell\Big\} \qquad \text{and} \qquad \overline{\mathcal{I}}_\ell \approx \Big\{1 \le i \le k : \big\langle \overline{R}_i,\overline{\sigma}\big\rangle < x_\ell\Big\}.$$

We now show $\mathcal{I}_\ell$ and $\overline{\mathcal{I}}_\ell$ are nearly identical: $|\mathcal{I}_\ell \cap \overline{\mathcal{I}}_\ell| \ge k_\ell - o(k_\ell)$.

Lemma 3.6.25. *Recall J_ℓ from* (3.96) *and assume that $|J_{\ell-1}| \le n^{1-\alpha_{\ell-1}}$ for $\alpha_{\ell-1}$ defined in* (3.95)*. Then,*

$$\left|\mathcal{I}_\ell \cap \overline{\mathcal{I}}_\ell\right| \ge k_\ell - O\left(n^{1-\alpha_{\ell-1}/4}\right)$$

with probability at least $1 - O\left(n^{-1/40+\epsilon}\right)$, where $\epsilon > 0$ is arbitrary.

Proof of Lemma 3.6.25. Recall $S_{\ell-1} = \sum_{0\le s\le \ell-1} n_s$ appearing in Lemma 3.6.22. We

find $x_\ell \in \mathbb{R}$ satisfying

$$k\Phi\left(\frac{1}{\sqrt{S_{\ell-1}}}\left(x_\ell - \sum_{0\le s\le \ell-1} n_s\sqrt{\frac{2k_s}{\pi k^2}}\right)\right) = k_\ell \tag{3.123}$$

as $\Phi(\cdot)$ is continuous. For this choice of x_ℓ, using Lemma 3.6.22 with $T = \ell - 1$ and applying Markov's inequality, we arrive at

$$\mathbb{P}\left(\left|\left|\{1\le i\le k : \langle R_i,\sigma\rangle < x_\ell\}\right| - k_\ell\right| > n^{\frac{79}{80}}\right) \le O\left(n^{-\frac{1}{40}+\epsilon}\right), \tag{3.124}$$

where $\epsilon > 0$ is arbitrary.

We now claim that as long as $\ell \le c\log_{10}\log_{10} n$ for $c > 0$ small enough,

$$k_\ell = \omega\Big(n^{\frac{79}{80}}\Big).$$

Recall k_ℓ from (3.11). We have

$$k_\ell = 2\left\lfloor \frac{1}{2} n f_\ell^3 \right\rfloor + 1 \ge n f_\ell^3 - 1 = n\cdot 10^{-3\cdot 2^\ell} - 1 \ge n\cdot 10^{-3\cdot 2^{c\log_{10}\log_{10} n}} - 1.$$

Above, we used the fact $\ell \le L \le c\log_{10}\log_{10} n$.

Rearranging, we have

$$n\cdot 10^{-3\cdot 2^{c\log_{10}\log_{10} n}} - 1 = \exp_{10}\left(\log_{10} n - 3\cdot\left(\log_{10} n\right)^{c\log_{10} 2}\right) - 1.$$

From here, it is evident that if $c\log_{10} 2 < 1$, then indeed $k_\ell = \omega(n^{79/80})$. Consequently, (3.124) yields

$$k_\ell + O\left(n^{\frac{79}{80}}\right) \ge \left|\{1\le i\le k : \langle R_i,\sigma\rangle < x_\ell\}\right| \ge k_\ell - O\left(n^{\frac{79}{80}}\right),$$

with probability $1 - O(n^{-1/40+\epsilon})$. Define next the sets

$$\mathrm{LG}_j \triangleq \Big\{1\le i\le k : \big|\langle R_i,\sigma\rangle - \langle \overline{R}_i,\overline{\sigma}\rangle\big| \ge n^{1/2-\beta_j}\Big\}, \tag{3.125}$$

where the inner product appearing in LG_j is taken over $\mathbb{R}^{\sum_{0\le s\le j} n_s}$ and

$$\beta_j = \alpha_j/4 = \alpha_0\cdot 10^{-j}/4, \quad \text{where} \quad \alpha_0 = 0.04. \tag{3.126}$$

Note that under the assumption $|J_{\ell-1}| \le n^{1-\alpha_{\ell-1}}$, Lemma 3.6.20 yields

$$\sum_{1\le i\le k}\big|\langle R_i,\sigma\rangle - \langle \overline{R}_i,\overline{\sigma}\rangle\big| \le Ck\sqrt{n}\left(\tau\log_{10} n + n^{-\alpha_{\ell-1}/2}\right),$$

(where the inner product is taken over $\mathbb{R}^{S_{\ell-1}}$) with probability at least $1-\exp(-k/3)$.

Note also from the definition of $\mathrm{LG}_{\ell-1}$ that

$$\sum_{1\le i\le k}\left|\langle R_i,\sigma\rangle-\left\langle \overline{R}_i,\overline{\sigma}\right\rangle\right|\ge|\mathrm{LG}_{\ell-1}|n^{1/2-\beta_{\ell-1}}.$$

Since $k=\Theta(n)$, we arrive at

$$\mathbb{P}\Big(\big|\mathrm{LG}_{\ell-1}\big|=O\left(n^{1+\beta_{\ell-1}-\alpha_{\ell-1}/2}\right)\Big)\ge 1-\exp\left(-k/2\right).$$

Next, introduce the sets

$$\mathcal{S}(x)\triangleq\Big\{1\le i\le k:\langle R_i,\sigma\rangle<x\Big\}\qquad\text{and}\qquad\overline{\mathcal{S}}(x)\triangleq\Big\{1\le i\le k:\left\langle\overline{R_i},\overline{\sigma}\right\rangle<x\Big\}.$$

In particular by (3.124) w.p. at least $1-O\big(n^{-1/40+\epsilon}\big)$,

$$k_\ell-O\left(n^{\frac{79}{80}}\right)\le\big|\mathcal{S}(x_\ell)\big|,\ \big|\overline{\mathcal{S}}(x_\ell)\big|\le k_\ell+O\left(n^{\frac{79}{80}}\right).$$

We now record some useful set inclusion properties (each holding w.p. $1-O\big(n^{-\frac{1}{40}+\epsilon}\big)$, which is suppressed for convenience).

- We claim

$$\big|\mathcal{I}_\ell\cap\mathcal{S}(x_\ell)\big|\ge k_\ell-O\left(n^{\frac{79}{80}}\right)\quad\text{and}\quad\big|\overline{\mathcal{I}}_\ell\cap\overline{\mathcal{S}}(x_\ell)\big|\ge k_\ell-O\left(n^{\frac{79}{80}}\right).$$

  To see this, let $\overline{x_\ell}=\max_{i\in\mathcal{I}_\ell}\langle R_i,\sigma\rangle$. Note that if $\overline{x_\ell}\le x_\ell$, then $\mathcal{I}_\ell\subset\mathcal{S}\big(x_\ell\big)$, yielding the conclusion as $|\mathcal{I}_\ell|=k_\ell$. If $\overline{x_\ell}>x_\ell$, then $\mathcal{S}\big(x_\ell\big)\subset\mathcal{I}_\ell$, and we have the claim since $|\mathcal{S}\big(x_\ell\big)|\ge k_\ell-O\big(n^{\frac{79}{80}}\big)$. The same argument applies also to $\overline{\mathcal{S}}(x_\ell)$ and $\overline{\mathcal{I}}_\ell$.

- Next, observe that if $i\in\mathcal{S}\left(x_\ell-n^{1/2-\beta_{\ell-1}}\right)\setminus\mathrm{LG}_{\ell-1}$ then $i\in\overline{\mathcal{S}}(x_\ell)$. Likewise, if $i\in\overline{\mathcal{S}}(x_\ell)\setminus\mathrm{LG}_{\ell-1}$, then $i\in\mathcal{S}\left(x_\ell+n^{1/2-\beta_{\ell-1}}\right)$. That is, except for the indices in $\mathrm{LG}_{\ell-1}$, we have

$$\mathcal{S}\left(x_\ell-n^{1/2-\beta_{\ell-1}}\right)\subset\overline{\mathcal{S}}(x_\ell)\subset\mathcal{S}\left(x_\ell+n^{1/2-\beta_{\ell-1}}\right).$$

  Recall now the relation between k_ℓ and x_ℓ from (3.123). Using the fact $\Phi(\cdot)$ is $1-$Lipschitz, we obtain

$$\left|k_\ell-k\Phi\left(\frac{1}{\sqrt{S_{\ell-1}}}\left(x_\ell-n^{\frac{1}{2}-\beta_{\ell-1}}-\sum_{0\le s\le\ell-1}n_s\sqrt{\frac{2k_s}{\pi k^2}}\right)\right)\right|\le k\frac{n^{\frac{1}{2}-\beta_{\ell-1}}}{\sqrt{S_{\ell-1}}}=O\left(n^{1-\beta_{\ell-1}}\right),$$

  as $k=\Theta(n)$ and $\Theta(n)=n_0\le\sum_{0\le s\le\ell-1}n_s=S_{\ell-1}\le n$ and thus $S_{\ell-1}=\Theta(n)$.

Consequently,

$$\begin{aligned}\left|\mathcal{S}\left(x_\ell - n^{1/2-\beta_{\ell-1}}\right)\right| &\geq k_\ell - O(n^{79/80}) - O\left(n^{1-\beta_{\ell-1}}\right)\\ \left|\mathcal{S}\left(x_\ell + n^{1/2-\beta_{\ell-1}}\right)\right| &\leq k_\ell + O(n^{79/80}) + O\left(n^{1-\beta_{\ell-1}}\right).\end{aligned}$$

Finally, observe that any subset of

$$\mathcal{S}\left(x_\ell - n^{1/2-\beta_{\ell-1}}\right) \setminus \mathrm{LG}_{\ell-1}$$

of cardinality at least

$$k_\ell - O\left(n^{\frac{79}{80}}\right) - O\left(n^{1-\beta_{\ell-1}}\right) - |\mathrm{LG}_{\ell-1}|$$

is necessarily contained in $\mathcal{I}_\ell \cap \overline{\mathcal{I}}_\ell$. Thus, we arrive at

$$\begin{aligned}\left|\mathcal{I}_\ell \cap \overline{\mathcal{I}}_\ell\right| &\geq k_\ell - O\left(n^{\frac{79}{80}}\right) - O\left(n^{1-\beta_{\ell-1}}\right) - |\mathrm{LG}_{\ell-1}|\\ &\geq k_\ell - O\left(n^{\frac{79}{80}}\right) - O\left(n^{1-\beta_{\ell-1}}\right) - O\left(n^{1+\beta_{\ell-1}-\alpha_{\ell-1}/2}\right)\\ &\geq k_\ell - O\left(n^{1-\alpha_{\ell-1}/4}\right)\end{aligned}$$

using (3.126) and (3.95). Noticing that this process is valid with probability at least

$$1 - O\left(n^{-1/40+\epsilon}\right) - \exp(-k/3)$$

with $k = \Theta(n)$, the proof of Lemma 3.6.25 is complete. □

Index Sets Being Nearly Identical Implies Next Block Being Nearly Identical. Denote

$$\sigma\big(k:\ell\big) \triangleq (\sigma_i : k \leq i \leq \ell) \in \{-1,1\}^{\ell-k+1}.$$

The last auxiliary result we need is the following.

Lemma 3.6.26. *Suppose that the algorithm run for $T-1$ rounds generating $\sigma, \overline{\sigma} \in \{-1,1\}^{\sum_{0\leq t\leq T-1} n_t}$. Consider the inner products $\langle R_i, \sigma\rangle$, $1 \leq i \leq k$ (taken on $\mathbb{R}^{\sum_{0\leq t\leq T-1} n_t}$) and let $\mathcal{I}_T$ with $|\mathcal{I}_T| = k_T$ be such that $i \in \mathcal{I}_T$ iff $\langle R_i, \sigma\rangle$ is among k_T smallest inner products. Similarly, define the set $\overline{\mathcal{I}}_T$ with $|\overline{\mathcal{I}}_T| = k_T$ for the collection $\langle \overline{R}_i, \overline{\sigma}\rangle$, $1 \leq i \leq k$; and the random variable*

$$\mathcal{T} \triangleq \left|\mathcal{I}_T \cap \overline{\mathcal{I}}_T\right|.$$

Then, conditional on $\mathcal{T} = D$,

$$d_H\Big(\sigma\big(S_{T-1}+1:S_T\big), \overline{\sigma}\big(S_{T-1}+1:S_T\big)\Big) \stackrel{d}{=} \mathrm{Bin}\left(n_T, \frac{1}{2} - \frac{1}{\pi}\sin^{-1}\Big(D\cos(\tau)/k_T\Big)\right).$$

Proof of Lemma 3.6.26. For convenience, drop τ appearing in $\overline{\mathcal{M}}(\tau)$. Observe that

the randomness in $\mathcal{I}_T$ and $\overline{\mathcal{I}}_T$ are due to $\mathcal{M}_{[k]:[s_{T-1}]}$ and $\overline{\mathcal{M}}_{[k]:[s_{T-1}]}$, respectively (that is, due to first $n_0+n_1+\cdots+n_{T-1}$ columns of corresponding matrices). Having fixed $\mathcal{I}_T$ and $\overline{\mathcal{I}}_T$, note that the next n_T entries of σ and $\overline{\sigma}$ are obtained by running the majority algorithm on the submatrices

$$\mathcal{M}\Big(\mathcal{I}_T : [S_{T-1}+1, S_T]\Big) \qquad \text{and} \qquad \overline{\mathcal{M}}\Big(\overline{\mathcal{I}}_T : [S_{T-1}+1, S_T]\Big),$$

respectively. Now, condition on $|\mathcal{I} \cap \overline{\mathcal{I}}| = D$. Define variables $A, B, \overline{A}$, and $\overline{B}$ as follows:

$$A \triangleq \sum_{1\le i\le D} X_i, \qquad \overline{A} \triangleq \sum_{1\le i\le D} \big(\cos(\tau)X_i + \sin(\tau)Y_i\big);$$

and

$$B \triangleq \sum_{D+1\le i\le k_T} X_i, \qquad \overline{B} \triangleq \sum_{D+1\le i\le k_T} X'_i,$$

where X_i, X'_i, Y_i are i.i.d. $\mathcal{N}(0,1)$. It is then clear that

$$d_H\Big(\sigma\big(S_{T-1}+1 : S_T\big), \overline{\sigma}\big(S_{T-1}+1 : S_T\big)\Big) \stackrel{d}{=} \mathrm{Bin}(n_T, p),$$

where

$$p \triangleq \mathbb{P}\Big(\mathrm{sgn}(A+B) \neq \mathrm{sgn}(\overline{A}+\overline{B})\Big) = 2\mathbb{P}\Big(A+B \ge 0, \overline{A}+\overline{B} \le 0\Big)$$

by symmetry. Observe that

$$\mathbb{E}[A+B] = \mathbb{E}[\overline{A}+\overline{B}] = 0 \qquad \text{and} \qquad \mathbb{E}[(A+B)(\overline{A}+\overline{B})] = \mathbb{E}[A\overline{A}] = D\cos(\tau).$$

From here, applying Lemma 3.6.2 to bivariate standard normal variables $k_T^{-\frac{1}{2}}(A+B)$ and $-k_T^{-\frac{1}{2}}(\overline{A}+\overline{B})$, we conclude

$$p = 2\left(\frac{1}{4} + \frac{1}{2\pi}\sin^{-1}(-D\cos(\tau)/k_T)\right) = \frac{1}{2} - \frac{1}{\pi}\sin^{-1}(D\cos(\tau)/k_T),$$

where we used the fact $\sin^{-1}(\cdot)$ is an odd function, completing the proof of Lemma 3.6.26. □

Equipped with all necessary auxiliary tools, we now complete the proof of Proposition 3.6.18.

Proof of Proposition 3.6.18. Let $T \le c\log_{10}\log_{10} n$ for some $c > 0$ small enough, recall α_ℓ from (3.95) and J_ℓ from (3.96). Note that applying Lemma 3.6.19 we immediately obtain

$$|J_0| = d_H\Big(\sigma(1:n_0), \overline{\sigma}(1:n_0)\Big) \stackrel{d}{=} \mathrm{Bin}\left(n_0, \frac{\tau}{\pi}\right).$$

In particular, applying a Chernoff bound, and recalling $n_0 = \Theta(n)$,

$$|J_0| = O(n\tau) = O\left(n^{1-2\alpha_0}\right) < n^{1-\alpha_0}$$

with probability at least $1 - \exp\bigl(-\Omega\left(n^{1-2\alpha_0}\right)\bigr)$.

We now proceed by inducting on ℓ, where the base case, $\ell = 0$, has been verified above. Assume now that

$$\mathbb{P}\Bigl(|J_{\ell-1}| \le n^{1-\alpha_{\ell-1}}\Bigr) \ge 1 - p_{\ell-1}. \tag{3.127}$$

Using Lemma 3.6.25, we obtain

$$\mathbb{P}\Bigl(\bigl|\mathcal{I}_\ell \cap \overline{\mathcal{I}}_\ell\bigr| \ge k_\ell - O\left(n^{1-\alpha_{\ell-1}/4}\right)\Bigr) \ge 1 - O\left(n^{-1/40+\epsilon}\right). \tag{3.128}$$

Now, conditional on $|\mathcal{I}_\ell \cap \overline{\mathcal{I}}_\ell| = D$, Lemma 3.6.26 implies that

$$\begin{aligned} |J_\ell| - |J_{\ell-1}| &= d_H\Bigl(\sigma(S_{\ell-1}+1 : S_\ell), \overline{\sigma}(S_{\ell-1}+1 : S_\ell)\Bigr) \\ &\stackrel{d}{=} \mathrm{Bin}\Bigl(n_\ell, \frac{1}{2} - \frac{1}{\pi}\sin^{-1}\Bigl(D\cos(\tau)/k_\ell\Bigr)\Bigr), \end{aligned} \tag{3.129}$$

where $S_T \triangleq \sum_{0 \le j \le T} n_j$ for any $T \in \mathbb{N}$.

Now, recall that $k_\ell \ge n f_\ell^3 - 1$ for $f_\ell = 10^{-2^\ell}$ (see (3.10) and (3.11)). Hence,

$$\frac{k_\ell - O\left(n^{1-\alpha_{\ell-1}/4}\right)}{k_\ell} = 1 - O\left(\frac{n^{1-\alpha_{\ell-1}/4}}{k_\ell}\right) \ge 1 - O\left(\frac{n^{-\alpha_{\ell-1}/4}}{10^{-3\cdot 2^\ell}}\right).$$

We now claim

$$1 - O\left(\frac{n^{-\alpha_{\ell-1}/4}}{10^{-3\cdot 2^\ell}}\right) \ge 1 - O\left(n^{-\alpha_{\ell-1}/4.5}\right)$$

for all large enough n, provided $c > 0$ is small enough. Ignoring the absolute constants, it suffices to verify

$$n^{-\frac{\alpha_{\ell-1}}{4.5}} \ge 10^{3\cdot 2^\ell} \cdot n^{-\frac{\alpha_{\ell-1}}{4}} \iff n^{\frac{\alpha_{\ell-1}}{36}} \ge 10^{3\cdot 2^\ell} \iff \frac{1}{36}\log_{10} n \cdot \alpha_{\ell-1} \ge 3 \cdot 2^\ell.$$

Recall that $\alpha_{\ell-1} = 0.4 \cdot 10^{-\ell}$. Thus, it suffices to verify

$$\frac{1}{270}\log_{10} n \ge 20^\ell.$$

Recalling $\ell \le c \log_{10}\log_{10} n$ for $c > 0$ small enough, we have

$$20^\ell \le 20^{c\log_{10}\log_{10} n} = (\log_{10} n)^{c''}, \qquad \text{where} \qquad c'' = c\log_{10} 20.$$

Finally, provided $c'' < 1$, we indeed have

$$\frac{1}{270}\log_{10} n \geq \log_{10}^{c''} n,$$

thus the claim.

We next employ this claim and the inequality, $\cos(\tau) \geq 1 - \tau^2/2$, which is valid for all τ. Using (3.128) it follows that there is an event of probability at least $1 - O\left(n^{-1/40+\epsilon}\right)$ such that on this event, $|J_\ell| - |J_{\ell-1}|$ is stochastically dominated by the binomial random variable

$$\mathrm{Bin}\left(n_\ell, \frac{1}{2} - \frac{1}{\pi}\sin^{-1}\Big(\big(1 - \Theta\big(n^{-\alpha_{\ell-1}/4.5}\big)\big)\big(1 - \tau^2/2\big)\Big)\right). \tag{3.130}$$

Using Taylor series $\sin^{-1}(1-x) = \frac{\pi}{2} - \sqrt{2x} + o(\sqrt{x})$, we obtain that the binomial variable appearing in (3.130) is stochastically dominated further by a binomial random variable $\mathrm{Bin}(n_\ell, q_\ell)$ where $q_\ell = \Theta(n^{-\alpha_{\ell-1}/9})$. Since $n_\ell \leq n$ per (3.13), we also have by Chernoff bound

$$\mathbb{P}\Big(\mathrm{Bin}\big(n_\ell, q_\ell\big) = O\left(n^{1-\alpha_{\ell-1}/9}\right)\Big) \geq 1 - \exp\Big(-\Omega\Big(n^{1-\alpha_{\ell-1}/9}\Big)\Big). \tag{3.131}$$

Combining (3.127), (3.129) and (3.131) via a union bound, we conclude that

$$\mathbb{P}\Big(|J_\ell| \leq n^{1-\alpha_{\ell-1}} + O\left(n^{1-\alpha_{\ell-1}/9}\right)\Big) \geq 1 - p_{\ell-1} - O\left(n^{-1/40+\epsilon}\right) - \exp\Big(-\Omega\Big(n^{1-\alpha_{\ell-1}/9}\Big)\Big). \tag{3.132}$$

We set

$$p_\ell \triangleq p_{\ell-1} + O\left(n^{-1/12+\epsilon}\right) + \exp\Big(-\Omega\Big(n^{1-\alpha_{\ell-1}/9}\Big)\Big) = p_{\ell-1} + O\left(n^{-1/40+\epsilon}\right). \tag{3.133}$$

We now ensure $|J_\ell| \leq n^{1-\alpha_\ell}$ w.h.p., where $\alpha_\ell = \alpha_{\ell-1}/10$. To that end, we first claim

$$n^{1-\alpha_{\ell-1}} + O\left(n^{1-\alpha_{\ell-1}/9}\right) \leq n^{1-\alpha_{\ell-1}} + n^{1-\alpha_{\ell-1}/9.5}. \tag{3.134}$$

To prove this, it suffices to verify $n^{\frac{\alpha_{\ell-1}}{9} - \frac{\alpha_{\ell-1}}{9.5}} = n^{\frac{\alpha_{\ell-1}}{171}} = \omega(1)$. Using the fact $\alpha_{\ell-1} = 0.1 \cdot 10^{-\ell}$ per (3.95) and $\ell \leq L \leq c\log_{10}\log_{10} n$ in the setting of Proposition 3.6.18, we have

$$n^{\frac{\alpha_{\ell-1}}{171}} = \exp_{10}\left(\frac{\alpha_{\ell-1}}{171}\log_{10} n\right) \geq \exp_{10}\left(\frac{10^{-c\log_{10}\log_{10} n}}{1710}\log_{10} n\right) = \exp_{10}\left(\frac{1}{1710}\big(\log_{10} n\big)^{1-c}\right),$$

which is indeed $\omega(1)$ if $c < 1$. This yields (3.134). Hence, it suffices to show $x + x^{1/9.5} \leq x^{1/10}$ for $x = n^{-\alpha_{\ell-1}}$. Now, assume $\ell \leq L \leq c\log_{10}\log_{10} n$ (where L is the number of

steps analyzed) for $c > 0$ small enough. Then

$$\begin{aligned} x &= \exp_{10}\Big(-\alpha_{\ell-1}\log_{10} n\Big) \\ &= \exp_{10}\Big(-0.04\cdot 10^{-(\ell-1)}\cdot \log_{10} n\Big) \\ &\le \exp_{10}\Big(-0.4\cdot 10^{-c\log_{10}\log_{10} n}\cdot \log_{10} n\Big) \\ &= \exp_{10}\Big(-0.4\cdot (\log_{10} n)^{1-c}\Big). \end{aligned}$$

We have $t + t^{1/9.5} < t^{1/10}$ for t sufficiently small (e.g. $0 \le t < 0.01$ suffices). Hence, provided $c < 1$, it is the case that for n sufficiently large, $x + x^{1/9.5} \le x^{1/10}$ for $x = n^{-\alpha_{\ell-1}}$ and for any $\ell \le c\log_{10}\log_{10} n$. Thus,

$$\mathbb{P}\Big(|J_\ell| \le n^{1-\alpha_\ell}\Big) \ge 1 - p_\ell \qquad \text{where} \qquad p_\ell = p_{\ell-1} + O\left(n^{-1/40+\epsilon}\right). \tag{3.135}$$

Since the inductive step from $\ell - 1 \to \ell$ (more concretely from (3.127) to (3.135)), increases the probability by $O\left(n^{-1/40+\epsilon}\right)$ and the whole process runs $\log_{10}\log_{10} n$ rounds, we complete the proof of Proposition 3.6.18. □

Proof of Theorem 3.3.8

Having established Proposition 3.6.18, we now finish the proof of Theorem 3.3.8. For simplicity, we omit floor/ceiling operators whenever convenient.

Proof of Theorem 3.3.8. Let $L = c\log_{10}\log_{10} n$ be as in Proposition 3.6.18 for $c > 0$ small enough. We now show that it suffices to analyze L rounds as opposed to the full implementation of $C\log_{10}\log_{10} n$ rounds, where $C > c > 0$. In particular, we claim

$$\sum_{c\log_{10}\log_{10} n \le j \le C\log_{10}\log_{10} n} n_j = o(n), \tag{3.136}$$

for any constant $c > 0$. For $N \triangleq C\log_{10}\log_{10} n$, the number of rounds, and $0 \le j \le N$; recall f_j from (3.10), n_j from (3.13), and k_j from (3.11). Applying now a

telescoping argument,

$$\begin{aligned}
\sum_{c\log_{10}\log_{10} n\le j\le C\log_{10}\log_{10} n} n_j &= \sum_{c\log_{10}\log_{10} n\le j\le C\log_{10}\log_{10} n}\left(\left\lfloor \frac{n}{A}\sum_{0\le i\le j} f_i\right\rfloor - \left\lfloor \frac{n}{A}\sum_{0\le i\le j-1} f_i\right\rfloor\right)\\
&= \left\lfloor \frac{n}{A}\underbrace{\sum_{0\le i\le C\log_{10}\log_{10} n} f_i}_{=A}\right\rfloor - \left\lfloor \frac{n}{A}\sum_{0\le i\le c\log_{10}\log_{10} n-1} f_i\right\rfloor\\
&\le n\left(1-\frac{1}{A}\sum_{0\le i\le c\log_{10}\log_{10} n} f_i\right)+1\\
&\le \frac{n}{A}\sum_{c\log_{10}\log_{10} n+1\le i\le C\log_{10}\log_{10} n} f_i + 1\\
&= O\left(n\log_{10}\log_{10} n\cdot 10^{-2^{c\log_{10}\log_{10} n}}\right)\\
&= O\left(n\cdot\log_{10}\log_{10} n\cdot 10^{-(\log_{10} n)^{c'}}\right)
\end{aligned}$$

for $c' = c\log_{10} 2 < 1$. We now verify

$$n\cdot\log_{10}\log_{10} n\cdot 10^{-(\log_{10} n)^{c'}} = o(n) \iff \log_{10}\log_{10} n\cdot 10^{-(\log_{10} n)^{c'}} = o(1).$$

Indeed,

$$\log_{10}\log_{10} n\cdot 10^{-(\log_{10} n)^{c'}} = \exp_{10}\left(\log_{10}\log_{10}\log_{10} n - \left(\log_{10} n\right)^{c'}\right) = o(1)$$

for any $c > 0$. This yields (3.136).

Finally, combining (3.136) with Proposition 3.6.18, we complete the proof of Theorem 3.3.8. □

3.6.7 Proof of Theorem 3.5.2

Proof of Theorem 3.5.2. The proof is quite similar to that of Theorem 3.2.4, hence we only point out the necessary modification. Note that the probability term that one considers (cf. Lemma 3.6.8) is

$$\mathbb{P}\Big[\big|\langle\sigma^{(i)}, X\rangle\big| \le \kappa\sqrt{n}, 1\le i\le m\Big]^{\alpha n},$$

where $X = (X_1, \ldots, X_n)$ has i.i.d. entries with $X_i \sim \mathcal{D}$. To apply Theorem 3.5.1, set

$$Y_j \triangleq \begin{pmatrix} \frac{1}{\sqrt{n}}\sigma_1(j)X_j \\ \frac{1}{\sqrt{n}}\sigma_2(j)X_j \\ \vdots \\ \frac{1}{\sqrt{n}}\sigma_m(j)X_j \end{pmatrix} \in \mathbb{R}^m. \tag{3.137}$$

Indeed $Y_j \in \mathbb{R}^m$, $1 \le j \le n$, is a collection of independent centered random vectors, and

$$\Big\{ \big|\big\langle \sigma^{(i)}, X \big\rangle\big| \le \kappa\sqrt{n}, 1 \le i \le m \Big\} = \big\{ S \in U \big\},$$

for $S = \sum_{j \le n} Y_j$ and $U = [-\kappa, \kappa]^n$. Furthermore, $\Sigma \triangleq \mathrm{Cov}(S) \in \mathbb{R}^{m \times m}$ is such that (a) $\Sigma_{ii} = 1$ for $1 \le i \le m$; and (b) $\Sigma_{ij} = \Sigma_{ji} = n^{-1}\big\langle \sigma^{(i)}, \sigma^{(j)} \big\rangle$ for $1 \le i < j \le m$. Next, observe that since $\Sigma \in \mathbb{R}^{m \times m}$ with $m = O_n(1)$, it follows that

$$\begin{aligned} \mathbb{E}\Big[\big\| \Sigma^{-\frac{1}{2}} Y_j \big\|_2^3 \Big] &\le \big\| \Sigma^{-\frac{1}{2}} \big\|^3 \mathbb{E}\Big[\big\| Y_j \big\|_2^3 \Big] \\ &= \big\| \Sigma^{-\frac{1}{2}} \big\|^3 \mathbb{E}\left[\Big(\frac{m}{n} X_j^2 \Big)^{3/2} \right] \\ &= O\big(n^{-\frac{3}{2}} \big). \end{aligned}$$

Applying now Theorem 3.5.1, we obtain

$$\mathbb{P}\big[S \in U \big] \le \mathbb{P}\big[Z \in U \big] + O\big(n^{-\frac{1}{2}} \big).$$

Consequently, for $Z \sim \mathcal{N}(0, \Sigma)$,

$$\begin{aligned} \mathbb{P}\big[S \in U \big]^{\alpha n} &\le \mathbb{P}\big[Z \in U \big]^{\alpha n} \Big(1 + O(n^{-\frac{1}{2}}) \Big)^{\alpha n} \\ &= \mathbb{P}\big[Z \in U \big]^{\alpha n} \exp\Big(\alpha n \ln\Big(1 + O\big(n^{-1/2} \big) \Big) \Big) \\ &= \mathbb{P}\big[Z \in U \big]^{\alpha n} \exp\Big(\Theta\big(\sqrt{n} \big) \Big), \end{aligned}$$

where in the last step we used the Taylor expansion, $\ln(1 - x) = -x + o(x)$ as $x \to 0$. Modifying Lemma 3.6.8 by taking the extra $e^{\Theta(\sqrt{n})}$ factor into account and then applying the *first moment method*, we establish Theorem 3.5.2. $\square$

Chapter 4

Computing the Partition Function of the Sherrington-Kirkpatrick Model is Hard on Average

4.1 Introduction

The subject of this chapter is the algorithmic hardness of the problem of exactly computing the partition function associated with the Sherrington-Kirkpatrick (SK) model of spin glasses, a mean field model that was first introduced by Sherrington and Kirkpatrick in 1975 [255], to propose a solvable model for the 'spin-glass' phase, an unusual magnetic behaviour predicted to occur in spatially random physical systems. The model is as follows. Fix a positive integer n, and consider n sites $i \in \{1, 2, \ldots, n\}$, a naming motivated from a site of a magnet. To each site i, assign a spin, $\sigma_i \in \{-1, 1\}$, and define the energy Hamiltonian $H(\boldsymbol{\sigma})$ for this spin configuration $\boldsymbol{\sigma} = (\sigma_i : 1 \leq i \leq n) \in \{-1, 1\}^n$ via $H(\boldsymbol{\sigma}) = \frac{\beta}{\sqrt{n}} \sum_{1 \leq i < j \leq n} J_{ij}\sigma_i\sigma_j$, where the parameters $\mathbf{J} = (J_{ij} : 1 \leq i < j \leq n) \in \mathbb{R}^{n(n-1)/2}$ are called spin-spin interactions (or shortly, couplings), and the parameter β is called the inverse temperature. The associated partition function is given by, $Z(\mathbf{J}, \beta) = \sum_{\boldsymbol{\sigma} \in \{-1,1\}^n} \exp\left(-\frac{\beta}{\sqrt{n}} \sum_{1 \leq i < j \leq n} J_{ij}\sigma_i\sigma_j\right)$. The SK model corresponds to the case, where the couplings J_{ij} are iid standard normal; and the partition function, $Z(\mathbf{J}, \beta)$ carries useful information about the underlying physical system [59]. The SK model is a *mean-field* model of spin glasses, namely the interaction between any two distinct sites, $1 \leq i < j \leq n$, is modeled with random coupling parameters J_{ij}, which do not depend on the spatial location of i and j. The rationale behind the scaling $\sqrt{n}$ is to ensure that the average energy per spin is roughly independent of n, and consequently, the free energy limit, $\lim_n n^{-1} \log Z(\mathbf{J}, \beta)$ is non-trivial. Namely, the limits

$$\lim_{n\to\infty} \frac{1}{n} \left(\max_{\boldsymbol{\sigma} \in \mathcal{B}_n} \frac{1}{\sqrt{n}} \sum_{1 \leq i < j \leq n} J_{ij}\sigma_i\sigma_j \right) \quad \text{and} \quad \lim_{n\to\infty} \frac{\log Z(\boldsymbol{J}, \beta)}{n}$$

exist and they are non-trivial: they remain bounded away from zero as $n \to \infty$. The computation of the free energy limit is a long journey on which we now elaborate. Without proving the existence of the limit, Parisi put forth a formula for the limiting value of the free energy per spin in his celebrated paper [232] using the non-rigorous replica method. The very existence of the limit was established rigorously much later by Guerra and Toninelli [155] using a simple, though very clever, interpolation argument. Later, Talagrand rigorously verified in the breakthrough paper [272] that Parisi's prediction is indeed correct. Parisi's formula for the free energy limit was extended to more general spin glass models by Panchenko [228, 230].

Despite the simplicity of its formulation, it turns out that the SK model is highly non-trivial to study, and analyzing the behaviour of a more elaborate model (such as a model where the spatial positions of the sites are incorporated, by modelling them as the vertices of $\mathbb{Z}^2$ and the couplings are modified to be position-dependent) is really difficult. For a more detailed discussion on these and related issues, see the monographs by Panchenko [229] and Talagrand [273].

In addition to its relevance in statistical physics and in spin glass theory, it is worth mentioning that the problem of computing the partition function has ties to Bayesian inference and machine learning as well, as demonstrated by the following example [31]. Consider the Rademacher spiked Wigner model where i) a signal σ^* is drawn uniformly at random from $\{-1, 1\}^n$, and ii) the learner sees the measurement

$$Y = \frac{\lambda}{n}\sigma^*(\sigma^*)^T + \frac{1}{\sqrt{n}}W \in \mathbb{R}^{n\times n}.$$

Here, $W \in \mathbb{R}^{n\times n}$ is a symmetric random matrix, whose upper triangular part consists of i.i.d. standard normal entries (in this case W is said to be a "GOE matrix", where GOE stands for the Gaussian orthogonal ensemble) and the parameter $\lambda > 0$ is called the signal-to-noise ratio (SNR). The goal of the learner is to recover σ^* (up to a sign flip). It is then natural to study the posterior distribution $\mathbb{P}[\sigma|Y]$ of σ, having measured Y. After some algebra [31, p.5], we arrive at

$$\mathbb{P}[\sigma|Y] \propto \prod_{1\le i<j\le n} \exp\left(\lambda Y_{ij}\sigma_i\sigma_j\right) = \exp\left(\lambda \sum_{1\le i<j\le n} Y_{ij}\sigma_i\sigma_j\right).$$

Ignoring the scaling $-n^{-\frac{1}{2}}$ present in our case, it is evident that the observation Y defines here the couplings for a Hamiltonian (with no external field), where the SNR parameter λ plays the role of the inverse temperature β. In particular, the posterior distribution $\mathbb{P}[\sigma|Y]$ enjoys the spin glass equations. Thus in this case, the partition function has a natural interpretation of being the normalizer for the associated **Gibbs distribution**.

In addition, the SK model also captures the limiting behaviour of much studied models in computer science, including the MaxCUT [87] and the MAXSAT [231].

Having mentioned the relevance of the problem of computing the partition function in statistical physics, in inference/machine learning, and in theoretical computer science; we now proceed with the settings we consider here in more details.

In the first part of this work, we focus on the SK model with the (random) external field, which was studied by Talagrand [273] (equation 1.61 therein), namely, the model, where the energy Hamiltonian is given by,

$$H(\boldsymbol{\sigma}) = \frac{\beta}{\sqrt{n}} \sum_{1 \le i < j \le n} J_{ij}\sigma_i\sigma_j + \sum_{i=1}^{n} A_i\sigma_i. \tag{4.1}$$

Here, the iid standard normal random variables $\mathbf{J} = (J_{ij} : 1 \le i < j \le n) \in \mathbb{R}^{n(n-1)/2}$ are the couplings, and the independent zero-mean normal random variables, $\mathbf{A} = (A_i : i \in [n]) \in \mathbb{R}^n$ incorporate the external field contribution. To address this model we study the following equivalent model with the energy Hamiltonian:

$$H(\boldsymbol{\sigma}) = \frac{\beta}{\sqrt{n}} \sum_{1 \le i < j \le n} J_{ij}\sigma_i\sigma_j + \sum_{i=1}^{n} B_i\sigma_i - \sum_{i=1}^{n} C_i\sigma_i, \tag{4.2}$$

where $\mathbf{J} = (J_{ij} : 1 \le i < j \le n) \in \mathbb{R}^{n(n-1)/2}$ are the couplings as above, $\mathbf{B} = (B_i : i \in [n]) \in \mathbb{R}^n$ and $\mathbf{C} = (C_i : i \in [n]) \in \mathbb{R}^n$ are independent zero-mean normal random variables, which we still refer to as external field components. Observe that, if $\mathcal{A}_1$ is an oracle, which, for input $(\mathbf{J}, \mathbf{A})$, computes the partition function for the model whose Hamiltonian is given by (4.1), then $\mathcal{A}_1$ with input $(\mathbf{J}, \mathbf{B} - \mathbf{C})$ computes the partition function of the model whose Hamiltonian is given by (4.2). Similarly, if $\mathcal{A}_2$ is an oracle, which, for input $(\mathbf{J}, \mathbf{B}, \mathbf{C})$ computes the partition function of the model in (4.2), then $\mathcal{A}_2$ with input, $(\mathbf{J}, \frac{\mathbf{A}+\mathbf{G}}{\mathbf{2}}, \frac{\mathbf{G}-\mathbf{A}}{\mathbf{2}})$, where $\mathbf{G} = (G_i : i \in [n])$ is an iid copy of the random vector $\mathbf{A} = (A_i : i \in [n])$, computes the partition function of the model in (4.1), recalling that if A_i and G_i are iid Gaussian random variables, then $A_i + G_i$ and $A_i - G_i$ are independent. In spite of being equivalent, the model in (4.2), however, is more convenient to work with, in particular, for establishing a a certain downward self-recursive formula which expresses the partition function of an n-spin SK model as a weighted sum of the partition functions of two $(n-1)$-spin SK models, with properly adjusted external field components.

The algorithmic problem is the problem of computing the partition function $Z(\mathbf{J}, \mathbf{B}, \mathbf{C})$ associated to the modified model in (4.2), when $(\mathbf{J}, \mathbf{B}, \mathbf{C}) \in \mathbb{R}^{n(n-1)/2+2n}$ is given as a (random) input. The (worst-case) algorithmic problem of computing $Z(\mathbf{J}, \mathbf{B}, \mathbf{C})$ for an arbitrary input $(\mathbf{J}, \mathbf{B}, \mathbf{C})$ is known to be #P-hard for a much broader class of statistical physics models and associated partition functions, see e.g. [37] and [175]. On the other hand, the classical reduction techniques that are used for establishing worst-case hardness do not seem to transfer to the problems with random inputs. The subject of this work is the case of Gaussian random inputs, $(\mathbf{J}, \mathbf{B}, \mathbf{C})$. The computational model that we adopt in the first part of the work is the finite-precision arithmetic for which the real-valued vector $(\mathbf{J}, \mathbf{B}, \mathbf{C})$ cannot be used as a formal algorithmic input. In order to handle this issue, we consider a model, where the algorithm designer first selects a level N of digital precision, and the values of J_{ij}, B_i, C_i, or more concretely, $\widehat{J}_{ij} = \exp(\frac{\beta J_{ij}}{\sqrt{n}})$, $\widehat{B}_i = \exp(B_i)$, and $\widehat{C}_i = \exp(C_i)$ are computed, up to this selected level N of digital precision: $\widehat{J}_{ij}^{[N]}$, $\widehat{B}_i^{[N]}$, and $\widehat{C}_i^{[N]}$, where

$x^{[N]} = 2^{-N}\lfloor 2^N x \rfloor$. The task of the algorithm designer is to exactly compute the partition function, associated with the input $(\widehat{J}_{ij}^{[N]} : 1 \leq i < j \leq n)$, $(\widehat{B}_i^{[N]} : 1 \leq i \leq n)$, and $(\widehat{C}_i^{[N]} : 1 \leq i \leq n)$ in polynomial (in n) time.

Under the aforementioned assumptions, our main result is as follows. Let $k > 0$ be any arbitrary constant. If there exists a polynomial time algorithm, which computes the partition function with input $(\widehat{J}_{ij}^{[N]} : 1 \leq i < j \leq n)$, $(\widehat{B}_i^{[N]} : 1 \leq i \leq n)$, and $(\widehat{C}_i^{[N]} : 1 \leq i \leq n)$ exactly with probability at least $1/n^k$, then $P = \#P$. Here, the probability is taken with respect to the randomness of $(\mathbf{J}, \mathbf{B}, \mathbf{C})$. To the best of our knowledge, this is the first result establishing formal algorithmic hardness of a computational problem arising in the field of spin glasses.

The approach we pursue here aims at capturing a *worst-case* to *average-case* reduction, and is similar to establishing the average-case hardness of other problems involving counting, such as the problem of computing the permanent of a matrix modulo p, with entries chosen independently and uniformly over finite field $\mathbb{Z}_p$. One recent application of this idea includes also the problem of counting cliques in Erdös-Rényi hypergraphs [53].

Lipton observed in [208] that, for a suitably chosen prime p, the permanent of a matrix can be expressed as a univariate polynomial, generated using integer multiples of a random uniform input. Hence, provided this polynomial can be recovered, the permanent of any *arbitrary* matrix can be computed. Therefore, the average-case hardness of computing the permanent of a matrix modulo p equals the worst-case hardness of the same problem, which is known to be #P-hard. Lipton proves his result, by assuming there exists an algorithm, which correctly computes the permanent for at least $1 - O(1/n)$ fraction of matrices over $\mathbb{Z}_p^{n \times n}$. Subsequent research weakened this assumption to the existence of an algorithm with constant probability of success [109], and finally, to the existence of an algorithm with inverse polynomial probability $(1/n^{O(1)})$ of success, Cai et al. [67], a regime, which is also our focus. The proof technique that we follow is similar to that of Cai et al. [67], and is built upon earlier ideas from Gemmell and Sudan [146], Feige and Lund [109], and Sudan [270].

More specifically, the argument of Cai et al. [67] is as follows. The permanent of a given matrix $M \in \mathbb{Z}_p^{n \times n}$ equals, via Laplace expansion, a weighted sum of the permanents of n minors $M_{11}, M_{21}, \ldots, M_{n1}$ of M, each of dimension $n - 1$. Then a certain matrix polynomial is constructed, whose value at k is equal to M_{k1}, by incorporating two random matrices, independently generated from the uniform distribution on $\mathbb{Z}_p^{(n-1)\times(n-1)}$. The permanent of this matrix polynomial is a univariate polynomial over a finite field with a known upper bound on its degree, and the problem boils down to recovering this polynomial from a list of pairs of numbers intersecting the graph of the polynomial at sufficiently many points. This, in fact, is a standard problem in coding theory, and the recovery of this polynomial is achieved by a list-decoding algorithm by Sudan [270], which is an improved version of Berlekamp-Welch decoder.

The method that we use follows the proof technique of Cai et al. [67], with several additional modifications. First, to avoid dealing with correlated random inputs, we reduce the problem of computing the partition function of the model in (4.2) to computing the partition function of a different object, where the underlying cuts and

polarities induced by the spin assignment $\boldsymbol{\sigma} \in \{-1,1\}^n$ are incorporated. Second, a downward self-recursion formula for computing the partition function, analogous to Laplace expansion for permanent, is established; and this is the rationale for using the aforementioned equivalent model whose Hamiltonian is given by (4.2). This is achieved by recursing downward with respect to the sign of σ_n, and expressing the partition function of an n-spin system, with a weighted sum of the partition functions of two $(n-1)$-spin systems, with appropriately adjusted external field components. Third, recalling that, we are interested in the case of random Gaussian inputs, we establish a probabilistic coupling between truncated version of log-normal distribution, and uniform distribution modulo a large prime p. Towards this goal, we establish that the log-normal distribution is "sufficiently" Lipschitz in a small interval and near-uniform modulo p. Finally, we also need to connect modulo p computation to the exact computation of the partition function, in the sense defined above, i.e., truncating the inputs up to a certain level N of digital precision, and computing the associated partition function. This is achieved by using a standard Chinese remaindering argument: Take prime numbers $p_1, \ldots, p_K$, compute $Z \pmod{p_i}$, for every i, and use this information to compute compute $Z \pmod{P}$ where $P = \prod_{k=1}^{K} p_k$, via Chinese remaindering. Provided $P > Z$, $Z \pmod{p}$ is precisely Z. The existence of sufficiently many such primes of appropriate size that we can work with is justified through the prime number theorem.

In the second part of this work, we focus on the same problem without the external field component, but this time under the *real-valued* computational model. We recall the model for convenience. First, generate iid standard normal random variables, J_{ij}; and let the elements of the sequence $\mathbf{J} = (J_{ij} : 1 \leq i < j \leq n) \in \mathbb{R}^{\frac{n(n-1)}{2}}$ be the couplings. For each spin configuration $\boldsymbol{\sigma} = (\sigma_i : 1 \leq i \leq n) \in \{-1,1\}^n$, define the associated energy Hamiltonian $H(\boldsymbol{\sigma}) = \sum_{i<j} J_{ij}\sigma_i\sigma_j$. The algorithmic question of interest is the exact computation of the associated partition function, namely the object,

$$Z(\mathbf{J}) = \sum_{\boldsymbol{\sigma}\in\{-1,1\}^n} \exp\left(-\sum_{1\leq i<j\leq n} J_{ij}\sigma_i\sigma_j\right) = \sum_{\boldsymbol{\sigma}\in\{-1,1\}^n} \exp(-H(\boldsymbol{\sigma})),$$

using the real-valued computational model, i.e. a computational engine operating over real-valued inputs, appropriately defined, as opposed to the previous setting, where the computational engine performs floating point operations. The input vector, namely the vector of real-valued couplings $\mathbf{J} \in \mathbb{R}^{n(n-1)/2}$, is given as a random input. Albeit the usual definition of the partition function involves also the inverse temperature parameter β, and a normalization factor by $\sqrt{n}$; we suppress these in order to keep the discussion simple.

The main result towards this direction is as follows. If there exists a polynomial time algorithm, which computes the partition function exactly with probability at least $3/4 + 1/\mathrm{poly}(n)$ under real-valued computational model, then $P = \#P$. Similar to the previous setting, the probability here is taken with respect to the randomness in the input of the algorithm, namely, with respect to the distribution of $\mathbf{J}$.

The techniques of the previous setting (finite-precision arithmetic) do not, however, transform to real-valued computational model, since the finite field structure $\mathbb{Z}_p$ utilized for the proof is lost, upon passing to real-valued computation model. We bypass this obstacle by building on an argument of Aaronson and Arkhipov [3], where they established the average-case hardness of the exact computation of the permanent of a random matrix with iid Gaussian entries. Informally, their argument is as follows. The basis is the $\#P$-hardness of exactly computing the permanent for arbitrary matrices. Suppose, for the sake of the contradiction, that a polynomial-time algorithm $\mathcal{A}$ exactly computing the permanent (with a certain probability of success) exists. Consider a "natural interpolation" $X(t) = tX + (1-t)Y$, $t \in [0,1]$, between a worst-case input $X \in \{0,1\}^{n\times n}$ and a random input $Y \in \mathbb{R}^{n\times n}$. Define the univariate polynomial $q(t)$ to be the permanent of $X(t)$. It then follows that $\deg(q) \leq n$ and $q(1)$ is the permanent of the worst-case input X. In particular, the permanent of X can be computed (in polynomial time) provided q can be reconstructed (in polynomial time). $\mathcal{A}$ is then used to create a list of sufficiently many noisy samples of $q(\cdot)$. With this, the problem boils down to recovering a low-degree univariate polynomial q from a list of its noisy samples. This is a natural problem in coding theory, and one can efficiently recover q using the Berlekamp-Welch decoder.

We close this section with the set of notational convention. The set of integers and positive integers are respectively denoted by $\mathbb{Z}$ and $\mathbb{Z}^+$. The set, $\{1,2,\ldots,n\}$ is denoted by $[n]$, and the set $\{0,1,\ldots,p-1\}$, namely the set of all residues modulo p, is denoted by $\mathbb{Z}_p$. Given a real number x, the largest integer not exceeding x is denoted by $\lfloor x \rfloor$. We say $a \equiv b \pmod{p}$, if p divides $a-b$, abbreviated as $p \mid a-b$. Given an $x > 0$, $\log x$ denotes logarithm of x, base 2. Given a (finite) set S, denote the number of elements (i.e., the cardinality) of S by $|S|$. Given a finite field $\mathbb{F}$, denote by $\mathbb{F}[x]$ the set of all (finite-degree) polynomials, whose coefficients are from $\mathbb{F}$. Namely, $f \in \mathbb{F}[x]$ if there is a positive integer n, and $a_0,\ldots,a_n \in \mathbb{F}$, such that for every $x \in \mathbb{F}$, $f(x) = \sum_{k=0}^{n} a_k x^k$. The degree of $f \in \mathbb{F}[x]$ is, $\deg(f) = \max\{0 \leq k \leq n : a_k \neq 0\}$. For two random variables X and Y, the total variation distance between (the distribution functions of) X and Y is denoted by $d_{TV}(X,Y)$. For any given vector $v \in \mathbb{R}^d$, we denote by $\|v\|$ the Euclidean norm of v, that is, $\sqrt{\sum_{i=1}^{d} v_i^2}$. $\Theta(\cdot), O(\cdot)$, $o(\cdot)$, and $\Omega(\cdot)$ are standard (asymptotic) order notations for comparing the growth of two sequences. Finally, we use the words oracle and algorithm interchangeably in the sequel, and denote them by $\mathcal{O}$ and $\mathcal{A}$. These objects will be assumed to exist for the sake of proof purposes.

4.2 Average-Case Hardness under Finite-Precision Arithmetic

4.2.1 Model and the Main Result

Our focus is on computing the partition function of the model, whose Hamiltonian for a given spin configuration $\boldsymbol{\sigma} \in \{-1,1\}^n$ at inverse temperature β is given by:

$$H(\boldsymbol{\sigma}) = \frac{\beta}{\sqrt{n}} \sum_{1 \le i < j \le n} J_{ij}\sigma_i\sigma_j + \sum_{i=1}^{n} B_i\sigma_i - \sum_{i=1}^{n} C_i\sigma_i,$$

where the random variables $\mathbf{J} = (J_{ij} : 1 \le i < j \le n) \in \mathbb{R}^{n(n-1)/2}$ are the couplings; and the random variables $\mathbf{B} = (B_i : i \in [n]) \in \mathbb{R}^n$, and $\mathbf{C} = (C_i : i \in [n]) \in \mathbb{R}^n$ are the external field components. For simplicity, we study the case, where $(\mathbf{J}, \mathbf{B}, \mathbf{C})$ consists of i.i.d. standard normal entries.

While our focus is on the case where the vector $(\mathbf{J}, \mathbf{B}, \mathbf{C})$ consists of i.i.d. standard normal entries, it is worth mentioning the following. We anticipate that our analysis will still remain valid under the following, more general, setting. Suppose J_{ij}, $1 \le i < j \le n$; B_i, $1 \le i \le n$; and C_i, $1 \le i \le n$ are independent normal random variables with zero-mean and possibly different variances, where the variances are strictly positive and are not too small (e.g., they are of order at least $\frac{1}{n^{O(1)}}$). In particular, variances are allowed to be distinct even within each $\mathbf{J}$, $\mathbf{B}$, and $\mathbf{C}$ as long as they are strictly positive and not too small: for instance, J_{ij} need not have the same variance across $1 \le i < j \le n$. In this more general setting, we still anticipate that our techniques and analysis apply.

The associated partition function (at the temperature $1/\beta$) reads as:

$$Z(\mathbf{J}, \mathbf{B}, \mathbf{C}) = \sum_{\boldsymbol{\sigma} \in \{-1,1\}^n} \exp(-H(\boldsymbol{\sigma})).$$

We now incorporate the cuts and polarities induced by $\boldsymbol{\sigma} \in \{-1,1\}^n$. Observe that,

$$\begin{aligned} H(\boldsymbol{\sigma}) = \frac{\beta}{\sqrt{n}} \sum_{i<j:\sigma_i=\sigma_j} J_{ij} + \sum_{i:\sigma_i=+1} B_i + \sum_{i:\sigma_i=-1} C_i \\ - \left(\frac{\beta}{\sqrt{n}} \sum_{i<j:\sigma_i\neq\sigma_j} J_{ij} + \sum_{i:\sigma_i=-1} B_i + \sum_{i:\sigma_i=+1} C_i \right), \end{aligned}$$

where the ranges for the indices are $1 \le i < j \le n$ and $1 \le i \le n$. For convenience, we will denote the first part above by $\Sigma^+_{\boldsymbol{\sigma}}$, and the second part inside the brackets by $\Sigma^-_{\boldsymbol{\sigma}}$. Observe that, the object, $\Sigma \triangleq \sum_i B_i + \sum_i C_i + \frac{\beta}{\sqrt{n}} \sum_{i<j} J_{ij} = \Sigma^+_{\boldsymbol{\sigma}} + \Sigma^-_{\boldsymbol{\sigma}}$, is independent of $\boldsymbol{\sigma}$, and trivially computable. Now, note that, $\Sigma - H(\boldsymbol{\sigma}) = 2\Sigma^-_{\boldsymbol{\sigma}}$, and

therefore,

$$Z(\mathbf{J},\mathbf{B},\mathbf{C}) = \sum_{\boldsymbol{\sigma}\in\{-1,1\}^n} \exp(-\Sigma)\exp(2\Sigma^-_{\boldsymbol{\sigma}}).$$

Namely, $Z(\mathbf{J},\mathbf{B},\mathbf{C})$ is computable, if and only if, $\sum_{\boldsymbol{\sigma}\in\{-1,1\}^n}\exp(2\Sigma^-_{\boldsymbol{\sigma}})$ is computable. The presence of the factor 2 is, again, a minor detail that we omit in the sequel, since our techniques transfer without any modification. Thus, our focus is on computing $\sum_{\boldsymbol{\sigma}\in\{-1,1\}^n}\exp(\Sigma^-_{\boldsymbol{\sigma}})$; and denoting $\exp(\beta J_{ij}/\sqrt{n})$ by $\widehat{J_{ij}}$, $\exp(B_i)$ by $\widehat{B_i}$, and $\exp(C_i)$ by $\widehat{C}_i$, the object we are interested in computing is given by,

$$Z(\widehat{\mathbf{J}},\widehat{\mathbf{B}},\widehat{\mathbf{C}}) = \sum_{\boldsymbol{\sigma}\in\{-1,1\}^n}\left(\prod_{i:\sigma_i=-}\widehat{B_i}\right)\left(\prod_{i:\sigma_i=+}\widehat{C}_i\right)\left(\prod_{i<j:\sigma_i\neq\sigma_j}\widehat{J_{ij}}\right).$$

Our focus is on algorithms, that can compute $Z(\widehat{\mathbf{J}},\widehat{\mathbf{B}},\widehat{\mathbf{C}})$ exactly, in the following sense. The algorithm designer first selects a certain level N of digital precision, and computes these numbers, up to the selected precision level. Given a real number $x\in\mathbb{R}$, let $x^{[N]} = 2^{-N}\lfloor 2^N x\rfloor$ be the number obtained by keeping only first N binary bits of x after the binary point. The computational goal of the algorithm designer is to compute $Z(\widehat{\mathbf{J}}^{[\mathbf{N}]},\widehat{\mathbf{B}}^{[\mathbf{N}]},\widehat{\mathbf{C}}^{[\mathbf{N}]})$ exactly, where $\widehat{\mathbf{J}}^{[\mathbf{N}]} = (\widehat{J_{ij}}^{[N]} : 1\le i<j\le n)$, $\widehat{\mathbf{B}}^{[\mathbf{N}]} = (\widehat{B_i}^{[N]} : i\in[n])$, and $\widehat{\mathbf{C}}^{[\mathbf{N}]} = (\widehat{C_i}^{[N]} : i\in[n])$.

We now switch to a model with integer inputs. For convenience, let $\widetilde{J_{ij}} = \lfloor 2^N\widehat{J_{ij}}\rfloor = 2^N\widehat{J_{ij}}^{[N]}$, $\widetilde{B_i} = \lfloor 2^N\widehat{B_i}\rfloor = 2^N\widehat{B_i}^{[N]}$, $\widetilde{C_i} = \lfloor 2^N\widehat{C_i}\rfloor = 2^N\widehat{C_i}^{[N]}$; and define $f(n,\boldsymbol{\sigma})$ to be

$$f(n,\boldsymbol{\sigma}) = \frac{n(n-1)}{2} - n - I_n(\boldsymbol{\sigma}), \tag{4.3}$$

where $I_n(\sigma) = |\{(i,j) : \sigma_i\neq\sigma_j, 1\le i<j\le n\}|$. Equipped with this, we will focus on computing the following object with integer-valued inputs,

$$Z_n(\widetilde{\mathbf{J}},\widetilde{\mathbf{B}},\widetilde{\mathbf{C}}) = \sum_{\boldsymbol{\sigma}\in\{-1,1\}^n} 2^{Nf(n,\boldsymbol{\sigma})}\left(\prod_{i:\sigma_i=-}\widetilde{B_i}\right)\left(\prod_{i:\sigma_i=+}\widetilde{C}_i\right)\left(\prod_{i<j:\sigma_i\neq\sigma_j}\widetilde{J_{ij}}\right), \tag{4.4}$$

where the subscript n highlights the dependence on n, indicating that the system consists of n spins. Observe that, $Z_n(\widetilde{\mathbf{J}},\widetilde{\mathbf{B}},\widetilde{\mathbf{C}}) = 2^{Nn(n-1)/2}Z(\widehat{\mathbf{J}}^{[\mathbf{N}]},\widehat{\mathbf{B}}^{[\mathbf{N}]},\widehat{\mathbf{C}}^{[\mathbf{N}]})$. As a sanity check, note that $|I_n(\boldsymbol{\sigma})| + n \le \max_{0<k<n} k(n-k)+n < n(n-1)/2$ for $n>6$, and every $\boldsymbol{\sigma}\in\{-1,1\}^n$. Thus the model is indeed integral-valued.

We now state our main result, for the average-case hardness of computing $Z_n(\widetilde{\mathbf{J}},\widetilde{\mathbf{B}},\widetilde{\mathbf{C}})$.

Theorem 4.2.1. *Let $k,\alpha>0$ be arbitrary fixed constants. Suppose that, the precision value N satisfies, $C(k)\log n \le N \le n^\alpha$, where $C(k)$ is a constant, depending only on k. Suppose that there exists a polynomial in n time algorithm $\mathcal{A}$, which on input $(\widetilde{\mathbf{J}},\widetilde{\mathbf{B}},\widetilde{\mathbf{C}})$ produces a value $Z_{\mathcal{A}}(\widetilde{\mathbf{J}},\widetilde{\mathbf{B}},\widetilde{\mathbf{C}})$ satisfying*

$$\mathbb{P}(Z_{\mathcal{A}}(\widetilde{\mathbf{J}},\widetilde{\mathbf{B}},\widetilde{\mathbf{C}}) = Z_n(\widetilde{\mathbf{J}},\widetilde{\mathbf{B}},\widetilde{\mathbf{C}})) \ge \frac{1}{n^k},$$

for all sufficiently large n, where $Z_n(\widetilde{\mathbf{J}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{C}})$ is defined in (4.4). Then, $P = \#P$.

Quantitatively, the constant $C(k)$ can be taken as $21k + 20 + \epsilon$, where $\epsilon > 0$ is arbitrary (and this constant $C(k)$ can potentially be improved). The probability in Theorem 4.2.1 is taken with respect to the randomness of $(\widetilde{\mathbf{J}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{C}})$, which, in turn, is derived from the randomness of $(\mathbf{J}, \mathbf{B}, \mathbf{C})$. The logarithmic lower bound on the number of bits is imposed to address the technical issues when establishing the near-uniformity of the random variables $(\widetilde{\mathbf{J}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{C}})$ modulo an appropriately chosen prime. The upper bound on the number of bits that we retain is for ensuring that the input to the algorithm is of polynomial length in n.

4.2.2 Proof of Theorem 4.2.1

For any given $\boldsymbol{\Xi} \in \mathbb{Z}^{n(n-1)/2+2n}$ (note that, any algorithm computing the partition function of an n-spin system with external field accepts an input of size $n(n-1)/2 + 2n$), let $Z_n(\boldsymbol{\Xi}, p_n) \in \mathbb{Z}_{p_n}$ denotes $Z_n(\boldsymbol{\Xi}) \pmod{p_n}$, and similarly let $Z_{\mathcal{A}}(\boldsymbol{\Xi}; p_n)$ denotes $Z_{\mathcal{A}}(\boldsymbol{\Xi}) \pmod{p_n}$. Let $\mathbf{U} \in \mathbb{Z}_{p_n}^{n(n-1)/2+2n}$ be a random vector, consisting of iid entries, drawn independently from uniform distribution on $\mathbb{Z}_{p_n}$. The following result is our main proposition, and establishes the average-case hardness of computing the partition function defined in (4.4) modulo p_n, when the entry to the algorithm is $\mathbf{U}$. This, together with a coupling argument will establish Theorem 4.2.1.

