Provable Algorithms for Learning and Variational Inference in Undirected Graphical Models

by

Frederic Koehler

Submitted to the Department of Mathematics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Mathematics and Statistics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author .....
Department of Mathematics
May 6, 2021

Certified by .....
Ankur Moitra
Associate Professor of Mathematics
Thesis Supervisor

Certified by .....
Elchanan Mossel
Professor of Mathematics
Thesis Supervisor

Accepted by .....
Jonathan Kelner
Chairman, Department Committee on Graduate Theses

Provable Algorithms for Learning and Variational Inference in Undirected Graphical Models

by

Frederic Koehler

Submitted to the Department of Mathematics
on May 6, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Mathematics and Statistics

Abstract

Graphical models are a general-purpose tool for modeling complex distributions in a way which facilitates probabilistic reasoning, with numerous applications across machine learning and the sciences. This thesis deals with algorithmic and statistical problems of learning a high-dimensional graphical model from samples, and related problems of performing inference on a known model, both areas of research which have been the subject of continued interest over the years. Our main contributions are the first computationally efficient algorithms for provably (1) learning a (possibly ill-conditioned) walk-summable Gaussian Graphical Model from samples, (2) learning a Restricted Boltzmann Machine (or other latent variable Ising model) from data, and (3) performing naive mean-field variational inference on an Ising model in the optimal density regime. These different problems illustrate a set of key principles, such as the diverse algorithmic applications of "pinning" variables in graphical models. We also show in some cases that these results are nearly optimal due to matching computational/cryptographic hardness results.

Thesis Supervisor: Ankur Moitra
Title: Associate Professor of Mathematics

Thesis Supervisor: Elchanan Mossel
Title: Professor of Mathematics

Acknowledgments

First, I would like to thank my thesis advisors, Elchanan Mossel and Ankur Moitra. Their ability to quickly see the connections between different areas of mathematical research has been both a major help and inspiration to me throughout my time at MIT. Their patient advising has also been invaluable when it came to decision making and answering the numerous other questions which came up during the PHD. I am also grateful to Guy Bresler for serving as a member of my thesis committee, as well as for allowing me to participate in his group meetings.

Doing research is much more enjoyable with other people and I am grateful to have worked with a number of brilliant collaborators during my PHD, including Vishesh Jain, Andrej Risteski, Viraj Mehta, Govind Ramnarayan, Ronen Eldan, Ofer Zeitouni, Jingbo Liu, Younhun Kim, Guy Bresler, Matthew Brennan, Dylan Foster, Tselil Schramm, Linus Hamilton, Jon Kelner, Raghu Meka, Sitan Chen, Morris Yau, Surbhi Goel, Adam Klivans, Enric Boix-Adsera, John Urschel, Adam Hesterberg, Erik Demaine, Jayson Lynch, Mehtaab Sawhney, and Dhruv Rohatgi.

I'm grateful to my officemates Jake Wellens, John Urschel, Dheeraj Nagaraj, Enric Boix-Adsera, Matthew Brennan, Sinho Chewi, and Austin Stromme for creating a great environment to brainstorm, do research, and decompress from the stresses of research life. I am particularly grateful to Matt for many inspirational conversations about mathematics and research, as well as for being a great personal friend.

I personally benefited a lot from being in the diverse environments of the MIT math department, IDSS, and the theory group at MIT. Especially, I am grateful to the classes and reading groups I participated where I learned a lot and which provided very helpful background for my own research. I'm also grateful to my undergraduate teachers and mentors and especially to Samir Khuller for patiently introducing me to math and algorithms research in the very beginning.

Outside of research, I'm grateful to my family, including my parents, brother, and sister, and friends for their support and camaraderie. In particular I want to thank Dylan McKay, Nadine Javier, David Jacobowitz, Matthew Smith, Jamila Pegues,

Tiffany Ho, Iniko Ntosake, Thomas Reeves, Samuel Cheng, Abhinav Khanna, Kiran Vodrahalli, Jason Altschuler, Mark Sellke, Monica Gonzalez, Joshua Lam, Ted Grunberg, David Qiu, Derek Leung, Vishal Patil, Sidhanth Mohanty, Julie Takagi, Paxton Turner, Julien Clancy, Christian Gaetz, Robin Elliott, Andy Senger, Thao Do, Jonathan Tidor, Boya Song, Yunfei Liu, and Scott Lawrence.

Contents

Chapter 1

Introduction

In this thesis, we present new algorithms and theoretical results for some basic algorithmic problems of learning and inference in high-dimensional graphical models. Graphical models are a powerful framework for modelling high-dimensional distributions in a way that is interpretable and enables sophisticated forms of inference and reasoning. One of the key tools for reasoning in graphical models is the *Markov property*, which allows us to formally reason about conditional independencies between different random variables in a way that plays a crucial role in applications (e.g. causal inference [153, 155]).

They are extensively used in a variety of disciplines including the natural and social sciences where, besides originating as fundamental models of magnets and statistical field theories (e.g. [151, 74]), they have been used in a vast number of other settings such as models for the structure of gene regulatory networks (e.g. [196, 137, 164, 18]), of connectivity and learning in the brain (e.g. [139, 95, 187]), and the flocking behavior of birds (e.g. [20]). In many contexts, the structure of interactions between different observed variables is unknown a priori and the goal is to infer this structure in a sample-efficient way from data. There has been decades of research on various formulations of this problem, both theoretically and empirically: for example, provable algorithms have been developed for learning tree-structured graphical models [44], for learning models on graphs of bounded tree-width [103], for learning Ising models on general graphs of bounded degree [34, 29, 189, 108] and in a variety of

other contexts like Gaussian graphical models (e.g. [135]). For the most part, the main interest has been on learning under the assumption that the underlying model is sparse. Sparsity is a natural assumption since many applications are in a sample-starved regime where the learning problem is information-theoretically impossible without sparsity. Sparse models are generally considered to be more interpretable than their dense counterparts since they satisfy conditional independence relations which make probabilistic reasoning easier.

In this thesis, we largely focus on two fundamental classes of graphical models: multivariate Gaussian distributions (referred to as a Gaussian Graphical Model (GGM) in this context), and their precise analogues on the hypercube, known as *Ising models* and long studied in statistical physics. We focus on two related problems: (1) learning a graphical model from data in a sample-efficient way, and (2) estimating the partition function (and related quantities) of a known model. For Gaussian distributions, the normalizing constant has a well-known exact formula ($\sqrt{2\pi \det \Sigma}$) so we focus on the problem (1): in Chapter 1, we give the first efficient algorithms which, for the large class of walk-summable GGMs, succeed in recovering the model from few samples regardless of the condition number of the model. For Ising models, the fully-observed learning problem has been extensively studied but relatively little was known about the situation when there are latent variables[1]: we establish general results for learning in the latent variable setting in Chapters 2 and Chapter 6. Methods for estimating the partition function (e.g. of the learned model, in order to do inference) are discussed in Chapters 2,3, and 4.

One of the goals of of this thesis is to illustrate some common themes which appear throughout sometimes seemingly unrelated problems in graphical models. We state a few of these themes explicitly here:

1. *Pinning as a way to tame strong correlations.* One tool used in the analysis of variational methods in Chapter 3 is a convenient Lemma used in the analysis of

[1]For a Gaussian distribution, latent variables do not significantly change the difficulty of learning because marginalizing out a coordinate in a Gaussian results again in a Gaussian; this is very much not true in an Ising model, where arbitrarily complex higher order interactions can in fact be created — see Chapter 2.

"correlation rounding" [141, 12, 156]. This Lemma is a particular instantiation of the following general principle: if many vectors (or random variables) are highly correlated, then selecting out a few of them and projecting orthogonal to them may greatly reduce the size (or correlation) of these vectors. We see the same kind of idea appear in a very different technical form in Chapter 5, where one of the key conceptual ideas is that conditioning on (or "pinning") a single variable in the graphical model can tame the correlations in nearby nodes dramatically, and in Chapter 2, where we see a similar effect when pinning a set of variables to plus. This idea is also one natural way to motivate the all-ones initialization used in Chapter 4.

2. *Inference as a tool in learning.* The problems of learning a graphical model and performing inference (in the sense of e.g. computing marginals) are often treated separately. Of course, there are exceptions like in Maximum-Likelihood Estimation where the problem of estimating the normalizing constant naturally appears. In this thesis we see a few applications of a different flavor: in Chapter 2, we see how ideas from "influence maximization" (in the sense of e.g. [106]) in social networks actually can be used for learning, by running a version of an influence maximization algorithm directly on data; in Chapter 6 we see how belief propagation can be a valuable tool for the *analysis* of a natural node-wise regression approach, and in Chapter 5 we see how some technical ideas used in the analysis of Gaussian belief propagation and Bayesian active learning can be applied to the structure learning problem. A trick used in Chapter 6 based on an idea in [90] goes in the *reverse direction*: using access to data to avoid having to compute a normalizing constant, by performing a logistic regression instead.

3. *Conditional correlation as a canonical measure of edge strength.* In order to learn a combinatorial structure like the graph of a graphical model from data, there usually needs to be a cutoff which lets us drop edges with interactions too weak to be seen from a reasonable amount of data; otherwise, even estimating

a two-variable graphical model will be impossible, because we cannot test from data if two random variables are *exactly* uncorrelated vs. very weakly correlated. For some parametric models, it's possible to specify the notion of edge nondegeneracy in terms of the parameters of the model, but for more complex situations like latent variable models, where there are multiple ways to parameterize the same distribution (see Chapter 2) this does not always make sense. The Markov property suggests a natural way to handle this issue: if the the conditional mutual information $I(X_i; X_j | X_{\sim i,j}) \neq 0$ (or conditional covariance $\mathrm{Cov}(X_i, X_j | X_{\sim i,j}) \neq 0$ related by Pinsker's inequality [47]) then an edge must be present between nodes i and j to satisfy the Markov property, so requiring a quantitative lower bound on this quantity is a natural way to define nondegeneracy of an edge, and this is used in both Chapter 5 and Chapter 6. The conditional mutual information also appears naturally in the aforementioned correlation rounding Lemma used in Chapter 3, where we use small average conditional mutual information as a proxy for the measure being well approximated by a product measure (i.e. a graphical model with no edges).

4. *Combinatorial algorithms as a statistical aid to convex programs.* In Chapter 5, we see that the standard convex program for sparse linear regression, the Lasso, performs poorly in some examples but that this can be helped by preprocessing with a single step of a combinatorial method (*forward selection*/Orthogonal Matching Pursuit); in Chapter 3 we see that the convex program analyzed can sometimes be made more efficient (computationally, and in terms of vertex query complexity) by simple subsampling of the graphical model. In Chapters 2 and Chapter 6 we see natural combinations of regression methods and combinatorial pruning methods used together. In Chapter 5 we see also an application in the reverse direction: by proving that Lasso succeeds with an appropriate preconditioning step, we morally obtain the corresponding result for the natural greedy forward-backward method by comparison to the Frank-Wolfe algorithm applied to the Lasso convex program.

1.1 Graphical Models: A Crash Course

In this section, we quickly overview the definition of the Markov property and discuss some relevant background in graphical models; the focus of this thesis is undirected graphical models, but we also explain their connection to, and how they can arise from, directed models. We have made this section largely self-contained but also relatively concise: the interested reader may want to refer to a more in-depth treatment of the material such as [118, 22, 113].

Undirected graphical models. Suppose that $G = (V, E)$ is an undirected graph and for every $v \in V$, we attach a random variable X_v. The random variable X_v can be valued in an arbitrary space; usually it will be a number, vector, or matrix. We let X without a subscript denote the collection $(X_v)_{v \in V}$: for example, if each X_v is a real-valued random variable, then X is a random vector in $\mathbb{R}^V$; in the general setting, we can think of X as a *random function* mapping $v \mapsto X_v$. Given a set $A \subset V$, we will let X_A denote $(X_v)_{v \in A}$.

In this setup, we say that X is a *Markov Random Field* (MRF) over G if it satisfies the following *Markov Property*: for any triple of sets $A, B, S \subset V$ such that all paths starting from a vertex in A and ending in a vertex in B pass through a vertex in S, we have that X_A is conditionally independent of X_B given X_S. In other words, the graph condition is that S is a separator between A and B in the graph G; we also note that this property is equivalent to the version where we only consider maximal A and B so that $S \cup A \cup B = V$. The pair (X, G) is also referred to as an *undirected graphical model*, and we refer to G by itself as the *structure graph* of the MRF. This general definition encapsulates a large number of probabilistic models of interest:

1. A discrete-time *Markov chain* is a Markov random field over the infinite path graph with vertex set $V = \mathbb{Z}_{\geq 0}$ and an edge between each adjacent number $(t, t+1)$. In this context, we think of the vertex set V as representing time. For

example, a simple random walk with Rademacher steps is the process

$$X_t = \sum_{t=1}^{\infty} \epsilon_t$$

where $\epsilon_1, \epsilon_2, \ldots$ is an infinite sequence of independent Rademacher random variables, i.e. random coin flips valued in $\{\pm 1\}$.

2. An *Ising model* over a graph G is a random vector $X \in \{\pm 1\}^V$ such that

$$\Pr(X = x) \propto \exp\left(\langle x, Jx\rangle/2 + \langle h, x\rangle\right)$$

where $J : \mathbb{R}^{n\times n}$ is the *interaction matrix* and the interactions respect the graph structure, i.e. there is an edge between i and j iff $J_{ij} \neq 0$. Here the notation $\propto$ indicates equality up to a constant normalizing factor, so that the probabilities sum to 1: this normalizing factor is also referred to as the *partition function* of the model, and can be explicitly written as

$$Z = \sum_{x\in\{\pm 1\}^n} \exp\left(\langle x, Jx\rangle/2 + \langle h, x\rangle\right).$$

We note that this general definition of an Ising model captures both classical lattice Ising models and also spin glass models like the Edwards-Anderson (EA) and Sherrington-Kirkpatrick (SK) model.

3. A *Gaussian Graphical Model* over G is a non-degenerate Gaussian distribution $N(\mu, \Sigma)$ which respects the graph structure in the sense that the *precision matrix* $\Theta = \Sigma^{-1}$ respects the graph structure, i.e. there is an edge between i and j iff $\Theta_{ij} \neq 0$. This is completely analogous to the Ising model as the density is of the form

$$p(x) \propto \exp\left(-\langle (x-\mu), \Theta(x-\mu)\rangle/2\right).$$

In fact, if we adopt measure theoretic notation and write

$$\frac{dp}{dq}(x) \propto \exp\left(-\langle (x-\mu), \Theta(x-\mu)\rangle/2\right) \tag{1.1}$$

then if the base measure q is the Lebesgue measure on $\mathbb{R}^n$, we get for p a general Gaussian distribution and if the base measure q is the uniform measure on $\{\pm 1\}^n$ we get for p a general Ising model.

The fact that the Ising model and Gaussian Graphical Model are both Markov random fields can be seen from the fact that their densities can be written in terms of clique potentials on the graph, i.e. in the form (1.2) which we discuss below.

With appropriate generalization, this definition can also encompass fields over a non-discrete vertex set V. For example, the continuous analogue/limit of the simple random walk is a *Brownian motion* which is field over $\mathbb{R}$, and there are higher-dimensional examples such as *Gaussian Free Fields* [167] and other statistical field theories, like the φ^4-field theory [74, 151]. However, these settings involve some considerable technical complications: for example, the two-dimensional Gaussian free field does not exist as a function $v \mapsto X_v$ but instead as a distribution which test functions can be integrated against.

Equivalent characterizations: Markov blanket and Hammersley-Clifford Theorem. An equivalent description of a Markov Random Field is in terms of the following local version of the Markov property: X_v is conditionally independent of all other nodes $X_{\sim(\mathcal{N}(v)\cup\{v\})}$ given its neighbors $X_{\mathcal{N}(v)}$ where $\mathcal{N}(v) := \{u : u \sim v\}$ is the graph-theoretic neighborhood of node v. This fact is implied from the (global) Markov property since the neighborhood separates v from the rest of the graph. The equivalence of these properties can be seen through a third and very useful characterization called the *Hammersley-Clifford Theorem* (see e.g. [19]), which for simplicity we state for X valued in a finite set with $\Pr(X = x) \neq 0$ for all x. The Theorem says that X is a Markov random field over G iff there exist potentials f_K

on the cliques K of G such that

$$\Pr(X = x) \propto \exp\left(\sum_{K \in G} f_K(x_K)\right). \tag{1.2}$$

A simple proof of this fact from the local Markov property proceeds by decomposing $\log \Pr(X = x)$ according to the *Efron-Stein decomposition* [148], which is closely related to the inclusion-exclusion proofs of this result in the literature [19]. Concretely, in the case of $x \in \{\pm 1\}^n$, the proof proceeds by writing $\log \Pr(X = x)$ out as a polynomial, i.e. writing $f(x) := \log \Pr(X = x) = \sum_{S \subset [n]} \hat{f}(S) \prod_{i \in S} x_i$ for some $\hat{f}(S) \in \mathbb{R}$, using that the parities $\left(\prod_{i \in S} x_i\right)_{S \subset n}$ form a basis for the space of functions on the hypercube $\{\pm 1\}^n \to \mathbb{R}$, which follows from the orthogonality relation for

$$\frac{1}{2^n} \sum_{x \in \{\pm 1\}^n} \left(\prod_{i \in S} x_i\right)\left(\prod_{j \in T} x_j\right) = \mathbb{1}(S = T),$$

which follows from the fact that for any $j \in (T \setminus S) \cup (S \setminus T)$ that the sumand is odd in x_j; see [148] for a more detailed discussion. Then by Bayes rule

$$\Pr(X_i = x_i | X_{\sim i} = x_{\sim i}) = \frac{\Pr(X = x)}{\Pr(X_{\sim i} = x_{\sim i})} = \frac{\exp\left(x_i \sum_{S : i \in S} \hat{f}(S) \prod_{j \in S \setminus \{i\}} x_j\right)}{\sum_{x_i'} \exp\left(x_i' \sum_{S : i \in S} \hat{f}(S) \prod_{j \in S \setminus \{i\}} x_j\right)}$$

and the local Markov property says that this must be a function only of x_i and $x_{\mathcal{N}(i)}$. Hence the function on $\{\pm 1\}^{n-1}$ defined by

$$g_i(x_{\sim i}) := \sum_{S : i \in S} \hat{f}(S) \prod_{j \in S \setminus \{i\}} x_j$$

can only depend on $x_i, x_{\mathcal{N}(i)}$, so for any other S with $i \in S$, $j \in S$, and $j \notin i \cup \{i\}\mathcal{N}(i)$ we have

$$\hat{f}(S) = \frac{1}{2^{n-1}} \sum_{x \in \{\pm 1\}^{n-1}} x_S g_i(x) = 0$$

as the summand is odd in x_j. Hence we get that for all S with $i \in S$ and $S \not\subset i \cup \mathcal{N}(i)$ that $\hat{f}(S) = 0$. Applying this argument for every node in the graph, we see that

$\hat{f}(S) = 0$ unless the set of vertices in S is a clique of G and this proves the Hammersley-Clifford Theorem by taking potentials $f_S(x_S) := \hat{f}(S) \prod_{i \in S} x_i$.

Finally, from the Hammersley-Clifford characterization (1.2) we can prove the global Markov property because if removing separator S splits the graph G into two sets of nodes A and B, then

$$\Pr(X = x | X_S = x_S) \propto \exp\left(\sum_{K:|A\cap K|>0} f_K(x)\right) \cdot \exp\left(\sum_{K:|B\cap K|>0} f_K(x)\right),$$

the first term on the rhs involves only x_A and x_S, the second term on the rhs involves on x_B and x_S, and so the conditional density factorizes proving conditional independence.

Directed graphical models and moralization. The *Markov Random Fields* we described above naturally live on an undirected graph. There are also natural notions of graphical models living on a directed graph which play an important role in some applications. The most common class of directed models are called *Bayes Networks*, which live on a *Directed Acyclic Graph* (DAG); they are characterized by an appropriate version of the Markov property: for any nodes v and u not a descendant of v in the DAG, we have that X_v is conditionally independent of X_u given its parents $X_{\mathrm{pa}(v)}$. Here $\mathrm{pa}(v)$ is the set of nodes which have an arc pointing into node v. This property is called the *local Markov property*; there is also an equivalent definition of Bayes networks in terms of a more global characterization involving the concept of d-separation, see [118, 153]. See also [118] for some other versions of directed graphical models, including versions which combine both directed and undirected edges.

The key connection between undirected and directed models is a process called *moralization*, which takes as input a directed graphical model (X, H) where X is a Bayes network over the directed graph H, and produces a corresponding undirected G such that X is a Markov random field over G. The process is simple: node v is connected to u in the undirected model if in the directed graph H either: (1) u is a parent of v, (2) u is a child of v, or (3) u is the parent of a child of v. We describe

the reason for this informally: including only the neighbors of (1) would make v independent of everything but its descendants, so we need to include nodes of type (2), i.e. the children as well; this separates v from the descendants of those children, but now the parents of the children may be indirectly useful in predicting X_v so we need to add the nodes of type (3).

We now make this precise. First observe that a directed graphical model admits the following factorization:

$$\Pr(X = x) = \prod_v \Pr(X_v = x_v \mid X_{\mathrm{pa}(v)=x_{\mathrm{pa}(v)}})$$

which follows by picking a topological ordering $v(1), \ldots, v(n)$ of the nodes in the directed graph, applying the the chain rule $\Pr(X = x) = \prod_{i=1}^n \Pr(X_{v(i)} \mid X_{v(<i)})$ and using the local Markov property stated above. Now let $\mathcal{N}(v)$ be the union of (1) the parents of node v, (2) the children of node v, and (3) the parents of the children of node v. By Bayes rule (abbreviating $\Pr(\cdot|X_S) = \Pr(\cdot|X_S = x_S)$ where the fixing of X_S is clear from context) we can compute the conditional law

$$\begin{aligned}
&\Pr(X_u = x_u \mid X_{\sim u} = x_{\sim u}) \\
&= \frac{\Pr(X = x)}{\sum_{x'_u} \Pr(X_{\sim u} = x_{\sim u}, X_u = x'_u)} \\
&= \left(\frac{\prod_v \Pr(X_v = x_v | X_{\mathrm{pa}(v)})}{\prod_{v: u \notin \mathrm{pa}(v)} \Pr(X_v = x_v | X_{pa(v)})} \right) \\
&\quad \left(\frac{1}{\sum_{x'_u} \Pr(X_u = x'_u | X_{pa(u)}) \prod_{v: u \in \mathrm{pa}(v)} \Pr(X_v = x_v | X_u = x'_u, X_{\mathrm{pa}(v) \setminus u})} \right) \\
&= \frac{\Pr(X_u = x_u | X_{pa(u)}) \prod_{v: u \in \mathrm{pa}(v)} \Pr(X_v = x_v | X_{pa(v)})}{\sum_{x'_u} \Pr(X_u = x'_u | X_{pa(u)}) \prod_{v: u \in \mathrm{pa}(v)} \Pr(X_v = x_v | X_u = x'_u, X_{\mathrm{pa}(v) \setminus u})}
\end{aligned}$$

and see that $\Pr(X_u = x_u \mid X_{\sim u} = x_{\sim u})$ depends only on x_u and (1) $x_{\mathrm{pa}(u)}$ i.e. the value of the parents of u, (2) x_v for $u \in \mathrm{pa}(v)$ i.e. the value of the children of u, and (3) $x_{\mathrm{pa}(v) \setminus u}$ for v with $u \in \mathrm{pa}(v)$, i.e. the value of the parents of the children of u.

Unlike for undirected graphical models, it usually is not possible to recover the directed graph structure of a Bayes network given samples from the distribution. (In

other words, the directed graph structure is not *identifiable* from observational data.) For example, the distribution of a pair of random variables (X_1, X_2) can be presented by a Bayes network with a single edge from X_1 to X_2, or with a single edge in the reverse direction. One solution to this problem is to perform list recovery, in the sense of finding all DAGs consistent with the observed data (see e.g. [8]). Another solution is to assume that the Bayes network is *causal*, in the sense that it describes not only the joint distribution of the random variables but also the behavior under *interventions*, where a random variable is forced to a particular value (this is different from conditioning). In this case, given the ability to observe the results of arbitrary interventions the network can be recovered, see [153, 155]. Since interventions are often expensive, it is desirable to minimize the number of interventions performed and one method used in practice is to learn the undirected structure first, see e.g. [125].

Conditional Laws and Related Approximations. An important role throughout this thesis is played by the condition law $p(X_i|X_{\sim i})$ for p a Markov random field. For simplicity, we again consider the case of discrete MRFs, though the following discussion generalizes straightforwardly to continuous ones. When $p(x) \propto \exp(f(x))$ we know by the local Markov property that

$$p(X_i = x_i|X_{\sim i} = x_{\sim i}) = \frac{\exp(f(x))}{\sum_{x_i'} \exp(f(x_i', x_{\sim i}))} \tag{1.3}$$

where the notation $f(x_i', x_{\sim i})$ means to apply f to the vector $(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)$. One of the most important aspects of (1.3) is that the right hand side can be evaluated efficiently (i.e. in polynomial time), as long as we can compute f in polynomial time and the state space of x_i' is polynomial size. This is generally not true of the overall pmf $p(x_i)$, which is very often computationally hard to compute, since the normalizing constant involves a sum over $\exp(O(n))$ many states (see e.g. [171] for a strong computational hardness result). Because of this distinction, learning algorithms based on pseudolikelihood methods or node-wise regression (i.e. trying to

predict one node in a graphical model from the others) are often nicer to work with than their "global" alternatives like Maximum-Likelihood Estimation, which may be computationally intractable.

The equation (1.3) also motivates some approximations to p. In particular, making an approximation that the rhs of (1.3) is a function only of x_i (which should be valid if it concentrates in the randomness over $x_{\sim i} \sim p$) suggests that p will behave roughly as a product measure and leads to the well-known *naive mean-field approximation* analyzed in Chapter 3. The failure of that approximation in some models leads to the consideration of alternatives; one such alternative, which is exact on tree models (and can be thought of as a natural generalization of (1.3) to subsets of nodes which form a tree), is the combination of the *Bethe approximation* and the corresponding *belief propagation* algorithm which plays a key role in Chapter 6; it also implicitly appears in Chapter 3 where we use that the partition function of the SK model is approximated by a corrected mean-field approximation called the TAP free energy [181]. In Gaussian Graphical Models, belief propagation ends up to be an iterative algorithm for solving linear systems and its convergence is guaranteed under a criterion known as *walk-summability* [131]: the analysis of such walk-summable GGMs (in our case, using electrical methods) plays a key role in Chapter 5.

Other technical preliminaries. Throughout this thesis, we make use of fundamental tools from high-dimensional probability like concentration inequalities — see [188] for a reference. Otherwise, we generally introduce the needed background (e.g. from electrical flows, statistical learning theory, approximation algorithms, statistics, information theory) where it is needed, inside of the relevant Chapter(s).

1.2 Overview of Chapters

In the first Section of the introduction, we gave a high-level description of the contents of this thesis (i.e. what is it about?). Here we give a more detailed overview of the chapter-by-chapter contents of this thesis, which is partly based on published

works [33, 97, 111, 105, 79]. First, we define a couple important terms more carefully. In this thesis, we will focus on two related problems in graphical models: learning and inference. In *learning*, our goal will generally be to estimate the structure graph G given m samples $X_1, \ldots, X_m$ which are i.i.d. (independent and identically distributed) copies of the random variable X. In *inference*, our goal will be to efficiently determine properties (e.g. marginals, conditional laws) of a graphical model given a description of it. (This should not be confused with the term *statistical inference* which encompasses learning as well.) The chapters of this work mostly focus on one or the other but often involve both tasks.

Chapter 2: Learning Latent Variable Models via Influence Maximization. In this chapter, we study how to learn Ising models when there are *latent variables*, i.e. we are given access to samples where only a fixed subset of observable nodes are revealed. A particular subcase of interest is the *Restricted Boltzmann Machine* [90] which lives on a bipartite graph. In the latent variable setting, recovering the graph structure of the entire model is not possible (there is not enough information), but we can learn the induced Markov random field on the revealed variables. We give some hardness results for this problem, based on a cryptographically hard problem known as learning a *sparse parity with noise*, and give positive results when the model is *ferromagnetic*, i.e. the external field and interaction parameters are nonnegative. We also discuss how methods for estimating the partition function based on algorithmic diagrammatic expansions [16] can be applied in this setting.

Chapter 3: Convex Hierarchies, Naive Mean-field Approximation, and Correlation Rounding. We consider a different approach to estimating the partition function, known as *variational methods*[2]. We strengthen results of [159, 17] by showing how to round from (a convex relaxation of) the Gibbs measure of an Ising model to a product measure with similar free energy, under the natural mean-field condition $\|J\|_F^2 = o(n)$ from [17]. As a consequence, we resolve a question about

[2]For a connection between variational methods and the aforementioned diagramatic expansions, see e.g. [181].

correlation rounding from the approximation algorithms literature [3] using rigorous results in spin glass theory [2]. We also generalize the results in a natural way to models with higher-order interactions.

Chapter 4: Landscape Analysis of Naive Mean-Field Approximation in Ferromagnetic Models. We continue the analysis of the naive mean-field approximation from the previous chapter, proving that in ferromagnetic models (as considered in Chapter 2) the standard heuristic method for maximizing the variational free energy provably succeeds given all-ones initialization, despite the fact that it is a first-order optimization method (like gradient descent) on a nonconvex objective.

Chapter 5: Learning GGMs without a Well-Condition Assumption. We return to the problem of learning graphical models, but this time focus on the most popular class of continuous models: Gaussian Graphical Models (GGMs). We observe that many graphical models with interesting large-scale behavior are not well-conditioned (i.e. their covariance matrices are not well-conditioned), and hence previous theoretical guarantees for learning GGMs are either not computationally efficient or are statistically suboptimal in the number of samples they require; we also confirm this in simulations. We show two different ways to resolve the problem: either by using a combinatorial forward-backward method to learn the model (in the spirit of Chapter 2), or by preconditioning the Lasso with a greedy pinning operation (in the spirit of Chapter 3).

Chapter 6: Learning RBMs with Bounded Weights. We return to the problem of learning Restricted Boltzmann Machines (RBMs) from samples. We study the algorithmic difficulty of this problem when we parameterize the complexity of the model by its weights, as is commonly done in e.g. generalization bounds [13]. We show nearly matching upper and lower bounds in this setting: a key insight in the proof of the upper bounds is a connection between RBM learning and feedforward neural network learning which comes from the *belief propagation* algorithm in probabilistic inference, a close relative of the mean-field iterative method analyzed in

Chapter 4. This also suggests that RBMs and related models could serve as natural distributional settings to study learning of feedforward networks, and we explore the consequences of this in a setting similar to Chapter 2, revisiting some classical ideas for using RBMs in supervised learning from [90].

Chapter 2

Learning Latent Variable Models via Influence Maximization

2.1 Introduction

2.1.1 Background

The presence of unobserved (or *latent*) variables is of fundamental importance in a wide range of applications. Latent variable models can capture much more complex dependencies among the observed variables than fully observed models, because the variables can influence each other through unobserved mechanisms. In this way, such models allow scientific theories that explain data in a more parsimonious way to be learned and tested. They can also be used to perform dimensionality reduction [91] and feature extraction [45] and thus serve as a basis for a variety of other machine learning tasks.

Despite their practical importance, the problem of learning graphical models with latent variables has seen much less progress than the fully-observed setting. In one application domain, phylogenetic reconstruction, there has been a lot of activity, e.g. [63, 49, 175, 176, 144], however these results are all quite specific to the setting of tree graphical models. Otherwise, the only works we are aware of are the following: Chadrasekaran et al. [42] studied Gaussian graphical models with latent variables

and sparsity and incoherence constraints. The marginal distribution on the observed variables is also a Gaussian graphical model, so it is straightforward to learn its distribution. However their focus was on discovering latent variables whose inclusion in the model "explains away" many of the observed dependencies. Anandkumar and Valluvan [5] were the first to give provable algorithms for learning discrete graphical models with latent variables, although they need rather strong conditions to do so. They require both that the graphical model is locally treelike and that it exhibits correlation decay.

In this Chapter (and also Chapter 6) we study Restricted Boltzmann Machines (or RBMs), a widely-used class of graphical models with latent variables that were popularized by Geoffrey Hinton in the mid 2000s. In fact, our results will extend straightforwardly to general Ising models with latent variables. An RBM has n_1 observed (or visible) variables $X_1, X_2, \ldots, X_n$ and n_2 latent (or hidden) variables $H_1, H_2, \ldots, H_m$ and is described by

(1) an $n_1 \times n_2$ interaction matrix W

(2) a length n vector $b^{(1)}$ and a length m vector $b^{(2)}$ of external fields/biases

Then for any $x \in \{\pm 1\}^{n_1}$ and $h \in \{\pm 1\}^{n_2}$, the probability that the model assigns to this configuration is given by:

$$\Pr(X = x, H = h) = \frac{1}{Z} \exp\left(x^T J h + \sum_{i=1}^{n_1} b_i^{(1)} x_i + \sum_{i=1}^{n_2} b_j^{(2)} h_j \right)$$

where Z is the partition function. We sometimes write $n = n_1 + n_2$ for the total number of nodes, consistent with our notation for general Ising models. It is often convenient to think about an RBM as a weighted bipartite graph whose nodes represent variables and whose weights are given by W. This family of models has found a number of applications including in collaborative filtering [162], topic modeling [92] and in deep learning where they are layered on top of each other to form deep belief networks [89]. As the number of layers grows, they can capture increasingly complex hierarchical dependencies among the observed variables.

We focus on the problem of learning RBMs from i.i.d. samples of the observed variables, with particular emphasis on the practically relevant case where the latent variables have low degree. What makes this challenging is that even though the variables in the RBM have only pairwise interactions, when the latent variables are marginalized out we can (and usually do) get higher-order interactions. Indeed, for general graphical models with latent variables and pairwise interactions, Bogdanov, Mossel and Vadhan [25] proved learning is hard (assuming $NP \neq RP$) by showing how the distribution on observed variables can simulate the uniform distribution on satisfying assignments of any given circuit. We note that this construction requires a large number (at least one for each gate) of interconnected latent variables and that the hard instances are highly complex because they come from a series of circuit manipulations. Beyond learning, Long and Servedio [126] proved that for RBMs a number of other related problems are hard, including approximating the partition function within an exponential factor and approximate inference and sampling.

The previous work leaves the following question unresolved: *Are there natural and well-motivated families of Ising models with latent variables that can be efficiently learned?* We will answer this question affirmatively in the case of *ferromagnetic* RBMs and (more generally) ferromagnetic Ising models with latent variables, which are defined as follows: A ferromagnetic RBM is one in which the interaction matrix and the vectors of external fields are nonnegative. On the other hand, we give a negative result showing that without ferromagneticity, even in the highly optimistic case when there are only a constant number of latent variables with bounded degree the problem is as hard as sparse parity with noise. This establishes a dichotomy that is just not present in the fully-observed setting.

Historically, ferromagneticity is a natural and well-studied property that plays a key role in many classic results in statistical physics and theoretical computer science. For example, the Lee-Yang theorem [121] shows that the complex zeros of the partition function of a ferromagnetic Ising model all lie on the imaginary axis — this property does not hold for general Ising models. Ferromagnetic Ising models are also one of the largest classes of graphical models for which there are efficient algorithms for sampling

and inference, which follows from the seminal work of Jerrum and Sinclair [100]. This makes them an appealing class of graphical models to be able to learn. In contrast, without ferromagneticity it is known that sampling and inference are computationally hard when the Gibbs measure on the corresponding infinite d-regular tree becomes non-unique [171].

2.1.2 Our Results

First we focus on learning ferromagnetic Restricted Boltzmann Machines with bounded degree. The idea behind our algorithm is simple: the observed variables that exert the most influence on some variable X_i ought to be X_i's two-hop neighbors. While this may seem intuitive, the most straightforward interpretation of this statement is false — the variable with the largest correlation with X_i may actually not be a two-hop neighbor. In addition, even if we correct the statement (e.g. by stating instead that there should be a neighbor with large influence), such facts about graphical models are often subtle and challenging to prove. Ultimately, we make use of the famous Griffiths-Hurst-Sherman correlation inequality [83] to prove that the discrete influence function

$$I_i(S) = \mathbb{E}\left[X_i | X_S = \{+1\}^{|S|}\right]$$

is submodular (see Theorem 5). The GHS inequality has found many applications in mathematical physics where it is an important ingredient in determining critical exponents at phase transitions. By recognizing that the concavity of magnetization is analogous to the properties of the multilinear extension of a submodular function [40], we are able to bring to bear tools from submodular maximization to learning graphical models with latent variables.

More precisely, we show that any set T that is sufficiently close to being a maximizer of I_i must contain the two-hop neighbors of X_i. We can thus use the greedy algorithm for maximizing a monotone submodular function [147] to reduce our problem of finding the two-hop neighbors of X_i to a set of constant size, where the constant depends on the maximum degree and upper and lower bounds on the strength

of non-zero interactions. It is information theoretically impossible to learn W, $b^{(1)}$ and $b^{(2)}$ uniquely, but we do something almost as good and learn a description of the distribution of the observed variables as a Markov Random Field (or MRF, see Definition 6):

Theorem 1 (Informal)**.** *There is a nearly quadratic time algorithm with logarithmic sample complexity for learning the distribution of observed variables (expressed as a Markov Random Field) for ferromagnetic Restricted Boltzmann Machines of bounded degree and upper and lower bounded interaction strength.*

The key part of this Theorem is the structure recovery guarantee, which learns the 2-hop neighborhoods of a node and is formalized in Theorem 9. Given this structure, learning the parameters of the model is straightforward from an algorithmic standpoint; we actually defer the discussion of this step to Chapter 6, because it is more related to the other material in that Chapter and in part because it uses tools which are also introduced in Chapter 5. We note that unlike earlier greedy algorithms for learning Ising models [29, 87] our dependence on the maximum degree is singly exponential and hence is nearly optimal [163]. In independent work, Lynn and Lee [127] also considered the problem of maximizing the influence but in a *known* Ising model. They gave a (conjecturally optimal) algorithm for solving this problem given an ℓ_1-constraint on the external field.

Our algorithm extends straightforwardly to general ferromagnetic Ising models with latent variables. In this more general setting, the two-hop neighborhood of a node i is replaced by an induced Markov blanket (i.e. neighborhood in the Markov Random Field), which informally corresponds to the set of observed nodes that separate i from the other observed nodes. We prove:

Theorem 2 (Informal)**.** *There is a nearly quadratic time algorithm with logarithmic sample complexity for learning the distribution of observed variables (expressed as a Markov Random Field) for ferromagnetic Ising model with latent variables, under the conditions that the interaction strengths are upper and lower bounded, the induced Markov blankets have bounded size and that the distance between any node i and any*

other node in its Markov blanket is bounded.

See Theorem 11 for the precise statement of the structure recovery guarantee, and again we defer to Chapter 6 for how to analyze learning the model given the structure. We remark that in our setting, the maximal Fourier coefficients of the induced MRF can be arbitrarily small, which is a serious obstacle to directly applying existing algorithms for learning MRFs (see Example 4). Our method also has the advantage of running in near-quadratic time whereas existing MRF algorithms would require runtime n^{d_H+1}, where d_H is the maximum hidden degree[1]. We also show how Lee-Yang properties that hold for ferromagnetic Ising models [124] carry over to the induced MRFs in the presence of latent variables, which allows us to approximate the partition function and perform inference efficiently. See Theorem 14 for the precise statement. Compared to the previous settings where provable guarantees were known, ours is the first to work even when there are long range correlations.

As we alluded to earlier, being ferromagnetic turns out to be the *key* property in avoiding computational intractability. More precisely, we show a rather surprising converse to the well-known fact that marginalizing out a latent variable produces a higher-order interaction among its neighbors. We show that marginalizing out a collection of latent variables can produce *any* desired higher-order interaction among their neighbors.

Theorem 3 (Informal)**.** *Every binary Markov Random Field of order d_H can be expressed as the distribution on observed variables of a Restricted Boltzmann Machine, where the maximum degree of any latent node is at most d_H.*

See Theorem 6 for the precise statement. Our approach to showing the equivalence between RBMs and MRFs is to show a non-zero correlation bound between the soft absolute value function that arises from marginalizing out latent variables and a parity function. We accomplish this through estimates of the Taylor expansion of special functions. With this in hand, we can match the largest degree terms in the energy function of an MRF and recurse.

[1]The induced MRF has order d_H, so these methods (e.g. [108]) need to solve regression problems on polynomials of degree d_H.

Apart its usefulness in proving hardness, this result also resolves a basic question about the representational power of RBMs. Towards the goal of understanding deep learning, a number of recent works have shown depth separations in feed-forward neural networks [180, 161, 60]. They explicitly construct (or show that there exists) a function that can be computed by a depth $d+1$ feed-forward neural network of small size, but with depth d would require exponential size. In fact, RBMs are the building block of another popular paradigm in deep learning: deep belief networks [89]. Towards understanding the representational power of RBMs, Martens et al. [133] showed that it is possible to approximately represent the uniform distribution on satisfying inputs to the parity function, and more generally any predicate depending only on the number of 1s, using a dense RBM. In practice, sparse RBMs are desirable because their dependencies are easier to interpret. The above theorem exactly characterizes what distributions can be represented this way: They are exactly the bounded order MRFs.

In any case, what this means for our lower bound is that without ferromagneticity, even RBMs with a constant number of latent variables of constant degree inherits the hardness results of learning MRFs [32, 108], that in turn follow from the popular assumption that learning sparse parities with noise is hard. For comparison, the technique used in [133] seems insufficient for this reduction — their method can only build certain noiseless functions.

Corollary 1 (Informal)**.** *If k-sparse noisy parity on n bits is hard to learn in time $n^{o(k)}$, then it is hard to learn a representation of the distribution on n observed variables (as any unnormalized function that can be efficiently computed) that is close to within total variation distance $1/3$ of a Restricted Boltzmann Machine where the maximum degree of any latent node is d_H in time $n^{o(d_H)}$. This is true even if the number of hidden nodes in the RBM is promised to be constant w.r.t. n.*

See Theorem 8 for the precise statement. Recall that it is impossible to learn the parameters of an RBM uniquely. Our result shows that learning merely a description of the distribution on the observed variables — i.e. a form of improper learning —

is hard too, even for RBMs with only a constant number of hidden variables. In contrast, previous lower bounds were for graphical models with many more latent variables than observed variables [25]. At the time it seemed plausible that there were large classes of graphical models with latent variables that could be efficiently learnable. But in light of how simple our hard examples are, it seems difficult to imagine any other natural and well-motivated class of graphical models with latent variables (without ferromagneticity) that is also easy to learn.

2.1.3 Further Discussion

There is an intriguing analogy between our results and the problem of learning juntas [143, 184]. While the general problem of learning k-juntas seems to be hard to solve in time $n^{o(k)}$ there are some special cases that can be solved much faster. Most notably, if the junta is monotone then there is a simple algorithm that works: Find all the coordinates with non-zero influence and solve the junta learning problem restricted to those coordinates. We can think of ferromagneticity as the natural analogue of monotonicity in the context of RBMs, since this property also prevents certain types of cancellations. Are there other contraints that one can impose on RBMs, perhaps inspired by ones that work for juntas, that make the problem much easier?

Another enticing question for future work is to study “deeper" versions of the problem, such as ferromagnetic deep belief networks. Are there new provable algorithms for classes of deep networks to be discovered? There is a growing literature on learning deep networks under various assumptions [9, 99, 200, 78], but the ability of ferromagnetic RBMs to express long-range correlations seems to make it a potentially more challenging problem to tackle.

2.2 Preliminaries

Recall from Chapter 1 that an Ising model is a probability distribution $\mu(J, h)$ on the hypercube $\{\pm 1\}^n$ under which

$$\Pr(X = x) = \mu(x) = \frac{1}{Z} \exp\Big(\frac{1}{2}\sum_{i,j} J_{ij} x_i x_j + \sum_i h_i x_i\Big).$$

Definition 1. *A* ferromagnetic Ising model with consistent external fields *is an Ising model such that $J_{ij} \geq 0$ for all i, j and such that $h_i \geq 0$. We will refer to this just as a* ferromagnetic Ising model *from now on. We will also refer to such a J as a* ferromagnetic interaction matrix.

We are particularly interested in Ising models with *hidden variables*; thus we introduce the well-known concept of a *Restricted Boltzmann Machine*. We will focus on the case of RBMs in the sequel, though everything can be generalized to Ising models with arbitrary sets of hidden nodes without much effort, as long as there are no large connected components of hidden nodes.

Definition 2. *Fix a vertex set V which is split into two disjoint parts as $V = V_1 \cup V_2$, and let $n_1 = |V_1|$ and $n_2 = |V_2|$. A* Restricted Boltzmann Machine *(or RBM) is a probability distribution on $\{\pm 1\}^{n_1} \times \{\pm 1\}^{n_2}$ under which*

$$\Pr(X = x, H = h) = \frac{1}{Z} \exp\left(x^T W h + \sum_{i=1}^{n} b_i^{(1)} x_i + \sum_{i=1}^{m} b_j^{(2)} h_j\right)$$

where $W : \mathbb{R}^{n_1 \times n_2}$ is the interaction matrix*, X is referred to as the* observed/visible nodes*, Y is referred to as the* latent/hidden nodes*, $b^{(1)}$ is the vector of external fields/biases of the observed nodes and $b^{(2)}$ is the vector of external fields for the hidden nodes.*

Clearly the joint distribution of a Restricted Boltzmann Machine is just a special case of a general Ising model. Therefore we say a Restricted Boltzmann Machine is *ferromagnetic* if $W_{ij} \geq 0, b_i^{(1)} \geq 0, b_i^{(2)} \geq 0$ which is consistent with our previous terminology.

2.3 Submodularity of Influence in Ising models

Definition 3. *Fix a ferromagnetic interaction matrix J. We define the* smooth influence function *for X_i to be*

$$\mathcal{I}_i(h) = \mathbb{E}_{X\sim\mu(J,h)}[X_i]$$

Definition 4. *Suppose $f : \mathbb{R}^n_{\geq 0} \to \mathbb{R}$ is a $\mathcal{C}^2$ function, i.e. it has continuous second partial derivatives. We say that f is a* smooth monotone submodular function *if*

1. *$\partial_i f \geq 0$ everywhere, and*

2. *$\partial_i \partial_j f \leq 0$ everywhere.*

We will see that smooth monotone submodularity of $\mathcal{I}_i$ in ferromagnetic Ising models follows from the following correlation inequality of Griffiths, Hurst and Sherman [83]:

Theorem 4 (GHS inequality, [83])**.** *Let J be the interaction matrix of a ferromagnetic Ising model on n nodes without external field. Then for any (not necessarily distinct) $1 \leq i, j, k, \ell \leq n$ we have*

$$\mathbb{E}[X_iX_jX_kX_\ell] - \mathbb{E}[X_iX_j]\mathbb{E}[X_kX_\ell] - \mathbb{E}[X_iX_k]\mathbb{E}[X_jX_\ell] - \mathbb{E}[X_iX_\ell]\mathbb{E}[X_jX_k] \\ +2\mathbb{E}[X_iX_\ell]\mathbb{E}[X_jX_\ell]\mathbb{E}[X_kX_\ell] \leq 0\,,$$

where the expectations are taken with respect to the Boltzmann distribution.

Corollary 2. *Let J be a ferromagnetic interaction matrix, i.e. $J_{ij} \geq 0$. Then for any $i \in [n]$, $\mathcal{I}_i(h) : \mathbb{R}^n_{\geq 0} \to \mathbb{R}$ is a smooth monotone submodular function.*

Proof. The equivalence of correlation inequalities and partial derivative inequalities is well-known (and is used in [83]); we include a proof only for completeness, since this precise statement does not appear in [83].

Let $Z(h)$ denote the partition function of the Ising model with interaction matrix J and external field h. Then observe that

$$\mathcal{I}_i(h) = \frac{\sum_x x_i \exp(x^T J x + h \cdot x)}{Z(h)} = \partial_i \log Z(h)\,,$$

so it suffices to prove that $\partial_j \partial_i \log Z(h) \geq 0$ for all i, j and $\partial_k \partial_j \partial_i \log Z(h) \leq 0$ for all i, j, k. First observe by computing partial derivatives that

$$\partial_j \partial_i \log Z(h) = \mathrm{Cov}(X_i, X_j) \geq 0\,,$$

where the covariance is taken with respect to $\mu(J, h)$ and the inequality follows from Griffiths inequality. One can similarly observe that

$$\partial_k \partial_j \partial_i \log Z(h) =$$
$$\mathbb{E}[X_i X_j X_k] - \mathbb{E}[X_i X_k]\mathbb{E}[X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j X_k] - \mathbb{E}[X_i X_j]\mathbb{E}[X_k] + 2\mathbb{E}[X_i]\mathbb{E}[X_j]\mathbb{E}[X_k]\,,$$

where the expectation is taken with respect to $\mu(J, h)$. We now eliminate the external field by the introduction of a *ghost vertex* X_{n+1} such that in the new Ising model, $J_{i(n+1)} = h_i$, J_{ij} is otherwise the same as before and there is no external field. In this new Ising model the marginal of $X_1, \ldots, X_n$ given $X_{n+1} = 1$ is the same as their distribution in the first Ising model, and the marginal given $X_{n+1} = -1$ is the same but with flipped signs. Letting $\mathbb{E}_\nu$ denote expectation with respect to this new Ising model, we see that

$$\begin{aligned}
&\mathbb{E}[X_i X_j X_k] - \mathbb{E}[X_i X_k]\mathbb{E}[X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j X_k] - \mathbb{E}[X_i X_j]\mathbb{E}[X_k] + 2\mathbb{E}[X_i]\mathbb{E}[X_j]\mathbb{E}[X_k] \\
&= \mathbb{E}_\nu[X_i X_j X_k X_\ell] - \mathbb{E}_\nu[X_i X_j]\mathbb{E}_\nu[X_k X_\ell] - \mathbb{E}_\nu[X_i X_k]\mathbb{E}_\nu[X_j X_\ell] - \mathbb{E}_\nu[X_i X_\ell]\mathbb{E}_\nu[X_j X_k] \\
&\qquad + 2\mathbb{E}_\nu[X_i X_\ell]\mathbb{E}_\nu[X_j X_\ell]\mathbb{E}_\nu[X_k X_\ell]\,,
\end{aligned}$$

where $\ell = n + 1$. Thus it suffices to verify that this last expression is at most zero, which follows from Theorem 4. □

Definition 5. *Fix a ferromagnetic Ising model* (J, h). *We define the* discrete influence

function *for X_i to be a function from $S \subset [n]$ to $\mathbb{R}$ given by*

$$I_i(S) = \mathbb{E}_{X\sim\mu(J,h)}\left[X_i | X_S = \{+1\}^{|S|}\right] = \mathbb{E}_{X\sim\mu(J,h+\infty\mathbf{1}_S)}[X_i]\,.$$

Theorem 5. *Fix a ferromagnetic Ising model (J,h). Then for every i, the discrete influence function $I_i(S)$ is a monotone submodular function.*

Proof. Since $I_i(S) = \mathbb{E}_{\mu(J,h+\infty\mathbf{1}_S)}[X_i]$, monotonicity follows immediately from Corollary 2. Similarly, submodularity follows because if $S \subset T$ and we let $h_S = h + \infty\cdot\mathbf{1}_S$ and likewise for h_T, then we obtain

$$I_i(S\cup\{j\}) - I_i(S) = \int_{h'_j=0}^{\infty} \partial_j \mathcal{I}_i(h_S + h'_j e_j) \geq \int_{h'_j=0}^{\infty} \partial_j \mathcal{I}_i(h_T + h'_j e_j) = I_i(T\cup\{j\}) - I_i(T)\,,$$

where the inequality follows point-wise, by integrating the inequality $\partial_k\partial_j\mathcal{I}_i \leq 0$ along any coordinate-wise non-decreasing path from $h_S + h'_j e_j$ to $h_T + h'_j e_j$. □

This submodularity has the following standard consequence, which will be very useful later.

Lemma 1. *Fix a ferromagnetic Ising model (J,h). Suppose $i \in [n]$ and $S, T \subset [n]$, and $I_i(T) > I_i(S)$. Then there exists $j \in T$ such that*

$$I_i(S\cup\{j\}) - I_i(S) \geq \frac{I_i(T) - I_i(S)}{|T\setminus S|}$$

Proof. This follows because

$$I_i(S\cup T) - I_i(S) \geq I_i(T) - I_i(S)$$

and by submodularity, since we can go from S to $S\cup T$ by adjoining elements of $T\setminus S$ one-by-one,

$$I_i(S\cup T) - I_i(S) \leq \sum_{j\in T\setminus S} I_i(S\cup\{j\}) - I_i(S) \leq |T\setminus S| \max_{j\in T\setminus S}(I_i(S\cup\{j\}) - I_i(S))$$

which completes the proof. □

2.4 Interreducibility Between RBMs and MRFs

First we recall the definition of (binary) Markov Random Fields from Chapter 1 and introduce some terminology for the *order* of the MRF:

Definition 6. *A* Markov Random Field *(or MRF) of order r is a probability distribution on $\{\pm 1\}^n$ such that*

$$\Pr(X = x) = \frac{1}{Z}\exp(f(x))$$

where f is a multivariate polynomial of degree r such that $f(0) = 0$, referred to as the potential. *The* structure graph *of a Markov random field has vertices $1, \ldots, n$ and connects vertex i and j if there is a monomial in $f(x)$ with non-zero coefficient involving both x_i and x_j.*

We will mostly be interested in Markov random fields where the structure graph has bounded degree, i.e. the size of the Markov blanket of each node is bounded. Now we observe that the marginal distribution on the observable variables of a Restricted Boltzmann machine is a Markov Random Field, of order at most the max degree of a hidden node. This is well known and was used for instance in [133], but we state and prove it for completeness:

Lemma 2. *Fix a Restricted Boltzmann Machine $(J, b^{(1)}, b^{(2)})$. Let w_j be the j^{th} column of W, i.e. the edge weights into hidden unit j. Then*

$$P(X = x) = \frac{1}{Z}\exp\left(\sum_{j=1}^{n_2}\rho(w_j \cdot x + b_j^{(2)}) + \sum_{i=1}^{n_1} b_i^{(1)} x_i\right)$$

where Z is the same as the partition function of the original RBM and $\rho(x) = \log(e^x + e^{-x})$ (this can be thought of as a "soft absolute value" function).

Proof. We show a slightly more general fact. Consider a general Markov Random Field of the form $\Pr(X = x) = \frac{1}{Z}\exp(f(x))$ where u is a vertex with only pairwise interactions, i.e.

$$f(x) = h_u x_u + \sum_{v \sim u} w_{uv} x_u x_v + g(x_{\sim u}).$$

We now compute the marginal distribution on the model when u is hidden. Observe that

$$\Pr(X_{\sim u} = x_{\sim u}) = \exp(g(x_{\sim u}))\frac{\sum_{x_u} \exp(h_u x_u + \sum_{v\sim u} w_{uv} x_u x_v)}{Z}$$

so if we let U denote the neighborhood of u and let

$$f_U(x_U) = \log \sum_{x_u} \exp(h_u x_u + \sum_{v\sim u} w_{uv} x_u x_v) = \rho(h_u + \sum_v w_{uv} \cdot x_v)$$

where $\rho(x) = \log(e^x + e^{-x})$ then

$$\Pr(X_{\sim u} = x_{\sim u}) = \frac{\exp(g(x_{\sim u}) + f_U(x_U))}{Z}$$

Applying this inductively gives the result of the lemma. □

Our main result in this section is a reduction in the other direction: We show that every MRF can be converted to an equivalent Restricted Boltzmann Machine. This is more difficult and to our knowledge was not known before. The key technical fact underlying the result is the following lemma, which shows that we can build an RBM with hidden nodes connected to the observed nodes in the set S with any desired correlation with a parity on S as long as the desired correlation is small. Then by building many of these hidden units we can capture the MRF potential exactly.

Lemma 3. *Fix $k \geq 0$ and let $\rho(x) = \log(e^x + e^{-x})$. Then there exist constants $\delta = \delta(k) > 0$ and $\gamma = \gamma(k) \in (0, \pi/2)$ such that for any δ' with $|\delta'| < \delta$ and $S \subset [n]$ with $|S| = k$, there exist w, h with $|w|_1 + h \leq \gamma$ such that*

$$\mathbb{E}_{X\sim\{\pm 1\}^n}[\rho(w \cdot X_S + h)\chi_S(X)] = \delta'$$

where the expectation is with respect to uniform measure.

Proof. This will follow by using the explicit formula for the taylor expansion of $\rho(x)$, which we will now derive. Recall $\rho'(x) = \tanh(x)$ and that tanh has an explicit power

series expansion with radius $\pi/2$ around 0:

$$\tanh(x) = \sum_{n=1}^{\infty} \frac{2^{2n}(2^{2n}-1)B_{2n}}{(2n)!} x^{2n-1}$$

with radius of convergence $\pi/2$, where $B_{2n} = \frac{(-1)^{n+1}2(2n!)}{(2\pi)^{2n}}\zeta(2n)$ are the even Bernoulli numbers. By integrating, we see

$$\rho(x) = \log 2 + \sum_{n=1}^{\infty} \frac{2^{2n}(2^{2n}-1)B_{2n}}{(2n)!(2n)} x^{2n}$$

with the same radius of convergence.