Proposition 4.2.2. *Let $k > 0$ be an arbitrary constant. Suppose $\mathcal{A}$ is a polynomial in n time algorithm, which for any positive integer n, any prime number $p_n \geq 9n^{2k+2}$, and any input $\mathbf{a} = (\mathbf{a}^{\mathbf{J}}, \mathbf{a}^{\mathbf{B}}, \mathbf{a}^{\mathbf{C}}) \in \mathbb{Z}_{p_n}^{n(n-1)/2+2n}$ produces some output $Z_{\mathcal{A}}(\mathbf{a}; p_n) \in \mathbb{Z}_{p_n}$; and satisfies*

$$\mathbb{P}(Z_{\mathcal{A}}(\mathbf{U}; p_n) = Z_n(\mathbf{U}; p_n)) \geq \frac{1}{n^k},$$

where $\mathbf{U} = (\mathbf{U}^{\mathbf{J}}, \mathbf{U}^{\mathbf{B}}, \mathbf{U}^{\mathbf{C}}) \in \mathbb{Z}_{p_n}^{n(n-1)/2+2n}$ consists of iid entries chosen uniformly at random from $\mathbb{Z}_{p_n}$, and the probability is taken with respect to the randomness in $\mathbf{U}$. Then, $P = \#P$.

We now provide an outline of the proof of Proposition 4.2.2, which is the main building block of the proof of Theorem 4.2.1.

Proof Outline for Proposition 4.2.2 The proof of Proposition 4.2.2 is based on the $\#P-$hardness of the algorithmic problem of exactly computing the partition function for arbitrary inputs, and is inspired by the proof of the average-case hardness of computing the permanent [67].

The argument is by contradiction. Suppose that such an algorithm, $\mathcal{A}$, exists. Analogous to the Laplace expansion for permanent, we first establish a downward self-recursion for the partition function with respect to the sign of σ_n. Specifically, we establish that the partition function of an $n-$spin system (modulo p_n) can be expressed as a weighted sum of the partition functions of two $(n-1)-$spin systems

with adjusted parameters:

$$Z_n(\mathbf{J},\mathbf{B},\mathbf{C};p_n) = C'_n Z_{n-1}(\mathbf{J}',\mathbf{B}^+,\mathbf{C}^+;p_n) + B'_n Z_{n-1}(\mathbf{J}',\mathbf{B}^-,\mathbf{C}^-;p_n),$$

for suitable $B'_n, C'_n \in \mathbb{Z}_{p_n}$, $\mathbf{B}^+,\mathbf{B}^-,\mathbf{C}^+,\mathbf{C}^- \in \mathbb{Z}_{p_n}^{n-1}$.

Next, we let $v_1 = (\mathbf{J}',\mathbf{B}^+,\mathbf{C}^+)$ and $v_2 = (\mathbf{J}',\mathbf{B}^-,\mathbf{C}^-)$. We then generate two i.i.d. random vectors K and M (each of whose coordinates are distributed uniformly on $\mathbb{Z}_{p_n}$), and through interpolation, construct a "vector polynomial"

$$D(x) \triangleq (2-x)v_1 + (x-1)v_2 + (x-1)(x-2)(K+xM).$$

Here, $D(x)$ admits a natural interpretation: it corresponds to the parameters (i.e. the couplings and the external field) of an $(n-1)-$spin system. Next, we define the univariate polynomial $\phi(x) = Z_{n-1}(D(x);p_n)$. Namely, $\phi(x)$ is the partition function (modulo p_n) associated to the $(n-1)$-spin system whose parameters are stored in the vector $D(x)$. One can verify that $d \triangleq \deg(\phi) < n^2$, $\phi(1) = Z_{n-1}(D(1);p_n) = Z_{n-1}(\mathbf{J}',\mathbf{B}^+,\mathbf{C}^+;p_n)$, and $\phi(2) = Z_{n-1}(D(2);p_n) = Z_{n-1}(\mathbf{J}',\mathbf{B}^-,\mathbf{C}^-;p_n)$. In particular, our object of interest, $Z_n(\mathbf{J},\mathbf{B},\mathbf{C};p_n)$, satisfies

$$Z_n(\mathbf{J},\mathbf{B},\mathbf{C};p_n) = C'_n\phi(1) + B'_n\phi(2).$$

Therefore, provided $\phi(\cdot)$ can be reconstructed, the partition function can be computed. We then define the set

$$\mathcal{D} \triangleq \{D(x) : x = 3,4,\ldots,p_n\}.$$

The extra randomness incorporated via K and M above ensures $\mathcal{D}$ consists of pairwise independent samples. We then use the algorithm $\mathcal{A}(\cdot)$ to obtain a list $(x,\mathcal{A}(D(x)))$ of pairs, where $x \in \mathcal{D}$. This list is nothing but a list of noisy samples of the polynomial $\phi(\cdot)$, whose degree is bounded from above by n^2. After a series of technical steps, we show how to identify $\phi(\cdot)$. These steps consist of lower bounding the number of correct evaluations of ϕ in the list above, as well as a list-decoding algorithm by Sudan [270]. This yields that under the hypothesis of Proposition 4.2.2, there is a (randomized) polynomial-time procedure, which for any $\mathbf{a} \in \mathbb{Z}_{p_n}^{n(n-1)/2+2n}$ computes $Z_n(\mathbf{a};p_n)$ with high probability. We then use the Gaussian tail estimate, $\mathbb{P}(\mathcal{N}(0,1) > t) = \exp(-\Theta(t^2))$, to upper bound the partition function (whp); and generate sufficiently many primes p_n of sufficient size for which the product $\prod_n p_n$ is larger than the partition function itself. The latter step uses the prime density. Chinese remainder theorem then allows the exact computation of the partition function. All of the steps and reductions above run in polynomial time, as we show; and by controlling the probability of error along the route, we establish the desired contradiction.

Taking Proposition 4.2.2 for granted, we now provide an outline of how one concludes the proof of Theorem 4.2.1.

Proof Outline for Theorem 4.2.1 Provided that we are equipped with Proposition 4.2.2, we conclude the proof of Theorem 4.2.1, again by relying on a contradiction argument. Suppose, for the sake of the contradiction, that such a (polynomial-time) algorithm, $\mathcal{A}$, as in Theorem 4.2.1 exists.

We establish the near-uniformity of the log-normal distribution modulo p_n, where p_n is a sufficiently large prime. This will allow us to show that the total variation distance between the parameters of the partition function we are interested in computing and the uniform distribution modulo p_n is small. In particular, considering a coupling between the parameters of interest and the uniform distribution, we show that $\mathcal{A}$ succeeds in computing the partition function modulo p_n in polynomial-time with inverse polynomial probability of success. This contradicts with Proposition 4.2.2; and thus concludes the proof of Theorem 4.2.1.

Equipped with this outline, we now proceed with the full proofs.

Proof. (of Proposition 4.2.2) We will use as basis the #P-hardness of computing the partition function, for arbitrary inputs. Namely, if there exists a polynomial time algorithm computing $Z(\mathbf{j}, \mathbf{b}, \mathbf{c})$ for any arbitrary input $\mathbf{j}, \mathbf{b}, \mathbf{c}$ with probability bounded away from zero as $n \to \infty$, then P=#P.

Let $q \geq 1/n^k$ be the success probability of $\mathcal{A}$, and $\mathbf{a} = (\mathbf{a}^{\mathbf{J}}, \mathbf{a}^{\mathbf{B}}, \mathbf{a}^{\mathbf{C}}) \in \mathbb{Z}_{p_n}^{n(n-1)/2+2n}$ be an *arbitrary* input, whose partition function we want to compute. For convenience, we drop $\mathbf{a}$, denote $(\mathbf{a}^{\mathbf{J}}, \mathbf{a}^{\mathbf{B}}, \mathbf{a}^{\mathbf{C}})$ by $(\mathbf{J}, \mathbf{B}, \mathbf{C})$. The following lemma establishes the downward self-recursive behaviour of the partition function (modulo p_n) by expressing the partition function of an n-spin system as a weighted sum of partition functions of two $(n-1)$-spin systems, with appropriately adjusted external field components.

Lemma 4.2.3. *The following identity holds:*

$$Z_n(\mathbf{J}, \mathbf{B}, \mathbf{C}; p_n) = C'_n Z_{n-1}(\mathbf{J}', \mathbf{B}^+, \mathbf{C}^+; p_n) + B'_n Z_{n-1}(\mathbf{J}', \mathbf{B}^-, \mathbf{C}^-; p_n),$$

where, $Z_n(\mathbf{J}, \mathbf{B}, \mathbf{C}; p_n) = Z_n(\mathbf{J}, \mathbf{B}, \mathbf{C}) \pmod{p_n}$ *with* Z_n *defined in (4.4);* $\mathbf{J}' \in \mathbb{Z}_{p_n}^{(n-1)(n-2)/2}$ *is such that* $J'_{ij} = J_{ij}$ *for every* $1 \leq i < j \leq n-1$*;* $\mathbf{B}^+, \mathbf{B}^-, \mathbf{C}^+, \mathbf{C}^- \in \mathbb{Z}_{p_n}^{n-1}$ *are such that* $B_i^+ = 2^{-N} B_i J_{in}$*,* $B_i^- = B_i$*,* $C_i^+ = C_i$*, and* $C_i^- = 2^{-N} C_i J_{in}$*, for every* $1 \leq i \leq n-1$*; and* $C'_n = C_n 2^{(n-2)N}$*,* $B'_n = B_n 2^{(n-2)N}$*.*

The proof of this lemma is provided in Section 4.5.2. Namely, provided we can compute $Z_{n-1}(\mathbf{J}', \mathbf{B}^+, \mathbf{C}^+; p_n)$ and $Z_{n-1}(\mathbf{J}', \mathbf{B}^-, \mathbf{C}^-; p_n)$, we can compute $Z_n(\mathbf{J}, \mathbf{B}, \mathbf{C}; p_n)$. Note that, since we are interested in modulo p_n computation, the number 2^{-N} is nothing but g^N, where $g \in \mathbb{Z}_{p_n}$ satisfies $2g \equiv 1 \pmod{p_n}$, that is, g is the multiplicative inverse of 2 modulo p_n.

Next, let $v_1 = (\mathbf{J}', \mathbf{B}^+, \mathbf{C}^+) \in \mathbb{Z}_{p_n}^T$, and $v_2 = (\mathbf{J}', \mathbf{B}^-, \mathbf{C}^-) \in \mathbb{Z}_{p_n}^T$; where the input dimension $(n-1)(n-2)/2 + 2(n-1)$ of the algorithm computing partition function for a model with $(n-1)$-spins is denoted by T for convenience. Now, we construct the vector polynomial

$$D(x) = (2-x)v_1 + (x-1)v_2 + (x-1)(x-2)(K + xM), \tag{4.5}$$

of dimension T, where K, M are iid random vectors, drawn from uniform distribution on $\mathbb{Z}_{p_n}^T$. The incorporation of this extra randomness is due to an earlier idea by Gemmell and Sudan [146].

Next, consider $\phi(x) = Z_{n-1}(D(x); p_n)$, namely, the partition function of an $(n-1)$-spin system, associated with the vector $D(x)$ (where the first $(n-1)(n-2)/2$ components correspond to couplings, the following $(n-1)$ components correspond to B_i's, and the last $(n-1)$ components correspond to C_i's), which is a univariate polynomial in x. We now upper bound the degree of $\phi(x)$. Note that,

$$d = \deg(\phi) \leq 3\left(\max_{\boldsymbol{\sigma}\in\{-1,1\}^{n-1}} |I(\boldsymbol{\sigma})| + n - 1\right) = 3\left(\max_{1\leq k\leq n-1} k(n-1-k) + n - 1\right) < n^2,$$

for n large. Observe also that, $\phi(1) = Z_{n-1}(D(1); p_n) = Z_{n-1}(\mathbf{J}', \mathbf{B}^+, \mathbf{C}^+; p_n)$, $\phi(2) = Z_n(D(2); p_n) = Z_{n-1}(\mathbf{J}', \mathbf{B}^-, \mathbf{C}^-; p_n)$, hence, $Z_n(\mathbf{J}, \mathbf{B}, \mathbf{C}; p_n) = C_n'\phi(1) + B_n'\phi(2)$. Therefore, provided that we can recover $\phi(\cdot)$, $Z_n(\mathbf{J}, \mathbf{B}, \mathbf{C}; p_n)$ can be computed. With this, we now turn our attention to recovering the polynomial $\phi(\cdot)$. Let $\mathcal{D}$ be a set of cardinality $p_n - 2$, defined as $\mathcal{D} = \{D(x) : x = 3, 4, \ldots, p_n\}$. We claim that, $\mathcal{D}$ consists of pairwise independent samples.

Lemma 4.2.4. *For every distinct $x_1, x_2 \in \{3, 4, \ldots, p_n\}$, the random vectors $D(x_1)$ and $D(x_2)$ are independent and uniformly distributed over $\mathbb{Z}_{p_n}^T$. That is, for every such x_1, x_2 and every $y_1, y_2 \in \mathbb{Z}_{p_n}^T$; it holds that $\mathbb{P}(D(x_1) = y_1) = 1/p_n^T = \mathbb{P}(D(x_2) = y_2)$, and*

$$\mathbb{P}(D(x_1) = y_1, D(x_2) = y_2) = 1/p_n^{2T} = \mathbb{P}(D(x_1) = y_1)\mathbb{P}(D(x_2) = y_2),$$

where the probability is taken with respect to the randomness in K and M.

The proof of this lemma is provided in Section 4.5.3. Now, we run $\mathcal{A}$ on $\mathcal{D}$, and will use the independence to deduce via Chebyshev's inequality that, with high probability, $\mathcal{A}$ runs correctly, on at least $q/2$ fraction of inputs in $\mathcal{D}$, where $q \geq 1/n^k$ is the success probability of our algorithm. This is encapsulated by the following lemma.

Lemma 4.2.5. *Let the random variable $\mathcal{N}$ be the number of points $D(x) \in \mathcal{D}$, such that $\mathcal{A}(D(x)) = \phi(x) = Z_{n-1}(D(x); p_n)$, namely $\mathcal{A}$ correctly computes the partition function at $D(x)$. Then,*

$$\mathbb{P}(\mathbb{N} \geq (p_n - 2)q/2) \geq 1 - \frac{1}{(p_n - 2)q^2},$$

where q is the success probability of $\mathcal{A}$, and the probability is taken with respect to the randomness in $\mathcal{D}$, which, in turn, is due to the randomness in K and M.

The proof of this lemma can be found in Section 4.5.4. Now, let $G(f) = \{(x, f(x)) : x = 1, 2, \ldots, p_n\}$ be the graph of a function $f \in \mathbb{Z}_{p_n}[x]$. Define the set $\mathcal{S} = \{(x, \mathcal{A}(D(x))) : x = 3, 4, \ldots, p_n\}$, and let $\mathcal{F}$ be a set of polynomials, defined as,

$$\mathcal{F} = \{f \in \mathbb{Z}_{p_n}[x] : \deg(f) < n^2, |G(f) \cap \mathcal{S}| \geq (p_n - 2)q/2\}.$$

Namely, $f \in \mathcal{F}$ if and only if, its coefficients are from $\mathbb{Z}_{p_n}$, it is of degree at most $n^2 - 1$; and its graph intersects the set $\mathcal{S}$ on at least $(p_n - 2)q/2$ points. Due to Lemma 4.2.5, we know that $\phi(x) \in \mathcal{F}$, with probability at least $1 - \frac{1}{(p_n-2)q^2}$. We now show that this set $\mathcal{F}$ of candidate polynomials contains at most polynomial in n many polynomials.

Lemma 4.2.6. *If $p_n \geq 9n^{2k+2}$, then $|\mathcal{F}| \leq 3/q$, where q is the success probability of $\mathcal{A}$. In particular, $\mathcal{F}$ contains at most polynomial in n many polynomials, since $q \geq 1/n^k$.*

The proof of this lemma is provided in Section 4.5.5.

In what remains, we will show how to explicitly construct all such polynomials, through a randomized algorithm, which succeeds with high probability. To that end, we use the following elegant result, due to Cai et al. [67].

Lemma 4.2.7. *There exists a randomized procedure running in polynomial time, through which, with high probability, one can generate a list $\mathcal{L} = (x_i, y_i)_{i=1}^{L}$ of L pairs, such that, $y_i = \phi(x_i)$, for at least t pairs from the list with distinct first coordinates, where $t > \sqrt{2Ld}$, with $d = \deg(\phi)$, and $\phi(x) = Z_{n-1}(D(x); p_n)$.*

The proof of this lemma is isolated from the argument of [67], and provided in Section 4.5.6 for completeness. Of course, these discussions are all based on the assumption that we condition on the high probability event that $\{\mathbb{N} \geq (p_n - 2)q/2\}$, where $\mathbb{N}$ is the random variable defined in Lemma 4.2.5.

Having obtained this list, we now turn our attention to finding all polynomials (where, by Lemma 4.2.6, there is at most polynomial in n many of those), whose graph intersects the list at at least t points with distinct first coordinates (for the specific values of t depending on the magnitude of p_n, see the proof of Lemma 4.2.7 in Section 4.5.6). For this, we use the following list-decoding algorithm of [270], introduced originally in the context of coding theory, which is an improved version of Berlekamp-Welch decoder.

Lemma 4.2.8. *(Theorem 5 in [270]) Given a sequence $\{(x_i, y_i)\}_{i=1}^{L}$ of L distinct pairs, where x_is and y_is are an element of a field $\mathbb{F}$, and integer parameters t and d, such that $t \geq d\lceil\sqrt{2(L+1)/d}\rceil - \lfloor d/2 \rfloor$, there exists an algorithm which can find all polynomials $f : \mathbb{F} \to \mathbb{F}$ of degree at most d, such that the number of points (x_i, y_i) satisfying $y_i = f(x_i)$ is at least t.*

The algorithm is a probabilistic polynomial time algorithm. For the sake of completeness, we briefly sketch his algorithm here. For weights $w_x, w_y \in \mathbb{Z}^+$, define (w_x, w_y)-weighted degree of a monomial $q_{ij}x^i y^j$ to be $iw_x + jw_y$. The (w_x, w_y)-weighted degree of a polynomial, $Q(x, y) = \sum_{(i,j)\in I} q_{ij}x^i y^j$ is defined to be $\max_{(i,j)\in I} iw_x + jw_y$. Let $m, \ell \in \mathbb{Z}^+$ be positive integers, to be determined. Construct a non-zero polynomial $Q(x, y) = \sum_{i,j} q_{ij}x^i y^j$, whose $(1, d)$-weighted degree is at most $m + \ell d$, and $Q(x_i, y_i) = 0$, for every $i \in [L]$. The number of coefficients q_{ij} of any such polynomial is at most, $\sum_{j=0}^{\ell} \sum_{i=0}^{m+(\ell-j)d} 1 = (m + 1)(\ell + 1) + d\ell(\ell + 1)/2$. Hence, provided $(m + 1)(\ell + 1) + d\ell(\ell + 1)/2 > L$, we have more unknowns (i.e., coefficients q_{ij}) than

equations, $Q(x_i, y_i) = 0$, for $i \in [L]$, and thus, such a $Q(x_i, y_i)$ exists, and moreover, can be found in polynomial time. Now, we look at the following univariate polynomial, $Q(x, f(x)) \in \mathbb{F}[x]$. This polynomial has degree, at most $m + \ell d$. Note that, for every i such that $f(x_i) = y_i$, $Q(x_i, f(x_i)) = Q(x_i, y_i) = 0$. Hence, provided that, m, ℓ are chosen, such that $m + \ell d < t$, it holds that, this polynomial has $t > m + \ell d = \deg Q(x, f(x))$ zeroes, hence, it must be identically zero. Now, viewing $Q(x, y)$ to be $Q_x(y)$, a polynomial in y, with coefficients from $\mathbb{F}[x]$, we have that whenever $Q_x(\xi) = 0$, it holds that, $(y - \xi)$ divides $Q_x(y)$, hence, for $\xi = f(x)$, we get $y - f(x) \mid Q(x, y)$. Provided that $Q(x, y)$ exists (which will be guaranteed by parameter assumptions) and can be reconstructed in polynomial in n time, it can also be factorized in probabilistic polynomial time [181], and $y - f(x)$ will be one of its irreducible factors. For a concrete choice of parameters, see [270]; or [67], which also has a brief and different exposition of the aforementioned ideas. We will use this result with $t > \sqrt{2Ld}$, where $d = \deg(\phi) < n^2$.

Now, we have a randomized procedure, which outputs a certain list $\mathcal{K}$ of at most $3/q$ polynomials, one of which is the correct $\phi(x) = Z_{n-1}(D(x); p_n)$. The idea for the remainder is as follows. We will find a point x, at which, all polynomials from the list $\mathcal{K}$ disagree. Towards this goal, define a set $\mathcal{T}$ of triples,

$$\mathcal{T} = \{(x, f(x), g(x)) : f(x) = g(x), x \in \mathbb{Z}_{p_n}, f, g \in \mathcal{K}\}.$$

We now use a double-counting argument. Note that, every pair (f, g) of distinct polynomials from the list $\mathcal{K}$ can agree on at most $n^2 - 1$ points. Since, the total number of such pairs (f, g) of distinct polynomials from $\mathcal{K}$ is less than $(3/q)^2$, we deduce $|\mathcal{T}| < 9n^{2k+2}$. Since $|\mathbb{Z}_{p_n}| > |\mathcal{T}|$, it follows that, there exists a v, such that, no triple, whose first coordinate is v belongs to $\mathcal{T}$. Clearly, this point v can be found in polynomial time, since p_n and the size of the list are polynomial in n. Thus, there is at least one point on which all polynomials from the list $\mathcal{K}$ disagree. It is possible now to identify $\phi(x) = Z_{n-1}(D(x); p_n)$, by evaluating $Z_{n-1}(D(v); p_n)$, since whp, $\phi(\cdot) \in \mathcal{K}$, and all polynomials from list $\mathcal{K}$ take distinct values at v. Provided $\phi(x)$ can be identified, we can compute $Z_n(\mathbf{J}, \mathbf{B}, \mathbf{C}; p_n)$, the original partition function of interest, simply via $C'_n\phi(1) + B'_n\phi(2)$, as mentioned in the beginning.

Therefore, $Z_n(\mathbf{J}, \mathbf{B}, \mathbf{C}; p_n)$ can be computed, provided that $Z_{n-1}(D(v); p_n)$ can be computed, a reduction from an $n-$spin system, to an $(n-1)-$spin system. Note that, the probability of error in this randomized reduction is upper bounded, via the union bound, by the sum of probabilities that, $\mathbb{N}$, defined in Lemma 4.2.5 is less than $(p_n - 2)q/2$, which is of probability at most $\frac{1}{(p_n-2)q^2}$, which is c/n^2 for some constant $c > 0$, independent of n; plus, the probability of failure during the construction of a list of L pairs $(x_i, y_i)_{i=1}^L$ with $t > \sqrt{2Ld}$, which, conditional on the high probability event, $\{\mathcal{N} \geq (p_n - 2)q/2\}$, is exponentially small in n; and finally, the probability that we encounter an error during generating the list of polynomials through factorization, per Lemma 4.2.8, which can again be made exponentially small in n. Thus, the overall probability of error for this reduction is c'/n^2, for some absolute constant $c' > 0$, independent of n. Next, select a large H and repeat the same downward reduction protocol $n \to n-1, n-1 \to n-2, \cdots, H+1 \to H$, such that the total

probability of error $\sum_{j=H}^{n} c'/j^2$ during the entire reduction is less than $1/2$ (note that, the reduction step, $n-1 \to n-2$ aims at computing $\phi(v) = Z_{n-1}(D(v); p_n)$, where v is the element of $\mathbb{Z}_{p_n}$ discussed earlier; and each step, we reduce the problem of recovering the associated polynomial to evaluating the partition function of a system with one less number of spins, at a single input point). Once the system has H spins, compute the partition function by hand. This procedure yields an algorithm computing $Z_n(\mathbf{J}, \mathbf{B}, \mathbf{C}; p_n)$, the partition function value we wanted to compute in the beginning of the proof of Proposition 4.2.2, with probability greater than $1/2$. Now, if we repeat this algorithm R times, and take the majority vote (i.e., the number that appeared the majority number of times), the probability of having a wrong answer appearing as majority vote is, by Chernoff bound, exponentially small in R. Taking R to be polynomial in n, we have that with probability at least $1 - e^{-\Omega(n)}$ this procedure correctly computes $Z_n(\mathbf{J}, \mathbf{B}, \mathbf{C}; p_n)$.

We have now established that, provided, there is a polynomial time algorithm $\mathcal{A}$, which exactly computes the partition function on $1/n^k$ fraction of inputs (from $\mathbb{Z}_{p_n}^{n(n-1)/2+2n}$), then there exists a (randomized) polynomial time procedure, for which, for every $\mathbf{a} \in \mathbb{Z}_{p_n}^{n(n-1)/2+2n}$ (including, in particular, the adversarially-chosen ones), it correctly evaluates $Z_n(\mathbf{a}) \pmod{p_n}$ with probability $1 - o(1)$. We now use this procedure to show, how to evaluate $Z_n(\mathbf{a})$ (without the mod operator). We use the Chinese Remainder Theorem, which, for convenience, is stated below.

Theorem 4.2.9. *Let $p_1, \ldots, p_k$ be distinct pairwise coprime positive integers, and $a_1, \ldots, a_k$ be integers. Then, there exists a unique integer $m \in \{0, 1, \ldots, P\}$ where $P = \prod_{\ell=1}^{k} p_\ell$, such that, $m \equiv a_i \pmod{p_i}$, for every $1 \leq i \leq k$.*

In particular, letting $P_i = P/p_i$, $m = \sum_{\ell=1}^{k} c_i P_i a_i \pmod{P}$ works, where $c_i \equiv P_i^{-1} \pmod{p_i}$. The number c_i can be computed by running Euclidean algorithm: Since $\gcd(P_i, p) = 1$, it follows from Bézout's identity that, there exists integers $c_i, b \in \mathbb{Z}$ such that, $c_i P_i + pb = 1$, and thus, $c_i P_i \equiv 1 \pmod{p}$. Now, we proceed as follows. Fix a positive integer m. If we can find a collection $\{p_1, \ldots, p_\ell\}$ of primes such that the corresponding product $P = \prod_{k=1}^{\ell} p_k$ exceeds m, then we can recover m, from $(r_i)_{i=1}^{\ell}$, where $r_i \in \mathbb{Z}_{p_i}$ is such that $m \equiv r_i \pmod{p_i}$, namely, r_i is the remainder obtained upon dividing m by p_i, for each i.

For this goal, we now establish a bound, where with high probability, the original partition function is less than this bound. Recall the standard Gaussian tail estimate, $\mathbb{P}(Z > t) = O(\exp(-t^2/2))$. Using this,

$$\mathbb{P}(e^{\beta n^{-1/2} J} > t) = O(\exp(-n \log^2(t)/(2\beta^2))),$$

which, for $t = n$, gives a bound, $o(n^{-2})$. Now, for external field contribution, we have $\mathbb{P}(e^B > t) \leq O(\exp(-\log^2(t)/(2\beta^2)))$ (also for C), which, for $t = n$, gives $O(n^{-\log n/(2\beta^2)})$, which is, again, $o(n^{-2})$. Hence, with high probability, the $n(n-1)/2 + 2n$-dimensional vector, $\mathbf{V} = (\mathbf{J}, \mathbf{B}, \mathbf{C})$ is such that, $\|\mathbf{V}\|_\infty \leq n$. Therefore, with high probability, the partition function is at most sum of 2^n terms, each of which is a product of at most n^2 terms (since, we have n terms for external field, and

at most $n^2/2$ terms for spin-spin couplings) each bounded by $2^N n$. This establishes, the partition function is at most $2^n(2^N n)^{n^2} = 2^{Nn^2+n^2\log_2 n+O(n)}$.

It now remains to show that, there exists sufficiently many prime numbers of appropriate size, that we can use for Chinese remaindering.

Lemma 4.2.10. *Let $k, \alpha > 0$ be a fixed constants, and N satisfies $\Omega(\log n) \leq N \leq n^\alpha$. The number of primes between $9n^{2k+2}$ and $2(2+\alpha+2k)Nn^{2k+2}\log n$ is at least Nn^{2k+2}, for all sufficiently large n.*

The proof of this lemma can be found in Section 4.5.7.

Having done this, we will find a sequence of Nn^{2k+2} primes via brute force search in polynomial time, since $N \leq n^\alpha$ for some constant α, with $p_j > \Omega(n^{2k+2})$. This will establish, $\prod_j p_j > \Omega((n^{2k+2})^{Nn^{2k+2}}) = \Omega(2^{Nn^{2k+2}(2k+2)\log n})$. Since the partition function is at most $2^{Nn^2+n^2\log_2 n+O(n)}$ and since $N = \Omega(\log_2 n)$, we therefore conclude that the product of primes we have selected is, whp, larger than the partition function itself, and therefore, by running $\mathcal{A}$ with each of these prime basis, and Chinese remaindering, we can compute the partition function exactly. Therefore, the proof of Proposition 4.2.2 is complete. □

We now establish that the density of log-Normal distribution is Lipschitz continuous within a finite interval, and will bound the Lipschitz constant, to establish a certain probabilistic coupling.Recall that $J_{ij}, 1 \leq i < j \leq n$ are i.i.d. standard normal and $\widehat{J}_{ij} = e^{\beta/\sqrt{n}J_{ij}}$. Let $f_{\widehat{J}}$ denote the common density of $\widehat{J}_{ij}$.

Lemma 4.2.11. *For every $0 < \delta < \Delta$ satisfying $\log\Delta > \beta^2$ and every $\delta \leq t, \tilde{t} \leq \Delta$, the following bound holds.*

$$\exp\left(-\frac{2n\log\Delta}{\beta^2\delta}|\tilde{t}-t|\right) \leq \frac{f_{\widehat{J}}(\tilde{t})}{f_{\widehat{J}}(t)} \leq \exp\left(\frac{2n\log\Delta}{\beta^2\delta}|\tilde{t}-t|\right). \tag{4.6}$$

The proof of this lemma is provided in Section 4.5.1. Furthermore, letting $\widehat{B}_i = e^{B_i}$ and $\widehat{C}_i = e^{C_i}$, and denoting the (common) densities by $f_{\widehat{B}}$ and $f_{\widehat{C}}$, we have that the same Lipschitz condition holds also for $f_{\widehat{B}}(t)$ and $f_{\widehat{C}}(t)$, and therefore, the result of Lemma 4.2.11 applies also to the exponentiated version of the external field components, see Remark 4.5.1.

The idea for the remaining part is as follows. We will establish that, the algorithmic inputs (obtained by exponentiating the real-valued inputs and truncating at an appropriate level N), are close to uniform distribution (modulo p_n), in total variation sense, which will establish the existence of a desired coupling to conclude the proof of Theorem 4.2.1. To that end, we now establish an auxiliary result, showing that the log-Normal distribution is nearly uniform, modulo p_n.

Lemma 4.2.12. *The following bound holds for every $A \in \{\widetilde{J}_{ij} : 1 \leq i < j \leq n\} \cup \{\widetilde{B}_i : i \in [n]\} \cup \{\widetilde{C}_i : i \in [n]\}$:*

$$\max_{0\leq\ell\leq p_n-1} |\mathbb{P}(A \equiv \ell \bmod (p_n)) - p_n^{-1}| = O(N^{-1}n^{-5k-4}).$$

The proof of this lemma is provided in Section 4.5.8. We now return to the proof of Theorem 4.2.1. Using Lemma 4.2.12, the total variation distance between any $A \in \{\widetilde{J}_{ij}, \widetilde{B}_i, \widetilde{C}_i\}$ and $U \sim \text{Unif}(\mathbb{Z}_{p_n})$ is at most, $O(p_n N^{-1} n^{-5k-4})$, which, using the trivial inequality $p_n \leq O(Nn^{3k+2})$, is $O(n^{-2k-2})$. We now use the following well-known maximal total variation coupling result.

Theorem 4.2.13. *Let the random variables X, Y have marginal distributions, μ and ν, and let $d_{TV}(\mu, \nu)$ denotes the total variation distance between μ and ν. Then, for any coupling (namely, any joint distribution with marginals of X and Y being μ and ν, respectively) of X and Y, it holds that, $\mathbb{P}(X = Y) \leq 1 - d_{TV}(\mu, \nu)$. Moreover, there is a coupling of X and Y, under which, we have the equality $\mathbb{P}(X = Y) = 1 - d_{TV}(\mu, \nu)$.*

Using this maximal coupling result, we now observe that, we can couple A (where, $A \in \{\widetilde{J_{ij}}, \widetilde{B_i}, \widetilde{C_i}\}$) with a random variable U, uniformly distributed on $\mathbb{Z}_{p_n}$, such that

$$\mathbb{P}(A = U) \geq 1 - O(n^{-2k-2}).$$

Now, let U_{ij}, U_i^B, and U_i^C be random variables, uniform over $\mathbb{Z}_{p_n}$, such that,

$$\mathbb{P}(\widetilde{J}_{ij} \neq U_{ij}) \leq O(n^{-2k-2}), \quad \mathbb{P}(\widetilde{B}_i \neq U_i^B) \leq O(n^{-2k-2}), \text{ and } \quad \mathbb{P}(\widetilde{C}_i \neq U_i^C) \leq O(n^{-2k-2}).$$

In particular, using union bound, we can couple $\boldsymbol{\Xi} = (\widetilde{\mathbf{J}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{C}})$, with a vector, $\mathbf{U} = (\mathbf{U^J}, \mathbf{U^B}, \mathbf{U^C}) \in \mathbb{Z}_{p_n}^{n(n-1)/2+2n}$, such that, $\mathbb{P}(\boldsymbol{\Xi} = \mathbf{U}) \geq 1 - O(n^{-2k})$. Now, we define several auxiliary events. Let $\mathcal{E}_1 = \{Z_n(\boldsymbol{\Xi}; p_n) = Z_n(\mathbf{U}; p_n)\}$, $\mathcal{E}_2 = \{Z_{\mathcal{A}}(\boldsymbol{\Xi}; p_n) = Z_{\mathcal{A}}(\mathbf{U}; p_n)\}$, and $\mathcal{E}_3 = \{Z_{\mathcal{A}}(\boldsymbol{\Xi}) = Z_n(\boldsymbol{\Xi})\}$. Observe that, due to the coupling, we have $\mathbb{P}(\mathcal{E}_1), \mathbb{P}(\mathcal{E}_2) \geq 1 - O(n^{-2k})$. Now, suppose, the statement of the Theorem 4.2.1 holds, and that, $\mathbb{P}(\mathcal{E}_3) \geq 1/n^k$. Observe that, $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \subseteq \{Z_n(\mathbf{U}; p_n) = Z_{\mathcal{A}}(\mathbf{U}; p_n)\}$. Hence, $\mathcal{A}$ satisfies,

$$\begin{aligned}
\mathbb{P}(Z_n(\mathbf{U}; p_n) = Z_{\mathcal{A}}(\mathbf{U}; p_n)) &\geq \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) \\
&= 1 - \mathbb{P}(\mathcal{E}_1^c \cup \mathcal{E}_2^c \cup \mathcal{E}_3^c) \\
&\geq \mathbb{P}(\mathcal{E}_3) - \mathbb{P}(\mathcal{E}_1^c) - \mathbb{P}(\mathcal{E}_2^c) \\
&\geq \frac{1}{n^k} - O(n^{-2k}) \geq \frac{1}{n^{k'}},
\end{aligned}$$

using union bound, where k' obeys: $k < k' < 2k$ and $n^{2k'+2} \log n = O(n^{3k+2})$. This contradicts with Proposition 4.2.2, with the probability of success taken to be as $1/n^{k'}$ for this value of k'.

4.3 Average-Case Hardness under Real-Valued Computational Model

In this section, we study the problem of exactly computing the partition function associated with the Sherrington-Kirkpatrick model, but this time under the real-valued

computation model, as opposed to the finite precision arithmetic model adopted in previous section.

Details on the Model of Computation More specifically, we assume that there exists a computational engine which can store arbitrary real-valued inputs, and operate on them. The allowed operations are the arithmetic operations (such as addition, subtraction, multiplication, and division), including also the computation of polynomials. Each such operation on real-valued inputs is assumed to require a unit time and is assumed to be of unit cost. An example of such a computation engine operating over real-valued inputs is the so-called Blum-Shub-Smale (BSS) machine [51, 50].

The techniques employed in the previous section do not extend to real-valued computational model, since it is not clear what the appropriate real-valued analogue of $\mathbb{Z}_p$ is.

4.3.1 Model and the Main Result

We start by incorporating the cuts induced by the spin assignment $\boldsymbol{\sigma} \in \{-1,1\}^n$, and reduce the problem to computing a partition function associated with the cuts, in a manner analogous to the previous setting. Let $\Sigma = \sum_{i<j} J_{ij} = \sum_{\sigma_i \neq \sigma_j} J_{ij} + \sum_{\sigma_i = \sigma_j} J_{ij}$. Note that, Σ is independent of the spin assignment $\boldsymbol{\sigma} \in \{-1,1\}^n$, and is computable in polynomial time. Observe also that, $\Sigma - H(\boldsymbol{\sigma}) = 2\sum_{\sigma_i \neq \sigma_j} J_{ij}$, where $H(\boldsymbol{\sigma}) = \sum_{i<j} J_{ij}\sigma_i\sigma_j$. Therefore,

$$Z(\mathbf{J}) = \sum_{\boldsymbol{\sigma}\in\{-1,1\}^n} \exp(-H(\boldsymbol{\sigma})) = \sum_{\boldsymbol{\sigma}\in\{-1,1\}^n} \exp(-\Sigma) \exp\left(2 \sum_{i<j:\sigma_i\neq\sigma_j} J_{ij}\right).$$

Letting $X_{ij} = e^{2J_{ij}}$, we observe that since $\exp(-\Sigma)$ is a trivially computable constant, it suffices to compute $\widehat{Z}(\mathbf{J})$, where

$$\widehat{Z}(\mathbf{J}) = \sum_{\boldsymbol{\sigma}\in\{-1,1\}^n} \prod_{i<j:\sigma_i\neq\sigma_j} X_{ij}.$$

Note that, $\widehat{Z}(\mathbf{J})$ involves X_{ij}, which are, in turn, derived from J_{ij}. For this reason, we will refer to this object as $\widehat{Z}(\mathbf{X})$, as well.

In what follows, we are interested in computing $\widehat{Z}(\mathbf{X})$, when $\mathbf{X}$ is given as a random input to the real-valued computational engine that we operate under. We now elaborate on this. Let $\mathbf{J} = (J_{ij} : 1 \leq i < j \leq n) \in \mathbb{R}^{n(n-1)/2}$ be a random vector with i.i.d. standard normal coordinates; and let $\mathbf{X} = (X_{ij} : 1 \leq i < j \leq n) \in \mathbb{R}^{n(n-1)/2}$ be the random vector with $X_{ij} = \exp(2J_{ij})$, for $1 \leq i < j \leq n$. The oracle receives $\mathbf{X}$ as its input, and the goal is to compute

$$\widehat{Z}(\mathbf{X}) \triangleq \sum_{\boldsymbol{\sigma}\in\{-1,1\}^n} \prod_{i<j:\sigma_i\neq\sigma_j} X_{ij}. \tag{4.7}$$

Our main result under this setting is as follows:

Theorem 4.3.1. *Let $\delta \geq 1/\text{poly}(n) > 0$ be an arbitrary real number, $\mathbf{J} = (J_{ij} : 1 \leq i < j \leq n) \in \mathbb{R}^{n(n-1)/2}$ with $J_{ij} \stackrel{d}{=} \mathcal{N}(0,1)$ i.i.d.; and $\mathbf{X} = (X_{ij} : 1 \leq i < j \leq n) \in \mathbb{R}^{n(n-1)/2}$ with $X_{ij} = \exp(2J_{ij})$ for $1 \leq i < j \leq n$. Suppose that $\mathcal{O}$ is an algorithm, such that:*

$$\mathbb{P}\left(\mathcal{O}(\mathbf{X}) = \widehat{Z}(\mathbf{X})\right) \geq \frac{3}{4} + \delta,$$

where $\widehat{Z}(\mathbf{X})$ is defined in (4.7). Then, $P = \#P$.

Note that the probability in Theorem 4.3.1 is taken with respect to the randomness in $\mathbf{X}$, which, in turn, stems from the randomness in the (i.i.d. standard normal) couplings $\mathbf{J}$. We recall here one more time that, the input to the algorithm $\mathcal{O}$ is real-valued, and that, the algorithm operates under a real-valued computational engine, e.g. using a Blum-Shub-Smale machine.

Proof Outline for Theorem 4.3.1 We provide an outline of the proof of Theorem 4.3.1 below. The basis of our proof is the $\#P-$hardness of the algorithmic problem of computing the partition function for arbitrary inputs. The argument is by contradiction. Assume, for the sake of the contradiction, that such an algorithm $\mathcal{O}(\cdot)$ exists. Let $\mathcal{Q} = (q_{ij} : 1 \leq i < j \leq n)$ be an arbitrary input of couplings, and let $\mathbf{a} = (a_{ij} : 1 \leq i < j \leq n)$ be such that $a_{ij} = \exp(q_{ij})$. Recall that it is $\#P-$hard to compute the associated partition function for *arbitrary* inputs, thus it is illustrative to think of $\mathbf{a}$ as a *worst-case input*. Define also $\mathbf{X} = (X_{ij} : 1 \leq i < j \leq n)$ with $X_{ij} = \exp(2J_{ij})$. We consider a standard interpolation $\mathbf{X}(t) = (1-t)\mathbf{X} + t\mathbf{a}$, $t \in [0,1]$, between a random input $\mathbf{X}$ and an *arbitrary* input $\mathbf{a}$.

We next define the (univariate) polynomial

$$f(t) = \widehat{Z}(\mathbf{X}(t)) = \sum_{\boldsymbol{\sigma} \in \{-1,1\}^n} \prod_{i<j : \sigma_i \neq \sigma_j} X_{ij}(t).$$

Namely, $f(t)$ corresponds to the "partition function" of an $n-$spin system whose parameters are stored in the vector $\mathbf{X}(t)$. It is not hard to see that $\deg(f) = n^2/2 + o(n)$, and $f(1) = \widehat{Z}(\mathbf{a})$. Our ultimate goal is to establish that f can be recovered from its noisy samples, and consequently, $f(1) = \widehat{Z}(\mathbf{a})$ can be computed (whp) in polynomial time, which by assumption is $\#P-$hard.

We next show that the total variation distance between $\mathbf{X}(0)$ and $\mathbf{X}(t)$ is small, when t is small. That is, the distributions of $\mathbf{X}(0)$ and $\mathbf{X}(t)$ are close, when t is small. Considering now a coupling between $\mathbf{X}(0)$ and $\mathbf{X}(t)$; we establish that $\mathcal{O}$ can also compute $f(t) = \widehat{Z}(\mathbf{X}(t))$ with probability at least $\frac{3}{4} + \frac{\delta}{2}$ (which is very close to the success probability $\frac{3}{4} + \delta$ that $\mathcal{O}(\cdot)$ has as per Theorem 4.3.1), when t is small.

Using $\mathcal{O}$, we then generate a list of sufficiently many (noisy) samples of the (low-degree) polynomial $f(\cdot)$. Provided that the number of points in the list at which f is evaluated correctly is sufficient, one can use the Berlekamp-Welch decoder to

reconstruct $f(\cdot)$. This will establish that $f(1) = \widehat{Z}(\mathbf{a})$ can be computed in polynomial time, with probability at least $\frac{1}{2} + \frac{\delta}{2}$.

Finally, we repeat this process R times (with R being at most polynomial in n) and take the majority vote. Using Chernoff bound, we have that $\widehat{Z}(\mathbf{a})$ can be computed, in polynomial time, with probability at least $1 - \exp(-\Omega(n))$, contradicting with the fact that $\mathbf{a}$ is a worst-case input.

We now proceed with the full proof.

4.3.2 Proof of Theorem 4.3.1

Let $\mathcal{Q} = (q_{ij} : 1 \leq i < j \leq n)$ be an arbitrary input (of couplings), so that it is $\#P-$hard to compute the associated partition function, $\widehat{Z}(\mathbf{a})$, which is

$$\widehat{Z}(\mathbf{a}) = \sum_{\boldsymbol{\sigma} \in \{-1,1\}^n} \prod_{i<j : \sigma_i \neq \sigma_j} a_{ij},$$

with $\mathbf{a} = (a_{ij} : 1 \leq i < j \leq n)$, where $a_{ij} = e^{q_{ij}}$. In particular, $a_{ij} > 0$ for any $1 \leq i < j \leq n$. Now, let $\mathbf{J}$ be a vector with iid standard normal components, and let $\mathbf{X} = (X_{ij} : 1 \leq i < j \leq n)$ be a vector, where $X_{ij} = e^{2J_{ij}}$ for every $1 \leq i < j \leq n$. Define $\mathbf{X}(t)$ via the following interpolation:

$$\mathbf{X}(t) = (1-t)\mathbf{X} + t\mathbf{a}, \quad 0 \leq t \leq 1. \tag{4.8}$$

Let $f(t)$ be

$$f(t) = \widehat{Z}(\mathbf{X}(t)) = \sum_{\boldsymbol{\sigma} \in \{-1,1\}^n} \prod_{i<j : \sigma_i \neq \sigma_j} \left((1-t)X_{ij} + ta_{ij}\right). \tag{4.9}$$

Note that, $f(t)$ is a univariate polynomial in t, with degree

$$\deg(f) = \max_{\boldsymbol{\sigma} \in \{-1,1\}^n} |\{(i,j) : 1 \leq i < j \leq n, \sigma_i \neq \sigma_j\}| = \frac{n^2}{2} + o(n),$$

and $f(1) = \widehat{Z}(\mathbf{a})$. Assuming the existence of an algorithm $\mathcal{O}(\cdot)$ whose probability of success is at least $\frac{3}{4} + \frac{1}{\text{poly}(n)}$, we will show the existence of a randomized polynomial time algorithm which, with probability $\frac{1}{2} + \frac{1}{\text{poly}(n)}$, recovers the polynomial $f(t)$. In particular repeating this algorithm R times to compute $f(1)$, where R is chosen to be polynomial in n; and taking majority vote, the probability that an incorrect value appears more than the half of time is exponentially small by Chernoff bound. Thus, one can compute $\widehat{Z}(\mathbf{a})$ with probability at least $1 - \exp(-\Omega(n))$.

Lemma 4.3.2. *Let $\mathbf{X}(t)$ be defined as above. Fix any $1 \leq i < j \leq n$, and let $X(t) \triangleq X_{ij}(t)$. Then, there exists an absolute constant $\mathcal{C}_{ij} > 0$, depending only on a_{ij}, such that, $d_{TV}(X(t), X(0)) \leq \mathcal{C}_{ij} t$ for every $t \in [0,1]$.*

An informal, information-theoretic way, of seeing the hypothesis of Lemma 4.3.2 is as follows. Using Pinsker's inequality [85, 236], we have

$d_{TV}(X_{ij}(t), X_{ij}(0)) \leq \kappa\sqrt{D(X_{ij}(t)\|X_{ij}(0))}$, where $D(\cdot\|\cdot)$ is the KL divergence, and $\kappa > 0$ is some absolute constant. Next, using the fact that, KL divergence locally looks like the chi-square divergence, which is essentially a weighted Euclidean ℓ_2 distance between two probability distributions, defined on the same probability space, $\chi^2(\cdot\|\cdot)$ (see, e.g., [236]), one expects for t small, $D(X_{ij}(t)\|X_{ij}(0)) \approx O(t^2)$, and thus, $d_{TV}(X_{ij}(t), X_{ij}(0)) \approx O(t)$.

The full proof of this lemma is deferred to Section 4.5.9.

We next state a tensorization inequality for the total variation distance.

Lemma 4.3.3. *Let $P_1, \ldots, P_\ell$ and $Q_1, \ldots, Q_\ell$ be probability measures, defined on the same sample space Ω. Then,*

$$d_{TV}\left(\otimes_{i=1}^{\ell} P_i, \otimes_{i=1}^{\ell} Q_i\right) \leq \sum_{i=1}^{\ell} d_{TV}(P_i, Q_i).$$

While this lemma is known, we provide a proof in Section 4.5.10 for completeness.

Using Lemma 4.3.2, together with the tensorization property above, we deduce $d_{TV}(\mathbf{X}(t), \mathbf{X}(0)) \leq \frac{Cn^2t}{2}$, where

$$C = \sum_{1 \leq i < j \leq n} \mathcal{C}_{ij},$$

the sum of the constants $\mathcal{C}_{ij}$ prescribed by Lemma 4.3.2.

Now, let $L = \lceil n^2/\delta \rceil$, and $\epsilon = \frac{\delta}{2Cn^2L}$. For every $k \in [L]$, we will evaluate $\widehat{Z}(\mathbf{X}(\epsilon k))$ via the oracle $\mathcal{O}(\cdot)$, and will use these values to reconstruct $f(t)$, from which, $f(1) = \widehat{Z}(\mathbf{a})$ can be computed. Note that, with this choice of L and ϵ, $d_{TV}(\mathbf{X}(\epsilon k), \mathbf{X}(0)) \leq \frac{\delta}{4}$, for every $k \in [L]$.

Fix an arbitrary $k \in [L]$, and consider a coupling between $\mathbf{X}(\epsilon k)$ and $\mathbf{X}(0)$, which maximizes $\mathbb{P}(\mathbf{X}(\epsilon k) = \mathbf{X}(0))$. Note that, in this case, $\mathbb{P}(\mathbf{X}(\epsilon k) = \mathbf{X}(0)) \geq 1 - d_{TV}(\mathbf{X}(\epsilon k), \mathbf{X}(0))$. Define the events $\mathcal{E}_1 = \{\mathcal{O}(\mathbf{X}(\epsilon k)) = \mathcal{O}(\mathbf{X}(0))\}$, $\mathcal{E}_2 = \{\mathcal{O}(\mathbf{X}(0)) = \widehat{Z}(\mathbf{X}(0))\}$, and finally, $\mathcal{E}_3 = \{\widehat{Z}(\mathbf{X}(0)) = \widehat{Z}(\mathbf{X}(\epsilon k))\}$. Clearly, $\mathbb{P}(\mathcal{E}_1^c), \mathbb{P}(\mathcal{E}_3^c) \leq d_{TV}(\mathbf{X}(\epsilon k), \mathbf{X}(0))$; and $\mathbb{P}(\mathcal{E}_2^c) \leq \frac{1}{4} - \delta$, since $\mathbf{X}(0) = (X_{ij} : 1 \leq i < j \leq n)$ with $X_{ij} = \exp(2J_{ij})$ with $J_{ij} \stackrel{d}{=} \mathcal{N}(0, 1)$. Since

$$\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \subseteq \{\mathcal{O}(\mathbf{X}(\epsilon k)) = \widehat{Z}(\mathbf{X}(\epsilon k))\},$$

it follows that,

$$\mathbb{P}\left(\mathcal{O}(\mathbf{X}(\epsilon k)) = \widehat{Z}(\mathbf{X}(\epsilon k))\right) \geq \frac{3}{4} + \delta - 2d_{TV}(\mathbf{X}(\epsilon k), \mathbf{X}(0)) \geq \frac{3}{4} + \frac{\delta}{2}.$$

Now, let $I_1, I_2, \ldots, I_L$ be Bernoulli random variables, where for each $k \in [L]$, $I_k = 1$ if and only if $\mathcal{O}(\mathbf{X}(\epsilon k)) = \widehat{Z}(\mathbf{X}(\epsilon k))$. Clearly, $\mathbb{P}(I_k = 1) \geq \frac{3}{4} + \frac{\delta}{2}$.

Lemma 4.3.4. *Let $X_1, X_2, \ldots, X_\ell$ be Bernoulli random variables (not necessarily independent), where there exists $0 < q < 1$, such that $\mathbb{E}[X_k] \geq q$, for every $k \in [\ell]$.*

Let $0 < \epsilon < q$ be arbitrary. Then,

$$\mathbb{P}\left(\frac{1}{\ell}\sum_{k=1}^{\ell} X_k > \epsilon\right) \geq \frac{q-\epsilon}{1-\epsilon}.$$

The proof of this lemma is provided in Section 4.5.11. In particular, letting $N = \sum_{k=1}^{L} I_k$, and using Lemma 4.3.4 with $\epsilon = \frac{1}{2} + \frac{\delta}{2}$ and $q = \frac{3}{4} + \frac{\delta}{2}$, we deduce

$$\mathbb{P}\left(N \geq \left(\frac{1}{2} + \frac{\delta}{2}\right) L\right) \geq \frac{1}{2} + \frac{\delta}{2}.$$

Let $\mathcal{L} = \{(x_k, y_k) : k \in [L]\}$ where $x_k = \epsilon k$, and $y_k = \mathcal{O}(\mathbf{X}(\epsilon k))$. The next result shows, provided $N \geq \left(\frac{1}{2} + \frac{\delta}{2}\right) L$, one can recover $f(t) = \widehat{Z}(\mathbf{X}(t))$ in polynomial time.

Theorem 4.3.5 (Berlekamp-Welch). *Let f be a univariate polynomial, with $\deg(f) = d$ over any field $\mathbb{F}$. Let $\mathcal{L} = \{(x_i, y_i) : 1 \leq i \leq L\}$ be a list such that, for at least t pairs of the list, where $t > \frac{L+d}{2}$, $y_i = f(x_i)$ holds. Then, there exists an algorithm which recovers f, using at most polynomial in L and d many field operations over $\mathbb{F}$.*

Note that, provided $N \geq \left(\frac{1}{2} + \frac{\delta}{2}\right) L$, the list $\mathcal{L}$ constructed above will satisfy the requirements of Berlekamp-Welch algorithm, and therefore, the value of $f(1) = \widehat{Z}(\mathbf{a})$ can be computed efficiently, with probability $\frac{1}{2} + \frac{\delta}{2}$, using at most polynomial in n many arithmetic operations over reals.

Now we repeat this process by R times, and take majority vote. The probability that, a wrong answer will appear as a majority vote, is exponentially small, using Chernoff bound. Taking R to be polynomial in n, we deduce this process efficiently computes $\widehat{Z}(\mathbf{a})$ with probability at least $1 - \exp(-\Omega(n))$, which is known to be a $\#P-$hard problem.

4.4 Conclusion and Future Work

In this work, we have studied the average-case hardness of the algorithmic problem of exactly computing the partition function associated with the Sherrington-Kirkpatrick model of spin glass with Gaussian couplings and random external input. We have established that, unless $P = \#P$, there does not exists a polynomial time algorithm which exactly computes the partition function on average. We have established our result by combining the approach of Cai et al. [67] for establishing the average-case hardness of computing the permanent of a (random) matrix, modulo a prime number p; with a probabilistic coupling between log-normal inputs and random uniform inputs over a finite field. To the best of our knowledge, ours is the first such result, pertaining the statistical physics models. We also note that, our approach is not limited to the case of Gaussian inputs: for random variables with sufficiently well-behaved density, for which, one can establish a coupling as in Lemma 4.2.12 to a prime of appropriate size, our techniques transfer.

Several future research directions are as follows. The proof sketch outlined in this work, as well as in the previous works [67, 109, 208] do not transfer to the several other fundamental open problems aiming at establishing similar hardness results related to SK model. One such fundamental problem is the problem of exactly computing a ground state, namely, the problem of finding a state $\boldsymbol{\sigma}^* \in \{-1,1\}^n$, such that, $H(\boldsymbol{\sigma}^*) = \max_{\boldsymbol{\sigma}\in\{-1,1\}^n} H(\boldsymbol{\sigma})$. Arora et al. [16] established that the problem of exactly computing a ground state is NP-hard in the worst case sense. Furthermore, Montanari [220] recently proposed a message-passing algorithm, which, for a fixed $\epsilon > 0$ finds a state $\boldsymbol{\sigma}_*$ such that $H(\boldsymbol{\sigma}_*) \geq (1-\epsilon)\max_{\boldsymbol{\sigma}\in\{-1,1\}^n} H(\boldsymbol{\sigma})$ with high probability, in a time at most $O(n^2)$, assuming a widely-believed structural conjecture in statistical physics. Namely, it is possible to efficiently approximate the ground state of SK model within a multiplicative factor of $1-\epsilon$. The proof techniques of Cai et al. [67], as well as Lipton's approach [208], do not, however, seem to be useful in addressing the average-case hardness of the algorithmic problem of exactly computing the ground state since the algebraic structure relating the problem into the recovery of a polynomial is lost, when one considers the maximization; and this problem remains open.

Another fundamental problem, which remains open, is the average-case hardness of the problem of computing the partition function approximately, namely, computing $Z(\mathbf{J},\beta)$ to within a multiplicative factor of $(1\pm\epsilon)$, which has been of interest in the field of approximation algorithms.

Yet another natural question is whether the assumption on the oracle $\mathcal{O}(\cdot)$ in Theorem 4.3.1 for the real-valued computational model, that is,

$$\mathbb{P}\left(\mathcal{O}(\mathbf{X}) = \widehat{Z}(\mathbf{X})\right) \geq \frac{3}{4} + \frac{1}{\text{poly}(n)}$$

can be weakened e.g., to $1/2+1/\text{poly}(n)$ or even to $1/\text{poly}(n)$, as handled in the finite-precision setting. As we have mentioned previously, our approach for establishing the average-case hardness of the problem of exact computation of the partition function under the finite-precision arithmetic model is in parallel with the line of research dealing with the average-case hardness of computing the permanent over a finite field. A typical result along these lines is obtained under the assumption that there exists an oracle which computes the permanent with a certain probability of success, q. The first such result, under the weakest assumption of $q = 1-1/3n$, is obtained by Lipton [208]. Subsequent research weakened this assumption to $q = 3/4 + 1/\text{poly}(n)$ by Gemmell et al. [145], then to $q = 1/2 + 1/\text{poly}(n)$ by Gemmell and Sudan [146]; and finally to $q = 1/\text{poly}(n)$, by Cai et al. [67].

The assumption on the success probability of the oracle that we have adopted in this work for the real-valued computational model is similar to that of Gemmell et al. [145], and thus, the most natural question is to ask, whether, at the very least, the technique of Gemmell and Sudan [146] can be applied. We now discuss that this seems to be a challenging task, and show where the extension fails.

The idea of Gemmell and Sudan, essentially, aims at reconstructing a certain polynomial (similar to (4.9)), which is observed through its noisy samples (e.g., similar to the list $\mathcal{L} = \{(x_k, \mathcal{O}(\epsilon k)) : k \in [L]\}$, that we have defined earlier), and is adapted

to our case as follows. Let $\mathbf{J} = (J_{ij} : 1 \leq i < j \leq n)$ and $\mathbf{J}' = (J'_{ij} : 1 \leq i < j \leq n)$ be two iid random vectors, each with iid standard normal components, and let $\mathbf{X} = (X_{ij} : 1 \leq i < j \leq n)$ and $\mathbf{X}' = (X'_{ij} : 1 \leq i < j \leq n)$, where $X_{ij} = e^{2J_{ij}}, X'_{ij} = e^{2J'_{ij}}$, for $1 \leq i < j \leq n$. Define:

$$\mathbf{X}(t) = t(1-t)\mathbf{X} + (1-t)\mathbf{X}' + t^2\mathbf{a}, \tag{4.10}$$

where $\mathbf{a}$ is a worst-case input. Note that, the sampling set $\{\mathbf{X}(t) : t \in [0,1]\}$ is defined more carefully, by incorporating an extra randomness via $\mathbf{X}'$ (cf. equation (4.8)). The purpose of this extra randomness in Gemmell and Sudan's work was to bring pairwise independence, that is to ensure the independence of $\mathbf{X}(t)$ and $\mathbf{X}(t')$ for $t \neq t'$, in order to be able to use a tighter concentration inequality (namely, Chebyshev's inequality) as a replacement of our Lemma 4.3.4 while obtaining a high probability guarantee on the constructed list. In their work, this is successful: $\mathbf{X}$ and $\mathbf{X}'$ consist of iid samples, drawn independently from uniform distribution over a finite field $\mathbb{F}_p$, in which case, it is not hard to show, $\mathbf{X}(t)$ and $\mathbf{X}(t')$ are always independent for $t \neq t'$. For us, however, this is no longer true: $\mathbf{X}$ and $\mathbf{X}'$ both consist of iid log-normal components, which breaks down uniformity and independence.

We leave the following problem open for future work: Let $\mathbf{J} = (J_{ij} : 1 \leq i < j \leq n) \in \mathbb{R}^{n(n-1)/2}$ be a random vector with $J_{ij} \stackrel{d}{=} \mathcal{N}(0,1)$, i.i.d; and let $\mathbf{X} = (X_{ij} : 1 \leq i < j \leq n) \in \mathbb{R}^{n(n-1)/2}$ with $X_{ij} = \exp(2J_{ij})$, $1 \leq i < j \leq n$, as in the setting of Theorem 4.3.1. Suppose that, there is an algorithm $\mathcal{A}(\cdot)$, such that

$$\mathbb{P}(\widehat{Z}(\mathbf{X}) = \mathcal{A}(\mathbf{X})) \geq \frac{1}{2} + \frac{1}{\text{poly(n)}},$$

and that, the algorithm operates over real-valued inputs. Then $P = \#P$. An even more challenging variant of this problem is to establish the same result, under a weaker assumption on the success probability of the algorithm:

$$\mathbb{P}(\widehat{Z}(\mathbf{X}) = \mathcal{A}(\mathbf{X})) \geq \frac{1}{\text{poly(n)}},$$

like we established under the finite-precision computational model.

As we have noted, our approach is not limited to the Gaussian inputs, so long as the distributions involved are well-behaved. The current method, however, does not address the case of couplings with iid Rademacher inputs, and the average-case hardness of the exact computation of partition function with iid Rademacher couplings remains open. It is not surprising though in light of the fact that the average-case hardness of the problem of computing the permanent of a matrix with 0/1 entries remains open, as well.

4.5 Appendix : Proofs of the Technical Lemmas

4.5.1 Proof of Lemma 4.2.11

Proof. The density of $\widehat{J}_{ij}$ is given by

$$\begin{aligned} f_{\widehat{J}}(t) &= \frac{d}{dt}\mathbb{P}\left(e^{\frac{\beta}{\sqrt{n}}J} \le t\right) \\ &= \frac{d}{dt}\mathbb{P}\left(J \le \sqrt{n}\frac{\log t}{\beta}\right) \\ &= \frac{\sqrt{n}}{\sqrt{2\pi}\beta t}e^{-n\frac{\log^2 t}{2\beta^2}}. \end{aligned}$$

Here J denotes the standard normal random variable. It is easy to see that

$$f_{\widehat{J}}(t) = O\left(\sqrt{n}t\right), \tag{4.11}$$

as $t \downarrow 0$ since $e^{\log^2 x}$ diverges faster than x^c for every constant c as $x \to \infty$. Also

$$f_{\widehat{J}}(t) = O\left(\frac{\sqrt{n}}{t^2}\right), \tag{4.12}$$

as $t \to \infty$. Both bounds are very crude of course, but suffice for our purposes.