We will need the standard fact that $B_{2n} \neq 0$ for any $n \geq 1$, which follows immediately from the equation $B_{2n} = \frac{(-1)^{n+1}2(2n!)}{(2\pi)^{2n}}\zeta(2n)$ and the fact that $\zeta(s) = \sum_{m=1}^{\infty} \frac{1}{m^s} > 0$ for $s > 1$.

Now we use that the Fourier expansion of $\rho(w \cdot X_S + h)$ can be found by taking the power series expansion of ρ, plugging in $x = w \cdot X_S + h$ and using the identity $X_i^2 = 1$ to reduce to the parity basis. Let $m = \lceil \frac{|S|}{2} \rceil$ and take $\gamma \in (0, \pi/2)$. By restricting to w, h such that $|w|_1 + |h| < \gamma$ we can write

$$\rho(w \cdot X_S + h) = \log 2 + \sum_{n=1}^{m} \frac{2^{2n}(2^{2n}-1)B_{2n}}{(2n)!(2n)} (w \cdot X_S + h)^{2n} + O(\gamma^{2m+2}).$$

Note that in the sum, only the top $n = m$ term contributes to the coefficient of χ_S. Observe that when $|S|$ is even[2],

$$[\chi_S](w \cdot X_S + h)^{2m} = |S|! \prod_{s \in S} w_S$$

and when $|S|$ is odd

$$[\chi_S](w \cdot X_S + h)^{2m} = |S|! h \prod_{s \in S} w_S.$$

In the case where $|S|$ is even, first consider the case where $w_s = \gamma/|S|$ for $s \in S$. We

[2]We use the notation $[\chi_S]f$ to denote the Fourier coefficient of χ_S in the Fourier expansion of f.

then see that

$$[\chi_S]\rho(w \cdot X_S + h) = \frac{2^{2m}(2^{2m}-1)B_{2m}|S|!}{(2m)!(2m)|S|^{2m}}\gamma^{2m} + O(\gamma^{2m+2})$$

and so as long as γ is sufficiently small, the coefficient is positive. Next observe that if we flip the sign of w_{s^*} for a single $s^* \in S$, then the sign of $[\chi_S](w \cdot X_S + h)^{2m}$ flips and so the sign of $\rho(w \cdot X_S + h)$ must also flip when γ is sufficiently small. Since this coefficient varies continuously as a function of w_{s^*}, we see by the intermediate value theorem we see that we can get the coefficient of χ_S to be any value in $[-\delta, \delta]$ for some $\delta > 0$.

The case when $|S|$ is odd is the same, except that we take $w_s = \gamma/(|S|+1)$ and vary h in $[-\gamma/(|S|+1), \gamma/(|S|+1)]$. □

Theorem 6. *Consider an arbitrary order r Markov random field on the hypercube $\{\pm 1\}^n$, i.e. a probability distribution of the form $\Pr(X = x) = (1/Z)\exp(f(x))$ where f is a polynomial of degree r. Suppose that the structure graph of the MRF has degree d and the coefficients of f are bounded by a constant M. There is an RBM with n observable nodes and parameters $(J, b^{(1)}, b^{(2)})$ with the following properties:*

(1) The induced MRF of the RBM equals the original MRF, i.e. the marginal law of the observed variables is the same as the distribution of the original MRF.

(2) There are at most $O_{d,M}(n)$ hidden units[3].

(3) The degree of every vertex in the hidden layer is at most r.

(4) The two-hop neighborhood of every observed node equals its original MRF-neighborhood. In particular the two-hop degree d_2 equals the degree d of the structure graph of the MRF.

Proof. By Lemma 2 this reduces to rewriting the MRF potential in term of a summation of $\rho(\cdot)$ terms coming from hidden units. We use the building block of Lemma 3

[3]This is a general upper bound; from the construction we see that if few Fourier coefficients are nonzero, then few hidden units are used.

and build the potential of the MRF in a top-down fashion. More precisely we can build any boolean function with Fourier mass supported on the first r Fourier levels as follows:

(a) For each of the degree r coefficients, use several copies of the parity building block to build a boolean function with the correct degree r Fourier coefficients.

(b) Now recurse to the lower level coefficients — if we use only the building block for $|S| \leq r-1$ we will not affect the degree r coefficients.

The end result is that any Markov random field of order r can be converted into a Restricted Boltzmann distribution with hidden nodes of degree at most r, such that the observed nodes have the same distribution as the same Markov random field. If all of the Fourier coefficients of the potential of the original MRF are bounded by M, then the number of hidden units we need to introduce is $O_{d,M}(n)$, because given the upper bound on d each visible unit is involved in only a constant number of hyperedges, and given the upper bound on d and M it takes only a constant number of copies of the building block to build each Fourier coefficient. □

2.5 The Learning Problem for RBMs

We consider the problem of learning a Restricted Boltzmann Machine given samples from its marginal distribution on the observed nodes X. Note that if we were also given samples from the joint distribution on (X, Y), then this would be the standard learning problem for Ising models as considered in e.g. [29, 108]. However, in our setting it is impossible to recover the underlying interaction matrix W because it is not uniquely determined, i.e. Restricted Boltzmann Machines are *unidentifiable* as the following examples illustrate:

Example 1. *Consider the Restricted Boltzmann machine with two observable nodes* $\{1, 2\}$ *and two hidden nodes labeled* $\{3, 4\}$ *such that* $W_{13} = 1, W_{23} = 1$ *and* $W_{14} = -1, W_{24} = 1$. *Then the marginal distribution on the observables is just independent*

Rademachers, so this Restricted Boltzmann machine is not distinguishable from a model with no connections at all.

The previous example used non-ferromagnetic interactions to demonstrate the nonidentifability of RBMs. However, even when the RBM is ferromagnetic the model remains highly nonidentifiable:

Example 2. *Consider a model with two observable nodes* $\{1, 2\}$*, no external fields, and any number of hidden units/connections. Since the marginal distribution on* X_1 *and* X_2 *each must be Rademacher by symmetry, the observable distribution is specified just by a single parameter, the covariance between* X_1 *and* X_2*. However even in the simplest case, where there is only a single hidden unit connected to both* X_1 *and* X_2*, there are two parameters in the model, the two edge weights and we clearly see that these edge weights are not uniquely determined by the distribution.*

Example 3 (Hidden Structure is Undetermined)**.** *Consider an RBM with three observable nodes* $\{1, 2, 3\}$*, a single hidden node connected to all of them with positive edge weights, and no external field. We know the observable distribution is an MRF so it is of the form*

$$\Pr(X = x) = \frac{1}{Z} \exp(W_{12}x_1x_2 + W_{13}x_1x_3 + W_{23}x_2x_3 + W_{123}x_1x_2x_3).$$

Perhaps surprisingly, in this model $W_{123} = 0$*. This can be seen from Lemma 2 and Taylor-expanding* ρ*, or simply by symmetry: the observable distribution is symmetric under the sign flip* $x \mapsto -x$ *and so necessarily* $W_{123} = 0$*. However, since there are only pairwise interactions in the potential it is easy to see (or we can apply Theorem 6) that there exists another RBM with only degree-*2 *hidden nodes that has exactly the same observable distribution.*

These examples illustrate (even in restricted setting) that we cannot hope to reconstruct J. Instead we consider the natural objectives from the perspective of viewing the observable distribution as a Markov Random Field: *structure learning* and learning the parameters of the Markov random field. We start with structure

learning, which can be viewed as the problem of learning the *two-hop neighborhoods* of the observed random variables — i.e. learning the square of the adjacency matrix of the bipartite structure graph.

Definition 7. *Suppose i is an observed node in an RBM $(J, b^{(1)}, b^{(2)})$. The* two-hop neighborhood *of i, denoted $\mathcal{N}_2(i)$, is the smallest set $S \subset [n] \setminus \{i\}$ such that conditioned on X_S, X_i is conditionally independent of X_j for all $j \in [n] \setminus (S \cup \{i\})$.*

Note that S is uniquely determined, because it is just the neighborhood of i when we view the observable distribution as a Markov Random Field.

Definition 8. *The* two-hop degree *d_2 of an RBM is the maximum size of $\mathcal{N}_2(i)$ over all observed nodes i.*

Observe that $\mathcal{N}_2(i)$ is always a subset of the graph-theoretic two-hop neighborhood of i, i.e. the smallest set S such that vertex i is separated from the other observable nodes in the structure graph of the RBM. However it may be a strict subset, as in Example 1. We will later show in Lemma 6 that the graph-theoretic two-hop neighborhood always agrees with $\mathcal{N}_2(i)$ in *ferromagnetic* RBMs.

In order to learn the two-hop structure of an RBM it will be necessary to have lower and upper bounds on the edge weights of the model, so we introduce the following notion of degeneracy. This is a standard assumption in the literature on learning Ising models [29, 189, 108]. In particular, a lower bound is needed because otherwise it would be impossible to distinguish a non-edge from an edge with an arbitrarily weak interaction. An upper bound is needed to ensure the distribution of any variable is not arbitrarily close to being deterministic.

Definition 9. *We say that an Ising model is is* (α, β)-nondegenerate[4] *if both:*

(1) For every i, j such that $|J_{ij}| \neq 0$, we have $|J_{ij}| > \alpha$.

(2) $\sum_j |J_{ij}| + |h_i| \leq \beta$ for every node i.

We say that an RBM is (α, β)-nondegenerate *if it is (α, β)-nondegenerate as an Ising model.*

[4]Observe that the notational convention follows [108] instead of [29], where β denotes the maximum edge weight.

2.5.1 Maximal Coefficients Can be Arbitrarily Small

In this subsection, we discuss some important obstacles to directly using regression-based methods (in particular [108]) for learning the parameters of a ferromagnetic RBM. By Lemma 2, we can cast the problem of learning $\mathcal{N}_2(i)$ for each node i as a structure learning problem on the induced MRF. In order to use the results of Klivans and Meka [108], we need to get bounds on the potential

$$p(x) = \sum_{j=1}^{n_2} \rho(w_j \cdot x + b_j^{(2)}) + \sum_{i=1}^{n_1} b_i^{(1)} x_i.$$

In particular we need a bound on the size of the coefficients of $\partial_i p$. For a function $p : \{\pm 1\}^{n_1} \to \mathbb{R}$, let $\|p\|_1$ denote the sum of the absolute values of its Fourier coefficients. Observe that

$$\left|\mathbb{E}_{X\sim\{\pm1\}^n}\left[\partial_i p \prod_{i\in S} X_i\right]\right| \leq \left|b_i^{(1)}\right| + \left|\mathbb{E}_{X\sim\{\pm1\}^n}\left[\sum_{j:w_{ij}\neq 0} \rho(w_j \cdot X + b_j^{(2)}) \prod_{i\in S} X_i\right]\right| \leq \left|b_i^{(1)}\right| + 2\beta\deg(i)$$
$$\leq 2\beta(\deg(i)+1)$$

which follows from Holder's inequality, since $|\rho(w_j \cdot X + b_j^{(2)})| \leq 2\beta$ and $|b_i^{(1)}| \leq \beta$. Furthermore the coefficient of X_S in $\partial_i p$ can be non-zero only when S is a subset of the two-hop neighborhood of i, which follows from the Markov property. Thus we conclude that

$$\|\partial_i p\|_1 \leq 2^{d_2+1}\beta(\deg(i)+1)$$

where d_2 is the maximum size of a node's two-hop neighborhood.

With this calculation in hand, the algorithm of Klivans and Meka [108] is able to estimate the *maximal* of the potential $p(x)$ to within ϵ additive error using roughly

$$\frac{e^{O(d_H 2^{d_2+1}\beta(d_V+1))}}{\epsilon^4}\log n$$

samples where d_H is the maximum degree of any hidden node and d_V is the maximum degree of any observed node. We could then apply Theorem 7.2 of [108] to learn the

two-hop neighborhoods in the RBM if we had an additional assumption that the induced MRF was η-identifiable:

Definition 1. *A Markov Random Field is η-identifiable if every maximum Fourier coefficent of its potential p has magnitude at least η.*

Unfortunately, even for MRFs induced by ferromagnetic RBMs and even under the assumption of (α, β)-nondegeneracy, η can be made to be arbitrarily small, as the following example shows:

Example 4 (Failure of η-identifiability in ferromagnetic RBMs)**.** *Consider an RBM on three observed nodes with spins X_1, X_2, X_3 and a single hidden node with spin H_1 connected to all of the observed nodes with edge weight $1/4$. On the hidden node let there be an external field $b_1^{(2)} = \epsilon$. When $\epsilon = 0$, we see (as in Example 3) that*

$$\Pr(X = x) = \frac{1}{Z} \exp(JX_1X_2 + JX_1X_3 + JX_2X_3)$$

for some constant J that is bounded away from zero. Hence the model is η-identifiable. However, for a small $\epsilon > 0$, one can see by Taylor series expansion that the coefficient of $X_1X_2X_3$ is nonzero, and by continuity it can be made arbitrarily small by decreasing ϵ. This does not affect the (α, β)-nondegeneracy of the model, but clearly the parameter η in η-identifiability goes to zero.

Thus existing guarantees for regression-based methods do not seem to be strong enough for our purposes. Moreover they would even require time n^{d_H+1} to run, where d_H is the hidden degree, since they solve a high-dimensional regression problem in the basis of all size d_H monomials. In contrast our approach for learning the two-hop neighborhoods not only works in spite of the fact that the maximal Fourier coefficients can be arbitrarily small, it also runs in nearly quadratic time (see Theorem 9).

2.5.2 Hardness for Improperly Learning RBMs

In this subsection we show that structure learning for general (i.e. possibly non-ferromagnetic) RBMs takes time $n^{\Omega(d_H)}$ under the conjectured hardness for learning

sparse parity with noise.

Definition 10. *The* k-sparse parity with noise *distribution is the following distribution on* (X, Y) *parameterized by a constant* $\eta \in (0, 1/2)$ *and an unknown subset* S *of size at most* k*:*

1. *Sample* $X \sim \mathrm{Unif}(\{-1, +1\}^n)$.
2. *With probability* $1/2 + \eta$*, set* $Y = \prod_{s \in S} X_s$*, and with probability* $1/2 - \eta$*, set* $Y = (-1) \prod_{s \in S} X_s$.

The learning problem for k-sparse parity with noise *is to learn* S *in polynomial time with high probability, given access to an oracle which generates samples of* (X, Y).

The important point is that the joint distribution of an $(r-1)$-sparse parity with noise (X, Y) is a Markov Random Field with order r interactions, and by Theorem 6 it is also the marginal distribution on the observables of an MRF with maximum hidden degree d_H, where the two-hop neighborhood of Y is exactly the set S. This means if we could learn the two-hop neighborhoods of an RBM in time $n^{o(d_H)}$ this would yield a $n^{o(k)}$ algorithm for learning k-sparse parities with noise, which is a long-standing open question in theoretical computer science and conjectured to be impossible. The best known algorithm of Valiant [184] runs in time $n^{0.8k}$. We summarize this observation in the following observation:

Observation 1. *If* k*-sparse parity with noise on* n *bits cannot be learned in time* $n^{o(k)}$*, then there is no algorithm which runs in time* $n^{o(d_H)}$ *and learns the two-hop neighborhood structure of a general RBM from samples of the distribution on its observed nodes.*

We will now furthermore show that this result applies even in the case of *improper learning*, where we do not aim to learn the structure but instead aim to learn a different distribution close to the RBM. For this purpose it is useful to recall the following equivalent[5] formulation of learning sparse parities as a hypothesis testing problem:

[5]It is clear that if we have an algorithm for the learning problem, we can use it for the hypothesis

Definition 11. *The* hypothesis testing problem for k-sparse parity with noise *is to distinguish with high probability*[6] *between the cases where* (X, Y) *is drawn from the uniform distribution on* $\{\pm 1\}^{n+1}$ *and where* (X, Y) *is drawn from the* k*-sparse parity with noise distribution for an unknown* S*.*

We now use this to show hardness for improper learning. First we show hardness in the case of algorithms returning a distribution $\mathcal{Q}$ with an (approximately) computable probability mass function.

Theorem 7. *If* k*-sparse parity with noise on* n *bits cannot be learned in time* $n^{o(k)}$*, then there is no algorithm that runs in time* $n^{o(d_H)} \cdot poly(1/\epsilon)$ *and returns a probability distribution* $\mathcal{Q}$ *such that:*

(1) It is possible to (approximately) compute the pmf $\mathcal{Q}(x, y)$ *for* $x, y \in \{\pm 1\}^n \times \{\pm 1\}$ *in polynomial time.*

(2) $\|\mathcal{Q} - \mathcal{P}\|_{\mathrm{TV}} < \epsilon$ *where* $\mathcal{P}$ *is the distribution on the observables of an RBM with hidden degree* d_H*.*

Proof. We show how to use $\mathcal{Q}$ to solve the hypothesis testing problem for sparse parity with noise. Recall that for any distributions $\mathcal{P}_1, \mathcal{P}_2$

$$\|\mathcal{P}_1 - \mathcal{P}_2\|_{\mathrm{TV}} = \mathbb{E}_{X \sim \mathcal{P}_1}\left[\frac{\mathcal{P}_1(X) - \mathcal{P}_2(X)}{\mathcal{P}_1(X)} \mathbb{1}[\mathcal{P}_1(X) \geq \mathcal{P}_2(X)]\right]$$

and observe that the quantity inside the expectation is always valued in $[0, 1]$. Therefore, with $\mathcal{P}_1 = \mathrm{Unif}(\{\pm 1\}^{n+1})$ and $\mathcal{P}_2 = Q$, we may use m samples from $\mathcal{P}_1$ and the above formula to approximate the TV between $\mathcal{Q}$ and the uniform distribution on $\{\pm 1\}^{n+1}$ within error $O(1/\sqrt{m})$ with high probability (by Hoeffding's inequality). Since the TV distance between the uniform distribution and any particular sparse parity with noise is $\Omega(\eta)$ (consider the tester that looks at whether $Y = \prod_{s \in S} X_s$),

testing problem (the algorithm will return some set S and we just have to test if the parity of X_S is correlated with Y). In the other direction, observe that if we pick a particular i and look at the marginal distribution on $(X_{\neq i}, Y)$ then if $i \in S$ this marginal distribution becomes uniform on $\{\pm 1\}^n$, whereas if $i \notin S$ this is just a sparse parity with noise on a smaller number of variables, so if we can hypothesis test we can efficiently determine for every i whether i lies in S.

[6] i.e. with probability of Type I and Type II error going to 0 sufficiently fast.

this lets us solve the hypothesis testing problem for sparse parity with noise. Thus, if the algorithm can find $\mathcal{Q}$ in time $n^{o(d_H)}$, then this violates the conjectured hardness of learning sparse parity with noise. $\square$

Remark 1. *We see from the proof of Theorem 6 that only a constant number of hidden nodes (in terms of n) are used in the construction of the sparse parity RBM, so the above result holds even if the RBM is promised to have $O_{d_H}(1)$ many hidden nodes.*

In fact, the hardness result extends even to the case when we have access only to an *unnormalized* probability distribution function.

Theorem 8. *If k-sparse parity with noise on n bits cannot be learned in time $n^{o(k)}$, then there is no algorithm which runs in time $n^{o(d_H)} \cdot poly(1/\epsilon)$ and returns a probability distribution $\mathcal{Q}$ such that:*

(1) $\|\mathcal{Q} - \mathcal{P}\|_{\mathrm{TV}} < \epsilon$ where $\mathcal{P}$ is the distribution on the observables of an RBM with hidden degree d_H.

(2) There exists a function $q(x, y)$ such that $\mathcal{Q}(x, y) = \frac{1}{C_q} q(x, y)$ and $q(x, y)$ is efficiently computable.

Proof. We again reduce from the hypothesis testing problem for sparse parity with noise. As before suppose $Z^{(1)}, \ldots, Z^{(m)}$ are iid samples from the uniform distribution on $\{\pm 1\}^{n+1}$; we will look at the statistics of $q(Z)$. Observe that if $\mathcal{Q}$ were the uniform distribution, then we would have $q(Z) = C_q 1/2^{n+1}$, whereas if $\mathcal{Q}$ were a sparse parity with noise we would have $q(Z) \propto e^{J_\eta \prod_{s \in S} Z_s}$ where J_η is a constant that corresponds to η.

Let $q_{1/3}$ be such that the number of $z^{(i)}$ with $q(Z^{(i)}) \leq q_{1/3}$ is at most $m/3$, and define $q_{2/3}$ similarly. Consider the quantity $V := \frac{q_{2/3} - q_{1/3}}{q_{1/3} + q_{2/3}}$. Under the uniform distribution V is concentrated around zero, whereas under a sparse parity distribution V is concentrated about $\frac{e^{J_\eta} - e^{-J_\eta}}{e^{J_\eta} + e^{-J_\eta}}$. The same is true under distributions which are close in TV to either distribution, since V is defined in terms of cumulative distribution function statistics. Therefore we can distinguish between independent bits and sparse parity with noise efficiently given access to q. $\square$

2.6 A Greedy Algorithm for Learning Ferromagnetic RBMs

We describe a simple and efficient greedy algorithm for learning the two-hop neighborhood of an observed node i from samples, if the RBM is ferromagnetic. This algorithm is much faster than is possible for general RBMs according to the lower bound of the previous subsection. Let $\widetilde{\mathbb{E}}$ denote the empirical expectation, and define the *empirical influence*

$$\widetilde{I}_i(S) = \widetilde{\mathbb{E}}[X_i | X_S = \{1\}^S] \,.$$

Let $\eta > 0$ be a real-valued parameter and $k \geq 1$ an integer parameter to be specified later.

Algorithm 1: GreedyNbhd(i)

---

1. Set $S_0 := \emptyset$.
2. For t from 0 to $k-1$:
   (a) Let $j_{t+1} := \arg\max_j \widetilde{I}_i(S_t \cup \{j\})$, where j ranges over all observed nodes.
   (b) Set $S_{t+1} := S_t \cup \{j_{t+1}\}$
3. Let $\widetilde{\mathcal{N}}_2 := \{j \in S_k : \widetilde{I}_i(S_k) - \widetilde{I}(S_k \setminus \{j\}) \geq \eta\}$.
4. Return $\widetilde{\mathcal{N}}_2$.

(α, β)-nondegeneracy has the following useful consequences:

Lemma 4. *Suppose X_i is the spin at vertex i in an (α, β)-nondegenerate Ising model. Then* $\min(\Pr(X_i = +), \Pr(X_i = -)) \geq \sigma(-2\beta)$, *where* $\sigma(x) = \frac{1}{1+e^{-x}}$.

Proof. We show the lower bound for $\Pr(X_i = +)$ since the two cases are symmetrical. By the law of total expectation, it suffices to show that for any fixing $x_{\neq i}$ of the other spins $X_{\neq i}$ that $\Pr(X_i = +|X_{\neq i} = x_{\neq i}) \geq \sigma(-2\beta)$, and this follows because

$$\Pr(X_i = +|X_{\neq i} = x_{\neq i}) = \frac{\exp(\sum_{j:j\neq i} J_{ij}x_j)}{\exp(\sum_{j:j\neq i} J_{ij}x_j) + \exp(-\sum_{j:j\neq i} J_{ij}x_j)} = \sigma\Big(2\sum_{j:j\neq i} J_{ij}x_j\Big) \geq \sigma(-2\beta).$$

□

Lemma 5. *Suppose X_i is the spin at vertex i in an (α, β)-nondegenerate Ising model and j is a neighbor of i. Then for any fixing $x_{\neq i,j}$ of the other spins $X_{i\neq j}$ of the Ising model, we have*

$$\big|\mathbb{E}[X_i|X_j = 1, X_{\neq i,j} = x_{\neq i,j}] - \mathbb{E}[X_i|X_j = -1, X_{\neq i,j} = x_{\neq i,j}]\big| \geq 2\alpha(1 - \tanh^2(\beta)).$$

Proof. Observe that

$$\mathbb{E}[X_i|X_{\neq i}] = \tanh\Big(\sum_{k:k\neq i} J_{ik}x_k\Big).$$

Since $\tanh'(x) = 1 - \tanh^2(x)$ and tanh is a monotone function, we see that if we let $x = -J_{ij} + \sum_{k:k\notin\{i,j\}} J_{ik}x_k$, then since $x \in [-\beta, \beta]$ we have

$$|\tanh(x + 2J_{ij}) - \tanh(x)| \geq 2|J_{ij}| \inf_{x\in[-\beta,\beta]}(1 - \tanh^2(x)) \geq 2\alpha(1 - \tanh^2(\beta)). \quad \square$$

The following lemma shows quantitatively that in a nondegenerate ferromagnetic RBM, the graph-theoretic two-hop neighborhood of a vertex i always equals $\mathcal{N}_2(i)$, the two-hop Markov blanket. It is immediate from the Markov property for the RBM as an Ising model that $\mathcal{N}_2(i)$ is contained in the graph-theoretic two-hop neighborhood, and the lemma implies the reverse inclusion.

Lemma 6. *Suppose node i is an observed node in a ferromagnetic (α, β)-nondegenerate RBM and denote by T the graph-theoretic two-hop neighborhood of i. If $S \subset [n]$ is a*

set of nodes such that $T \not\subset S$, then for any $j \in T \setminus S$, we have

$$I_i(S \cup \{j\}) - I_i(S) \geq 2\alpha^2\sigma(-2\beta)(1 - \tanh(\beta))^2 .$$

Proof. Fix $j \in \mathcal{N}_\in(i) \setminus S$ and let k be a hidden node which is a mutual neighbor of i, j. Now observe by submodularity it suffices to prove the lower bound when $S = [n] \setminus \{i, j, k\}$. Then

$$\begin{aligned}
I_i(S \cup \{j\}) - I_i(S) &= \mathbb{E}[X_i | X_S = 1^S, X_j = 1] - \mathbb{E}[X_i | X_S = 1^S] \\
&= \mathbb{E}[X_i | X_S = 1^S, X_j = 1] - \mathbb{E}[X_i | X_S = 1^S, X_j = 1] \Pr(X_j = 1 | X_S = 1^S) \\
&\quad - \mathbb{E}[X_i | X_S = 1^S, X_j = -1] \Pr(X_j = -1 | X_S = 1^s) \\
&= \Pr(X_j = -1 | X_S = 1^S)(\mathbb{E}[X_i | X_S = 1^S, X_j = 1] - \mathbb{E}[X_i | X_S = 1^S, X_j = -1]) \\
&\geq \sigma(-2\beta)(\mathbb{E}[X_i | X_S = 1^S, X_j = 1] - \mathbb{E}[X_i | X_S = 1^S, X_j = -1]) .
\end{aligned}$$

Furthermore when $S = [n] \setminus \{i, j, k\}$ we know that X_i and X_j are independent conditioned on k, so

$$\begin{aligned}
&\mathbb{E}[X_i | X_S = 1^S, X_j = 1] - \mathbb{E}[X_i | X_S = 1^S, X_j = -1] \\
&= \mathbb{E}[X_i | X_S = 1^S, X_k = 1](\Pr(X_k = 1 | X_S = 1^S, X_j = 1] - \Pr(X_k = 1 | X_S = 1^S, X_j = -1]) \\
&\quad + \mathbb{E}[X_i | X_S = 1^S, X_k = -1](\Pr(X_k = -1 | X_S = 1^S, X_j = 1] - \Pr(X_k = -1 | X_S = 1^S, X_j = -1]) \\
&= \mathbb{E}[X_i | X_S = 1^S, X_k = 1](\Pr(X_k = 1 | X_S = 1^S, X_j = 1] - \Pr(X_k = 1 | X_S = 1^S, X_j = -1]) \\
&\quad - \mathbb{E}[X_i | X_S = 1^S, X_k = -1](\Pr(X_k = 1 | X_S = 1^S, X_j = 1] - \Pr(X_k = 1 | X_S = 1^S, X_j = -1]) \\
&= (\mathbb{E}[X_i | X_S = 1^S, X_k = 1] - \mathbb{E}[X_i | X_S = 1^S, X_k = -1]) \\
&\quad \cdot (\Pr(X_k = 1 | X_S = 1^S, X_j = 1] - \Pr(X_k = 1 | X_S = 1^S, X_j = -1]) \\
&\geq 2\alpha^2(1 - \tanh(\beta))^2 ,
\end{aligned}$$

where the last inequality is by Lemma 5. □

As the first step in analyzing our algorithm, we first determine a sufficient number of samples to compute $\widetilde{I}_i(S)$ to a specified precision for all small sets S.