We have for every $t, \tilde{t} > 0$

$$|\log f_{\widehat{J}}(\tilde{t}) - \log f_{\widehat{J}}(t)| \le |\log(\tilde{t}) - \log t| + \frac{n}{2\beta^2}|\log^2(\tilde{t}) - \log^2(t)|.$$

Now since $|\frac{d\log t}{dt}| = 1/t \le 1/\delta$ for $t \ge \delta$, we obtain that in the range $0 < \delta \le t, \tilde{t} \le \Delta$

$$\begin{aligned} |\log f_{\widehat{J}}(\tilde{t}) - \log f_{\widehat{J}}(t)| &\le (1/\delta)|\tilde{t} - t|, \\ |\log^2(\tilde{t}) - \log^2(t)| &\le \frac{2\log\Delta}{\delta}|\tilde{t} - t|. \end{aligned}$$

Applying these bounds, exponentiating, and using the assumption on the lower bound on $\log\Delta$ and $n > \beta^2$, we obtain

$$\exp\left(-\frac{2n\log\Delta}{\beta^2\delta}|\tilde{t} - t|\right) \le \frac{f_{\widehat{J}}(\tilde{t})}{f_{\widehat{J}}(t)} \le \exp\left(\frac{2n\log\Delta}{\beta^2\delta}|\tilde{t} - t|\right). \tag{4.13}$$

□

Remark 4.5.1. *Let $\widehat{B}_i = e^{B_i}$ and $\widehat{C}_i = e^{C_i}$; and denote the (common) densities by $f_{\widehat{B}}$ and $f_{\widehat{C}}$. $f_{\widehat{B}}(t) = f_{\widehat{C}}(t) = \frac{1}{\sqrt{2\pi}t}\exp\left(-\frac{\log^2 t}{2}\right)$, and therefore, as $t \downarrow 0$, $f_{\widehat{B}}(t) = O(t) = O(\sqrt{n}t)$, and furthermore, as $t \to \infty$, $f_{\widehat{B}}(t) = O(1/t^2) = O(\sqrt{n}/t^2)$. Similarly, the same Lipschitz condition holds, also for $f_{\widehat{B}}(t)$ and $f_{\widehat{C}}(t)$, and therefore, the result of Lemma 4.2.11 applies also to the exponentiated version of the external*

field components. Note also that, we still have the same asymptotic behaviour, even if the external field components B_i and C_i have a constant variance, different than 1.

4.5.2 Proof of Lemma 4.2.3

Proof. We begin by deriving a downward self recursion formula for $I_n(\boldsymbol{\sigma}) = |\{(i,j) : 1 \leq i < j \leq n, \sigma_i \neq \sigma_j\}|$. Note that, for a given spin configuration $\boldsymbol{\sigma} \in \{-1,1\}^n$ if $\sigma_n = +1$, then $I_n(\boldsymbol{\sigma}) = I_{n-1}(\boldsymbol{\sigma}) + |\{i : \sigma_i = -1, 1 \leq i \leq n-1\}|$, where we take the projection of $\boldsymbol{\sigma}$ onto its first $(n-1)$ coordinates. Similarly, if $\sigma_n = -1$, then $I_n(\boldsymbol{\sigma}) = I_{n-1}(\boldsymbol{\sigma}) + |\{i : \sigma_i = +1, 1 \leq i \leq n-1\}|$. For a given spin configuration $\boldsymbol{\sigma}$, and dimension $n-1$, recalling the definition of $f(n,\boldsymbol{\sigma})$ in (4.3), we observe that for $\sigma_n = +1$

$$f(n,\boldsymbol{\sigma}) - f(n-1,\boldsymbol{\sigma}) = (n-2) - |\{i : \sigma_i = -1, 1 \leq i \leq n-1\}|,$$

and similarly, for $\sigma_n = -1$,

$$f(n,\boldsymbol{\sigma}) - f(n-1,\boldsymbol{\sigma}) = (n-2) - |\{i : \sigma_i = +1, 1 \leq i \leq n-1\}|.$$

Now, observe that, using the relation between $f(n,\boldsymbol{\sigma})$ and $f(n-1,\boldsymbol{\sigma})$ with respect to polarity of σ_n, we have:

$$\begin{aligned}
Z_n(\mathbf{J},\mathbf{B},\mathbf{C};p_n) &= C_n 2^{(n-2)N} \sum_{\substack{\boldsymbol{\sigma}\in\{-1,1\}^{n-1}\\ \sigma_n=+1}} 2^{Nf(n-1,\boldsymbol{\sigma})} \left[\left(\prod_{\substack{1\leq i\leq n-1\\ \sigma_i=-}} 2^{-N} B_i J_{in} \right) \right. \\
&\quad \left. \cdot \left(\prod_{\substack{1\leq i\leq n-1\\ \sigma_i=+}} C_i \right) \left(\prod_{\substack{1\leq i<j\leq n-1\\ \sigma_i\neq\sigma_j}} J_{ij} \right) \right] \\
&\quad + B_n 2^{(n-2)N} \sum_{\substack{\boldsymbol{\sigma}\in\{-1,1\}^{n-1}\\ \sigma_n=-1}} 2^{Nf(n-1,\boldsymbol{\sigma})} \left[\left(\prod_{\substack{1\leq i\leq n-1\\ \sigma_i=-}} B_i \right) \right. \\
&\quad \left. \cdot \left(\prod_{\substack{1\leq i\leq n-1\\ \sigma_i=+}} 2^{-N} C_i J_{in} \right) \left(\prod_{\substack{1\leq i<j\leq n-1\\ \sigma_i\neq\sigma_j}} J_{ij} \right) \right] \\
&= C'_n Z_{n-1}(\mathbf{J}',\mathbf{B}^+,\mathbf{C}^+;p_n) + B'_n Z_{n-1}(\mathbf{J}',\mathbf{B}^-,\mathbf{C}^-;p_n).
\end{aligned}$$

□

4.5.3 Proof of Lemma 4.2.4

Proof. Fix an $1 \leq i \leq T$, and let ξ_i denote the i^{th} component of an arbitrary vector ξ. Note that, the event $\{D(x_1) = y_1, D(x_2) = y_2\}$ implies:

$$(y_1)_i = (2 - x_1)(v_1)_i + (x_1 - 1)(v_2)_i + (x_1 - 1)(x_1 - 2)(K_i + x_1 M_i)$$
$$(y_2)_i = (2 - x_2)(v_1)_i + (x_2 - 1)(v_2)_i + (x_2 - 1)(x_2 - 2)(K_i + x_2 M_i).$$

Since this is a pair of equations with two unknowns (namely, K_i and M_i), it has a unique solution, which holds with probability $1/p_n^2$ (note that, $x_1, x_2 \notin \{1, 2\}$, hence for $i = 1, 2$, $(x_i - 1)(x_i - 2)$ terms are not zero, and thus their modulo p_n inverse exists). Finally, using independence across $i \in \{1, 2, \ldots, T\}$, we get $\mathbb{P}(D(x_1) = y_1, D(x_2) = y_2) = 1/p_n^{2T}$. For $\mathbb{P}(D(x_1) = y_1)$, it is not hard to show by conditioning that, this event has probability $1/p_n^T$. □

4.5.4 Proof of Lemma 4.2.5

Proof. Let $\mathbb{N}_x \in \{0, 1\}$, $x = 3, 4, \ldots, p_n$, be random variables, where $\mathbb{N}_x = 1$ iff $\mathcal{A}(D(x)) = \phi(x) = Z_{n-1}(D(x); p_n)$. Namely, $\mathbb{N}_x \sim \text{Ber}(q)$. Note that, $\mathbb{N} = \sum_{x=3}^{p_n} \mathbb{N}_x$. Let $Z = \mathbb{N}/(p_n - 2)$. We have $\mathbb{E}[Z] = q$. Hence,

$$\begin{aligned}\mathbb{P}(\mathbb{N} < (p_n - 2)q/2) = \mathbb{P}\left(\frac{\sum_{x=3}^{p_n} \mathbb{N}_x}{p_n - 2} < q/2\right) &= \mathbb{P}(Z - \mathbb{E}[Z] < -q/2) \\ &\leq \mathbb{P}(|Z - \mathbb{E}[Z]| > q/2) \\ &\leq \frac{\text{Var}(Z)}{(q/2)^2} \leq \frac{1}{(p_n - 2)q^2},\end{aligned}$$

by Chebyshev's inequality, and the trivial inequality, $4q - 4q^2 \leq 1$. Note that, since we only have pairwise independence as opposed to iid, a Chernoff-type bound do not apply. □

4.5.5 Proof of Lemma 4.2.6

Proof. Assume the contrary, and take a subset $\mathcal{F}' \subseteq \mathcal{F}$ with $|\mathcal{F}'| = \lceil 3/q \rceil$. Let, $G_{\mathcal{S}}(f) = \{i : (i, f(i)) \in G(f) \cap \mathcal{S}\}$. Note that, $\bigcup_{f \in \mathcal{F}} G_{\mathcal{S}}(f) \subseteq \{3, 4, \ldots, p_n\}$, and furthermore, for any distinct $f, f' \in \mathcal{F}'$, it holds that, $|G_{\mathcal{S}}(f) \cap G_{\mathcal{S}}(f')| \leq n^2 - 1$. Indeed, if not, define $\widehat{f} = f - f'$, and observe that $\deg(\widehat{f}) \leq n^2 - 1$. If $|G_{\mathcal{S}}(f) \cap G_{\mathcal{S}}(f')| \geq n^2$, then, on at least n^2 values of i, $f(i) = f'(i)$, and thus, $\widehat{f}(i) = 0$, yielding that $\widehat{f}$ has at least n^2 distinct zeroes (modulo p_n), a contradiction to the

degree of $\widehat{f}$. Now, using inclusion-exclusion principle,

$$\begin{aligned} p_n - 2 \geq \left| \bigcup_{f \in \mathcal{F}'} G_{\mathcal{S}}(f) \right| &\geq \sum_{f \in \mathcal{F}'} |G_{\mathcal{S}}(f)| - \sum_{f,f' \in \mathcal{F}', f \neq f'} |G_{\mathcal{S}}(f) \cap G_{\mathcal{S}}(f')| \\ &\geq \lceil \frac{3}{q} \rceil \frac{(p_n - 2)q}{2} - \frac{1}{2} \lceil \frac{3}{q} \rceil (\lceil \frac{3}{q} \rceil - 1)(n^2 - 1) \\ &= \frac{1}{2} \lceil \frac{3}{q} \rceil \left((p_n - 2)q - (\lceil 3/q \rceil - 1)(n^2 - 1) \right) \\ &\geq (p_n - 2) + \frac{p_n - 2}{2} - \frac{3}{2q} (\lceil 3/q \rceil - 1)(n^2 - 1). \end{aligned}$$

However, contradicting with this inequality, we claim that in fact $p_n - 2 > \frac{3}{q}(\lceil 3/q \rceil - 1)(n^2 - 1)$. Since $\lceil 3/q \rceil < 3/q + 1$, it is sufficient to show that, $p_n - 2 > \frac{9}{q^2}(n^2 - 1)$. Since $q \geq 1/n^k$, we have $\frac{9}{q^2}(n^2 - 1) \leq 9n^{2k}(n^2 - 1) = 9n^{2k+2} - 9n^{2k} < p_n - 2$, for n large (for any k). Hence, we arrive at a contradiction. □

4.5.6 Proof of Lemma 4.2.7

Proof. We condition on the high probability event, $\{\mathbb{N} \geq (p_n - 2)q/2\}$, where $\mathbb{N}$ is the random variable defined in Lemma 4.2.5. We divide the construction, into two cases, depending on the magnitude of p_n that we are working at.

First, suppose $9n^{2k+2} \leq p_n \leq 161n^{3k+2}$. Apply $\mathcal{A}$ on $D(x)$, for every $x = 3, 4, \ldots, p_n$ (which, due to magnitude constraint on p_n, takes at most polynomial in n many operations). By Lemma 4.2.5, with probability at least $1 - \frac{1}{(p_n - 2)q^2}$, $\mathcal{A}(D(x)) = \phi(x) = Z_{n-1}(D(x); p_n)$ for at least $\frac{(p_n - 2)q}{2}$ points. Now, since $q \geq 1/n^k$, we have a list $(x_i, y_i)_{i=1}^{L}$ (where $L = p_n - 2$ and $y_i = \mathcal{A}(D(x))$), and there is a polynomial f of degree d less than n^2 (namely, $\phi(x) = Z_{n-1}(D(x); p_n)$), such that, the graph of f intersects the list at at least $t = \frac{p_n - 2}{2n^k}$ points. As $p_n \geq 9n^{2k+2}$, it holds that $t > \sqrt{2Ld}$. Clearly, for all such pairs, the first coordinates are all distinct.

Next, suppose $p \geq 161n^{3k+2}$. In this case, it is not clear, whether running the algorithm on $\{D(x) : x = 3, 4, \ldots, p_n\}$ takes polynomial in n many calls to $\mathcal{A}$. To handle this issue, we apply the following resampling procedure (where the choice of numbers is to make sure the argument works). Select $L = 40n^{2k+2}$ numbers $x_1, x_2, \ldots, x_L$, uniformly and independently from $\{3, 4, \ldots, p_n\}$. Our goal is to find a lower bound on the number of x_i's, for which with high probability we have at least a certain number of distinct x_i's, on which $\mathcal{A}$ run correctly. We claim that, with high probability, we will end up with at least $9n^{k+2}$ distinct x_i's on which $\mathcal{A}(D(x_i)) = Z_{n-1}(D(x_i); p_n)$. We argue as follows. Define a collection $\{E_j : 1 \leq j \leq L\}$ of events,

$$E_j = \{x_j \neq x_i, \text{ for } i \leq j - 1, \mathcal{A}(D(x_j)) = \phi(x_j) = Z_{n-1}(D(x_j); p_n)\}.$$

Namely, E_j is the event that, (x_j, y_j) is a 'nice' sample, in the sense that, x_j is distinct from all preceding x_i's, and $y_j = \phi(x_j) = Z_{n-1}(D(x_j); p_n)$. Now, we can change the perspective slightly, and imagine that, (x_j, y_j) is samples from a set, where $x_j \in$

$\{3, 4, \ldots, p_n\}$, and $y_j = \mathcal{A}(D(x_j))$. Recall that, among the set $\{D(x) : x = 3, \ldots, p_n\}$, the algorithm computes the partition function on at least $\frac{(p_n-2)q}{2} \geq \frac{p_n-2}{2n^k}$ locations (conditional on the high probability event $\{\mathcal{N} \geq (p_n - 2)q/2\}$ of Lemma 4.2.5). Note that,

$$\mathbb{P}(E_j) \geq \frac{\frac{p_n-2}{2n^k} - L}{p_n - 2} = \frac{1}{2n^k} - \frac{40n^{2k+2}}{161n^{3k+2} - 2} \geq \frac{1}{4n^k},$$

since, the worst case for E_j is that, all preceding chosen entries are distinct, leaving less number of choices for x_j, and we repeat the procedure L times. With this, we now claim that with high probability, at least $9n^{k+2}$ of events $(E_j)_{j=1}^{L}$ occur. To see this, we note that, the event of interest ($9n^{k+2}$ of events $(E_j : j \in L)$ occur), is stochastically dominated by the event that, a binomial random variable $\mathrm{Bin}(L, 1/4n^k)$, whose expectation is $L/4n^k = 10n^{k+2}$ is at least $L = 9n^{k+2}$, which, by a standard Chernoff bound, is exponentially small. At the end , we have a list of $L = 40n^{2k+2}$ pairs, $(x_i, y_i)_{i=1}^{L}$, on which we have at least $t \geq 9n^{k+2}$ correct evaluations (whp), where $t \geq 9n^{k+2} > \sqrt{2Ld}$ with $d = n^2$. □

4.5.7 Proof of Lemma 4.2.10

Proof. Suppose, this is false, and the number of primes between $9n^{2k+2}$ and $2(2 + \alpha + 2k)Nn^{2k+2} \log n$ is at most Nn^{2k+2}, for all large n. Recall that, prime number theorem (PNT) states,

$$\lim_{m\to\infty} \frac{\pi(m)}{m/\log m} = 1,$$

where $\pi(m) = \sum_{p\leq m, p \text{ prime}} 1$ is the prime counting function. Now we have, for $m \triangleq 2(2 + \alpha + 2k)Nn^{2k+2} \log n$, $\pi(m) \leq Nn^{2k+2} + 9n^{2k+2} = Nn^{2k+2}(1 + o(1))$. Now, using $N \leq n^\alpha$, we have, $\log m \leq (2 + \alpha + 2k + o(1)) \log n$, and therefore,

$$\frac{m}{\log m} \geq \frac{2(2 + \alpha + 2k)Nn^{2k+2} \log n}{(2 + \alpha + 2k + o(1)) \log n} = 2(1 - o(1))Nn^{2k+2},$$

and since $\pi(m) \leq Nn^{2k+2}(1 + o(1))$, we get a contradiction with PNT, for n large enough. □

4.5.8 Proof of Lemma 4.2.12

Proof. We have for every $\ell \in [0, p_n - 1]$

$$\mathbb{P}(A = \ell \bmod (p_n)) = \sum_{m\in\mathbb{Z}} \int_{\frac{mp_n+\ell}{2^N}}^{\frac{mp_n+\ell+1}{2^N}} f_X(t)\, dt.$$

We now let,

$$M^*(n) = \frac{n^{5k+9/2}N2^N}{p_n} \quad \text{and} \quad M_*(n) = \frac{2^N}{Nn^{5k/2+3}p_n}.$$

Note the following bound on the size of $p_n = o(Nn^{3k+2})$, due to Lemma 4.2.10. We now consider separately the case $m \in [M_*(n), M^*(n)-1]$ and $m \notin [M_*(n), M^*(n)-1]$. For $m \in [M_*(n), M^*(n)-1]$ applying Lemma 4.2.11 with

$$\delta = \frac{M_*(N)p_n}{2^N}$$
$$\Delta = \frac{M^*(N)p_n}{2^N},$$

we have for very t and $\tilde{t}$ such that

$$2^N\tilde{t} \in [mp_n + \ell, mp_n + \ell + 1]$$
$$2^N t \in [mp_n, mp_n + 1]$$

$$\frac{f_X\left(\tilde{t}\right)}{f_X\left(t\right)} \leq \exp\left(\frac{2n2^N \log\left(\frac{M^*(n)p_n}{2^N}\right)}{\beta^2 M_*(n)p_n}|\tilde{t}-t|\right).$$

Since $|\tilde{t}-t| \leq p_n/2^N$, we obtain

$$\frac{f_X\left(\tilde{t}\right)}{f_X\left(t\right)} \leq \exp\left(\frac{2n \log\left(\frac{M^*(n)p_n}{2^N}\right)}{\beta^2 M_*(n)}\right).$$

Applying the value of and $M^*(n)$ we have $\log\left(\frac{M^*(n)p_n}{2^N}\right) = O(\log n)$. Given an upper bound $p_n = O(Nn^{3k+2})$, we have that the exponent is

$$O\left(\frac{n\log n}{M_*(n)}\right) = O\left(\frac{n^{11k/2+6}N^2}{2^N}\right).$$

We now claim

$$O\left(\frac{n^{11k/2+6}N^2}{2^N}\right) = O\left(N^{-1}n^{-5k-4}\right).$$

To show this, it suffices to show

$$N^3 n^{\frac{21k}{2}+10} = O\left(2^N\right)$$

We now verify this: note that this is true if there exists a constant $\mathcal{C} > 0$ such that for all sufficiently large n (and N):

$$2^N > \mathcal{C}N^3 n^{\frac{21k}{2}+10} \iff N > \log\mathcal{C} + 3\log N + \left(\frac{21k}{2}+10\right)\log n.$$

Fix $\mathcal{C} > 0$ arbitrary, and observe that

$$N - 3\log N - \log \mathcal{C} > \frac{N}{2}$$

for all N that is sufficiently large. Thus it suffices to ensure

$$N > (21k + 20)\log n.$$

But due to the hypothesis on N stating $N \geq C(k)\log n$ with $C(k) = 21k + 20 + \epsilon$, for some $\epsilon > 0$, we indeed have this. Thus the term above is

$$O\left(N^{-1}n^{-5k-4}\right).$$

We obtain a bound

$$\exp\left(O\left(N^{-1}n^{-5k-4}\right)\right) = 1 + O\left(N^{-1}n^{-5k-4}\right).$$

Similarly, we obtain for the same range of $t, \tilde{t}$

$$\frac{f_X\left(\tilde{t}\right)}{f_X\left(t\right)} \geq 1 - O\left(N^{-1}n^{-5k-4}\right).$$

Thus

$$\begin{aligned}
&|\sum_{M_*(n)\leq m\leq M^*(n)}\left(\int_{\frac{mp_n+\ell}{2^N}}^{\frac{mp_n+\ell+1}{2^N}} f_X(t)dt - \int_{\frac{mp_n}{2^N}}^{\frac{mp_n+1}{2^N}} f_X(t)dt\right)| \\
&|\sum_{M_*(n)\leq m\leq M^*(n)}\int_{\frac{mp_n}{2^N}}^{\frac{mp_n+1}{2^N}}\left(f_X\left(t+\frac{\ell}{2^N}\right) - f_X(t)\right)dt| \\
&\leq O\left(N^{-1}n^{-5k-4}\right)\sum_{M_*(n)\leq m\leq M^*(n)}\int_{\frac{mp_n}{2^N}}^{\frac{mp_n+1}{2^N}} f_X(t)dt \\
&= O\left(N^{-1}n^{-5k-4}\right),
\end{aligned}$$

as the sum above is at most the integral of the density function, and thus at most 1.

We now consider the case $m \leq M_*(n)$. We have applying (4.11)

$$\int_0^{\frac{M_*(n)p_n}{2^N}} f_X(t)dt = O\left(\left(\frac{M_*(n)p_n}{2^N}\right)^2\sqrt{n}\right)$$

which applying the value of $M_*(n)$ is $O\left(N^{-2}n^{-5k-6+1/2}\right) = O\left(N^{-1}n^{-5k-4}\right)$.

Finally, suppose $m \geq M^*(n)$. Applying (4.12)

$$\int_{t\geq\frac{M^*(n)p_n}{2^N}} f_X(t)dt = O\left(\frac{\sqrt{n}}{\frac{M^*(n)p_n}{2^N}}\right) = O(N^{-1}n^{-5k-4}).$$

We conclude that

$$\max_{0\le \ell\le p_n-1} |\mathbb{P}(A_{ij}=\ell \bmod (p_n)) - \mathbb{P}(A_{ij}=0 \bmod (p_n))| = O(N^{-1}n^{-5k-4}).$$

Thus

$$\begin{aligned}
&\mathbb{P}(A_{ij}=\ell \bmod (p_n)) - p_n^{-1} \\
&= \frac{p_n\mathbb{P}(A_{ij}=\ell \bmod (p_n)) - \sum_\ell \mathbb{P}(A_{ij}=\ell \bmod (p_n))}{p_n} \\
&\le \frac{p_n\left(\mathbb{P}(A_{ij}=0 \bmod (p_n)) + O(N^{-1}n^{-5k-4})\right) - p_n\left(\mathbb{P}(A_{ij}=0 \bmod (p_n)) - O(N^{-1}n^{-5k-4})\right)}{p_n} \\
&= O(N^{-1}n^{-5k-4}),
\end{aligned}$$

completing the proof of the lemma. A lower bound $O(N^{-1}n^{-5k-4})$ is shown similarly. □

4.5.9 Proof of Lemma 4.3.2

Proof. Let $X(\lambda) = (1-\lambda)e^J + \lambda a$, with $a > 0$, and $J \stackrel{d}{=} \mathbb{N}(0,4)$. Note that, the density of J is $f_J(t) = \frac{1}{\sqrt{8\pi}}\exp(-t^2/8)$, for every $t \in \mathbb{R}$. Fix a $\lambda_0 \in (0,1)$. Note that since the total variation distance is upper bounded by one, it suffices to establish that there exists a constant $\mathcal{C}_{\lambda_0}$, such that

$$d_{TV}(X(\lambda), X(0)) \le \mathcal{C}_{\lambda_0}\lambda, \qquad \forall \lambda \in [0, \lambda_0].$$

We begin with a calculation of the density of $X(\lambda)$. Note that, the density of $X(\lambda)$ is supported on $[\lambda a, \infty)$. Fix a $t \ge \lambda a$. Observe that, $\mathbb{P}\left(X(\lambda) \le t\right) = \mathbb{P}\left(J \le \log\left(\frac{t-\lambda a}{1-\lambda}\right)\right)$, and thus, differentiation with respect to t yield the density of $X(\lambda)$ to be:

$$f_{X(\lambda)}(t) = \frac{1}{\sqrt{8\pi}(t-\lambda a)}\exp\left(-\frac{1}{8}\log\left(\frac{t-\lambda a}{1-\lambda}\right)^2\right), \qquad \forall t \ge \lambda a.$$

Now let

$$f_X(t) = \frac{1}{\sqrt{8\pi}t}\exp\left(-\frac{1}{8}\log(t)^2\right),$$

be the density of the log-normal e^J, with $J \stackrel{d}{=} \mathbb{N}(0,4)$. Observe that, wherever it is defined,

$$f_{X(\lambda)}(t) = \frac{1}{1-\lambda}f_X\left(\frac{t-\lambda a}{1-\lambda}\right).$$

Recall next the definition of the TV distance, for two continuous random variables Y, Z with densities f_Y and f_Z, respectively: $d_{TV}(Y,Z) = \frac{1}{2}\int |f_Y(t) - f_Z(t)|\, dt$. In

particular, we need to control the following quantity:

$$d_{TV}(X(\lambda), X(0)) = \frac{1}{2}\int_{-\infty}^{\infty} |f_{X(\lambda)}(t) - f_X(t)|\, dt \tag{4.14}$$

$$= \frac{1}{2}\int_{0}^{\lambda a} f_X(t)\, dt + \frac{1}{2}\int_{\lambda a}^{\infty} |f_{X(\lambda)}(t) - f_X(t)|\, dt \tag{4.15}$$

$$\leq \frac{1}{2}\mathcal{M}_1 \lambda a + \frac{1}{2}\int_{\lambda a}^{\infty} |f_{X(\lambda)}(t) - f_X(t)|\, dt \tag{4.16}$$

where $\mathcal{M}_1 = \sup_{t\in\mathbb{R}} f_X(t)$, which is easily found to be finite. With this, we now focus on bounding the second term:

$$\begin{aligned}
|f_{X(\lambda)}(t) - f_X(t)| &= \left| \frac{1}{1-\lambda} f_X\left(\frac{t-\lambda a}{1-\lambda}\right) - f_X(t) \right| \\
&\leq \left| \frac{1}{1-\lambda} f_X\left(\frac{t-\lambda a}{1-\lambda}\right) - \frac{1}{1-\lambda} f_X(t) \right| + \left| \frac{1}{1-\lambda} f_X(t) - f_X(t) \right| \\
&\leq \frac{1}{1-\lambda_0} \left| f_X\left(\frac{t-\lambda a}{1-\lambda}\right) - f_X(t) \right| + \frac{\lambda}{1-\lambda_0} f_X(t),
\end{aligned}$$

where the first inequality uses the triangle inequality, and the second inequality uses the fact that $\lambda \leq \lambda_0 < 1$. We then have:

$$\int_{\lambda a}^{\infty} |f_{X(\lambda)}(t) - f_X(t)|\, dt \leq \frac{1}{1-\lambda_0}\int_{\lambda a}^{\infty} \left| f_X\left(\frac{t-\lambda a}{1-\lambda}\right) - f_X(t) \right|\, dt + \frac{\lambda}{1-\lambda_0}\int_{\lambda a}^{\infty} f_X(t)\, dt \tag{4.17}$$

$$\leq \frac{1}{1-\lambda_0}\int_{\lambda a}^{\infty} \left| f_X\left(\frac{t-\lambda a}{1-\lambda}\right) - f_X(t) \right|\, dt + \frac{\lambda}{1-\lambda_0}, \tag{4.18}$$

using the fact that $f_X(t)$ is a legitimate density, and thus, $f_X(t) \geq 0$ and $\int_0^{\infty} f_X(t) = 1$. Combining everything we have thus far, in particular, Equations (4.16) and (4.18); we arrive at:

$$d_{TV}(X(\lambda), X(0)) \leq \lambda\left(\frac{1}{2(1-\lambda_0)} + \frac{a}{2}\mathcal{M}_1\right) + \frac{1}{2(1-\lambda_0)}\int_{\lambda a}^{\infty} \left| f_X\left(\frac{t-\lambda a}{1-\lambda}\right) - f_X(t) \right|\, dt, \tag{4.19}$$

where $\mathcal{M}_1 = \sup_{t\in\mathbb{R}} f_X(t)$, is the maximum value of the log-normal density, which is a finite absolute constant. The remaining task is to bound the integral in Equation (4.19). Now, let

$$I_{t,\lambda} = \left(\min\left(\frac{t-\lambda a}{1-\lambda}, t\right), \max\left(\frac{t-\lambda a}{1-\lambda}, t\right) \right).$$

We now make the following observation:

$$\frac{t-\lambda a}{1-\lambda} \geq t \iff t - \lambda a \geq t - \lambda t \iff t \geq a.$$

Namely, we have that for $t \geq a$:

$$I_{t,\lambda} = \left(t, \frac{t-\lambda a}{1-\lambda}\right). \tag{4.20}$$

By the mean-value theorem, and the fact that $\lambda \leq \lambda_0 < 1$, we have:

$$\left| f_X\left(\frac{t-\lambda a}{1-\lambda}\right) - f_X(t)\right| = \left|\frac{t-\lambda a}{1-\lambda} - t\right| \cdot |f_X'(\xi)|, \qquad \exists \xi \in I_{t,\lambda} \tag{4.21}$$

$$\leq \frac{\lambda}{1-\lambda_0}|t-a| \sup_{\xi \in I_{t,\lambda}} |f_X'(\xi)|. \tag{4.22}$$

Now, we study the derivative $f_X'(t)$ of the log-normal density, which computes easily as:

$$f_X'(t) = -\frac{\exp(-\frac{1}{8}\log(t)^2)(4+\log t)}{8\sqrt{2\pi}t^2}.$$

Note that, as $t \to 0$, $-(4+\log t) = \log(1/t)(1+o(1))$, and thus, as $t \to 0$,

$$f_X'(t) = \frac{1+o(1)}{8\sqrt{2\pi}} \exp\left(-\frac{1}{8}\log(1/t)^2 + \log(\log(1/t)) + 2\log(1/t)\right) = o(1).$$

A similar conclusion holds also as $t \to \infty$. Inspecting the graph of this function, we encounter the following features:

- $f_X'(t) \geq 0$ on $[0, e^{-4}]$, and $f_X(t) < 0$ on (e^{-4}, ∞).
- There exists a $T_1 \in (0, e^{-4})$, such that $f_X'(t)$ is increasing on $(0, e^{-4})$, and decreasing on (T_1, e^{-4}).
- There exists a $T_2 \in (e^{-4}, \infty)$ such that, $f_X'(t)$ is decreasing on (e^{-4}, T_2), and is increasing on (T_2, ∞).

In particular, $\sup_{t\in\mathbb{R}} |f_X'(t)| \leq \max\{f_X(T_1), -f_X(T_2)\} \triangleq \mathcal{M}_2$ (an absolute constant), recalling that $f_X(T_2) < 0$. Now, as long as $t \geq \max(a, T_2)$, and recalling Equation (4.20), since $|f_X'(t)|$ is decreasing on $I_{t,\lambda} = (t, \frac{t-\lambda}{1-\lambda})$ (since $f_X'(t)$ is increasing, and negative on this interval, we have the aforestated condition for $|f_X'(t)|$), we have that $\sup_{\xi\in I_{t,\lambda}} |f_X'(\xi)| = |f_X'(t)|$. We now upper bound the integral:

$$\int_{\lambda a}^{\infty} \left| f_X\left(\frac{t-\lambda a}{1-\lambda}\right) - f_X(t)\right| \, dt,$$

by splitting into two pieces: $t \in [\lambda a, \max(a, T_2)]$, and $t \in (\max(a, T_2), \infty)$. Recalling Equation (4.22), we have:

$$\int_{\lambda a}^{\infty} \left| f_X\left(\frac{t-\lambda a}{1-\lambda}\right) - f_X(t)\right| \, dt \leq \frac{\lambda}{1-\lambda_0} \int_{\lambda a}^{\infty} |t-a| \sup_{\xi \in I_{t,\lambda}} |f_X'(\xi)| \, dt.$$

Now, investigating right-hand-side, we have:

$$\begin{aligned}
&\frac{\lambda}{1-\lambda_0}\left(\int_{\lambda a}^{\max(a,T_2)} |t-a| \sup_{\xi\in I_{t,\lambda}} |f_X'(\xi)|\, dt + \int_{\max(a,T_2)}^{\infty} |t-a| \sup_{\xi\in I_{t,\lambda}} |f_X'(\xi)|\, dt\right) \\
&\leq \frac{\lambda}{1-\lambda_0}\left(\int_{\lambda a}^{\max(a,T_2)} |t-a|\mathcal{M}_2\, dt + \int_{\max(a,T_2)}^{\infty} |t-a|\cdot|f_X'(t)|\, dt\right) \\
&\leq \frac{\lambda}{1-\lambda_0}\left(\mathcal{C}_1(a) + \int_{\max(a,T_2)}^{\infty} |t-a|\cdot|f_X'(t)|\, dt\right),
\end{aligned}$$

using the fact that, $\int_{\lambda a}^{\max(a,T_2)} |t-a|\mathcal{M}_2\, dt$ is upper bounded by some absolute constant $\mathcal{C}_1(a)$, depending only on a (by simply considering integral from 0 to avoid λ dependency, and the fact that $\mathcal{M}_2$ is finite). For the second integral, observe that:

$$\begin{aligned}
\int_{\max(a,T_2)}^{\infty} |t-a|\cdot|f_X'(t)|; dt &= \int_{\max(a,T_2)}^{\infty} (t-a)\cdot\frac{\exp(-\frac{1}{8}\log(t)^2)(4+\log(t))}{8\sqrt{2\pi}t^2}\, dt \\
&\leq \frac{1}{8\sqrt{2\pi}}\int_{\max(a,T_2)}^{\infty}\frac{\exp(-\frac{1}{8}\log(t)^2)(4+\log t)}{t}\, dt = \mathcal{C}_2(a) < \infty,
\end{aligned}$$

using the fact that the integrand is equal to,

$$\exp\left(-\frac{1}{8}\log(t)^2 + \log(4+\log(t)) - \log(t)\right),$$

which is

$$\exp\left(-\frac{1}{8}\log(t)^2 + O(\log t)\right),$$

as $t\to\infty$. Combining these lines, we therefore have,

$$\int_{\lambda a}^{\infty}\left|f_X\left(\frac{t-\lambda a}{1-\lambda} - f_X(t)\right|\right)\, dt \leq \frac{\lambda}{1-\lambda_0}(\mathcal{C}_1(a)+\mathcal{C}_2(a)),$$

where $\mathcal{C}_1(a)$ and $\mathcal{C}_2(a)$ are two finite constants, depending only on a. Finally, recalling Equation (4.19), we then have:

$$\begin{aligned}
d_{TV}(X(\lambda),X(0)) &\leq \lambda\left(\frac{1}{2(1-\lambda_0)} + \frac{1}{2}\mathcal{M}_1\right) + \frac{\lambda}{2(1-\lambda_0)^2}(\mathcal{C}_1(a)+\mathcal{C}_2(a)) \\
&= \lambda\left(\frac{1}{2(1-\lambda_0)} + \frac{1}{2}\mathcal{M}_1 + \frac{1}{2(1-\lambda_0)^2}(\mathcal{C}_1(a)+\mathcal{C}_2(a))\right) \\
&\triangleq \lambda\mathcal{C}_{\lambda_0}
\end{aligned}$$

for every $\lambda\in[0,\lambda_0]$, as claimed earlier. Finally, taking $\mathcal{C}_{ij} = \max\{\mathcal{C}_{\lambda_0}, 1/\lambda_0\}$, we have $d_{TV}(X(\lambda),X(0)) \leq \mathcal{C}_{ij}\lambda$ for every $\lambda\in[0,1]$.

□

4.5.10 Proof of Lemma 4.3.3

Proof. Recall the following coupling interpretation of total variation distance:

$$d_{TV}(P,Q) = \inf\{\mathbb{P}(X \neq Y) : (X,Y) \text{ is such that } X \stackrel{d}{=} P, Y \stackrel{d}{=} Q\}.$$

Now, let $P_1, \ldots, P_\ell$ and $Q_1, \ldots, Q_\ell$ be measures defined on a sample space Ω. Suppose $X_1, \ldots, X_\ell$ are independent random variables with $X_i \stackrel{d}{=} P_i$ for $1 \leq i \leq \ell$; and $Y_1, \ldots, Y_\ell$ are independent random variables with $Y_i \stackrel{d}{=} Q_i$ for $1 \leq i \leq \ell$. Consider the vectors, $\mathbf{X} = (X_1, \ldots, X_\ell)$ and $\mathbf{Y} = (Y_1, \ldots, Y_\ell)$. Observe that, $\mathbf{X} \stackrel{d}{=} \otimes_{k=1}^{\ell} P_k$ and $\mathbf{Y} \stackrel{d}{=} \otimes_{k=1}^{\ell} Q_k$. Note that,

$$\{\mathbf{X} \neq \mathbf{Y}\} \subseteq \bigcup_{k=1}^{\ell} \{X_k \neq Y_k\}.$$

Now, using union bound, we have:

$$d_{TV}\left(\otimes_{k=1}^{\ell} P_k, \otimes_{k=1}^{\ell} Q_k\right) \leq \mathbb{P}(\mathbf{X} \neq \mathbf{Y}) \leq \sum_{k=1}^{\ell} \mathbb{P}(X_k \neq Y_k).$$

Now, recalling

$$d_{TV}(P_k, Q_k) = \inf_{(X_k, Y_k): X_k \stackrel{d}{=} P_k, Y_k \stackrel{d}{=} Q_k} \mathbb{P}(X_k \neq Y_k),$$

and taking infimums on the right hand side, we immediately obtain:

$$d_{TV}\left(\otimes_{k=1}^{\ell} P_k, \otimes_{k=1}^{\ell} Q_k\right) \leq \sum_{k=1}^{\ell} d_{TV}(P_k, Q_k),$$

as claimed. □

4.5.11 Proof of Lemma 4.3.4

Proof. Letting $Y_k = -X_k$ with $\mathbb{E}[Y_k] \leq -q$, we have:

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{\ell}\sum_{k=1}^{\ell} X_k > \epsilon\right) &= 1 - \mathbb{P}\left(\frac{1}{\ell}\sum_{k=1}^{\ell} Y_k \geq -\epsilon\right) \\
&= 1 - \mathbb{P}\left(\frac{1}{\ell}\sum_{k=1}^{\ell} (1 + Y_k) \geq 1 - \epsilon\right) \\
&\geq 1 - \frac{1-q}{1-\epsilon} = \frac{q-\epsilon}{1-\epsilon}
\end{aligned}$$

since for $Y = \frac{1}{\ell}\sum_{k=1}^{\ell}(1+Y_k) \geq 0$, it holds that $\mathbb{E}[Y] \leq 1-q$, and therefore by Markov inequality, we have $\mathbb{P}(Y \geq 1-\epsilon) \leq \frac{\mathbb{E}[Y]}{1-\epsilon} \leq \frac{1-q}{1-\epsilon}$. □

Chapter 5

Self-Regularity of Non-Negative Output Weights for Overparameterized Two-Layer Neural Networks

5.1 Introduction

Neural network (NN) architectures achieved a great deal of success in practice. An ever-growing list of their applications includes image recognition [165], image classification [192], speech recognition [219], natural language processing [79], game playing [257] and more. Despite this great empirical success, however, a rigorous understanding of these networks is still an ongoing quest.

A common paradigm in classical statistics is that *overparameterized* models, that is, models with more parameters than necessary, pick on the idiosyncrasies of the training data itself—dubbed as *overfitting*; and as a consequence, tend to predict *poorly* on the unseen data—called poor *generalization*. The aforementioned success of the NN architectures, however, stands in the face of this conventional wisdom; and a growing body of recent literature, starting from [294], has demonstrated exactly the opposite effect for a broad class of NN models: even though the number of parameters, such as the number of hidden units (neurons), of a NN significantly exceeds the sample size, and a perfect (zero) *in-training error* is achieved (commonly called as *data interpolation*); they still retain a good generalization ability. Some partial and certainly very incomplete list of references to this point are found in [99, 203, 156, 152, 46, 17]. Defying statistical intuition even further, it was established empirically in [46] that beyond a certain point, increasing the number of parameters increases out of sample accuracy.

Explaining this conundrum is arguably one of the most vexing current problems in the field of theoretical machine learning. Standard Vapnik-Chervonenkis (VC) theory do not help explaining the good generalization ability of overparameterized NN models, since the VC-dimension of these networks grows (at least) linearly in the number

of parameters [162, 42]. These findings fueled significant research efforts aiming at understanding the generalization ability of such networks. One such line of research is the algorithm-independent front; and is through the lens of controlling the norm of the matrices carrying weights [225, 41, 207, 153, 104], PAC-Bayes theory [224, 223], and compression-based bounds [20], among others. A major drawback of these approaches, however, is that they require certain norm constraints on the weights considered; therefore making their guarantees *a posteriori* in nature: whether or not the weights of the NN are bounded (hence a good generalization holds) can be determined only after the training process is complete. An alternative line of research (detailed below) focuses on the end results of the algorithms, and potentially yields *a priori* guarantees: for instance, relatively recently, Arora et al. gave in [18] *a priori* guarantees for the solution found by the *gradient descent* algorithm under random initialization.

A predominant explanation of the aforementioned phenomenon (that the overparameterization does not hurt the generalization ability of the NN architectures) which has emerged recently is based on the idea of *self-regularization*. Specifically, it is argued that even though there is an abundance of parameter choices perfectly fitting *(interpolating)* the data (and thus achieving zero in-training error); the algorithms used in training the models, such as the *gradient descent* and its many variants such as *stochastic gradient descent*, *mirror descent*, etc., tend to find solutions which are regularized according to some additional criteria, such as small norms, thus introducing algorithm dependent *inductive bias*. Namely, the algorithms implemented for minimizing training error "prefer" certain kinds of solutions. The use of these solutions for model building in particular is believed to result in low generalization errors. Thus a significant research effort (as was partially mentioned above) was devoted to the analysis of the end results of the implementation of such algorithms. This line of research include the analysis of the end results of the gradient descent [64, 115], stochastic gradient descent [160, 65, 203, 69], as well as the stochastic gradient Langevin dynamics [222].

In this work, we consider two-layer NN models (5.1)—also known as *shallow* architectures—consisting of an arbitrary number $\overline{m} \in \mathbb{N}$ of hidden units and sigmoid, rectified linear unit (ReLU), or binary step activations—activations that are arguably among the most popular practical choices—and investigate the following question: *to what extent a low training error itself places a restriction on the weights of the learned NN?* We take an algorithm-independent route; and establish the following "picture", under the assumption that the output weights $a = (a_i : 1 \leq i \leq \overline{m}) \in \mathbb{R}^{\overline{m}}$ of the "learned" NN are *non-negative*. When the number N of training samples is at least an explicit (low-degree) polynomial function in d, $N = d^{O(1)}$, the norm $\|a\|_1$ of the output weights $a \in \mathbb{R}^{\overline{m}}_{\geq 0}$ of **any** NN model achieving a small training error is well-controlled: $\|a\|_1 = O(1)$, with high probability over the training data set. In particular, for the ReLU and step networks, we obtain a near-linear sample complexity bound, $N = \Theta(d \log d)$ for such a result to hold. Note that a condition such as the non-negativity of a_i is necessary in a strict sense for such a bound on $\|a\|_1$. Indeed, notice that by growing the width $\overline{m}$ arbitrarily and appropriately choosing alternating signs for the new weights a_i; one can introduce cancellations and make $\|a\|_1$ to explode; while keeping the training error unchanged.

Our results are established using elementary tools, in particular through an ϵ-net argument (Definition 5.1.2). Notably, our results (a) are independent of the number $\overline{m}$ of the hidden units (which can potentially be quite large), (b) are oblivious to the way the training is done (that is, independent of the choice of the training algorithm); and (c) are valid under quite mild distributional assumptions on the input/label pairs $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$. In particular, the coordinates of X need not be independent.

Moreover, a bounded outer norm for such network models implies a well-controlled *fat-shattering dimension* (FSD) [40]—a measure of the complexity of the model class achieving a low training error. In Section 5.3, we leverage our outer norm bounds and the FSD to establish generalization guarantees for the networks that we investigate. This work presents extensions of certain results in our unpublished preprint [107].

Preliminaries

We commence this section with a list of notational convention that we follow throughout.

Notation The set of reals, non-negative reals, and positive integers are denoted respectively by $\mathbb{R}, \mathbb{R}_{\geq 0}$, and $\mathbb{N}$. For any set S, $|S|$ denotes its cardinality. For any $N \in \mathbb{N}$, $[N] \triangleq \{1, 2, \ldots, N\}$. For any $v \in \mathbb{R}^n$, its ℓ_p norm is denoted by $\|v\|_p$. $B_2(0, R) \triangleq \left\{x \in \mathbb{R}^d : \|x\|_2 \leq R\right\}$ denotes the Euclidean ball (of radius R), and $\mathbb{S}^{d-1} \triangleq \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ denotes the Euclidean unit sphere. For $u, v \in \mathbb{R}^n$, their Euclidean inner product is denoted by $u^T v$. For any $r \in \mathbb{R}$, $\exp(r)$ denotes e^r; and $\ln(r)$ denotes the logarithm of r base e. For any "event" E; $\mathbb{1}\{E\} = 1$ when E is true; and $\mathbb{1}\{E\} = 0$ when E is false. $\texttt{SGM}(x)$ denotes the sigmoid activation function, $1/(1 + \exp(-x))$; $\texttt{ReLU}(x)$ denotes the ReLU activation function, $\max\{x, 0\}$; and $\texttt{Step}(x)$ denotes the (binary) step activation, $\mathbb{1}\{x > 0\}$. $X \overset{d}{=} \mathcal{N}(0, \Sigma)$ if X is a zero-mean multivariate normal vector with covariance Σ. A random variable U is symmetric around zero if U and $-U$ have the same distribution, that is $U \overset{d}{=} -U$. For any random variable U, (if finite) its moment generating function (MGF) at $s \in \mathbb{R}$, $\mathbb{E}[\exp(sU)]$, is denoted by $M_U(s)$. Finally, $\Theta(\cdot), o(\cdot), O(\cdot)$ are the standard asymptotic order notations.

Setup A two-layer $\texttt{NN}$ $(a, W) \in \mathbb{R}^{\overline{m}} \times \mathbb{R}^{\overline{m} \times d}$ with $\overline{m}$ hidden units (neurons) computes, for each $X \in \mathbb{R}^d$,

$$\sum_{1 \leq j \leq \overline{m}} a_j \sigma\left(w_j^T X\right). \tag{5.1}$$

Here, $\sigma(\cdot)$ is the activation; $w_j \in \mathbb{R}^d$, the j^{th} row of W, carries the weights of neuron j; and $a = (a_j : 1 \leq j \leq \overline{m}) \in \mathbb{R}^{\overline{m}}$ carries the output weights. $\|a\|_1$ is referred to as the *outer norm*. The problem of "learning" such two-layer $\texttt{NN}$ models with nonlinear activations under no restrictions on the signs of weights has been previously studied, see e.g. [210, 261, 262, 260, 280] (and the references therein) for the case of quadratic activation function, $\sigma(x) = x^2$.

In this work, we investigate the self-regularization for the aforementioned architectures under the assumption that $a_j \geq 0$ for $j \in [\overline{m}]$. This non-negativity assumption

appears often in the theoretical study of this model: see [143, 93, 204] for generic $a \in \mathbb{R}^{\overline{m}}_{\geq 0}$; and [98, 250, 295, 150] for the case a_j are equal to the same positive number.

Our study of NN models under the non-negativity assumption is also partly motivated from an applied point of view, in that, non-negativity is inherent to many data sets appearing in practice, including audio data and data on muscular activity [259, 1], and it allows interpretability. Furthermore, non-negativity is also a commonly used assumption in the context of matrix factorization, termed as the *non-negative matrix factorization problem (NMF)*: given a matrix $M \in \mathbb{R}^{n \times m}$ with non-negative entries and an integer $r \geq 1$, the goal of the NMF is to find matrices $A \in \mathbb{R}^{n \times r}$ and $W \in \mathbb{R}^{r \times m}$ with non-negative entries such that the product AW is as "close" to M as possible; as quantified, e.g., by the Frobenius norm. This problem is a fundamental problem appearing in many practical applications, including information retrieval, document clustering, image segmentation, demography and chemometrics, see [19] and the references therein. Moreover, NMF is also related to the neural network models that we consider herein with a non-negative activation $\sigma(\cdot)$: observe that in the context of NN models we consider, given data (X_i, Y_i), $1 \leq i \leq N$, the goal of the learner is to find a $(a, W) \in \mathbb{R}^{\overline{m}} \times \mathbb{R}^{\overline{m} \times d}$ such that Y_i and $a^T \sigma(WX_i)$ are as close as possible, as quantified by the ℓ_2 norm (here, σ acts coordinate-wise to the vector WX). In addition to its key role in the NMF problem; the non-negativity was also argued as a natural assumption for representing objects in the seminal papers by Lee and Seung [196, 195]; and also has roots in biology, in particular in the context of neuronal firing rates, see [171], and the references therein.

In the sequel, $d \in \mathbb{N}$ is reserved for the input dimension; and $\overline{m} \in \mathbb{N}$ is reserved for the number of neurons. We consider herein two-layer NN models with sigmoid, $\texttt{SGM}(x)$; rectified linear unit, $\texttt{ReLU}(x)$; and binary step, $\texttt{Step}(x)$, activation functions. We refer to these as sigmoid, ReLU; and step networks, respectively. The sigmoid and the ReLU are arguably among the most popular practical choices. The step function, on the other hand, is one of the initial activations considered in the NN literature, and is inspired from a biological point of view: it resembles the firing pattern of a neuron, an initial motivation for studying NN architectures.

Given the data $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $1 \leq i \leq N$, consider the problem of finding a two-layer NN $(a, W) \in \mathbb{R}^{\overline{m}} \times \mathbb{R}^{\overline{m} \times d}$ which "fits" the data as accurately as possible. This is achieved by solving the so-called *empirical risk minimization* problem, where the accuracy is quantified by the *training error*

$$\widehat{\mathcal{L}}(a, W) \triangleq \frac{1}{N} \sum_{1 \leq i \leq N} \left(Y_i - \sum_{1 \leq j \leq \overline{m}} a_j \sigma\left(w_j^T X_i\right) \right)^2. \tag{5.2}$$

One then runs a training algorithm, e.g., the gradient descent algorithm or one of its variants (such as stochastic gradient descent or mirror descent), to find an (a, W) with a small $\widehat{\mathcal{L}}(a, W)$.

Distributional assumption We study the case where the input/label pairs (X_i, Y_i), $1 \leq i \leq N$, are i.i.d. samples of a distribution on $\mathbb{R}^d \times \mathbb{R}$ (which is potentially unknown to the learner). For our outer norm bounds, we assume that their distribution satisfies the following.

- We assume the input $X \in \mathbb{R}^d$ satisfies $\mathbb{P}\left(\|X\|_2^2 \leq Cd\right) \geq 1 - \exp\left(-\Theta(d)\right)$ for some constant $C > 0$.
- We assume the label Y is such that $\mathbb{E}[|Y|] \triangleq M < \infty$.

Later in Section 5.3 when we study generalization guarantees, we consider a stronger assumption on labels: we assume the labels Y are bounded, that is, for some $M > 0$, $|Y| \leq M$ almost surely.

These assumptions are quite mild. For instance, $X \in \mathbb{R}^d$ need not have i.i.d. coordinates. Moreover, most real data sets indeed have bounded labels [99]; and this bounded label assumption is employed extensively in literature, see e.g. [144, 18, 96, 151, 206]. Our next assumption regards the number N of training samples.

Assumption 5.1.1. *Throughout, we assume that the sample size N satisfies $N \leq \exp(cd)$ for some $c > 0$.*

Assumption 5.1.1 is required for technical reasons: observe that since $\mathbb{P}\left(\|X_i\|_2^2 > Cd\right) \leq \exp(-\Theta(d))$; it holds, by a union bound, that

$$\mathbb{P}\Big(\|X_i\|_2^2 \leq Cd, 1 \leq i \leq N\Big) \geq 1 - N\exp(-\Theta(d)).$$

For this bound to be non-vacuous, N should at most be $\exp(cd)$ for a small enough $c > 0$. This assumption, again, is very benign due to obvious practical reasons. Moreover, in fact, it suffices to have $N \geq \text{poly}(d)$ for our results to hold.

Nets and Covering Numbers The crux of our proofs is the so-called *$\epsilon-$net argument* [113, 281, 282]. This (rather elementary) argument is also known as the *covering number argument*; and has been employed extensively in the literature; including compressed sensing, machine learning and probability theory.

Definition 5.1.2. *Let $\epsilon > 0$. Given a metric space (X, ρ), a subset $\mathcal{N}_\epsilon \subset X$ is called an* $\epsilon-$net *of X if, for every $x \in X$, there is a $y \in \mathcal{N}_\epsilon$ such that $\rho(x, y) \leq \epsilon$. The smallest cardinality of such an $\mathcal{N}_\epsilon$, if finite, is called the* covering number *of X, denoted by $\mathcal{N}(X, \epsilon)$.*

The next result, verbatim from [282, Corollary 4.2.13], is an upper bound on the covering number of the Euclidean ball.

Theorem 5.1.3. *For $R \geq 1$ and any $\epsilon > 0$, $\mathcal{N}\left(B_2(0, R), \epsilon\right) \leq (3R/\epsilon)^d$; and $\mathcal{N}\left(\mathbb{S}^{d-1}, \epsilon\right) \leq (3/\epsilon)^d$.*

Organization of Chapter. The rest of the chapter is organized as follows. Our main results on the self-regularity of output weights are presented in Section 5.2. In particular, see Sections 5.2.1, 5.2.2, and 5.2.3 for the cases of sigmoid, ReLU, and step networks, respectively. By leveraging our outer norm bounds and employing earlier results on the fat shattering dimension, we establish in Section 5.3 generalization guarantees. We outline several future directions in Section 5.4. Finally, we present our proofs in Section 5.5.

5.2 Outer Norm Bounds

In this section, we establish the self-regularity of the output weights for the aforementioned networks. That is, we establish that the outer norms of sigmoid, ReLU, and step networks with non-negative output weights achieving a small training error (5.2) on polynomially many data is $O(1)$.

5.2.1 Self-Regularity for the Sigmoid Networks

Our first focus in on the sigmoid networks. This object, for each $X \in \mathbb{R}^d$, computes the function (5.1) with $\sigma(x) = \texttt{SGM}(x) = (1+\exp(-x))^{-1}$. Our first main result establishes an outer norm bound for this architecture.

Theorem 5.2.1. *Let $\delta, M, R > 0$; and $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i \in [N]$ be i.i.d. data with $\mathbb{E}[|Y_i|] = M < \infty$; where N satisfies Assumption 5.1.1. For any $\overline{m} \in \mathbb{N}$, define*

$$\mathcal{S}(\overline{m}, \delta, R) = \left\{(a, W) \in \mathbb{R}^{\overline{m}}_{\geq 0} \times \mathbb{R}^{\overline{m} \times d} : \max_{1 \leq j \leq \overline{m}} \|w_j\|_2 \leq R, \widehat{\mathcal{L}}(a, W) \leq \delta^2\right\},$$

where $\widehat{\mathcal{L}}(\cdot)$ is defined in (5.2) with $\sigma(\cdot) = \texttt{SGM}(\cdot)$. Suppose, in addition, that the random variable $w^T X \in \mathbb{R}$ is symmetric around zero for every $w \in \mathbb{R}^d$. Then,

$$\begin{aligned} &\mathbb{P}\left(\sup_{(a,W) \in \mathcal{S}(\delta, R)} \|a\|_1 \leq 3(1+e)(\delta + 2M)\right) \\ &\geq 1 - \left(3R\sqrt{Cd}\right)^d \exp\left(-\Theta(N)\right) - N\exp\left(-\Theta(d)\right) - o_N(1), \end{aligned} \tag{5.3}$$

where $\mathcal{S}(\delta, R) \triangleq \bigcup_{\overline{m} \in \mathbb{N}} \mathcal{S}(\overline{m}, \delta, R)$.

Corollary 5.2.2. *Let $R = \exp\left(d^{O(1)}\right)$. Then, under the assumptions of Theorem 5.2.1; it holds, w.h.p., that $\sup_{(a,W) \in \mathcal{S}(\delta, R)} \|a\|_1 \leq 3(1+e)(\delta + 2M)$, if $N \geq d^{O(1)}$.*

The proof of Theorem 5.2.1 is provided in Section 5.5.1.

Above, $o_N(1)$ is a function which depends only on the distribution of Y and N; and tends to zero as $N \to \infty$. Several remarks are now in order. Theorem 5.2.1 states that *any* two-layer sigmoid NN which (a) consists of internal weights w_j bounded in norm by an exponentially large (in d) quantity and non-negative output weights; and (b) achieves a small training error on a *sufficiently large* data set, has a *well-controlled*

outer norm. It is worth noting that Theorem 5.2.1 is oblivious to how the training is done: this result not only applies to the weights obtained, say, via the *gradient descent* algorithm; but applies to *any* weights (subject to the aforementioned assumptions) achieving a small training loss.

Moreover, the upper bound established in Theorem 5.2.1 is also oblivious to the number $\overline{m}$ of the neurons of the `NN` used for fitting, provided that a small training error is attained. (Here, it is worth noting that our results are contingent upon a small training error of δ^2; and in practice, one needs a large number $\overline{m}$ of neurons to ensure a small training error. Namely, $\overline{m}$, in some sense, controls $\widehat{\mathcal{L}}(a, W)$. We do not focus on the question of whether such a small training error is attainable; but rather establish self-regularity for any (a, W) with $\widehat{\mathcal{L}}(a, W) \leq \delta^2$.) In particular, adopting a teacher/student setting as in [152] where the input/label pairs (X_i, Y_i) are generated by a teacher `NN`; the output norm of any student `NN`—which may potentially be significantly overparameterized with respect to the teacher `NN`—is still well-controlled, provided the assumptions of Theorem 5.2.1 are satisfied. The extra requirement that $w^T X$ is symmetric is quite mild: it holds for many data distributions, e.g., for $X \overset{d}{=} \mathcal{N}(0, \Sigma)$ where Σ is an arbitrary positive semidefinite matrix.

The $o_N(1)$ term is due to a certain high probability event $\mathcal{E}_0$, see (5.10) in the proof. The probability of this event is controlled through the weak law of large numbers; and the $o_N(1)$ term can be improved explicitly (a) to $O(1/N)$ if $\mathbb{E}[Y^2] < \infty$; and (b) to $\exp(-\Theta(N))$ if Y_i satisfy the large deviations bounds (which holds, for instance, when the moment generating function of Y_i exists in a neighbourhood around zero). Moreover, if Y is (almost surely) bounded (which holds for real data sets, as noted earlier), then it can be dropped altogether.

Furthermore, Corollary 5.2.2—which follows immediately from Theorem 5.2.1—asserts that even under the mild assumption $R = \exp\left(d^{O(1)}\right)$ (i.e., the weights w_j are *unbounded* from a practical perspective), $\sum_j a_j$ is still $O(1)$, provided that that the number N of data is polynomial in d.

Moreover, an inspection of the proof of Theorem 5.2.1 reveals the following. The constant $3(1 + e)$ can be improved to any constant greater than four with slightly more work. Moreover, the thesis of Theorem 5.2.1 still remains valid (with appropriately modified constants) for any non-negative activation which is continuous at the origin and whose value at the origin is positive. This includes the softplus activation $\ln(1 + e^x)$ [148], the Gaussian activation, $\exp(-x^2)$; among others.

5.2.2 Self-Regularity for the ReLU Networks

Our next focus is on the ReLU networks. This object, for each input $X \in \mathbb{R}^d$, computes the function (5.1) with $\sigma(x) = \texttt{ReLU}(x) = \max\{x, 0\} = \frac{1}{2}(x + |x|)$.

We first observe that the `ReLU` function is positive homogeneous: for any $c \geq 0$ and $x \in \mathbb{R}$, $\texttt{ReLU}(cx) = c \cdot \texttt{ReLU}(x)$. For this reason, we assume, without loss of generality, that $\|w_j\|_2 = 1$ for $1 \leq j \leq \overline{m}$. Indeed, if $w_j \neq 0$, one can simply "push" its norm outside; whereas if $w_j = 0$, then one can replace it with any unit norm vector and set $a_j = 0$ instead.

It is worth noting that since the ReLU case requires no explicit assumptions on $\|w_j\|_2$, an outer bound for this case is a somewhat stronger conclusion than an outer bound for the case of sigmoid activation.

Equipped with this, we now present our next result.

Theorem 5.2.3. *Let $\delta, M > 0$; and $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i \in [N]$ be i.i.d. data with $\mathbb{E}[|Y_i|] = M < \infty$; where N satisfies Assumption 5.1.1. For any $\overline{m} \in \mathbb{N}$, define*

$$\mathcal{G}(\overline{m}, \delta) = \left\{(a, W) \in \mathbb{R}^{\overline{m}}_{\geq 0} \times \mathbb{R}^{\overline{m} \times d} : \|w_j\|_2 = 1, 1 \leq j \leq \overline{m}; \widehat{\mathcal{L}}(a, W) \leq \delta^2\right\},$$

where $\widehat{\mathcal{L}}(\cdot)$ is defined in (5.2) with $\sigma(\cdot) = \texttt{ReLU}(\cdot)$. Suppose, in addition, that for $Y_w \triangleq w^T X$, (a) there exists a $\boldsymbol{\mu}^* > 0$ such that $\mathbb{E}[\texttt{ReLU}(Y_w)] \geq \boldsymbol{\mu}^*$ for any $w \in \mathbb{R}^d$ with $\|w\|_2 = 1$; and (b) for some $s > 0$, $M_1(s)$ and $M_2(s)$ are independent of d and are finite; where $M_1(s) \triangleq \sup_{w:\|w\|_2=1} M_{Y_w}(s)$ and $M_2(s) \triangleq \sup_{w:\|w\|_2=1} M_{Y_w}(-s)$. Then,

$$\begin{aligned} &\mathbb{P}\left(\sup_{(a,W)\in\mathcal{G}(\delta)} \|a\|_1 \leq 4(\delta + 2M)(\boldsymbol{\mu}^*)^{-1}\right) \\ &\geq 1 - \left(\frac{12\sqrt{Cd}}{\boldsymbol{\mu}^*}\right)^d \exp\left(-\Theta(N)\right) - N\exp\left(-\Theta(d)\right) - o_N(1), \end{aligned} \tag{5.4}$$

where $\mathcal{G}(\delta) \triangleq \bigcup_{\overline{m}\in\mathbb{N}} \mathcal{G}(\overline{m}, \delta)$.

The proof of Theorem 5.2.3 is provided in Section 5.5.2.

In particular, it suffices to have a near-linear number of samples, $N = \Theta(d \log d)$, to obtain a good uniform control over $\|a\|_1$. As mentioned above, we managed to bypass the dependence on the term R appearing in Theorem 5.2.1 by leveraging the fact that `ReLU` is a positive homogenenous function.