Lemma 7. *Let $\delta, \epsilon > 0$ and $k \geq 0$. If we are given M samples from a ferromagnetic Restricted Boltzmann Machine and $M \geq 2^{2k+1}(1/\epsilon^2)(\log(n) + k\log(en/k))\log(4/\delta)$, then with probability at least $1 - \delta$, for all $S \subset [n]$ such that $|S| \leq k$*

$$|I_i(S) - \widetilde{I}_i(S)| < \epsilon.$$

Proof. First observe that

$$\Pr(X_S = 1^S) \geq 2^{-|S|}$$

because in a ferromagnetic model (which by our definition has nonnegative external fields), $X_S = 1^S$ is the most likely state to observe for X_S. This inequality can also be proved by applying Griffith's inequality iteratively. Also observe that the total number of sets S we consider is $\sum_{j=0}^{k} \binom{n}{j} \leq (en/k)^k$. For each S, let M_S be the number of samples where $X_S = 1^S$. Then by Hoeffding's inequality,

$$\Pr(M_S - \mathbb{E}M_S < -t) \leq e^{-2t^2/M}.$$

In particular, since $\mathbb{E}M_S \geq 2^{-k}M$ as long as $|S| \leq k$,

$$\Pr(M_S < 2^{-k-1}M) \leq e^{-2M2^{-2k-2}}$$

Now by the usual rejection sampling argument, those samples which have $X_S = 1^S$ are independent and identically distributed samples from the conditional law. (One way to see this is that we can think of each sample as equivalently being generated by first sampling X_S, then sampling the rest of the spins conditioned on X_S). Therefore, by another application of Hoeffding's inequality, for a particular choice of i, S we have

$$\Pr(|\widetilde{I}_i(S) - I_i(S)| \geq \epsilon | M_S) \leq 2e^{-2M_S\epsilon^2}.$$

Now by the law of total expectation

$$\Pr(|\widetilde{I}_i(S) - I_i(S)| \geq \epsilon) = \mathbb{E}[\Pr(|\widetilde{I}_i(S) - I_i(S)| \geq \epsilon | M_S)]$$

$$
\begin{aligned}
&\leq 2\mathbb{E}[e^{-2M_S\epsilon^2}] \\
&= 2\mathbb{E}[(\mathbb{1}_{M_S<2^{-k-1}M} + \mathbb{1}_{M_S\geq 2^{-k-1}M})e^{-2M_S\epsilon^2}] \\
&\leq 2e^{-2M2^{-2k-2}} + 2e^{-2(2^{-k-1}M)\epsilon^2} \\
&\leq 4e^{-M2^{-2k-1}\epsilon^2}\,.
\end{aligned}
$$

And by the union bound, the probability that $|\widetilde{I}_i(S) - I_i(S)| \geq \epsilon$ for some i, S is at most

$$
n(en/k)^k 4e^{-M2^{-2k-1}\epsilon^2}\,.
$$

Therefore if we take $M \geq 2^{2k+1}(1/\epsilon^2)(\log(n)+k\log(en/k))\log(4/\delta)$ the result follows.

□

We also analyze the standard greedy algorithm for submodular maximization under noise; this corresponds to Steps 1-2 of the algorithm.

Lemma 8. *Suppose $t \geq 0$ is an integer, $f(S)$ is a monotone submodular function and $\widetilde{f}(S)$ is an approximation to f such that $|f(S) - \widetilde{f}(S)| < \epsilon$ for some uniform $\epsilon > 0$ and all S such that $|S| \leq t$. Let $S_0 = \emptyset$ and suppose S_{i+1} is formed by greedily adding to S_i the element j which maximizes $\widetilde{f}(S_i \cup \{j\})$. Then for any set T, we have*

$$
f(T) - f(S_t) \leq (1 - 1/|T|)^t f(T) + |T|\epsilon\,.
$$

Proof. Consider going from S_t to S_{t+1}. By Lemma 1, there exists some j^* such that

$$
f(S_t \cup \{j^*\}) - f(S_t) \geq \frac{f(T) - f(S_t)}{|T|}\,.
$$

Therefore for the j which is chosen to form S_{t+1}, we know

$$
(f(T)-f(S_t))-(f(T)-f(S_{t+1})) = f(S_{t+1})-f(S_t) = f(S_t\cup\{j\})-f(S_t) \geq \frac{f(T) - f(S_t)}{|T|}-\epsilon\,.
$$

Rearranging, we see that

$$
f(T) - f(S_{t+1}) \leq (1 - 1/|T|)(f(T) - f(S_t)) + \epsilon
$$

and the result follows by iterating this inequality (note that the sum of the epsilon terms forms a geometric series). □

Theorem 9. *Let $\delta > 0$. Suppose $X^{(1)}, \ldots, X^{(M)}$ are samples from the observable distribution of a ferromagnetic Restricted Boltzmann machine which is (α, β)-nondegenerate, and has two-hop degree d_2. Then if*

$$M \geq 2^{2k+3}(d_2/\eta)^2(\log(n) + k\log(en/k))\log(4/\delta)$$

where we set

$$\eta = \alpha^2\sigma(-2\beta)(1-\tanh(\beta))^2, \qquad k = d_2\log(4/\eta),$$

for every i algorithm GreedyNbhd *returns $\mathcal{N}_2(i)$, with probability at least $1-\delta$. Furthermore the total runtime is $O(Mkn^2) = e^{O(\beta d_2 - \log(\alpha))}n^2\log(n)$.*

Proof. Apply Lemma 7 with $\epsilon = \eta/(4d_2)$; then for our choice of M we have that $|\widetilde{I}_i(S) - I_i(S)| < \eta/(4d_2)$ for all S with $|S| \leq k$. Then applying Lemma 8 and using our choice of k with the inequality $1 + x \leq e^x$, we have

$$I_i(\mathcal{N}_2(i)) - I_i(S_k) \leq (1-1/d_2)^k + \eta/4 \leq \eta/2. \tag{2.1}$$

Suppose S_k does not contain the two-hop neighborhood of i. then we can take any of the two-hop neighbors $j \in \mathcal{N}_2(i) \setminus S_k$ and see that

$$I_i(\mathcal{N}_2(i)) - I_i(S_k) \geq I_i(S_k \cup \{j\}) - I_i(S_k) \geq 2\alpha^2\sigma(-2\beta)(1-\tanh(\beta))^2 = 2\eta$$

where the first inequality follows since $\mathcal{N}_2(i)$ is the global maximizer of I_i among all subsets of the observed nodes (by monotonicity and the Markov property), and the second inequality is Lemma 6. This contradicts (2.1), therefore S_k does contain the entire two-hop neighborhood of i.

It remains to show that Step 3 of the algorithm leaves in $\widetilde{\mathcal{N}}_2$ exactly the elements of S which are in the two-hop neighborhood. Since $|\widetilde{I}_i(S) - I_i(S)| < \eta/(4d_2)$ for

every set S with $|S| \le k$, this is straightforward: if j is a two-hop neighbor, then by Lemma 6 and triangle inequality we see that

$$|\widetilde{I}_i(S_k) - \widetilde{I}_i(S_k \setminus \{j\})| \ge 2\eta - \eta/2 > \eta$$

If j is not a two-hop neighbor, then $I_i(S_k) - I_i(S_k \setminus \{j\}) = 0$ by the Markov property, so by triangle inequality $|\widetilde{I}_i(S_k) - \widetilde{I}_i(S_k \setminus \{j\})| \le \eta/2 < \eta$. Thus for each i, the returned $\widetilde{\mathcal{N}}_2$ is the true two-hop neighborhood of vertex i.

To analyze the runtime, observe that the loop goes through at most k steps, and each iteration of the loop takes time $O(nM)$ to consider each j and compute $\widetilde{I}(S_t \cup \{j\})$ from samples, and we run GreedyNbhd from each of the n vertices. □

2.6.1 Improving the Sample Complexity

We consider the following algorithm for learning the two-hop neighborhood of an RBM, which is inspired by the approach of [34] for learning Ising models and MRFs (without hidden nodes). As we will show this algorithm has better sample complexity than the previous one, but sacrifices speed in order to achieve this: it runs in time $O(n^{d_2+1}\log(n))$. This leaves open the question of whether there is a *statistical-computational gap* inherent in the RBM-learning problem. As before, $\eta > 0$ is a parameter we will specify later.

Algorithm 2: SearchNbhd(i)

---

1. Let $\mathcal{F}$ be the family of subsets of n of size at most d_2 such that $S \in \mathcal{F}$ when for every j,
$$\widetilde{I}_i(S \cup \{j\}) - \widetilde{I}_i(S) \le \eta.$$
2. Return $\arg\min_{S \in \mathcal{F}} |S|$.

Theorem 10. *Algorithm* SEARCHNBHD *returns the correct neighborhood with probability at least* $1-\delta$ *given*

$$M \geq 2^{2d_2+3}(1/\eta)^2(\log(n) + d_2\log(en/d_2))\log(4/\delta)$$

samples, when $\eta = \alpha^2\sigma(-2\beta)(1-\tanh(\beta))^2$. *The algorithm runs in time* $O(n^{d_2+1}M)$.

Proof. Apply Lemma 7 with $\epsilon = \eta/4$ and $k = d_2$; then for our choice of M we have with probability at least $1-\delta$ that $|\widetilde{I}_i(S) - I_i(S)| < \eta/4$ for all S with $|S| \leq d_2$. Then, as in the proof of Theorem 9 we can apply the triangle inequality and Lemma 6 to show that $\mathcal{F}$ contains only supersets of the two-hop neighborhood, and that $\mathcal{N}_2$ lies in $\mathcal{F}$; hence $\mathcal{N}_2$ is the unique smallest set in $\mathcal{F}$ and so the output of SEARCHNBHD(i) is correct for every i. □

Note that the sample complexity is $e^{O(\beta+d_2-\log\alpha)}\log n$. This straightforwardly implies a bound for the special case of learning Ising models of bounded degree d without hidden nodes (which can be built as RBMs using a single vertex for each edge of the original model) which also has sample complexity $e^{O(\beta+d-\log\alpha)}\log n$ in terms of the edge weights of the original Ising model. Then we see by the result of [163] that for the special case of learning Ising models, this algorithm is essentially information-theoretically optimal (up to constants).

2.6.2 Learning Ferromagnetic Ising Models with Arbitrary Latent Variables

In this subsection we show how our learning algorithms can be generalized beyond the RBM setting to ferromagnetic Ising models with an arbitrary set of hidden nodes — i.e. the interaction matrix can connect pairs of observed nodes and pairs of hidden

nodes too. The marginal distribution on the observed nodes still induces a Markov Random Field, although it no longer has as simple a closed form as in Lemma 2.

In this setting, our goal is to learn the (induced) *Markov blanket* of every observed node i, which we continue to denote by $\mathcal{N}_2(i)$, and we let d_2 denote the maximum size of $\mathcal{N}_2(i)$ among all observed nodes i. The only new ingredient we need is the following generalization of Lemma 6:

Lemma 9. *Suppose i and j are nodes in an (α, β)-nondegenerate ferromagnetic Ising model. Suppose $S \subset [n]$ is a set of nodes which* do not *separate i and j: then*

$$I_i(S \cup \{j\}) - I_i(S) \geq 2\sigma(-2\beta)\alpha^k(1 - \tanh^2(\beta))^k.$$

where k is the length of the shortest path from i to j which does not go through S.

Proof. Suppose that $v_1, \ldots, v_k$ is the path from i to j so $v_1 = i$ and $v_k = j$. Then by submodularity it suffices to prove the lower bound when $S = [n] \setminus \{v_1, \ldots, v_k\}$. Since

$$\begin{aligned}\Pr(X_i = 1 | X_S = 1^S) = \Pr(X_i = 1 | X_j = 1, X_S = 1^S) \Pr(X_j = 1 | X_S = 1^S) \\ + \Pr(X_i = 1 | X_j = -1, X_S = 1^S) \Pr(X_j = -1 | X_S = 1^S)\end{aligned}$$

and $I_i(S) = 2\Pr(X_i = 1 | X_S = 1^S) - 1$ and $I_i(S \cup \{j\}) = 2\Pr(X_i = 1 | X_j = 1, X_S = 1^S) - 1$, we see

$$\begin{aligned}&\frac{1}{2}(I_i(S \cup \{j\}) - I_i(S)) \\ &= \Pr(X_j = -1 | X_S = 1^S)(\Pr(X_i = 1 | X_j = 1, X_S = 1^S) - \Pr(X_i = 1 | X_j = -1, X_S = 1^S)) \\ &\geq \sigma(-2\beta)(\Pr(X_i = 1 | X_j = 1, X_S = 1^S) - \Pr(X_i = 1 | X_j = -1, X_S = 1^S))\end{aligned}$$

by Lemma 55. Conditioned on $X_S = 1^S$, the Ising model we are considering reduces to an Ising model on a linear graph, so applying the below Lemma 10 proves the result. □

Lemma 10. *Let $X_1, \ldots, X_n$ be the spins on an (α, β)-nondegenerate ferromagnetic*

Ising model on a linear graph with vertices labeled in order as 1 *to* n*. Then*

$$\Pr(X_1 = 1 | X_n = 1) - \Pr(X_1 = 1 | X_n = -1) \geq (\alpha(1 - \tanh^2(\beta)))^{n-1}$$

Proof. We prove this by induction on n. When $n = 1$ the difference is clearly 1. In general, using that X_1, X_n are conditionally independent given X_{n-1} we see

$$\begin{aligned}
&\Pr(X_1 = 1 | X_n = 1) - \Pr(X_1 = 1 | X_n = -1) \\
&= \Pr(X_1 = 1 | X_{n-1} = 1)(\Pr(X_{n-1} = 1 | X_n = 1) - \Pr(X_{n-1} = 1 | X_{n-1} = -1)) \\
&\quad + \Pr(X_1 = 1 | X_{n-1} = -1)(\Pr(X_{n-1} = -1 | X_n = 1) - \Pr(X_{n-1} = -1 | X_{n-1} = -1)) \\
&= (\Pr(X_1 = 1 | X_{n-1} = 1) - \Pr(X_1 = 1 | X_{n-1} = -1)) \\
&\quad \cdot (\Pr(X_{n-1} = 1 | X_n = 1) - \Pr(X_{n-1} = 1 | X_{n-1} = -1)) \\
&\geq (\alpha(1 - \tanh^2(\beta)))^{n-1}
\end{aligned}$$

by the induction hypothesis and Lemma 5 □

As in the RBM case, Lemma 9 shows in particular that $\mathcal{N}_2(i)$ equals its obvious graph-theoretic analogue: the set of nodes j such that i and j are connected by a path whose intermediate nodes are all latent. We also get the following natural generalization of Theorem 9 for recovering $\mathcal{N}_2(i)$:

Theorem 11. *Let $\delta > 0$. Suppose $X^{(1)}, \ldots, X^{(M)}$ are samples from the observable distribution of an Ising model with hidden nodes which is (α, β)-nondegenerate. Suppose also that d_2 is known such that $d_2 \geq |\mathcal{N}_2(i)|$ for all observed nodes i and that for every i and $j \in \mathcal{N}_2(i)$, there is a path of length at most ℓ from i to j. Then if*

$$M \geq 2^{2k+3}(d_2/\eta)^2(\log(n) + k\log(en/k))\log(4/\delta)$$

where we set

$$\eta = \alpha^\ell \sigma(-2\beta)(1 - \tanh(\beta))^\ell, \qquad k = d_2 \log(4/\eta),$$

for every i algorithm GreedyNbhd *returns $\mathcal{N}_2(i)$, with probability at least $1 - \delta$.*

Furthermore the total runtime is $O(Mkn^2) = e^{O(\beta \ell d_2 - \ell \log(\alpha))} n^2 \log(n)$.

Proof. This is the same as proof of Theorem 9, except that we replace the use of Lemma 6 by Lemma 9. □

The corresponding analogue of Theorem 10 follows as well by using Lemma 9.

2.7 Inference on the Induced MRF via the Lee-Yang Property

We first recall various results from [124], whose approach is based on Barvinok's approach [15] for approximating the log-partition function. The basic idea is to Taylor expand $\log Z$ around the point of infinite external field, where $\log Z$ is easy to compute because only one spin configuration contributes. A *Lee-Yang property*[7] can be used to prove that the Taylor expansion is accurate.

Definition 12 (Lee-Yang property)**.** *Let* $P(z_1, \ldots, z_n)$ *be a multilinear polynomial with real coefficients.* P *has the* Lee-Yang property *if for any choice of complex numbers* $\lambda_1, \ldots, \lambda_n$ *such that* $|\lambda_i| \leq 1$ *for all* i *and* $|\lambda_i| < 1$ *for at least one* i*, we have that* $P(\lambda_1, \ldots, \lambda_n) \neq 0$.

Typically the polynomial P arises as the partition function of a Markov Random Field, where the λ_i are a re-parameterization of the external field. This is illustrated in the classical Lee-Yang theorem [121]:

Theorem 12 (Lee-Yang Theorem [121])**.** *Suppose* $J_{ij} \geq 0$ *and*

$$P(\lambda_1, \ldots, \lambda_n) := \sum_{x \in \{\pm 1\}^n} \exp(\frac{1}{2} \sum_{i,j} J_{ij} x_i x_j) \prod_{i: x_i = 1} \lambda_i,$$

so that $\left(\prod_{i=1}^n \lambda_i^{-1/2}\right) P(\lambda_1, \ldots, \lambda_n)$ *for positive real* λ_i *is the partition function of a ferromagnetic Ising model with external field* $h_i = \frac{1}{2} \log \lambda_i$*. Then* P *extends to complex* λ_i *as a multilinear polynomial with the Lee-Yang property.*

[7] Here we are following the terminology of [124]. There is an unrelated "Lee-Yang property" which appears in the literature on Lee-Yang for general real-valued spins.

The Lee-Yang property translates back to the following statement about the partition function:

Corollary 3. *Suppose* $Z(h) = \sum_{x\in\{\pm 1\}^n} \exp(\frac{1}{2}\sum_{i,j} J_{ij}x_ix_j + \sum_i h_ix_i)$ *is the partition function of a ferromagnetic Ising model with consistent non-positive external fields, i.e.* $h_i \leq 0$ *for all* i*. If we extend* Z *to complex* h*, then* $Z(h) \neq 0$ *for any* h *with* $\Re(h_i) \leq 0$ *for all* i *and* $\Re(h_i) < 0$ *for at least one* i*.*

Proof. This follows from the by taking $\lambda_i = e^{2h_i}$ so

$$Z(h) = \left(\prod_{i=1}^{n} \lambda_i^{-1/2}\right) P(\lambda_1, \ldots, \lambda_n) = e^{-\sum_{i=1}^n h_i} P(\lambda_1, \ldots, \lambda_n),$$

and using the non-vanishing of P by the previous Theorem. □

As we see the λ_i with $|\lambda_i| \leq 1$ correspond to non-positive external fields, whereas previously we assumed the external fields were non-negative. However the partition function is invariant to the global sign flip $x \mapsto -x$ so this is equivalent; this choice is made so we expand P around 0 instead of ∞. The following Lemma bounds the error made when we do this Taylor expansion.

Lemma 11 (Lemma 2.1 of [124])**.** *Suppose that*

$$Z(\lambda) = C \sum_{x\in\{\pm 1\}^n} \exp\left(\sum_{e\in E} f_e(x_e)\right) \lambda^{\#\{v : x_v = 1\}} \tag{2.2}$$

where E *is the set of edges of a hypergraph and each* f_e *is a real-valued function. Suppose* $0 < \epsilon < \frac{1}{4}$ *and*

$$m \geq \frac{|\lambda|}{1-|\lambda|}\left(\log(4n/\epsilon) + \log(\frac{1}{1-|\lambda|})\right)$$

and the values of $\frac{d^j}{d\lambda_j} Z(\lambda)|_{\lambda=0}$ *are given for* $j = 0, \ldots, m$*. Finally, suppose the Lee-Yang property holds for* $Z(\lambda)$ *as a univariate polynomial. Then for any* λ *with* $|\lambda| < 1$*, there is an algorithm which computes an additive* $\epsilon/4$*-approximation to* $\log Z(\lambda)$ *in polynomial time.*

This lemma does not specify a way to compute the needed values of $\frac{d^j}{d\lambda^j} Z(\lambda)|_{\lambda=0}$. However, for $j = 0$ this is easy to compute, because the only non-zero in the sum is when x is the all-1s vector. For $j \geq 1$, this is provided by Theorem 3.1 of [124] (building on the work of [152]) as long as the underlying hypergraph of the MRF has bounded degree. Recall that the *degree* of a vertex in a hypergraph is the number of hyperedges containing it.

Theorem 13 (Theorem 3.1 of [124])**.** *Fix $C > 0, d \in \mathbb{N}$. Suppose we are given as input an n-vertex hypergraph with edge set E of maximum degree d and maximum hyperedge size r, and $Z(\lambda)$ is defined as in (2.2). Then for any $\epsilon > 0$ there exists a deterministic $\text{poly}_{C,d,r}(n/\epsilon)$ time algorithm to compute $\frac{d^j}{d\lambda^j} Z(\lambda)|_{\lambda=0}$ for $j = 1, \ldots, m$ where $m = \lceil C \log(n/\epsilon) \rceil$.*

Finally, we describe how to apply these results to sample from the MRF induced by an RBM. The key is that, from the proof of Lemma 2, we see that the induced MRF has the same partition function as the original Ising model, so it inherits the Lee-Yang property guaranteed by Theorem 12:

Lemma 12. *Fix a ferromagnetic RBM with consistent non-positive external fields on the hidden nodes (i.e. $b_i^{(2)} \leq 0$) and with external field $b_i^{(1)} := s_i^0 + s_i$ with $s_i^0, s_i \leq 0$ on observed node i. Hence (by Lemma 2) the induced MRF has potential $g(x) + s \cdot x$ for some polynomial $g : \{\pm 1\}^{n_1} \to \mathbb{R}$ not depending on s, such that*

$$\Pr(X = x) = \frac{1}{Z(s_0 + s)} \exp(g(x) + s \cdot x)$$

for $x \in \{\pm 1\}^{n_1}$ where $Z(h^0 + h)$ is the partition function of the RBM. Let

$$P(\lambda_1, \ldots, \lambda_n) := \sum_{x \in \{\pm 1\}^{n_1}} \exp(g(x)) \prod_{i : x_i = 1} \lambda_i.$$

Then P has the Lee-Yang property.

Proof. As before, we see that if $\lambda_i = e^{2s_i}$ then

$$Z(s_0 + s) = \left(\prod_{i=1}^{n_1} \lambda_i^{-1/2} \right) P(\lambda_1, \ldots, \lambda_{n_1}) = e^{-\sum i=1^{n_1} s_i} P(\lambda_1, \ldots, \lambda_{n_1}).$$

We prove the theorem by induction on n_1, the number of observed nodes. If all of the λ_i equal 0 then it is clear that $P \neq 0$ as the sum is over only a single non-zero term. If there is at least one λ_i such that $\lambda_i = 0$, then $P(\lambda_1, \ldots, \lambda_{n_1})$ agrees with the P associated to the smaller RBM formed by conditioning $X_i = -1$, hence the non-vanishing follows by the induction hypothesis. Otherwise if all of the λ_i are non-zero, then we know by Corollary 3 that $Z(s_0 + s) \neq 0$ and we deduce that $P(\lambda) \neq 0$. □

Combining these results, we obtain the following theorem:

Theorem 14. *Fix $C > 0$ and a maximum degree d_2. Then for any ferromagnetic RBM in which $b_i^{(1)} \leq -C$ for all i, there is a deterministic polynomial time algorithm which given any $0 < \epsilon < 1/4$ and the description of the induced MRF, computes $\log Z$ within additive error $\epsilon/4$ where Z is the partition function of the induced MRF.*

Proof. By assumption we know a function f such that

$$\Pr(X = x) = \frac{1}{Z_f} \exp(f(x)).$$

If we take ω^* such that $\frac{1}{2} \log \omega^* = -H$, we see

$$Z_f = \left(\prod_{i=1}^{n_1} \omega^* \right)^{-1/2} Q(\omega^*, \ldots, \omega^*)$$

where

$$Q(\omega_i) = \sum_{x \in \{\pm 1\}^{n_1}} \exp(f(x) + H \sum x_i) \prod_{i:x_i=1} \omega_i.$$

Comparing f and Q to g and P from Lemma 12, which we apply with $s^0 = b^{(1)} + C$, we see that Q differs from P only by a multiplicative constant (corresponding to $e^{\hat{g}(\emptyset) - \hat{f}(\emptyset)}$) so Q also has the Lee-Yang property. Therefore we can compute $Q(\omega^*, \ldots, \omega^*)$ and so Z_f efficiently by the results of Lemma 12 and Theorem 13. □

The significance of accurately estimating $\log Z$ is that it allows for the performance of various inference tasks which are otherwise computationally intractable. For example, we can estimate to high precision the likelihood of observing any particular output from the MRF, since

$$\log \Pr(X = x) = p(x) - \log Z,$$

where $p(x)$ is the potential of the MRF. Hence the $\epsilon/4$ approximation to $\log Z$ from Theorem 14 implies an $\epsilon/4$ approximation to $\log \Pr(X = x)$, i.e. a PTAS for estimating $\Pr(X = x)$.

Chapter 3

Convex Hierarchies, Naive Mean-field Approximation, and Correlation Rounding

From Chapter 1, we recall that of the most widely studied probabilistic models in statistical physics and machine learning is the *Ising model*, which is a probability distribution on the hypercube $\{\pm 1\}^n$ of the form

$$P[X = x] := \frac{1}{Z} \exp\left(\sum_{i<j} J_{i,j} x_i x_j\right) = \frac{1}{Z} \exp\left(\frac{1}{2} x^T J x\right),$$

where $\{J_{i,j}\}_{i,j\in\{1,\ldots,n\}}$ are the entries of an arbitrary real, symmetric matrix with zeros on the diagonal. The distribution P is also referred to as the *Boltzmann distribution* or *Gibbs measure*. The key quantity of interest is the normalizing constant

$$Z := \sum_{x\in\{\pm 1\}^n} \exp\left(\sum_{i<j} J_{i,j} x_i x_j\right),$$

known as the *partition function* of the Ising model, and its logarithm, $\mathcal{F} := \log Z$, known as the *free energy*. The reason these are important is that one can easily extract from them many other quantities of interest, most notably the values of the

marginals (probabilities like $P[X_i = x_i]$), phase transitions in the behavior of the distribution (e.g. existence of long-range correlations), and many others.

Although originally introduced in statistical physics, Ising models and their generalizations have also found a wide range of applications in many different areas like statistics, computer science, combinatorics and machine learning (see the references and discussion in [17, 28, 191]). Consequently, various different algorithmic and analytic approaches to computing and/or approximating the free energy have been developed.

We should note at the outset that the partition function is both analytically and computationally intractable: closed form expressions for the partition function are extremely hard to derive (even for the Ising model on the standard 3-dimensional lattice), and approximating the partition function multiplicatively is NP-hard, even in the case of graphs with degrees bounded by a small constant (see [171]).

Nevertheless, there are a plethora of approaches to approximating the partition function – both for the purposes of deriving structural results, and for designing efficient algorithms. A major group of approaches consists of the so-called *variational methods*, which proceed by writing a variational expression for the free energy, and then modifying the resulting optimization problem in some way so as to make it tractable. More concretely, one can write the free energy using the *Gibbs variational principle* as

$$\mathcal{F} = \max_{\mu} \left[\sum_{i<j} J_{ij} \mathbb{E}_{\mu}[X_i X_j] + H(\mu) \right], \tag{3.1}$$

where μ ranges over all probability distributions on the Boolean hypercube. This can be seen by noting that

$$\mathbf{KL}(\mu||P) = \mathcal{F} - \sum_{i<j} J_{ij} \mathbb{E}_{\mu}[X_i X_j] - H(\mu) \tag{3.2}$$

and recalling that $\mathbf{KL}(\mu||P) \geq 0$ with equality if and only if $\mu = P$.

Of course, the polytope of distributions μ is intractable to optimize over. Two popular approaches for handling this are:

- *(Naive) Mean-field approximation*: instead of optimizing over all distributions, one optimizes over *product distributions*, thereby obtaining a lower bound on $\mathcal{F}$. In other words, we define the (naive mean-field) *variational free energy* by

$$\mathcal{F}^* := \max_{x \in [-1,1]^n} \left[\sum_{i<j} J_{ij} x_i x_j + \sum_i H\left(\frac{x_i + 1}{2}\right) \right].$$

  Indeed, if $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)$ is the optimizer in the above definition, then the product distribution ν on the Boolean hypercube with the i^{th} coordinate having expectation $\bar{x}_i$ minimizes $\mathbf{KL}(\mu||P)$ among all product distributions μ.

  This approach originated in the physics literature where it was used to great success in several cases, but from the point of view of algorithms it is *a priori* problematic: it's not clear this problem is any easier to solve, as the resulting optimization problem is highly non-convex.

- *Moment-based convex relaxations*: instead of optimizing over distributions, one optimizes over a "relaxation" (enlarging) of the polytope of distributions, thereby obtaining an upper bound on $\mathcal{F}$. There are systematic ways to do this, giving rise to *hierarchies* of convex relaxations (see, e.g. [12]). This approach is very natural and common in theoretical computer science, since the optimization problem is convex, hence efficiently solvable, although quantifying the quality of the relaxation is usually more difficult.

A priori these two approaches seem unrelated – indeed, the way they modify the variational problem is almost opposite. In this paper, we provide a unified perspective on these two approaches: for example, we show that the tight parameter regime where mean-field approximation and Sherali-Adams based approaches (even for classical MAX-k-CSP) give nontrivial guarantees is *identical*.

More precisely, we prove the following results:

1. **Simple and optimal mean-field bounds via rounding**: We obtain the optimal bounds on the quality of the mean-field approximation in a simple and elegant way. In particular, we show that there is a simple *rounding* procedure

which directly extracts a product distribution from the true Gibbs measure, and whose output is easy to analyze. More precisely, a recent result due to [97] proves that the mean-field approximation to $\mathcal{F}$ is within an additive error[1] of $O(n^{2/3}\|J\|_F^{2/3}\log^{1/3}(n\|J\|_F))$. We improve this and show:

Theorem 15. *Fix an Ising model J on n vertices. Then,*

$$\mathcal{F} - \mathcal{F}^* \leq 3n^{2/3}\|J\|_F^{2/3}.$$

We note that [97] prove this inequality is tight up to constants. This also recovers the result of [17] which shows the error is $o(n)$ when $\|J\|_F^2 = o(n)$. The technique also gives a structural result showing that certain conditional marginals are approximate fixpoints of the mean-field equations. The full results are in Section 3.3.

2. **Subexponential algorithms for approximating $\mathcal{F}$ up to the computational intractability limit**: Our proof of the above theorem is algorithmic, except that it assumes access to the true Gibbs measure. To fix this, we instead apply our rounding scheme to a convex relaxation proposed by [159] based on the Sherali-Adams hierarchy. The algorithm we get as a result runs in subexponential time so long as $\|J\|_F^2 = o(n)$; this condition for subexponentiality is tight under Gap-ETH. More precisely:

Theorem 16. *We can approximate $\mathcal{F}$ up to an additive factor of $o(n)$ in time $2^{o(n)}$ if $\|J\|_F^2 = o(n)$. Moreover, we can also output a product distribution achieving this approximation. On the other hand, for $\|J\|_F^2 = \Theta(n)$, it is Gap-ETH-hard to approximate $\mathcal{F}$ up to an additive factor of $o(n)$ in subexponential time.*

We also describe how to accelerate the algorithm on dense graphs using random subsampling. The full results are in Section 3.6.

[1]Here, $\|J\|_F := \sqrt{\sum_{i,j} J_{i,j}^2}$ is the *Frobenius norm* of the matrix J.

3. **Optimality of correlation rounding**: The rounding we use in the proof of the above theorems relies crucially on the *correlation rounding* technique introduced in [12]. This procedure was designed specifically to tackle dense and spectrally well-behaved instances of constraint satisfaction problems, as well as to derive subexponential algorithms for unique games. In order to better understand the efficacy of correlation rounding, Allen, O'Donnell, and Zhou [3] introduced a conjecture on the number of variables one needs to condition on in an arbitrary distribution, in order to guarantee that the remaining pairs of variables have average covariance at most ϵ. The current best result of [156] gives a bound of $O(1/\epsilon^2)$; [3] conjectured that this can be decreased to $O(1/\epsilon)$. We refute this conjecture in essentially the strongest possible sense. Namely, we show:

   Theorem 17. *There exists an absolute constant $C > 0$, a sequence of pairs (t_n, n) going to infinity, and a family of probability distributions (the SK spin glass) such that for* any set T *with $|T| \leq t_n$,*

$$\mathbb{E}_{(i,j)\sim\binom{[n]}{2}} \left[|\mathrm{Cov}(X_i, X_j)| \, | (X_k)_{k\in T} \right] \geq \frac{C}{\sqrt{t_n}}.$$

   We prove this theorem by combining our techniques with rigorous results on the Sherrington-Kirkpatrick spin glass. The full results are in Section 3.4.

4. **Generalization of all results to k-MRFs:** We give natural and tight generalizations of these results to order k Markov Random Fields. In general, we show that the tight regime for $o(n)$ additive error for both mean-field and sub-exponential time algorithms (under Gap-ETH) is $\|J\|_F^2 = o(n^{3-k})$, and show tightness of the higher-order correlation rounding guarantee. The full results are in Section 3.5.

3.1 Background and related work

3.1.1 The mean-field approximation

Owing to its simplicity, the mean field approximation has long been used in statistical physics (see [151] for a textbook treatment) and also in Bayesian statistics [7, 101, 191], where it is one of the prototypical examples of a *variational method*. It has the attractive property that it always gives a lower bound for the free energy.

The critical points of $\mathcal{F}^*$ have a fixpoint interpretation as the solutions to the *mean-field equation*, $x = \tanh^{\otimes n}(Jx)$. However, iterating this equation is known to converge to the mean-field solution only in high-temperature regimes such as Dobrushin uniqueness; as soon as we leave this regime, the iteration may fail to converge to the optimum even in simple models (Curie-Weiss) – see [97]. In contrast, [59] established a *structural result*, without relying on a high-temperature assumption, showing that the Gibbs measure can be decomposed into approximate fixpoints of the mean-field equations. In Section 3.3.1 we derive a similar result using the correlation rounding decomposition.

It is well known [62] that the mean field approximation is very accurate for the Curie-Weiss model (the Ising model on the complete graph) at all temperatures. On the other hand, it is also known [52] that for very sparse graphs like trees of bounded arity, this is not the case. In recent years, considerable effort has gone into bounding the error of the mean-field approximation on more general graphs; we refer the reader to [17, 97] for a detailed discussion and comparison of results in this direction. If one only wishes to show that the mean-field approximation asymptotically gives the correct free energy density $\mathcal{F}/n$ and does not care about the rate of convergence, then the breakthrough result is due to [17], who provided an exponential improvement over previous work of [28] to identify the regime where this happens.

Theorem 18 ([17])**.** *Let $(J_n)_{n=1}^{\infty}$ be a sequence of Ising models indexed by the number of vertices. if $\|J_n\|_F^2 = o(n)$, then $\mathcal{F}_{J_n} - \mathcal{F}_{J_n}^* = o(n)$.*

This result is tight – there are simple examples of models with $\|J_n\|_F^2 = \Theta(n)$

where $\mathcal{F}_{J_n} - \mathcal{F}^*_{J_n} = \Omega(n)$. On the other hand, if one also cares about the rate of convergence, then this result is not the best known. Here, improving on previous bounds of [28], [17], and [57], it was shown by [97] that:

Theorem 19 ([97]). *Fix an Ising model J on n vertices. Then,*

$$\mathcal{F} - \mathcal{F}^* \leq 200 n^{2/3} \|J\|_F^{2/3} \log^{1/3}(n\|J\|_F + e).$$

As stated earlier, our first main result Theorem 15 removes the logarithmic term in Theorem 19, thereby completely subsuming both of the theorems stated above. A more general version of this theorem, valid for higher-order Markov random fields on arbitrary finite alphabets, is Theorem 25 below.

3.1.2 Algorithms for dense CSPs

At first glance, the condition that $\|J\|_F^2 = o(n)$ may seem a little odd. To demystify it, consider the anti-ferromagnetic Ising model corresponding[2] to MAX-CUT on a graph with m edges which has $J_{ij} = -\frac{\beta n}{m}$ for each $(i,j) \in E$. If M is the optimum fraction of edges cut, then

$$\frac{1}{n\beta} \log Z \in \left[\mathrm{M} - \frac{1}{\beta}, \mathrm{M} + \frac{1}{\beta} \right], \qquad \|J\|_F^2 = \Theta\left(\beta^2 \frac{n^2}{m} \right), \tag{3.3}$$

so the requirement that $\|J\|_F^2 = o(n)$ is the same as requiring $m = \omega(n)$. In other words, our algorithms operate in the regime where the average degree is super-constant and the objective is to approximate MAX-CUT within factor $(1-\epsilon)$. Thus, they can be viewed as free-energy generalizations of optimization problems on dense graphs.

We briefly survey relevant work on approximation algorithms for dense graphs. The main emphasis in the literature has been on the case when $m = \Theta(n^2)$ for which PTASs have been developed, for instance the weak regularity lemma based algorithm

[2]The scaling here is chosen so that if the MAX-CUT is γn edges with $\gamma > 1/2$, then the two terms in (3.1) are of the same scale.

of [68], the greedy algorithms of [134], and the Sherali-Adams based approach of [51]. On the other hand, if $m = \Theta(n^{2-\epsilon})$ for any $\epsilon > 0$ then no PTAS for even MAX-CUT is possible [50].

The work most relevant to ours is the improved analysis of the Sherali-Adams relaxation due to [199] based on correlation rounding. Surprisingly, although there are many methods to approximate MAX-CUT when $m = \Theta(n^2)$ as mentioned above, to our knowledge *none of the algorithms except for Sherali-Adams* are guaranteed to give sub-exponential time algorithms down to $m = \omega(n)$; for example, the method of [68] is only sub-exponential time for $m = \omega(n \log n)$ and this is a fundamental barrier with their technique. This sub-exponential time guarantee for Sherali-Adams in this regime is not explicitly stated in [199] or anywhere else, as far as we are aware, but is straightforward to show even from the older correlation rounding guarantee of [156] (see Section 3.6). The correct generalization of this guarantee for MAX-k-CSP was essentially pointed out in [65] but once again, their algorithm misses the tight regime (achievable by Sherali-Adams) by poly-logarithmic factors. Our result recovers the tight regime (i.e. $\omega(n^{k-1})$ constraints) in this setting as well, while also generalizing to the free energy (see Section 3.6).

For computing the free energy, the two most relevant works are [159] and [97]: the first work does not make any connection to mean-field approximation and proves a slightly weaker guarantee for Sherali-Adams than the current work; the second work uses a regularity based approach to compute the mean-field approximation, and gets similar guarantees to the algorithm of this work but misses the correct sub-exponential time regime by log factors.

3.1.3 Correlation rounding, and a refutation of the Allen-O'Donnell-Zhou conjecture

Let $X_1, \dots, X_n$ be a collection of jointly distributed random variables, each of which takes values in $\{\pm 1\}$. There are two possibilities for such a collection:

- The average covariance of the collection, defined to be $\mathbb{E}_{(i,j)\sim\binom{[n]}{2}}|\mathrm{Cov}(X_i, X_j)|$,

is small.

- The average covariance of the collection is not small: in this case, we expect a random coordinate X_j to contain significant information about many of the other random variables in $X_1, \ldots, X_n$, so that we might intuitively conjecture that conditioning on the random variables X_j for all j in a 'small' random subset T of $[n]$ makes the average covariance sufficiently small.

This intuition is indeed true and has been quantitatively formalized by several works in the theoretical computer science community, including [12, 85, 156, 199]. We note that similar ideas have appeared independently in the statistical physics literature under the name of 'pinning'; for example, see [96] and references therein, as well as recent work [46]. To the best of our knowledge, the historically first statement of (essentially) the Theorem below was given in [141].

Theorem 20 ([156, 12, 141])**.** *Let $X_1, \ldots, X_n$ be a collection of $\{\pm 1\}$-valued random variables, and let $0 < \epsilon \leq 1$. Then, for some integer $0 \leq t \leq O(1/\epsilon^2)$:*

$$\mathbb{E}_{T \sim \binom{V}{t}} \mathbb{E}_{(i,j) \sim \binom{[n]}{2}} \left[\left| \mathrm{Cov}(X_i, X_j) \right| \left| (X_k)_{k \in T} \right. \right] \leq \epsilon.$$

The above theorem is at the heart of the so-called *correlation rounding* technique for the Sherali-Adams and SOS convex relaxation hierarchies, which has been used to provide state-of-the-art approximation algorithms for many classic NP-hard problems and their variants; we refer the reader to the references above for much more on this. As we will see below, it will also be key to our proof of Theorem 15.

Recently, it was conjectured by Allen, O'Donnell and Zhou [3] that the upper bound on t in Theorem 20 can be improved significantly. More precisely, they conjectured that:

Conjecture 1 (Conjecture A in [3])**.** *Theorem 20 holds with $0 \leq t \leq O(1/\epsilon)$.*

Their motivation for this conjecture was twofold:

- On a technical level, the proof of Theorem 20 in [156] proceeds by first showing that for some integer $0 \leq t \leq O(1/\epsilon^2)$

$$\mathbb{E}_{T \sim \binom{V}{t}} \mathbb{E}_{(i,j) \sim \binom{[n]}{2}} \left[|I(X_i, X_j)| \, | (X_k)_{k \in T} \right] \leq \epsilon^2,$$

  where $I(X, Y)$ denotes the *mutual information* between X and Y, and then using the standard inequality $|\mathrm{Cov}(X, Y)| \leq \sqrt{2I(X, Y)}$; we will present a generalized version of this proof from [132, 199] later. Essentially, they conjectured that one could surmount the quadratic loss without passing through mutual information.

- From a complexity-theoretic point of view, the best known lower bounds for dense MAX-CSP problems (such as [1, 132]) leave open the possibility that MAX-CUT on n vertices can be computed to within ϵn^2 additive error in time $n^{O(1/\epsilon)}$, whereas the best known algorithms all require time at least $2^{O(1/\epsilon^2)}$. If Conjecture 1 were true, the running time of the Sherali-Adams based approach would have improved to $n^{O(1/\epsilon)}$ time for $\epsilon n^2 \|J\|_\infty$ error (which, for dense graphs, is close to matching the lower bound of [132]).

[3] prove Conjecture 1 for the special case when the random variables $X_1, \ldots, X_n$ are the leaves of a certain type of information flow tree known as the caterpillar graph. In addition, [132] showed a similar improvement for correlation rounding in the MAX k-CSP problem, when promised that there exists an assignment satisfying all of the constraints. As described in the introduction (Theorem 17), we use ideas from statistical physics to refute Conjecture 1 in essentially the strongest possible form by showing that Theorem 20 does not hold with $0 \leq t \leq o(1/\epsilon^2)$.

3.2 Technical tools

3.2.1 Hierarchies of convex relaxations

Computing the free energy of an Ising model has as a special case the problem MAX-QP/MAX-2CSP, because if we let $J_\beta = \beta J$ then

$$\lim_{\beta\to\infty} \frac{1}{\beta} \log Z(J_\beta) = \lim_{\beta\to\infty} \sup_{\mu} \left(\frac{1}{2}\mathbb{E}[X^T J X] + \frac{1}{\beta} H(\mu) \right) = \max_{x\in\{\pm 1\}^n} x^T J x. \tag{3.4}$$

As with many other problems in combinatorial optimization, this is a maximization problems on the Boolean hypercube, i.e. as a problem of the form

$$\max_{x\in\{\pm 1\}^n} f(x).$$

These problems are often NP-hard to solve exactly, but *convex hierarchies* give a principled way to write down a natural family of convex relaxations which are efficiently solvable and give increasingly better approximations to the true value. First, one re-expresses the problem as an optimization problem over the convex polytope of *probability distributions* using that

$$\max_{x\in\{\pm 1\}^n} f(x) = \max_{\mu\in\mathcal{P}(\{\pm 1\}^n)} \mathbb{E}_\mu[f(x)];$$

the advantage of this reformulation is that the objective is now linear in the variable μ. Second, one relaxes $\mathcal{P}(\{\pm 1\}^n)$ to a larger convex set of *pseudo-distributions* which are more tractable to optimize over. The tightness of relaxation is controlled by a parameter r (known as the *level* or *number of rounds* of the hierarchy); as the parameter r increases, the relaxation becomes tighter with the level n relaxation corresponding to the original optimization problem.

Different hierarchies correspond to different choices of the space of pseudo-distributions; two of the most popular are the *Sherali-Adams (SA) hierarchy* and the *Sum-of-Squares (SOS)/Laserre hierarchy*. In the *Sherali-Adams hierarchy*, we define a *level r-pseudodistribution* to be given by the following variables and constraints:

1. For every $S \subset [n]$ with $|S| = r$, a valid joint distribution μ_S over $\{\pm 1\}^S$.

2. *Compatability conditions*, which require that for every $U \subseteq [n]$ with $|U| \leq r$ and every $S, S' \subseteq [n]$ with $|S| = |S'| = r$ and $U \subset S \cap S'$, $\mu_S|_U = \mu_{S'}|_U$.

Observe that, by linearity, this data defines a unique *pseudo-expectation operator*[3] $\tilde{\mathbb{E}}$ from real polynomials of degree at most r to $\mathbb{R}$.

Let SA_r denote the set of level r-pseudodistributions on the hypercube. Then for $r \geq \deg(f)$, we can write down $\max_{\mu \in SA_r} \tilde{\mathbb{E}}_\mu[f(x)]$ as a linear program with $2^r \binom{n}{r}$ many variables and a number of constraints which is polynomial in the number of variables. By *strong duality* for linear programs, we can also think of the value of the level r SA relaxation as corresponding to the best upper bound derivable on $\sup_\mu \mathbb{E}_\mu[f(x)]$ in a limited proof system, which captures e.g. case analysis on sets of size at most r.

In addition to this standard setup, since the variational formulation for $\log Z$ has an entropy term, we will need a proxy for it when we use the Sherali-Adams hierarchy. The particular proxy we will use was introduced by [159] – further details are in Section 3.6.

3.2.2 The correlation rounding lemma

As mentioned in the introduction, our proof of Theorem 15 will depend crucially on the correlation rounding lemma. Here, we present a general higher-order version of this lemma due to [132], building on previous work of [156] and [199].

Definition 2. *The* multivariate total correlation *of a collection of random variables $X_1, \ldots, X_n$ is defined to be*

$$C(X_1; \cdots; X_n) = \mathbf{KL}\left(\mu(X_{1,\ldots,n}) || \mu(X_1) \times \cdots \times \mu(X_n)\right).$$

[3] This operator may behave very differently from a true expectation. For example, it's possible that $\tilde{\mathbb{E}}[f^2] < 0$ for some f. The SOS hierarchy is formed by additionally requiring $\tilde{\mathbb{E}}[f^2] \geq 0$ for all low-degree f.

From the definition of **KL** divergence, it follows that

$$C(X_1; \cdots ; X_n) = \left(\sum_{i=1}^{n} H(X_i) \right) - H(X_1, \ldots, X_n).$$

By using conditional distributions/ conditional entropies, we may define the *conditional multivariate total correlation* in the obvious way. Note that in the two-variable case, the total correlation is the same as the *mutual information* $I(X_1; X_2)$.

Theorem 21 (Correlation rounding lemma, [132]). *Let $X_1, \ldots, X_n$ be a collection of $\{\pm 1\}$-valued random variables. Then, for any $k, \ell \in [n]$, there exists some $t \leq \ell$ such that:*

$$\mathbb{E}_{S \sim \binom{V}{t}} \mathbb{E}_{F \sim \binom{V}{k}} [C(X_F | X_S)] \leq \frac{k^2 \log(2)}{\ell}.$$

Remark 2. *The same conclusion holds for general random variables $X_1, \ldots, X_n$ with the factor $\log 2$ replaced by $\frac{\sum_{i=1}^{n} H(X_i)}{n}$. Also, the guarantee holds for general level $(\ell + k)$-pseudodistributions.*

For the reader's convenience, we provide a complete proof of this result in Section 3.8, correcting certain errors which have been persistent in the literature.

3.2.3 The Sherrington-Kirkpatrick model

The famous Sherrington-Kirkpatrick (SK) spin glass model was introduced in the work [107] as a solvable model of disordered systems. The Gibbs measure of the SK spin glass on n vertices (without external field) is a random probability distribution on $\{\pm 1\}^n$ given by:

$$\Pr(X = x) := \frac{1}{Z_n(\beta)} \exp\left(\frac{\beta}{\sqrt{n}} \sum_{1 \leq i < j \leq n} J_{ij} X_i X_j \right),$$

where $J_{ij} \sim N(0, 1)$ are i.i.d. standard Gaussians and β is a fixed parameter referred to as the *inverse temperature*. In [107], a prediction, now known as the *replica-symmetric prediction*, was made for the limiting value of $\frac{1}{n} \log Z_n(\beta)$ as $n \to \infty$. It was soon realized that this prediction could not be correct for all values of β;

finding and understanding the correct prediction led physicists to the development of a sophisticated theory of (mean-field) spin glasses [139]. In particular, physicists discovered that the SK spin glass exhibits two *phases* depending on the value of β (here we are only considering the $h = 0$ case of no external field):

1. *Replica Symmetry* (RS, $\beta < 1$). This is the regime where the original prediction for the limiting value of $\frac{1}{n} \log Z_n(\beta)$ is correct. Moreover, the Gibbs measure exhibits a number of unusual properties: for example, the marginal law on any small subset of the coordinates converges to a product distribution as $n \to \infty$ ([178]).

2. *(Full) Replica Symmetry Breaking* (fRSB, $\beta > 1$). In this phase, the limit of $\frac{1}{n} \log Z_n(\beta)$ does not have a simple closed form; however, there is a remarkable variational expression for the limiting value known as the *Parisi formula*. Moreover, the Gibbs measure is understood to be shattered into multiple clusters with the geometry of an *ultrametric space*. (We leave a more precise description of this complex situation to the references.)

In the replica symmetric phase (with no external field), the prediction for the limiting value of $\frac{1}{n} \log Z_n(\beta)$ was rigorously confirmed by the work of [2]. Furthermore, they proved their result for general distributions of the disorder J_{ij}, giving what is known as a *universality* result.

Theorem 22 ([2])**.** *Let $\epsilon > 0$. For the SK spin glass at inverse temperature $\beta < 1$,*

$$\Pr\left(\left|\frac{1}{n} \log Z_n(\beta) - (\log 2 + \beta^2/4)\right| \geq \epsilon\right) \to 0$$

as $n \to \infty$. Moreover, this also holds if the J_{ij} are i.i.d. samples from any *distribution with finite moments, mean* 0 *and variance* 1*.*

This is the only result we will need from the spin glass literature. For an account of more recent developments, including the proofs of the Parisi formula and ultrametricity conjecture, we refer the reader to the books [149, 178, 179].

3.3 Mean-field approximation via correlation rounding: proof of Theorem 15

First we recall a couple of lemmas which are essentially used in all works on correlation rounding. Recall that for two probability distributions P and Q on the same finite space Ω, the total variation distance between P and Q is defined by $\mathbf{TV}(P,Q) := \sup_{A\subseteq\Omega} \left|\sum_{a\in A}(P(a)-Q(a))\right|$.

Lemma 13 (Lemma 5.1, [12])**.** *Let X and Y be jointly distributed random variables valued in $\{\pm 1\}$. Let P_X, P_Y denote the marginal distributions of X and Y, and let $P_{X,Y}$ denote their joint distribution. Then,*

$$|\mathrm{Cov}(X,Y)| = 2\mathbf{TV}(P_{X,Y}, P_X \times P_Y).$$

From this, one can observe the following consequence of correlation rounding:

Lemma 14. *Let $X_1,\ldots,X_n$ be a collection of $\{\pm 1\}$-valued random variables. Then, for any $\ell \in [n]$, there exists some $S \subset [n]$ with $|S| \le \ell$ such that:*

$$\mathbb{E}_{X_S}\mathbb{E}_{\{u,v\}\in\binom{V}{2}}\left[\mathrm{Cov}(X_u, X_v|X_S)^2\right] \le \frac{8\log 2}{\ell}.$$

Proof. This is standard and we include the proof for completeness. We begin by applying Theorem 21 with ℓ; let S denote the resulting set of size at most ℓ. By Pinsker's inequality, we have

$$2\mathbf{TV}^2\left(\mu(X_{u,v}|X_S=x_s), (\mu(X_u|X_S=x_s)\times\mu(X_v|X_S=x_s))\right) \le C(X_u;X_v|X_S=x_s),$$

for any $x_s \in \{\pm 1\}^{|S|}$. Therefore, by taking the expectation on both sides, we get:

$$2\mathbb{E}_{X_S}\mathbf{TV}^2\left(\mu(X_{u,v}|X_S), (\mu(X_u|X_S)\times\mu(X_v|X_S))\right) \le C(X_u;X_v|X_S).$$

By averaging over the choice of $\{u, v\} \in \binom{V}{2}$, we get

$$\mathbb{E}_{E=\{u,v\}\sim\binom{V}{2}}\mathbb{E}_{X_S}\left[\mathbf{TV}^2\left(\mu(X_{u,v}|X_S), (\mu(X_u|X_S)\times\mu(X_v|X_S))\right)\right] \le \mathbb{E}_{E\sim\binom{V}{2}}\left[\frac{C(X_E|X_S)}{2}\right] \le \frac{2\log 2}{\ell},$$

where the second inequality follows by the choice of S and Theorem 21. Finally, Lemma 13 shows that for any $x_s \in \{\pm 1\}^{|S|}$,

$$|\mathrm{Cov}(X_u, X_v|X_S = x_S)| \le 2\mathbf{TV}\left(\mu(X_{u,v}|X_S = x_S), (\mu(X_u|X_S = x_S)\times\mu(X_v|X_S = x_S))\right),$$

from which we obtain our desired conclusion:

$$\mathbb{E}_{X_S}\mathbb{E}_{\{u,v\}\in\binom{V}{2}}\left[\mathrm{Cov}(X_u, X_v|X_S)^2\right] \le \frac{8\log 2}{\ell}. \tag{3.5}$$

□

Finally, we recall the *maximum-entropy principle* characterizing product distributions:

Lemma 15. *Let μ denote a probability distribution on the finite space $\Omega_1 \times \cdots \times \Omega_n$. Let ν denote the product distribution on $\Omega_1 \times \cdots \times \Omega_n$ whose marginal distribution on Ω_i is the same as that of μ for all $i \in [n]$. Then, $H(\mu) \le H(\nu)$.*

Proof. This is a direct application of the chain rule and tensorization for entropy. Indeed, let $X := (X_1, \ldots, X_n) \sim \mu$. Then,

$$H(\mu) = H(X) \le H(X_1) + \cdots + H(X_n) = H(\nu).$$

□

We are now ready to prove Theorem 15.