Analogous to Theorem 5.2.1, the bound established in Theorem 5.2.3 is also oblivious to (a) how the training is done, and (b) the number $\overline{m}$ of neurons. In particular, even potentially overparameterized networks have a well-controlled outer norm; provided that they achieve a small training error on a sufficient number N of data. The additional distributional requirements are still mild. For instance, when $X \overset{d}{=} \mathcal{N}(0, I_d)$, $w^T X \overset{d}{=} \mathcal{N}(0, 1)$ for any w with $\|w\|_2 = 1$; and $\boldsymbol{\mu}^*$ can be taken to be $1/\sqrt{2\pi}$. The requirement (b) ensures the existence of the moment generating function in a neighborhood around zero, hence the large deviations bounds are applicable. The same remarks on $o_N(1)$ term following Theorem 5.2.1 also apply here: it can be improved to $O(1/N)$ or $\exp(-\Theta(N))$ under slightly stronger assumptions on Y_i.

5.2.3 Self-Regularity for the Step Networks

Our final focus is on the step networks. This object, for each $X \in \mathbb{R}^d$, computes (5.1) with $\sigma(x) = \texttt{Step}(x) = \mathbb{1}\{x > 0\}$ (which is the Heaviside step function).

Like the ReLU case, $\texttt{Step}(x)$ is also homogeneous: for every $c > 0$, $\texttt{Step}(cx) = \texttt{Step}(x)$. For this reason, we assume, without loss of generality, $\|w_j\|_2 = 1$, $1 \leq j \leq \overline{m}$.

Theorem 5.2.4. *Let $\delta, M > 0$; and $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i \in [N]$ be i.i.d. data with $\mathbb{E}[|Y_i|] = M < \infty$; where N satisfies Assumption 5.1.1. For any $\overline{m} \in \mathbb{N}$, define*

$$\mathcal{H}(\overline{m}, \delta) = \left\{(a, W) \in \mathbb{R}^{\overline{m}}_{\geq 0} \times \mathbb{R}^{\overline{m} \times d} : \|w_j\|_2 = 1, 1 \leq j \leq \overline{m}; \widehat{\mathcal{L}}(a, W) \leq \delta^2\right\},$$

with $\widehat{\mathcal{L}}(\cdot)$ as in (5.2) with $\sigma(\cdot) = \texttt{Step}(\cdot)$. Moreover, assume that for some $\eta > 0$, $\inf_{w:\|w\|_2=1} \mathbb{P}\left(w^T X \geq \eta\right) \geq \eta$. Then,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{(a,W)\in\mathcal{H}(\delta)} \|a\|_1 \leq 2(\delta + 2M)\eta^{-1}\right) \\
&\geq 1 - \left(\frac{6\sqrt{Cd}}{\eta}\right)^d \exp\left(-\Theta(N)\right) - N\exp\left(-\Theta(d)\right) - o_N(1)
\end{aligned}$$

where $\mathcal{H}(\delta) \triangleq \bigcup_{\overline{m}\in\mathbb{N}} \mathcal{H}(\overline{m}, \delta)$.

The proof of Theorem 5.2.4 is quite similar to that of Theorems 5.2.1 and 5.2.3; and is provided in Section 5.5.3 for completeness. Furthermore; main remarks following Theorems 5.2.1 and 5.2.3—in particular, independence from $\overline{m}$ as well as the training algorithm—apply here, as well.

The extra condition on the distribution ensures that the collection $\left\{\mathbb{P}(w^T X \geq \eta) : \|w\|_2 = 1\right\}$ is uniformly bounded away from zero. This is again quite mild, as demonstrated by the following example. Suppose $Y_w \triangleq w^T X$ is centered and equidistributed for w with $\|w\|_2 = 1$. (Observe that this is indeed the case, e.g. when $X \stackrel{d}{=} \mathcal{N}(0, I_d)$.) Then as long as $\mathrm{Var}(Y_w) > 0$ the extra requirement per Theorem 5.2.4 is satisfied. Indeed, for this case $\mathbb{P}(Y_w > 0) > 0$. Hence, using the continuity of probabilities

$$\mathbb{P}(Y_w > 0) = \mathbb{P}(w^T X > 0) = \lim_{t\to\infty} \mathbb{P}\left(w^T X > t^{-1}\right) > 0,$$

one ensures the existence of such an η. In the case where $X \stackrel{d}{=} \mathcal{N}(0, I_d)$, one can concretely take $\eta = 0.3$.

5.3 Generalization Guarantees via Outer Norm Bounds

5.3.1 The Learning Setting

In this section, we leverage the outer norm bounds we established in Theorems 5.2.1-5.2.4 to provide generalization guarantees for the neural network architectures having non-negative output weights that we investigated.

Our approach is through a quantity called the *fat-shattering dimension* (FSD) of such networks introduced by Kearns and Schapire [187]. This quantity is essentially a scale-sensitive measure of the complexity of the "class" (appropriately defined) that the network architecture being considered belongs to. We introduce the FSD formally in Definition 5.5.1 found in Section 5.5.4. For more information on the FSD, we refer

the interested reader to the original paper by Kearns and Schapire [187]; as well as earlier papers by Bartlett, Long, and Williamson [43], and Bartlett [40].

In what follows, we prove our promised generalization guarantee (Theorem 5.3.1 below) by combining the prior results on the FSD of such networks with our outer norm bounds. Bartlett provides in [40] upper bounds on the FSD of certain function classes H. He then leverages these bounds to give good generalization guarantees. One of the classes he studies is precisely the class of two-layer NN with a **bounded outer norm** (as we do). In particular, he establishes in [40, Corollary 24] (which is restated as Theorem 5.5.2 below) that the class of two-layer networks with **bounded outer norm** has a well-controlled FSD: informally, it has "low complexity". He then leverages the FSD bounds to devise good generalization guarantees for the architectures that he investigates. It is worth noting, however, that he establishes this link in the context of *classification* setting, $Y \in \{\pm 1\}$. Since we assume $a_j \geq 0$, and the activations we study are non-negative, this does not apply to our case: the outputs of the networks we study are always non-negative. Nevertheless, we by-pass this by combining our outer norm bounds (Theorems 5.2.1-5.2.4), Theorem 5.5.2, as well as building upon several other prior results tailored for the *regression* setting.

We next recall the learning setting for convenience. Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \mathbb{R}$ for the input/label pairs (X, Y); and let $(X_i, Y_i) \sim \mathcal{D}$, $1 \leq i \leq N$, be the i.i.d. training data. The goal of the learner is to find a NN $(a, W) \in \mathbb{R}^{\overline{m}} \times \mathbb{R}^{\overline{m} \times d}$ with $\overline{m}$ hidden units (neurons) and activation $\sigma(\cdot)$ which "explains" the data (X_i, Y_i), $1 \leq i \leq N$, as accurately as possible, often by solving the *empirical risk minimization* problem, $\min_{a,W} \widehat{\mathcal{L}}(a, W)$ (5.2). The "learned" network is then used for predicting the unseen data. The generalization ability of the "learned" network $(a, W) \in \mathbb{R}^{\overline{m}} \times \mathbb{R}^{\overline{m} \times d}$ is quantified by the so-called *generalization error* (also known as the *population risk*)

$$\mathcal{L}(a, W) \triangleq \mathbb{E}_{(X,Y)\sim\mathcal{D}} \left[\left(Y - \sum_{1 \leq j \leq \overline{m}} a_j \sigma\left(w_j^T X\right) \right)^2 \right]. \tag{5.5}$$

Here, the expectation is taken w.r.t. to a fresh sample $(X, Y) \sim \mathcal{D}$, which is independent of the training data. The "gap" $\left| \widehat{\mathcal{L}}(a, W) - \mathcal{L}(a, W) \right|$ between the training error and the generalization error is called the *generalization gap*.

In what follows, we focus our attention on the generalization ability (5.5) of the learned networks (a, W) that achieved a small training error, $\widehat{\mathcal{L}}(a, W) \leq \delta^2$ (5.2), on a polynomial (in d) number of data. The details of the training process (such as the algorithm used for training) are immaterial to us; and our results apply to **any** NN (a, W) provided it achieved a small training error, $\widehat{\mathcal{L}}(a, W) \leq \delta^2$.

In this section, we also assume that the labels Y are bounded: $\mathcal{D}$ is such that for some $M > 0$, $|Y| \leq M$ almost surely. This is necessary, as the prior results we employ from Haussler [164] and Bartlett, Long, and Williamson [43] (in particular, see Theorem 5.5.5 below) apply only to the case where the labels are bounded. For this reason, the $o_N(1)$ terms present in Theorems 5.2.1-5.2.3 disappear, see the remarks following each theorem.

5.3.2 The Generalization Guarantees

Equipped with our outer norm bounds (Theorems 5.2.1-5.2.4), and Theorem 5.5.2, we now provide the promised generalization guarantees for the aforementioned networks whose output weights a_i are non-negative. To that end, let $\alpha, M, \mathcal{M}, A > 0$ be certain parameters (elaborated below); and set

$$\xi\left(\alpha, M, \mathcal{M}, A\right) \triangleq \frac{2}{\ln 2} \cdot \frac{c \cdot 128^2 \cdot \mathcal{M}^6 A^6 \cdot \max\{\mathcal{M}A, 2M\}^2}{\alpha^2} \cdot \ln\left(\frac{128\mathcal{M}^3 A^3 \max\{\mathcal{M}A, 2M\}}{\alpha}\right), \tag{5.6}$$

where $c > 0$ is the absolute constant appearing in Theorem 5.5.2. Our result is as follows.

Theorem 5.3.1. *Let $\alpha, \delta, M, R > 0$, and $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $1 \leq i \leq N$, be i.i.d. samples drawn from an arbitrary distribution $\mathcal{D}$ on $\mathbb{R}^d \times \mathbb{R}$ with $|Y| \leq M$ almost surely; where N satisfies Assumption 5.1.1. For the ξ term defined in (5.6), set*

$$\zeta(\alpha, M, A, N) \triangleq \exp\Big(\xi(\alpha, M, 2, A) \cdot d \cdot \ln^2\left(\frac{2304 \cdot N \cdot A^2 \cdot \max\{2A, M\}}{\alpha}\right) - \frac{\alpha^2 \cdot N}{64 \cdot \max\{2A, M\}^2}\Big). \tag{5.7}$$

(a) **(Sigmoid Networks)** *Under the assumptions of Theorem 5.2.1, with probability at least*

$$1 - \zeta(\alpha, M, 3(1+e)(\delta+2M), N) - \Big(3R\sqrt{Cd}\Big)^d \exp\left(-\Theta(N)\right) - N\exp\left(-\Theta(d)\right)$$

over $(X_i, Y_i) \sim \mathcal{D}$, $1 \leq i \leq N$, it holds that

$$\sup_{(a,W)\in\mathcal{S}(\delta,R)} \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\left(Y - \sum_{j=1}^{\overline{m}} a_j \mathtt{SGM}\left(w_j^T X\right)\right)^2\right]$$

is at most $\alpha + \delta^2$, provided

$$N \geq c \cdot 2^{21} \cdot \frac{A^6 \cdot \max\{A, M\}^2}{\alpha^2} \cdot d$$

and $\alpha \leq 2^{11} \cdot A^3 \cdot \max\{A, M\}$ with $A = 3(1+e)(\delta + 2M)$. Here, $\mathcal{S}(\delta, R)$ is the set introduced in Theorem 5.2.1.

(b) **(ReLU Networks)** *Under the assumptions of Theorem 5.2.3 and assuming additionally $\left(\boldsymbol{\mu}^*\right)^{-1} = \exp\Big(o(d)\Big)$, with probability at least*

$$1 - \zeta\left(\alpha, M, \frac{4\sqrt{Cd}(\delta+2M)}{\boldsymbol{\mu}^*}, N\right) - \left(\frac{12\sqrt{Cd}}{\boldsymbol{\mu}^*}\right)^d \exp\left(-\Theta(N)\right) - N\exp\left(-\Theta(d)\right)$$

over $(X_i, Y_i) \sim \mathcal{D}$, $1 \leq i \leq N$, *it holds that*

$$\sup_{(a,W)\in\mathcal{G}(\delta)} \mathbb{E}_{(X,Y)\sim\mathcal{D}} \left[\left(Y - \sum_{j=1}^{\overline{m}} a_j \texttt{ReLU}\left(w_j^T X\right) \right)^2 \right]$$

is at most $\alpha + \delta^2 + e^{-\Theta(d)}$ *provided*

$$N \geq c \cdot 2^{21} \cdot \frac{A^6 \cdot \max\{A, M\}^2}{\alpha^2} \cdot d$$

and $\alpha \leq 2^{11} \cdot A^3 \cdot \max\{A, M\}$ *with* $A = \frac{4\sqrt{Cd}(\delta+2M)}{\boldsymbol{\mu}^*}$. *Here,* $\mathcal{G}(\delta)$ *is the set introduced in Theorem 5.2.3.*

(c) **(Step Networks)** *Under the assumptions of Theorem 5.2.4, with probability at least*

$$1 - \zeta\left(\alpha, M, \frac{2(\delta+2M)}{\eta}, N\right) - \left(\frac{6\sqrt{Cd}}{\eta}\right)^d \exp\left(-\Theta(N)\right) - N\exp\left(-\Theta(d)\right)$$

over $(X_i, Y_i) \sim \mathcal{D}$, $1 \leq i \leq N$, *it holds that*

$$\sup_{(a,W)\in\mathcal{H}(\delta)} \mathbb{E}_{(X,Y)\sim\mathcal{D}} \left[\left(Y - \sum_{j=1}^{\overline{m}} a_j \texttt{Step}\left(w_j^T X\right) \right)^2 \right],$$

is at most $\alpha + \delta^2$, *provided*

$$N \geq c \cdot 2^{21} \cdot \frac{A^6 \cdot \max\{A, M\}^2}{\alpha^2} \cdot d$$

and $\alpha \leq 2^{11} \cdot A^3 \cdot \max\{A, M\}$ *with* $A = \frac{2(\delta+2M)}{\eta}$. *Here,* $\mathcal{H}(\delta)$ *is the set introduced in Theorem 5.2.4.*

Theorem 5.3.1 is established by combining various individual results established in separate works [164, 43, 12, 40] together with our outer norm bounds. See Section 5.5.4 for its proof.

We next comment on the performance parameters appearing in Theorem 5.3.1. The parameter α controls the so-called *generalization gap*: the gap between the training error and the generalization error. The parameter δ controls the training error: we study those (a, W) with $\widehat{\mathcal{L}}(a, W) \leq \delta^2$. The parameter M is an (almost sure) upper bound on the labels; whereas R is an (quite mild) upper bound on internal weights required for the technical reasons, only for the case of sigmoid networks, see Theorem 5.2.1, Corollary 5.2.2; and the remarks following them.

The term $\zeta(\alpha, M, A, N)$ is a probability term appearing in the uniform convergence result (Proposition 5.5.3) that we employ. This proposition provides a control for the

generalization gap uniformly over all two-layer neural networks with *bounded outer norm* like we investigate herein.

In what follows, we think of the parameter α as a sufficiently small, though constant, quantity. It is worth noting that in the high-dimensional regime $d \to \infty$ (which is a legitimate assumption for many existing guarantees in the field of machine learning) and for $N \geq d^{O(1)}$; $\xi(\alpha, M, A) = O(1)$ (with respect to d), provided that $A = O(1)$ like we establish earlier. Namely, this object is simply a constant in d. Furthermore, while we made no attempts in simplifying it, it can potentially be improved. In the sigmoid and step cases, the value of A that we consider is indeed $O(1)$. For the ReLU case, however, the situation is more involved; and a certain scaling which makes $A = \text{poly}(d)$ is necessary, as we elaborate soon. Soon in Section 5.3.3, we investigate the probability term $\zeta(\alpha, M, A, N)$ appearing in (5.7). We show that provided N is sufficiently large (while remaining polynomial in d), the ζ term behaves like $\exp\left(-d^{O(1)}\right)$, thus it is indeed $o_d(1)$. Moreover, our analysis will also reveal that the dependence of N on d is quite mild; and is in fact near-linear in some important cases of interest. To that end, we now inspect the lower bound on N appearing in Theorem 5.3.1; and claim that it scales polynomially in d. Note that in the high-dimensional regime where $d \to \infty$ while all other parameters are kept as constants in d, this immediately follows. (It is worth noting that that this assumption in a sense is also necessary so that the high probability guarantees hold.) In the case d is of constant order itself, all lower bounds can, in principle, be perceived as potentially high degree polynomials in d. As an example back-of-the-envelope calculation, if $d = 100$ and $\alpha = 0.01$, then $2^{21}/\alpha^2$ is roughly $2 \cdot 10^{10}$, which is of order d^5. Thus, it suffices to have, e.g. $N = \Omega(d^6)$ in this case. See below for a much more elaborate sample complexity analysis in the high-dimensional regime, $d \to \infty$.

Theorem 5.5.2 as well as the uniform generalization gap guarantee, Proposition 5.5.3, apply to activations with a bounded output; whereas the output of ReLU is potentially unbounded. In our proof, we bypass this by considering an auxiliary activation $\texttt{S-ReLU}(\cdot)$, which is a "saturated" version of the ReLU. Specifically, we let $\texttt{S-ReLU}(x) = 0$ for $x \leq 0$, $\texttt{S-ReLU}(x) = x$ for $0 < x \leq 1$; and $\texttt{S-ReLU}(x) = 1$ for $x \geq 1$. We then rescale w_j to have $\|w_j\|_2 = 1/\sqrt{Cd}$ and multiply A by $\sqrt{Cd}$ (we therefore consider $A = 4\sqrt{Cd}(\delta + 2M)/\boldsymbol{\mu}^*$, $\sqrt{Cd}$ times the bound appearing in Theorem 5.2.3). Note that this step is indeed valid due to the homogeneity of the ReLU activation, see also Section 5.2.2. Since $\|X\|_2 \leq \sqrt{Cd}$ with probability at least $1 - \exp(-\Theta(d))$ and since $|w_j^T X| \leq 1$ for $\|w_j\|_2 = 1/\sqrt{Cd}$ and $\|X\|_2 \leq \sqrt{Cd}$ by Cauchy-Schwarz inequality; the output of this activation will, w.h.p., coincide with that of the ReLU activation. We then control the difference between the generalization errors for a pair of two-layer neural networks having the same architecture, the same number $\overline{m} \in \mathbb{N}$ of hidden units, the same weights (a, W); but different activations (one with $\texttt{ReLU}(\cdot)$ and the other with $\texttt{S-ReLU}(\cdot)$). This done by a conditioning argument. See the proof for further details.

Similar to what we have noted previously for our outer norm bounds, Theorem 5.3.1 is also oblivious to (a) how the training is done and (b) the number $\overline{m}$ of hidden units as long as $a_i \geq 0$, and $\widehat{\mathcal{L}}(a, W) \leq \delta^2$ for the learned network. More-

over, similar to prior cases, the extra conditional expectation requirement (5.22) is quite mild.

Our next focus is on the sample complexity required by Theorem 5.3.1. We show that they are indeed polynomial in d. Furthermore for some very important cases, they are even near-linear.

5.3.3 Sample Complexity Analysis

While the required sample complexity N can simply be inferred from Theorem 5.3.1, we spell out the implied scaling analysis below for convenience. In what follows, all asymptotic notations are w.r.t. the natural parameter d (namely the dimension) of the problem in the regime $d \to \infty$; and our goal is to ensure that the corresponding probability term is $1 - o_d(1)$ for an appropriate function $o_d(1)$. (It is worth noting though that our bounds will be in fact much stronger, e.g. $1 - \exp(-d^{O(1)})$.)

To that end, recall the term (5.6) with $\mathcal{M} = 2$ appearing in Theorem 5.3.1:

$$\xi(\alpha, M, 2, A) = \frac{2^{23} \cdot c \cdot A^6 \cdot \max\{A, M\}^2}{\ln 2 \cdot \alpha^2} \cdot \ln\left(\frac{2^{11} \cdot A^3 \cdot \max\{A, M\}}{\alpha}\right). \tag{5.8}$$

Sigmoid and Step Networks

First, the outer norm bounds we establish indicate $A = O(1)$. Hence, the "A parameter" considered in parts (a) and (c) of Theorem 5.3.1 are $O(1)$. Moreover, $M = O(1)$ (since it is not sound for the real-valued label Y to grow with dimension d). Treating α as a constant in d, we then obtain $\xi(\alpha, M, A) = O(1)$ for the term appearing in (5.8). Hence, in order to ensure that the probability term ζ appearing in (5.7) is $o_d(1)$, a necessary and sufficient condition is $N = \Omega\left(d \ln^2 N\right)$. We claim that it suffices to have

$$N = \Omega\left(d \ln^2 d\right). \tag{5.9}$$

Indeed, if N satisfies (5.9), then provided N remains polynomial in d, $N = \mathrm{poly}(d)$, it holds that

$$\ln^2 N = O\left(\ln^2 d\right) \implies d \ln^2 N = O\left(d \ln^2 d\right) = O(N).$$

We now investigate the sample complexity required by the corresponding outer norm bounds for the case of sigmoid and step networks.

Sigmoid networks Note, in this case, that the dominant contribution to the probability term appearing in Theorem 5.2.1/Theorem 5.3.1(a) (other than ξ term) is $(3R\sqrt{Cd})^d \exp(-\Theta(N))$. Suppose first that $R = d^K$ where $K = O(1)$ (namely R remains polynomial in d). Then $(3R\sqrt{Cd})^d \exp(-\Theta(N)) = \exp\Big(-\Theta(N) + d\left(K + \frac{1}{2}\right)\ln d + d\ln(3\sqrt{C})\Big) = \exp\Big(-\Theta(N) + \Theta(d\ln d) + o(d\ln d)\Big)$. Provided $N = \Omega(d \ln d)$, this bound is indeed $o_d(1)$. Taking the maximum between this and (5.9), we obtain that it suffices to have $N = \Omega(d \ln^2 d)$, which is near-linear.

Suppose next that $R = \exp(d^K)$, like in Corollary 5.2.2. Then provided $K > 0$, $(3R\sqrt{Cd})^d \exp(-\Theta(N)) = \exp\Big(-\Theta(N)+d^{K+1}+\frac{1}{2}d\ln d+d\ln(3\sqrt{C})\Big) = \exp\Big(-\Theta(N)+ d^{K+1}+o\left(d^{K+1}\right)\Big)$. Hence, provided $N = \Omega(d^{K+1})$, this bound is indeed $o_d(1)$. Taking the maximum between this and (5.9), we obtain that it suffices to have $N = \Omega(d^{K+1})$, which is polynomial in d.

Step networks Treating the distributional parameter η appearing in Theorem 5.2.4/Theorem 5.2.3 as a constant in d, we have $\exp\left(-\Theta(N)\right)\left(\frac{6\sqrt{Cd}}{\eta}\right)^d = \exp\left(-\Theta(N) + \frac{1}{2}d\ln d + d\ln\left(\frac{6\sqrt{C}}{\eta}\right)\right) = \exp\left(-\Theta(N) + \Theta(d\ln d) + o(d\ln d)\right)$. Thus, provided $N = \Omega(d\ln d)$, this bound is indeed $o_d(1)$. Taking the maximum between this and (5.9), we obtain that it suffices to have $N = \Omega(d\ln^2 d)$, which, again, is near-linear.

ReLU Networks

The situtation is more involved for the case of ReLU networks. We first study the ξ term (5.8). Treating $M, \alpha, C, \delta, \boldsymbol{\mu^*} = O(1)$ (in d), $\xi\left(\alpha, M, 2, \frac{4\sqrt{C}(\delta+2M)}{\boldsymbol{\mu^*}}\sqrt{d}\right) = \Theta\left(d^4\ln d\right)$. Hence, $\zeta\left(\alpha, M, \frac{4\sqrt{Cd}(\delta+2M)}{\boldsymbol{\mu^*}}, N\right) = \exp\Big(\Theta\Big(d^4\cdot\ln d\cdot d\cdot\ln^2(Nd)\Big)-\Theta\left(\frac{N}{d}\right)\Big) = \exp\Big(\Theta\Big(d^5\cdot\ln d\cdot\ln^2(Nd)\Big)-\Theta\left(\frac{N}{d}\right)\Big) = \exp\Big(\Theta\Big(d^5\cdot\ln^3 d\Big)-\Theta\left(\frac{N}{d}\right)\Big)$, where we used the fact $\ln(Nd) = \Theta(\ln d)$ if $N = \mathrm{poly}(d)$. Thus, provided $N = \Omega\left(d^6\ln^3 d\right)$, this bound is indeed $o_d(1)$. Inspecting next the term $(12\sqrt{Cd}/\boldsymbol{\mu^*})^d\exp(-\Theta(N))$ appearing in the probability bound, we observe as long as $N = \Omega(d\ln d)$, this term is also $o_d(1)$. Taking the maximum of these two, it suffices to have $N = \Omega\left(d^6\ln^3 d\right)$. This, again, is a polynomial in d; albeit having a slightly worse degree (of six).

5.4 Conclusion and Future Directions

We have studied two-layer NN models with sigmoid, ReLU, and step activations; and established that the *outer norm* of any such NN achieving a small training loss on a polynomially (in d) many data and having non-negative output weights is *well-controlled*. Our results are independent of the width $\overline{m}$ of the network and the training algorithm; and are valid under very mild distributional assumptions on input/label pairs. We then leveraged the outer norm bounds we established to obtain good generalization guarantees for the networks we investigated. Our generalization results are obtained by employing earlier results on the fat-shattering dimension of such networks, and have good sample complexity bounds as we have discussed. In particular, for certain important cases of interest, we obtain near-linear sample guarantees.

We now provide future directions. As was already mentioned, our approach operates under mild distributional requirements; and can potentially handle different distributions as well as other activations, provided (rather natural) certain properties of these objects we leveraged remain in place.

A very important question is to which extent our approach applies to deeper networks. In what follows, we give a very brief argument demonstrating that for such an extension, one needs much more stringent regularity assumptions on the internal weights. Consider, as an example, a ReLU network with three hidden layers. Observe that the outputs of the neurons at the first hidden layer are non-negative as $\texttt{ReLU}(x) \geq 0$ for all $x \in \mathbb{R}$. Let us now focus on its second hidden layer, which takes weighted sums of the outputs of the first hidden layer. If all the weights in the second layer are negative, then upon passing to ReLU, one obtains all zeroes, forcing the final output to be zero. Now, let us assume, instead, that the weights of the second layer are such that the input to the ReLU functions are positive, though arbitrarily close to zero (this can potentially be achieved, e.g., by taking many small negative weights and few large positive weights in a way that ensures proper cancellation). If this holds, then even if the outer norm, $\|a\|_1$, is very large, one still obtains a bounded output at the end of the network. As demonstrated by this conceptual example, one indeed needs more stringent assumptions on the internal weights so as to address larger depth. At the present time, we are unable to have a complete resolution of necessary and sufficient assumptions for addressing deeper architectures (while maintaining the position that these assumptions must also be sound from a practical point of view).

Yet another important direction pertains to the non-negativity of the weights, and a crucial question is whether this assumption can be relaxed. We now provide a brief argument demonstrating that in full generality, this is not necessarily the case. Namely, strictly speaking, the non-negativity assumption is necessary. We focus on the so-called "teacher/student" setting, a setting that has been quite popular recently, see, e.g. [152]. In this setting, given i.i.d. input data $X_i \in \mathbb{R}^d$, $1 \leq i \leq N$, a teacher network $(a^*, W^*) \in \mathbb{R}^{m^*} \times \mathbb{R}^{m^* \times d}$ with $m^* \in \mathbb{N}$ neurons and activation $\sigma(\cdot)$ generates the labels Y_i. That is, $Y_i = \sum_{1 \leq j \leq m^*} a_j^* \sigma\left((w_j^*)^T X_i\right)$. A student network with an $\overline{m} \in \mathbb{N}$ number of hidden units (where $\overline{m}$ is not necessarily equal to m^*) is then "trained" by minimizing the objective function (5.2) on the data (X_i, Y_i), $1 \leq i \leq N$; and the resulting network is then used for predicting the unseen data. We now construct a wider student network interpolating the data whose vector of output weights has arbitrarily large norm, by introducing many cancellations. Fix $z \in \mathbb{N}$, a non-zero $v \in \mathbb{R}^d$; and $\nu > 0$. Construct a new network $\left(\overline{a}, \overline{W}\right)$ on $m^* + 2z$ neurons as follows. Set $\overline{a_j} = a_j^*$ and $\overline{W_j} = W_j^*$ for $1 \leq j \leq m^*$. For any $m^* + 1 \leq j \leq m^* + 2z$, set $\overline{a_j} = \nu$ if j is even, and $-\nu$, if j is odd. At the same time, set $\overline{W_j} = v$ for $m^* + 1 \leq j \leq m^* + 2z$. This network interpolates the data while $\|\overline{a}\|_1 = \|a^*\|_1 + 2z\nu$. Hence, $\|\overline{a}\|_1$ can be made arbitrarily large by amplifying z and/or $\nu > 0$. In particular, in full generality, such a non-negativity assumption is indeed necessary. It is worth noting, however, that the example above is a somewhat tailored one involving many dependencies/cancellations. It might still be possible to establish similar bounds for the case of potentially negative weights under more stringent constraints on them which prevent such cancellations.

5.5 Proofs

In this section, we provide complete proofs of all of our results.

5.5.1 Proof of Theorem 5.2.1

Proof of Theorem 5.2.1. Observe that

$$\mathbb{P}(\mathcal{E}_0) \geq 1 - o_N(1) \quad \text{for} \quad \mathcal{E}_0 \triangleq \{\textstyle\sum_{i=1}^N |Y_i| \leq 2MN\}, \tag{5.10}$$

using the weak law of large numbers. Next, let $(a, W) \in \mathcal{S}(\delta, R)$. Then there exists an $\overline{m} \in \mathbb{N}$ such that $(a, W) \in \mathcal{S}(\overline{m}, \delta, R)$. Applying Cauchy-Schwarz inequality, $\sum_{1\leq i\leq N} \left| Y_i - \sum_{1\leq j\leq \overline{m}} a_j \texttt{SGM}\left(w_j^T X_i\right)\right| \leq N\delta$. Next, by the triangle inequality and the fact $\sum_i |Y_i| \leq 2MN$ on $\mathcal{E}_0$,

$$\sum_{1\leq i\leq N} \sum_{1\leq j\leq \overline{m}} a_j \texttt{SGM}\left(w_j^T X_i\right) \leq N(\delta + 2M), \tag{5.11}$$

on the event $\mathcal{E}_0$. Now, let $\mathcal{N}_\epsilon$ be a *minimal* $\epsilon-$net for $B_2(0, R)$, $\epsilon > 0$ to be tuned appropriately. Using Theorem 5.1.3, one can ensure $|\mathcal{N}_\epsilon| \leq (3R/\epsilon)^d$. Next, fix any $\widehat{w} \in \mathcal{N}_\epsilon$, and set $\overline{Z_i} \triangleq \widehat{w}^T X_i$, $1 \leq i \leq N$. Since $\overline{Z_i}$ is symmetric, $\mathbb{P}(\overline{Z_i} \geq 0) = \mathbb{P}(-\overline{Z_i} \leq 0) = \mathbb{P}(\overline{Z_i} \leq 0)$, implying $\mathbb{P}(\overline{Z_i} \geq 0) \geq \frac{1}{2}$. Define now $Z_i \triangleq \mathbb{1}\left\{\overline{Z_i} \geq 0\right\}$. Since Z_i "stochastically dominates" Bernoulli(1/2), we have $\mathbb{P}\left(\sum_{1\leq i\leq N} Z_i \geq N/3\right) \geq \mathbb{P}\left(\text{Binomial}\left(N, 1/2\right) \geq N/3\right) \geq 1 - \exp\left(-\Theta(N)\right)$. The last inequality is due to standard large deviations bounds. Taking a union bound over the net $\mathcal{N}_\epsilon$, we obtain

$$\begin{aligned} &\mathbb{P}(\mathcal{E}_1) \geq 1 - (3R/\epsilon)^d \exp\left(-\Theta(N)\right), \\ &\qquad \text{where} \quad \mathcal{E}_1 \triangleq \bigcap_{\widehat{w}\in\mathcal{N}_\epsilon} \left\{ \sum_{1\leq i\leq N} \mathbb{1}\left\{\widehat{w}^T X_i \geq 0\right\} \geq N/3 \right\}. \end{aligned} \tag{5.12}$$

Furthermore, another union bound over the data yields

$$\begin{aligned} &\mathbb{P}(\mathcal{E}_2) \geq 1 - N \exp\left(-\Theta(d)\right), \\ &\qquad \text{where} \quad \mathcal{E}_2 \triangleq \left\{\|X_i\|_2^2 \leq Cd, 1 \leq i \leq N\right\}. \end{aligned} \tag{5.13}$$

We now choose $\epsilon = 1/\sqrt{Cd}$. We claim that on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, it is the case that for every $w \in B_2(0, R)$; $\sum_{1\leq i\leq N} \mathbb{1}\left\{w^T X_i \geq -1\right\} \geq \frac{N}{3}$. Let $w \in B_2(0, R)$, and $\widehat{w} \in \mathcal{N}_\epsilon$ be such that $\|w - \widehat{w}\|_2 \leq \epsilon = (Cd)^{-1/2}$. Using Cauchy-Schwarz inequality, $\left|\widehat{w}^T X_i - w^T X_i\right| \leq \|X_i\|_2 (Cd)^{-1/2} \leq 1$, where $\|X_i\|_2 \leq \sqrt{Cd}$ due to the event $\mathcal{E}_2$ (5.13). In particular, if $\widehat{w}^T X_i \geq 0$, then $w^T X_i \geq -1$. Hence $\sum_{1\leq i\leq N} \mathbb{1}\left\{w^T X_i \geq -1\right\} \geq \sum_{1\leq i\leq N} \mathbb{1}\left\{\widehat{w}^T X_i \geq 0\right\} \geq \frac{N}{3}$. Using now the fact $a_j \geq 0$, and $\texttt{SGM}(\cdot) \geq 0$ for the sig-

moid activation, we arrive at

$$\sum_{1\le j\le \overline{m}} a_j \sum_{1\le i\le N} \mathtt{SGM}\left(w_j^T X_i\right) \ge \frac{N}{3}\cdot \mathtt{SGM}(-1)\cdot \sum_{1\le j\le \overline{m}} a_j. \tag{5.14}$$

We now combine the facts $\mathtt{SGM}(-1) = (1+e)^{-1}$, (5.11) and (5.14), to obtain that on the event $\mathcal{E}_0\cap\mathcal{E}_1\cap\mathcal{E}_2$, $\sum_{1\le j\le\overline{m}} a_j \le 3(1+e)(\delta+2M)$. Since the event $\mathcal{E}_0\cap\mathcal{E}_1\cap\mathcal{E}_2$ holds with probability at least $1-\left(3R\sqrt{Cd}\right)^d \exp\left(-\Theta(N)\right) - N\exp\left(-\Theta(d)\right) - o_N(1)$ by a union bound, the proof is complete. □

5.5.2 Proof of Theorem 5.2.3

Proof of Theorem 5.2.3. Recall from (5.10) the event $\mathcal{E}_0 = \{\sum_{1\le i\le N}|Y_i| \le 2MN\}$ where $\mathbb{P}(\mathcal{E}_0)\ge 1-o_N(1)$.

Let $(a,W)\in\mathcal{G}(\delta)$. Then, for some $\overline{m}\in\mathbb{N}$, $(a,W)\in\mathcal{G}(\overline{m},\delta)$. Using Cauchy-Schwarz inequality and the triangle inequality like in the beginning of the proof of Theorem 5.2.1; we first establish that on the event $\mathcal{E}_0$, the following holds:

$$\sum_{1\le i\le N}\sum_{1\le j\le\overline{m}} a_j \mathtt{ReLU}\left(w_j^T X_i\right) \le N(\delta+2M). \tag{5.15}$$

Next, let $\mathcal{N}_\epsilon$ be a *minimal* $\epsilon-$net for $\mathbb{S}^{d-1}$, $\epsilon>0$ to be tuned. Using Theorem 5.1.3, one can ensure $|\mathcal{N}_\epsilon|\le(3/\epsilon)^d$. Fix any $\widehat{w}\in\mathcal{N}_\epsilon$; and consider the i.i.d. random variables $Y_{\widehat{w},i}\triangleq \mathtt{ReLU}\left(\widehat{w}^T X_i\right)$, $i\in[N]$, whose mean is bounded from below by $\boldsymbol{\mu}^*$ due to condition (a). Note that the event $\left\{\sum_{1\le i\le N} Y_{\widehat{w},i}\le \frac{1}{2}\boldsymbol{\mu}^* N\right\}$ is then a large deviations event. Moreover, the condition (b) on the distribution of $Y_{\widehat{w},i}$ ensures the existence of the log-MGF in a neighborhood around zero; hence a large deviations bound via Chernoff's inequality [58] is applicable. Applying Chernoff's inequality, we thus obtain $\mathbb{P}\left(\sum_{1\le i\le N} Y_{\widehat{w},i}\ge\frac{1}{2}\boldsymbol{\mu}^*N\right)\ge 1-\exp\left(-\Theta(N)\right)$. Due to the distributional assumption, the lower bound is uniform in $\widehat{w}\in\mathcal{N}_\epsilon$.

Taking now a union bound over $\widehat{w}\in\mathcal{N}_\epsilon$, we obtain $\mathbb{P}(\mathcal{E}_1)\ge 1-(3/\epsilon)^d\exp\left(-\Theta(N)\right)$ where $\mathcal{E}_1\triangleq\bigcap_{\widehat{w}\in\mathcal{N}_\epsilon}\left\{\sum_{1\le i\le N}\mathtt{ReLU}\left(\widehat{w}^T X_i\right)\ge\frac{1}{2}\boldsymbol{\mu}^*N\right\}$. Another union bound over data X_i, $1\le i\le N$, yields that $\mathbb{P}(\mathcal{E}_2)\ge 1-N\exp\left(-\Theta(d)\right)$ where $\mathcal{E}_2\triangleq\{\|X_i\|_2^2\le Cd, 1\le i\le N\}$.

Choose $\epsilon\triangleq\frac{\boldsymbol{\mu}^*}{4\sqrt{Cd}}$, and assume in the remainder that we are on the event $\mathcal{E}_1\cap\mathcal{E}_2$. Next, observe that $\mathtt{ReLU}$ is $1-$Lipschitz: $|\mathtt{ReLU}(x)-\mathtt{ReLU}(y)| = \left|\frac{x+|x|}{2}-\frac{y+|y|}{2}\right|\le|x-y|$, using triangle inequality twice. Now, fix *any* $w\in\mathbb{S}^{d-1}$. Let $\widehat{w}\in\mathcal{N}_\epsilon$ be the member of the net closest to w. Using the Lipschitz property, and the Cauchy-Schwarz, we obtain $\left|\mathtt{ReLU}\left(w^T X_i\right)-\mathtt{ReLU}\left(\widehat{w}^T X_i\right)\right|\le\left|w^T X_i-\widehat{w}^T X_i\right|\le\|w-\widehat{w}\|_2\cdot\|X_i\|_2\le\frac{\boldsymbol{\mu}^*}{4}$. Consequently, $\mathtt{ReLU}\left(w^T X_i\right)\ge\mathtt{ReLU}\left(\widehat{w}^T X_i\right)-\frac{\boldsymbol{\mu}^*}{4}$. Summing this over $1\le i\le N$, we have $\sum_{1\le i\le N}\mathtt{ReLU}\left(w^T X_i\right)\ge\sum_{1\le i\le N}\mathtt{ReLU}\left(\widehat{w}^T X_i\right)-\frac{\boldsymbol{\mu}^*}{4}N\ge\frac{\boldsymbol{\mu}^*}{4}N$. Using $a_j\ge 0$, we

obtain by taking w_j in place of w:

$$\sum_{1\le j\le \overline{m}} a_j \sum_{1\le i\le N} \texttt{ReLU}\left(w_j^T X_i\right) \ge \frac{\boldsymbol{\mu}^*}{4} N \sum_{1\le j\le \overline{m}} a_j. \tag{5.16}$$

Combining (5.15) and (5.16), we obtain that on the event $\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2$, $\sum_{1\le j\le \overline{m}} a_j \le 4(\delta + 2M)\left(\boldsymbol{\mu}^*\right)^{-1}$. Since the event $\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2$ holds with probability at least $1 - \left(12\sqrt{Cd}(\boldsymbol{\mu}^*)^{-1}\right)^d \exp\left(-\Theta(N)\right) - N\exp\left(-\Theta(d)\right) - o_N(1)$ via a union bound, we complete the proof. □

5.5.3 Proof of Theorem 5.2.4

Proof of Theorem 5.2.4. The proof is quite similar to that of the proof of Theorems 5.2.1 and 5.2.3, and is provided for completeness.

Again, recall from (5.10) the event $\mathcal{E}_0 = \{\sum_{1\le i\le N} |Y_i| \le 2MN\}$ where $\mathbb{P}(\mathcal{E}_0) \ge 1 - o_N(1)$. Then, take an $(a, W) \in \mathcal{H}(\delta)$. There exists an $\overline{m} \in \mathbb{N}$ such that $(a, W) \in \mathcal{H}(\overline{m}, \delta)$. Using again Cauchy-Schwarz inequality and the triangle inequality like in the beginning of the proof of Theorems 5.2.1/ 5.2.3; we have that on the event $\mathcal{E}_0$, the following holds:

$$\sum_{1\le i\le N} \left| Y_i - \sum_{1\le j\le \overline{m}} a_j \texttt{Step}\left(w_j^T X_i\right) \right| \le N\delta.$$

This, together with (a) the fact that the labels are bounded, $|Y_i| \le M$; and (b) the triangle inequality; then yields

$$\sum_{1\le i\le N} \sum_{1\le j\le \overline{m}} a_j \texttt{Step}\left(w_j^T X_i\right) \le N\left(\delta + M\right). \tag{5.17}$$

Let $\mathcal{N}_\epsilon$ be a *minimal* $\epsilon-$net for $\mathbb{S}^{d-1}$, where $\epsilon > 0$ to be tuned appropriately. Using Theorem 5.1.3, one can ensure $|\mathcal{N}_\epsilon| \le (3/\epsilon)^d$.

Next, fix any $\widehat{w} \in \mathcal{N}_\epsilon$; and set $Z_i \triangleq \mathbb{1}\left\{\widehat{w}^T X_i \ge \eta\right\}$, $1 \le i \le N$ (where we drop the dependence of Z_i on $\widehat{w}$ for convenience). Evidently, Z_i is an i.i.d. collection of Bernoulli random variables, with $\mathbb{E}[Z_i] \ge \eta$ (due to the assumption on the distribution of X). Hence, using standard concentration results, $\mathbb{P}\left(\sum_{1\le i\le N} Z_i \ge N\eta/2\right) \ge 1 - \exp\left(-\Theta(N)\right)$. Moreover, the lower bound is, again, uniform in $\widehat{w}$ via an exact same stochastic domination argument, like in the proof of Theorem 5.2.1.

Taking now a union bound over the net $\mathcal{N}_\epsilon$,

$$\mathbb{P}(\mathcal{E}_1) \ge 1 - (3/\epsilon)^d \exp\left(-\Theta(N)\right), \tag{5.18}$$

where

$$\mathcal{E}_1 \triangleq \bigcap_{\widehat{w}\in\mathcal{N}_\epsilon} \left\{ \sum_{1\le i\le N} \mathbb{1}\left\{\widehat{w}^T X_i \ge \eta\right\} \ge N\eta/2 \right\}.$$

Furthermore, another union bound over data, $1 \leq i \leq N$, yields

$$\mathbb{P}(\mathcal{E}_2) \geq 1 - N \exp\left(-\Theta(d)\right), \tag{5.19}$$

where

$$\mathcal{E}_2 \triangleq \left\{\|X_i\|_2^2 \leq Cd, 1 \leq i \leq N\right\}.$$

We now choose $\epsilon = \frac{\eta}{2\sqrt{Cd}}$; and assume in the remainder that we are on the event $\mathcal{E}_1 \cap \mathcal{E}_2$.

Fix any $w \in \mathbb{S}^{d-1}$; and let $\widehat{w} \in \mathcal{N}_\epsilon$ be such that $\|w - \widehat{w}\|_2 \leq \frac{\eta}{2\sqrt{Cd}}$. Using Cauchy-Schwarz inequality, $\left|\widehat{w}^T X_i - w^T X_i\right| \leq \|w - \widehat{w}\|_2 \|X_i\|_2 \leq \eta/2$, for every $i \in [N]$, since the event we are on is a subset of $\mathcal{E}_2$ in (5.19). Observe now that $\{\widehat{w}^T X \geq \eta\} \subseteq \{w^T X \geq \eta/2\}$. Thus, on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, it holds that

$$\sum_{1 \leq i \leq N} \mathbb{1}\{w^T X_i \geq \eta/2\} \geq \sum_{1 \leq i \leq N} \mathbb{1}\{\widehat{w}^T X_i \geq \eta/2\} \geq N\eta/2.$$

Since $w \in \mathbb{S}^{d-1}$ is arbitrary, and $\texttt{Step}(w^T X_i) = 1$ if $w^T X_i \geq \eta/2 > 0$, we arrive at

$$\sum_{1 \leq j \leq \overline{m}} a_j \sum_{1 \leq i \leq N} \texttt{Step}\left(w_j^T X_i\right) \geq \frac{N\eta}{2} \sum_{1 \leq j \leq \overline{m}} a_j. \tag{5.20}$$

We now combine (5.17) and (5.20) to arrive at the conclusion that on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, it holds

$$\sum_{1 \leq j \leq \overline{m}} a_j \leq 2(\delta + M)\eta^{-1}.$$

Finally, we combine (5.18) (with $\epsilon = \frac{\eta}{2\sqrt{Cd}}$) and (5.19) via a union bound; and arrive at the conclusion that $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - \left(6\sqrt{Cd} \cdot (\eta)^{-1}\right)^d \exp\left(-\Theta(N)\right) - N \exp\left(-\Theta(d)\right) - o_N(1)$. This concludes the proof. □

5.5.4 Proof of Theorem 5.3.1

In this section, we establish Theorem 5.3.1. We build upon earlier results by Bartlett [40] and Bartlett, Long, and Williamson [43].

The FSD of the Networks with a Bounded Outer Norm We now recall the definition of the fat-shattering dimension (FSD), verbatim from [40], for convenience.

Definition 5.5.1. *Let X be an input space, H be a class of real-valued functions defined on X (that is, H consists of functions $f : X \to \mathbb{R}$). Fix a $\gamma > 0$, which is a certain* scale *parameter. We say that a sequence $(x_1, x_2, \ldots, x_m)$ of m points from X is γ-shattered by H if there is an $r = (r_1, \ldots, r_m) \in \mathbb{R}^m$ such that, for all $b = (b_1, \ldots, b_m) \in \{-1, 1\}^m$ there is an $h \in H$ satisfying $(h(x_i) - r_i)b_i \geq \gamma$. Define*

the fat-shattering dimension of H as the function

$$\mathrm{FSD}_H(\gamma) \triangleq \max\Big\{m : H \text{ } \gamma\text{-shatters some } x \in X^m\Big\}. \tag{5.21}$$

We next record the following result.

Theorem 5.5.2. *[40, Corollary 24] Let $\mathcal{M} > 0$, and $\sigma : \mathbb{R} \to [-\mathcal{M}/2, \mathcal{M}/2]$ be a non-decreasing function. Define a class F of functions on $\mathbb{R}^d$ by*

$$F \triangleq \left\{X \mapsto \sigma\left(w^T X + w_0\right) : w \in \mathbb{R}^d, w_0 \in \mathbb{R}\right\}$$

and let $H(A) \triangleq \left\{\sum_{1 \le j \le \overline{m}} a_j f_j : \overline{m} \in \mathbb{N}, f_j \in F, \|a\|_1 \le A\right\}$ where $A \ge 1$. Then for every $\gamma \le \mathcal{M}A$,

$$\mathrm{FSD}_{H(A)}(\gamma) \le \frac{c\mathcal{M}^2 A^2 d}{\gamma^2} \ln\left(\frac{\mathcal{M}A}{\gamma}\right)$$

for some universal constant $c > 0$.

Here $a = (a_j : 1 \le j \le \overline{m}) \in \mathbb{R}^{\overline{m}}$ is the vector of output weights, $\|a\|_1$ is the outer norm; and $\gamma > 0$ is a certain *scale* parameter. Observe that $H(A)$ is precisely the class of two-layer `NN` with activation function $\sigma(\cdot)$ whose outer norm is at most A. Per Theorem 5.5.2, the FSD of the class of two-layer networks with **bounded outer norm** is upper bounded by an explicit quantity.

Our next proposition provides a control for the generalization gap uniformly over all two-layer `NN` models with bounded outer norm.

Proposition 5.5.3. *Let $M, \mathcal{M}, A > 0$; $\sigma : \mathbb{R} \to [-\mathcal{M}/2, \mathcal{M}/2]$ be a non-decreasing activation function; and $\mathcal{D}$ be an arbitrary distribution on $\mathbb{R}^d \times \mathbb{R}$ for the input/label pairs (X, Y) where $|Y| \le M$ almost surely. Recall the class $H(A)$ of two-layer neural networks with activation σ and outer norm at most A from Theorem 5.5.2; and let (X_i, Y_i), $1 \le i \le N$, be i.i.d. samples drawn from $\mathcal{D}$. Then for any $\alpha > 0$, with probability at least*

$$1 - 4\exp\left(\xi(\alpha, M, \mathcal{M}, A) \cdot d \cdot \ln^2\left(\frac{576 N \mathcal{M}^2 A^2 \max\{\mathcal{M}A, 2M\}}{\alpha}\right) - \frac{\alpha^2 N}{64 \max\{\mathcal{M}A, 2M\}^2}\right)$$

over the draw of the training data (X_i, Y_i), $1 \le i \le N$, it holds that

$$\sup_{\varphi \in H(A)} \left| \frac{1}{N} \sum_{i=1}^{N} \Big(\varphi(X_i) - Y_i\Big)^2 - \mathbb{E}_{(X,Y)\sim\mathcal{D}}\Big[\Big(\varphi(X) - Y\Big)^2\Big] \right|$$

is at most α, provided

$$N \ge 64 \cdot 128c \cdot \frac{\mathcal{M}^6 A^6 \max\{\mathcal{M}A, 2M\}^2}{\alpha^2} d.$$

Here, $c, c' > 0$ are absolute constants, the term ξ is introduced in (5.6) and the

expectation is taken with respect to a fresh sample $(X, Y) \sim \mathcal{D}$ *independent of* (X_i, Y_i), $1 \leq i \leq N$.

It is worth noting that while we made no attempts for simplifying the constants appearing throughout Proposition 5.5.3, we believe that they can be improved. Proposition 5.5.3 is established by combining various existing auxiliary results from literature. However, its proof requires certain extra notation. In order to avoid notational clutter, we provide the proof in Section 5.5.5. From now on, we assume Proposition 5.5.3 is at our disposal.

We finally provide a technical lemma to be used in the proof for the ReLU case.

Lemma 5.5.4. *Suppose that the distribution of* $X \in \mathbb{R}^d$ *satisfies the assumptions of Theorem 5.2.3. Then,*

$$\begin{aligned}\lambda(d) &\triangleq \sup_{w \in \mathbb{R}^d : \|w\|_2 = 1/\sqrt{Cd}} \mathbb{E}\Big[\big|w^T X\big|^2 \mathbb{1}\left\{\|X\|_2^2 > Cd\right\}\Big] \\ &\leq \exp\left(-\Theta(d)\right). \end{aligned} \tag{5.22}$$

The scaling $\|w\|_2 = 1/\sqrt{Cd}$ is required for technical reasons for the proof of the part (b) of Theorem 5.3.1. Proof of Lemma 5.5.4 uses routine manipulations; and is provided next.

Proof of Lemma 5.5.4. Set

$$\bar{\lambda}(d) := \sup_{w \in \mathbb{R}^d : \|w\|_2 = 1} \mathbb{E}\Big[\big|w^T X\big|^2 \mathbb{1}\left\{\|X\|_2^2 > Cd\right\}\Big].$$

Clearly $\bar{\lambda}(d) = Cd\lambda(d)$. Since $C = O(1)$, it suffices to prove $\bar{\lambda}(d) \leq \exp\left(-\Theta(d)\right)$. Next, fix a $w \in \mathbb{R}^d$ with $\|w\|_2 = 1$. Observe that using the inequality $e^x \geq 1 + x$, we obtain $e^{rw^T X} + e^{-rw^T X} \geq r\Big|w^T X\Big|$ for any $r \geq 0$. Using the chain of inequalities $8\left(a^4 + b^4\right) \geq 4\left(a^2 + b^2\right)^2 \geq (a+b)^4$, both due to Cauchy-Schwarz, we thus obtain $\frac{8}{r^4}\Big(e^{4rw^T X} + e^{-4rw^T X}\Big) \geq \Big|w^T X\Big|^4$. Now, take $r = s/4$ and then take the expectation of both sides to obtain

$$\frac{2048}{s^4}\Big(M_1(s) + M_2(s)\Big) \geq \mathbb{E}\Big[\Big|w^T X\Big|^4\Big], \tag{5.23}$$

where $M_1(s)$ and $M_2(s)$ are defined in Theorem 5.2.3. Thus,

$$\begin{aligned}
&\mathbb{E}\Big[\big|w^T X\big|^2 \mathbb{1}\left\{\|X\|_2^2 > Cd\right\}\Big]^2 \\
&\leq \mathbb{E}\Big[\Big|w^T X\Big|^4\Big]\mathbb{E}\Big[\mathbb{1}\left\{\|X\|_2^2 > Cd\right\}^2\Big] && (5.24)\\
&= \mathbb{E}\Big[\Big|w^T X\Big|^4\Big]\mathbb{P}\Big(\|X\|_2^2 > Cd\Big) && (5.25)\\
&\leq \frac{2048}{s^4}\cdot\Big(M_1(s)+M_2(s)\Big)\cdot\exp\Big(-\Theta(d)\Big) && (5.26)\\
&\leq \exp\Big(-\Theta(d)\Big), && (5.27)
\end{aligned}$$

where (5.24) uses Cauchy-Schwarz inequality; (5.25) uses the fact $\mathbb{E}[\mathbb{1}\{E\}^2] = \mathbb{P}(E)$ valid for any event E; (5.26) uses (5.23) and the fact $\mathbb{P}\left(\|X\|_2^2 > Cd\right) \leq \exp(-\Theta(d))$; and finally (5.27) uses the condition (b) on the distribution of X stated in Theorem 5.2.3. Taking square roots and taking the supremum over all $\|w\|_2 = 1$, we obtain $\overline{\lambda}(d) \leq \exp(-\Theta(d))$; establishing Lemma 5.5.4. □

Equipped with Proposition 5.5.3 and Lemma 5.5.4, we now complete the proof of Theorem 5.3.1.

Proof of Theorem 5.3.1. Throughout the proof, we assume that N is a sufficiently large polynomial in d and satisfies Assumption 5.1.1. Moreover, since the labels are bounded, $|Y| \leq M$ almost surely, the $o_N(1)$ terms in Theorems 5.2.1-5.2.4 disappear, as noted previously.

For the case of sigmoid and step activations, $\mathcal{M}$ can be taken as 2. Thus, for the ξ term appearing in Proposition 5.5.3, we simply employ $\xi(\alpha, M, 2, A)$.

Part (a) Define the class $\overline{\mathcal{S}}(\delta, R) = \left\{X \mapsto \sum_{1\leq j\leq \overline{m}} a_j \texttt{SGM}\left(w_j^T X\right) : (a, W) \in \mathcal{S}(\delta, R)\right\}$, where $\mathcal{S}(\delta, R)$ is introduced in Theorem 5.2.1. Note, by the definition of $\mathcal{S}(\delta, R)$, that

$$\begin{aligned}
&\sup_{(a,W)\in\mathcal{S}(\delta,R)} \widehat{\mathcal{L}}\,(a, W) \\
&= \sup_{(a,W)\in\mathcal{S}(\delta,R)} \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \sum_{j=1}^{\overline{m}} a_j \texttt{SGM}\left(w_j^T X_i\right)^2\right)^2 \leq \delta^2.
\end{aligned}$$

Applying Theorem 5.2.1, we find that provided $N \geq \text{poly}(d)$, $\overline{\mathcal{S}}(\delta, R) \subset H(A)$ with probability bounded by (5.3), where $H(A)$ is the class defined in Theorem 5.5.2 with $\sigma(\cdot) = \texttt{SGM}(\cdot)$ and $A = 3(1+e)(\delta + 2M)$.

Finally, we (a) set $\mathcal{M} = 2$ in Proposition 5.5.3; (b) then consider $\xi(\alpha, M, 2, A)$; and (c) set $\zeta(\alpha, M, A, N)$ as in (5.7). Combining now Theorem 5.2.1 and Proposition 5.5.3 via a union bound, we establish the desired conclusion.

Part (b) As the output of the ReLU is not bounded, the situation is more involved.

First, recall from Theorem 5.2.3 the set $\mathcal{G}(\overline{m},\delta) \triangleq \Big\{(a,W) \in \mathbb{R}^{\overline{m}}_{\geq 0} \times \mathbb{R}^{\overline{m}\times d} : \|w_j\|_2 = 1, 1 \leq j \leq \overline{m}, \widehat{\mathcal{L}}\,(a,W) \leq \delta^2\Big\}$ and $\mathcal{G}(\delta) \triangleq \bigcup_{\overline{m}\in\mathbb{N}} \mathcal{G}(\overline{m},\delta)$. By Theorem 5.2.3, it holds that with probability bounded by (5.4), for any $(a,W) \in \mathcal{G}(\delta)$, $\|a\|_1 \leq 4(\delta+2M)(\boldsymbol{\mu}^*)^{-1}$. Using the homogeneity of the ReLU activation, we instead rescale w_j by $1/\sqrt{Cd}$; and consider throughout the sets

$$\widetilde{\mathcal{G}}(\overline{m},\delta) \triangleq \Big\{(a,W) \in \mathbb{R}^{\overline{m}}_{\geq 0} \times \mathbb{R}^{\overline{m}\times d} : \|w_j\|_2 = \frac{1}{\sqrt{Cd}}; \\ j \in [\overline{m}], \widehat{\mathcal{L}}\,(a,W) \leq \delta^2\Big\} \quad \text{and} \quad \widetilde{\mathcal{G}}(\delta) \triangleq \bigcup_{\overline{m}\in\mathbb{N}} \mathcal{G}(\overline{m},\delta). \tag{5.28}$$

Then, with probability at least

$$1 - \left(\frac{12\sqrt{Cd}}{\boldsymbol{\mu}^*}\right)^d \exp\left(-\Theta(N)\right) - N\exp\left(-\Theta(d)\right), \tag{5.29}$$

it holds that

$$\sup_{(a,W)\in\widetilde{G}(\delta)} \|a\|_1 \leq \frac{4\sqrt{Cd}(\delta+2M)}{\boldsymbol{\mu}^*}. \tag{5.30}$$

We now define an activation function, $\texttt{S-ReLU}(\cdot)$, which is a "saturated" version of the ReLU: $\texttt{S-ReLU}(x) = 0$ for $x < 0$, is x for $0 \leq x < 1$; and is 1 for $x \geq 1$. Next, using a union bound over data (X_i, Y_i), $1 \leq i \leq N$, $\mathbb{P}\Big(\|X_i\|_2^2 \leq Cd, 1 \leq i \leq N\Big) \geq 1 - N\exp\left(-\Theta(d)\right)$. Hence by Cauchy-Schwarz inequality, $\mathbb{P}\left(\sup_{\|w\|_2=\frac{1}{\sqrt{Cd}}} \left|w^T X_i\right| \leq 1, 1 \leq i \leq N\right) \geq 1 - N\exp\left(-\Theta(d)\right)$. Consequently, w.p. at least $1 - N\exp(-\Theta(d))$ over (X_i, Y_i); it holds that for all $(a,W) \in \widetilde{G}(\delta)$

$$\frac{1}{N}\sum_{1\leq i\leq N}\left(Y_i - \sum_{1\leq j\leq \overline{m}} a_j \texttt{ReLU}\left(w_j^T X_i\right)\right)^2 = \frac{1}{N}\sum_{1\leq i\leq N}\left(Y_i - \sum_{1\leq j\leq \overline{m}} a_j \texttt{S-ReLU}\left(w_j^T X_i\right)\right)^2 \leq \delta^2. \tag{5.31}$$

Define next the class

$$\overline{\mathcal{G}}(\delta) \triangleq \left\{X \mapsto \sum_{1\leq j\leq \overline{m}} a_j \texttt{S-ReLU}\left(w_j^T X\right) : (a,W) \in \widetilde{\mathcal{G}}(\delta)\right\}. \tag{5.32}$$

Note that, this set consists of all two-layer neural networks with (a) activation $\texttt{S-ReLU}(\cdot)$, the saturated version of $\texttt{ReLU}(\cdot)$; and (b) weights trained on the $\texttt{ReLU}(\cdot)$ network.

By Theorem 5.2.3 and (5.30), we find that provided $N \geq \text{poly}(d)$, $\overline{\mathcal{G}}(\delta) \subset H(A)$ with probability given by (5.29), where $H(A)$ is the class defined in Theorem 5.5.2 with $\sigma(\cdot) = \texttt{S-ReLU}(\cdot)$ and $A = 4\sqrt{Cd}(\delta+2M)/\boldsymbol{\mu}^*$.

Observe that S-ReLU is a non-decreasing activation with bounded range. Hence, Proposition 5.5.3 applies: one can simply take $\mathcal{M} = 2$. We now apply Proposition 5.5.3 with $\mathcal{M} = 2$, and $A = 4\sqrt{Cd}(\delta + 2M)(\boldsymbol{\mu}^*)^{-1}$ as in (5.30). Combining the probability bound (5.29) and the one in Proposition 5.5.3 by a union bound, we find that for every $\alpha > 0$, with probability at least

$$1 - \zeta\left(\alpha, M, \frac{4\sqrt{Cd}(\delta + 2M)}{\boldsymbol{\mu}^*}, N\right) - \left(\frac{12\sqrt{Cd}}{\boldsymbol{\mu}^*}\right)^d \exp\left(-\Theta(N)\right) - N\exp\left(-\Theta(d)\right) \tag{5.33}$$

(where ξ is introduced in (5.7)) over training data (X_i, Y_i), $1 \le i \le N$, it holds that

$$\sup_{\varphi \in \overline{\mathcal{G}}(\delta)} \left| \frac{1}{N}\sum_{i=1}^{N} \left(Y_i - \varphi\left(X_i\right)\right)^2 - \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\left(Y - \varphi(X)\right)^2\right] \right|,$$

is at most α, where $\overline{\mathcal{G}}(\delta)$ introduced in (5.32). Recalling also (5.31) holding w.p. $1 - N\exp(-\Theta(d))$; we conclude that

$$\sup_{(a,W)\in\widetilde{\mathcal{G}}(\delta)} \mathbb{E}_{(X,Y)\sim\mathcal{D}} \left[\left(Y - \sum_{j=1}^{\overline{m}} a_j \texttt{S-ReLU}\left(w_j^T X\right)\right)^2\right], \tag{5.34}$$

is at most $\alpha + \delta^2$, with probability at least

$$1 - \zeta\left(\alpha, M, \frac{4\sqrt{Cd}(\delta + 2M)}{\boldsymbol{\mu}^*}, N\right) - \left(\frac{12\sqrt{Cd}}{\boldsymbol{\mu}^*}\right)^d \exp\left(-\Theta(N)\right) - 2N\exp\left(-\Theta(d)\right). \tag{5.35}$$

We next fix an $(a, W) \in \widetilde{\mathcal{G}}(\delta)$, and study the quantity

$$\begin{aligned}\Delta(a, W) \triangleq \Bigg| &\mathbb{E}_{(X,Y)\sim\mathcal{D}} \left[\left(Y - \sum_{j=1}^{\overline{m}} a_j \texttt{ReLU}\left(w_j^T X\right)\right)^2\right] \\ &- \mathbb{E}_{(X,Y)\sim\mathcal{D}} \left[\left(Y - \sum_{j=1}^{\overline{m}} a_j \texttt{S-ReLU}\left(w_j^T X\right)\right)^2\right]\Bigg|. \end{aligned} \tag{5.36}$$

This quantity is nothing but the difference of generalization errors between two networks of same architecture, same number $\overline{m}$ of hidden units and same weights (a, W); but different activations, $\texttt{ReLU}(\cdot)$ and $\texttt{S-ReLU}(\cdot)$.