Proof of Theorem 15. Let $\epsilon > 0$ be some parameter which will be optimized later. We begin by applying Lemma 14 with $\ell = 1/(\epsilon^2 \log 2)$ (for clarity of exposition, we

will omit floors and ceilings since they do not make any essential difference); let S denote the resulting set of size at most ℓ. Let μ denote the Boltzmann distribution, and recall that the Gibbs variational principle Eq. (3.1) states that

$$\mathcal{F} = \mathbb{E}_\mu \left[X^T J X \right] + H_\mu(X).$$

Let ν_{x_S} denote the product distribution on $\{\pm 1\}^n$ for which $\mathbb{E}_{\nu_{x_S}}[X_i] = \mathbb{E}[X_i | X_S = x_S]$. Then, using the chain rule for entropy, we see that

$$\begin{aligned}
\mathcal{F} &= \sum_{i<j} J_{i,j} \mathbb{E}_\mu[X_i X_j] + H_\mu(X) \\
&= \sum_{i<j} J_{i,j} \mathbb{E}_\mu[X_i X_j] + H_\mu(X|X_S) + H_\mu(X_S) \\
&= \mathbb{E}_{x_S} \left[\sum_{i<j} J_{ij} \mathbb{E}_\mu[X_i X_j | X_S = x_S] + H_\mu(X | X_S = x_S) \right] + H_\mu(X_S) \\
&\leq \mathbb{E}_{x_S} \left[\sum_{i<j} J_{ij} \mathbb{E}_\mu[X_i X_j | X_S = x_S] + H_\mu(X | X_S = x_S) \right] + 1/\epsilon^2 \\
&\leq \mathbb{E}_{x_S} \left[\sum_{i<j} J_{ij} \mathbb{E}_\mu[X_i X_j | X_S = x_S]] + H_{\nu_{x_s}}(X) \right] + 1/\epsilon^2,
\end{aligned}$$

where in the fourth line, we have used that $|S| \leq \ell = 1/(\epsilon^2 \log 2)$, and in the last line, we have used Lemma 15. From Lemma 14 and the Cauchy-Schwarz inequality, it follows that

$$\begin{aligned}
\mathbb{E}_{X_S} \left[\sum_{i<j} J_{ij} \mathbb{E}_\mu[X_i X_j | X_S] \right] &= \mathbb{E}_{x_S} \left[\sum_{i<j} J_{ij} \left(\mathrm{Cov}(X_i, X_j | X_S = x_S) + \mathbb{E}_\mu[X_i | X_S = x_S][X_j | X_S = x_S] \right) \right] \\
&= \sum_{i<j} J_{i,j} \mathbb{E}_{X_S}[\mathrm{Cov}(X_i, X_j | X_S)] + \mathbb{E}_{X_S} \sum_{i,j} J_{i,j} \mathbb{E}_{\nu_{X_S}}[X_i X_j] \\
&\leq \sqrt{\sum_{i<j} J_{i,j}^2} \sqrt{2 \binom{|V|}{2} \mathbb{E}_{X_S} \mathbb{E}_{E \in \binom{V}{2}} [\mathrm{Cov}(X_u, X_v | X_S)^2]} + \mathbb{E}_{X_S} \sum_{i<j} J_{i,j} \mathbb{E}_{\nu_{X_S}}[X_i X \\
&\leq 2\epsilon n \|J\|_F + \mathbb{E}_{X_S} \sum_{i<j} J_{i,j} \mathbb{E}_{\nu_{X_S}}[X_i X_j].
\end{aligned}$$

To summarize, we have shown that

$$\mathcal{F} \leq \mathbb{E}_{x_S}\left[\sum_{i<j} J_{i,j}\mathbb{E}_{\nu_{x_S}}[X_iX_j] + H_{\nu_{x_S}}(X)\right] + 2\epsilon n\|J\|_F + \frac{1}{\epsilon^2}.$$

In particular, there exists some choice of x_S, such that with $\nu := \nu_{x_S}$, we have

$$\mathcal{F} \leq \left[\sum_{i<j} J_{ij}\mathbb{E}_{\nu}[X_iX_j] + H_{\nu}(X)\right] + 2\epsilon n\|J\|_F + 1/\epsilon^2.$$

Finally, by setting $\epsilon = \frac{1}{n^{1/3}\|J\|_F^{1/3}}$ we get the desired conclusion:

$$\mathcal{F} \leq E_{\nu}\left[\sum_{i<j} J_{ij}X_iX_j + H(X)\right] + 3n^{2/3}\|J\|_F^{2/3}.$$

□

Remark 3. *For the choice of ϵ in the above proof to make sense, we require that $\ell = 1/(\epsilon^2 \log 2) \leq n$, which translates to $\|J\|_F^{2/3} \leq n^{1/3}\log 2$. However, the above bound also holds if $\|J\|_F^{2/3} > n^{1/3}\log 2$ since in this case, our error term equals $3\log 2n > 2n$, whereas there is a trivial upper bound of $n\log 2$ on $\mathcal{F} - \mathcal{F}^*$, obtained by considering the product distribution supported at the point $\arg\max_{x\in\{\pm 1\}^n}\{\sum_{ij} J_{ij}x_ix_j\}$.*

3.3.1 Aside: correlation rounding and the mean-field equation

The above proof shows that for the product measure $\nu := \nu_{x_S}$, $\mathcal{F}_\nu$ is close to $\mathcal{F}$. This shows indirectly, by considering the maximizer of $\mathcal{F}^*$, that there exists a product distribution with marginals that are an *exact solution* to the mean-field equation $x = \tanh^{\otimes n}(Jx)$ which is close to the Gibbs distribution in **KL** distance. In this subsection, we show that the marginals output by correlation rounding are already an *approximate solution* to the mean-field equation, given an additional assumption on J. It will be easier to prove the result if we generalize to Ising models with external field h, in which case $\Pr(X = x) = \frac{1}{Z}\exp(\frac{1}{2}x^TJx + h\cdot x)$ and the mean-field equation is $x = \tanh^{\otimes n}(Jx + h)$. We recall the following lemma:

Lemma 16 ([59])**.** *Let Z be an arbitrary random variable such that $|Z| \leq L$ almost surely for $L \geq 1$. Then*

$$|\tanh(\mathbb{E}Z) - \mathbb{E}\tanh(Z)| \leq 20L \cdot \mathbb{E}|\tanh(Z) - \mathbb{E}\tanh(Z)|$$

Lemma 17. *Let $X_1, \ldots, X_n$ be the spins of an Ising model with interaction matrix J and external field h and let $L = \max_i(\|J_i\|_1 + |h_i|)$. Then*

$$\frac{1}{n}\sum_i |\tanh(J_i \cdot X + h_i) - \tanh(\mathbb{E}J_i \cdot X + h_i)|^2 \leq 400L^2\|J\|_F\sqrt{\mathbb{E}_{j,k}\mathrm{Cov}(X_j, X_k)^2}$$

Proof. Since $\mathbb{E}[X_i|X_{\sim i}] = \tanh(J_i \cdot X)$, we find by Lemma 16 that

$$\begin{aligned}
|\tanh(J_i \cdot X + h_i) - \tanh(\mathbb{E}J_i \cdot X + h_i)|^2 &\leq 400L^2(\mathbb{E}|\tanh(J_i \cdot X + h_i) - \mathbb{E}|\tanh(J_i \cdot X + h_i)|)^2 \\
&\leq 400L^2\mathrm{Var}(\tanh(J_i \cdot X + h_i)) \\
&\leq 400L^2\mathrm{Cov}(\tanh(J_i \cdot X + h_i), J_i \cdot X) \\
&= 400L^2\sum_j J_{ij}\mathrm{Cov}(X_i, X_j)
\end{aligned}$$

where the second inequality is by Jensen's inequality, the third inequality follows because tanh is increasing and 1-lipschitz and h_i is deterministic, and the final equality follows because $\mathrm{Cov}(\mathbb{E}[X|Y], Y) = \mathrm{Cov}(X, Y)$. Summing over i, we find

$$\begin{aligned}
\frac{1}{n}\sum_i |\tanh(J_i \cdot X + h_i) - \tanh(\mathbb{E}J_i \cdot X + h_i)|^2 &\leq \frac{1}{n}400L^2\sum_{i,j} J_{ij}\mathrm{Cov}(X_i, X_j) \\
&\leq 400L^2\|J\|_F\sqrt{\mathbb{E}_{j,k}\mathrm{Cov}(X_j, X_k)^2}
\end{aligned}$$

where the last inequality is Cauchy-Schwarz. □

Finally, correlation rounding controls the average covariance, giving us our desired result – after conditioning, the marginals approximately satisfy the mean-field equation.

Theorem 23. *Let $X_1, \ldots, X_n$ be the spins of an Ising model with interaction matrix*

J and external field h, and let $L = \max_i(\|J_i\|_1 + |h_i|)$. Fix ℓ and let S with $|S| \le \ell$ be the set given by Lemma 14. Let $Y_i = \mathbb{E}[X_i|X_S]$. Then

$$\mathbb{E}_{X_S}\left[\frac{1}{n}\sum_{i=1}^{n}|Y_i - \tanh(J_i \cdot Y + h_i)|^2\right] = O\left(\frac{\ell}{n} + \frac{1}{\sqrt{\ell}}L^2\|J\|_F\right)$$

Proof. We split the sum as

$$\begin{aligned}\sum_{i=1}^{n}|Y_i - \tanh(J_i \cdot Y + h_i)|^2 &= \sum_{i\in S}|Y_i - \tanh(J_i \cdot Y + h_i)|^2 + \sum_{i\notin S}|Y_i - \tanh(J_i \cdot Y + h_i)|^2 \\ &\le \ell + \sum_{i\notin S}|Y_i - \tanh(J_i \cdot Y + h_i)|^2\end{aligned}$$

To bound the latter sum, recall that conditioning on $X_S = x_S$ gives an Ising model on the remaining spins $X_S \neq x_S$. Therefore the result follows from Lemma 17 by taking the expectation over X_S and applying Jensen's inequality, which lets us we bound the covariance by Lemma 14. □

To interpret this, suppose that $L = O(1)$. Then (as in our previous uses of correlation rounding) we see that if $\|J\|_F^2 = o(n)$, then there exists an ℓ such that the error is $o(1)$. The proof of this result generalizes in a straightforward way to k-MRFs. When the alphabet is binary we have (in the notation of Section 3.5) that if $(\nabla f)_i = \partial_i f$ is the discrete derivative, then the same result holds where the mean-field equation is $x = \tanh^{\otimes n}(\nabla f(x) + \nabla h(x))$, and where $L = \max_i(\|\partial_i f\|_\infty + \|h_i\|_\infty)$.

3.4 Correlation rounding is tight for spin glasses: proof of Theorem 17

We define the following universal constant, which we already know an upper bound on by Theorem 20:

$$\kappa_* := \limsup_{t\to\infty} \sup_{\substack{\mu\in\mathcal{P}(\{\pm1\}^n)\\ n\ge t}} \min_{S:|S|\le t} \sqrt{t}\ \mathbb{E}_{(i,j)\sim\binom{[n]}{2}}[|\mathrm{Cov}(X_i, X_j|X_S)|].$$

If Conjecture 1 were true, then we would have $\kappa_* = 0$ – indeed, the conjecture says that the expected conditional covariance decays like $O(1/t)$, even for a random choice of the conditioning set S. We will instead show an explicit positive lower bound on κ_*, thereby disproving the conjecture.

We begin by proving a variant of Theorem 15, which gives a bound on the error of the mean-field approximation in terms of the constant κ_*.

Lemma 18. *Let $\{J_n\}_{n\geq 1}$ be a sequence of Ising models indexed by the number of vertices. Let $\mathcal{F}_n$ (resp. $\mathcal{F}_n^*$) denote the free energy (resp. variational free energy) of J_n. Suppose that $\kappa_*^2 \limsup_{n\to\infty} n\|J_n\|_\infty^2 < 16$. Then,*

$$\limsup_{n\to\infty} \frac{\mathcal{F}_n - \mathcal{F}_n^*}{n^{4/3}\|J_n\|_\infty^{2/3}} \leq \frac{3}{\sqrt[3]{4}}\kappa_*^{2/3}.$$

Proof. Let $\{t_n\}_{n\geq 1}$ be a sequence of natural numbers going to infinity, which will be specified later; our choice will be such that $t_n \leq n$ for all n. For the Ising model J_n, let

$$S_n := \operatorname*{arg\,min}_{S\subseteq[n],|S|\leq t_n} \sqrt{t_n}\, \mathbb{E}_{(i,j)\sim\binom{[n]}{2}}[|\mathrm{Cov}(X_i, X_j|X_S)|],$$

and let κ_n denote the minimum value i.e. the value of the objective corresponding to S_n. By repeating the first part of the proof of Theorem 15, we get

$$\begin{aligned}\mathcal{F}_n &\leq \mathbb{E}_{x_{S_n}}\left[\sum_{ij}(J_n)_{ij}\mathbb{E}_\mu[X_iX_j|X_{S_n} = x_{S_n}]] + H_{\nu_{x_{S_n}}}(X)\right] + t_n \\ &\leq \sum_{i,j}(J_n)_{i,j}[\mathrm{Cov}(X_i, X_j|X_{S_n})] + \mathbb{E}_{x_{S_n}}\left[\sum_{i,j}(J_n)_{i,j}\mathbb{E}_{\nu_{x_{S_n}}}[X_iX_j] + H_{\nu_{x_{S_n}}}(X)\right] + t_n.\end{aligned}$$

As opposed to the proof of Theorem 15 where we used the Cauchy-Schwarz inequality, here we simply estimate the first term by

$$\sum_{i,j}(J_n)_{i,j}[|\mathrm{Cov}(X_i, X_j|X_{S_n})|] \leq 2\binom{n}{2}\frac{\kappa_n\|J_n\|_\infty}{\sqrt{t_n}}.$$

Finally, set

$$t_n = \min\left\{\frac{n^{4/3}\kappa_n^{2/3}\|J_n\|_\infty^{2/3}}{\sqrt[3]{4}}, n\right\};$$

note that $t_n < n$ for all sufficiently large n by assumption, along with the fact that $\limsup_{n\to\infty}\kappa_n \le \kappa_*$. It follows that for all n sufficiently large,

$$\mathcal{F}_n - \mathcal{F}_n^* \le \frac{3}{\sqrt[3]{4}} n^{4/3}\|J_n\|_\infty^{2/3}\kappa_n^{2/3};$$

dividing both sides by $n^{4/3}\|J_n\|_\infty^{2/3}$, taking the lim sup as $n \to \infty$, and using $\limsup_{n\to\infty}\kappa_n \le \kappa_*$ yields the desired conclusion. □

To complete the proof of Theorem 17, we will exhibit a sequence of Ising models J_n for which $\limsup_{n\to\infty} n\|J_n\|_\infty^2$ is finite and $\limsup_{n\to\infty}(\mathcal{F}_n - \mathcal{F}_n^*)/\left(n^{4/3}\|J_n\|_\infty^{2/3}\right)$ is positive. Specifically, we will show that this is true for a 'typical' growing sequence of the Rademacher SK-spin glass. First, we need the following lemma, which formalizes that the naive mean-field approximation fails in the SK model; in the physics literature, this was already argued in the original paper of Thouless, Anderson and Palmer [181] (and their argument is similar to the proof of the following Lemma, so it is essentially rigorous).

Lemma 19. *Fix $\beta \in [0, 1/2)$. Let $\mathcal{F}_n(\beta)$ denote the (random) free energy of the SK spin glass on n vertices with parameter β, and let $\mathcal{F}_n^*(\beta)$ denote its variational free energy. Then,*

$$\mathcal{F}_n(\beta) - \mathcal{F}_n^*(\beta) \ge n\beta^2/4 - o(n)$$

asymptotically almost surely (a.a.s) i.e. with probability going to 1 as $n \to \infty$. This holds under the same universality regime as Theorem 22.

Proof. We prove this by calculating $\mathcal{F}_n(\beta)$ and $\mathcal{F}_n^*(\beta)$. Since $\beta < 1$, we know from Theorem 22 that a.a.s.

$$\frac{\mathcal{F}_n(\beta)}{n} = \log 2 + \frac{\beta^2}{4} + o_n(1).$$

It remains to compute $\mathcal{F}_n^*(\beta)$. By definition,

$$\mathcal{F}_n^*(\beta) = \sup_{x\in[-1,1]^n} \left(\frac{\beta}{2} x^T J x + \sum_i H\left(\frac{1+x_i}{2}\right)\right).$$

We claim that a.a.s., this optimization problem is concave – indeed, direct calculation shows that for all $x \in [-1,1]$

$$\frac{d^2}{dx^2} H\left(\frac{1+x}{2}\right) \leq -1,$$

whereas the random matrix theory [6] of Wigner matrices shows

$$\|J\| \leq 2 + o_{n\to\infty}(1)$$

a.a.s. Since the Hessian of first term is J, this proves the claim since $0 \leq \beta < 1/2$.

Finally, since the gradient of the objective function

$$\frac{\beta}{2} x^T J x + \sum_i H\left(\frac{1+x_i}{2}\right)$$

clearly vanishes at the point $x_i = 0$ for all $i \in [n]$, it follows that this point is the global maximizer a.a.s, so that $\mathcal{F}_n^*(\beta) = n \log 2$ a.a.s. □

By combining the previous two lemmas, we can prove the following theorem which, in particular, implies Theorem 17.

Theorem 24. *Let κ_* be the universal constant defined at the start of this section.*

$$\kappa_* \geq \frac{\sqrt{27}}{16}.$$

Proof. From Lemma 19 applied to the Rademacher SK spin glass with parameter $\beta \in [0, 1/2)$ i.e. $(J_n)_{ij} = \pm\beta/\sqrt{n}$ independently with probability $1/2$, we obtain a sequence of Ising models $\{J_n\}_{n\geq 1}$ indexed by the number of vertices for which the following holds:

- $\|J_n\|_\infty = \frac{\beta}{\sqrt{n}}$ i.e. $\limsup_{n\to\infty} n\|J_n\|_\infty^2 = \beta^2$
- $\limsup_{n\to\infty} \frac{\mathcal{F}_n - \mathcal{F}_n^*}{n\beta^2} \geq \frac{1}{4}$ i.e. $\limsup_{n\to\infty} \frac{\mathcal{F}_n - \mathcal{F}_n^*}{n^{4/3}\|J_n\|_\infty^{2/3}} \geq \frac{\beta^{4/3}}{4}$.

In view of Lemma 18, there are two possibilities:

- $\kappa_*^2\beta^2 \geq 16$ for some $\beta \in [0, 1/2)$, so that $\kappa_* \geq 4$, or
- $\kappa_*^2\beta^2 < 16$ for all $\beta \in [0, 1/2]$, in which case we have

$$\frac{\beta^{4/3}}{4} \leq \frac{3}{\sqrt[3]{4}}\kappa_*^{2/3}$$

for all $\beta \in [0, 1/2)$, so that $\kappa_* \geq \sqrt{27}/16$.

□

3.5 Mean-field approximation for k-MRFs

In this section, we prove a much more general bound for mean-field approximation, extending our result Theorem 15 to order k Markov random fields (MRFs) over general finite alphabets. Our bound has only a mild dependence on the alphabet size q and is tight for every fixed k, q.

Definition 3. *An* order k Markov random field *(k-MRF) on n vertices over the finite alphabet Σ is a probability distribution on the space Σ^n of the form*

$$\Pr(X = x) = \frac{1}{Z}e^{f(x)+h(x)},$$

where the interaction term $f(x)$ can be written as a sum of hyperedge potentials on hyperedges of size k i.e.

$$f(x) = \sum_{E\subseteq[n],|E|=k} f_E(x_E),$$

and the external field $h(x)$ *is the sum of the external fields at each vertex i.e.*

$$h(x) = \sum_{i=1}^{n} h_i(x_i).$$

In analogy with the Ising model case, we will denote $\sup_{x_E} |f_E(x_E)|$ by $\|f_E\|_\infty$ and $\sum_{E\subseteq[n],|E|=k} \|f_E\|_\infty^2$ by $\|J\|_F^2$. The exact same proof as the Ising case gives the following variational principle for the free energy $\mathcal{F} := \log Z$:

$$\mathcal{F} = \sup_\mu \left[\mathbb{E}_\mu[f(x) + h(x)] + H(\mu)\right], \tag{3.6}$$

where the supremum ranges over all probability distributions on Σ^n. By restricting the variational problem to product distributions over Σ^n, we obtain the variational free energy $\mathcal{F}^*$ as before.

Theorem 25. *For any k-MRF on n vertices over an alphabet of size q,*

$$\mathcal{F} - \mathcal{F}^* \le 3\left(\frac{k\log q}{\sqrt{k!}} n^{k/2} \|J\|_F\right)^{2/3}.$$

The proof of this theorem is essentially the same as that of Theorem 15 with appropriate modifications. We will need the following simple lemma.

Lemma 20. *Let μ and ν are two probability distributions on the same space Ω. Then for any function $f\colon \Omega \to \mathbb{R}$ such that $|f(X)| \le M$ a.s. under both μ and ν, we have*

$$|\mathbb{E}_{X\sim\mu}[f(X)] - \mathbb{E}_{Y\sim\nu}[f(Y)]| \le 2M\mathbf{TV}(\mu,\nu).$$

Proof. By a standard characterization of $\mathbf{TV}$, we can couple X and Y so that $\Pr(X \neq Y) = \mathbf{TV}(\mu,\nu)$. Since $|f(X) - f(Y)| \le 2M$ a.s, we are done. □

Proof of Theorem 25. Let $\epsilon > 0$ be some parameter which will be optimized later. We begin by applying Theorem 21 with $\ell = 1/(\epsilon^2 \log q)$; let S be the resulting set of size at most ℓ. Let μ denote the Boltzmann distribution. For each assignment $x_S \in \Sigma^{|S|}$ to the variables in S, let ν_{x_S} denote the product measure on Σ^n for which $\mathbb{E}_{\nu_{x_S}}[X_i] = \mathbb{E}[X_i | X_S = x_S]$. Then, using the variational principle, the same computation as in the binary Ising model case shows that

$$\mathcal{F} \le \mathbb{E}_{x_S}\left[\mathbb{E}_\mu[f(X)|X_S = x_S] + \mathbb{E}_{\nu_{x_S}}[h(X)] + H_{\nu_{x_S}}(X)\right] + \ell \log q.$$

As before, we decompose the first term as

$$\mathbb{E}_{x_S}\left[\mathbb{E}_\mu[f(X)|X_S = x_S]\right] = \mathbb{E}_{x_S}\left[\mathbb{E}_{\nu_{x_S}}[f(X)] + \left[\mathbb{E}_\mu[f(X)|X_S = x_S] - \mathbb{E}_{\nu_{x_S}}[f(X)]\right]\right].$$

Since $f(X) = \sum_{E\in\binom{[n]}{k}} f(X_E)$, it follows by Lemma 20 that

$$\mathbb{E}_{x_S}\left|\mathbb{E}_\mu[f(X)|X_S = x_S] - \mathbb{E}_{\nu_{x_S}}[f(X)]\right| \le 2\binom{n}{k}\mathbb{E}_{x_S}\mathbb{E}_{E\sim\binom{[n]}{k}}\left[\|f_E\|_\infty \mathbf{TV}\left((\mu|X_S = x_S)|_{X_E}, \nu_{x_S}|_{X_E}\right)\right].$$

By the Cauchy-Schwarz inequality, the right hand side is bounded by

$$2\sqrt{\binom{n}{k}}\|J\|_F\sqrt{\mathbb{E}_{x_S}\mathbb{E}_{E\sim\binom{[n]}{k}}\mathbf{TV}^2\left((\mu|X_S = x_S)|_{X_E}, \nu_{x_S}|_{X_E}\right)},$$

whereas by Pinsker's inequality and the choice of S, we have

$$\begin{aligned}\sqrt{\mathbb{E}_{x_S}\mathbb{E}_{E\sim\binom{[n]}{k}}\mathbf{TV}^2\left((\mu|X_S = x_S)|_{X_E}, \nu_{x_S}|_{X_E}\right)} &\le \sqrt{\mathbb{E}_{E\sim\binom{[n]}{k}} C\left((\mu|X_S)|_{X_E}, \nu_{X_S}|_{X_E}\right)} \\ &\le \frac{k\sqrt{\log q}}{\sqrt{\ell}}.\end{aligned}$$

To summarize, there exists some x_S such that the associated product distribution $\nu := \nu_{x_S}$ satisfies

$$\mathcal{F} \le \mathbb{E}_\nu[f(x) + h(x)] + H(\nu) + 2k\epsilon\sqrt{\binom{n}{k}}\|J\|_F \log q + \frac{1}{\epsilon^2}. \tag{3.7}$$

Using $\binom{n}{k} \le n^k/k!$ and optimizing the value of ϵ completes the proof. □

3.5.1 Tightness of Theorem 25

In our formulation, there is a natural way to lift a k-MRF to an ℓ-MRF for any $k \le \ell$ by the following averaging procedure. Given a k-MRF specified by the collection

$(f_E)_{E\in\binom{[n]}{k}}$, we define the collection of functions $(g_F)_{F\in\binom{[n]}{\ell}}$ by

$$g_F(x_F) := \frac{1}{\binom{n-k}{\ell-k}} \sum_{E\subset F, |E|=k} f_E(x_E).$$

This scaling is chosen so that for any x,

$$\sum_{F\in\binom{[n]}{\ell}} g_F(x_F) = \sum_{F\in\binom{[n]}{\ell}} \frac{1}{\binom{n-k}{\ell-k}} \sum_{E\subset F, |E|=k} f_E(x_E) = \sum_{E\in\binom{[n]}{k}} f_E(x_E).$$

Hence, both the k-MRF and the ℓ-MRF correspond to the same distribution over Σ^n, and thus have the same mean-field error. On the other hand, it follows from the triangle inequality that

$$\sum_{F\in\binom{[n]}{\ell}} \|g_F(x_F)\|_\infty^2 \le \left(\frac{\binom{\ell}{k}}{\binom{n-k}{\ell-k}}\right)^2 \sum_{F\in\binom{[n]}{\ell}} \sum_{|E|=k, E\subset F} \|f_E(x_E)\|_\infty^2 = \frac{\binom{\ell}{k}^2}{\binom{n-k}{\ell-k}} \sum_{E\in\binom{[n]}{k}} \|f_E(x_E)\|_\infty^2$$

In particular, denoting $\sum_{F\in\binom{[n]}{\ell}} \|g_F(x_F)\|_\infty^2$ by $\|J_\ell\|_F^2$ and $\sum_{E\in\binom{[n]}{k}} \|f_E(x_E)\|_\infty^2$ by $\|J_k\|_F^2$, we see that for k and ℓ fixed,

$$\|J_\ell\|_F^{2/3} \le \frac{C_{k,\ell}}{n^{\ell-k}} \|J_k\|_F^2,$$

so that

$$n^{\ell/3}\|J_\ell\|_F^{2/3} \le C_{k,\ell} n^{k/3} \|J_k\|_F^{2/3}.$$

Therefore by lifting any of the tight examples for Theorem 15, we get a corresponding tightness result for k-MRFs:

Theorem 26. *For fixed k and q, Theorem 25 is tight up to constants. In other words, there exists an absolute constant $c_{k,q} > 0$ such that for infinitely many k-MRFs on an alphabet of size q,*

$$\mathcal{F} - \mathcal{F}^* \ge c_{k,q} \left(n^{k/2}\|J\|_F\right)^{2/3}.$$

Remark 4. *This tightness guarantee for mean-field also shows that Theorem 21 is tight up to constants for any fixed k. No more general form of Conjecture 1 was*

given for higher-order models, but combining the lifting result with the construction from Theorem 17 gives an analogous tightness result in terms of average TV-distance between product and joint distributions, ruling out improved bounds.

3.6 Algorithmic results: proof of Theorem 16

We now show how to go from the proof of our bounds on the quality of mean-field approximation to concrete algorithms; this is a relatively straightforward application of the Sherali-Adams relaxation. The only serious difficulty is to find a good proxy for the entropy that is suitable for use with pseudo-distributions; this was solved in [159] by introducing the following *pseudo-entropy functional* for level $(r+1)$ pseudo-distributions:

$$\tilde{H}_r(\mu) = \min_{S:|S|\leq r} \left[H(X_S) + \sum_i H(X_i|X_S) \right]. \tag{3.8}$$

By the chain rule for entropy, we see that for any r and for any true probability distribution μ, $H(\mu) \leq \tilde{H}_r(\mu)$. Moreover, essentially the standard proof of the concavity of entropy shows that for any r, $\tilde{H}_r(\mu)$ is a concave function of the pseudo-distribution μ (Lemma 8 of [159]). Then, we can write the Sherali-Adams relaxation to Eq. (3.6) as

$$\mathcal{F}_{SA,r+k} := \max_{\mu \in SA_{r+k}} \tilde{E}[f(X) + h(X)] + \tilde{H}_r(\mu). \tag{3.9}$$

Note that by considering the Boltzmann distribution μ in the above optimization problem, and using that $H(\mu) \leq \tilde{H}_r(\mu)$, it follows that $\mathcal{F}_{SA,r+k} \geq \mathcal{F}$.

Combining this relaxation with correlation rounding gives Algorithm SA-MeanField for finding good mean-field solutions.

Remark 5. *Instead of searching over all $S \subseteq [n]$ with $|S| \leq r$, we may greedily select S vertex by vertex, stopping when the average total correlation $\mathbb{E}_E[C(X_E|X_S)]$ satisfies the guarantee of Theorem 21. That this works follows from a slightly modified analysis of correlation rounding.*

Theorem 27. *Let $H(p)$ denote the entropy of* $\mathrm{Ber}(p)$. *We have the following running time and performance guarantees for Algorithm* SA-MeanField.

Algorithm 1 SA-MeanField

1. Find a pseudo-distribution μ maximizing Eq. (3.9) within ϵ additive error. This can be done efficiently using (for example) the ellipsoid method.
2. For every $S \subseteq [n]$ with $|S| \leq r$ and for every $x_S \in \Sigma^S$, let ν_{S,x_S} be the product distribution given by matching the first moments of μ conditioned on $X_S = x_S$.
3. Return the ν_{S,x_S} which maximizes $\mathbb{E}_\nu[f(X) + h(X)] + H(\nu)$.

---

1. *The running time is*

$$2^{O(nH((r+k)/n)+(r+k)\log q)} + \operatorname{poly}\log(1/\epsilon).$$

2. *The product distribution ν returned by the algorithm satisfies*

$$0 \leq \mathcal{F} - \mathcal{F}_\nu \leq \sqrt{\frac{4\log q}{r}}\frac{kn^{k/2}\|J\|_F}{\sqrt{k!}} + r\log q + \epsilon,$$

where

$$\mathcal{F}_\nu := \mathbb{E}_\nu[f(X) + h(X)] + H(\nu).$$

3. *We also have the following guarantee for the pseudo-distribution μ computed in the first step:*

$$0 \leq \mathcal{F}_{SA,r+k}(\mu) - \mathcal{F} \leq \sqrt{\frac{4\log q}{r}}\frac{kn^{k/2}\|J\|_F}{\sqrt{k!}} + \epsilon,$$

where

$$\mathcal{F}_{SA,r+k}(\mu) := \tilde{E}_\mu[f(X) + h(X)] + \tilde{H}_r(\mu).$$

Proof. The runtime is dominated by the first step, where we solve a convex program with at most $q^{r+k}\binom{n}{r+k}$ many variables and poly $\left(q^{r+k}\binom{n}{r+k}\right)$ many LP constraints. Therefore, by standard guarantees for the ellipsoid method [84] we can solve Eq. (3.9) within ϵ additive error in time poly $\left(q^{r+k}\binom{n}{r+k}, \log(1/\epsilon)\right)$. Using the standard bound

(which follows from sub-additivity of entropy)

$$\log \binom{n}{r+k} \leq nH\left(\frac{r+k}{n}\right),$$

this quantity is at most poly $\left(2^{O(nH((r+k)/n)+(r+k)\log q)}, \log(1/\epsilon)\right)$. Finally, we use the AM-GM inequality to separate the $2^{O(nH((r+k)/n)+(r+k)\log q)}$ term in the bound.