For convenience, denote

$$\varphi_{SR}(X) \triangleq \sum_{1\le j\le\overline{m}} a_j \texttt{S-ReLU}\left(w_j^T X\right)$$

and

$$\varphi_R(X) \triangleq \sum_{1\le j\le \overline{m}} a_j \texttt{ReLU}\left(w_j^T X\right).$$

In what follows, we employ the simple observation that since $a_j \ge 0$ and $0 \le \texttt{S-ReLU}(x) \le 1$, $0 \le \varphi_{\rm SR}(X) \le \|a\|_1$.

Suppressing the subscript $(X,Y)\sim\mathcal{D}$ from the expectations, we have

$$\begin{aligned}
&\Delta(a,W)\\
&= \left|\mathbb{E}\left[(Y-\varphi_{SR}(X))^2\right] - \mathbb{E}\left[(Y-\varphi_R(X))^2\right]\right| && (5.37)\\
&= \left|\mathbb{E}\left[2Y\varphi_R(X) - 2Y\varphi_{SR}(X)\right] + \mathbb{E}\left[\varphi_{SR}(X)^2 - \varphi_R(X)^2\right]\right| && (5.38)\\
&\le \left|\mathbb{E}\left[2Y\varphi_R(X) - 2Y\varphi_{SR}(X)\right]\right| + \left|\mathbb{E}\left[\varphi_{SR}(X)^2 - \varphi_R(X)^2\right]\right| && (5.39)\\
&\le \mathbb{E}\Big[\Big|2Y\varphi_R(X) - 2Y\varphi_{SR}(X)\Big|\Big] + \mathbb{E}\Big[\Big|\varphi_{SR}(X)^2 - \varphi_R(X)^2\Big|\Big]. && (5.40)
\end{aligned}$$

Above, (5.37) follows by the definition of $\Delta(a,W)$ per (5.36); (5.38) follows after simple algebra; (5.39) follows by the triangle inequality; and (5.40) follows by the Jensen's inequality.

We next study two individual terms appearing in (5.40) separately, while keeping in mind that $(a,W)\in\widetilde{\mathcal{G}}(\delta)$ implies $a_j\ge 0$ for $1\le j\le\overline{m}$ and $\|w_j\|_2 = 1/\sqrt{Cd}$ for $1\le j\le\overline{m}$. For convenience, set $\overline{\mathcal{E}} \triangleq \{\|X\|_2^2 > Cd\}$. We have

$$\begin{aligned}
&\mathbb{E}\Big[\Big|2Y\varphi_R(X) - 2Y\varphi_{SR}(X)\Big|\Big]\\
&\le 2M\mathbb{E}\Big[\Big|\varphi_R(X) - \varphi_{SR}(X)\Big|\Big] && (5.41)\\
&= 2M\Big(\mathbb{E}\Big[|\varphi_R(X) - \varphi_{SR}(X)|\,\Big|\overline{\mathcal{E}}^c\Big]\mathbb{P}\left(\overline{\mathcal{E}}^c\right)\Big)\\
&+ 2M\Big(\mathbb{E}\Big[\Big|\varphi_R(X) - \varphi_{SR}(X)\Big|\mathbb{1}\Big\{\overline{\mathcal{E}}\Big\}\Big] && (5.42)\\
&\le 2M\mathbb{E}\Big[\Big|\varphi_R(X) - \varphi_{SR}(X)\Big|\mathbb{1}\Big\{\overline{\mathcal{E}}\Big\}\Big] && (5.43)\\
&\le 2M\Big(\mathbb{E}\Big[\varphi_R(X)\mathbb{1}\Big\{\overline{\mathcal{E}}\Big\}\Big] + \mathbb{E}\Big[\varphi_{SR}(X)\mathbb{1}\Big\{\overline{\mathcal{E}}\Big\}\Big]\Big) && (5.44)\\
&\le 2Me^{-\Theta(d)}\|a\|_1\left(\sqrt{\lambda(d)}+1\right). && (5.45)
\end{aligned}$$

Here, (5.41) uses the fact $|Y|\le M$ almost surely; (5.42) is by the law of total expectation; (5.43) uses the fact that on the event $\|X\|_2^2\le Cd$, $\varphi_R(X) = \varphi_{SR}(X)$ since $\|w_j\|_2 = 1/\sqrt{Cd}$; (5.44) uses the triangle inequality; and finally (5.45) uses the facts $0\le\texttt{S-ReLU}(x)\le 1$ for every x, $a_j\ge 0$ for every $1\le j\le\overline{m}$; $\texttt{ReLU}(x)\le|x|$; and $\mathbb{E}\Big[\Big|w_j^TX\Big|\mathbb{1}\{\overline{\mathcal{E}}\}\Big] \le \sqrt{\mathbb{E}\Big[\Big|w_j^TX\Big|^2\mathbb{1}\Big\{\overline{\mathcal{E}}\Big\}\Big]\cdot\mathbb{E}\Big[\mathbb{1}\Big\{\overline{\mathcal{E}}\Big\}\Big]} \le e^{-\Theta(d)}\sqrt{\lambda(d)}$ using Lemma 5.5.4 and Cauchy-Schwarz inequality. Here, $\lambda(d)$ is the function defined in (5.22).

We now study the second term in (5.40). Observe that

$$
\begin{aligned}
&\mathbb{E}\Big[\Big|\varphi_{SR}(X)^2-\varphi_R(X)^2\Big|\Big]\\
&=\mathbb{E}\Big[\Big|\varphi_{SR}(X)^2-\varphi_R(X)^2\Big|\,\Big|\overline{\mathcal{E}}^c\Big]\mathbb{P}\left(\overline{\mathcal{E}}^c\right)+\mathbb{E}\Big[\Big|\varphi_{SR}(X)^2-\varphi_R(X)^2\Big|\mathbb{1}\Big\{\overline{\mathcal{E}}\Big\}\Big] &(5.46)\\
&=\mathbb{E}\Big[\Big|\varphi_{SR}(X)^2-\varphi_R(X)^2\Big|\mathbb{1}\Big\{\overline{\mathcal{E}}\Big\}\Big] &(5.47)\\
&\le\Big(\mathbb{E}\Big[\varphi_{SR}(X)^2\mathbb{1}\Big\{\overline{\mathcal{E}}\Big\}\Big]+\mathbb{E}\Big[\varphi_R(X)^2\mathbb{1}\Big\{\overline{\mathcal{E}}\Big\}\Big]\Big) &(5.48)\\
&\le\Big(e^{-\Theta(d)}\cdot\|a\|_1^2+\sum_{1\le j\le\overline{m}}a_j^2\mathbb{E}\Big[\mathtt{ReLU}\left(w_j^TX\right)^2\mathbb{1}\{\overline{\mathcal{E}}\}\Big]\\
&+2\sum_{1\le j_1<j_2\le\overline{m}}a_{j_1}a_{j_2}\mathbb{E}\Big[\mathtt{ReLU}\left(w_{j_1}^TX\right)\mathtt{ReLU}\left(w_{j_2}^TX\right)\mathbb{1}\{\overline{\mathcal{E}}\}\Big]\Big) &(5.49)\\
&\le e^{-\Theta(d)}\|a\|_1^2+\lambda(d)\sum_{1\le j\le\overline{m}}a_j^2+2\lambda(d)\sum_{1\le j_1<j_2\le\overline{m}}a_{j_1}a_{j_2} &(5.50)\\
&=e^{-\Theta(d)}\|a\|_1^2+\lambda(d)\|a\|_1^2. &(5.51)
\end{aligned}
$$

Indeed, (5.46) is again by the law of total expectation; (5.47) uses the fact that on $\|X\|_2^2\le Cd$, $\varphi_{SR}(X)=\varphi_R(X)$ since $\|w_j\|_2=1/\sqrt{Cd}$; (5.48) uses triangle inequality; (5.49) is obtained by opening the parantheses while using $a_i\ge 0$, $0\le\mathtt{S\text{-}ReLU}(x)\le 1$; (5.50) uses the fact $a_j\ge 0$, Lemma 5.5.4 as well as the Cauchy-Schwarz inequality

$$
\begin{aligned}
&\mathbb{E}\Big[\mathtt{ReLU}\left(w_{j_1}^TX\right)\mathtt{ReLU}\left(w_{j_2}^TX\right)\mathbb{1}\{\overline{\mathcal{E}}\}\Big]\\
&\le\sqrt{\mathbb{E}\Big[\mathtt{ReLU}^2\left(w_{j_1}^TX\right)\mathbb{1}\{\overline{\mathcal{E}}\}\Big]}\cdot\sqrt{\mathbb{E}\Big[\mathtt{ReLU}^2\left(w_{j_2}^TX\right)\mathbb{1}\{\overline{\mathcal{E}}\}\Big]}
\end{aligned}
$$

which is at most $\lambda(d)$ as $\mathtt{ReLU}(x)\le|x|$. Finally, (5.51) is obtained by just noticing that for $a_j\ge 0$, $\|a\|_1^2=\left(\sum_{1\le j\le\overline{m}}a_j\right)^2=\sum_{1\le j\le\overline{m}}a_j^2+2\sum_{1\le j_1<j_2\le\overline{m}}a_{j_1}a_{j_2}$. We now combine (5.45) and (5.51) to upper bound the right hand side of (5.40) and arrive at $\Delta(a,W)\le 2Me^{-\Theta(d)}\|a\|_1\left(\sqrt{\lambda(d)}+1\right)+\|a\|_1^2e^{-\Theta(d)}$, where we used the fact $\lambda(d)\le\exp(-\Theta(d))$ by (5.22). Since $\|a\|_1\le 4\sqrt{Cd}(\delta+2M)/\boldsymbol{\mu}^*$ on $\widetilde{\mathcal{G}}(\delta)$ as recorded in (5.30), we obtain that $\sup_{(a,W)\in\widetilde{\mathcal{G}}(\delta)}\Delta(a,W)$ is at most

$$
\sup_{(a,W)\in\widetilde{\mathcal{G}}(\delta)}\Delta(a,W)\le e^{-\Theta(d)}\left(\frac{8M\sqrt{C}(\delta+2M)}{\boldsymbol{\mu}^*}\times\sqrt{d}\left(\sqrt{\lambda(d)}+1\right)+\frac{16C(\delta+2M)^2}{{\boldsymbol{\mu}^*}^2}d\right). \tag{5.52}
$$

Recall that $\lambda(d)\le\exp(-\Theta(d))$ by (5.22). Note that as long as $M,C,\delta,(\boldsymbol{\mu}^*)^{-1}=\exp(o(d))$ as well, the term on the right hand side of (5.52) is $e^{-\Theta(d)}$.

We finally combine (5.34), (5.36); and (5.52) to obtain

$$\sup_{(a,W)\in\mathcal{G}(\delta)} \mathbb{E}_{(X,Y)\sim\mathcal{D}} \left[\left(Y - \sum_{j=1}^{\overline{m}} a_j \texttt{ReLU}\left(w_j^T X\right) \right)^2 \right]$$

is at most $\alpha + \delta^2 + e^{-\Theta(d)}$ with probability at least

$$1 - \zeta\left(\alpha, M, \frac{4\sqrt{Cd}(\delta + 2M)}{\boldsymbol{\mu}^*}, N\right) - \left(\frac{12\sqrt{Cd}}{\boldsymbol{\mu}^*}\right)^d \exp\left(-\Theta(N)\right) - 2N\exp\left(-\Theta(d)\right),$$

as shown in (5.35). This concludes the proof of Part (b).

Part (c) This is quite similar to Part (a). Define the class

$$\overline{\mathcal{H}}(\delta) \triangleq \left\{ X \mapsto \sum_{1\le j\le \overline{m}} a_j \texttt{Step}\left(w_j^T X\right) : (a, W) \in \mathcal{H}(\delta) \right\}$$

where $\mathcal{H}(\delta)$ is introduced in Theorem 5.2.4. Note, by definition, that

$$\sup_{(a,W)\in\mathcal{H}(\delta)} \widehat{\mathcal{L}}\left(a, W\right) = \sup_{(a,W)\in\mathcal{H}(\delta)} \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \sum_{j=1}^{\overline{m}} a_j \texttt{Step}\left(w_j^T X_i\right)^2 \right) \le \delta^2.$$

Applying Theorem 5.2.4, we find that provided $N \ge \text{poly}(d)$, $\overline{\mathcal{H}}(\delta) \subset H(A)$ w.h.p. , where $H(A)$ is the class defined in Theorem 5.5.2 with $\sigma(\cdot) = \texttt{Step}(\cdot)$ and $A = 2(\delta + 2M)/\eta$.

Like in the previous case, we then (a) set $\mathcal{M} = 2$ in Proposition 5.5.3; (b) then let $\xi(\alpha, M, 2)$ to be $\xi(\alpha, M, 2, A)$; and (c) set $\zeta(\alpha, M, A, N)$ as in (5.7). Combining now Theorem 5.2.4 and Proposition 5.5.3 via a union bound, we establish the desired conclusion. □

5.5.5 Proof of Proposition 5.5.3

In this section, we prove Proposition 5.5.3. For the results we cite from [43], the numbers recorded below are from the version accessed at `http://phillong.info/publications/fatshat.pdf`[1].

Some Extra Notation on Covering Numbers We introduce several quantities verbatim from [43]. Let W be an arbitrary set, and $f : W \to \mathbb{R}$ be any function. For any $w = (w_1, \ldots, w_N) \in W^N$, denote by $f|_w$ the N-tuple $(f(w_1), f(w_2), \ldots, f(w_N)) \in$

[1]See the archived version at `http://web.archive.org/web/20200921180645/http://phillong.info/publications/fatshat.pdf` if the link above is expired.

$\mathbb{R}^N$. For a class $\mathcal{C}$ of functions $f : W \to \mathbb{R}$, let $\mathcal{C}|_w \subseteq \mathbb{R}^N$ denotes the set

$$\mathcal{C}|_w \triangleq \Big\{ f|_w : f \in \mathcal{C} \Big\} = \Big\{ \Big(f(w_1), \ldots, f(w_N) \Big) : f \in \mathcal{C} \Big\} \subseteq \mathbb{R}^N. \tag{5.53}$$

Next, recall the covering numbers from Definition 5.1.2. Throughout this section, and in particular the proof of Theorem 5.3.1, we take the metric ρ appearing in Definition 5.1.2 to be the normalized ℓ_1 distance: for any $w, \bar{w} \in \mathbb{R}^N$, set $\rho(w, \bar{w}) = \frac{1}{N} \sum_{1 \le i \le N} |w_i - \bar{w}_i|$. For any $U \subseteq \mathbb{R}^N$, denote by $\mathcal{N}(\epsilon, U)$ the covering number of U (at scale ϵ) with respect to the metric ρ above. That is, $\mathcal{N}(\epsilon, U)$ is the cardinality of the smallest $\mathcal{N}_\epsilon \subset U$ (if finite) such that for every $w \in U$, there exists a $\bar{w} \in N_\epsilon$ with $\rho(w, \bar{w}) = \frac{1}{N} \sum_{1 \le i \le N} |w_i - \bar{w}_i| \le \epsilon$. (It is worth noting that here we flipped the order of arguments in $\mathcal{N}$ appearing in Definition 5.1.2. The rationale for this is to be consistent with the notation of Bartlett et al. [43].)

Throughout this section, we often consider the following special case of $\mathcal{N}(\cdot, \cdot)$: we employ $\mathcal{N}(\cdot, \mathcal{C}|_w)$ for appropriate classes $\mathcal{C}$ of functions where w is an element of the Euclidean space $\mathbb{R}^N$ for some N.

Proof of Proposition 5.5.3. We first provide a result established originally in [164, Theorem 3,p. 107].

Theorem 5.5.5. *Let X, Y be sets; G be a PH-permissible class of $[0, T]$-valued functions defined on $Z \triangleq X \times Y$ where $T \in \mathbb{R}^+$, and P be any distribution on Z. Suppose Z_i, $1 \le i \le N$, are i.i.d. samples from P. Then for any $\alpha > 0$, with probability at least*

$$1 - 4 \left(\sup_{z \in Z^{2N}} \mathcal{N} \left(\frac{\alpha}{16}, G\Big|_z \right) \right) \cdot \exp\left(-\alpha^2 N / 64 T^2 \right)$$

over data Z_i, $1 \le i \le N$, it holds that

$$\sup_{g \in G} \left| \frac{1}{N} \sum_{i=1}^{N} g(Z_i) - \mathbb{E}_{Z \sim P}[g(Z)] \right| \le \alpha,$$

where $\mathbb{E}[g(Z)]$ is taken with respect to a fresh sample (namely a sample drawn from P, and independent of Z_i).

The version we record above is verbatim from [43, Theorem 13]. (The parameters M and m in [43] are replaced, respectively, with the parameters T and N above.)

Here, *PH-permissible* refers to a rather mild measurability constraint[2], see [164, Section 9.2]. The precise details of this technicality are immaterial to us; and it is satisfied for our purposes. Moreover, $\mathcal{N}(\cdot, \cdot)$ is the covering numbers quantity defined above.

In what follows, we apply Theorem 5.5.5. Specifically, we take $X = \mathbb{R}^d$, $Y = [0, M]$ (recall that the labels are bounded almost surely by M) thus $Z = \mathbb{R}^d \times [0, M]$ and we

[2]The letters H and P stand, respectively, for Haussler and Pollard—who gave a preliminary version of Theorem 5.5.5.

set P to simply be $\mathcal{D}$, the distribution from which the data are drawn. We then set

$$G \triangleq \Big\{(\varphi(X)-Y)^2 : X \in \mathbb{R}^d, Y \in [0,M], \varphi(\cdot) \in H(A)\Big\}, \tag{5.54}$$

and take T to be $\max\{\mathcal{M}A, 2M\}^2$ (see below).This is nothing but the ℓ_2 error obtained for predicting the label Y with $\varphi(X)$, with X being the input and $\varphi(\cdot)$ being the "predictor". With these, we obtain immediately

$$\sup_{\varphi \in H(A)} \left| \frac{1}{N} \sum_{i=1}^{N} \Big(\varphi(X_i) - Y_i\Big)^2 - \mathbb{E}_{(X,Y)\sim\mathcal{D}}\Big[(\varphi(X)-Y)^2\Big] \right| \tag{5.55}$$

is at most α; with probability at least

$$1 - 4\left(\sup_{z \in Z^{2N}} \mathcal{N}\left(\frac{\alpha}{16}, G\Big|_z\right)\right) \cdot \exp\left(-\frac{\alpha^2 N}{64 \max\{\mathcal{M}A, 2M\}^2}\right) \tag{5.56}$$

over data $Z_i = (X_i, Y_i) \sim \mathcal{D}$, $1 \le i \le N$. Above, we used the facts (a) $|Y| \le M$ almost surely; and (b) for any $\varphi \in H(A)$, it is the case $\varphi(X) = \sum_{1 \le j \le \overline{m}} a_j \sigma\left(w_j^T X\right)$ (for an $\overline{m} \in \mathbb{N}$ and $w_j \in \mathbb{R}^d$, $1 \le j \le \overline{m}$), where $\|a\|_1 \le A$ and $\sup_{x \in \mathbb{R}} |\sigma(x)| \le \mathcal{M}/2$. These together with the triangle inequality yield $-\mathcal{M}A/2 \le \varphi(X) \le \mathcal{M}A/2$ and $-M \le Y \le M$. Hence, $-\max\{\mathcal{M}A/2, M\} \le \varphi(X), Y \le \max\{\mathcal{M}A/2, M\}$; implying $\Big(\varphi(X) - Y\Big)^2 \le \max\{\mathcal{M}A, 2M\}^2$. Thus, T can be taken as $\max\{\mathcal{M}A, 2M\}^2$. We next study covering number quantity $\sup_{z \in Z^{2N}} \mathcal{N}\left(\alpha/16, G|_z\right)$ appearing in (5.56). For this, we rely on the following result taken verbatim from [43, Lemma 17].

Lemma 5.5.6. *Let X be a set, and F be a set of functions from X to $[0,1]$. Then for any $\epsilon > 0$ and any $N \in \mathbb{N}$, if $a \le 0$ and $b \ge 1$, we have*

$$\sup_{z \in (X \times [a,b])^N} \mathcal{N}\left(\epsilon, (\ell_F)\Big|_z\right) \le \sup_{x \in X^N} \mathcal{N}\left(\frac{\epsilon}{3|b-a|}, F\Big|_x\right).$$

Here, $\ell_f(x,y) = (f(x)-y)^2$, $\ell_F = \{\ell_f : f \in F\}$, and for $z = (z_1, \ldots, z_N)$ (where $z_i = (x_i, y_i)$), $(\ell_F)|_z = \{(\ell_f(x_i,y_i)^2 : 1 \le i \le N) : f \in F\}$, which is the notation introduced in (5.53) with $\mathcal{C} := \ell_F$ and $w := z$.

We take $F = H(A)$ and $\ell_F = G$ to arrive at

$$\sup_{z \in Z^{2N}} \mathcal{N}\left(\frac{\alpha}{16}, G\Big|_z\right) \le \sup_{x \in \left(\mathbb{R}^d\right)^{2N}} \mathcal{N}\left(\frac{\alpha}{32\mathcal{M}A\max\{\mathcal{M}A, 2M\}}, H(A)|_x\right). \tag{5.57}$$

Here, in addition to inserting $\alpha/16$, we also rescaled ϵ so as to reflect the fact that the functions in $H(A)$ take values in $[0, \mathcal{M}A]$. (While all the bounds established by Bartlett et al. in [43] assume the output space to be $[0,1]$, they extend in a straightforward manner to any output spaces of form $[L,U]$ by rescaling corresponding parameters. This is already noted in the beginning of [43, Section 6].)

We next record yet another result by Bartlett et al. [43, Corollary 16].

Lemma 5.5.7. *Let F be a class of $[0,1]$-valued functions defined on X, $0 < \epsilon < 1/2$ and $2N \geq \mathrm{FSD}_F(\epsilon/4)$. Then,*

$$\sup_{x \in X^N} \mathcal{N}\left(\epsilon, F|_x\right) \leq \exp\left(\frac{2}{\ln 2} \mathrm{FSD}_F(\epsilon/4) \ln^2 \frac{9N}{\epsilon}\right),$$

where the quantity $\mathrm{FSD}_F(\cdot)$ stands for the fat-shattering dimension introduced in (5.21).

(While we again skip the proof of this lemma, it is worth noting that it is obtained by combining two earlier results by Alon et al. [12, Lemmas 14,15].)

Taking now $X = \mathbb{R}^d$ and $F = H(A)$ in Lemma 5.5.7; rescaling ϵ to $\frac{\epsilon}{\mathcal{M}A}$; and then plugging $\epsilon = \frac{\alpha}{32\mathcal{M}A \max\{\mathcal{M}A, 2M\}}$ as in (5.57), we obtain

$$\begin{aligned}
\sup_{x \in \left(\mathbb{R}^d\right)^{2N}} \mathcal{N}\left(\frac{\alpha}{32\mathcal{M}A \max\{\mathcal{M}A, 2M\}}, H(A)|_x\right) \\
\leq \exp\left(\frac{2}{\ln 2} \mathrm{FSD}_{H(A)}\left(\frac{\alpha}{128\mathcal{M}^2 A^2 \max\{\mathcal{M}A, 2M\}}\right)\right. \\
\left. \cdot \ln^2\left(\frac{576 N \mathcal{M}^2 A^2 \max\{\mathcal{M}A, 2M\}}{\alpha}\right)\right). \quad (5.58)
\end{aligned}$$

We finally apply Theorem 5.5.2 above to upper bound the FSD term appearing in (5.58). Provided

$$\frac{\alpha}{128\mathcal{M}^2 A^2 \max\{\mathcal{M}A, 2M\}} \leq \mathcal{M}A$$

that is $\alpha \leq 128\mathcal{M}^3 A^3 \max\{\mathcal{M}A, 2M\}$; it holds that

$$\begin{aligned}
&\mathrm{FSD}_{H(A)}\left(\frac{\alpha}{128\mathcal{M}^2 A^2 \max\{\mathcal{M}A, 2M\}}\right) \\
&\leq \frac{128^2 c \mathcal{M}^6 A^6 \max\{\mathcal{M}A, 2M\}^2}{\alpha^2} d \\
&\cdot \ln\left(\frac{128\mathcal{M}^3 A^3 \max\{\mathcal{M}A, 2M\}}{\alpha}\right)
\end{aligned}$$

where $c > 0$ is the absolute constant appearing in Theorem 5.5.2.

Finally, combining this bound with chain of equations (5.56), (5.57), (5.58) we complete the proof. □

Chapter 6

Stationary Points of Two-Layer Quadratic Networks and the Global Optimality of the Gradient Descent Algorithm

6.1 Introduction

Neural network architectures are demonstrated to be extremely powerful in practical tasks such as natural language processing [79], image recognition [165], image classification [192], speech recognition [219], and game playing [257]; and is becoming popular in other areas, such as applied mathematics [72, 286], clinical diagnosis [86]; and so on. Despite this empirical success, a rigorous mathematical understanding of these architectures is still an ongoing quest.

While it is NP-hard to train such architectures in the worst-case setting, it has been observed empirically that the gradient descent, albeit being a simple first-order local procedure, is rather successful in training such networks. This is somewhat surprising due to the highly non-convex nature of the associated objective function. Our main motivation in this work is to provide further insights into the optimization landscape and generalization abilities of these networks.

6.1.1 Model, Contributions, and Comparison with the Prior Work

In this section, we introduce the model considered in this work, describe our contributions and discuss the relevant literature.

Model. In this work, we consider a shallow neural network architecture with one hidden layer of width m. Namely, the network consists of m neurons. We study it under the realizable model assumption, that is, the labels are generated by a teacher network with ground truth weight matrix $W^* \in \mathbb{R}^{m \times d}$ whose j^{th} row $W_j^* \in \mathbb{R}^d$ car-

ries the weights of j^{th} neuron and $m \geq d$. We assume that the input data $X \in \mathbb{R}^d$ consists of i.i.d. centered sub-Gaussian coordinates. It is worth noting that such shallow architectures with planted weights and Gaussian input data have been explored extensively in the literature, see e.g. [101, 205, 276, 297, 263, 64].

Our focus is in particular on networks with quadratic activation, studied also by Soltanolkotabi, Javanmard and Lee [264]; and Du and Lee [98], among others. This object, an instance of what is known as a *polynomial network* [210], computes for every input data $X \in \mathbb{R}^d$ the function:

$$f(W^*; X) = \sum_{j=1}^{m} \left\langle W_j^*, X \right\rangle^2 = \|W^* X\|_2^2. \tag{6.1}$$

We note that albeit being a stylized activation function, blocks of quadratic activations can be stacked together to approximate deeper networks with sigmoid activations as shown by Livni et al. [210]; and furthermore this activation serves as a second order approximation of general non-linear activations as noted by Venturi et al. [279]. Thus, we study the quadratic networks as an attempt to gain further insights on more complex networks.

Let $X_i \in \mathbb{R}^d$, $1 \leq i \leq N$ be an i.i.d. collection of input data, and let $Y_i = f(W^*; X_i)$ be the corresponding label generated per (6.1). The goal of the learner is as follows: given the training data $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $1 \leq i \leq N$, find a weight matrix $W \in \mathbb{R}^{m \times d}$ that explains the input-output relationship on the training data set in the best possible way, often by solving the so-called "*empirical risk minimization*" (ERM) optimization problem

$$\min_{W \in \mathbb{R}^{m \times d}} \widehat{\mathcal{L}}(W) \quad \text{where} \quad \widehat{\mathcal{L}}(W) \triangleq \frac{1}{N} \sum_{1 \leq i \leq N} \Big(Y_i - f(W; X_i) \Big)^2; \tag{6.2}$$

and understand its generalization ability, quantified by the "*generalization error*" (also known as the "*population risk*" associated with any solution candidate $W \in \mathbb{R}^{m \times d}$) that is given by

$$\mathcal{L}(W) \triangleq \mathbb{E}\Big[\Big(f(W^*; X) - f(W; X) \Big)^2 \Big], \tag{6.3}$$

where the expectation is with respect to a "fresh" sample X, which has the same distribution as $X_i, 1 \leq i \leq N$, but is independent from the sample. The landscape of the loss function $\widehat{\mathcal{L}}(\cdot)$ is non-convex, therefore rendering the optimization problem (potentially) difficult. Nevertheless, the gradient descent algorithm, despite being a simple first-order procedure, is rather successful in training neural networks in general: it appears to find a $W \in \mathbb{R}^{m \times d}$ with near-optimal $\widehat{\mathcal{L}}(W)$. Our partial motivation is to investigate this phenomenon in the case where the activation function is quadratic.

Contributions. We first study the landscape of risk functions and quantify an *"energy barrier"* separating rank-deficient matrices from the full-rank planted weights.

Specifically, if $W^* \in \mathbb{R}^{m\times d}$ is full-rank, namely rank d (recall $m \geq d$), then the risk function for any rank-deficient W is bounded away from zero by an explicit constant—independent of d—controlled by the smallest singular value $\sigma_{\min}(W^*)$ of W^*, as well as the second and the fourth moments of the data. See Theorem 6.2.1 for the population and Theorem 6.2.2 for the empirical versions of this result. (Theorem 6.2.2 holds w.h.p. with respect to the observed sample.)

Next, we study the *full-rank stationary points* of the risk functions and the performance of the *gradient descent* algorithm. We first establish that when W^* is full rank, any full-rank stationary point W of the risk functions is necessarily global minimum, and that any such W is of form $W = QW^*$ where $Q \in \mathbb{R}^{m\times m}$ is orthonormal. See Theorem 6.2.3 for the population; and Theorem 6.2.4 for the empirical versions. Namely, W is a global optimum up to a rotation. We then establish that all "approximate" stationary points (appropriately defined) W of $\widehat{\mathcal{L}}(\cdot)$ below the aforementioned "energy barrier" are "nearly" global optimum. Furthermore, we establish that if the number N of samples is poly(d), then the weights W of any full-rank "approximate" stationary point are uniformly close to W^*. As a corollary, gradient descent with initialization below the "energy barrier" recovers in time poly (ϵ^{-1}, d) a solution W for which the weights are "ϵ-close" to the planted weights. Consequently, the generalization error $\mathcal{L}(W)$ for this solution W is at most ϵ. The bound on $\mathcal{L}(W)$ is derived by controlling the condition number of a certain matrix whose i.i.d. rows consists of tensorized data $X_i^{\otimes 2}$; using a recently developed machinery in [107] studying the spectrum of expected covariance matrices of tensorized data. See Theorem 6.2.5 for the population; and Theorem 6.2.6 for the empirical version.

Subsequently, we study the question whether one can find the initialization of the gradient descent algorithm below the aforementioned energy barrier. We answer affirmatively this question in the context of randomly generated $W^* \in \mathbb{R}^{m\times d}$, and establish in Theorem 6.2.8 that as long as the network is sufficiently overparameterized, specifically $m > Cd^2$, for some sufficiently large constant C, it is possible to initialize W_0 such that w.h.p. the risk associated to W_0 is below the required threshold. This is achieved by using tools from random matrix theory, specifically a semicircle law for Wishart matrices which shows the spectrum of $(W^*)^T W^*$ is tightly concentrated [25]. See Theorem 6.2.7 for the population; and Theorem 6.2.8 for the empirical version. It is worth noting that neural networks with random weights is an active area of research by itself due to the relationship with random feature methods. For example, Rahimi and Recht showed in [243] that shallow architectures trained by choosing the internal weights randomly and optimizing only over the output weights return a classifier with reasonable generalization performance at accelerated training speed. Random shallow networks were also shown to well-approximate dynamical systems [154]; have been successfully employed in the context of extreme learning machines [173]; and were studied in the context of random matrix theory, see [233] and references therein.

Our next focus is on the *sample complexity* for generalization. While we study the landscape of the empirical risk, it is not by any means certain that any (potentially not full-rank) optimizer of $\min_W \widehat{\mathcal{L}}(W)$ also achieve zero generalization error. We give necessary and sufficient conditions on the samples $X_i, 1 \leq i \leq N$ so that

any minimizer has indeed zero generalization error in our setting. We show that, if $\text{span}(X_i X_i^T : 1 \leq i \leq N)$ is the space of all $d \times d$-dimensional real symmetric matrices, then any global minimum of the empirical risk is necessarily a global optimizer of the population risk, and thus, has zero generalization error. Note that, this geometric condition is not *retrospective* in manner: it can be checked ahead of the optimization procedure by computing $\text{span}(X_i X_i^T : 1 \leq i \leq N)$. Conversely, we show that if the span condition above is not met then there exists a global minimum W of the empirical risk function which induces a strictly positive generalization error. This is established in Theorem 6.2.9.

To complement our analysis, we then ask the following question: what is the "*critical number*" N^* of the training samples, under which the (random) data $X_i, 1 \leq i \leq N$ enjoys the aforementioned span condition? We prove this number to be $N^* = d(d+1)/2$, under a very mild assumption that the coordinates of $X_i \in \mathbb{R}^d$ are jointly continuous. This is shown in Theorem 6.2.10. Finally in Theorem 6.2.11, we show that when $N < N^*$, not only that there exists W with zero empirical risk and strictly positive generalization error, we also bound this error from below by an amount very similar to the bound for rank-deficient matrices discussed in our earlier Theorem 6.2.2.

We end with a comment on overparameterization and generalization. A common paradigm in statistical learning theory is that, overparameterized models, that is, models with more parameters than necessary, while being capable of interpolating the training data, tend to generalize poorly because of overfitting to the proposed model. Yet, it has been observed empirically that neural networks tend to not suffer from this complication [294]: despite being overparameterized, they seem to have a good generalization performance, provided the interpolation barrier is exceeded. In Theorem 6.2.9 (a) we establish the following result which sheds some light on this phenomenon for the case of shallow neural networks with quadratic activations: suppose that the data enjoys the aforementioned geometric condition. Then, any interpolator achieves zero generalization error, even when the interpolator is a neural network with a potentially larger number $\widehat{m}$ of internal nodes compared to the one that generated the data, namely by using a weight matrix $W \in \mathbb{R}^{\widehat{m} \times d}$ where $\widehat{m} \geq m$. In other words, the model does not overfit when a much larger width of the interpolator is chosen at the learning state.

Comparison with [264] and [98]. We now make a comparison with two very related prior work, also studying the quadratic activations. We start with the work by Soltanolkotabi, Javanmard and Lee [264]. In [264, Theorem 2.2], the authors study the empirical risk landscape of a slightly more general version of our model: $Y_i = \sum_{j=1}^m v_j^* \langle W_j^*, X_i \rangle^2$, assuming $\text{rank}(W^*) = d$ like us, and assuming all non-zero entries of v^* have the same sign. Thus our model is the special case where all entries of v^* are equal to unity. The authors establish that as long as $d \leq N \leq cd^2$ for some small fixed constant c, every local minima of the empirical risk function is also a global minima (namely, there exists no spurious local minima), and furthermore, every saddle point has a direction of negative curvature. As a result they show that gradient

descent with an arbitrary initialization converges to a globally optimum solution of the ERM problem (6.2). In particular, their result does not require the initialization point to be below some risk value (the energy barrier), like in our case. Nevertheless, our results show that one needs not to worry about saddle points below the energy barrier as none exists per our Theorem 6.2.2. Importantly, though, the regime $N < cd^2$ for small c that [264, Theorem 2.2] applies is below the *provable sample complexity value* $N^* = d(d+1)/2$ when the data are drawn from a continuous distribution as per our Theorem 6.2.10. In particular, as we establish when $N < N^*$, the ERM problem (6.2) admits global optimum solutions with zero empirical risk value, but with generalization error bounded away from zero. Thus, the regime $N < N^*$ does not correspond to the regime where solving the ERM has a guaranteed control on the generalization error. The same theorem in [264] also studies the approximate stationary points, and shows that for any such point W, the associated empirical risk, $\widehat{\mathcal{L}}(W)$, is also small. Our Theorem 6.2.6, though, takes a step further and shows that not only the empirical risk is small but the recovered W is close to planted weights W^*; and therefore it has small generalization error $\mathcal{L}(W)$, by explicitly bounding the generalization error from above.

It is also worth noting that albeit not being our focus in the present work, [264, Theorem 2.1] also studies the landscape of the empirical risk when a quadratic network model $X \mapsto \sum_{j=1}^{m} v_j^* \left\langle W_j^*, X \right\rangle^2$ is used for interpolating arbitrary input/label pairs $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $1 \leq i \leq N$, that is, without making an assumption that the labels are generated according to a network with planted weights. They establish similar landscape results; namely, the absence of spurious local minima, and the fact that every saddle point has a direction of negative curvature, as long as the output weights v^* has at least d positive and at least d negative entries (consequently, the width m has to be at least $2d$). While this result does not assume any rank condition on W the assumption on the minimum number of positive and minimum number of negative entries such as the one above is somewhat unnatural.

Yet another closely related work studying quadratic activations is the paper by Du and Lee [98], which focuses on the shallow architectures with all unity output weights as we do. This paper establishes that for any smooth and convex loss $\ell(\cdot, \cdot)$, the landscape of the regularized loss function $\frac{1}{N}\sum_{i=1}^{N} \ell(f(W; X_i), Y_i) + \frac{\lambda}{2}\|W\|_F^2$ still admits aforementioned favorable geometric characteristics. Furthermore, since the learned weights are of bounded Frobenius norm due to the norm penalty $\|W\|_F^2$ imposed on objective, they retain good generalization via Rademacher complexity considerations. While this work addresses the training and generalization error when the norm of W is controlled during training; it does not carry out approximate stationarity analysis like Soltanolkotabi et al. [264] and we do; and does not study their associated loss/-generalization like in our case. Even though they show that optimal solutions to the optimization problem incorporating bounded norms generalize well; it remains unclear from their analysis whether the approximate stationary points of this objective also have a well-controlled norm.

It is worth mentioning that the two main directions that we undertake in this work were not explored in neither of these two prior work. These include the di-

rection pertaining to the initialization (Theorems 6.2.7 and 6.2.8); and the direction pertaining to the sample complexity (Theorems 6.2.9-6.2.11). The latter direction relates to an interesting interpolation/overparameterization property which we have discussed before. We will return later to this direction in Section 6.2.3 after we present Theorem 6.2.9.

Further relevant prior work. As noted in the introduction, neural networks achieved remarkable empirical success which fueled research starting from the expressive ability of these networks, going as early as Barron [39]. More recent works along this front focused on deeper and sparser models, see e.g. [218, 275, 105, 252, 235, 66]. In particular, the expressive power of such network architectures is relatively well-understood. Another issue pertaining such architectures is their computational tractability: Blum and Rivest established in [49] that it is NP-complete to train a very simple, 3-node, network; whose nodes compute a linear thresholding function. Despite this worst-case result, it has been observed empirically that local search algorithms such as gradient descent (GD), are rather successful in training. While several authors, including [253, 176, 149], devised provable training algorithms for such networks; these algorithms unfortunately are based on methods other than the gradient descent; thus not shedding any light on its apparent empirical success.

On a parallel front, many papers studied the behaviour of the GD by analyzing the trajectory of it or its stochastic variant (SGD), under certain stylistic assumptions on the data as well as the network. These assumptions include Gaussian inputs, shallow networks (with or without the convolutional structure) and the existence of planted weights (the so-called teacher network) generating the labels. Some partial and certainly very incomplete references to this end include [276, 64, 65, 296, 263, 205, 101]. Later work relaxed the distributional assumptions. For instance, [100] studied the problem of learning a convolutional unit with ReLU with no specific distributional assumption on input, and established the convergence of SGD with rate depending on the smoothness of the input distribution and the closeness of the patches. Several other works along this line, in particular under the presence of overparameterization, are the works by Du et al. [99, 102].

Yet another line of research on the optimization front, rather than analyzing the trajectory of the GD, focuses on the mean-field analysis: empirical distribution of the parameters of network with infinitely many internal nodes can be described as a Wasserstein gradient flow, thus some tools from the theory of optimal transport can be used, see e.g. [284, 247, 74, 265, 258]. Albeit explaining the story to some extent for infinitely wide networks, it remains unclear whether these techniques provide results for a more realistic network model with finitely many internal nodes.

As noted earlier, the optimization landscape of such networks is usually highly non-convex. More recent research on such non-convex objectives showed that if the landscape has certain favorable geometric properties such as the absence of spurious local minima and the existence of direction with negative curvature for every saddle point, local methods can escape the saddle points and converge to the global minima. Examples of this line of research on loss functions include [142, 202, 197, 179, 97].

Motivated by this front of research, many papers analyzed geometric properties of the optimization landscape, including [237, 157, 75, 158, 185, 159, 266, 114, 298, 226, 143, 249, 267, 299, 227, 279, 98, 264].

We now touch upon yet another very important focus, that is the *generalization ability* of such networks: how well a solution found, e.g. by GD, predicts an unseen data? A common paradigm in statistical learning theory that was mentioned previously is that overparameterized models tend to generalize poorly. Yet, neural networks tend to not suffer from this complication [294]. Since the VC-dimension of these networks grow (at least) linear in the number of parameters [162, 42], standard Vapnik-Chervonenkis theory do not help explaining the good generalization ability under presence of overparameterization. This has been studied, among others, through the lens of the norm of weight matrices [225, 41, 207, 153, 104, 290]; PAC-Bayes theory [224, 223], and compression-based bounds [20]. A main drawback is that these papers require some sort of constraints on the weights and are mostly *a posteriori*: whether or not a good generalization takes place can be determined only when the training process is finished. A recent work by Arora et al. [18] provided an *a priori* guarantee for the solution found by the GD. Our result regarding the generalization guarantee described in Theorem 6.2.11 also provides simple a priori guarantee on the generalization.

A Follow-up Work. After our work appeared on arXiv, a follow-up work was done by Mannelli, Vanden-Eijnden, and Zdeborová [251]. In this paper, the authors consider the same architecture (namely, a shallow network with quadratic activations) under the so-called teacher/student setting; and study the landscape of the empirical risk as well as the performance of the gradient flow, the continuous-time analogue of the gradient descent. Importantly, they consider also the regime where the number m^* of the hidden units of the teacher network is less than the dimension d (whereas our focus is on $m^* \geq d$). In particular when $m^* = 1$, the width m of the student network is at least d, and the data consists of i.i.d. standard normal entries; they prove the following. In the limit as $d \to \infty$, if $n > 2d$ then with positive probability the only minimizer of the empirical risk is the matrix A^* of teacher weights itself; whereas for $n < 2d$, the empirical risk admits spurious minima with probability tending to one. (Namely, the geometry of the empirical risk undergoes a phase transition as $\alpha \triangleq n/d$ crosses $\alpha_c = 2$). Moreover, they also prove that for $m \geq d$, the gradient flow converges to a global minima of the empirical risk and to the global minimum of the population risk (which is A^*); and characterize the rate of convergence for the latter case. (It is worth noting that running gradient flow on the population risk can be perceived as running it on the empirical risk in the limit of the large number n of samples.)

Chapter organization. In Section 6.2.1 we present our main results on the landscape of the risk functions, including our energy barrier result for rank-deficient matrices, our result about the absence of full-rank stationary points of the risk function except the globally optimum points; and our result on the convergence of gradient

descent. In Section 6.2.2, we present our results regarding randomly generated weight matrices W^* and sufficient conditions for good initializations. In Section 6.2.3, we study the critical number of training samples guaranteeing good generalization property. We collect useful auxiliary lemmas in Section 6.3; and provide the proofs of all of our results in Section 6.4.

Notation. The set of reals, positive reals; and the set $\{1, 2, \ldots, k\}$ are denoted respectively by $\mathbb{R}$, $\mathbb{R}_+$, and $[k]$. For any matrix A, its smallest and largest singular values, spectrum, trace, Frobenius and the spectral norm are denoted respectively by $\sigma_{\min}(A)$, $\sigma_{\max}(A)$, $\sigma(A)$, $\mathrm{trace}(A)$, $\|A\|_F$, and $\|A\|_2$. I_n denotes the $n \times n$ identity matrix. Planted weights are denoted with an asterisk, e.g. W^*. $\exp(\alpha)$ denotes e^α. Given any $v \in \mathbb{R}^n$, $\|v\|_2$ denotes its Euclidean ℓ_2 norm $\sqrt{\sum_{1 \le i \le n} v_i^2}$. Given two vectors $x, y \in \mathbb{R}^n$, their Euclidean inner product $\sum_{1 \le i \le n} x_i y_i$ is denoted by $\langle x, y \rangle$. Given a collection $Z_1, \ldots, Z_k$ of objects of the same kind (e.g., vectors or matrices), $\mathrm{span}(Z_i : i \in [k])$ is the set, $\left\{ \sum_{j=1}^k \alpha_j Z_j : \alpha_j \in \mathbb{R} \right\}$. We say a random variable X is "centered" if $\mathbb{E}[X] = 0$. $\Theta(\cdot), O(\cdot)$, $o(\cdot)$, and $\Omega(\cdot)$ are standard (asymptotic) order notations for comparing the growth of two sequences. $\widehat{\mathcal{L}}(\cdot)$, $\nabla \widehat{\mathcal{L}}(\cdot)$, $\mathcal{L}$, and $\nabla \mathcal{L}$ denote respectively the empirical risk, its gradient; the population risk, and its gradient.

6.2 Main Results

Our main results are now in order.

6.2.1 Optimization Landscape

Existence of an Energy Barrier

Our first result shows the presence of an energy barrier in the landscape of the population risk $\mathcal{L}(\cdot)$ below which any rank-deficient $W \in \mathbb{R}^{m \times d}$ ceases to exist.

Theorem 6.2.1. *Suppose that $X \in \mathbb{R}^d$ has i.i.d. centered coordinates with variance μ_2, (finite) fourth moment μ_4, $\mathrm{rank}(W^*) = d$, and let $\mathcal{L}(W)$ be defined as* (6.3).

(a) (Lower bound) It holds that

$$\min_{W \in \mathbb{R}^{m \times d} : \mathrm{rank}(W) < d} \mathcal{L}(W) \ge \min\left\{ \mu_4 - \mu_2^2, 2\mu_2^2 \right\} \cdot \sigma_{\min}(W^*)^4.$$

(b) (Tightness) There exists a matrix $W \in \mathbb{R}^{m \times d}$ such that $\mathrm{rank}(W) \le d - 1$ and

$$\mathcal{L}(W) \le \max\left\{ \mu_4, 3\mu_2^2 \right\} \cdot \sigma_{\min}(W^*)^4.$$

The proof of Theorem 6.2.1 is deferred to Section 6.4.2. Two remarks are in order.

First, the hypothesis of Theorem 6.2.1 holds under mild distributional assumptions on the coordinates of data: a finite fourth moment and zero mean suffices.

Second, part (b) of Theorem 6.2.1 implies that our lower bound on the energy value is tight up to a multiplicative constant determined by the moments of the data. That is, there exists a W with $\mathrm{rank}(W) \leq d-1$ such that $\mathcal{L}(W) = \Theta(\sigma_{\min}(W^*)^4)$, where the asymptotic $\Theta(\cdot)$ hides the constants μ_2 and μ_4.

Our next result is an analogue of Theorem 6.2.1 for the empirical risk $\widehat{\mathcal{L}}(\cdot)$; and it establishes the presence of a similar energy barrier in the landscape of the empirical risk $\widehat{\mathcal{L}}(\cdot)$ below which any rank-deficient $W \in \mathbb{R}^{m\times d}$ ceases to exist, with high probability.

Theorem 6.2.2. *Let $K > 0$ be an arbitrary constant; and $X_i \in \mathbb{R}^d$, $1 \leq i \leq N$ be a collection of i.i.d. random vectors each having centered i.i.d. sub-Gaussian coordinates. That is, for some $C > 0$, $\mathbb{P}(|X(j)| > t) \leq \exp(-Ct^2)$ for every $t \geq 0$, $j \in [d]$. Suppose, furthermore, that for every $M > 0$, the distribution of $X(j)$, conditional on $|X(j)| \leq M$, is centered: $\mathbb{E}\Big[X(j)\Big||X(j)| \leq M\Big] = 0$. Let $Y_i = f(W^*; X_i)$, $1 \leq i \leq N$ be the corresponding label generated by a planted teacher network per (6.1), where $\mathrm{rank}(W^*) = d$ and $\|W^*\|_F \leq d^K$. Then, for some absolute constants $C_3, C' > 0$, with probability at least*

$$1 - \exp(-C'd) - \left(9d^{4K+9}\right)^{d^2-1} \cdot \exp\left(-C_3 N d^{-4-4K}\right) - Nde^{-Cd}$$

it holds that

$$\min_{W\in\mathbb{R}^{m\times d}:\mathrm{rank}(W)\leq d-1} \widehat{\mathcal{L}}(W) \triangleq \min_{W\in\mathbb{R}^{m\times d}:\mathrm{rank}(W)\leq d-1} \frac{1}{N}\sum_{1\leq i\leq N} (Y_i - f(W;X_i))^2 \geq \frac{1}{2}C_5\sigma_{\min}(W^*)^4.$$

Here,

$$C_5 = \min\Big\{\mu_4(1/2) - \mu_2(1/2)^2, 2\mu_2(1/2)^2\Big\}, \quad \textit{where} \quad \mu_t(1/2) = \mathbb{E}\Big[X(j)^t\Big||X(j)| \leq d^{1/2}\Big].$$

Furthermore, if the dimension d of data is constant ($d = O(1)$); then with probability $1 - O(1/N)$,

$$\min_{W\in\mathbb{R}^{m\times d}:\mathrm{rank}(W)\leq d-1} \widehat{\mathcal{L}}(W) \geq \frac{1}{2}\overline{C_5}\sigma_{\min}(W^*)^4,$$

where

$$\overline{C_5} = \min\Big\{\mu_4 - \mu_2^2, 2\mu_2^2\Big\} \quad \textit{and} \quad \mu_n = \mathbb{E}\Big[X(j)^n\Big].$$

The proof of Theorem 6.2.2 is provided in Section 6.4.9. Several remarks are now in order.

Assuming d is large, Theorem 6.2.2 shows that with high probability, $\widehat{\mathcal{L}}(W)$ is bounded away from zero by an explicit constant for any W that is rank-deficient, provided $N = d^{O(1)}$; where the $O(1)$ term depends on K. Furthermore, provided N is a sufficiently large polynomial-in-d quantity, the probability estimate is of form $1 - \exp(-\Theta(d))$ which is exponential in the dimension.

Note that one indeed needs a "finite d correction" for the case when the data is low-dimensional, $d = O(1)$: for $d = O(1)$, the term $Nd\exp(-\Theta(d))$ makes the

probability estimate vacuous. The constant $\overline{C_5}$ appearing in this case is precisely the same constant appearing in Theorem 6.2.1; and in particular no conditioning is required. Furthermore, while we establish the probability estimate to be $1 - O(1/N)$ for simplicity; it can be improved: it appears, from our analysis (which uses Chebyshev's inequality), that for any $\alpha > 0$, one can show the probability estimate to be $1 - O(N^{-\alpha})$. Furthermore, using more elaborate tools (such as concentration for heavy-tailed variables, in particular for i.i.d. averages of fourth moments of sub-Gaussian variables), this estimate can potentially be improved to $1 - \exp(-\bar{c_0}N^{c_0})$ for suitable constants $c_0, \bar{c_0} > 0$. Finally, our analysis yields also that the $1 - O(1/N)$ probability estimate still remains valid even when X_i has centered i.i.d coordinates with a finite eight moment. That is, the sub-Gaussianity assumption can be relaxed. Indeed, the expression $\widehat{\mathcal{L}}(W)$ is an i.i.d. sum of form $N^{-1}\sum_{1\leq i\leq N}(X_i^T M X_i)^2$ for a suitable matrix M, and one needs the finiteness of $\mathbb{E}[(X^T M X)^4]$ to apply Chebyshev's inequality: this quantity is finite provided $\mathbb{E}[X_i(j)^8] < \infty$ and $\|M\|_F = O(1)$. We do not pursue these extensions for keeping the presentation clear.

The assumption that the conditional mean of $X_i(j)$ is zero is benign: it holds, e.g., for random variables whose distributions are symmetric around zero. For instance, any normal distribution with zero-mean satisfies this assumption. Furthermore, an inspection of its proof reveals that Theorem 6.2.2 still remains valid even when the coordinates of the data have heavier tails: our techniques apply also to the tails of form $\mathbb{P}(|X_i(j)| > t) \leq \exp(-Ct^\alpha)$ where $C, \alpha > 0$ are arbitrary. We use $\alpha = 2$ throughout for simplicity.

An inspection of the proofs of Theorems 6.2.1 and 6.2.2 yield that the landscapes of the corresponding risks still admit an energy barrier, even if we consider the same network architecture with planted weight matrix $W^* \in \mathbb{R}^{m\times d}$, and quadratic activation function having lower order terms, that is, the activation $\alpha x^2 + \beta x + \gamma$, with $\alpha \neq 0$. In this case, in addition to $\sigma_{\min}(W^*)$ and the corresponding moments of the data; the coefficient α also appears in the barrier expression. In particular, Theorem 6.2.1 still remains valid with $\min\{\mu_4 - \mu_2^2, 2\mu_2^2\} \cdot \sigma_{\min}(W^*)^4$ replaced with $\alpha^2 \cdot \min\{\mu_4 - \mu_2^2, 2\mu_2^2\} \cdot \sigma_{\min}(W^*)^4$; and Theorem 6.2.2 still remains valid with $\frac{1}{2}C_5\sigma_{\min}(W^*)^4$ replaced with $\frac{\alpha^2}{2}C_5\sigma_{\min}(W^*)^4$.

Global Optimality of Full-Rank Stationary Points

We now establish that if W is a full-rank stationary point of the population risk, $\mathcal{L}(\cdot)$, then W is necessarily a global minimum.

Theorem 6.2.3. *Suppose $W^* \in \mathbb{R}^{m\times d}$ with $\operatorname{rank}(W^*) = d$. Suppose $X \in \mathbb{R}^d$ has centered i.i.d. coordinates with $\mathbb{E}[X_i^2] = \mu_2$, $\mathbb{E}[X_i^4] = \mu_4$; and $\operatorname{Var}(X_i^2) > 0$. Let $W \in \mathbb{R}^{m\times d}$ be a stationary point of the population risk with full-rank, that is, $\nabla\mathcal{L}(W) = \mathbb{E}[\nabla(f(W^*;X) - f(W;X))^2] = 0$, and $\operatorname{rank}(W) = d$. Then, $W = QW^*$ for some orthogonal matrix Q, and that, $\mathcal{L}(W) = 0$.*

The proof of Theorem 6.2.3 is deferred to Section 6.4.6.

Our next result is an analogue of Theorem 6.2.3 for the empirical risk, $\widehat{\mathcal{L}}(\cdot)$, and shows that if $N \geq d(d+1)/2$ and W is any full-rank stationary point of the empirical

risk then W is necessarily a global minimum.

Theorem 6.2.4. *Let $X_i \in \mathbb{R}^d$, $1 \le i \le N$; $W^* \in \mathbb{R}^{m\times d}$ with $\operatorname{rank}(W^*) = d$, and suppose W is a full-rank stationary point of the empirical risk: $\operatorname{rank}(W) = d$, and $\nabla_W \widehat{\mathcal{L}}(W) = 0$. Then, $\widehat{\mathcal{L}}(W) = 0$. Furthermore; if X_i are i.i.d. random vectors having a jointly continuous distribution and $N \ge d(d+1)/2$, then with probability one $W = QW^*$ for some orthogonal matrix $Q \in \mathbb{R}^{m\times m}$.*

The proof of Theorem 6.2.4 is given in Section 6.4.10.

Note that an implication of Theorems 6.2.3 and 6.2.4 is that the corresponding losses admit no full-rank saddle points. Namely, the landscape of the corresponding losses have fairly benign properties below the aforementioned energy barrier. We soon show how this implies the convergence of gradient descent in the next section.

Convergence of Gradient Descent

We now combine Theorems 6.2.1 and 6.2.3 to obtain the following conclusion on the performance of the gradient descent algorithm for the population risk. Suppose that the algorithm is initialized at a point W_0 having a small population risk $\mathcal{L}(W_0)$, in particular lower than the smallest risk value achieved by the rank-deficient matrices. Then with a properly chosen step size, the algorithm converges to a global optimum: it generates a trajectory $\{W_k\}_{k\ge 0}$ of weights such that $\lim_{k\to\infty} \mathcal{L}(W_k) = \min_W \mathcal{L}(W) = 0$.

Theorem 6.2.5. *Let $W_0 \in \mathbb{R}^{m\times d}$ be a matrix of weights, with the property that*

$$\mathcal{L}(W_0) < \min_{W\in\mathbb{R}^{m\times d}:\operatorname{rank}(W)<d} \mathcal{L}(W).$$

Define

$$L = \sup\left\{\|\nabla^2 \mathcal{L}(W)\| : \mathcal{L}(W) \le \mathcal{L}(W_0)\right\},$$

where by $\|\nabla^2\mathcal{L}(W)\|$ we denote the spectral norm of the matrix $\nabla^2\mathcal{L}(W)$. Then, $L < \infty$ and the gradient descent algorithm with initialization $W_0 \in \mathbb{R}^{m\times d}$ and any fixed step size of $0 < \eta < 1/2L$ generates a trajectory $\{W_k\}_{k\ge 0}$ of weights such that $\lim_{k\to\infty} \mathcal{L}(W_k) = \min_W \mathcal{L}(W) = 0$.

The proof of Theorem 6.2.5 is provided in Section 6.4.7.

Our next focus is on the performance of the gradient descent algorithm for the empirical risk. By combining Theorems 6.2.2 and 6.2.4 we obtain the following conclusions. Suppose that the gradient descent algorithm is initialized at a point with a sufficiently small empirical risk, in particular lower than the smallest risk value achieved by rank-deficient matrices (i.e. the energy barrier); and fix an $\epsilon > 0$. Then, with a properly chosen step size; it finds an approximate stationary point W (that is, a $W \in \mathbb{R}^{m\times d}$ with a small $\|\nabla\widehat{\mathcal{L}}(W)\|_F$) in time poly $(\epsilon^{-1}, \sigma_{\min}(W^*)^{-1}, d)$ for which the weights W^TW are uniformly "ϵ-close" to the planted weights $(W^*)^TW^*$, and consequently the generalization error $\mathcal{L}(W)$ is at most ϵ. Furthermore, the algorithm converges to a global optimum of the empirical risk minimization problem

$\min_W \widehat{\mathcal{L}}(W)$, which is zero; thus recovering planted weights, due to the absence of spurious stationary points within the set of full-rank matrices.

Theorem 6.2.6. *Let $\epsilon, K > 0$ be arbitrary. Suppose that $X_i \in \mathbb{R}^d$, $1 \le i \le N$ satisfies the assumptions in Theorem 6.2.2; $W_0 \in \mathbb{R}^{m\times d}$ is a matrix of weights with the property*

$$\widehat{\mathcal{L}}(W_0) < \frac{1}{2} C_5 \sigma_{\min}(W^*)^4,$$

where C_5 is the constant defined in Theorem 6.2.2; and $\|W^*\|_F \le d^K$. Define

$$L = L(W_0) \triangleq \sup\left\{\|\nabla^2 \widehat{\mathcal{L}}(W)\| : \widehat{\mathcal{L}}(W) \le \widehat{\mathcal{L}}(W_0)\right\}$$

where by $\|\nabla^2 \widehat{\mathcal{L}}(W)\|$ we denote the spectral norm of the (Hessian) matrix $\nabla^2 \widehat{\mathcal{L}}(W)$; and

$$\zeta \triangleq \min\left\{\frac{\epsilon \cdot \sigma_{\min}(W^*)^2}{32 d^{4K+4}}, \frac{\epsilon^2 \cdot \sigma_{\min}(W^*)^2}{(C')^2 d^{4K+15}}, \frac{\epsilon \cdot \sigma_{\min}(W^*)}{2(C')^2 \mu_2^2 d^{4K+16}}\right\}, \tag{6.4}$$

where $C' > 0$ is some absolute constant, and $\mu_2 = \mathbb{E}[X(j)^2]$. Then, with probability at least

$$1 - \exp\left(-c' N^{1/4}\right) - \left(9 d^{4K+9}\right)^{d^2-1} \cdot \exp\left(-C_4 N d^{-4-4K}\right) - N d \exp(-C d),$$

(where $c', C, C_4 > 0$ are also absolute constants)

(a) For any W with $\widehat{\mathcal{L}}(W) < \frac{1}{2} C_5 \sigma_{\min}(W^*)^4$, $\|W|_F \le d^{K+1}$. Moreover, $L = \mathrm{poly}(d) < \infty$.

(b) Running gradient descent algorithm starting from W_0 with a step size of $0 < \eta < 1/2L$ generates a full-rank $W \in \mathbb{R}^{m\times d}$ with $\|\nabla \widehat{\mathcal{L}}(W)\|_F \le \zeta$ in time $\mathrm{poly}\left(\epsilon^{-1}, \sigma_{\min}(W^*)^{-1}, d\right)$. Furthermore, for this W, $\widehat{\mathcal{L}}(W) \le \epsilon$.

(c) For W in (b)*, it holds that $\|W^T W - (W^*)^T W^*\|_F \le \epsilon$; and the generalization error $\mathcal{L}(W)$ is at most ϵ, provided $N \ge d^{58/3}$.*

Furthermore, suppose the data dimension d is constant, $d = O(1)$, and $W_0 \in \mathbb{R}^{m\times d}$ is a matrix of weights with the property $\widehat{\mathcal{L}}(W_0) < \frac{1}{2}\overline{C_5}\sigma_{\min}(W^*)^4$ for the same constant $\overline{C_5}$ appearing in Theorem 6.2.2. Then, for ζ chosen per (6.4)*, with probability at least $1 - O(1/N)$,*

(a') For any W with $\widehat{\mathcal{L}}(W) \le \frac{1}{2}\overline{C_5}\sigma_{\min}(W^*)^4$ it is the case $\|W\|_F = O(1)$. Moreover, $L = O(1) < \infty$.

(b') Running gradient descent algorithm starting from W_0 with a step size of $0 < \eta < 1/2L$ generates a full-rank W with $\|\nabla \widehat{\mathcal{L}}(W)\| \le \zeta$ and $\widehat{\mathcal{L}}(W) \le \epsilon$;

(c') For any W in (b)*, $\max\left\{\left\|W^T W - (W^*)^T W^*\right\|_F, \mathcal{L}(W)\right\} \le \epsilon$.*

The proof of Theorem 6.2.6 is provided in Section 6.4.11. Several remarks are now in order.

As a simple corollary to (b), we thus obtain that for d large enough, the gradient descent algorithm with initialization $W_0 \in \mathbb{R}^{m\times d}$ and a step size of $0 < \eta < 1/2L$ generates a trajectory $\{W_k\}_{k\geq 0}$ of weights such that $\lim_{k\to\infty} \widehat{\mathcal{L}}(W_k) = \min_W \widehat{\mathcal{L}}(W) = 0$. This yields an analogue of Theorem 6.2.5 for the empirical risk, $\widehat{\mathcal{L}}(\cdot)$.

Note that, when N is a sufficiently large, polynomial-in-d function, the probability estimate for the first part is essentially $1 - \exp(-\Theta(d))$. Furthermore, we note that the exponent $1/4$ in the probability estimate and the sample bound $d^{58/3}$ are required only for part (c), and can potentially be improved. In particular, the exponent can be improved to one for parts (a),(b) and (d).

Note again that analogous to Theorem 6.2.2, one needs a correction for "finite d case" since the term $Nd\exp(-\Theta(d))$ makes the probability estimate vacuous for $d = O(1)$. Furthermore, the remarks following Theorem 6.2.2 still remain valid. In particular, the choice of $1 - O(1/N)$ is for simplicity; and the estimate can be improved, almost immediately, to $1 - O(N^{-\alpha})$ for any $\alpha > 0$; and to $1 - \exp(-\bar{c_0}N^{c_0})$ for some $c_0, \bar{c_0} > 0$ using more delicate concentration tools.

We now provide an important remark pertaining (c): as we show in the proof, provided N grows at least polynomially in d; with probability $1 - \exp(-C'N^{1/4})$ over X_i it holds that for any W having a small risk, $\widehat{\mathcal{L}}(W)$; W^TW is close to $(W^*)^TW^*$: $\|W^TW - (W^*)^TW^*\|_F$ is small. Consequently $\mathcal{L}(W)$ is small. This is one of the additional technical results of our work; and is achieved by controlling the condition number of a certain matrix whose i.i.d. rows consist of the tensorized data $X_i^{\otimes 2}$. The proof uses a recent work analyzing the spectrum of expected covariance matrices of tensorized data [107].

The above results concern the performance of gradient descent assuming the initialization is proper, i.e. it is below the aforementioned energy barrier. One can then naturally ask whether such an initialization is indeed possible in some generic context. In the next section, we address this question of proper initialization when the (planted) weights are generated randomly, in order to complement Theorems 6.2.5 and 6.2.6. We establish that such a proper initialization is indeed possible by providing a deterministic initialization guarantee, which with high probability beats the aforementioned energy barrier.

6.2.2 On Initialization: Randomly Generated Planted Weights

Our results in the previous section showed that provided the initialization of the gradient descent method occurs below the critical energy, the algorithm converges to the global minimum. This raises the question whether such an initialization can be found in a constructive way.

In this section, we show that the answer is yes in the setting of randomly generated weights of the ground truth matrix W^*. Specifically, we provide a way to properly initialize such networks under the assumption that the (planted) weight matrix $W^* \in \mathbb{R}^{m\times d}$ has arbitrary i.i.d. centered entries with unit variance and finite fourth moment;

and the data has centered i.i.d. sub-Gaussian coordinates. (It is worth mentioning that similar to before, the sub-Gaussianity assumption on the data is required only for the case of empirical risk and the corresponding population risk result holds under a milder distributional assumption, see Theorem 6.2.7 below.) Our result is valid provided that the network is sufficiently overparameterized: $m > Cd^2$ for some large constant C. Note that this implies W^* is a tall matrix sending $\mathbb{R}^d$ into $\mathbb{R}^m$. The rationale behind this approach is as follows. The value of the risk is determined by the spectrum of $\Delta \triangleq W^T W - (W^*)^T W^*$ and the moments of the data distribution. Under our randomness assumption, the so-called Wishart matrix $(W^*)^T W^*$ is tightly concentrated around a multiple of the identity if m is sufficiently large. Hence one can control the spectrum of Δ, and therefore the loss functions ($\mathcal{L}$ and $\widehat{\mathcal{L}}(\cdot)$) by properly choosing the initialization W.

We now state the main results of this section, starting with the population risk version.

Theorem 6.2.7. *Suppose that the data $X \in \mathbb{R}^d$ consists of i.i.d. centered coordinates with $\mathrm{Var}(X_i^2) > 0$ and $\mathbb{E}[X_i^4] < \infty$. Recall $\mathcal{L}(W)$ from (6.3).*

(a) (Gaussian case) Suppose that the planted weight matrix $W^* \in \mathbb{R}^{m \times d}$ has i.i.d. standard normal entries. Let the initial weight matrix $W_0 \in \mathbb{R}^{m \times d}$ be defined by $(W_0)_{i,i} = \sqrt{m + 4d}$ for $1 \le i \le d$, and $(W_0)_{i,j} = 0$ otherwise, that is $W_0^T W_0 = \gamma I_d$ with $\gamma = m + 4d$. Then, provided $m > Cd^2$ for a sufficiently large absolute constant $C > 0$,

$$\mathcal{L}(W_0) < \min_{W \in \mathbb{R}^{m \times d} : \mathrm{rank}(W) < d} \mathcal{L}(W),$$

with probability at least $1 - \exp(-\Omega(d))$, where the probability is with respect to the draw of W^*.

(b) (General case) Suppose the planted weight matrix $W^* \in \mathbb{R}^{m \times d}$ has centered i.i.d. entries with unit variance and finite fourth moment. Let the initial weight matrix $W_0 \in \mathbb{R}^{m \times d}$ be defined by $(W_0)_{i,i} = \sqrt{m}$ for $1 \le i \le d$, and $(W_0)_{i,j} = 0$ otherwise, that is $W_0^T W_0 = m I_d$. Then, provided $m > Cd^2$ for a sufficiently large absolute constant $C > 0$,

$$\mathcal{L}(W_0) < \min_{W \in \mathbb{R}^{m \times d} : \mathrm{rank}(W) < d} \mathcal{L}(W),$$

with high probability, as $d \to \infty$, where the probability is with respect to the draw of W^*.

The proof of this theorem is provided in Section 6.4.8.

Note that, the part (a) of Theorem 6.2.7 gives an explicit rate for probability, in the case when the i.i.d. entries of the planted weight matrix W^* are standard normal, and is based on a non-asymptotic concentration result for the spectrum of such matrices. The extension in part (b) is based on a semicircle law obtained by Bai and Yin [25].

The corresponding result for the empirical risk is provided below.

Theorem 6.2.8. *Suppose that the planted weight matrix $W^* \in \mathbb{R}^{m\times d}$ has centered i.i.d. entries with unit variance and finite fourth moment; the (i.i.d.) data $X_i \in \mathbb{R}^d$, $1 \le i \le N$, has i.i.d. centered sub-Gaussian coordinates (namely for some $C > 0$, $\mathbb{P}(|X_i(j)| > t) \le \exp(-Ct^2)$ for any $t > 0$, $1 \le i \le N$ and $1 \le j \le d$); and the $W_0 \in \mathbb{R}^{m\times d}$ satisfies $(W_0)_{ii} = \sqrt{m}$ for $i \in [d]$ and $(W_0)_{ij} = 0$ for $i \ne j$, that is $W_0^T W_0 = mI_d \in \mathbb{R}^{d\times d}$. Then for some absolute constants $C, C' > 0$ with probability at least*

$$1 - \exp\left(-C'\frac{N}{d^5 m}\right) - Nd\exp(-Cd) - o_d(1),$$

it is the case that for the constant C_5 defined in Theorem 6.2.2,

$$\widehat{\mathcal{L}}(W_0) < \frac{1}{2}C_5\sigma_{\min}(W^*)^4$$

provided $m > C''d^2$ for a sufficiently large constant $C'' > 0$.

The proof of Theorem 6.2.8 is provided in Section 6.4.12.

It is worth noting that unlike earlier Theorems 6.2.2 and 6.2.6; Theorems 6.2.7 and 6.2.8 do not have a separate statement for the case of finite d ($d = O(1)$): in order to ensure the concentration property for the Wishart ensemble takes place, one should consider the regime $d \to \infty$. That is, our initialization results do not hold for the regime where d is constant.

With this, we now turn our attention to the number of training samples required to learn such models.

6.2.3 Critical Number of Training Samples

The focus of previous sections is on landscape results pertaining the empirical risk minimization problem. One can then naturally ask the following question: what is the smallest number of samples required to claim that a small empirical risk "controls" also the generalization error?

In this section, we focus on this condition; namely our focus is on the number of training samples required for bounding the generalization error. We identify a necessary and sufficient condition on the training data under which any minimizer of the empirical risk (which, in the case we consider of planted weights, simply interpolates the data) has zero generalization error. We obtain our results for potentially overparameterized interpolators, that is for networks of potentially larger width than the width of the original network generating the labels. Furthermore we identify the smallest number N^* of training samples, such that (randomly generated) training data $X_1, \dots, X_N$ satisfies the aforementioned condition, so long as $N \ge N^*$.

A Necessary and Sufficient Geometric Condition on the Training Data

We start by providing a necessary and sufficient (geometric) condition on the training data under which any minimizer of the empirical risk (which, in the case of planted weights, necessarily interpolates the data) has zero generalization error.

Theorem 6.2.9. *Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be a set of data.*

(a) Suppose

$$\operatorname{span}\{X_i X_i^T : 1 \leq i \leq N\} = \mathcal{S},$$

where $\mathcal{S}$ is the set of all $d \times d$ symmetric real-valued matrices. Let $\widehat{m} \in \mathbb{N}$ be arbitrary. Then for any $W \in \mathbb{R}^{\widehat{m} \times d}$ interpolating the data, that is satisfying $f(W^*; X_i) = f(W; X_i)$ for every $i \in [N]$, it holds that $W^T W = (W^*)^T W^*$. In particular, if $\widehat{m} \geq m$, then for some matrix $Q \in \mathbb{R}^{\widehat{m} \times m}$ with orthonormal columns (i.e. $Q^T Q = I_m$), $W = QW^*$, and if $m \geq \widehat{m}$, then for some matrix $Q' \in \mathbb{R}^{m \times \widehat{m}}$ with orthonormal columns (i.e. $(Q')^T Q' = I_{\widehat{m}}$), $W^* = Q'W$.

(b) Suppose,

$$\operatorname{span}\{X_i X_i^T : 1 \leq i \leq N\},$$

is a strict subset of $\mathcal{S}$. Then, for any $W^* \in \mathbb{R}^{m \times d}$ with $\operatorname{rank}(W^*) = d$ and any positive integer $\widehat{m} \geq d$, there exists a $W \in \mathbb{R}^{\widehat{m} \times d}$ such that $W^T W \neq (W^*)^T W^*$, while W interpolates the data, that is, $f(W^*; X_i) = f(W; X_i)$ for all $i \in [N]$. In particular, for this $W \in \mathbb{R}^{m \times d}$, $\mathcal{L}(W) > 0$, where $\mathcal{L}$ is defined with respect to any jointly continuous distribution on $\mathbb{R}^d$.

The proof of Theorem 6.2.9 is deferred to Section 6.4.13.

Several remarks are now in order. The condition stated in Theorem 6.2.9 is not retrospective in manner: it can be checked ahead of the optimization process. Next, there are no randomness assumptions in the setting of Theorem 6.2.9 (except the last part regarding the population risk), and it provides a purely geometric necessary and sufficient condition: as long as $\operatorname{span}(X_i X_i^T : i \in [N])$ is the space of all symmetric matrices (in $\mathbb{R}^{d \times d}$) we have that any (global) minimizer of the empirical risk has zero generalization error. Conversely, in the absence of this geometric condition, there are optimizers $W \in \mathbb{R}^{m \times d}$ of the empirical risk $\widehat{\mathcal{L}}(\cdot)$ such that while $\widehat{\mathcal{L}}(W) = 0$, the generalization error of W is bounded away from zero, and $W^T W \neq (W^*)^T W^*$. It is also worth recalling that in the case when W does not interpolate the data but has a rather small training error, the result of Theorem 6.2.6(c) allows one to control $\|W^T W - (W^*)^T W^*\|_F$, and consequently the generalization error $\mathcal{L}(W)$. Soon in Theorem 6.2.11, we give a more refined version of Theorem 6.2.9, with a concrete lower bound on $\mathcal{L}(W)$, in the setting where the training data is generated randomly.

We further highlight the presence of the parameter $\widehat{m} \in \mathbb{N}$. In particular, part (a) of Theorem 6.2.9 states that provided the span condition is satisfied, any neural network with $\widehat{m}$ internal nodes interpolating the data has necessarily zero generalization error, regardless of whether $\widehat{m}$ is equal to m, in particular, even when $\widehat{m} \geq m$. This, in fact, is an instance of an interesting phenomenon empirically observed about neural networks, which somewhat challenges one of the main paradigms in statistical learning theory: overparameterizartion does not hurt generalization performance of neural networks once the data is interpolated. Namely beyond the interpolation threshold, one retains good generalization property.

It is worth noting that Theorem 6.2.9 still remains valid under a slightly more general setup where each node $j \in [m]$ has an associated positive but otherwise arbitrary output weight $a_j^* \in \mathbb{R}_+$. That is, the network computes, instead, the function

$\sum_{1\le j\le m} a_j^* \left\langle W_j^*, X\right\rangle^2$ with $a_j^* > 0$. Indeed, in this case, the (output) weights a_j^* can be "pushed" inside the matrix W_j^*: one can set $\widehat{W}_j \triangleq \sqrt{a_j^*} W_j^* \in \mathbb{R}^d$ and observe that for any $X \in \mathbb{R}^d$, $\sum_{1\le j\le m} a_j^* \left\langle W_j^*, X\right\rangle^2 = \sum_{1\le j\le m} \left\langle \widehat{W}_j, X\right\rangle^2$. Considering $\widehat{W} \in \mathbb{R}^{m\times d}$ instead, we are indeed under the setting of Theorem 6.2.9.

Randomized Data Enjoys the Geometric Condition

We now identify the smallest number N^* of training samples, such that (randomly generated) training data $X_1, \dots, X_N$ satisfies the aforementioned geometric condition almost surely; as soon as $N \ge N^*$.

Theorem 6.2.10. *Let $N^* = d(d+1)/2$, and $X_1, \dots, X_N \in \mathbb{R}^d$ be i.i.d. random vectors with jointly continuous distribution. Then,*

(a) If $N \ge N^*$, then $\mathbb{P}(\mathrm{span}(X_i X_i^T : i \in [N]) = \mathcal{S}) = 1$.

(b) If $N < N^*$, then for arbitrary $Z_1, \dots, Z_N \in \mathbb{R}^d$, $\mathrm{span}(Z_i Z_i^T : i \in [N]) \subsetneq \mathcal{S}$.

The proof of Theorem 6.2.10 is deferred to Section 6.4.14.

The critical number N^* is obtained to be $d(d+1)/2$ since $\dim(\mathcal{S}) = \binom{d}{2} + d = d(d+1)/2$. Note also that, with this observation, part (b) of Theorem 6.2.10 is trivial, since we do not have enough number of matrices to span the space $\mathcal{S}$.

Sample Complexity Bound for the Planted Network Model

Combining Theorems 6.2.9 and 6.2.10, we arrive at the following sample complexity result.

Theorem 6.2.11. *Let $X_i, 1 \le i \le N$ be i.i.d. with a jointly continuous distribution on $\mathbb{R}^d$. Let the corresponding labels $(Y_i)_{i=1}^N$ be generated via $Y_i = f(W^*; X_i)$, with $W^* \in \mathbb{R}^{m\times d}$ with $\mathrm{rank}(W^*) = d$.*

(a) Suppose $N \ge N^*$, and $\widehat{m} \in \mathbb{N}$. Then with probability one over the training data $X_1, \dots, X_N$, if $W \in \mathbb{R}^{\widehat{m}\times d}$ is such that $f(W; X_i) = Y_i$ for every $i \in [N]$, then $f(W; X) = f(W^*; X)$ for every $X \in \mathbb{R}^d$.

(b) Suppose $X_i, 1 \le i \le N$ are i.i.d. random vectors with i.i.d. centered coordinates having variance μ_2 and finite fourth moment μ_4. Suppose that $N < N^*$. Then there exists a $W \in \mathbb{R}^{m\times d}$ such that $f(W; X_i) = Y_i$ for every $i \in [N]$, yet the generalization error satisfies

$$\mathcal{L}(W) \ge \min\{\mu_4 - \mu_2^2, 2\mu_2^2\} \cdot \sigma_{\min}(W^*)^4.$$

The proof of Theorem 6.2.11 is deferred to Section 6.4.15.

We highlight that the lower bound arising in Theorem 6.2.11 (b) is very similar to the energy barrier bounds obtained earlier for rank-deficient matrices in Theorem 6.2.2 and Theorem 6.2.1 (a). Note also that the interpolating network in part (a) can

potentially be larger than the original network generating the data: any large network, despite being overparameterized, still generalizes well, provided it interpolates on a training set enjoying the aforementioned geometric condition.

Theorem 6.2.11 provides the necessary and sufficient number of data points for training a shallow neural network with quadratic activation function so as to guarantee good (perfect) generalization property.

6.3 Preliminaries

We collect herein several useful auxiliary results that we employ in our proofs.

6.3.1 An Analytical Expression for the Population Risk

Towards proving our energy barrier results, Theorem 6.2.2 and Theorem 6.2.1, we start with providing an analytical expression for the population risk $\mathcal{L}(W)$ of any $W \in \mathbb{R}^{m\times d}$ in terms of how close it is to the planted weight matrix $W^* \in \mathbb{R}^{m\times d}$.

We recall that a random vector X in $\mathbb{R}^d$ is defined to have jointly continuous distribution if there exists a measurable function $f : \mathbb{R}^d \to \mathbb{R}$ such that for any $i \in [N]$ and Borel set $\mathcal{B} \subseteq \mathbb{R}^d$,

$$\mathbb{P}(X \in \mathcal{B}) = \int_{\mathcal{B}} f(x_1, \ldots, x_d)\, d\lambda(x_1, \ldots, x_d),$$

where λ is the Lebesgue measure on $\mathbb{R}^d$.

Theorem 6.3.1. *Let $W^* \in \mathbb{R}^{m\times d}$, $f(W^*; X)$ be the function computed by (6.1); and $f(W; X)$ be similarly the function computed by (6.1) for $W \in \mathbb{R}^{m\times d}$. Recall,*

$$\mathcal{L}(W) = \mathbb{E}[(f(W^*; X) - f(W; X))^2],$$

where the expectation is with respect to the distribution of $X \in \mathbb{R}^d$.

(a) Suppose the distribution of X is jointly continuous. Then $\mathcal{L}(W) = 0$, that is, $f(W^*; X) = f(W; X)$ almost surely with respect to X, if and only if $W = QW^*$ for some orthonormal matrix $Q \in \mathbb{R}^{m\times m}$.

Suppose now that the coordinates of $X \in \mathbb{R}^d$ are i.i.d. with $\mathbb{E}[X_i] = 0, \mathbb{E}[X_i^2] = \mu_2$, and $\mathbb{E}[X_i^4] = \mu_4$.

(b) It holds that:

$$\mathcal{L}(W) = \mu_2^2 \cdot \mathrm{trace}(A)^2 + 2\mu_2^2 \cdot \mathrm{trace}(A^2) + (\mu_4 - 3\mu_2^2) \cdot \mathrm{trace}(A \circ A),$$

where $A = (W^*)^T W^* - W^T W \in \mathbb{R}^{d\times d}$, and $A \circ A$ is the Hadamard product of A with itself. In particular, if $X \in \mathbb{R}^d$ has i.i.d. standard normal coordinates, we obtain $\mathcal{L}(W) = \mathrm{trace}(A)^2 + 2\mathrm{trace}(A^2)$.

(c) The following bounds hold:

$$\mu_2^2 \cdot \mathrm{trace}(A)^2 + \min\left\{\mu_4 - \mu_2^2, 2\mu_2^2\right\} \cdot \mathrm{trace}(A^2) \leq \mathcal{L}(W),$$

and

$$\mu_2^2 \cdot \mathrm{trace}(A)^2 + \max\left\{\mu_4 - \mu_2^2, 2\mu_2^2\right\} \cdot \mathrm{trace}(A^2) \geq \mathcal{L}(W).$$

The proof of Theorem 6.3.1 is provideed in Section 6.4.1.

In a nutshell, Theorem 6.3.1 states that the population risk $\mathcal{L}(W)$ of any $W \in \mathbb{R}^d$ is completely determined by how close it is to the planted weights W^* as measured by the matrix $A = (W^*)^T W^* - W^T W$; and the second and fourth moments of the data. This is not surprising: $\mathcal{L}(W)$ is essentially a function of the first four moments of the data, and the difference of the quadratic forms generated by W and W^*, which is precisely encapsulated by the matrix A. Note also that the characterization of the "optimal orbit" per part (a) is not surprising either: any matrix W with the property $W = QW^*$ where $Q \in \mathbb{R}^{m\times m}$ is an orthonormal matrix, that is, $Q^T Q = I_m$, has the property that $f(W; X) = \|WX\|_2^2 = X^T W^T W X = f(W^*; X)$ for any data $X \in \mathbb{R}^d$. Part (a) then says the the reverse is true as well, provided that the distribution of X is jointly continuous. Note also that for X with centered i.i.d. entries the thesis of part (a) follows also from part (c): $\mathcal{L}(W) = 0$ implies that $\mathrm{trace}(A^2) = 0$, which, together with the fact that A is symmetric, then yields $A = 0$, that is, $W^T W = (W^*)^T W^*$.

6.3.2 Useful Lemmas and Results from Linear Algebra and Random Matrix Theory

Our next result is a simple norm bound for the ensemble $X_i \in \mathbb{R}^d$, $1 \leq i \leq N$ with sub-Gaussian coordinates.

Lemma 6.3.2. *Let $X_i \in \mathbb{R}^d$, $1 \leq i \leq N$ be an i.i.d. collection of random vectors with centered i.i.d. sub-Gaussian coordinates, that is, for some constant $C > 0$, $\mathbb{P}(|X_i(j)| > t) \leq \exp(-Ct^2)$ for every $i \in [N], j \in [d]$, and $t \geq 0$. Then,*

$$\mathbb{P}\left(\|X_i\|_\infty \leq d^{K_1}, 1 \leq i \leq N\right) \geq 1 - Nd\exp(-Cd^{2K_1}).$$

The proof of Lemma 6.3.2 is provided in Section 6.4.3.

Our energy barrier result Theorem 6.2.2 for the empirical risk is proven by establishing the emergence of a barrier for a **single** rank-deficient $A \in \mathbb{R}^{d\times d}$, together with a covering numbers argument.

Lemma 6.3.3. *Let $K_1 > 0$ be an arbitrary constant; and $X_i \in \mathbb{R}^d$, $1 \leq i \leq N$ be a collection of i.i.d. data with centered i.i.d. sub-Gaussian coordinates where for any $M > 0$, the mean of $|X_1(1)|$ conditional on $|X_1(1)| \leq M$ is zero; and let $Y_i = f(W^*; X_i)$ be the corresponding label generated by a neural network with planted weights $W^* \in \mathbb{R}^{m\times d}$ as per (6.1), where $\|W^*\|_F \leq d^{K_2}$. Fix any $A \in \mathbb{R}^{d\times d}$, where*

$\|A\|_F \le d^{2K_2}$, $\mathrm{rank}(A) \le d-1$, *and* $A \succeq 0$. *Define the event*

$$\mathcal{E}(A) \triangleq \left\{ \frac{1}{N} \sum_{1 \le i \le N} \left(Y_i - X_i^T A X_i\right)^2 \ge \frac{1}{2} C_5(K_1) \sigma_{\min}(W^*)^4 \right\},$$

where

$$C_5(K_1) \triangleq \min\Big\{ \mu_4(K_1) - \mu_2(K_1)^2, 2\mu_2(K_1)^2 \Big\}$$

for $\mu_n(K) = \mathbb{E}\Big[X_1(1)^n \Big| |X_1(1)| \le d^K \Big]$. *Then, there exists a constant* $C' > 0$ *(independent of* W*, and depending only on data distribution,* K_1*, and* W^**) such that*

$$\mathbb{P}\Big(\mathcal{E}(A)^c \Big| \|X_i\|_\infty \le d^{K_1}, 1 \le i \le N \Big) \le \exp\left(-C_3 \frac{N}{d^{4K_1+4K_2+2}} \right).$$

In particular,

$$\mathbb{P}\big(\mathcal{E}(A)\big) \ge 1 - \exp\left(-C_3 \frac{N}{d^{4K_1+4K_2+2}} \right) - N d e^{-C d^{2K_1}},$$

where $C > 0$ *is the same constant as in Lemma 6.3.2.*

The parameter K_1 appearing in Lemma 6.3.3 controls the amount of *truncation* applied on training data; and K_2 controls the norm of the planted weight matrix. The proof of Lemma 6.3.3 is provided in Section 6.4.4.

The next result is a covering number bound, adopted from [68, Lemma 3.1] with minor modifications.

Lemma 6.3.4. *Let*

$$S_R \triangleq \left\{ A \in \mathbb{R}^{d \times d} : \mathrm{rank}(A) \le r, A \succeq 0, \|A\|_F \le R \right\}.$$

Then there exists an $\epsilon-$*net* $\bar{S_R}$ *for* S_R *in Frobenius norm (that is, for every* $A \in S_R$ *there exists a* $\widehat{A} \in \bar{S_R}$ *such that* $\|A - \widehat{A}\|_F \le \epsilon$*) such that*

$$|\bar{S_R}| \le \left(\frac{9R}{\epsilon} \right)^{dr+r}.$$

The proof of Lemma 6.3.4 is provided in Section 6.4.5.

Some of our results use the following well-known results. These results are verbatim from the literature and provided herein without proof.

Theorem 6.3.5. *([70]) Let* ℓ *be an arbitrary positive integer; and* $P : \mathbb{R}^\ell \to \mathbb{R}$ *be a polynomial. Then, either* P *is identically* 0*, or* $\{x \in \mathbb{R}^\ell : P(x) = 0\}$ *has zero Lebesgue measure, namely,* $P(x)$ *is non-zero almost everywhere.*

Theorem 6.3.6. *([170, Theorem 7.3.11]) For two matrices* $A \in \mathbb{R}^{p \times n}$ *and* $B \in \mathbb{R}^{q \times n}$ *where* $q \le p$*;* $A^T A = B^T B$ *holds if and only if* $A = QB$ *for some matrix* $Q \in \mathbb{R}^{p \times q}$ *with orthonormal columns.*

Our results regarding the initialization guarantees use the several auxiliary results from random matrix theory: The spectrum of tall random matrices are essentially concentrated:

Theorem 6.3.7. ([281, Corollary 5.35]) Let A be an $m \times d$ matrix with independent standard normal entries. For every $t \geq 0$, with probability at least $1 - 2\exp(-t^2/2)$, we have:

$$\sqrt{m} - \sqrt{d} - t \leq \sigma_{\min}(A) \leq \sigma_{\max}(A) \leq \sqrt{m} + \sqrt{d} + t.$$

Theorem 6.3.8. *([24],[281, Theorem 5.31])*

Let $A = A_{N,n}$ be an $N \times n$ random matrix whose entries are independent copies of a random variable with zero mean, unit variance, and finite fourth moment. Suppose that the dimensions N and n grow to infinity while the aspect ratio n/N converges to a constant in $[0, 1]$. Then

$$\sigma_{\min}(A) = \sqrt{N} - \sqrt{n} + o(\sqrt{n}), \quad \textit{and} \quad \sigma_{\min}(A) = \sqrt{N} + \sqrt{n} + o(\sqrt{n}),$$

almost surely.

The following concentration result, recorded herein verbatim from Vershynin [281], will be useful for our approximate stationarity analysis.

Theorem 6.3.9. *([281, Theorem 5.44]) Let A be an $N \times n$ matrix whose rows A_i are independent random vectors in $\mathbb{R}^n$ with the common second moment matrix $\Sigma = \mathbb{E}[A_i A_i^T]$. Let m be a number such that $\|A_i\|_2 \leq \sqrt{m}$ almost surely for all i. Then, for every $t \geq 0$, the following inequality holds with probability at least $1 - n \cdot \exp(-ct^2)$:*

$$\left\| \frac{1}{N} A^T A - \Sigma \right\| \leq \max\left(\|\Sigma\|^{1/2} \delta, \delta^2 \right) \quad \textit{where} \quad \delta = t\sqrt{m/N}.$$

Here, $c > 0$ is an absolute constant.

Finally, we make use of the matrix-operator version of the Hölder's inequality:

Theorem 6.3.10. *([48, p.95]) For any matrix $U \in \mathbb{R}^{k \times \ell}$, let $\|U\|_{\sigma_p}$ denotes the ℓ_p norm of the vector*

$$(\sigma_1(U), \ldots, \sigma_{\min\{k,\ell\}}(U))$$

of singular values of U. Then, for any $p, q > 0$ with $\frac{1}{p} + \frac{1}{q} = 1$, it holds that

$$|\langle U, V \rangle| = |\mathrm{trace}(U^T V)| \leq \|U\|_{\sigma_p} \|V\|_{\sigma_q}.$$

6.4 Proofs

In this section, we present the proofs of the main results of this work.

The order of the proofs presented herein is slightly different from the order of the corresponding results in the main body, in that none of the proofs below (with one exception that we detail below) use a proof presented later than itself. That

is, whenever we present the proof of a result below, it is ensured that if this proof requires another result as a building block, this building block is shown earlier. The rationale behind this is to avoid any potential confusion and to ensure that no cyclic reasoning is present.

With this arrangement, only Theorem 6.2.4 uses results presented later in this section (more precisely, it uses Theorems 6.2.9 and 6.2.10); and it can be checked directly that there is no cyclic reasoning in the proof of Theorem 6.2.4.

6.4.1 Proof of Theorem 6.3.1

First, we have

$$f(W;X) - f(W^*;X) = X^T((W^*)^T W^* - W^T W)X \triangleq X^T A X, \tag{6.5}$$

where $A = (W^*)^T W^* - W^T W \in \mathbb{R}^{d\times d}$ is a symmetric matrix. Note also that,

$$\text{trace}(A)^2 = \sum_{i=1}^{d} A_{ii}^2 + 2\sum_{i<j} A_{ii}A_{jj}, \tag{6.6}$$

and

$$\text{trace}(A^2) = \text{trace}(A^T A) = \|A\|_F^2 = \sum_{i,j} A_{ij}^2 = \sum_{i=1}^{d} A_{ii}^2 + 2\sum_{i<j} A_{ij}^2, \tag{6.7}$$

where A^2 is equal to $A^T A$, as A is symmetric.

(a) Recall Theorem 6.3.5. In particular, if $\mathcal{L}(W) = 0$, then we have $P(X) = X^T A X = 0$ almost surely. Since $P(\cdot) : \mathbb{R}^d \to \mathbb{R}$ a polynomial, it then follows that $P(X) = 0$ identically. Now, since A is symmetric, it has real eigenvalues, called $\lambda_1, \ldots, \lambda_d$ with corresponding (real) eigenvectors $\xi_1, \ldots, \xi_d$. Now, taking $X = \xi_i$, we have $X^T A X = \xi_i^T A \xi = \lambda_i \langle \xi_i, \xi_i \rangle = 0$. Since $\xi_i \neq 0$, we get $\lambda_i = 0$ for any i. Finally, since $A = Q\Lambda Q^T$, it must necessarily be the case that $A = 0$. Hence, $W^T W = (W^*)^T W^*$, which imply $W = QW^*$ for some $Q \in \mathbb{R}^{m\times m}$ orthonormal, per Theorem 6.3.6.

(b) Using Equation (6.5), we first have

$$\mathcal{L}(W) = \sum_{1\leq i,j,i',j'\leq d} A_{ij}A_{i',j'}\mathbb{E}\left[X_i X_j X_{i'} X_{j'}\right].$$

Note that if $|\{i,j,i',j'\}| \in \{3,4\}$, then $\mathbb{E}\left[X_i X_j X_{i'} X_{j'}\right] = 0$, since X has centered i.i.d. coordinates. Keeping this in mind, and carrying out the algebra we then

get:

$$\mathcal{L}(W) = \sum_{i=1}^{d} A_{ii}^2 \mathbb{E}\left[X_i^4\right] + 2\sum_{i<j} A_{ii}A_{jj}\mathbb{E}\left[X_i^2\right]\mathbb{E}\left[X_j^2\right] + 4\sum_{i<j} A_{ij}^2 \mathbb{E}\left[X_i^2\right]\mathbb{E}\left[X_j^2\right]$$
$$= \mu_4 \sum_{i=1}^{d} A_{ii}^2 + 2\mu_2^2 \sum_{i<j} A_{ii}A_{jj} + 4\mu_2^2 \sum_{i<j} A_{ij}^2.$$

Using now Equations (6.6) and (6.7), we get:

$$\mathcal{L}(W) = (\mu_4 - 3\mu_2^2)\cdot \operatorname{trace}(A \circ A) + \mu_2^2 \cdot \operatorname{trace}(A)^2 + 2\mu_2^2 \cdot \operatorname{trace}(A^2),$$

since $A_{ii}^2 = (A \circ A)_{ii}$.

(c) Define k to be such that $\mu_4 - \mu_2^2 = 2k\mu_2^2$, namely, k is related to measures of dispersion pertaining X_i: $\sqrt{2k}$ is the coefficient of variation and $(2k+1)$ is the kurtosis associated to the random variable X_i. With this, we have:

$$\mathcal{L}(W) = \mu_2^2 \cdot \operatorname{trace}(A)^2 + 2\mu_2^2 \left(k\sum_{i=1}^{d} A_{ii}^2 + 2\sum_{i<j} A_{ij}^2 \right).$$

From here, the desired conclusion follows since

$$\mu_2^2 \cdot \operatorname{trace}(A)^2 + 2\min\{k,1\}\mu_2^2 \left(\sum_{i=1}^{d} A_{ii}^2 + 2\sum_{i<j} A_{ij}^2 \right) \leq \mathcal{L}(W),$$

and

$$\mu_2^2 \cdot \operatorname{trace}(A)^2 + 2\max\{k,1\}\mu_2^2 \left(\sum_{i=1}^{d} A_{ii}^2 + 2\sum_{i<j} A_{ij}^2 \right) \geq \mathcal{L}(W),$$

together with Equation (6.7).

6.4.2 Proof of Theorem 6.2.1

(a) Note first that using Theorem 6.3.1 part (c), we have:

$$\mathcal{L}(W) \geq \min\{\operatorname{Var}(X_i^2), 2\mathbb{E}\left[X_i^2\right]^2\}\operatorname{trace}(A^2).$$

Now, fix any $W \in \mathbb{R}^{m\times d}$ with $\operatorname{rank}(W) < d$. Let $a_1 \geq \cdots \geq a_d$ be the eigenvalues of $(W^*)^T W^*$; $b_1 \geq \cdots \geq b_d$ be the eigenvalues $-W^T W$; and $\lambda_1 \geq \cdots \geq \lambda_d$ be the eigenvalues of $(W^*)^T W^* - W^T W$. Since W is rank-deficient, we have $b_1 = 0$. Furthermore, $a_d = \sigma_{\min}(W^*)^2$, since the eigenvalues of $(W^*)^T W^*$ are precisely the squares of the singular values of W^*. Now, recall the (Courant-Fischer) variational characterization of the eigenvalues [170]. If M is a $d \times d$ matrix with

eigenvalues $c_1 \geq \cdots \geq c_d$, then:

$$c_1 = \max_{x:\|x\|_2=1} x^T M x \quad \text{and} \quad c_d = \min_{x:\|x\|_2=1} x^T M x.$$

With this, fix an $x \in \mathbb{R}^d$ with $\|x\|_2 = 1$. Then,

$$x^T((W^*)^T W^* - W^T W)x \geq \min_{x:\|x\|_2=1} x^T(W^*)^T W^* x + x^T(-W^T W)x = a_d + x^T(-W^T W)x.$$

Since this inequality holds for every x with $\|x\|_2 = 1$, we can take the max over all x, and arrive at,

$$\lambda_1 = \max_{x:\|x\|_2=1} x^T((W^*)^T W^* - W^T W)x \geq a_d + b_1 = a_d \geq \sigma_{\min}(W^*)^2.$$

Now, since $\lambda_1^2, \ldots, \lambda_d^2$ are precisely the eigenvalues of A^2, we have $\mathrm{trace}(A^2) = \sum_{i=1}^d \lambda_i^2 \geq \lambda_1^2$. Hence, for any W with $\mathrm{rank}(W) < d$, it holds that:

$$\mathcal{L}(W) \geq \min\left\{\mathrm{Var}(X_i^2), 2\mathbb{E}\left[X_i^2\right]^2\right\}\lambda_1^2.$$

Finally, since $\lambda_1^2 \geq \sigma_{\min}(W^*)^4$, the desired conclusion follows by taking the minimum over all rank-deficient W.

(b) Let the eigenvalues of $(W^*)^T W^*$ be denoted by $\lambda_1^*, \ldots, \lambda_d^*$, with the corresponding orthogonal eigenvectors $q_1^*, \ldots, q_d^*$. Namely, diagonalize $(W^*)^T W^*$ as $Q^* \Lambda^* (Q^*)^T$ where the columns of $Q^* \in \mathbb{R}^{d\times d}$ are $q_1^*, \ldots, q_d^*$, and $\Lambda^* \in \mathbb{R}^{d\times d}$ is a diagonal matrix with $(\Lambda^*)_{i,i} = \lambda_i^*$ for every $1 \leq i \leq d$. Let

$$\overline{W} = \sum_{j=1}^{d-1} \sqrt{\lambda_j^*} q_j^* (q_j^*)^T \in \mathbb{R}^{d\times d}.$$

Observe that, $\overline{W}^T \overline{W} = Q^* \overline{\Lambda} Q^*$, where $\overline{\Lambda} \in \mathbb{R}^{d\times d}$ is a diagonal matrix with $(\overline{\Lambda})_{i,i} = (\Lambda^*)_{i,i}$ for every $1 \leq i \leq d-1$, and $(\overline{\Lambda})_{d,d} = 0$; and that, $\mathrm{rank}(\overline{W}) = d-1$. Now, let $\overline{W_1}, \ldots, \overline{W_d} \in \mathbb{R}^d$ be the rows of $\overline{W}$, and fix a $j \in [d]$ such that $\overline{W_j} \neq 0$.

Having constructed a $\overline{W} \in \mathbb{R}^{d\times d}$, we now prescribe $W \in \mathbb{R}^{m\times d}$ as follows. For $1 \leq i \leq d$, $i \neq j$, let $W_i = \overline{W_i}$, where W_i is the i^{th} row of W. Then set $W_j = \frac{1}{2}\overline{W_j}$, and for every $d+1 \leq i \leq m$, set $W_i = \frac{\sqrt{3}}{2\sqrt{m-d}}\overline{W_j}$. For this matrix, we now claim

$$W^T W = \overline{W}^T \overline{W}.$$

To see this, fix an $X \in \mathbb{R}^d$, and recall that $X^T W^T W X - X^T \overline{W}^T \overline{W} X =$

$\|WX\|_2^2 - \|\overline{W}X\|_2^2$. We now compute this quantity more explicitly:

$$\begin{aligned}
\|WX\|_2^2 - \|\overline{W}X\|_2^2 &= \sum_{k=1}^{d} \langle W_k, X\rangle^2 - \sum_{k=1}^{m} \left\langle \overline{W}_k, X\right\rangle^2 \\
&= \sum_{k=1,k\neq j}^{d} \langle W_k, X\rangle^2 + \langle W_j, X\rangle^2 \\
&\quad - \sum_{k=1,k\neq j}^{d} \langle W_k, X\rangle^2 - \left\langle \frac{1}{2}W_j, X\right\rangle^2 - \sum_{k=d+1}^{m} \left\langle \frac{\sqrt{3}}{2\sqrt{m-d}}W_j, X\right\rangle^2 \\
&= \langle W_j, X\rangle^2 - \frac{1}{4}\langle W_j, X\rangle^2 - \frac{3}{4(m-d)}(m-d)\langle W_j, X\rangle^2 = 0.
\end{aligned}$$

Hence, for every $X \in \mathbb{R}^d$, we have:

$$X^T W^T W X = X^T \overline{W}^T \overline{W} X.$$

Now let $\Xi = W^T W - \overline{W}^T \overline{W}$. Note that $\Xi \in \mathbb{R}^{d\times d}$ is symmetric, and $X^T \Xi X = 0$ for every $X \in \mathbb{R}^d$. Now, taking X to be e_i, that is, the i^{th} element of the standard basis for the Euclidean space $\mathbb{R}^d$, we deduce $\Xi_{i,i} = 0$ for every $i \in [d]$. For the off-diagonal entries, let $X = e_i + e_j$. Then, $X^T \Xi X = \Xi_{i,i} + \Xi_{i,j} + \Xi_{j,i} + \Xi_{j,j} = 0$, which, together with the fact that the diagonal entries of Ξ are zero, imply $\Xi_{i,j} = -\Xi_{j,i}$; namely Ξ is skew-symmetric. Finally, since Ξ is also symmetric we have $\Xi_{i,j} = \Xi_{j,i}$, which then implies for every $i, j \in [d]$, $\Xi_{i,j} = 0$, that is, $\Xi = 0$, and thus, $W^T W = \overline{W}^T \overline{W}$.

Hence, we have for $W \in \mathbb{R}^{m\times d}$ with $\mathrm{rank}(W) = d - 1$,

$$W^T W - (W^*)^T W^* = Q^* \Lambda' (Q^*)^T,$$

with $(\Lambda')_{i,i} = 0$ for every $1 \leq i \leq d-1$; and $(\Lambda')_{d,d} = -\lambda_d^*$. Namely, the spectrum of the matrix $A = (W^*)^T W^* - W^T W$ contains only two values: 0 with multiplicity $d-1$, and λ_d^* with multiplicity one. In particular,

$$\mathrm{trace}(A) = \lambda_d^* \quad \text{and} \quad \mathrm{trace}(A^2) = (\lambda_d^*)^2.$$

Using now the upper bound provided by Theorem (6.3.1) part (c) yields the desired claim. Therefore, the energy band lower bound is tight, up to a multiplicative constant.

6.4.3 Proof of Lemma 6.3.2

For any fixed $i \in [N], j \in [d]$, note that using sub-Gaussian property one has $\mathbb{P}(|X_i(j)| > d^{K_1}) \leq \exp(-Cd^{2K_1})$, thus $\mathbb{P}(\exists i \in [N], j \in [d] : |X_i(j)| > d^{K_1}) \leq Nd\exp(-Cd^{2K_1})$, using union bound, which yields the conclusion.

6.4.4 Proof of Lemma 6.3.3

Let

$$\mathcal{E}_1 \triangleq \Big\{ \|X_i\|_\infty \leq d^{K_1}, 1 \leq i \leq N \Big\}.$$

By Lemma 6.3.2, $\mathbb{P}(\mathcal{E}_1) \geq 1 - Nd\exp(-Cd^{2K_1})$. Now, note that

$$\mathbb{P}(\mathcal{E}(A)^c) = \mathbb{P}(\mathcal{E}(A)^c|\mathcal{E}_1)\mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}(A)^c|\mathcal{E}_1^c)\mathbb{P}(\mathcal{E}_1^c) \leq \mathbb{P}(\mathcal{E}(A)^c|\mathcal{E}_1) + N\exp(-Cd^{2K_1}). \tag{6.8}$$

We now study $\mathbb{P}(\mathcal{E}(A)^c|\mathcal{E}_1)$, hence assume we condition on $\mathcal{E}_1$ from now on. Triangle inequality yields

$$|Y_i - X_i^T A X_i| \leq |X_i^T A X_i| + |X_i^T (W^*)^T W^* X_i|.$$

Observe now that

$$\|X_i X_i\|_F^2 = \mathrm{trace}(X_i X_i^T X_i X_i^T) = \|X_i\|_2^2 \mathrm{trace}(X_i X_i^T) = \|X_i\|_2^4,$$

which implies (conditional on $\mathcal{E}_1$)

$$\|X_i X_i^T\|_F = \|X_i\|_2^2 \leq d^{2K_1+1}.$$

Now, Cauchy-Schwarz inequality with respect to inner product $\langle U, V \rangle \triangleq \mathrm{trace}(U^T V)$ yields

$$|X_i^T A X_i| = \left\langle A, X_i X_i^T \right\rangle \leq \|A\|_F \|X_i X_i^T\|_F \leq d^{2K_1+2K_2+1},$$

for every $i \in [N]$, using $\|A\|_F \leq d^{2K_2}$.

Next, let $A^* = (W^*)^T W^* \in \mathbb{R}^{d \times d}$, and let $\eta_1^*, \ldots, \eta_d^*$ be the eigenvalues of A^*, all non-negative. Observe that

$$\|W^*\|_F^2 = \mathrm{trace}(A^*) = \sum_{1 \leq j \leq d} \eta_j^* \leq d^{2K_2}.$$

Now note that $(\eta_1^*)^2, (\eta_2^*)^2, \ldots, (\eta_d^*)^2$ are the eigenvalues of $(A^*)^2 = (A^*)^T A^*$. With this reasoning, we have

$$\|A^*\|_F^2 = \mathrm{trace}((A^*)^T A^*) = \mathrm{trace}((A^*)^2) = \sum_{1 \leq j \leq d} (\eta_j^*)^2 \leq \left(\sum_{1 \leq j \leq d} \eta_j^* \right)^2 \leq d^{4K_2}.$$

Consequently, $\|A^*\|_F \leq d^{2K_2}$, and therefore, the exact same reasoning yields

$$|X_i^T (W^*)^T W^* X_i| = X_i^T A^* X_i \leq d^{2K_1+2K_2+1},$$

for every $i \in [N]$. Hence, conditional on $\mathcal{E}_1$, it holds that for every $i \in [N]$:

$$\left(X_i^T A X_i - X_i^T (W^*)^T W^* X_i \right)^2 \leq 4d^{4K_1+4K_2+2}.$$

We now apply concentration to i.i.d. sum

$$\frac{1}{N}\sum_{1\le i\le N}\left(X_i^TAX_i - X_i^T(W^*)^TW^*X_i\right)^2$$

is a sum of bounded random variables that are at most $4d^{4K_1+4K_2+2}$.

Now, recalling the distributional assumption on the data, we have that conditional on $\|X_i\|_\infty \le d^{K_1}$, the data still has i.i.d. centered coordinates. In particular, the "energy barrier" result for the population risk as per Theorem 6.2.1 applies:

$$\mathbb{E}\left[\left(X^TAX - X^T(W^*)^TW^*X\right)^2\Big|\mathcal{E}_1\right] \ge C_5(K_1)\sigma_{\min}(W^*)^4,$$

where

$$C_5(K_1) = \min\{\mu_4(K_1) - \mu_2(K_1)^2, 2\mu_2(K_1)^2\},$$

is controlled by the conditional moments of data coordinates.

Finally applying Hoeffding's inequality for bounded random variables we arrive at

$$\frac{1}{N}\sum_{1\le i\le N}\left(X_i^TAX_i - X_i^T(W^*)^TW^*X_i\right)^2 \ge \frac{1}{2}C_5\sigma_{\min}(W^*)^4,$$

with probability at least $1-\exp\left(-C_3Nd^{-4K_1-4K_2-2}\right)$. Namely,

$$\mathbb{P}\Big(\mathcal{E}(A)^c\Big|\mathcal{E}_1\Big) \le \exp(-C_3Nd^{-4K_1-4K_2-2}).$$

Returning to (6.8), this yields

$$\mathbb{P}(\mathcal{E}_A) \ge 1 - \exp(-C_3Nd^{-4K_1-4K_2-2}) - Nd\exp(-Cd^{2K_1}).$$

This completes the proof of Lemma 6.3.3.

6.4.5 Proof of Lemma 6.3.4

The proof is almost verbatim from [68, Lemma 3.1], and included herein for completeness.

Note that any $A \in \mathbb{R}^{d\times d}$, $A \succeq 0$ and $\mathrm{rank}(A) = r$ decomposes as $A = Q\Lambda Q^T$, where $Q \in \mathbb{R}^{d\times r}$ satisfying $Q^TQ = I_d$, and $\Lambda \in \mathbb{R}^{r\times r}$, a diagonal matrix with non-negative diagonal entries. Notice, furthermore, that $\|A\|_F = \|\Lambda\|_F \le R$ as Q is orthonormal. With this, we now construct an appropriate net covering the set of all permissible Q and Σ.

Let D be the set of all $r\times r$ diagonal matrices with non-negative diagonal entries with Frobenius norm at most R. Let $\bar{D}$ be an $\frac{\epsilon}{3}-$net for D in Frobenius norm. Using standard results (see, e.g. [281, Lemma 5.2]), we have

$$|\bar{D}| \le \left(\frac{9R}{\epsilon}\right)^r.$$

Now let $O_{d,r} = \{Q \in \mathbb{R}^{d\times r} : Q^TQ = I_d\}$. To cover $O_{d,r}$ we use a more convenient norm $\|\cdot\|_{1,2}$ defined as

$$\|X\|_{1,2} = \max_i \|X_i\|_2,$$

where X_i is the i^{th} column of X. Define $Q_{d,r} = \{X \in \mathbb{R}^{d\times r} : \|X\|_{1,2} \leq 1\}$. Note that $O_{d,r} \subset Q_{d,r}$. Furthermore, observe also that $Q_{d,r}$ has an $\epsilon-$net of cardinality at most $(3/\epsilon)^{dr}$. With this, we now take $\bar{O}_{d,r}$ to be an $\frac{\epsilon}{3R}-$net for $O_{d,r}$. Consider now the set

$$\bar{S}_R \triangleq \{\bar{Q}\bar{\Lambda}\bar{Q}^T : \bar{Q} \in \bar{O}_{d,r}, \bar{\Lambda} \in \bar{D}\}.$$

Clearly,

$$|\bar{S}_R| \leq |\bar{O}_{d,r}||\bar{D}| \leq (9R/\epsilon)^{dr+r}.$$

We now claim $\bar{S}_R$ is indeed an $\epsilon-$net for S_R in Frobenius norm. To prove this, take an arbitrary $A \in S_R$, and let $A = Q\Lambda Q^T$. There exists a $\bar{Q} \in \bar{O}_{d,r}$, and a $\bar{\Sigma} \in \bar{D}$ such that $\|\Sigma - \bar{\Sigma}\|_F \leq \epsilon/3$, and $\|Q - \bar{Q}\|_{1,2} \leq \epsilon/3R$. Now, let $\bar{A} = \bar{Q}\bar{\Sigma}\bar{Q}^T$. Observe that using triangle inequality

$$\begin{aligned}\|\bar{A} - A\|_F &= \|Q\Lambda Q^T - \bar{Q}\bar{\Lambda}\bar{Q}^T\|_F \\ &\leq \|Q\Lambda Q^T - \bar{Q}\Lambda Q^T\|_F + \|\bar{Q}\Lambda Q^T - \bar{Q}\bar{\Lambda}Q^T\|_F + \|\bar{Q}\bar{\Lambda}Q^T - \bar{Q}\bar{\Lambda}\bar{Q}^T\|_F.\end{aligned}$$

For the first term, notee that since Q is orthonormal, $\|(Q-\bar{Q})\Lambda Q^T\|_F = \|(Q-\bar{Q})\Lambda\|_F$. Next,

$$\|(Q-\bar{Q})\Lambda\|_F^2 = \sum_{1\leq i\leq d} \Lambda_{ii}^2\|Q_i - \bar{Q}_i\|_2^2 \leq \|Q-\bar{Q}\|_{1,2}^2\|\Sigma\|_F^2 \leq (\epsilon/3)^2,$$

using $\|Q-\bar{Q}\|_{1,2} \leq \epsilon/3R$ and $\|\Sigma\|_F \leq R$. Thus, $\|Q\Lambda Q^T - \bar{Q}\Lambda Q^T\|_F \leq \epsilon/3$. Similarly, we also have $\|\bar{Q}\bar{\Lambda}Q^T - \bar{Q}\bar{\Lambda}\bar{Q}^T\|_F \leq \epsilon/3$. Finally, $\|\bar{Q}\Lambda Q^T - \bar{Q}\bar{\Lambda}Q^T\|_F = \|\Lambda Q^T - \bar{\Lambda}Q^T\|_F = \|\Lambda - \bar{\Lambda}\|_F \leq \epsilon/3$ using again the facts that Q and $\bar{Q}$ are both orthonormal. This concludes that $\|\bar{A} - A\|_F \leq \epsilon$; thus $|\bar{S}_R|$ is indeed an $\epsilon-$net for S_R, in Frobenius norm, of cardinality at most $(9R/\epsilon)^{dr+r}$.

As a side remark observe that we gain an extra factor of 2 in the exponent owing to the fact that A is positive semidefinite (otherwise the bound would be $(9R/\epsilon)^{2dr+r}$).

6.4.6 Proof of Theorem 6.2.3

We first establish the following proposition, for any W, which is a stationary point of the population risk.

Proposition 6.4.1. *Let $\mathcal{D}^* \in \mathbb{R}^{d\times d}$ be a diagonal matrix with $\mathcal{D}^*_{ii} = ((W^*)^TW^*)_{ii}$, and define $\mathcal{D} \in \mathbb{R}^{d\times d}$ analogously. Then, $W \in \mathbb{R}^{m\times d}$ enjoys the "stationarity equation":*

$$\begin{aligned}&(\mu_4 - 3\mu_2^2)W\mathcal{D}^* + \mu_2^2 W\|W^*\|_F^2 + 2\mu_2^2(W(W^*)^TW^*) \\ &= (\mu_4 - 3\mu_2^2)W\mathcal{D} + \mu_2^2 W\|W\|_F^2 + 2\mu_2^2(W(W^TW)).\end{aligned}$$

Proof. To that end, fix a $k_0 \in [m]$ and $\ell_0 \in [d]$. Note that, $\nabla_{k_0,\ell_0}\mathcal{L}(W) = \mathbb{E}\left[\nabla_{k_0,\ell_0}(f(W^*;X) - f(W;X)\right.$ using dominated convergence theorem. Next, $\mathbb{E}\left[\nabla_{k_0,\ell_0}(f(W^*;X) - f(W;X))^2\right] = 0$ implies that, for every $k_0 \in [m]$ and $\ell_0 \in [d]$:

$$\sum_{j=1}^{m} \mathbb{E}\left[\left\langle W_j^*, X\right\rangle^2 \left\langle W_{k_0}, X\right\rangle X_{\ell_0}\right] = \sum_{j=1}^{m} \mathbb{E}\left[\left\langle W_j, X\right\rangle^2 \left\langle W_{k_0}, X\right\rangle X_{\ell_0}\right].$$

Note next that, $\sum_{j=1}^{m} \mathbb{E}\left[\left\langle W_j^*, X\right\rangle^2 \left\langle W_{k_0}, X\right\rangle X_{\ell_0}\right]$ computes as,

$$\mu_4 \sum_{j=1}^{m} (W_{j,\ell_0}^*)^2 W_{k_0,\ell_0} + \mu_2^2 \sum_{j=1}^{m} \sum_{1\le\ell\le d, \ell\neq\ell_0} W_{k_0,\ell_0}(W_{j,\ell}^*)^2 + 2\mu_2^2 \sum_{j=1}^{m} \sum_{1\le\ell\le d, \ell\neq\ell_0} W_{k_0,\ell} W_{j,\ell}^* W_{j,\ell_0}^*.$$

We now put this object into a more convenient form. Notice that the expression above is

$$(\mu_4 - 3\mu_2^2) A_{k_0,\ell_0} + \mu_2^2 B_{k_0,\ell_0} + 2\mu_2^2 C_{k_0,\ell_0},$$

where

$$A_{k_0,\ell_0} = W_{k_0,\ell_0} \sum_{j=1}^{m} (W_{j,\ell_0}^*)^2 \quad \text{and} \quad B_{k_0,\ell_0} = \sum_{j=1}^{m}\sum_{\ell=1}^{d} W_{k_0,\ell_0}(W_{j,\ell}^*)^2 \quad \text{and} \quad C_{k_0,\ell_0} = \sum_{j=1}^{m}\sum_{\ell=1}^{d} W_{k_0,\ell} W_{j,\ell_0}^* W_{j,\ell}^*.$$

Observe that, $B_{k_0,\ell_0} = W_{k_0,\ell_0}\|W^*\|_F^2$. We now study A_{k_0,ℓ_0} and C_{k_0,ℓ_0} more carefully. Observe that $\sum_{j=1}^{m}(W_{j,\ell_0}^*)^2 = ((W^*)^T W^*)_{\ell_0,\ell_0}$. Now, let $\mathcal{D}^* \in \mathbb{R}d \times d$ be a diagonal matrix where $(\mathcal{D}^*)_{ij} = ((W^*)^T W^*)_{ii}$, if $i = j$; and 0 otherwise. We then have $A_{k_0,\ell_0} = (W\mathcal{D}^*)_{k_0,\ell_0}$. We now study C_{k_0,ℓ_0}. Recall that W_i^* is the i^{th} row W^*. Observe that, $\sum_{j=1}^{m} W_{j,\ell_0}^* W_{j,\ell}^* = ((W^*)^T W^*)_{\ell_0,\ell}$. Hence,

$$\sum_{j=1}^{m}\sum_{\ell=1}^{d} W_{k_0,\ell} W_{j,\ell_0}^* W_{j,\ell}^* = \sum_{\ell=1}^{d}\sum_{j=1}^{m} W_{k_0,\ell} W_{j,\ell_0}^* W_{j,\ell}^* = \sum_{\ell=1}^{d} W_{k_0,\ell}((W^*)^T W^*)_{\ell_0,\ell} = (W((W^*)^T W^*))_{k_0,\ell_0},$$

that is, $C_{k_0,\ell_0} = (W((W^*)^T W^*))_{k_0,\ell_0}$. Combining everything, we have that for every $k_0 \in [m]$ and $\ell_0 \in [d]$:

$$\sum_{j=1}^{m} \mathbb{E}\left[\left\langle W_j^*, X\right\rangle^2 \left\langle W_{k_0}, X\right\rangle X_{\ell_0}\right] = (\mu_4 - 3\mu_2^2)(W\mathcal{D}^*)_{k_0,\ell_0} + \mu_2^2 W_{k_0,\ell_0}\|W^*\|_F^2 + 2\mu_2^2(W((W^*)^T W^*))_{k_0,\ell_0}.$$

In particular, stationarity yields:

$$(\mu_4 - 3\mu_2^2) W\mathcal{D}^* + \mu_2^2 W\|W^*\|_F^2 + 2\mu_2^2(W((W^*)^T W^*)) = (\mu_4 - 3\mu_2^2) W\mathcal{D} + \mu_2^2 W\|W\|_F^2 + 2\mu_2^2 W(W^T W), \tag{6.9}$$

where the $d \times d$ diagonal matrix $\mathcal{D}$ is defined as $\mathcal{D}_{ii} = (W^T W)_{ii}$; and entrywise equalities are converted into equality of two matrices by varying $k_0 \in [m]$ and $\ell_0 \in [d]$. □

Having now established the Proposition 6.4.1 for the "stationarity equation", we now study its implications for any full-rank W.

Let $W \in \mathbb{R}^{m\times d}$ be a stationary point with $\text{rank}(W) = d$. We first establish $\|W\|_F = \|W^*\|_F$. Since $W \in \mathbb{R}^{m\times d}$ is a stationary point, it holds that for every $(k_0, \ell_0) \in [m] \times [d]$, $\nabla_{k_0,\ell_0}\mathcal{L}(W) = 0$. In particular, Equation (6.9) holds.

Recalling now that W is full rank, it follows from the rank-nullity theorem that $\ker(W)$ is trivial, that is, $\ker(W) = \{0\}$. Hence, for matrices M_1, M_2 (with matching dimensions), whenever $WM_1 = WM_2$ holds, we deduce $M_1 = M_2$, since each column of $M_1 - M_2$ is contained in $\ker(W)$. Thus, Equation (6.9) then yields:

$$(\mu_4 - 3\mu_2^2)\mathcal{D}^* + \mu_2^2\|W^*\|_F^2 I_d + 2\mu_2^2(W^*)^T W^* = (\mu_4 - 3\mu_2^2)\mathcal{D} + \mu_2^2\|W\|_F^2 I_d + 2\mu_2^2 W^T W. \tag{6.10}$$

Next, note that $\text{trace}(\mathcal{D}^*) = \sum_{i=1}^d ((W^*)^T W^*)_{ii} = \text{trace}((W^*)^T W^*) = \|W^*\|_F^2$, and similarly, $\text{trace}(\mathcal{D}) = \|W\|_F^2$. In particular, taking traces of both sides in Equation (6.10), we get

$$(\mu_4 - \mu_2^2)\|W^*\|_F^2 + \mu_2^2 d\|W^*\|_F^2 = (\mu_4 - \mu_2^2)\|W\|_F^2 + \mu_2^2 d\|W\|_F^2,$$

implying that $\|W^*\|_F^2 = \|W\|_F^2$. Incorporating this into Equation (6.10), we then arrive at:

$$(\mu_4 - 3\mu_2^2)\mathcal{D}^* + 2\mu_2^2(W^*)^T W^* = (\mu_4 - 3\mu_2^2)\mathcal{D} + 2\mu_2^2 W^T W.$$

Now, suppose $i \in [d]$. Note that inspecting (i, i) coordinate above, we get:

$$(\mu_4 - 3\mu_2^2)((W^*)^T W^*)_{ii} + 2\mu_2^2((W^*)^T W^*)_{ii} = (\mu_4 - 3\mu_2^2)(W^T W)_{ii} + 2\mu_2^2(W^T W)_{ii}.$$

Since $\mu_4 - \mu_2^2 = \text{Var}(X_i^2) > 0$, we then get

$$((W^*)^T W^*)_{ii} = (W^T W)_{ii}.$$

Now, focus on off-diagonal entries, by fixing $i \neq j$. Observe that since $\text{Var}(X_i^2) > 0$, it also holds $\mathbb{E}[X_i^2] = \mu_2 > 0$. Now note that, $\mathcal{D}^*_{ij} = \mathcal{D}_{ij} = 0$ in this case. We then have,

$$2\mu_2((W^*)^T W^*)_{ij} = 2\mu_2(W^T W)_{ij} \Rightarrow (W^*)^T W^* = W^T W.$$

We conclude that the matrix $(W^*)^T W^* - W^T W$ is a zero matrix. Hence, $W = QW^*$ for some orthonormal $Q \in \mathbb{R}^{m\times m}$ per Theorem 6.3.6, and $\mathcal{L}(W) = 0$.

6.4.7 Proof of Theorem 6.2.5

Let $\{W_t\}_{t\geq 0}$ be a sequence of $m \times d$ matrices corresponding to the weights along the trajectory of gradient descent, that is, $W_t \in \mathbb{R}^{m\times d}$ is the weight matrix at iteration t of the algorithm. We first show $L < \infty$. To see this, recall Theorem 6.3.1 (c): $\mathcal{L}(W) \geq \mu_2^2 \cdot \text{trace}(A)^2$, where $\text{trace}(A) = \|W\|_F^2 - \|W^*\|_F^2$. In particular, this yields

$\mu_2^2(\|W\|_F^2 - \|W^*\|_F^2)^2 \leq \mathcal{L}(W)$. Hence, for any W with $\mathcal{L}(W) \leq \mathcal{L}(W_0)$, it holds that

$$\|W\|_F \leq \left(\frac{\sqrt{\mathcal{L}(W_0)}}{\mu_2} + \|W^*\|_F^2 \right)^{1/2} < \infty.$$

Namely, the (Frobenius) norm of the weights of any W with $\mathcal{L}(W) \leq \mathcal{L}(W_0)$ remains uniformly bounded from above. This, in turn, yields that the (spectral norm of the) Hessian of the objective function remains uniformly bound from above for any such W, since the objective is a polynomial function of W, which is precisely what we denote by L.