For 2., note that $0 \leq \mathcal{F} - \mathcal{F}_\nu$ follows from the Gibbs variational principle, so we only need to show the right inequality. We will deduce this from the stronger (since $\mathcal{F}_{SA,r+2} \geq \mathcal{F}$) statement

$$\mathcal{F}_{SA,r+2} - \mathcal{F}_\nu \leq \sqrt{\frac{4\log q}{r}} \frac{kn^{k/2}\|J\|_F}{\sqrt{k!}} + r\log q + \epsilon, \tag{3.10}$$

which itself follows from

$$\mathcal{F}_{SA,r+2}(\mu) - \mathcal{F}_\nu \leq \sqrt{\frac{4\log q}{r}} \frac{kn^{k/2}\|J\|_F}{\sqrt{k!}} + r\log q, \tag{3.11}$$

where μ is the $r+2$ pseudo-distribution returned in the first step. Now, note that Eq. (3.11) follows by exactly the same proof as for Theorem 25 (in particular, Eq. (3.7)) using the fact that an $r+k$ pseudo-distribution suffices to give the correlation rounding guarantee on sets of size at most r, and recalling that in Eq. (3.7), $\epsilon = 1/\sqrt{r\log q}$.

Finally, 3. follows from Eq. (3.11), noting additionally that we can avoiding losing the term $r\log q$ (equivalently, the term $1/\epsilon^2$ in Eq. (3.7)), if we round instead to the mixture of product distributions given by $\sum_{x_S} P(x_S)\nu_{S,x_S}$. □

In particular, we obtain the following more general and precise version of Theorem 16.

Corollary 4. *Fix k and q. If $\|J_n\|_F \leq c_{k,q} f(n) n^{3/2-k/2}$, where $f(n) \to 0$ as $n \to \infty$ and $c_{k,q} > 0$ is some constant depending only on k and q, then $\mathcal{F}_n$ can be approximated to within $\sqrt{f(n)}n$ additive error in (sub-exponential) time $2^{-O\left(n\sqrt{f(n)}\log f(n)\right)}$ by Algorithm* SA-MeanField. *Moreover, the algorithm outputs a product distribution*

achieving this approximation.

3.6.1 Faster algorithms using random subsampling

Until now, the algorithms we considered have been deterministic. However, in dense instances there is a major advantage to using randomness: we can accurately estimate $\mathcal{F}$ by looking at a *vanishingly small portion* of the entire input instance. In [98] the following structural guarantee is given, relating the free energy of small random induced subgraphs to that of the original model: Fix a k-MRF on the vertex set $[n]$ with interaction functions $(f_E)_{E\in\binom{[n]}{k}}$, and denote its free energy by $\mathcal{F}$. Consider a random subset Q of $[n]$ of size $|Q| = s$. Consider also the k-MRF on the vertex set Q whose interaction functions are given by

$$\left(\frac{n^{k-1}f_E}{s^{k-1}}\right)_{E\in\binom{Q}{k}}.$$

We will denote the free energy of this k-MRF by $\mathcal{F}_Q$.

Theorem 28 (Theorem 4, [98])**.** *Let $\epsilon > 0$ and suppose $s \geq 10^6\omega$, where $\omega := k^7\log(1/\epsilon)/\epsilon^8$. Then, with probability at least 39/40:*

$$\left|\mathcal{F} - \frac{n}{s}\mathcal{F}_Q\right| \leq C_q k^4\epsilon\left(n^{k/2}\|J\|_F + \epsilon n^k\|J\|_\infty + \omega n/s\right),$$

where $\|J\|_\infty := \sup_E \|f_E\|_\infty$.

Note that for the (rescaled) sampled k-MRF, it follows from Markov's inequality that

$$\|J_Q\|_F^2 \leq 10\frac{n^{2k-2}\binom{s}{k}}{s^{2k-2}\binom{n}{k}}\|J\|_F^2 \leq 10e^k\left(\frac{n}{s}\right)^{k-2}\|J\|_F^2$$

with probability at least 9/10. Whenever this happens, Theorem 27 shows that we can estimate $n\mathcal{F}_Q/s$ to within additive error

$$\sqrt{\frac{40\log q}{r}}\frac{ke^{k/2}n^{k/2}\|J\|_F}{\sqrt{k!}} + \frac{n\varepsilon}{s} \leq 10\sqrt{\frac{\log q}{r}}n^{k/2}\|J\|_F + \frac{n\varepsilon}{s}$$

in time $2^{O(sH((r+k)/s)+(r+k)\log q)} + \operatorname{poly}\log(1/\varepsilon)$. Taking $r = 1/(\epsilon^2 \log q)$ and $\varepsilon = \epsilon$, it follows that with probability at least $7/8$, we can find an estimate $\hat{\mathcal{F}}$ to $\mathcal{F}$ in *constant* time $2^{O_{k,q}\left(\frac{1}{\epsilon^2}\log(\frac{1}{\epsilon})\right)}$ such that

$$\left|\mathcal{F} - \hat{\mathcal{F}}\right| \leq C_q k^4 \epsilon \left(n^{k/2}\|J\|_F + \epsilon n^k \|J\|_\infty + \omega n/s\right).$$

Given an error probability $\delta > 0$, by repeating the above procedure independently $O(\log(1/\delta))$ many times and returning the median estimate, the standard Chernoff bound allows us to obtain the following.

Theorem 29. *Let $\delta, \epsilon > 0$ and suppose $s \geq 10^6\omega$, where $\omega := k^7 \log(1/\epsilon)/\epsilon^8$. Then, the above algorithm runs in time $2^{O_{k,q}\left(\frac{1}{\epsilon^2}\log(\frac{1}{\epsilon})\right)}\log(1/\delta)$ and returns an estimate $\hat{\mathcal{F}}$ such that:*

$$\left|\mathcal{F} - \hat{\mathcal{F}}\right| \leq C_q k^4 \epsilon \left(n^{k/2}\|J\|_F + \epsilon n^k \|J\|_\infty + \omega n/s\right)$$

with probability at least $1 - \delta$.

3.6.2 Algorithmic tightness under Gap-ETH

It's natural to ask if the tradeoff between graph density (more precisely, $\|J\|_F$) and runtime in our algorithm is optimal. It turns out that under a variant of the *Exponential Time Hypothesis*, this is indeed true. The variant we need is the following conjecture known as ETHA or Gap-ETH [132]:

Conjecture 2 (Gap-ETH)**.** *There exist constants $\epsilon, c > 0$ such that no algorithm running in time $O(2^{cn})$ can distinguish between a satisfiable 3-SAT formula and a 3-SAT formula with at most $1 - \epsilon$ fraction of satisfiable clauses. Here, n denotes the number of clauses.*

One of the motivations for this conjecture is that under the ordinary ETH, the quasilinear-length PCP of Dinur [55] shows that there exists some $\epsilon > 0$ such that no algorithm running in time $\Omega(2^{n/\operatorname{poly}\log(n)})$ can distinguish between a satisfiable 3-SAT formula and one with at most $1 - \epsilon$ fraction of satisfiable clauses; if this PCP were

of linear-length, then one could deduce Gap-ETH from ETH. Under Gap-ETH, one immediately finds that $\|J\|_F^2 = o(n)$ is the tight regime for approximating $\mathcal{F}/n$ with sub-exponential time algorithms.

Proposition 1. *Under Gap-ETH, the following holds for some $\epsilon > 0$:*

1. *There exist a constant $c > 0$ and an infinite family of graphs with $\Theta(n)$ many edges on which it takes time at least 2^{cn} to approximate MAX-CUT within multiplicative error $(1 - \epsilon)$.*

2. *There exist a constant $c > 0$ and an infinite family of Ising models with $\|J\|_F^2 = \Theta_\epsilon(n)$ on which it takes time at least 2^{cn} to approximate $\mathcal{F}$ within additive error ϵn.*

Proof. 1. This follows directly from the statement of Gap-ETH and the existence of an L-reduction from MAX-3SAT to MAX-CUT [150].

2. This follows from (1) by defining the corresponding anti-ferromagnetic Ising model and sufficiently high inverse temperature β, which gives an approximation guarantee for MAX-CUT as in Eq. (3.3). □

Remark 6. *Complexity-theoretic bounds straightforwardly imply lower bounds on the number of Sherali-Adams rounds needed; for example Proposition 1 implies that for these graphs $\Omega(n)$ rounds of Sherali-Adams are needed to approximate MAX-CUT; if, on the contrary, only $o(n)$ rounds sufficed, then solving the LP would give a $2^{nH(o(n)/n)} = 2^{o(n)}$ time algorithm (see Theorem 27).*

We can further apply reductions from [65] to get additional tightness results; they originally stated their results under the assumption of ETH, but the same reductions can be applied from Gap-ETH as well and give the following cleaner results.

Theorem 30 ([65])**.** *Under Gap-ETH, there is some $\epsilon > 0$ for which the following holds.*

1. *Consider an arbitrary sequence d_n with $d_n = o(n)$. Then there does not exist any algorithm which approximates MAX-CUT within multiplicative error $(1-\epsilon)$ in time $2^{o(n/d_n)}$ on all graphs of average degree at least d_n.*

2. *There exist a constant $c > 0$ and an infinite family of k-SAT instances with $\Theta_k(n^{k-1})$ many clauses (all of which are distinct) on which it takes time at least 2^{cn} to approximate MAX-k-SAT within multiplicative error $(1-\epsilon)$.*

As with Proposition 1, these translate immediately to lower bounds for computing partition functions by picking a sufficiently large inverse temperature β:

Corollary 5. *Under Gap-ETH, there is some $\epsilon > 0$ such that*

1. *Fix any sequence $d_n = o(n)$. There is no algorithm which computes $\mathcal{F}$ within additive ϵn error in time $2^{o(d_n)}$ on Ising models where $\|J\|_F^2 \leq d_n$.*

2. *For any fixed $k \geq 2$, there exist a constant $c > 0$ and an infinite family of binary k-MRFs with $\|J\|_F = \Theta_k(n^{3/2-k/2})$ on which it takes time at least 2^{cn} to approximate $\mathcal{F}$ within ϵn additive error.*

Proof. (1) follows directly from Theorem 30 using the same reduction as in Proposition 1. A slight generalization of this argument also shows (2): consider $\epsilon > 0$ and a family of k-SAT instances on n variables and $m_n = \Theta_k(n^{k-1})$ (distinct) clauses as in part (2) of Theorem 30. For the reduction, we start from the k-SAT instance with n variables and m distinct clauses, and define for each $E \in \binom{[n]}{k}$

$$f_E(x_E) := \frac{\beta n}{m} \#\{\text{clauses depending only on the variables in } E \text{ which are satisfied by } x_E\},$$

where β is a sufficiently large constant (depending on ϵ) to be specified later. Hence,

$$\|J\|_F^2 := \sum_E \|f_E\|_\infty^2 \leq \frac{\beta^2 n^2}{m} 2^{2k}$$

since there are at most 2^k distinct clauses supported on x_E and at most m subsets E which support a clause. Therefore, if we assume that (2) is false, then for any $c > 0$, we can compute the free energy of this model within additive error n in time at most 2^{cn} as long as

$$\frac{\beta^2 n^2}{m} 2^{2k} = \Theta_k(n^{3-k}),$$

which is true since $m = \Theta_k(n^{k-1})$ by assumption. On the other hand, since

$$\sum_E f_E(x_E) = \frac{\beta n}{m} \#\{\text{satisfied clauses for assignment } x\},$$

and since there is at least one assignment x for which the number of clauses satisfied is at least $m(1 - 2^{-k})$, it follows that if we take $\beta = 1/4\epsilon$, then an n-additive approximation for the partition function gives an ϵn-additive approximation for the k-SAT instances (by returning the approximation to the partition function multiplied by $m/n\beta$), thereby contradicting part (2) of Theorem 30.

□

3.7 Conclusion

We presented a unified perspective on two major variational approaches to calculating the free energy that hitherto seemed completely disparate: mean-field approximations and convex relaxations. This view has both analytic benefits (we derived bounds on the quality of mean-field approximations) and algorithmic benefits (we derived algorithms for approximating the free energy up to the intractability limit).

We conclude with several open problems, and discuss some recent related developments which occurred after the original publication of this work:

1. As mentioned earlier, there is a straightforward example showing that up to a constant, the exponent $\frac{2}{3}$ is optimal in Theorem 15 for the natural univariate quantity $(n\|J\|_F)$. However, this example does not rule out other bounds of the form $O(n^{1-\alpha}\|J\|_F^{2\alpha})$ for $\alpha \in [0, 1]$. As there is always a trivial bound $O(n)$ for the mean-field approximation (consider the optimal point-mass distribution), we may assume that $\|J\|_F = o(n^{1/2})$ and ask about the supremum of all α such that an upper bound of this form holds. The Curie-Weiss model at critical temperature shows that we cannot take α to be 0 without introducing additional logarithmic factors in the upper bound. Progress in this direction has (since the original publication of this work) been made in [10, 58] though the precise

answer is not known.

2. It's possible that the fRSB phase of the SK spin glass is more difficult to correlation-round than the RS phase. Is one of these spin glass models *extremal*, in the sense that they can be used to get the optimal value of κ_*? If not, what do the extremal distributions look like?

3. How many rounds do convex hierarchies (Sherali-Adams, Sum-of-Squares) need to correctly estimate the value of the free energy and ground state of the SK spin glass? (By computing the ground state, we mean to drop the entropy and just consider the MAX-QP problem.) Are $\Omega(n)$ rounds required? The question of optimizing the zero-temperature SK model was previously asked by Andrea Montanari and (since the original publication of this work) the recent work [73] shows that a super-constant number of rounds are needed for SOS to solve that problem (see also previous works cited there). The work [11] gave initial evidence based on low-degree polynomials that a nearly-linear number of rounds may be needed.

3.8 Appendix: Proof of Theorem 21

We will make use of the following information theoretic notion:

Definition 4. *The* multivariate mutual information *of a collection of random variables $X_1, \ldots, X_n$ is defined to be*

$$I(X_1; \cdots; X_n) = \sum_{m=1}^{n} (-1)^{m-1} \sum_{S \subset \binom{n}{m}} H(X_S).$$

Note that when $n = 2$, this corresponds to the usual notion of mutual information between two random variables. We may also define the *conditional multivariate mutual information* by using the conditional entropy in the above equation; note that

the chain rule for entropy shows immediately that

$$I(X_1;\cdots;X_n) = I(X_1;\cdots;X_{n-1}) - I(X_1;\cdots;X_{n-1}|X_n).$$

We will deduce Theorem 21 from the following lemma, which is slightly stronger. Our statement and proof correct two errors found in [132, 199]: missing sign terms in the relation between $C(X_S)$ and $I(X_S)$, and use of an invalid version of identity Eq. (3.12) below which sums over tuples instead of sets.

Lemma 21. *Let $X_1, \ldots, X_n$ be a collection of $\{\pm 1\}$-valued random variables. Then, for any $k, \ell \in [n]$, there exists some $t \leq \ell$ such that:*

$$\mathbb{E}_{S\sim\binom{V}{t}}\mathbb{E}_{F\sim\binom{V-S}{k}}[C(X_F|X_S)] \leq \frac{k^2\log(2)}{\ell}.$$

Proof. We begin by showing that

$$\mathbb{E}_{F\sim\binom{V}{k}}[C(X_F|X_S)] = \sum_{r=2}^{k}\binom{k}{r}(-1)^r\mathbb{E}_{R\sim\binom{V}{r}}[I(X_R|X_S)]. \tag{3.12}$$

For simplicity, we will prove the unconditional version of this identity. The same proof gives the conditional version as well. We start by noting that:

$$C(X_1;\cdots;X_n) = \sum_{R\subset[n],|R|\geq 2}(-1)^{|R|}I(X_R).$$

Therefore,

$$\begin{aligned}\sum_{F\subseteq\binom{V}{k}} C(X_F) &= \sum_{S\subseteq\binom{V}{k}}\sum_{R\subseteq F,|R|\geq 2}(-1)^{|R|}I(X_R)\\ &= \sum_{r=2}^{k}\sum_{R\subseteq\binom{V}{r}}\binom{|V|-r}{|V|-k}(-1)^r I(X_R),\end{aligned}$$

and dividing both sides by $\binom{|V|}{k}$ gives:

$$\begin{aligned}\mathbb{E}_{F\sim\binom{V}{k}}[C(X_F)] &= \sum_{r=2}^{k}\binom{|V|-r}{k-r}\binom{|V|}{r}\binom{|V|}{k}^{-1}(-1)^r\mathbb{E}_{R\sim\binom{V}{r}}[I(X_R)] \\ &= \sum_{r=2}^{k}\binom{k}{r}(-1)^r\mathbb{E}_{R\sim\binom{V}{r}}[I(X_R)],\end{aligned}$$

as desired.

Next, we consider the key quantity:

$$Q := \sum_{t=0}^{\ell}\mathbb{E}_{S\sim\binom{V}{t}}\mathbb{E}_{F\sim\binom{V-S}{k}}[C(X_F|X_S)] = \sum_{r=2}^{k}\binom{k}{r}(-1)^r\sum_{t=0}^{\ell}\mathbb{E}_{S\sim\binom{V}{t}}\mathbb{E}_{R\sim\binom{V-S}{r}}[I(X_R|X_S)],$$

where the second equality follows from Eq. (3.12). By the chain rule for mutual information, we have the telescoping sum:

$$\begin{aligned}\sum_{t=0}^{\ell}\mathbb{E}_{S\sim\binom{V}{t}}\mathbb{E}_{R\sim\binom{V-S}{r}}[I(X_R|X_S)] &= \sum_{t=0}^{\ell}\left(\mathbb{E}_{S\sim\binom{V}{t}}\mathbb{E}_{E\sim\binom{V-S}{r-1}}[I(X_E|X_S)] - \mathbb{E}_{S\sim\binom{V}{t+1}}\mathbb{E}_{E\sim\binom{V-S}{r-1}}[I(X_E|X_S)]\right) \\ &= \mathbb{E}_{E\sim\binom{V}{r-1}}[I(X_E)] - \mathbb{E}_{S\sim\binom{V}{\ell+1}}\mathbb{E}_{E\sim\binom{V-S}{r-1}}[I(X_E|X_S)],\end{aligned}$$

so that

$$\begin{aligned}Q &= \sum_{r=2}^{k}\binom{k}{r}(-1)^r\left(\mathbb{E}_{E\sim\binom{V}{r-1}}[I(X_E)] - \mathbb{E}_{S\sim\binom{V}{\ell+1}}\mathbb{E}_{E\sim\binom{V-S}{r-1}}[I(X_E|X_S)]\right) \\ &\leq \binom{k}{2}\mathbb{E}_{i\sim V}[H(X_i)] + \sum_{r=3}^{k}\binom{k}{r}(-1)^r\left(\mathbb{E}_{E\sim\binom{V}{r-1}}[I(X_E)] - \mathbb{E}_{S\sim\binom{V}{\ell+1}}\mathbb{E}_{E\sim\binom{V-S}{r-1}}[I(X_E|X_S)]\right),\end{aligned}$$

where in the second line, we have separated out the $r = 2$ term, and dropped the nonpositive term $-\binom{k}{2}\mathbb{E}_{S\sim\binom{V}{\ell+1}}\mathbb{E}_{i\sim V-S}[H(X_i|X_S)]$.

Now, recall that

$$\binom{k}{r} = \binom{k-1}{r-1} + \binom{k-2}{r-1} + \cdots + \binom{r-1}{r-1}.$$

Hence,

$$
\begin{aligned}
Q &\leq \binom{k}{2}\mathbb{E}_{i\sim V}[H(X_i)] - \sum_{d=2}^{k-1}\sum_{r=3}^{d+1}(-1)^{r-1}\binom{d}{r-1}\left(\mathbb{E}_{E\sim\binom{V}{r-1}}[I(X_E)] - \mathbb{E}_{S\sim\binom{V}{\ell+1}}\mathbb{E}_{E\sim\binom{V-S}{r-1}}[I(X_E|X_S)]\right)\\
&= \binom{k}{2}\mathbb{E}_{i\sim V}[H(X_i)] - \sum_{d=2}^{k-1}\left(\mathbb{E}_{F\sim\binom{V}{d}}[C(X_F)] - \mathbb{E}_{S\sim\binom{V}{\ell+1}}\mathbb{E}_{F\sim\binom{V-S}{d}}[C(X_F|X_S)]\right)\\
&\leq \binom{k}{2}\mathbb{E}_{i\sim V}[H(X_i)] + \sum_{d=2}^{k-1}\mathbb{E}_{S\sim\binom{V}{\ell+1}}\mathbb{E}_{F\sim\binom{V-S}{d}}[C(X_F)]\\
&\leq \binom{k}{2}\mathbb{E}_{i\sim V}[H(X_i)] + \sum_{d=2}^{k-1}\mathbb{E}_{S\sim\binom{V}{\ell+1}}\mathbb{E}_{F\sim\binom{V-S}{d}}\left[\sum_{i\in F}H(X_i)\right]\\
&\leq \left(\binom{k}{2} + \sum_{d=2}^{k-1}d\right)\mathbb{E}_{i\sim V}[H(X_i)]\\
&\leq k^2\log(2),
\end{aligned}
$$

where we have used Eq. (3.12) in the second line. Recalling the definition of Q, we see that there exists some $t \in \{0, 1, \ldots, \ell\}$ such that

$$\mathbb{E}_{S\sim\binom{V}{t}}\mathbb{E}_{F\sim\binom{V-S}{k}}[C(X_F|X_S)] \leq \frac{k^2\log(2)}{\ell}.$$

□

In order to deduce Theorem 21 from this lemma, we need the following two simple properties of the total correlation.

- For any $F, S \subseteq [n]$, $C(X_F|X_S) = C(X_{F\cap S^c}|X_S)$. This follows since by the chain rule for entropy

$$
\begin{aligned}
C(X_F|X_S) &= \sum_{j\in F}H(X_j|X_S) - H(X_F|X_S)\\
&= \sum_{j\in F\cap S^c}H(X_j|X_S) - H(X_{F\cap S}|X_S) - H(X_{F\cap S^c}|X_S)\\
&= \sum_{j\in F\cap S^c}H(X_j|X_S) - H(X_{F\cap S^c}|X_S)\\
&= C(X_{F\cap S^c}|X_S).
\end{aligned}
$$

- For any $S \subseteq [n]$ and $F \subseteq E \subseteq [n]$, $C(X_F|X_S) \leq C(X_E|X_S)$. Indeed, by the chain rule for entropy, and since conditioning decreases entropy

$$\begin{aligned} C(X_E|X_S) &= \sum_{i \in E} H(X_i|X_S) - H(X_E|X_S) \\ &= \left[\sum_{i \in F} H(X_i|X_S) - H(X_F|X_S)\right] + \left[\sum_{i \in E \setminus F} H(X_i|X_S) - H(X_{E \setminus F}|X_{S \cup F})\right] \\ &\geq C(X_F|X_S) + C(X_{E \setminus F|X_{S \cup F}}). \end{aligned}$$

Proof of Theorem 21. Fix an arbitrary $S \in \binom{V}{t}$. We will show that

$$\mathbb{E}_{F \sim \binom{V}{k}}[C(X_F|X_S)] \leq \mathbb{E}_{E \sim \binom{V-S}{k}}[C(X_E|X_S)], \tag{3.13}$$

which combined with Lemma 21 proves the claim. To prove Eq. (3.13), consider a coupling where we first sample $F \sim \binom{V}{k}$ and then choose E uniformly at random from those subsets $T \in \binom{V-S}{k}$ for which $F \cap S^c \subset T$. Then by symmetry, the marginal law on E is uniform on $\binom{V-S}{k}$. Under this coupling, using the above two properties of the total correlation, we have

$$C(X_F|X_S) = C(X_{F \cap S^C}|X_S) \leq C(X_E|X_S);$$

taking the expectation over F and E proves Eq. (3.13), and hence the result. □

Chapter 4

Landscape Analysis of Naive Mean-field Approximation in Ferromagnetic Models

In this chapter, we continue the discussion of the naive mean-field free energy introduced in Chapter 3. Here we analyze the behavior of the natural first-order method for computing the naive mean-field free energy in ferromagnetic (a.k.a. attractive) models on arbitrary graphs — the same class of models considered in Chapter 2. The main result is that in ferromagnetic models, iterating the naive mean field equations from all-ones initialization provably solves the variational problem, even though it is nonconvex. Iterating the equations is the standard heuristic for solving such variational problems in general; we note that this problem can also provably be solved in polynomial time using submodular optimization, see e.g. [115] for this and more general results. We also note that a well-studied optimization method related to the TAP free energy is called "approximate message passing" (AMP) and it applies to quite different situations from the ones considered here (e.g. to the SK model and similar dense random models); see [43] and references within. A longer discussion of the above, as well as a related result for belief propagation which builds upon the work of [52], is given in [111].

Definition 13. *An Ising model is* ferromagnetic *(with consistent field) if $J_{ij} \geq 0$ for all i and $h_i \geq 0$ for every i. (We also can allow $h_i \leq 0$ for all i, but this is equivalent after flipping signs.)*

As above, we recall the *(naive) mean-field approximation* to the free energy is given by maximizing the functional

$$\Phi_{MF}(x) := \frac{1}{2} x^T J x + h \cdot x + \sum_i H\left(Ber\left(\frac{1+x_i}{2}\right)\right) \tag{4.1}$$

where $H(Ber(p)) = -p \log p - (1-p)\log(1-p)$ is the entropy of a Bernoulli random variable. By considering the first-order optimality conditions for (4.1), one arises at the *mean-field equations*

$$x = \tanh^{\otimes n}(J \cdot x + h) \tag{4.2}$$

where $\tanh^{\otimes n}$ denotes entry-wise application of tanh. The *mean-field iteration* is the natural iterative algorithm which starts with some x_0 and applies (4.2) iteratively to search for a fixed point; it is a natural variant of gradient descent for this problem.

Theorem 31. *Fix an arbitrary ferromagnetic Ising model parameterized by J, h and let x^* be a global maximizer of Φ_{MF}. Initializing with $x^{(0)} = \vec{1}$ and defining $x^{(1)}, x^{(2)}, \ldots$ by iterating the mean-field equations, we have that*[1] *for every $t \geq 1$,*

$$0 \leq \Phi_{MF}(x^*) - \Phi_{MF}(x^{(t)}) \leq \min\left\{\frac{\|J\|_1 + \|h\|_1}{t}, 2\left(\frac{\|J\|_1 + \|h\|_1}{t}\right)^{4/3}\right\}.$$

This result cannot hold for arbitrary Ising models, as even approximating the mean-field free energy is NP-hard in general Ising models with anti-ferromagnetic interactions [97].

[1] In this theorem and throughout, we use the notation $\|J\|_1, \|J\|_\infty$ to refer to the corresponding ℓ_1, ℓ_∞ norms of J when viewed as a vector of entries.

4.1 Convergence of Mean-Field Iteration

In this section, we give the proof of Theorem 31 by analyzing the mean-field iteration. Organizationally, we split this theorem into two (corresponding to the two seperate bounds implied by the min): we prove the first bound in the theorem as Theorem 32 and the second $O(1/t^{4/3})$ bound as Theorem 33.

4.1.1 Main convergence bound

In this section we prove the first ($O(1/t)$) bound appearing in Theorem 31, the bound which is better for small t; we consider this to be the more significant bound because it gives a meaningful convergence result even when $t = O(1)$. A key observation in the proof is that the functional Φ_{MF} is actually concave on a certain subset of the space of product distributions, and that the iteration stays in this region because the iteration is monotone w.r.t. the partial order structure; this allows us to show progress at each step.

For the analysis of mean-field iteration, it will be very helpful to split the updates up into two steps:

$$y^{(t+1)} := Jx^{(t)} + h$$
$$x^{(t+1)} := \tanh^{\otimes n}(y^{(t+1)}).$$

Lemma 22. *A global maximizer of Φ_{MF} is in $[0,1]^n$.*

Proof. For any x, if $|x|$ denotes the coordinate wise absolute value then we observe $\Phi_{MF}(x) \leq \Phi_{MF}(|x|)$ since J, h are entrywise nonnegative and the entropy term is preserved. Therefore if x is a global maximizer then so is $|x|$, and by compactness of $[-1,1]^n$ there exists at least one global maximizer. □

Lemma 23. *There exists at most one critical point of Φ_{MF} in $(0,1]^n$.*

Proof. Suppose there exist two critical points y and z. Recall that being a critical point is equivalent to solving the mean-field equation $y = \tanh^{\otimes n}(Jy + h)$. Consider

the line through y and z; this line intersects the boundary region $[0,1]^n \setminus (0,1]^n$ at some point; we parameterize the line as $x(t)$ so that $x(0)$ is on this boundary, i.e. $x(0)_i = 0$ for some i, $x(t_1) = y$ and $x(t_2) = z$. Without loss of generality we assume that $t_1 < t_2$. Now we consider the behavior of the function

$$g(t) := \tanh(J_i \cdot x(t) + h_i) - x(t)_i$$

on this line. Observe that by definition $g(0) = \tanh(J_i \cdot x(0) + h_i) - 0 \geq 0$ and $g(t_1) = 0$. It follows from strict concavity that $g(t_2) < 0$ since $t_2 > t_1$, so z cannot be a fixed point, which gives a contradiction. □

Based on these lemmas, we define x^* to be the global maximizer of Φ_{MF} in $[0,1]^n$. Define $S := \{x \in (0,1]^n : x_i \geq x_i^*\}$.