We now run gradient descent with a step size of $\eta < 1/2L$: a second order Taylor expansion reveals that

$$\mathcal{L}(W_1) - \mathcal{L}(W_0) \leq -\eta \|\nabla \mathcal{L}(W_0)\|_2^2/2,$$

where $\nabla \mathcal{L}(W)$ is the gradient of the population risk, evaluated at W.

In particular, $\mathcal{L}(W_1) \leq \mathcal{L}(W_0)$, and furthermore, $\|\nabla^2 \mathcal{L}(W_1)\| \leq L$, where $\|\nabla^2 \mathcal{L}(W)\|$ is the spectral norm of the Hessian matrix $\nabla^2 \mathcal{L}(W)$. From here, we induct on k: induction argument reveals we can retain a step size of $\eta < 1/2L$, and furthermore we deduce that the gradient descent trajectory $\{W_k\}_{k \geq 0}$ is such that: (i) $\mathcal{L}(W_k) \geq \mathcal{L}(W_{k+1})$, for every $k \geq 0$, and furthermore, (ii) it holds for every $k \geq 0$:

$$\mathcal{L}(W_{k+1}) - \mathcal{L}(W_k) \leq -\eta \|\nabla \mathcal{L}(W_k)\|_2^2/2.$$

We now establish that $\|\nabla \mathcal{L}(W_k)\|_2 \to 0$ as $k \to \infty$. Note that the objective function is lower bounded (by zero). If the gradient is non-vanishing then (by passing to a subsequence, if necessary) each step reduces the value of the objective function at least by a certain amount, that is (uniformly) bounded away from zero. But this contradicts with the fact that the objective is lower bounded. Thus we deduce

$$\lim_{k \to \infty} \|\nabla \mathcal{L}(W_k)\|_2 = 0.$$

Now, recall that the trajectory is such that $\mathcal{L}(W_k) \geq \mathcal{L}(W_{k+1})$, and that, $\|\nabla \mathcal{L}(W_k)\|_2 \to 0$ as $k \to \infty$. Suppose that the initial value, $\mathcal{L}(W_0)$, is such that

$$\mathcal{L}(W_0) < \min_{W \in \mathbb{R}^{m \times d} : \mathrm{rank}(W) < d} \mathcal{L}(W).$$

In particular, for every $k \in \mathbb{Z}^+$,

$$\mathcal{L}(W_k) \leq \mathcal{L}(W_0) < \min_{W \in \mathbb{R}^{m \times d} : \mathrm{rank}(W) < d} \mathcal{L}(W). \tag{6.11}$$

and therefore $W_k \in \mathbb{R}^{m \times d}$ is full-rank, for all k, per Theorem 6.2.1. We now establish

$$\lim_{k \to \infty} \mathcal{L}(W_k) = 0.$$

To see this, observe that the sequence $\{\mathcal{L}(W_k)\}_{k\geq 0}$ is monotonic (non-increasing), and furthermore, is bounded by zero from below. Hence,

$$\lim_{k\to\infty} \mathcal{L}(W_k) \triangleq \ell$$

exists [248, Theorem 3.14]. We now show $\ell = 0$.

Since the weights remain bounded along the trajectory, it follows that there exists a subsequence $\{W_{k_n}\}_{n\in\mathbb{N}}$ with a limit, that is, $W_{k_n} \to W^\infty$ as $n \to \infty$, where $W^\infty \in \mathbb{R}^{m\times d}$. Now, the continuity of $\nabla\mathcal{L}$, together with the continuity of the norm $\|\cdot\|_2$, imply that $\|\nabla\mathcal{L}(W^\infty)\|_2 = 0$. Furthermore, continuity of $\mathcal{L}(\cdot)$ then implies $\mathcal{L}(W^\infty) = \ell$. Now, since W_{k_n}'s are such that $\mathcal{L}(W_{k_n}) \leq \mathcal{L}(W_0)$ for all $n \in \mathbb{N}$, and $\mathcal{L}(W_0)$ is stricly smaller than the rank-deficient energy barrier, by taking limits as $k \to \infty$ and using (6.11), we conclude that W^∞ is full rank. Since W^∞ is also a stationary point of the loss, by Theorem 6.2.3, we deduce $\mathcal{L}(W^\infty) = 0$, which yields $\ell = 0$, as desired.

6.4.8 Proof of Theorem 6.2.7

Part (a)

Let $t = \sqrt{d}$. Then, using Theorem 6.3.7, it holds that with probability $1-2\exp(-d/2)$:

$$\begin{aligned}\sqrt{m} - 2\sqrt{d} \leq \sigma_{\min}(W^*) \quad &\leq \sigma_{\max}(W^*) \quad \leq \sqrt{m} + 2\sqrt{d} \\ \Rightarrow m + 4d - 4\sqrt{md} \leq \lambda_{\min}((W^*)^T W^*) \quad &\leq \lambda_{max}((W^*)^T W^*) \leq m + 4d + 4\sqrt{md}.\end{aligned}$$

Recall that $\sigma(A)$ denotes the spectrum of A, i.e., $\sigma(A) = \{\lambda : \lambda \text{ is an eigenvalue of } A\}$. We claim then the spectrum of $\gamma I - A$ is $\gamma - \sigma(A)$. To see this, simply note the following line of reasoning:

$$\gamma - \lambda \in \sigma(\gamma I - A) \iff \det((\gamma-\lambda)I - (\gamma I - A)) = 0 \Leftrightarrow \det(\lambda I - A) = 0 \Leftrightarrow \lambda \in \sigma(A).$$

Now, let $W_0 \in \mathbb{R}^{m\times d}$ be such that $W_0^T W_0 = \gamma I$ with $\gamma = m + 4d$. In particular, if $\lambda_1 \leq \cdots \leq \lambda_d$ are the eigenvalues of $\gamma I - (W^*)^T W^*$ with $\gamma = m + 4d$; then, it holds that:

$$-4\sqrt{md} \leq \lambda_1 \leq \cdots \leq \lambda_d \leq 4\sqrt{md}.$$

Now, recall by Theorem 6.3.1 (c) that,

$$\mathcal{L}(W_0) \leq \mu_2^2 \left(\sum_{i=1}^d \lambda_i\right)^2 + \max\left\{\mathrm{Var}(X_i^2), 2\mathbb{E}\left[X_i^2\right]^2\right\}\left(\sum_{i=1}^d \lambda_i^2\right),$$

where $\sigma(W_0^T W_0 - (W^*)^T W^*) = \{\lambda_1, \ldots, \lambda_d\}$. For the second term, we immediately have $\sum_{i=1}^d \lambda_i^2 \leq 16md^2$.

For the first term, note first that, if $\lambda_1' \leq \cdots \leq \lambda_d'$ are the eigenvalues of $(W^*)^T W^*$,

then

$$\sum_{k=1}^{d}\lambda'_k = \text{trace}((W^*)^T W^*) = \sum_{i=1}^{m}\sum_{j=1}^{d}(W^*_{ij})^2 \Rightarrow \sum_{k=1}^{d}(\lambda'_k - m) = \sum_{i=1}^{m}\sum_{j=1}^{d}((W^*_{ij})^2 - 1),$$

where $W^*_{ij} \overset{d}{=} N(0,1)$ i.i.d.. Note also that, $(W^*_{ij})^2 - 1$ is a centered random variable, and has sub-exponential tail, see [281, Lemma 5.14]. Now, letting $Z_{ij} = (W^*_{ij})^2 - 1$, and applying the Bernstein-type inequality [281, Proposition 5.16], we have that for some absolute constants $K, c > 0$, it holds:

$$\mathbb{P}\left(\left|\sum_{i=1}^{m}\sum_{j=1}^{d} Z_{ij}\right| > d\sqrt{m}\right) \leq 2\exp\left(-c\min\left(\frac{d}{K^2}, \frac{d\sqrt{m}}{K}\right)\right) \leq 2\exp(-cd/K^2) = \exp(-\Omega(d)),$$

for m sufficiently large. In particular, with probability at least $1 - \exp(-\Omega(d))$, it therefore holds that,

$$\left|\sum_{k=1}^{d}(\lambda'_k - m)\right| \leq d\sqrt{m}.$$

Finally, using triangle inequality,

$$\left|\sum_{k=1}^{d}\lambda_k\right| = \left|\sum_{k=1}^{d}(\lambda'_k - (m+4d))\right| \leq \left|\sum_{k=1}^{d}(\lambda'_k - m)\right| + 4d^2 \leq d\sqrt{m} + 4d^2,$$

with probability $1 - \exp(-\Omega(d))$. After squaring, we obtain that $\left(\sum_{i=1}^{d}\lambda_i\right)^2 \leq 16d^4 + 8d^3\sqrt{m} + d^2 m$. In particular, we get:

$$\begin{aligned}\mathcal{L}(W_0) &\leq \mu_2^2\left(\sum_{i=1}^{d}\lambda_i\right)^2 + \max\left\{\text{Var}(X_i^2), 2\mathbb{E}\left[X_i^2\right]^2\right\}\left(\sum_{i=1}^{d}\lambda_i^2\right)\\ &\leq \mu_2^2(16d^4 + 8d^3\sqrt{m} + md^2) + \max\left\{\text{Var}(X_i^2), 2\mathbb{E}\left[X_i^2\right]^2\right\}16md^2.\end{aligned}$$

Using now the overparameterization $m > Cd^2$, we further have:

$$\mathbb{E}\left[X_i^2\right]^2(16d^4 + 8d^3\sqrt{m} + md^2) + \max\left\{\text{Var}(X_i^2), 2\mathbb{E}\left[X_i^2\right]^2\right\}16md^2 \leq \mathcal{C}'(C)m^2,$$

where

$$\mathcal{C}'(C) = \mathbb{E}\left[X_i^2\right]^2\left(\frac{16}{C^2} + \frac{8}{C^{3/2}} + \frac{1}{C}\right) + \frac{16}{C}\max\left\{\text{Var}(X_i^2), 2\mathbb{E}\left[X_i^2\right]^2\right\}.$$

Note that for the constant $\mathcal{C}'(C)$,

$$\mathcal{C}'(C) \to 0 \quad \text{as} \quad C \to \infty.$$

Next, observe that, $\sqrt{m} - 2\sqrt{d} \geq \frac{1}{2}\sqrt{m}$ for m large (in the regime $m > Cd^2$, with C large enough). Thus, using what we have established in Theorem 6.2.1, we arrive at:

$$\begin{aligned}\min_{W \in \mathbb{R}^{m\times d}: \mathrm{rank}(W) < d} \mathcal{L}(W) &> \min\left\{\mathrm{Var}(X_i^2), 2\mathbb{E}\left[X_i^2\right]^2\right\} \sigma_{\min}(W^*)^4 \\ &\geq \min\left\{\mathrm{Var}(X_i^2), 2\mathbb{E}\left[X_i^2\right]^2\right\} (\sqrt{m} - 2\sqrt{d})^4 \\ &\geq \frac{1}{16} \min\left\{\mathrm{Var}(X_i^2), 2\mathbb{E}\left[X_i^2\right]^2\right\} m^2.\end{aligned}$$

Finally, observe also that if $\mathrm{Var}(X_i^2) > 0$, then $\mathbb{E}\left[X_i^2\right] > 0$ as well: indeed observe that if $\mathbb{E}\left[X_i^2\right] = 0$, then $X_i = 0$ almost surely, for which $\mathrm{Var}(X_i^2) = 0$. In particular, $\min\left\{\mathrm{Var}(X_i^2), 2\mathbb{E}\left[X_i^2\right]^2\right\} > 0$. Equipped with this, we then observe that provided:

$$\frac{1}{16} \min\left\{\mathrm{Var}(X_i^2), 2\mathbb{E}\left[X_i^2\right]^2\right\} > \mathcal{C}'(C) = \mathbb{E}\left[X_i^2\right]^2 \left(\frac{16}{C^2} + \frac{8}{C^{3/2}} + \frac{1}{C}\right) + \frac{16}{C} \max\left\{\mathrm{Var}(X_i^2), 2\mathbb{E}\left[X_i^2\right]^2\right\},$$

that is, provided $C > 0$ is sufficiently large, we are done.

Part (b)

Note that, the result of Bai and Yin [25] asserts that if $\{\mu_1, \ldots, \mu_d\}$ are the eigenvalues of

$$\mathcal{A} \triangleq \frac{1}{2\sqrt{md}}((W^*)^T W^* - mI_d),$$

and if we define the empirical measure

$$F^{\mathcal{A}}(x) = \frac{1}{d}\left|\{i : \mu_i \leq x\}\right|$$

then in the regime $d \to +\infty$, $d/m \to 0$, it holds that:

$$F^{\mathcal{A}}(x) \to \omega(x),$$

almost surely, where $\omega(x)$ is the semicircle law; and moreover

$$\frac{1}{d}\sum_{i=1}^{d} \mu_i^2 \to \int x^2 \, d\omega(x) \triangleq \chi_2$$

namely, χ_2 is respectively the second moment under semicircle law, whp. Now, define the same quantities as in proof of part (a), where this time $W_0^T W_0 = mI_d$, and $\{\lambda_1, \ldots, \lambda_d\} = \sigma((W^*)^T W^* - mI_d)$. In particular, we still retain the inequality per Theorem 6.3.1 (c):

$$\mathcal{L}(W_0) \leq \mu_2^2 \left(\sum_{i=1}^{d} \lambda_i\right)^2 + \max\left\{\mathrm{Var}(X_i^2), 2\mathbb{E}\left[X_i^2\right]^2\right\} \left(\sum_{i=1}^{d} \lambda_i^2\right).$$

Note that $\lambda_i = 2\sqrt{md}\mu_i$. Hence, we obtain

$$\sum_{i=1}^{d} \lambda_i^2 < (4 + o(1))md^2\chi_2$$

whp. We now control $\sum_{i=1}^{d} \lambda_i$ using central limit theorem (CLT). Observe that,

$$\sum_{i=1}^{d} \lambda_i = \text{trace}((W^*)^T W^* - mI_d) = \sum_{i=1}^{m}\sum_{j=1}^{d}((W_{ij}^*)^2 - 1).$$

Now, note that

$$\sigma_*^2 \triangleq \text{Var}((W_{ij}^*)^2 - 1) = \text{Var}((W_{ij}^*)^2) < \mathbb{E}\left[(W_{ij}^*)^4\right] < \infty.$$

We now use CLT, as $d \to \infty$ and $m/d \to \infty$. To that end, let $1/2 > \epsilon > 0$ be fixed. Observe now that, for any arbitrary $M > 0$, and sufficiently large d,

$$\left\{-1 \le \frac{1}{\sigma_*\sqrt{md}d^\epsilon}\sum_{i=1}^{m}\sum_{j=1}^{d}((W_{ij}^*)^2 - 1) \le 1\right\} \supset \left\{-M \le \frac{1}{\sigma_*\sqrt{md}}\sum_{i=1}^{m}\sum_{j=1}^{d}((W_{ij}^*)^2 - 1) \le M\right\}.$$

In particular, using central limit theorem, we deduce

$$\liminf_{d\to\infty} \mathbb{P}\left(-1 \le \frac{1}{\sigma_*\sqrt{md}d^\epsilon}\sum_{i=1}^{m}\sum_{j=1}^{d}((W_{ij}^*)^2 - 1) \le 1\right) \ge \mathbb{P}(Z \in [-M, M]),$$

where Z is a standard normal random variable. Now since $M > 0$ is arbitrary, we have, by sending $M \to +\infty$, we obtain

$$\liminf_{d\to\infty} \mathbb{P}\left(-1 \le \frac{1}{\sigma_*\sqrt{md}d^\epsilon}\sum_{i=1}^{m}\sum_{j=1}^{d}((W_{ij}^*)^2 - 1) \le 1\right) \ge 1,$$

and we then conclude

$$\lim_{d\to\infty} \mathbb{P}\left(-1 \le \frac{1}{\sigma_*\sqrt{md}d^\epsilon}\sum_{i=1}^{m}\sum_{j=1}^{d}((W_{ij}^*)^2 - 1) \le 1\right) = 1.$$

Hence,

$$\left|\sum_{i=1}^{d} \lambda_i\right| \le \sigma_*\sqrt{md}d^\epsilon,$$

with probability $1 - o_d(1)$, for d sufficiently large.

Moreover,

$$\sigma_{\min}(W^*)^4 \ge \frac{1}{16}m^2,$$

for m large, using yet another result of Bai and Yin, see Theorem 6.3.8. From here, carrying the exact same analysis as in part (a) we obtain provided $m > Cd^2$ for some large constant $C > 0$, and d sufficiently large the following holds with probability $1 - o_d(1)$:

$$\mathcal{L}(W_0) < \min_{W \in \mathbb{R}^{m \times d} : \mathrm{rank}(W) < d} \mathcal{L}(W),$$

where W_0 is prescribed such that $W_0^T W_0 = m I_d$.

6.4.9 Proof of Theorem 6.2.2

First, let

$$\mathcal{S}_1 \triangleq \left\{ W \in \mathbb{R}^{m \times d} : \mathrm{rank}(W) < d, \widehat{\mathcal{L}}(W) < \frac{1}{2} C_5 \sigma_{\min}(W^*)^4 \right\}.$$

We start with the following claim.

Claim 6.4.2. *In the setting of Theorem 6.2.2 the following holds. With probability at least $1 - \exp(-C'd)$ (where $C' > 0$ is some absolute constant) it holds that for any $W \in \mathbb{R}^{m \times d}$ with $\widehat{\mathcal{L}}(W) \le \frac{1}{2} C_5 \sigma_{\min}(W^*)^4$,*

$$\|W\|_F \le d^{K+1},$$

provided $N \ge C''' d$ for some absolute constant $C''' > 0$.

Proof of Claim 6.4.2. For convenience, let $\widehat{\mathcal{L}}_0 \triangleq \frac{1}{2} C_5 \sigma_{\min}(W^*)^4$, and for the random data vector $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ let $\sigma^2 = \mathbb{E}[X_1^2]$. Recall that X has i.i.d. centered coordinates with a sub-Gaussian coordinate distribution.

We have the following, where the implication is due to Cauchy-Schwarz:

$$\widehat{\mathcal{L}}_0 \ge \frac{1}{N} \sum_{1 \le i \le N} (Y_i - f(X_i; W))^2 \Rightarrow (\widehat{\mathcal{L}}_0)^{1/2} \ge \left| \frac{1}{N} \sum_{1 \le i \le N} (Y_i - f(X_i; W)) \right|$$

We now establish that with probability at least $1 - 2\exp(-t^2 d)$, the following holds, provided $N \ge C(t/\epsilon)^2 d$: **for every** $W \in \mathbb{R}^{m \times d}$,

$$\left| \frac{1}{N} \sum_{1 \le i \le N} X_i^T W^T W X_i - \sigma^2 \|W\|_F^2 \right| \le \epsilon \sigma^2 \|W\|_F^2.$$

To see this, we begin by noticing $X_i^T W^T W X_i = \mathrm{trace}(X_i^T W^T W X_i) = \left\langle W^T W, X_i X_i^T \right\rangle$. Using this we have

$$\left| \frac{1}{N} \sum_{1 \le i \le N} X_i^T W^T W X_i - \sigma^2 \|W\|_F^2 \right| = \left| \left\langle W^T W, \frac{1}{N} \sum_{1 \le i \le N} X_i X_i^T - \sigma^2 I_d \right\rangle \right|.$$

We now use Hölder's inequality (Theorem 6.3.10) with $p = 1, q = \infty$, $U = W^T W$ and $V = \frac{1}{N}\sum_i X_i X_i^T - \sigma^2 I_d$. This yields

$$\left| \left\langle W^T W, \frac{1}{N} \sum_{1 \le i \le N} X_i X_i^T - \sigma^2 I_d \right\rangle \right| \le \|W\|_F^2 \left\| \frac{1}{N} \sum_{1 \le i \le N} X_i X_i^T - \sigma^2 I_d \right\|.$$

Observing now $\mathbb{E}[X_i X_i^T] = \sigma^2 I_d$, we have

$$\left\| \frac{1}{N} \sum_{1 \le i \le N} X_i X_i^T - \sigma^2 I_d \right\| \le \epsilon \sigma^2$$

with probability at least $1 - 2\exp(-t^2 d)$ provided $N \ge C(t/\epsilon)^2 d$, using the concentration result on sample covariance matrix from Vershynin [281, Corollary 5.50]. Hence, on this high probability event, the following holds:

$$\frac{1}{N} \sum_{1 \le i \le N} X_i^T (W^*) T W^* X_i \le \sigma^2 (1+\epsilon) \|W^*\|_F^2 \quad \text{and} \quad \frac{1}{N} \sum_{1 \le i \le N} X_i^T W^T W X_i \ge \sigma^2 (1-\epsilon) \|W\|_F^2$$

Hence,

$$\widehat{\mathcal{L}}_0 \ge \frac{1}{N} \sum_{1 \le i \le N} (X_i^T W^T W X_i - X_i^T (W^*)^T W^* X_i) \ge \sigma^2 (1-\epsilon) \|W\|_F^2 - \sigma^2 (1+\epsilon) \|W^*\|_F^2.$$

This yields, for any W with $\widehat{\mathcal{L}}(W) \le \widehat{\mathcal{L}}_0$,

$$\|W\|_F \le \left(\frac{(\widehat{\mathcal{L}}_0)^{1/2}}{\sigma^2 (1-\epsilon)} + \frac{1+\epsilon}{1-\epsilon} \|W^*\|_F^2 \right)^{1/2}$$

with probability at least $1 - 2\exp(-t^2 d)$. Now, observe that

$$\sigma_{\min}(W^*)^2 = \lambda_{\min}((W^*)^T W^*) \le \operatorname{trace}((W^*)^T W^*) \le \|W^*\|_F^2 \le d^{2K}.$$

Furthermore, $C_5 = O(1)$. This yields

$$\widehat{\mathcal{L}}_0 = \frac{1}{2} C_5 \sigma_{\min}(W^*)^4 = O(d^{4K}). \tag{6.12}$$

We now take $\epsilon = 1/2$ above, and conclude that

$$\|W\|_F \le \left(\frac{(\widehat{\mathcal{L}}_0)^{1/2}}{\sigma^2 (1-\epsilon)} + \frac{1+\epsilon}{1-\epsilon} \|W^*\|_F^2 \right)^{1/2} \le d^{K+1}$$

for d large enough; with probability at least $1 - 2\exp(-t^2 d)$, which is $1 - \exp(-C' d)$ for some absolute constant $C' > 0$. □

Having established Claim 6.4.2, we now return to the proof of Theorem 6.2.2. Let

$$\mathcal{S}_2 \triangleq \left\{ W \in \mathbb{R}^{m\times d} : \operatorname{rank}(W) < d, \widehat{\mathcal{L}}(W) < \frac{1}{2} C_5 \sigma_{\min}(W^*)^4, \|W\|_F \leq d^{K+1} \right\}.$$

A consequence of Claim 6.4.2 is that $\mathbb{P}(\mathcal{S}_1 = \mathcal{S}_2) \geq 1 - \exp(-C'd)$. We now establish

Claim 6.4.3.

$$\mathbb{P}(\mathcal{S}_2 = \varnothing) \geq 1 - \left(9d^{2+4K+7}\right)^{d^2-1} \cdot \exp\left(-C_3 N d^{-4-4K}\right) - Nde^{-Cd}.$$

Note that combining Claims 6.4.2 and 6.4.3 through a union bound yields

$$\inf_{W\in\mathbb{R}^{m\times d}:\operatorname{rank}(W)<d} \widehat{\mathcal{L}}(W) \geq \frac{1}{2} C_5 \sigma_{\min}(W^*)^4,$$

with probability at least

$$1 - \exp(-C'd) - \left(9d^{9+4K}\right)^{d^2-1} \exp\left(-C_3 N d^{-4-4K}\right) - Nde^{-Cd},$$

therefore establishing Theorem 6.2.2.

Proof of Claim 6.4.3. Let $A = W^T W \in \mathbb{R}^{d\times d}$. We claim $\|A\|_F \leq d^{2K+2}$. To see this, note that $\|A\|_F^2 = \operatorname{trace}(A^T A) = \operatorname{trace}(A^2)$. Let $\theta_1, \ldots, \theta_d$ be the eigenvalues of A, all non-negative as $A \succeq 0$; and $\theta_1^2, \ldots, \theta_d^2$ are the eigenvalues of A^2. With this,

$$\operatorname{trace}(A^2) = \sum_{1\leq i\leq d} \theta_i^2 \leq \left(\sum_{1\leq i\leq d} \lambda_i \right)^2 = \operatorname{trace}(A)^2.$$

Hence, $\|A\|_F \leq \operatorname{trace}(A) = \|W\|_F^2 \leq d^{2K+2}$, as requested.

Next, let

$$S_R = \left\{ A \in \mathbb{R}^{d\times d} : \operatorname{rank}(A) \leq d-1, A \succeq 0, \|A\|_F \leq R \right\};$$

and let $\bar{S}_\epsilon$ be an $\epsilon-$net for $S_{d^{2K+2}}$ in Frobenius norm, where ϵ to be tuned appropriately later. Using Lemma 6.3.4 we have

$$\left|\bar{S}_\epsilon\right| \leq \left(\frac{9d^{2K+2}}{\epsilon} \right)^{d^2-1}.$$

Now, applying Lemma 6.3.3 with $K_1 = \frac{1}{2}$ and $K_2 = K$ and taking a union bound

across the net $\bar{S}_\epsilon$, we arrive at the following conclusion:

$$\mathbb{P}\left(\bigcup_{A\in\bar{S}_\epsilon}\underbrace{\left\{\frac{1}{N}\sum_{1\le i\le N}\Big(Y_i-X_i^TAX_i\Big)^2<\frac{1}{2}C_5\sigma_{\min}(W^*)^4\right\}}_{\mathcal{E}(A)^c \text{ from Lemma 6.3.3}}\Big|\|X_i\|_\infty\le\sqrt{d},1\le i\le N\right)$$
$$\le\left(\frac{9d^{2K+2}}{\epsilon}\right)^{d^2-1}\exp\Big(-C_3Nd^{-4+4K}\Big),$$

where

$$C_5=\min\Big\{\mu_4(1/2)-\mu_2(1/2)^2,2\mu_2(1/2)^2\Big\}\quad\text{and}\quad\mu_n(K)=\mathbb{E}\Big[X_i^n\Big||X_i|\le d^K\Big].$$

Now, since $\mathbb{P}\Big(\|X_i\|_\infty<\sqrt{d},1\le i\le N\Big)\ge 1-Nd\exp(-Cd)$ by Lemma 6.3.2, we obtain

$$\mathbb{P}\left(\bigcap_{A\in\bar{S}_\epsilon}\left\{\frac{1}{N}\sum_{1\le i\le N}\Big(Y_i-X_i^TAX_i\Big)^2\ge\frac{1}{2}C_5\sigma_{\min}(W^*)^4\right\}\right)$$
$$\ge 1-\left(\frac{9d^{2K+2}}{\epsilon}\right)^{d^2-1}\cdot\exp\left(-C_3\frac{N}{d^{4+4K}}\right)-Nd\exp\Big(-Cd\Big).$$

In the remainder of the proof, suppose for every $A\in\bar{S}_\epsilon$,

$$\frac{1}{N}\sum_{1\le i\le N}(Y_i-X_i^TAX_i)^2\ge\frac{1}{2}C_5\sigma_{\min}(W^*)^4,$$

and $\|X_i\|_\infty\le d^{1/2}$, $i\in[N]$, which collectively hold with probability at least

$$1-\left(\frac{9d^{2K+2}}{\epsilon}\right)^{d^2-1}\cdot\exp\left(-C_3\frac{N}{d^{4+4K}}\right)-2Nd\exp(-Cd).$$

Now, let $W\in\mathbb{R}^{m\times d}$ with $\|W\|_F\le d^{K+1}$, $\operatorname{rank}(W)\le d-1$. Let $A=W^TW$ (thus $\|A\|_F\le d^{2K+2}$) and $\widehat{A}\in\bar{S}_\epsilon$ be such that $\|A-\widehat{A}\|_F\le\epsilon$. We now estimate

$$\Delta\triangleq\left|\frac{1}{N}\sum_{1\le i\le N}(Y_i-X_i^TAX_i)^2-\frac{1}{N}\sum_{1\le i\le N}(Y_i-X_i^T\widehat{A}X_i)^2\right|.$$

For notational convenience, let $A^* = (W^*)^T W^*$. Now

$$\begin{aligned}\Delta &\leq \frac{1}{N}\sum_{1\leq i\leq N}\left|(X_i^T(A-A^*)X_i)^2 - (X_i^T(\widehat{A}-A^*)X_i)^2\right| \\ &= \frac{1}{N}\sum_{1\leq i\leq N}\left|X_i^T(A-\widehat{A})X_i\right|\cdot\left|X_i^T(A+\widehat{A}-2A^*)X_i\right|.\end{aligned}$$

Now, using Cauchy-Schwarz (for inner product $\langle M, N\rangle \triangleq \text{trace}(M^T N)$)

$$|X_i^T(A-\widehat{A})X_i| = |\left\langle A-\widehat{A}, X_i X_i^T\right\rangle| \leq \|A-\widehat{A}\|_F \cdot \|X_i\|_2^2,$$

using $\|X_i X_i^T\|_F = \|X_i\|_2^2$. In particular, we obtain

$$|X_i^T(A-\widehat{A})X_i| \leq \epsilon d^2.$$

For the term $|X_i^T(A+\widehat{A}-2A^*)X_i|$, we observe that triangle inequality yields

$$\|A+\widehat{A}-2A^*\|_F \leq 4d^{2K+2}.$$

Thus

$$\left|X_i^T(A+\widehat{A}-2A^*)X_i\right| \leq 4d^{2K+4}.$$

Using these, we obtain

$$\left|\widehat{\mathcal{L}}(W) - \frac{1}{N}\sum_{1\leq i\leq N}(Y_i - X_i^T\widehat{A}X_i)^2\right| \leq 4\epsilon d^{6+2K} = O(d^{-1}) = o_d(1),$$

taking $\epsilon = d^{-7-2K}$. Using finally the fact that

$$\frac{1}{N}\sum_{1\leq i\leq N}(Y_i - X_i^T\widehat{A}X_i)^2$$

is bounded away from zero across the net $\bar{S}_\epsilon$, we conclude the proof of Claim 6.4.3. □

Since it was already noted that Claims 6.4.2 and 6.4.3 together yield Theorem 6.2.2, we complete the proof of Theorem 6.2.2.

Case of Constant d: $d = O(1)$.

We now carry out a separate analysis for the case of constant d ($d = O(1)$). We only point out the necessary modifications, while hiding factors depending on the constant d under asymptotic notations.

In what follows, we use the fact that if X is a sub-Gaussian random variable; then $\mathbb{E}\left[|X|^p\right] < \infty$ for every $p \geq 1$. For a proof, see [281, Lemma 5.5]; which establishes a stronger conclusion that $\mathbb{E}\left[|X|^p\right]^{1/p} = O\left(\sqrt{p}\right)$ for every $p \geq 1$.

Modifying Claim 6.4.2. Claim 6.4.2 modifies to the following: with probability at least $1 - O(1/N)$, it holds that for any $W \in \mathbb{R}^{m \times d}$ with $\widehat{\mathcal{L}}(W) \le \frac{1}{2}\overline{C_5}\sigma_{\min}(W^*)^4$,

$$\|W\|_F = O(1).$$

We now sketch the proof of this modified claim. For a matrix $M \in \mathbb{R}^{d \times d}$, denote by $\|M\|_\infty := \max_{1 \le i,j \le d} |M_{ij}|$; and let $\epsilon > 0$ be arbitrary. Then, we show that over the randomness in $X_i \in \mathbb{R}^d$, $1 \le i \le N$, with probability at least $1 - O(1/N)$,

$$\left\| \frac{1}{N} \sum_{1 \le i \le N} X_i X_i^T - \sigma^2 I_d \right\|_\infty \le \epsilon.$$

Indeed, fix an $\epsilon > 0$. Then for any $1 \le j \le d$,

$$\mathbb{P}\left(\left| \frac{1}{N} \sum_{1 \le i \le N} X_i(j)^2 - \sigma^2 \right| \le \epsilon \right) \ge 1 - O(1/N).$$

by Chebyshev's inequality (here, $X_i = (X_i(j) : 1 \le j \le d) \in \mathbb{R}^d$). Here, we used in particular the fact $\mathbb{E}[X_i(j)^4] < \infty$. Likewise, for any $1 \le j < j' \le d$,

$$\mathbb{P}\left(\left| \frac{1}{N} \sum_{1 \le i \le N} X_i(j) X_i(j') \right| \le \epsilon \right) \ge 1 - O(1/N).$$

Taking a union bound over these $d(d+1)/2$ events corresponding to the entries of matrix $N^{-1}\sum_{1 \le i \le N} X_i X_i^T$, we are done. (Note that the number of events, $d(d+1)/2$, is $O(1)$. This, and other factors depending on ϵ are hidden under $O(\cdot)$.) Using the trivial bound $\|M\| \le \|M\|_F^2$ valid for any matrix $M \in \mathbb{R}^{d \times d}$, we arrive at the operator norm bound

$$\left\| \frac{1}{N} \sum_{1 \le i \le N} X_i X_i^T - \sigma^2 I_d \right\| \le \epsilon^2 d^2.$$

Finally, taking $\epsilon = 1/(2d^2) = O(1)$, we finish the proof of modified Claim 6.4.2.

Modifying Claim 6.4.3. Claim 6.4.3 now modifies to $\mathbb{P}(\mathcal{S}_2 = \varnothing) \ge 1 - O(1/N)$; and this modified version is shown as follows. Note that for any $\epsilon = O(1)$, the size of the "net" we consider is $O(1)$. We claim that with probability $1 - O(1/N)$, it holds that for any $A \in \bar{S}_\epsilon$,

$$\frac{1}{N} \sum_{1 \le i \le N} \Big(Y_i - X_i^T A X_i \Big)^2 \ge \frac{2}{3}\overline{C_5}\sigma_{\min}(W^*)^4,$$

where

$$\overline{C_5} = \min\Big\{ \mu_4 - \mu_2^2, 2\mu_2^2 \Big\} \quad \text{and} \quad \mu_n = \mathbb{E}\Big[X_1(1)^n \Big].$$

To show this, fix an $A \in \bar{S}_\epsilon$. Now, instead of Lemma 6.3.3; one can apply Chebyshev's inequality:

$$\frac{1}{N}\sum_{1\leq i\leq N}\left(X_i^T AX_i - X_i^T\left(W^*\right)^T W^* X_i\right)^2 \geq \frac{2}{3}\mathbb{E}\Big[\left(X^T AX - X^T\left(W^*\right)^T W^* X\right)^2\Big],$$

with probability at least $1-O(1/N)$. Here, we in particular used the fact $\mathbb{E}[X_i(j)^8] < \infty$. Since

$$\mathbb{E}\Big[\left(X^T AX - X^T\left(W^*\right)^T W^* X\right)^2\Big] \geq \overline{C_5}\sigma_{\min}(W^*)^4$$

by Theorem 6.2.1, we establish the claim by taking a union bound over the net $\bar{S}_\epsilon$ which has $O(1)$ cardinality. The rest of the argument for Claim 6.4.3 remains (nearly) intact. In particular, the bound $\|A-\widehat{A}\|_F \leq \epsilon$ remains intact; and $\|A+\widehat{A}-2A^*\|_F$ is now $O(1)$. Finally, keeping in mind that $\frac{1}{N}\sum_{1\leq i\leq N}\|X_i\|_2^4 = O(1)$ with probability $1-O(1/N)$, we complete the proof by taking ϵ small enough.

Putting these together like in the proof of Theorem 6.2.2, we complete the proof.

6.4.10 Proof of Theorem 6.2.4

We start by computing $\nabla\widehat{\mathcal{L}}(W)$. Taking derivatives with respect to j^{th} row W_j of $W \in \mathbb{R}^{m\times d}$, we arrive at

$$\nabla_{W_j}\widehat{\mathcal{L}}(W) = \frac{4}{N}\sum_{1\leq i\leq N}\left(\sum_{1\leq j\leq m}\langle W_j, X_i\rangle^2 - Y_i\right)\langle W_j, X_i\rangle X_i.$$

Interpreting these gradients as a row vector and aggregating into a matrix, we then have

$$\nabla_W\widehat{\mathcal{L}}(W) = W\left(\frac{4}{N}\sum_{1\leq i\leq N}\left(\sum_{1\leq j\leq m}\langle W_j, X_i\rangle^2 - Y_i\right)X_i X_i^T\right).$$

Assume now that $\text{rank}(W) = d$, and $\nabla\widehat{\mathcal{L}}(W) = 0$. We then arrive at

$$\frac{1}{N}\sum_{1\leq i\leq N}\left(\sum_{1\leq j\leq m}\langle W_j, X_i\rangle^2 - Y_i\right)X_i X_i^T = 0.$$

We now claim that $\widehat{\mathcal{L}}(W) = 0$. To see this, we take a route similar to [264, Lemma 6.1]. Let $M \triangleq W^T W$, and consider the function

$$f(M) \triangleq \frac{1}{N}\sum_{1\leq i\leq N}(Y_i - X_i^T M X_i)^2.$$

Observe that $f(\cdot)$ is quadratic in M. Thus, any $\widehat{M}$ with $\nabla f(\widehat{M}) = 0$, that is

$$\frac{1}{N}\sum_{1\le i\le N}(X_i^T\widehat{M}X_i - Y_i)X_iX_i^T = 0$$

it is the case that $\widehat{M}$ is a global optimum of f. In particular for any $M \in \mathbb{R}^{d\times d}$, $f(M) \ge f(\widehat{M})$. Now, take any $\bar{W} \in \mathbb{R}^{m\times d}$, and observe that $\widehat{\mathcal{L}}(\bar{W}) = f(\bar{W}^T\bar{W})$. Since $\nabla f(W^TW) = 0$, it follows that

$$\widehat{\mathcal{L}}(\bar{W}) = f(\bar{W}^T\bar{W}) \ge f(W^TW) = \widehat{\mathcal{L}}(W).$$

Namely, W is indeed a global optimizer of $\widehat{\mathcal{L}}(\cdot)$. Since $W = W^*$ makes the cost zero, we obtain $\widehat{\mathcal{L}}(W) = 0$.

Now, using Theorem 6.2.10, we obtain that $\text{span}(X_iX_i^T : 1 \le i \le N)$ is the set of all $d \times d$ symmetric matrices; with probability one, provided $N \ge d(d+1)/2$. In this case, using Theorem 6.2.9, we conclude that $W^TW = (W^*)^TW^*$, concluding the proof.

6.4.11 Proof of Theorem 6.2.6

Part (a)

Note that by Claim 6.4.2, it follows that with probability at least $1-\exp(-C'd)$, it is the case that for any W with $\widehat{\mathcal{L}}(W) \le \widehat{\mathcal{L}}(W_0) < \frac{1}{2}C_5\sigma_{\min}(W^*)^4$, $\|W\|_F \le d^{K+1}$. Now let

$$\mathcal{E}_1 \triangleq \left\{\sup_{W:\widehat{\mathcal{L}}(W)\le\widehat{\mathcal{L}}_0}\|W\|_F \le d^{K+1}\right\} \tag{6.13}$$

thus $\mathbb{P}(\mathcal{E}_1) \ge 1-\exp(-C'd)$ and

$$\mathcal{E}_2 \triangleq \Big\{\|X_i\|_\infty \le d^{1/2}, 1 \le i \le N\Big\}, \tag{6.14}$$

such that $\mathbb{P}(\mathcal{E}_2) \ge 1 - Nd\exp(-Cd)$ per Lemma 6.3.2.

Note that the $\|\nabla^2\widehat{\mathcal{L}}(W)\| = \text{poly}(\|W\|_F, \|X_1\|, \ldots, \|X_N\|)$. Thus on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, which holds with probability at least $1 - Nd\exp(-Cd) - \exp(-C'd)$, we have that

$$L = \sup\Big\{\|\nabla^2\widehat{\mathcal{L}}(W)\| : \widehat{\mathcal{L}}(W) \le \widehat{\mathcal{L}}_0\Big\} = \text{poly}(d) < +\infty$$

as claimed.

Part (b)

Suppose that the event $\mathcal{E}_1 \cap \mathcal{E}_2$ (where $\mathcal{E}_1$ and $\mathcal{E}_2$ are defined respectively in (6.13) and (6.14)) takes place. We run the gradient descent with a step size of $\eta < 1/2L$: a

second order Taylor expansion reveals that

$$\widehat{\mathcal{L}}(W_1) - \widehat{\mathcal{L}}(W_0) \leq -\eta \|\nabla \widehat{\mathcal{L}}(W_0)\|_F^2/2$$

where $\nabla \widehat{\mathcal{L}}(W)$ is the gradient of the empirical risk evaluated at W. In particular, $\widehat{\mathcal{L}}(W_1) \leq \widehat{\mathcal{L}}(W_0)$. Since $\mathcal{E}_1$ takes place, we conclude $\|\nabla^2 \widehat{\mathcal{L}}(W_1)\| \leq L = \text{poly}(d)$, where $\|\nabla^2 \widehat{\mathcal{L}}(W)\|$ is the spectral norm of the Hessian matrix $\nabla^2 \widehat{\mathcal{L}}(W)$. From here, we induct on k: induction argument reveals that we can retain a step size of $\eta < 1/2L$ (thus $\eta = \text{poly}(d)$), and furthermore along the trajectory $\{W_k\}_{k \geq 0}$, it holds:

$$\widehat{\mathcal{L}}(W_{k+1}) - \widehat{\mathcal{L}}(W_k) \leq -\eta \|\nabla \widehat{\mathcal{L}}(W_k)\|_F^2/2.$$

Now let T be the first time for which $\|\nabla \widehat{\mathcal{L}}(W)\|_F \leq \zeta$, namely the horizon required to arrive at an $\zeta-$stationary point. In what follows, we carry out our analysis in terms of ζ. At the end, we incorporate the bound (6.4) on ζ.

We claim $T = \text{poly}\left(\zeta^{-1}, d, \sigma_{\min}(W^*)^{-1}\right)$.

To see this, note that from the definition of T, it holds that $\|\nabla \widehat{\mathcal{L}}(W_t)\|_F \geq \zeta$ as $t \leq T-1$. Now, a telescoping argument together with $\eta = 1/\text{poly}(d)$ reveals

$$\widehat{\mathcal{L}}(W_T) - \widehat{\mathcal{L}}(W_0) \leq -T(\text{poly}(d))^{-1}\zeta^2.$$

Using now $\widehat{\mathcal{L}}(W_T) \geq 0$, we conclude $\widehat{\mathcal{L}}(W_0) \geq T\zeta^2 \text{poly}(d)$. Since $\widehat{\mathcal{L}}(W_0) = \widehat{\mathcal{L}}_0$ is at most polynomial in d as per (6.12), we conclude $T = \text{poly}(\zeta^{-1}, d)$.

We now turn our attention to bounding its risk. Let $r_i \triangleq Y_i - X_i^T W^T W X_i$. Note that $\widehat{\mathcal{L}}(W) = \frac{1}{N}\sum_{1 \leq i \leq N} r_i^2$. Now,

$$\begin{aligned}
\widehat{\mathcal{L}}(W) &= \frac{1}{N}\sum_{1 \leq i \leq N} r_i (X_i^T (W^*)^T W^* X_i - X_i^T W^T W X_i) \\
&= \left\langle W^T W - (W^*)^T W^*, \frac{1}{N}\sum_{1 \leq i \leq N} r_i X_i X_i^T \right\rangle.
\end{aligned}$$

Using Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\widehat{\mathcal{L}}(W) &= \left| \left\langle W^T W - (W^*)^T W^*, \frac{1}{N}\sum_{1 \leq i \leq N} r_i X_i X_i^T \right\rangle \right| \\
&\leq \left\| W^T W - (W^*)^T W^* \right\|_F \cdot \left\| \frac{1}{N}\sum_{1 \leq i \leq N} r_i X_i X_i^T \right\|_F.
\end{aligned}$$

Next, $\|W^T W\|_F^2 = \text{trace}((W^T W)^2) \leq (\text{trace}(W^T W))^2 = \|W\|_F^4$, using the fact that $W^T W \succeq 0$. In particular, on the event $\mathcal{E}_1$ defined as per (6.13), we conclude that $\|W\|_F \leq d^{K+1}$, and therefore $\|W^T W\|_F \leq d^3$. This, together with $\|W^*\|_F \leq d^K$ and

triangle inequality then yields

$$\left\|W^TW - (W^*)^TW^*\right\|_F \leq 2d^{2K+2},$$

with probability at least $1 - \exp(-C'd)$. Hence, on this event

$$\widehat{\mathcal{L}}(W) \leq 2d^{2K+2}\left\|\frac{1}{N}\sum_{1\leq i\leq N} r_iX_iX_i^T\right\|_F. \tag{6.15}$$

With this, we now turn our attention to bounding

$$\left\|\frac{1}{N}\sum_{1\leq i\leq N} r_iX_iX_i^T\right\|_F.$$

We establish that for the event

$$\mathcal{E}_3 \triangleq \left\{\inf_{\substack{W\in\mathbb{R}^{m\times d}:\sigma_{\min}(W)<\frac{1}{2}\sigma_{\min}(W^*)\\ \|W\|_F\leq d^{K+1}}} \widehat{\mathcal{L}}(W) \geq \frac{1}{2}C_5\sigma_{\min}(W^*)^4\right\}, \tag{6.16}$$

it is the case that

$$\mathbb{P}(\mathcal{E}_3) \geq 1 - \left(9d^{4K+9}\right)^{d^2-1}\cdot\exp\left(-C_4Nd^{-4-4K}\right) - Nd\exp(-Cd). \tag{6.17}$$

This is almost a straightforward modification of the proof of earlier energy barrier result Theorem 6.2.2, and we only point out required modifications. Take any $W \in \mathbb{R}^{m\times d}$ with $\sigma_{\min}(W) < \frac{1}{2}\sigma_{\min}(W^*)$. In particular,

$$\lambda_{\min}(W^TW) = \sigma_{\min}(W)^2 < \frac{1}{4}\sigma_{\min}(W^*)^2.$$

Inspecting now the proof of Theorem 6.2.1(a), we obtain that for such a W,

$$\mathbb{E}\Big[(X^TW^TWX - X^T(W^*)^TW^*X)^2\Big|\|X\|_\infty \leq d^{1/2}\Big] \geq \frac{3}{4}C_5\sigma_{\min}(W^*)^4,$$

and consequently, modifying Lemma 6.3.3, we have that

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{N}\sum_{1\leq i\leq N}(Y_i - X_i^TW^TWX_i)^2 \geq \frac{1}{2}C_5\sigma_{\min}(W^*)^4\Big|\|X_i\|_\infty \leq \sqrt{d}, 1\leq i\leq N\right)\\ &\geq 1 - \exp\Big(-C'Nd^{-4-4K}\Big).\end{aligned}$$

Using now a covering numbers bound, in an exact same manner as in the proof of

Theorem 6.2.2, we conclude that

$$\inf_{\substack{W\in\mathbb{R}^{m\times d}:\sigma_{\min}(W)<\frac{1}{2}\sigma_{\min}(W^*)\\ \|W\|_F\leq d^{K+1}}}\widehat{\mathcal{L}}(W)\geq\frac{1}{2}C_5\sigma_{\min}(W^*)^4$$

with probability at least

$$1-(9d^{4K+9})^{d^2-1}\cdot\exp\left(-C_4Nd^{-4-4K}\right)-Nd\exp(-Cd).$$

Now suppose in the remainder of this part that the event $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3$ which is

$$\left\{\sup_{W:\widehat{\mathcal{L}}(W)\leq\widehat{\mathcal{L}}_0}\|W\|_F\leq d^{K+1}\right\}\bigcap\left\{\|X_i\|_\infty\leq d^{1/2},1\leq i\leq N\right\}$$
$$\bigcap\left\{\inf_{\substack{W\in\mathbb{R}^{m\times d}:\sigma_{\min}(W)<\frac{1}{2}\sigma_{\min}(W^*)\\ \|W\|_F\leq d^{K+1}}}\widehat{\mathcal{L}}(W)\geq\frac{1}{2}C_5\sigma_{\min}(W^*)^4\right\}$$

holds true. In particular, for any W with risk less than $\frac{1}{2}C_5\sigma_{\min}(W^*)^4$, we have $\sigma_{\min}(W)>\frac{1}{2}\sigma_{\min}(W^*)>0$ (in particular, any such W is invertible). Now, take any $\zeta-$stationary point W generated by the gradient descent. Due to the event $\mathcal{E}_3$, and the fact $\widehat{\mathcal{L}}(W)<\widehat{\mathcal{L}}_0$ proven earlier; it holds that $\text{rank}(W)=d$, and from the definition of $\zeta-$stationarity, we have

$$\|\nabla\widehat{\mathcal{L}}(W)\|_F\leq\zeta.$$

Inspecting the proof of Theorem 6.2.4, we observe that

$$\nabla\widehat{\mathcal{L}}(W)=4W\left(\frac{1}{N}\sum_{1\leq i\leq N}r_iX_iX_i^T\right).$$

Thus we arrive at

$$\left\|W\left(\frac{1}{N}\sum_{1\leq i\leq N}r_iX_iX_i^T\right)\right\|_F\leq 4\zeta.$$

Let

$$B\triangleq W\left(\frac{1}{N}\sum_{1\leq i\leq N}r_iX_iX_i^T\right).$$

Note now that

$$\frac{1}{N}\sum_{1\leq i\leq N}r_iX_iX_i^T=(W^TW)^{-1}W^TB.$$

Next, we have

$$\|(W^TW)^{-1}\|_2=\frac{1}{\sigma_{\min}(W^TW)}=\frac{1}{\sigma_{\min}(W)^2}<\frac{4}{\sigma_{\min}(W^*)^2},$$

due to conditioning on $\mathcal{E}_3$ (6.16) above. Furthermore,

$$\|W^T\|_2 = \|W\|_2 = \sqrt{\lambda_{\max}(W^T W)} \leq \sqrt{\mathrm{trace}(W^T W)} = \|W\|_F \leq d^{K+1}.$$

We now combine these finding.

$$\begin{aligned}\left\| \frac{1}{N} \sum_{1 \leq i \leq N} r_i X_i X_i^T \right\|_F &= \|(W^T W)^{-1} W^T B\|_F \\ &\leq \|(W^T W)^{-1}\|_2 \|W^T B\|_F \\ &\leq \|(W^T W)^{-1}\|_2 \|W^T\|_2 \|B\|_F \\ &\leq 16\zeta \sigma_{\min}(W^*)^{-2} d^{K+1}.\end{aligned}$$

We now use the bounds on $\mathbb{P}(\mathcal{E}_1)$ as per (6.13), on $\mathbb{P}(\mathcal{E}_2)$ as per (6.14), and on $\mathbb{P}(\mathcal{E}_3)$ as per (6.17); to control $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3)$. We conclude by the union bound that with probability at least

$$1 - \exp(-C'd) - (9d^{4K+9})^{d^2-1} \cdot \exp\left(-C_4 N d^{-4-4K}\right) - Nd\exp(-Cd),$$

it holds that for any W with $\|\nabla \widehat{\mathcal{L}}(W)\|_F \leq \zeta$, its empirical risk is controlled as per (6.15):

$$\widehat{\mathcal{L}}(W) \leq 32\zeta \sigma_{\min}(W^*)^{-2} d^{4K+4}. \tag{6.18}$$

Finally, since

$$\zeta \leq \frac{\epsilon}{32} \sigma_{\min}(W^*)^2 d^{-4K-4}$$

per (6.4), we deduce $\widehat{\mathcal{L}}(W) \leq \epsilon$, as claimed. The running time is polynomial in ζ^{-1} and d; and therefore is polynomial in ϵ^{-1}, $\sigma_{\min}(W^*)^{-1}$; and d. This completes the proof of Part (b).

Part (c)

Let $W \in \mathbb{R}^{m \times d}$ be such that $\widehat{\mathcal{L}}(W) \leq \kappa$. Define the matrix

$$M \triangleq W^T W - (W^*)^T W^*.$$

We will bound $\|M\|_F$, which will ensure weights $W^T W$ are uniformly close to ground truth weights defined $(W^*)^T W^*$. We start by conditioning: assume in the remainder that the event $\mathcal{E}_2$ in (6.14) stating $\|X_i\|_\infty \leq d^{1/2}$, for every $i \in [N]$ is true: this holds with probability at least $1 - Nd\exp(-Cd)$, as per Lemma 6.3.2.

Note that

$$\widehat{\mathcal{L}}(W) = \frac{1}{N} \sum_{1 \leq i \leq N} (X_i^T M X_i)^2.$$

To this end, consider a matrix $\Xi \in \mathbb{R}^{N \times d(d+1)/2}$, consisting of i.i.d. rows where i^{th} row

of Ξ is $\mathcal{R}_i \triangleq (X_i(1)^2, \ldots, X_i(d)^2, X_i(k)X_i(\ell) : 1 \leq k < \ell \leq d) \in \mathbb{R}^{d(d+1)/2}$. Next, let

$$\Sigma = \mathbb{E}[\mathcal{R}_i \mathcal{R}_i^T] \in \mathbb{R}^{\frac{d(d+1)}{2} \times \frac{d(d+1)}{2}},$$

where $\mathcal{R}_i$ is the i^{th} row of matrix Ξ. Furthermore, let $\mathcal{M} \in \mathbb{R}^{d(d+1)/2}$ be a vector consisting of entries $M_{11}, \ldots, M_{dd}$; and $2M_{ij}$, $1 \leq i < j \leq d$. With this notation, if $v = \Xi\mathcal{M} \in \mathbb{R}^{N \times 1}$, then we have

$$\widehat{\mathcal{L}}(W) = \|v\|_2^2/N \Rightarrow \|v\|_2^2 \leq N\kappa,$$

since $\widehat{\mathcal{L}}(W) \leq \kappa$ by assumption.

Next, we have

$$\mathcal{M} = (\Xi^T\Xi)^{-1}\Xi^T v \Rightarrow \|\mathcal{M}\|_2^2 \leq \|(\Xi^T\Xi)^{-1}\|_2^2 \|\Xi^T v\|_2^2. \tag{6.19}$$

We start with the second term. Recall that $\|v\|_2 \leq \sqrt{N\kappa}$, and we condition on $\|X_i\|_\infty < d^{1/2}$, $1 \leq i \leq N$. Next, using Cauchy-Schwarz inequality,

$$\left|(\Xi^T v)_i\right| \leq \|v\|_2 \sqrt{Nd} \leq N d^{1/2}\sqrt{\kappa}. \tag{6.20}$$

Hence,

$$\|\Xi^T v\|_2^2 \leq N^2 d^3 \kappa. \tag{6.21}$$

We now control $\|(\Xi^T\Xi)^{-1}\|_2^2$. This is done in a manner similar to the proof of [107, Theorem 3.2]. The main tool is the result Theorem 6.3.9 for concentration of the spectrum of random matrices with i.i.d. non-isotropic rows. The parameter setting we operate under is provided below.

Parameter	Value
m	d^2
t	$N^{1/8}$
δ	$N^{-3/8}d$
γ	$\max(\|\Sigma\|^{1/2}\delta, \delta^2)$

Start by verifying that since we condition on $\|X_i\|_\infty < d^{1/2}$, it is indeed the case that ℓ_2-norm of each row of Ξ is at most d, thus the value of m above works.

We now claim $\gamma = \|\Sigma\|^{1/2}\delta$. To prove this it suffices to show

$$N > \|\Sigma\|^{-4/3} d^{\frac{8}{3}}.$$

Using [107, Theorem 5.1] (also see Remark 6.4.4 below) with $k = 2$, we obtain $\sigma_{\min}(\Sigma) \geq cd^{-4}$, for some absolute constant $c > 0$ depending only on the data coordinate distribution. Consequently,

$$\|\Sigma\|^{-4/3} \leq \sigma_{\min}(\Sigma)^{-4/3} \leq c^{-4/3} d^{16/3} \Rightarrow \|\Sigma\|^{-4/3} d^{\frac{8}{3}} < c^{-4/3} d^8,$$

which is below sample size N, as requested. Therefore, $\gamma = \|\Sigma\|^{1/2}\delta$.

We now claim

$$\frac{1}{2}\sigma_{\min}(\Sigma) > \gamma = \|\Sigma\|^{1/2} N^{-\frac{3}{8}} d.$$

This is equivalent to establishing

$$N > 2^{8/3} \frac{\|\Sigma\|^{4/3} d^{\frac{8}{3}}}{\sigma_{\min}(\Sigma)^{8/3}}.$$

Using again [107, Theorem 5.1], we have $\|\Sigma\| < f d^4$ for some absolute constant $f > 0$. This yields

$$2^{8/3} \frac{\|\Sigma\|^{4/3} d^{\frac{8}{3}}}{\sigma_{\min}(\Sigma)^{8/3}} < C' d^{\frac{56}{3}}$$

for some absolute constant $C' > 0$, which again holds for our case as $N > d^{18+\frac{4}{3}}$.

The rest is verbatim from [107, p.45]: we now apply Theorem 6.3.9. With probability at least $1 - d^2 \exp(-cN^{1/4})$ (here $c > 0$ is an absolute constant), it holds that:

$$\left\| \frac{1}{N} \Xi^T \Xi - \Sigma \right\| \leq \gamma. \tag{6.22}$$

Now, for $D = d(d+1)/2$:

$$\left\| \frac{1}{N} \Xi^T \Xi - \Sigma \right\| \leq \gamma \iff \forall v \in \mathbb{R}^D, \left| \| \frac{1}{\sqrt{N}} \Xi v \|_2^2 - v^T \Sigma v \right| \leq \gamma \|v\|_2^2,$$

which implies, for every v on the sphere $\mathbb{S}^{D-1} = \{v \in \mathbb{S}^D : \|v\|_2 = 1\}$,

$$\frac{1}{N} \|\Xi v\|_2^2 \geq v^T \Sigma v - \gamma \Rightarrow \frac{1}{N} \inf_{v:\|v\|=1} \|\Xi v\|_2^2 \geq \inf_{v:\|v\|=1} v^T \Sigma v - \gamma.$$

Now, using the Courant-Fischer variational characterization of the smallest singular value [170], we obtain

$$\sigma_{\min}(\Xi) \geq N(\sigma_{\min}(\Sigma) - \gamma) > \frac{N}{2} \sigma_{\min}(\Sigma), \tag{6.23}$$

with probability at least $1 - \exp(-c' N^{1/4})$, where $c' > 0$ is a positive absolute constant smaller than c.

We now return to (6.19), to specifically bound $\|(\Xi^T \Xi)^{-1}\|$. Let A be any matrix A. Note that, $\|A^{-1}\| = \sigma_{\min}(A)^{-1}$. Indeed, taking the singular value decomposition $A = U \Sigma V^T$, and observing, $A^{-1} = (V^T)^{-1} \Sigma^{-1} U^{-1}$ we obtain $\|A^{-1}\| = \max_i(\sigma_i(A))^{-1} = \sigma_{\min}(A)^{-1}$. This, together with (6.23), yields:

$$\|(\Xi^T \Xi)^{-1}\| \leq \frac{2}{N \sigma_{\min}(\Sigma)}, \tag{6.24}$$

with probability at least $1 - \exp(-c' N^{1/4})$.

We now have all ingredients to execute the bound in (6.19). Combining Equations

(6.21) and (6.24), we get:

$$\begin{aligned}\mathcal{M} = (\Xi^T\Xi)^{-1}\Xi^T v \Rightarrow \|\mathcal{M}\|_2^2 &\leq \|(\Xi^T\Xi)^{-1}\|_2^2 \cdot \|\Xi^T v\|_2^2 \\ &\leq \underbrace{\frac{4}{N^2\sigma_{\min}(\Sigma)^2}}_{\text{from (6.24)}} \cdot \underbrace{N^2 d^3 \kappa}_{\text{from (6.21)}} \\ &= 4\kappa\sigma_{\min}(\Sigma)^{-2}d^3 \leq 4C\kappa d^{11},\end{aligned}$$

for some constant $C > 0$. Using (6.18) from Part (b) above, we have that κ can be taken

$$32\zeta\sigma_{\min}(W^*)^{-2}d^{4K+4}$$

with probability at least

$$1 - \exp(-C'd) - (9d^{4K+9})^{d^2-1} \cdot \exp(-C_4 N d^{-4-4K}) - Nd\exp(-Cd).$$

Since $\|\mathcal{M}\|_2^2 \leq 4C\kappa d^{11}$ with probability at least $1 - \exp(-c'N^{1/4})$, we have that

$$\|\mathcal{M}\|_2 \leq C'\sqrt{\zeta}d^{15/2+2K}\sigma_{\min}(W^*)^{-1}$$

with probability at least

$$1 - \exp(-c'N^{1/4}) - \Big(9d^{4K+9}\Big)^{d^2-1} \cdot \exp\Big(-C_4 N d^{-4-4K}\Big) - Nd\exp(-Cd),$$

by the union bound. As $\|M\|_F \leq \|\mathcal{M}\|_2$; and

$$\sqrt{\zeta} \leq \frac{\epsilon}{C'}d^{-15/2-2K}\sigma_{\min}(W^*)$$

per (6.4), we arrive at $\big\|W^TW - (W^*)^TW^*\big\|_F \leq \epsilon$ as claimed.

We now show the generalization ability. For any $W \in \mathbb{R}^{m\times d}$, using auxiliary result, Theorem 6.3.1(c), we have

$$\mathcal{L}(W) \leq \mu_2^2 \cdot \text{trace}(M) + \max\{\mu_4 - \mu_2^2, 2\mu_2^2\} \cdot \text{trace}(M^2),$$

where $M = W^TW - (W^*)^TW^* \in \mathbb{R}^{d\times d}$. Now note that $\text{trace}(M)^2 = |\sum_{1\leq i\leq d} M_{ii}|^2 \leq d\sum_{1\leq i\leq d} M_{ii}^2 \leq d\|M\|_F^2$ by Cauchy-Schwarz. Furthermore $\text{trace}(M^2) = \text{trace}(M^TM) = \|M\|_F^2$. Thus,

$$\mathcal{L}(W) \leq \|M\|_F^2\Big(d\mu_2^2 + \max\{\mu_4 - \mu_2^2, 2\mu_2^2\}\Big) \leq 2d\mu_2^2\|M\|_F^2,$$

for d large. Since $\|M\|_F^2 \leq \|\mathcal{M}\|_2^2 \leq (C')^2\zeta d^{15+4K}\sigma_{\min}(W^*)^{-2}$; we obtain

$$\mathcal{L}(W) \leq \zeta \cdot 2(C')^2\mu_2^2 d^{16+4K}\sigma_{\min}(W^*)^{-2}.$$

Finally, since

$$\zeta \leq \frac{\epsilon}{2(C')^2\mu_2^2} d^{-16-4K} \sigma_{\min}(W^*)^2$$

per (6.4) we conclude the proof of generalization bound, that is $\mathcal{L}(W) \leq \epsilon$.

Remark 6.4.4. *The argument presented above uses [107, Theorem 5.1]. While that result is stated for distributions supported on $[-1,1]^d$, it still applies under the weaker assumption that the distribution has finite moments of all orders, see [107, Remark 5.5].*

Case of Constant d: $d = O(1)$

We provide a very brief sketch for the argument in the case $d = O(1)$. The argument is quite similar to the one in Theorem 6.2.2. Similar to the analysis (of $d = O(1)$ case) conducted for Theorem 6.2.2, we use the fact that if X has a sub-Gaussian random variable, then $\mathbb{E}\left[|X|^p\right]^{1/p} = O\left(\sqrt{p}\right)$ for every $p \geq 1$; and in particular, $\mathbb{E}\left[|X|^p\right] < \infty$ for all $p \geq 1$; see [281, Lemma 5.5] for a more precise statement.

Next, the upper bound on the energy value now modifies to $\frac{1}{2}\overline{C_5}\sigma_{\min}(W^*)^4$.

Part (a)**.** Note that part (a) for the case of general d follows from earlier Claim 6.4.2. For the case when d is constant, part (a) now follows from modified Claim 6.4.2, provided under Theorem 6.2.2 for the case $d = O(1)$. That is, for the event

$$\mathcal{E}_1 \triangleq \left\{ \sup_{W:\widehat{\mathcal{L}}(W) \leq \frac{1}{2}\overline{C_5}\sigma_{\min}(W^*)^4} \|W\|_F = O(1) \right\}$$

it is the case $\mathbb{P}(\mathcal{E}_1) \geq 1 - O(1/N)$, where $\overline{C_5}$ is the constant appearing in Theorem 6.2.2.

Furthermore,

$$\left\| \nabla^2 \widehat{\mathcal{L}}(W) \right\| = \text{Poly}\left(\|W\|_F, \frac{1}{N} \sum_{1 \leq i \leq N} \|X_i\|_2^D \right)$$

for some absolute constant $D > 0$; and for any constant $D > 0$,

$$\frac{1}{N} \sum_{1 \leq i \leq N} \|X_i\|_2^D = O(1)$$

with probability at least $1 - O(1/N)$ (where we used, in particular, the fact $\mathbb{E}\left[X_i(j)^{2D}\right] < \infty$).

Combining these, we find that

$$L \triangleq \sup\left\{ \left\| \nabla^2 \widehat{\mathcal{L}}(W) \right\| : \widehat{\mathcal{L}}(W) \leq \frac{1}{2}\overline{C_5}\sigma_{\min}(W^*)^4 \right\} = O(1)$$

with probability at least $1 - O(1/N)$.

Part (b)**.** The analysis for time horizon T remains intact. Furthermore, the entire analysis leading to (6.15) remains (nearly) intact, and this equation now modifies to

$$\widehat{\mathcal{L}}(W) \leq O(1) \cdot \left\| \frac{1}{N} \sum_{1 \leq i \leq N} r_i X_i X_i^T \right\|_F,$$

since the Frobenius norm terms involved (all of which are polynomials in d) are now $O(1)$. The event $\mathcal{E}_3$ appearing in (6.16) modifies now to

$$\mathcal{E}_3 \triangleq \left\{ \inf_{\substack{W \in \mathbb{R}^{m \times d}: \sigma_{\min}(W) < \frac{1}{2}\sigma_{\min}(W^*) \\ \|W\|_F \leq d^{K+1}}} \widehat{\mathcal{L}}(W) \geq \frac{1}{2}\overline{C_5}\sigma_{\min}(W^*)^4 \right\}$$

which holds with probability at least $1 - O(1/N)$ (the modifications are exactly the same as those noted in Theorem 6.2.2 for the case $d = O(1)$). The rest of the analysis in Part (b) is exactly the same: combining modified versions of events $\mathcal{E}_1$ and $\mathcal{E}_3$ (note that there is no need to incorporate the event $\mathcal{E}_2$ which for the case of general d is required for truncation) via a union bound and recalling the equation (6.4) on ζ, it follows that with probability at least $1 - O(1/N)$, $\widehat{\mathcal{L}}(W) \leq \epsilon$.

Part (c)**.** The modification for this part is as follows. First, we do not condition on $\mathcal{E}_2$ like above. Instead, we will apply Chebyshev's inequality (elaborated below). The entire analysis leading up to (6.20) remains the same. Note now that $\Xi^T v \in \mathbb{R}^{d(d+1)/2}$. For each coordinate $\left(\Xi^T v\right)_i$ of this vector, we have, using Chebyshev's inequality,

$$\left|(\Xi^T v)_i\right| \leq O(\sqrt{N}) \cdot \|v\|_2 \leq N \cdot \sqrt{\kappa} \cdot O(1),$$

with probability $1 - O(1/N)$. Taking a union bound over $d(d+1)/2 = O(1)$ coordinates yields that with probability $1 - O(1/N)$ this remains true over all $1 \leq i \leq d(d+1)/2$.

Next, to control $\left\| \left(\Xi^T \Xi\right)^{-1} \right\|_2^2$ we do not need a delicate concentration result (such as Theorem 6.3.9) like above. Instead, we take the following route.

Fix $\epsilon > 0$, to be tuned. Recall the notation $\mathcal{R}_i$ from above, where $\mathcal{R}_i$ is the i^{th} row of matrix $\Xi \in \mathbb{R}^{N \times d(d+1)/2}$; and recall that $\mathcal{R}_i$, $1 \leq i \leq N$ are i.i.d. random vectors. Using the outer product reprensentation of matrix multiplication as above, we have

$$\frac{1}{N}\Xi^T \Xi = \frac{1}{N} \sum_{1 \leq i \leq N} \mathcal{R}_i \mathcal{R}_i^T \in \mathbb{R}^{\frac{d(d+1)}{2} \times \frac{d(d+1)}{2}}.$$

Consequently,

$$\mathbb{E}\left[\frac{1}{N}\Xi^T \Xi\right] = \mathbb{E}\left[\mathcal{R}_i \mathcal{R}_i^T\right] = \Sigma \in \mathbb{R}^{\frac{d(d+1)}{2} \times \frac{d(d+1)}{2}}$$

where $\mathbb{E}[\cdot]$ acts entrywise.

Since $d = O(1)$; a simple application of Chebyshev's inequality together with a

union bound over $\Theta(d^4)$ entries (of $N^{-1}\Xi^T\Xi$) yields

$$\max_{1\le i,j\le d(d+1)/2}\left|\left(\frac{1}{N}\Xi^T\Xi-\Sigma\right)_{ij}\right|\le\epsilon$$

with probability at least $1-O(1/N)$. Using now $\|M\|\le\|M\|_F^2$ valid for any matrix M, we obtain

$$\left\|\frac{1}{N}\Xi^T\Xi-\Sigma\right\|\le\epsilon^2 d^4$$

with probability $1-O(1/N)$. Similar to above, $\sigma_{\min}(\Sigma)=\Omega(1)$. Furthermore, the analysis starting from (6.22) and leading to (6.23) remains intact (with γ replaced with $\epsilon^2 d^4$, $\epsilon>0$ to be tuned). In particular, for ϵ sufficiently small, it is the case that with probability at least $1-O(1/N)$,

$$\sigma_{\min}(\Xi)>\frac{N}{2}\sigma_{\min}(\Sigma).$$

The rest of the analysis remains intact, except the probability bounds are now modified to $1-O(1/N)$. Finally, recalling the bound (6.4) on ζ, we find that with probability $1-O(1/N)$,

$$\left\|W^TW-(W^*)^TW^*\right\|_F\le\epsilon\quad\text{and}\quad\mathcal{L}(W)\le\epsilon.$$

6.4.12 Proof of Theorem 6.2.8

Let $W_0^TW_0=mI_d$, and let $\{\lambda_1,\dots,\lambda_d\}=\sigma((W^*)^TW^*-mI_d)$. In what follows below, recall the quantities from the proof of Theorem 6.2.7(b): $\sigma_*\triangleq\mathrm{Var}((W_{ij}^*)^2-1)$, $\chi_2\triangleq\int x^2\,d\omega(x)$, where $\omega(x)$ is the semicircle law. Fix now an arbitrary $\epsilon>0$ and a $K>0$.

We start by defininining several auxiliary events:

$$\begin{aligned}
\mathcal{E}_1&\triangleq\left\{\sum_{1\le i\le d}\lambda_i^2<4(1+o(1))md^2\chi_2\right\},\\
\mathcal{E}_2&\triangleq\left\{\left|\sum_{1\le i\le d}\lambda_i\right|<\sigma_*\sqrt{md}d^\epsilon\right\},\\
\mathcal{E}_3&\triangleq\left\{\sigma_{\min}(W^*)^4\ge\frac{1}{16}m^2\right\},\\
\mathcal{E}_4&\triangleq\left\{\|X_i\|_\infty\le d^{1/2},1\le i\le N\right\}.
\end{aligned}$$

Note that from the proof of Theorem 6.2.7(b), we have $\mathbb{P}(\mathcal{E}_i)\ge1-o_d(1)$ for $i=1,2,3$;

and from union bound and sub-Gaussianity of X, $\mathbb{P}(\mathcal{E}_4) \geq 1 - N\exp(-Cd)$. Thus,

$$\mathbb{P}\left(\bigcap_{1\leq i\leq 4}\mathcal{E}_i\right) \geq 1 - o_d(1) - N\exp\left(-Cd\right).$$

In what follows, suppose we condition on the event $\bigcap_{1\leq i\leq 4}\mathcal{E}_i$. Note that in this conditional universe, it is still the case that X_i, $1 \leq i \leq N$ are i.i.d. random vectors with centered i.i.d. coordinates. Using now Hölder's inequality (Theorem 6.3.10) with $p = 1, q = \infty$, $U = X_iX_i^T$ and $V = (W^*)^TW^* - mI_d$, we arrive at

$$\begin{aligned}|X_i^T((W^*)^TW^* - mI_d)X_i| &= \left|\left\langle X_iX_i^T, (W^*)^TW^* - mI_d\right\rangle\right| \\ &\leq \|(W^*)^TW^* - mI_d\|\text{trace}(X_iX_i^T) \\ &\leq 2\sqrt{md}d^2,\end{aligned}$$

where we use the fact that $\text{trace}(X_iX_i^T) = \|X_i\|_2^2 \leq d^2$ (recall the conditioning on $\mathcal{E}_4$). Using Hoeffding's inequality, we have

$$\widehat{\mathcal{L}}(W_0) = \frac{1}{N}\sum_{1\leq i\leq N}\Big(X_i^T(W^*)^TW^*X_i - X_i^TW_0^TW_0X_i\Big)^2 \leq \frac{3}{2}\mathcal{L}(W_0),$$

with probability at least

$$1 - \exp\Big(-C'Nd^{-5}m^{-1}\Big),$$

where

$$\mathcal{L}(W_0) = \mathbb{E}\Big[\big(X^T(W^*)^TW^*X - X^TW_0^TW_0X\big)^2\,\Big|\,\|X\|_\infty \leq d^{1/2}\Big].$$

Namely, $\mathcal{L}(W_0)$ is the "population risk" in the "conditional universe".

Next, in this conditional space, using Theorem 6.3.1(c), we arrive at

$$\mathcal{L}(W_0) \leq \mu_2(1/2)^2\left|\sum_{1\leq i\leq d}\lambda_i\right|^2 + \max\Big\{\mu_4(1/2) - \mu_2(1/2)^2, 2\mu_2(1/2)^2\Big\}\left(\sum_{1\leq i\leq d}\lambda_i^2\right).$$

Finally, carrying out the exact same analysis as in the end of the proof of Theorem 6.2.7, we deduce

$$\widehat{\mathcal{L}}(W_0) < \frac{1}{2}C_5\sigma_{\min}(W^*)^4,$$

provided $m > C''d^2$ for a large enough constant C'', namely provided that the network is sufficiently overparameterized.

6.4.13 Proof of Theorem 6.2.9

(a) Let $\text{span}(X_iX_i^T : i \in [N]) = \mathcal{S}$, the set of all $d \times d$ symmetric matrices, and let $M \in \mathcal{S}$ be such that for any i, $X_i^TMX_i = 0$. We will establish $M = 0$. Let $1 \leq k, \ell \leq d$ be two fixed indices. To that end, let $\theta_i^{(k,\ell)} \in \mathbb{R}$ be such

that, $\sum_{i=1}^{N} \theta_i^{(k,\ell)} X_i X_i^T = e_k e_\ell^T + e_\ell e_k^T$, where the column vectors $e_k, e_\ell \in \mathbb{R}^d$ are respectively the k^{th} and ℓ^{th} elements of the standard basis for $\mathbb{R}^d$. Such $\theta_i^{(k,\ell)}$ indeed exist, due to the spanning property. Observe that $2M_{k,\ell} = e_k^T M e_\ell + e_\ell^T M e_k = \mathrm{tr}(e_k^T M e_\ell + e_\ell^T M e_k)$. Now, using the fact that $\mathrm{tr}(ABC) = \mathrm{tr}(BCA) = \mathrm{tr}(CAB)$ for every matrices A, B, C (with matching dimensions), we have:

$$2M_{k,\ell} = \mathrm{tr}(Me_\ell e_k^T + Me_k e_\ell^T) = \mathrm{tr}\left(\sum_{i=1}^{N} \theta_i^{(k,\ell)} M X_i X_i^T\right) = \sum_{i=1}^{N} \theta_i^{(k,\ell)} \mathrm{tr}(X_i^T M X_i) = 0,$$

for every $k, \ell \in [d]$. Finally, if W is such that $\widehat{\mathcal{L}}(W) = 0$, then $X_i^T M X_i = 0$ for any i, where $M = (W^*)^T W^* - W^T W$. Hence, provided that the geometric condition holds, we have $M = 0$, that is, $W^T W = (W^*)^T W^*$. From here, the final conclusion follows per Theorem 6.3.6. Since $W^T W = (W^*)^T W^*$, W clearly has zero generalization error, i.e. $\mathcal{L}(W) = 0$.

(b) Our goal is to construct a $W \in \mathbb{R}^{m \times d}$ with $f(W^*; X_i) = f(W; X_i)$, for every $i \in [N]$, whereas $W^T W \neq (W^*)^T W^*$. Consider the inner product $\langle A, B \rangle = \mathrm{trace}(AB)$, in the space of all symmetric $d \times d$ matrices. Find $0 \neq M \in \mathbb{R}^{d \times d}$ a symmetric matrix, such that, $M \in \mathrm{span}^{\perp}(X_i X_i^T : i \in [N])$, that is, $X_i^T M X_i = 0$ for every $i \in [N]$. We can find such M satisfying $\|M\|_2 = 1$. Consider the linear matrix function $M(\delta) = (W^*)^T W^* + \delta M$. Note that, $M(\delta)$ is symmetric for every δ. We claim that under the hypothesis of the theorem, there exists a $\delta_0 > 0$ such that $M(\delta)$ is positive semidefinite for every $\delta \in [0, \delta_0]$, and that there exists $W_\delta \in \mathbb{R}^{m \times d}$ with $W_\delta^T W_\delta = M(\delta)$, for all $\delta \in [0, \delta_0]$. Observe that, since $\mathrm{rank}(W^*) = d$, then $(W^*)^T W^* \in \mathbb{R}^{d \times d}$ with $\mathrm{rank}((W^*)^T W^*) = d$. Therefore, the eigenvalues $\lambda_1^*, \ldots, \lambda_d^*$ of $(W^*)^T W^*$ are all positive. In particular $\{\lambda_i^* : i \in [d]\} \subset [\delta_1, \infty)$, with $\delta_1 = \sigma_{\min}(W^*)^2$. Now, let $\mu_1(\delta), \ldots, \mu_d(\delta)$ be the eigenvalues of $M(\delta)$. Using Weyl's inequality [170], we have $|\mu_i(\delta) - \lambda_i^*| \leq \delta \|M\|_2 = \delta$, for every i. In particular, taking $\delta \leq \delta_1$, we deduce for every $i \in [d]$, it holds that $\mu_i(\delta) \geq \lambda_i^* - \delta_1 \geq 0$, that is, $\{\mu_i(\delta) : i \in [d]\} \subset [0, \infty)$. In particular, we also have $M(\delta)$ is symmetric, and thus, it is PSD. Thus, there exists a $\overline{W_\delta} \in \mathbb{R}^{d \times d}$ such that $\overline{W_\delta}^T \overline{W_\delta} = M(\delta)$. Now, using the same idea as in the proof of Theorem 6.2.1 part (c), we then deduce that for any $\widehat{m} \geq d$, there exists a matrix $W_\delta \in \mathbb{R}^{\widehat{m} \times d}$ such that $W_\delta^T W_\delta = \overline{W_\delta}^T \overline{W_\delta} = M(\delta)$. In particular, for this W_δ, if $f(W_\delta, X)$ is the function computed by the neural network with weight matrix $W_\delta \in \mathbb{R}^{\widehat{m} \times d}$, then on the training data $(X_i : i \in [N])$, $f(W_\delta; X_i) = X_i^T W_\delta^T W_\delta X_i = X_i^T (W^*)^T W^* X_i = f(W^*; X_i)$, since $X_i^T M X_i = 0$ for all $i \in [N]$. At the same time $W_\delta^T W_\delta - (W^*)^T W^* = \delta M \neq 0$, since $\delta \neq 0$ and $M \neq 0$, and therefore $W_\delta^T W_\delta \neq (W^*)^T W^*$.

Finally, to show $\mathcal{L}(W_\delta) > 0$, we argue as follows. Suppose $\mathcal{L}(W_\delta) = 0$. Then, by Theorem 6.3.5, it follows that $\psi(X) = X^T A X = 0$ identically, where $A = W_\delta^T W_\delta - (W^*)^T W^*$. Now, letting $\xi_1, \ldots, \xi_d$ to be the eigenvectors of A (with corresponding eigenvalues $\lambda_1, \ldots, \lambda_d$), we obtain $\xi_i^T A \xi_i = \lambda_i \xi_i^T \xi_i = \lambda_i \|\xi_i\|_2^2 = 0$, we namely obtain $\lambda_i = 0$ for every $i \in [d]$. Finally, since A is symmetric, and

hence admits a diagonalization of form $A = \mathcal{Q}\Lambda\mathcal{Q}$ with diagonal entries of Λ being zero, we deduce A is identically zero, which contradicts with the fact that $A = \delta M$, which is a non-zero matrix.

6.4.14 Proof of Theorem 6.2.10

Recall that, $\mathcal{S} = \{M \in \mathbb{R}^{d\times d} : M^T = M\}$. Note that, this space has dimension $\binom{d}{2}+d$: for any $1 \leq k \leq \ell \leq d$, it is easy to see that the matrices $e_k e_\ell^T + e_\ell e_k^T$ are linearly independent; and there are precisely $\binom{d}{2} + d$ such matrices. With this in mind, the statement of part (b) is immediate.

We now prove the part (a) of the theorem. For any X_i, let $X_i(j)$ be the j^{th} coordinate of X_i, with $j \in [d]$; and let $\mathcal{Y}_i$ be a $d(d+1)/2-$dimensional vector, obtained by retaining $X_i(1)^2, \ldots, X_i(d)^2$; and the products, $X_i(k)X_i(\ell)$ with $1 \leq k < \ell \leq d$. Now, let $\mathcal{X}$ be an $n \times d(d+1)/2$ matrix, whose rows are $\mathcal{Y}_1, \ldots, \mathcal{Y}_n$. Our goal is to establish,

$$\mathbb{P}[\det(\mathcal{X}) = 0] = 0,$$

when $n = d(d+1)/2$, where the probability is taken with respect to the randomness in $X_1, \ldots, X_n$ (in particular, this yields for $n \geq d(d+1)/2$, $\mathbb{P}(\text{rank}(\mathcal{X}) = d(d+1)/2)$, almost surely). Now, recalling Theorem 6.3.5, it then suffices to show that $\det(\mathcal{X})$ is not identically zero, when viewed as a polynomial in $X_i(j)$ with $i \in [N]$, $j \in [d]$.

We now prove part (b) by providing a deterministic construction (of the matrix $\mathcal{X}$) under which $\det(\mathcal{X}) \neq 0$. Let $p_1 < \cdots < p_d$ be distinct prime numbers. For every $1 \leq t \leq N$, set:

$$X_t = (p_1^{t-1}, \ldots, p_d^{t-1})^T \in \mathbb{R}^d.$$

In particular, $X_1 = (1, 1, \ldots, 1)^T \in \mathbb{R}^d$, which then implies $\mathcal{Y}_1$ is a vector of all ones. Now, we study $\mathcal{Y}_2$. The entries of $\mathcal{Y}_2$, called $z_1, \ldots, z_{d(d+1)/2}$, are of form p_i^2 with $i \in [d]$; or $p_i p_j$, where $1 \leq i < j \leq d$. By the fundamental theorem of arithmetic, we have $p_i p_j = p_k p_\ell \Rightarrow \{p_i, p_j\} = \{p_k, p_\ell\}$; and therefore, $z_1, \ldots, z_{d(d+1)/2}$ are pairwise distinct. With this construction, the matrix $\mathcal{X}$ is a Vandermonde matrix with determinant:

$$\prod_{1\leq k<\ell\leq d(d+1)/2} (z_k - z_\ell).$$

Since $z_k \neq z_\ell$ for every $k \neq \ell$ (from the construction on $\mathcal{Y}_2$, which, in turn, is constructed from X_2), this determinant is non-zero, proving the claim.

6.4.15 Proof of Theorem 6.2.11

(a) Note that, if $N \geq N^*$, then combining parts (a) of Theorems 6.2.9 and 6.2.10, we have that with probability one, $\text{span}(X_i X_i^T : i \in [N]) = \mathcal{S}$, which, together with $\widehat{\mathcal{L}}(W) = 0$, imply that,

$$\mathbb{P}(E \neq \varnothing) = 0,$$

where $E = \{W \in \mathbb{R}^{m\times d} : W^TW \neq (W^*)^TW^*; \widehat{\mathcal{L}}(W) = 0\}$, from which the desired conclusion follows.

(b) Assume W is taken as in proof of Theorem 6.2.9 (b), that is,

$$A = (W^*)^TW^* - W^TW = \delta M \quad \text{where} \quad \delta = \sigma_{\min}(W^*)^2 \quad \text{and} \quad \|M\| = 1,$$

with $M^T = M$. Let $\{\lambda_1, \ldots, \lambda_d\}$ be the spectrum of the matrix δM. Using now Theorem 6.3.1 (c), we have the lower bound

$$\begin{aligned}
\mathcal{L}(W) &\geq \mathbb{E}\left[X_i(j)^2\right]^2 \operatorname{trace}(A)^2 + \min\left\{\operatorname{Var}(X_i(j)^2), 2\mathbb{E}\left[X_i(j)^2\right]^2\right\} \cdot \operatorname{trace}(A^2) \\
&\geq \min\left\{\operatorname{Var}(X_i(j)^2), 2\mathbb{E}\left[X_i(j)^2\right]^2\right\} \left(\sum_{i=1}^{d} \lambda_i^2\right) \\
&\geq \min\left\{\operatorname{Var}(X_i(j)^2), 2\mathbb{E}\left[X_i(j)^2\right]^2\right\} \lambda_{\max}(\delta M)^2,
\end{aligned}$$

since $\operatorname{trace}(A^2) = \sum_{i=1}^{d} \lambda_i^2$. Finally, since $\lambda_{\max}(\delta M)^2 = \delta^2 = \sigma_{\min}(W^*)^4$ (as the spectral norm of M is one), we arrive at the desired conclusion.

Appendix A

MATLAB Code for Verifying Lemma 3.6.1

We verify Lemma 3.6.1 numerically using the following MATLAB code.

Our experiments demonstrate that the functions $f_2(\beta, \alpha), f_3(\beta, \alpha)$ appear to be minimized when β is close to one. For this reason, we restrict our attention to $\beta \in [0.9, 0.999]$ and generate $\beta = 0.9 : \text{sp} : 0.999$ with $\text{sp} = 10^{-3}$. We take $K = 1$ as in the rest of the chapter, and set $\alpha = 1.667$. In order the compute the probability term, we do not resort to any Monte Carlo simulations. Instead, we employ MATLAB's built-in `mvncdf` function to compute the associated "box" probability (for dimensions 2 and 3). (The function, `mvncdf`, computes rectangular probabilities for multivariate Gaussian distribution using numerical integration, see [2] for a more elaborate description.) In particular, the only potential source of error is the error encountered at the numerical integration step. A feature of the `mvncdf` function is that the error guarantee in probability calculation is available. In particular, an inspection of our plots reveals that $f_3(\beta, 1.677)$ is minimized for $\beta \approx 0.978$; and for this choice of β, the probability term is approximately 0.6205 whereas the error estimate is of order 10^{-8}.

```
close all, clear all;
sp  = 1e-3; %spacing
beta = 0:sp:0.999;
L = length(beta);
K = 1;
alpha = 1.67;
%\varphi_count
phi_count_2 = 1+binent((1-beta)./2);
phi_count_3 = 1+(1-beta)./2 ...
    +binent((1-beta)./2)+((1+beta)./2).*binent((1-beta)./(2.*(1+beta)));
%probability term
mu_2 = zeros(1,2);
mu_3 = zeros(1,3); %mean
box_low_2 = (-K)*ones(1,2);
box_low_3 = (-K)*ones(1,3); %lower limits -K for probability box
box_high_3 = K*ones(1,3); %upper limits K for probability box
box_high_2 = K*ones(1,2);
```

```
f_2 = zeros(1,L);
f_3 = zeros(1,L);
phi_probs_2 = zeros(1,L);
phi_probs_3 = zeros(1,L);
for i=1:L
    phi_probs_2(i) = ...
        mvncdf(box_low_2,box_high_2,mu_2,(1-beta(i))*eye(2) + ...
        beta(i)*ones(2,2));
    phi_probs_3(i) = ...
        mvncdf(box_low_3,box_high_3,mu_3,(1-beta(i))*eye(3) + ...
        beta(i)*ones(3,3)); %evaluate probability
    f_2(i)   = phi_count_2(i) + alpha*log2(phi_probs_2(i));
    f_3(i)   = phi_count_3(i) + alpha*log2(phi_probs_3(i)); ...
        %construct f_3
end
figure
title('3-OGP')
plot(beta,f_3),
hold on
plot(beta,f_2,'g')
ylabel('$f_3(\beta),f_2(\beta)$','Interpreter','latex');
xlabel('$\beta$','Interpreter','latex');
legend('f_3','f_2','Zero')
RefLine = refline([0,0]);
RefLine.Color = 'r';
disp(['The minima of f_3 is ',num2str(min(f_3)),'.'])

figure
title('2-OGP')
plot(beta,f_2),
ylabel('$f_2(\beta)$','Interpreter','latex');
xlabel('$\beta$','Interpreter','latex');
RefLine = refline([0,0]);
RefLine.Color = 'r';
disp(['The minima of f_2 is ',num2str(min(f_2)),'.'])

function ent = binent(x)
    ent = -x .* log2(x)-(1-x) .* log2(1-x);
end
```

Bibliography

[1] Non-negative matrix factorization. `https://en.wikipedia.org/wiki/Non-negative_matrix_factorization`. Accessed: 2021-03-01.

[2] `mvncdf` multivariate normal cumulative distribution function. `http://web.archive.org/https://www.mathworks.com/help/stats/mvncdf.html`. Accessed: 2021-07-03.

[3] Scott Aaronson and Alex Arkhipov. The computational complexity of linear optics. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 333–342. ACM, 2011.

[4] Emmanuel Abbe, Enric Boix Adsera, Matthew Brennan, Guy Bresler, and Dheeraj Nagaraj. The staircase property: How hierarchical structure can guide deep learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[5] Emmanuel Abbe, Shuangping Li, and Allan Sly. Binary perceptron: efficient algorithms can find solutions in a rare well-connected cluster. *arXiv preprint arXiv:2111.03084*, 2021.

[6] Emmanuel Abbe, Shuangping Li, and Allan Sly. Proof of the contiguity conjecture and lognormal limit for the symmetric perceptron. *arXiv preprint arXiv:2102.13069*, 2021.

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

[8] Dimitris Achlioptas, Amin Coja-Oghlan, and Federico Ricci-Tersenghi. On the solution-space geometry of random constraint satisfaction problems. *Random Structures & Algorithms*, 38(3):251–268, 2011.

[9] Dimitris Achlioptas and Federico Ricci-Tersenghi. On the solution-space geometry of random constraint satisfaction problems. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 130–139, 2006.

[10] Louigi Addario-Berry, Luc Devroye, Gábor Lugosi, and Roberto I Oliveira. Local optima of the sherrington-kirkpatrick hamiltonian. *Journal of Mathematical Physics*, 60(4):043301, 2019.

[11] Ahmed El Alaoui and Mark Sellke. Algorithmic pure states for the negative spherical perceptron. *arXiv preprint arXiv:2010.15811*, 2020.

[12] Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

[13] Noga Alon and Joel H Spencer. *The probabilistic method*. John Wiley & Sons, 2016.

[14] Dylan J Altschuler and Jonathan Niles-Weed. The discrepancy of random rectangular matrices. *Random Structures & Algorithms*, 2021.

[15] Ryan Alweiss, Yang P Liu, and Mehtaab Sawhney. Discrepancy minimization via a self-balancing walk. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 14–20, 2021.

[16] Sanjeev Arora, Eli Berger, Hazan Elad, Guy Kindler, and Muli Safra. On non-approximability for quadratic programs. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pages 206–215. IEEE, 2005.

[17] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pages 8139–8148, 2019.

[18] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

[19] Sanjeev Arora, Rong Ge, Ravi Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization—provably. *SIAM Journal on Computing*, 45(4):1582–1611, 2016.

[20] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. *arXiv preprint arXiv:1802.05296*, 2018.

[21] Gerard Ben Arous, Reza Gheissari, Aukosh Jagannath, et al. Algorithmic thresholds for tensor pca. *Annals of Probability*, 48(4):2052–2087, 2020.

[22] Gérard Ben Arous and Aukosh Jagannath. Shattering versus metastability in spin glasses. *arXiv preprint arXiv:2104.08299*, 2021.

[23] Benjamin Aubin, Will Perkins, and Lenka Zdeborova. Storage capacity in symmetric binary perceptrons. *Journal of Physics A: Mathematical and Theoretical*, 52(29):294003, 2019.

[24] Z. D. Bai and Y. Q. Yin. Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix. *The Annals of Probability*, 21(3):1275 – 1294, 1993.

[25] Zhidong D Bai and Yong Q Yin. Convergence to the semicircle law. *The Annals of Probability*, 16(2):863–875, 1988.

[26] Carlo Baldassi. Generalization learning in a perceptron with binary synapses. *Journal of Statistical Physics*, 136(5):902–916, 2009.

[27] Carlo Baldassi and Alfredo Braunstein. A max-sum algorithm for training discrete neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(8):P08008, 2015.

[28] Carlo Baldassi, Alfredo Braunstein, Nicolas Brunel, and Riccardo Zecchina. Efficient supervised learning in networks with binary synapses. *Proceedings of the National Academy of Sciences*, 104(26):11079–11084, 2007.

[29] Carlo Baldassi, Riccardo Della Vecchia, Carlo Lucibello, and Riccardo Zecchina. Clustering of solutions in the symmetric binary perceptron. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(7):073303, 2020.

[30] Carlo Baldassi, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina. Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses. *Physical review letters*, 115(12):128101, 2015.

[31] Afonso S Bandeira, Amelia Perry, and Alexander S Wein. Notes on computational-to-statistical gaps: predictions using statistical physics. *arXiv preprint arXiv:1803.11132*, 2018.

[32] Nikhil Bansal. Constructive algorithms for discrepancy minimization. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 3–10. IEEE, 2010.

[33] Nikhil Bansal, Haotian Jiang, Raghu Meka, Sahil Singla, and Makrand Sinha. Online discrepancy minimization for stochastic arrivals. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2842–2861. SIAM, 2021.

[34] Nikhil Bansal, Haotian Jiang, Sahil Singla, and Makrand Sinha. Online vector balancing and geometric discrepancy. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1139–1152, 2020.

[35] Nikhil Bansal and Joel H Spencer. On-line balancing of random inputs. *Random Structures & Algorithms*, 57(4):879–891, 2020.

[36] Nikhil Bansal and Joel H. Spencer. On-line balancing of random inputs. *Random Structures and Algorithms*, 57(4):879–891, December 2020.

[37] Francisco Barahona. On the computational complexity of ising spin glass models. *Journal of Physics A: Mathematical and General*, 15(10):3241, 1982.

[38] Boaz Barak, Samuel Hopkins, Jonathan Kelner, Pravesh K Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. *SIAM Journal on Computing*, 48(2):687–735, 2019.

[39] Andrew R Barron. Approximation and estimation bounds for artificial neural networks. *Machine learning*, 14(1):115–133, 1994.

[40] Peter L Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536, 1998.

[41] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

[42] Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

[43] Peter L Bartlett, Philip M Long, and Robert C Williamson. Fat-shattering and the learnability of real-valued functions. *journal of computer and system sciences*, 52(3):434–452, 1996.

[44] Heiko Bauke and Stephan Mertens. Universality in the level statistics of disordered systems. *Physical Review E*, 70(2):025102, 2004.

[45] Mohsen Bayati, David Gamarnik, and Prasad Tetali. Combinatorial approach to the interpolation method and scaling limits in sparse random graphs. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 105–114, 2010.

[46] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[47] Quentin Berthet and Philippe Rigollet. Computational lower bounds for sparse pca. *arXiv preprint arXiv:1304.0828*, 2013.

[48] Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

[49] Avrim Blum and Ronald L Rivest. Training a 3-node neural network is np-complete. In *Advances in neural information processing systems*, pages 494–501, 1989.

[50] Lenore Blum, Felipe Cucker, Michael Shub, and Steve Smale. *Complexity and real computation*. Springer Science & Business Media, 2012.

[51] Lenore Blum, Mike Shub, and Steve Smale. On a theory of computation over the real numbers; np completeness, recursive functions and universal machines. In *[Proceedings 1988] 29th Annual Symposium on Foundations of Computer Science*, pages 387–397. IEEE, 1988.

[52] Stefan Boettcher and Stephan Mertens. Analysis of the karmarkar-karp differencing algorithm. *The European Physical Journal B*, 65(1):131, 2008.

[53] Enric Boix-Adserà, Matthew Brennan, and Guy Bresler. The average-case complexity of counting cliques in Erdős–Rényi hypergraphs. *SIAM Journal on Computing*, (0):FOCS19–39, 2021.

[54] Erwin Bolthausen, Shuta Nakajima, Nike Sun, and Changji Xu. Gardner formula for Ising perceptron models at small densities. *arXiv preprint arXiv:2111.02855*, 2021.

[55] Christian Borgs, Jennifer Chayes, Stephan Mertens, and Chandra Nair. Proof of the local rem conjecture for number partitioning. i: Constant energy scales. *Random Structures & Algorithms*, 34(2):217–240, 2009.

[56] Christian Borgs, Jennifer Chayes, Stephan Mertens, and Chandra Nair. Proof of the local rem conjecture for number partitioning. ii. growing energy scales. *Random Structures & Algorithms*, 34(2):241–284, 2009.

[57] Christian Borgs, Jennifer Chayes, and Boris Pittel. Phase transition and finite-size scaling for the integer partitioning problem. *Random Structures & Algorithms*, 19(3-4):247–288, 2001.

[58] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[59] Anton Bovier and Irina Kurkova. A short course on mean field spin glasses. In *Spin Glasses: Statics and Dynamics*, pages 3–44. Springer, 2009.

[60] Alfredo Braunstein and Riccardo Zecchina. Learning by message passing in networks of discrete synapses. *Physical review letters*, 96(3):030201, 2006.

[61] Matthew Brennan and Guy Bresler. Optimal average-case reductions to sparse pca: From weak assumptions to strong hardness. *arXiv preprint arXiv:1902.07380*, 2019.

[62] Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. *arXiv preprint arXiv:1806.07508*, 2018.

[63] Guy Bresler and Brice Huang. The algorithmic phase transition of random k-sat for low degree polynomials. *arXiv preprint arXiv:2106.02129*, 2021.

[64] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 605–614. JMLR. org, 2017.

[65] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*, 2017.

[66] Helmut Bölcskei, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. Optimal approximation with sparsely connected deep neural networks. *SIAM Journal on Mathematics of Data Science*, 1(1):8–45, 2019.

[67] Jin-Yi Cai, Aduri Pavan, and D Sivakumar. On the hardness of permanent. In *Annual Symposium on Theoretical Aspects of Computer Science*, pages 90–99. Springer, 1999.

[68] Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[69] Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, pages 10836–10846, 2019.

[70] Richard Caron and Tim Traynor. The zero set of a polynomial. *WSMR Report*, pages 05–02, 2005.

[71] Karthekeyan Chandrasekaran and Santosh S Vempala. Integer feasibility of random polytopes: random integer programs. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 449–458, 2014.

[72] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, pages 6571–6583, 2018.

[73] Wei-Kuo Chen, David Gamarnik, Dmitry Panchenko, Mustazee Rahman, et al. Suboptimality of local algorithms for a class of max-cut problems. *Annals of Probability*, 47(3):1587–1618, 2019.

[74] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3036–3046, 2018.

[75] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204, 2015.

[76] Edward Grady Coffman and George S Lueker. *Probabilistic analysis of packing and partitioning algorithms*. Wiley-Interscience, 1991.

[77] Amin Coja-Oghlan and Charilaos Efthymiou. On independent sets in random graphs. *Random Structures & Algorithms*, 47(3):436–486, 2015.

[78] Amin Coja-Oghlan, Amir Haqshenas, and Samuel Hetterich. Walksat stalls well below satisfiability. *SIAM Journal on Discrete Mathematics*, 31(2):1160–1173, 2017.

[79] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[80] David Conlon and Asaf Ferber. Lower bounds for multicolor ramsey numbers. *Advances in Mathematics*, 378:107528, 2020.

[81] David Conlon, Jacob Fox, and Benny Sudakov. Recent developments in graph ramsey theory. *Surveys in combinatorics*, 424:49–118, 2015.

[82] Kevin P Costello. Balancing gaussian vectors. *Israel Journal of Mathematics*, 172(1):145–156, 2009.

[83] Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3):326–334, 1965.

[84] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[85] Imre Csiszar and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[86] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342, 2018.

[87] Amir Dembo, Andrea Montanari, and Subhabrata Sen. Extremal cuts of sparse random graphs. *The Annals of Probability*, 45(2):1190–1217, 2017.

[88] Bernard Derrida. Random-energy model: Limit of a family of disordered models. *Physical Review Letters*, 45(2):79, 1980.

[89] Bernard Derrida. Random-energy model: An exactly solvable model of disordered systems. *Physical Review B*, 24(5):2613, 1981.

[90] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the balanced binary stochastic block model. *Information and Inference: A Journal of the IMA*, 6(2):125–170, 2017.

[91] Yash Deshpande and Andrea Montanari. Information-theoretically optimal sparse pca. In *2014 IEEE International Symposium on Information Theory*, pages 2197–2201. IEEE, 2014.

[92] Yash Deshpande and Andrea Montanari. Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems. In *Conference on Learning Theory*, pages 523–562, 2015.

[93] Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, and Nikos Zarifis. Algorithms and sq lower bounds for pac learning one-hidden-layer relu networks. In *Conference on Learning Theory*, pages 1514–1539. PMLR, 2020.

[94] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017.

[95] Jian Ding and Nike Sun. Capacity lower bound for the Ising perceptron. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 816–827, 2019.

[96] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

[97] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. In *Advances in neural information processing systems*, pages 1067–1077, 2017.

[98] Simon S Du and Jason D Lee. On the power of over-parametrization in neural networks with quadratic activation. *arXiv preprint arXiv:1803.01206*, 2018.

[99] Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018.

[100] Simon S Du, Jason D Lee, and Yuandong Tian. When is a convolutional filter easy to learn? *arXiv preprint arXiv:1709.06129*, 2017.

[101] Simon S Du, Jason D Lee, Yuandong Tian, Barnabas Poczos, and Aarti Singh. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. *arXiv preprint arXiv:1712.00779*, 2017.

[102] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[103] Martin Dyer, Alan Frieze, and Mark Jerrum. On counting independent sets in sparse graphs. *SIAM Journal on Computing*, 31(5):1527–1541, 2002.

[104] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

[105] Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940, 2016.

[106] Ronen Eldan and Mohit Singh. Efficient algorithms for discrepancy minimization in convex sets. *Random Structures & Algorithms*, 53(2):289–307, 2018.

[107] Matt Emschwiller, David Gamarnik, Eren C Kızıldağ, and Ilias Zadik. Neural networks and polynomial regression. demystifying the overparametrization phenomena. *arXiv preprint arXiv:2003.10523*, 2020.

[108] Paul Erdös and George Szekeres. A combinatorial problem in geometry. *Compositio mathematica*, 2:463–470, 1935.

[109] Uriel Feige and Carsten Lund. On the hardness of computing the permanent of random matrices. In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*, pages 643–654. ACM, 1992.

[110] J. Feigenbaum and L. Fortnow. On the random-self-reducibility of complete sets. In *[1991] Proceedings of the Sixth Annual Structure in Complexity Theory Conference*, pages 124–132, 1991.

[111] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *Journal of the ACM (JACM)*, 64(2):1–37, 2017.

[112] Vitaly Feldman, Will Perkins, and Santosh Vempala. On the complexity of random satisfiability problems with planted solutions. *SIAM Journal on Computing*, 47(4):1294–1338, 2018.

[113] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Birkhäuser Basel, 2013.

[114] C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016.

[115] Spencer Frei, Yuan Cao, and Quanquan Gu. Algorithm-dependent generalization bounds for overparameterized deep residual networks. In *Advances in Neural Information Processing Systems*, pages 14797–14807, 2019.

[116] Ehud Friedgut. Sharp thresholds of graph properties, and the k-SAT problem. *Journal of the American mathematical Society*, 12(4):1017–1054, 1999.

[117] Alan M Frieze. On the independence number of random graphs. *Discrete Mathematics*, 81(2):171–175, 1990.

[118] Alan M Frieze and T Łuczak. On the independence and chromatic numbers of random regular graphs. *Journal of Combinatorial Theory, Series B*, 54(1):123–132, 1992.

[119] David Gamarnik. The overlap gap property: A topological barrier to optimizing over random structures. *Proceedings of the National Academy of Sciences*, 118(41), 2021.

[120] David Gamarnik and Aukosh Jagannath. The overlap gap property and approximate message passing algorithms for p-spin models. *The Annals of Probability*, 49(1):180–205, 2021.

[121] David Gamarnik, Aukosh Jagannath, and Subhabrata Sen. The overlap gap property in principal submatrix recovery. *arXiv preprint arXiv:1908.09959*, 2019.

[122] David Gamarnik, Aukosh Jagannath, and Alexander S Wein. Low-degree hardness of random optimization problems. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 131–140. IEEE, 2020.

[123] David Gamarnik, Aukosh Jagannath, and Alexander S Wein. Low-degree hardness of random optimization problems. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, 2020.

[124] David Gamarnik, Aukosh Jagannath, and Alexander S Wein. Circuit lower bounds for the p-spin optimization problem. *arXiv preprint arXiv:2109.01342*, 2021.

[125] David Gamarnik and Eren C Kizildağ. High-dimensional linear regression and phase retrieval via pslq integer relation algorithm. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 1437–1441. IEEE, 2019.

[126] David Gamarnik and Eren C Kızıldağ. Algorithmic obstructions in the random number partitioning problem. *arXiv preprint arXiv:2103.01369*, 2021.

[127] David Gamarnik, Eren C Kızıldağ, and Ilias Zadik. Stationary points of shallow neural networks with quadratic activation function. *arXiv preprint arXiv:1912.01599*, 2019.

[128] David Gamarnik, Eren C Kızıldağ, and Ilias Zadik. Inference in high-dimensional linear regression via lattice basis reduction and integer relation detection. *IEEE Transactions on Information Theory*, 67(12):8109–8139, 2021.

[129] David Gamarnik, Eren C Kızıldağ, and Ilias Zadik. Self-regularity of output weights for overparameterized two-layer neural networks. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 819–824. IEEE, 2021.

[130] David Gamarnik, Eren C Kızıldağ, and Ilias Zadik. Self-regularity of non-negative output weights for overparameterized two-layer neural networks. *IEEE Transactions on Signal Processing*, 70:1310–1319, 2022.

[131] David Gamarnik, Eren C. Kızıldağ, Will Perkins, and Changji Xu. Algorithms and barriers in the symmetric binary perceptron model. *arXiv preprint arXiv:2203.15667*, 2022.

[132] David Gamarnik and Eren C. Kızıldağ. Computing the partition function of the Sherrington–Kirkpatrick model is hard on average. *The Annals of Applied Probability*, 31(3):1474 – 1504, 2021.

[133] David Gamarnik and Eren C. Kızıldağ. Computing the partition function of the Sherrington–Kirkpatrick model is hard on average. *The Annals of Applied Probability*, 31(3):1474 – 1504, 2021.

[134] David Gamarnik, Quan Li, et al. Finding a large submatrix of a gaussian random matrix. *The Annals of Statistics*, 46(6A):2511–2561, 2018.

[135] David Gamarnik and Madhu Sudan. Limits of local algorithms over sparse random graphs. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 369–376, 2014.

[136] David Gamarnik and Madhu Sudan. Limits of local algorithms over sparse random graphs. *Ann. Probab.*, 45(4):2353–2376, 07 2017.

[137] David Gamarnik and Madhu Sudan. Performance of sequential local algorithms for the random nae-k-sat problem. *SIAM Journal on Computing*, 46(2):590–619, 2017.

[138] Elizabeth Gardner. Maximum storage capacity in neural networks. *EPL (Europhysics Letters)*, 4(4):481, 1987.

[139] Elizabeth Gardner. The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, 21(1):257, 1988.

[140] Elizabeth Gardner and Bernard Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and general*, 21(1):271, 1988.

[141] Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., USA, 1990.

[142] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[143] Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

[144] Rong Ge, Runzhe Wang, and Haoyu Zhao. Mildly overparametrized neural nets can memorize training data efficiently. *arXiv preprint arXiv:1909.11837*, 2019.

[145] Peter Gemmell, Richard Lipton, Ronitt Rubinfeld, Madhu Sudan, and Avi Wigderson. Self-testing/correcting for polynomials and for approximate functions. In *STOC*, volume 91, pages 32–42. Citeseer, 1991.

[146] Peter Gemmell and Madhu Sudan. Highly resilient correctors for polynomials. *Information processing letters*, 43(4):169–174, 1992.

[147] Ian P Gent and Toby Walsh. Phase transitions and annealed theories: Number partitioning as a case study'. In *ECAI*, pages 170–174. PITMAN, 1996.

[148] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323, 2011.

[149] Surbhi Goel, Varun Kanade, Adam Klivans, and Justin Thaler. Reliably learning the relu in polynomial time. *arXiv preprint arXiv:1611.10258*, 2016.

[150] Surbhi Goel, Adam Klivans, and Raghu Meka. Learning one convolutional layer with overlapping patches. In *International Conference on Machine Learning*, pages 1783–1791. PMLR, 2018.

[151] Surbhi Goel and Adam R Klivans. Learning neural networks with two nonlinear layers in polynomial time. In *Conference on Learning Theory*, pages 1470–1499. PMLR, 2019.

[152] Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. In *Advances in Neural Information Processing Systems*, pages 6979–6989, 2019.

[153] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *arXiv preprint arXiv:1712.06541*, 2017.

[154] Lukas Gonon, Lyudmila Grigoryeva, and Juan-Pablo Ortega. Approximation bounds for random neural networks and reservoir systems. *arXiv preprint arXiv:2002.05933*, 2020.

[155] Francesco Guerra and Fabio Lucio Toninelli. The thermodynamic limit in mean field spin glass models. *Communications in Mathematical Physics*, 230(1):71–79, 2002.

[156] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018.

[157] Benjamin Haeffele, Eric Young, and Rene Vidal. Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing. In *International conference on machine learning*, pages 2007–2015, 2014.

[158] Benjamin D Haeffele and René Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

[159] Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

[160] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

[161] Christopher Harshaw, Fredrik Sävje, Daniel Spielman, and Peng Zhang. Balancing covariates in randomized experiments using the gram-schmidt walk. *arXiv preprint arXiv:1911.03071*, 2019.

[162] Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension bounds for piecewise linear neural networks. In *Conference on Learning Theory*, pages 1064–1068, 2017.

[163] Hamed Hatami, László Lovász, and Balázs Szegedy. Limits of locally–globally convergent graph sequences. *Geometric and Functional Analysis*, 24(1):269–296, 2014.

[164] David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.

[165] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[166] AJ Hoffman and HW Wielandt. The variation of the spectrum of a normal matrix. *Duke Mathematical Journal*, 20(1):37–39, 1953.

[167] Samuel B Hopkins, Pravesh K Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The power of sum-of-squares for detecting hidden structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 720–731. IEEE, 2017.

[168] Samuel B Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *Conference on Learning Theory*, pages 956–1006, 2015.

[169] Samuel Brink Klevit Hopkins. Statistical inference and the sum of squares method. 2018.

[170] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge University Press, 2012.

[171] Patrik O Hoyer. Non-negative sparse coding. In *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, pages 557–565. IEEE, 2002.

[172] Brice Huang and Mark Sellke. Tight lipschitz hardness for optimizing mean field spin glasses. *arXiv preprint arXiv:2110.07847*, 2021.

[173] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[174] Haiping Huang, KY Michael Wong, and Yoshiyuki Kabashima. Entropy landscape of solutions in the binary perceptron problem. *Journal of Physics A: Mathematical and Theoretical*, 46(37):375002, 2013.

[175] Sorin Istrail. Statistical mechanics, three-dimensionality and np-completeness: I. universality of intracatability for the partition function of the ising model across non-planar surfaces. In *STOC*, pages 87–96, 2000.

[176] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

[177] Mark Jerrum. Large cliques elude the metropolis process. *Random Structures & Algorithms*, 3(4):347–359, 1992.

[178] Mark Jerrum. *Counting, sampling and integrating: algorithms and complexity*. Springer Science & Business Media, 2003.

[179] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1724–1732. JMLR. org, 2017.

[180] Roger David Joseph and Louise Hay. The number of orthants in n-space intersected by an s-dimensional subspace. Technical report, CORNELL AERONAUTICAL LAB INC BUFFALO NY, 1960.

[181] Erich Kaltofen. Polynomial factorization 1987–1991. In *Latin American Symposium on Theoretical Informatics*, pages 294–313. Springer, 1992.

[182] Narendra Karmarkar and Richard M Karp. *The differencing method of set partitioning*. Computer Science Division (EECS), University of California Berkeley, 1982.

[183] Narendra Karmarkar, Richard M Karp, George S Lueker, and Andrew M Odlyzko. Probabilistic analysis of optimum partitioning. *Journal of Applied Probability*, pages 626–645, 1986.

[184] Richard M Karp. The probabilistic analysis of some combinatorial search algorithms. 1976.

[185] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in neural information processing systems*, pages 586–594, 2016.

[186] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

[187] Michael J Kearns and Robert E Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

[188] Jeong Han Kim and James R Roche. Covering cubes by random half cubes, with applications to binary neural networks. *Journal of Computer and System Sciences*, 56(2):223–252, 1998.

[189] Pravesh K Kothari, Ryuhei Mori, Ryan O'Donnell, and David Witmer. Sum of squares lower bounds for refuting any csp. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 132–145, 2017.

[190] Werner Krauth and Marc Mézard. Storage capacity of memory networks with binary couplings. *Journal de Physique*, 50(20):3057–3066, 1989.

[191] Abba M Krieger, David Azriel, and Adam Kapelner. Nearly random designs with greatly improved balance. *Biometrika*, 106(3):695–701, 2019.

[192] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[193] Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. *arXiv preprint arXiv:1907.11636*, 2019.

[194] Joseph Lauer and Nicholas Wormald. Large independent sets in regular graphs of large girth. *Journal of Combinatorial Theory, Series B*, 97(6):999–1009, 2007.

[195] Daniel Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2001.

[196] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[197] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257, 2016.

[198] Hanno Lefmann. A note on ramsey numbers. *Studia Sci. Math. Hungar*, 22(1-4):445–446, 1987.

[199] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Mmse of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 680–687. IEEE, 2015.

[200] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Phase transitions in sparse pca. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 1635–1639. IEEE, 2015.

[201] Avi Levy, Harishchandra Ramadas, and Thomas Rothvoss. Deterministic discrepancy minimization via the multiplicative weight update method. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 380–391. Springer, 2017.

[202] Kfir Y Levy. The power of normalization: Faster evasion of saddle points. *arXiv preprint arXiv:1611.04831*, 2016.

[203] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

[204] Yuanzhi Li, Tengyu Ma, and Hongyang R Zhang. Learning over-parametrized two-layer neural networks beyond ntk. In *Conference on Learning Theory*, pages 2613–2682. PMLR, 2020.

[205] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pages 597–607, 2017.

[206] Zhiyuan Li, Yi Zhang, and Sanjeev Arora. Why are convolutional nets more sample-efficient than fully-connected nets? *arXiv preprint arXiv:2010.08515*, 2020.

[207] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. *arXiv preprint arXiv:1711.01530*, 2017.

[208] Richard J Lipton. New directions in testing. *Distributed computing and cryptography*, 2:191–202, 1989.

[209] Yang P Liu, Ashwin Sah, and Mehtaab Sawhney. A Gaussian fixed point random walk. *arXiv preprint arXiv:2104.07009*, 2021.

[210] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in neural information processing systems*, pages 855–863, 2014.

[211] Shachar Lovett and Raghu Meka. Constructive discrepancy minimization by walking on the edges. *SIAM Journal on Computing*, 44(5):1573–1582, 2015.

[212] George S Lueker. A note on the average-case behavior of a simple differencing method for partitioning. *Operations Research Letters*, 6(6):285–287, 1987.

[213] Jiri Matousek. *Geometric discrepancy: An illustrated guide*, volume 18. Springer Science & Business Media, 1999.

[214] Raghu Meka, Aaron Potechin, and Avi Wigderson. Sum-of-squares lower bounds for planted clique. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 87–96, 2015.

[215] Ralph Merkle and Martin Hellman. Hiding information and signatures in trapdoor knapsacks. *IEEE transactions on Information Theory*, 24(5):525–530, 1978.

[216] Stephan Mertens. Phase transition in the number partitioning problem. *Physical Review Letters*, 81(20):4281, 1998.

[217] Marc Mézard, Thierry Mora, and Riccardo Zecchina. Clustering of solutions in the random satisfiability problem. *Physical Review Letters*, 94(19):197205, 2005.

[218] Hrushikesh Mhaskar, Qianli Liao, and Tomaso Poggio. Learning functions: when is deep better than shallow. *arXiv preprint arXiv:1603.00988*, 2016.

[219] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton. Acoustic modeling using deep belief networks. *IEEE transactions on audio, speech, and language processing*, 20(1):14–22, 2011.

[220] Andrea Montanari. Optimization of the sherrington-kirkpatrick hamiltonian. *arXiv preprint arXiv:1812.10897*, 2018.

[221] Andrea Montanari. Optimization of the sherrington-kirkpatrick hamiltonian. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1417–1433. IEEE, 2019.

[222] Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638. PMLR, 2018.

[223] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

[224] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

[225] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401, 2015.

[226] Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2603–2612. JMLR. org, 2017.

[227] Quynh Nguyen and Matthias Hein. The loss surface and expressivity of deep convolutional neural networks. 2018.

[228] Dmitry Panchenko. The parisi ultrametricity conjecture. *Annals of Mathematics*, pages 383–393, 2013.

[229] Dmitry Panchenko. *The Sherrington-Kirkpatrick model*. Springer Science & Business Media, 2013.

[230] Dmitry Panchenko. The parisi formula for mixed p-spin models. *The Annals of Probability*, 42(3):946–958, 2014.

[231] Dmitry Panchenko. On the k-sat model with large number of clauses. *Random Structures & Algorithms*, 52(3):536–542, 2018.

[232] Giorgio Parisi. Infinite number of order parameters for spin-glasses. *Physical Review Letters*, 43(23):1754, 1979.

[233] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, pages 2637–2646, 2017.

[234] Will Perkins and Changji Xu. Frozen 1-rsb structure of the symmetric ising perceptron. *arXiv preprint arXiv:2102.05163*, 2021.

[235] Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

[236] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and 6.441 (MIT)*, pages 2012–2016, 2016.

[237] Timothy Poston, C-N Lee, Y Choie, and Yonghoon Kwon. Local minima and back propagation. In *IJCNN-91-Seattle International Joint Conference on Neural Networks*, volume 2, pages 173–176. IEEE, 1991.

[238] Aditya Potukuchi. Discrepancy in random hypergraph models. *arXiv preprint arXiv:1811.01491*, 2018.

[239] Aditya Potukuchi. A Spectral Bound on Hypergraph Discrepancy. In Artur Czumaj, Anuj Dawar, and Emanuela Merelli, editors, *47th International Colloquium on Automata, Languages, and Programming (ICALP 2020)*, volume 168 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 93:1–93:14, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

[240] Prasad Raghavendra, Tselil Schramm, and David Steurer. High-dimensional estimation via sum-of-squares proofs. *arXiv preprint arXiv:1807.11419*, 6, 2018.

[241] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

[242] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 21, 2008.

[243] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in neural information processing systems*, pages 1313–1320, 2009.

[244] Mustazee Rahman and Bálint Virág. Local algorithms for independent sets are half-optimal. *Ann. Probab.*, 45(3):1543–1577, 05 2017.

[245] Mustazee Rahman and Balint Virag. Local algorithms for independent sets are half-optimal. *The Annals of Probability*, 45(3):1543–1577, 2017.

[246] Thomas Rothvoss. Constructive discrepancy minimization for convex sets. *SIAM Journal on Computing*, 46(1):224–234, 2017.

[247] Grant M Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *arXiv preprint arXiv:1805.00915*, 2018.

[248] Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

[249] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. *arXiv preprint arXiv:1712.08968*, 2017.

[250] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International Conference on Machine Learning*, pages 4433–4441. PMLR, 2018.

[251] Stefano Sarao Mannelli, Eric Vanden-Eijnden, and Lenka Zdeborová. Optimization and generalization of shallow neural networks with quadratic activation functions. *Advances in Neural Information Processing Systems*, 33:13445–13455, 2020.

[252] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *arXiv preprint arXiv:1708.06633*, 2017.

[253] Hanie Sedghi and Anima Anandkumar. Provable methods for training neural networks with sparse connectivity. *arXiv preprint arXiv:1412.2693*, 2014.

[254] Mariya Shcherbina and Brunello Tirozzi. Rigorous solution of the Gardner problem. *Communications in mathematical physics*, 234(3):383–422, 2003.

[255] David Sherrington and Scott Kirkpatrick. Solvable model of a spin-glass. *Physical review letters*, 35(26):1792, 1975.

[256] Zbynek Sidák. On multivariate normal probabilities of rectangles: their dependence on correlations. *The Annals of Mathematical Statistics*, 39(5):1425–1434, 1968.

[257] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

[258] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 2019.

[259] Paris Smaragdis and Shrikant Venkataramani. A neural network alternative to non-negative audio models. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 86–90. IEEE, 2017.

[260] Mohammadreza Soltani. *Provable Algorithms for Nonlinear Models in Machine Learning and Signal Processing*. PhD thesis, 2019. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2021-05-20.

[261] Mohammadreza Soltani and Chinmay Hegde. Towards provable learning of polynomial neural networks using low-rank matrix estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 1417–1426, 2018.

[262] Mohammadreza Soltani and Chinmay Hegde. Fast and provable algorithms for learning two-layer polynomial neural networks. *IEEE Transactions on Signal Processing*, 67(13):3361–3371, 2019.

[263] Mahdi Soltanolkotabi. Learning relus via gradient descent. In *Advances in Neural Information Processing Systems*, pages 2007–2017, 2017.

[264] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

[265] Mei Song, Andrea Montanari, and P Nguyen. A mean field view of the landscape of two-layers neural networks. *Proceedings of the National Academy of Sciences*, 115:E7665–E7671, 2018.

[266] Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

[267] Daniel Soudry and Elad Hoffer. Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv preprint arXiv:1702.05777*, 2017.

[268] Joel Spencer. Six standard deviations suffice. *Transactions of the American mathematical society*, 289(2):679–706, 1985.

[269] Mihailo Stojnic. Another look at the Gardner problem. *arXiv preprint arXiv:1306.3979*, 2013.

[270] Madhu Sudan. Maximum likelihood decoding of reed solomon codes. In *Proceedings of 37th Conference on Foundations of Computer Science*, pages 164–172. IEEE, 1996.

[271] Michel Talagrand. Intersecting random half cubes. *Random Structures & Algorithms*, 15(3-4):436–449, 1999.

[272] Michel Talagrand. The parisi formula. *Annals of mathematics*, pages 221–263, 2006.

[273] Michel Talagrand. *Mean field models for spin glasses: Volume I: Basic examples*, volume 54. Springer Science & Business Media, 2010.

[274] Michel Talagrand. *Mean Field Models for Spin Glasses: Advanced replica-symmetry and low temperature*. Springer, 2011.

[275] Matus Telgarsky. Benefits of depth in neural networks. *arXiv preprint arXiv:1602.04485*, 2016.

[276] Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3404–3413. JMLR. org, 2017.

[277] Li-Hui Tsai. Asymptotic analysis of an algorithm for balanced parallel processor scheduling. *SIAM Journal on Computing*, 21(1):59–64, 1992.

[278] Paxton Turner, Raghu Meka, and Philippe Rigollet. Balancing gaussian vectors in high dimension. In *Conference on Learning Theory*, pages 3455–3486. PMLR, 2020.

[279] Luca Venturi, Afonso S Bandeira, and Joan Bruna. Spurious valleys in two-layer neural network optimization landscapes. *arXiv preprint arXiv:1802.06384*, 2018.

[280] Luca Venturi, Afonso S Bandeira, and Joan Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. *Journal of Machine Learning Research*, 20:133, 2019.

[281] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[282] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[283] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[284] Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. On the margin theory of feedforward neural networks. *arXiv preprint arXiv:1810.05369*, 2018.

[285] Alexander S Wein. Optimal low-degree hardness of maximum independent set. *arXiv preprint arXiv:2010.06563*, 2020.

[286] E Weinan, Jiequn Han, and Arnulf Jentzen. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Communications in Mathematics and Statistics*, 5(4):349–380, 2017.

[287] James G Wendel. A problem in geometric probability. *Mathematica Scandinavica*, 11(1):109–111, 1962.

[288] David Williams. *Probability with martingales*. Cambridge university press, 1991.

[289] Robert O Winder. Single stage threshold logic. In *2nd Annual Symposium on Switching Circuit Theory and Logical Design (SWCT 1961)*, pages 321–332. IEEE, 1961.

[290] Lei Wu, Zhanxing Zhu, et al. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.

[291] Changji Xu. Sharp threshold for the Ising perceptron model. *arXiv preprint arXiv:1905.05978*, 2019.

[292] Benjamin Yakir. The differencing algorithm ldm for partitioning: a proof of a conjecture of karmarkar and karp. *Mathematics of Operations Research*, 21(1):85–99, 1996.

[293] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

[294] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[295] Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer relu networks via gradient descent. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1524–1534. PMLR, 2019.

[296] Kai Zhong, Zhao Song, and Inderjit S Dhillon. Learning non-overlapping convolutional neural networks with multiple kernels. *arXiv preprint arXiv:1711.03440*, 2017.

[297] Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4140–4149. JMLR. org, 2017.

[298] Pan Zhou and Jiashi Feng. The landscape of deep learning algorithms. *arXiv preprint arXiv:1705.07038*, 2017.

[299] Yi Zhou and Yingbin Liang. Critical points of neural networks: Analytical forms and landscape properties. *arXiv preprint arXiv:1710.11205*, 2017.