Lemma 24. *The mean-field free energy functional Φ_{MF} is concave on S.*

Proof. First we claim that Φ_{MF} is concave at x^*. If x^* is on the interior of $[0,1]^n$, then this follows from the second-order optimality condition. From the mean-field equations (first order optimality condition) we see that it's impossible that there are any coordinates such that $x_i^* = 1$, and that if the graph is connected and there is a single coordinate such that $x_i^* = 0$, that the entire vector $x^* = 0$. If $x^* = 0$, then the maximum eigenvalue of J must be 1, so the free energy functional is globally concave – otherwise, by the Perron-Frobenius theorem there exists a eigenvector of J with all nonnegative entries and with eigenvalue greater than 1, from which we see that $x^* = 0$ cannot be the global optimum.

Now, it is easy to see that Φ_{MF} is concave on all of S, becuase if $0 \leq x \leq y$ coordinate-wise then $\nabla^2 \Phi_{MF}(x) \succeq \nabla^2 \Phi_{MF}(y)$, which follows because

$$\nabla^2 \Phi_{MF}(x) - \nabla^2 \Phi_{MF}(y) = (1/4) \sum_i (H''((1+x)/2) - H''((1+y)/2)) e_i e_i^T \succeq 0.$$

since $H''((1+x)/2) = \frac{-2}{1-x^2}$. □

Theorem 32 (Main bound in Theorem 31)**.** *Suppose that $x_0 \in S$ and define $(x^{(t)}, y^{(t)})_{t=1}^{\infty}$ by iterating the mean-field equations. Then for every t, $x^{(t)} \in S$. Furthermore*

$$\Phi_{MF}(x^*) - \Phi_{MF}(x^{(t)}) \leq \frac{\|J\|_1 + \|h\|_1}{t}.$$

Proof. To show that $x^{(t)} \in S$, observe that the mean-field iteration is monotone: if $x \leq x'$, then $\tanh^{\otimes n}(Jx + h) \leq \tanh^{\otimes n}(Jx' + h)$. Therefore, because $x^* \leq x_0$ we see that $x^* = \tanh^{\otimes n}(Jx^* + h) \leq \tanh^{\otimes n}(Jx^{(0)} + h) = x^{(1)}$ and so on iteratively.

To prove the convergence bound, first note that $\frac{\partial}{\partial x_i}\Phi_{MF}(x) = J_i \cdot x + h_i - \tanh^{-1}(x_i)$ and then observe by Lemma 24 and concavity that

$$\begin{aligned}
\Phi_{MF}(x^*) - \Phi_{MF}(x^{(t)}) &\leq \langle \nabla \Phi_{MF}(x^{(t)}), x^* - x_t \rangle \\
&\leq \|\nabla \Phi_{MF}(x^{(t)})\|_1 \\
&= \sum_i |\tanh^{-1}(x_i^{(t)}) - (Jx^{(t)} + h)_i| = \sum_i y_i^{(t)} - y_i^{(t+1)}
\end{aligned}$$

where the second inequality was by Hölder's inequality and $\|x^* - x^{(t)}\|_\infty \leq 1$, and the last equality follows from the definition of $y^{(t)}$ and because $y^{(t+1)} \leq y^{(t)}$ coordinate-wise. We can also see that $\Phi_{MF}(x^{(t)})$ is a monotonically increasing function of t by considering the path between $x^{(t)}$ and $x^{(t+1)}$ which updates one coordinate at a time, as the gradient always has non-positive entries along this path. Therefore if we sum over t we find that

$$\Phi_{MF}(x^*) - \Phi_{MF}(x^{(T)}) \leq \frac{1}{T}\sum_{t=1}^{T}(\Phi_{MF}(x^*) - \Phi_{MF}(x^{(t)})) \leq \frac{1}{T}\sum_{i=1}^{n}(y_i^{(1)} - y_i^{(T+1)}) \leq \frac{\|J\|_1 + \|h\|_1}{T}$$

since $y_i^{(T+1)} \geq 0$ and $y_i^{(1)} \leq \sum_j J_{ij} + h_i \leq \|J_i\|_1 + h_i$. □

The following simple example shows that the above result is not too far from optimal, in the sense that an asymptotic rate of $o(1/t^2)$ is impossible. We take advantage of the fact that when the model is completely symmetrical, the behavior of the update can be reduced to a 1-dimensional recursion, which is a standard trick (see e.g. [138, 151]).

Example 5. *Consider any d-regular graph with no external field and edge weight* $\beta = 1/d$*, which corresponds to the naive mean field prediction for the critical temperature. By symmetry, analyzing the mean field iteration reduces to the 1d recursion* $x \mapsto \tanh(x)$ *which behaves like* $x \mapsto x - x^3/3$ *near the fixed point* $x = 0$*. Solving this recurrence, we see that* x *converges to* 0 *at rate* $\Theta(1/\sqrt{t})$*. In terms of* x*, the estimated mean field free energy is* $(n/2)x^2 + nH(\frac{1+x}{2})$*, so by expanding we see that the estimated free energy converges at a* $\Theta(1/t^2)$ *rate in this example.*

4.1.2 Faster Asymptotic Rate

The above theorem and lower bound leave a gap between $O(1/t)$ and $\Omega(1/t^2)$ for the asymptotic rate of the mean-field iteration. This section is devoted to showing that for large t, we can obtain an improved asymptotic rate of $O(1/t^{4/3})$ for the mean-field iteration using a slightly more involved variant of the argument from the previous section. The key insight is that we can obtain some control of $\|x - x^*\|_\infty$ by consider the behavior of higher-order terms when expanding around x^*, and this can be used to get better bounds on the convergence in objective.

Lemma 25. *Suppose that* $x \in S$*. Then*

$$\|\nabla \Phi_{MF}(x)\|_1 \geq \frac{\|x - x^*\|_4^4}{\|x - x^*\|_\infty}$$

where x^* *is as above, the global maximizer of* Φ_{MF} *in* $[0,1]^n$*.*

Proof. Recall that

$$\nabla \Phi_{MF}(x) = Jx + h - \sum_i \tanh^{-1}(x_i) e_i.$$

Since x^* is a critical point and local maximum, so $\nabla \Phi_{MF}(x^*) = 0$ and $\nabla^2 \Phi_{MF}(x^*) \preceq 0$, then using that $\frac{d^2}{dx^2} \tanh^{-1}(x) = \frac{2x}{(1-x^2)^2}$, we see that by applying the fundamental

theorem of calculus twice that

$$\begin{aligned}\nabla\Phi_{MF}(x) &= J(x-x^*) - \sum_i e_i(\tanh^{-1}(x_i) - \tanh^{-1}(x_i^*)) \\ &= \nabla^2\Phi_{MF}(x^*)(x-x^*) - \sum_i e_i \int_{x_i^*}^{x_i}\int_{x_i^*}^{z} \frac{2y}{(1-y^2)^2} dydz\end{aligned}$$

and so

$$\begin{aligned}\langle x^*-x, \nabla\Phi_{MF}(x)\rangle &\geq \sum_i (x_i - x_i^*)\int_{x_i^*}^{x_i}\int_{x_i^*}^{z} \frac{2y}{(1-y^2)^2} dydz \\ &\geq \sum_i (x_i - x_i^*)\int_{x_i^*}^{x_i}\int_{x_i^*}^{z} 2y dydz \\ &= \sum_i (x_i - x_i^*)(x_i^3/3 - (x_i^*)^3/3 - (x_i - x_i^*)(x_i^*)^2) \\ &= \sum_i (x_i - x_i^*)^2(x_i^2 + x_i x_i^*) \geq \sum_i (x_i - x_i^*)^4\end{aligned}$$

where in the last inequality we used $x_i \geq x_i^* \geq 0$. Finally the result follows combining the above with $\langle x^*-x, \nabla\Phi_{MF}(x)\rangle \leq \|x^*-x\|_\infty \|\nabla\Phi_{MF}\|_1$ by Hölder's inequality. □

Theorem 33 (Second bound in Theorem 31)**.** *Suppose that $x_0 \in S$ and define $(x_t, y_t)_{t=1}^\infty$ by iterating the mean-field equations. Then for every t, $x_t \in S$. Furthermore for any $t \geq 1$,*

$$\|x_t - x^*\|_\infty^3 \leq \frac{\|J\|_1 + \|h\|_1}{t}$$

and

$$\Phi_{MF}(x^*) - \Phi_{MF}(x_{2t}) \leq \left(\frac{\|J\|_1 + \|h\|_1}{t}\right)^{4/3}.$$

Proof. From Lemma 25 we see that

$$\|x - x^*\|_\infty^3 \leq \frac{\|x-x^*\|_4^4}{\|x-x^*\|_\infty} \leq \|\nabla\Phi_{MF}(x)\|_1$$

and so as in the proof of Theorem 32 we see that for any T,

$$\|x_T - x^*\|_\infty^3 \leq \frac{1}{T}\sum_{t=1}^T \|x_t - x^*\|_\infty^3 \leq \frac{1}{T}\sum_{t=1}^T \|\nabla\Phi_{MF}(x_t)\|_1 = \frac{1}{T}\sum_{i=1}^n (y_{1,i} - y_{T+1,i}) = \frac{\|J\|_1 + \|h\|_1}{T}.$$

Therefore for any $t' > T$ we see by convexity and Hölder's inequality

$$\begin{aligned}\Phi_{MF}(x^*) - \Phi_{MF}(x_t) \leq \langle \nabla \Phi_{MF}(x_t), x^* - x_t \rangle &\leq \left(\frac{\|J\|_1 + \|h\|_1}{T} \right)^{1/3} \|\nabla \Phi_{MF}(x_t)\|_1 \\ &= \left(\frac{\|J\|_1 + \|h\|_1}{T} \right)^{1/3} \sum_i |\tanh^{-1}(x_{t,i}) - (Jx_t + h)_i| \\ &= \left(\frac{\|J\|_1 + \|h\|_1}{T} \right)^{1/3} \sum_i (y_{t,i} - y_{t+1,i})\end{aligned}$$

and summing this over $t' = T+1$ to $2T$ and telescoping we see that

$$\begin{aligned}\Phi_{MF}(x^*) - \Phi_{MF}(x_{2T}) \leq \frac{1}{T} \sum_{t'=T+1}^{2T} (\Phi_{MF}(x^*) - \Phi_{MF}(x_{t'})) &\leq \left(\frac{\|J\|_1 + \|h\|_1}{T} \right)^{1/3} \sum_i (y_{T,i} - y_{2T,i}) \\ &\leq \left(\frac{\|J\|_1 + \|h\|_1}{T} \right)^{4/3}\end{aligned}$$

which proves the result. □

Chapter 5

Learning GGMs without a Well-Conditioning Assumption

5.1 Introduction

A Gaussian Graphical Model (GGM) in n dimensions is a probability distribution with density

$$p_X(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\big(-(x-\mu)^T \Sigma^{-1} (x-\mu)/2 \big)$$

where μ is the mean and Σ is the covariance matrix. In other words, it is just a multivariate Gaussian. What makes the Gaussian interesting as a graphical model is that its *Markov Random Field* structure, in the sense of Chapter 1, is encoded entirely by the support of the *precision matrix* $\Theta = \Sigma^{-1}$, and so it is very useful to recover this graphical structure.

GGMs have wide-ranging applications in machine learning and the natural and social sciences where they are one of the most popular ways to model statistical relationships between observed variables — see e.g. [196, 137, 95, 187] among many other references. It is important to note that in most of the settings in which GGMs are applied, the number of observed samples is small compared to the dimensionality of the data. This means that in practice, it is only possible to learn the GGM

in a meaningful sense under some sort of sparsity assumption. We will make the assumption that the rows and columns of Θ are d-sparse – i.e., the case where the dependency graph G has maximum degree at most d.

From a theoretical standpoint, there is vast literature on learning sparse GGMs under various assumptions. Many approaches focus on *sparsistency* – where the goal is to learn the sparsity pattern of Θ assuming some sort of lower bound on the strength of non-zero interactions. This is a natural objective because once the sparsity pattern is known, estimating the entries of Θ is straightforward (e.g. one can use ordinary least squares); because of this, the problems of learning GGMs and sparse linear regression are very closely related. A popular approach to learning GGMs is the Graphical Lasso[1] [67] which solves the following convex program:

$$\max_{\Theta \succ 0} \left(\log\det(\Theta) - \langle \widehat{\Sigma}, \Theta \rangle - \lambda \|\Theta\|_1 \right)$$

where $\widehat{\Sigma}$ is the empirical covariance matrix and $\|\Theta\|_1$ is the ℓ_1 norm of the matrix as a vector.

It is known that if Θ satisfies various conditions, which typically include an assumption similar to or stronger than the restricted eigenvalue (RE) condition (a condition which, in particular, lower bounds the smallest eigenvalue of any $2d \times 2d$ principal submatrix of Σ) then Graphical Lasso and related ℓ_1 methods can succeed in recovering the graph structure (see e.g. [135, 203]). For the Graphical Lasso itself, under some incoherence assumptions on the precision matrix (stronger than RE), it has been shown [157] that the sparsity pattern of the precision matrix can be accurately recovered from $O((1/\alpha^2)d^2 \log(n))$ samples; here α is an *incoherence parameter* and we are omitting the dependence on some additional terms. We emphasize that this is only the best known theoretical guarantee — the performance in real life often seems better than this pessimistic bound.

Another popular approach to learning GGMs is the CLIME estimator which solves

[1]We note that [67] did not introduce this objective (see discussion there), but rather an *optimization procedure* used to maximize it, and Graphical Lasso technically refers to this specific optimization procedure.

the following linear program:

$$\min_{\Theta} \|\Theta\|_1 \text{ s.t. } \|\widehat{\Sigma}\Theta - I\|_\infty \leq \lambda$$

The analysis of CLIME assumes a bound M on the maximum ℓ_1-norm of any row of the inverse covariance (given that the X_i's are standardized to unit variance). This is also a type of condition number assumption, although with respect to a different geometry than RE: more precisely, since

$$M = \max_{\|u\|_\infty \leq 1} \|\Theta u\|_\infty$$

it can be thought of as the condition number of Σ when viewed as an operator mapping $\ell_\infty \to \ell_\infty$; this can be smaller than the normal Euclidean condition number. CLIME succeeds at structure recovery when given

$$m \gtrsim CM^4 \log n$$

samples (here for simplicity we are assuming the entries Θ_{ij} are either zero or bounded away from zero by an absolute constant c so that $M = \Omega(d)$).

While these works show that sparse GGMs can be estimated when the number of samples is logarithmic in the dimension, there is an important caveat in their guarantees. They all need to assume that Θ is well-conditioned, and differ mainly in the strength of their assumption: roughly speaking, one of the stronger assumptions[2] used in this literature is that Θ is well-conditioned in the usual sense, and the weakest is the $\ell_\infty \to \ell_\infty$ condition number bound assumed by CLIME. This is often heuristically justified by the belief that a small condition number is information-theoretically required for structure recovery to be possible. However, and as we will discuss later, recent works have pointed out that this is actually not the case — the correct information-theoretic condition is *significantly weaker* than even the assumption which CLIME

[2]Indeed, there are even stronger assumptions such as quantitative versions of *faithfulness* which we do not discuss but are needed to prove the correctness of the popular PC algorithm [102].

makes. Indeed, the fact that bounded condition number is not the right assumption for structure recovery is hinted at by the fact that it does not behave nicely under benign operations like rescaling individual variables. In the high-dimensional setting, bounded condition number can be a somewhat strong condition: in particular, this assumption is violated by simple and natural models (e.g. a graphical model on a path such as a time series), where these bounds turn out to be polynomial in the dimension.

In this paper, we study some fundamental classes of GGMs and show how to learn them efficiently in the low-sample regime under the correct information-theoretic assumption, even when they are ill-conditioned. We also complement our results with examples that break both previous algorithms and our own algorithms for learning general sparse GGMs. This leaves open the interesting question (raised in [140], and closely related to similar questions about sparse linear regression [201]) of whether some sparse GGMs may be computationally hard to learn with so few samples. Finally, we show experimentally that popular approaches, like the Graphical Lasso and CLIME, do in fact need a polynomial in n number of samples even in some relatively benign examples (and where our algorithm does succeed).

Our work was motivated by a recent paper of Misra, Vuffray and Lokhov [140] which studied the question of how many samples are needed *information-theoretically* to learn sparse GGMs in the ill-conditioned case. They required only the following natural non-degeneracy condition (which also appeared in [4, 192]): that for every i, j with $\Theta_{ij} \neq 0$, we have a lower bound on the *conditional partial correlation*[3] below:

$$\kappa \leq \frac{|\Theta_{ij}|}{\sqrt{\Theta_{ii}\Theta_{jj}}} = \frac{|\mathrm{Cov}(X_i, X_j \mid X_{\sim i,j})|}{\sqrt{\mathrm{Var}(X_i|X_{\sim i,j})\mathrm{Var}(X_j|X_{\sim i,j})}}.$$

Intuitively, this assumption means that if we have already observed all of the coordinates of X except for X_i and X_j, then the remaining randomness over X_i and X_j has a correlation coefficient of at least κ. This condition is the correct one because: (1)

[3]Here $X_{\sim i}$ (resp. $X_{\sim i,j}$) denotes the random vector formed by deleting coordinate i (resp. i, j) of X; please see Preliminaries for further details and formal definitions of conditional variances, covariances.

the absence of an edge in the GGM exactly corresponds to zero partial correlation in the above sense, (2) it has the correct symmetries — it is not affected by rescaling of any coordinate, and (3) it is the same condition which is needed in the classical, low-dimensional OLS regression t-test [104] to successfully reject the null hypothesis that the true coefficient of X_j is zero when regressing X_i off of $X_{\sim i}$.

Crucially, this assumption could be much weaker than any condition number bound, because it allows for the random variables to be strongly correlated. Here is a basic example (from [140]): suppose we have three Gaussians X_1, X_2 and X_3 where X_1 is heavily correlated with X_2. In this case, the condition number of Σ will explode as X_1 and X_2 become more correlated. Nevertheless, it remains possible to test if there is a κ-nondegenerate edge between X_1 and X_3, as long as we correctly adjust for the effect of X_2. In contrast, if we were unaware of the value of X_2, it would be very difficult to test for the same edge between X_1 and X_3, because the X_1 and X_2 edge contributes a very large amount of variance to X_1.

The work of [140] exhibited an algorithm achieving this requirement — more precisely, they showed that it is always possible to estimate the graph structure with

$$m \geq C \frac{d}{\kappa^2} \log n$$

samples *without requiring any additional assumptions*, clarifying that further condition number assumptions are indeed unnecessary. On the other hand, the result of [192] gives an information-theoretic lower bound[4] of $\Omega((1/\kappa^2) \log n)$ on the sample complexity for structure recovery. To summarize, the upper bound of [140] differs from the lower bound of [192] by exactly a factor of d (it is unknown what the optimal dependence is) and otherwise is optimal.

However, the algorithm of [140] runs in time $n^{O(d)}$, making it impossible to run except for small instances. This is because their algorithm is based on a reduction to a sequence of sparse linear regression problems that can all be ill-conditioned. It is

[4] A subtle point arises when interpreting this bound, because d and κ are closely related quantities (see e.g. Lemma 31 below). In the lower bound constructions of [192] they have $d = O(1/\kappa)$ and the term dominating their bound depends only on κ.

believed that such problems exhibit wide gaps between what is possible information theoretically and what is possible efficiently. For instance, it is known that the general *sparse linear regression* problem under fixed design is **NP**-hard[5] [146, 201]. Misra et al. solve the sparse linear regression problems using exhaustive search over d-size neighborhoods (hence the $n^{O(d)}$ time). This leads to the main question we study:

Can we get efficient and practical algorithms for learning GGMs (run-time $\ll n^{o(d)}$) in some natural, but still ill-conditioned, cases?

5.2 Results and Technical Overview

We show that for some popular and widely-used classes of GGMs—*attractive GGMs* and *walk-summable GGMs* — it is possible to achieve both logarithmic sample complexity (the truly high-dimensional setting) and computational efficiency, even when Θ is ill-conditioned.

Attractive GGMs

First we study the class of attractive GGMs, in which the off-diagonal entries of Θ are non-positive. In terms of the correlation structure, this means that all partial correlations are nonnegative. There are several practical motivations for studying attractive GGMs: in phylogenetic applications, observed variables are often positively dependent because of shared ancestry [206]; in various copula models that are popular in finance, we posit a latent global market variable that also leads to positive dependence [145]; see also [193] for more discussion.

A well-studied special case (which essentially captures all attractive GGMs — see Lemma 40) is the discrete Gaussian Free Field (GFF), in which case Θ is the generalized Laplacian associated to a weighted graph. This is a natural model because the Laplacian encourages "smoothness" with respect to the graph structure: see e.g. [167]; for this reason, the GFF is an important modeling tool in active and semi-supervised learning (see [205, 204, 128]); the GFF also arises in nature from a number

[5]For proper learning, where the algorithm is required to output a d-sparse estimator.

of diverse phenomena in random walks, statistical physics, and random surfaces [54, 66, 167].

In the GFF setting, Θ will be ill-conditioned, even in the weak $\ell_\infty \to \ell_\infty$ sense, whenever some pair of vertices have large *effective resistance* between them (e.g., paths, rectangular grids, etc.,); informally, it happens when the graph has many sparse cuts.

We show experimentally (in Appendix 5.11) that simple examples, like the union of a long path and some small cliques, do indeed foil the Graphical Lasso and other popular methods. Intuitively, this is because GFFs on a path exhibit long-range correlations that violate the assumptions used in current works — our examples show that the assumptions made in the literature are to some extent necessary for these algorithms. This analysis reveals a blind spot of the Graphical Lasso: It performs poorly in the presence of long dependency chains, which could lead to missing some important statistical relationships in applications.

We propose the following simple algorithm and show that it succeeds in learning the graph structure of attractive GGMs. This algorithm, called GreedyAndPrune, does the following to learn the neighborhood of node i:

1. Set $S = \emptyset$ and let $\nu > 0$ be a thresholding parameter.

2. (Greedy/OMP step) Repeat the following T times: set j to be the the minimizer of $\widehat{\mathrm{Var}}(X_i|X_S, X_j)$ and add j to S.

3. (Pruning step) For each $j \in S$: if $\widehat{\mathrm{Var}}(X_i|X_S) > (1-\nu)\widehat{\mathrm{Var}}(X_i|X_{S\setminus\{j\}})$, remove j from S.

4. Return S as the neighborhood of node i.

where $\widehat{\mathrm{Var}}$ indicates the variance is estimated from sample, using Ordinary Least Squares. A more detailed description of the algorithm is given in the Appendix. In the literature, this is called a *forward-backward method* [118].

Theorem 34 (Informal version of Theorem 40)**.** *Fix a κ-nondegenerate attractive GGM. The* GreedyAndPrune *algorithm runs in polynomial time and returns the true neighborhood of every node i with high probability with $m \geq C(d/\kappa^2)\log(1/\kappa)\log(n)$ samples, where C is a universal constant.*

Our algorithm matches the sample complexity of the previous best (inefficient) algorithms for this setting [4, 140] and obtains, up to log factors, the optimal dependence on κ for fixed d.

Analysis for Attractive GGMs The main intuition behind the algorithm and the crux of our analysis is the following: For attractive GGMs the conditional variance of a variable X_i when we condition on a set X_S is a monotonically decreasing and supermodular function of S. This fact was previously observed in the GFF setting (independently in [128, 129]) with relatively involved proofs; we give a new, short proof of this fact using just basic linear algebra. Other works such as [30, 48] have considered supermodularity in somewhat related regression settings, but with important differences (see Further Discussion).

Given the supermodularity result, we next need to address the issue that we don't have access to actual conditional variances, but only their empirical estimates. To achieve the efficient sample complexity of Theorem 34 we carefully analyze the alignment between the true decrement of conditional variance in one step, $\mathrm{Var}(X_i|X_S) - \mathrm{Var}(X_i|X_{S\cup\{j\}})$ and the noisy empirical decrement $\widehat{\mathrm{Var}}(X_i|X_S) - \widehat{\mathrm{Var}}(X_i|X_{S\cup\{j\}})$. A subtle obstacle is that we need to control the differences $\widehat{\mathrm{Var}}(X_i|X_S) - \widehat{\mathrm{Var}}(X_i|X_{S\cup\{j\}})$ without assuming too much accuracy on the estimates $\widehat{\mathrm{Var}}(X_i|X_S)$ themselves. Fortunately, this can be shown using matrix concentration, combined with some tools from classical low-dimensional regression tests [104].

To complete the analysis, we need a new structural result for attractive GGMs which bounds the conditional variance after the first step of greedy, so that only a bounded number of iterations of greedy are required to learn a superset of the neighborhood. We prove this by reducing to the setting of discrete GFFs, where we can use an electrical argument based on effective resistances. Formally, we prove the

following new structural result for walk-summable GGMs:

Lemma 26 (Lemma 34 of the Appendix)**.** *Suppose that i is a node with $d \geq 1$ neighbors in an attractive or walk-summable GGM. Then there exists a neighbor j such that*

$$\mathrm{Var}(X_i|X_j) \leq \frac{4d}{\Theta_{ii}} = 4d \cdot \mathrm{Var}(X_i|X_{\sim i}).$$

Previous work on Learning Attractive GGMs. Some prior work on learning attractive GGMs have focused on the Maximum Likelihood Estimator (MLE). This was shown to exist and be unique using connections to total positivity in [170, 117], but we are not aware of any sample complexity guarantees in the context of structure learning. It also is likely broken by the same examples (see Section 5.11) as the Graphical Lasso (since the constrained MLE is the same as the Graphical Lasso with zero regularization and a non-negativity constraint). Finally, the recent work [193] studied adaptive estimators for learning GGMs, but only for the case where the model is well-conditioned.

Optimal Information-Theoretic Bounds. The previous literature leaves open the information-theoretically optimal sample complexity for learning attractive GGMs. We resolve this question: a simple estimator based on ℓ_0-constrained least squares, which we refer to as SearchAndValidate, achieves sample complexity matching the information-theoretic lower bounds of [192] (whose instances can easily be made attractive) up to constants:

Theorem 35 (Informal version of Theorem 44)**.** *In a κ-nondegenerate attractive GGM, as long as $m = \Omega((1/\kappa^2)\log(n))$, with high probability Algorithm* SearchAndValidate *returns the true neighborhood of every node i. This algorithm runs in time $O(n^{d+1})$.*

The results of [192] imply that $\Omega((1/\kappa^2)\log(n))$ samples are required even to distinguish the empty graph from a graph with a single κ-nondegenerate edge in an unknown location. This bound does not depend on d, which may appear surprising. This is possible because $d \leq 1/\kappa^2$ in κ-nondegenerate attractive GGMs — see

Lemma 31. We also give a version of the above result for general models with sample complexity $O(d\log(n)/\kappa^2)$ and time complexity $O(n^{d+1})$, giving a faster alternative to [140] with the same sample complexity guarantee.

Theorem 35 is proved by a careful analysis of the signal-vs-entropy tradeoff between choosing the correct support (which is best in expectation) and an incorrect support with k disagreements for each k. To do this we again need to study structural properties of the GGM; we establish something similar to a "margin condition" in empirical process theory [186]. Precisely analyzing the differences in empirical risk again builds upon some classical ideas in regression testing [104].

This result also identifies an important barrier to improving the information theoretic lower bound of [192], as their lower-bound instances can easily be made attractive. If this bound is not tight for general GGMs, it appears significantly new ideas will be needed to separate the sample complexity of learning attractive and non-attractive GGMs — they must rely upon the ability of negative correlations to create nontrivial cancellations.

Walk-Summable GGMs

While attractive GGMs are natural in some contexts, in others they are not. For example, in Genome Wide Association Schemes (GWASs), genes typically have inhibitory effects too. This leads us to another popular and well-studied class of GGMs: the *walk-summable* models. These were originally introduced by Maliutov, Johnson, and Willsky [131] to explain the convergence properties of Gaussian Belief Propagation observed in practice (see also [194]).

All attractive GGMs are walk-summable, as are other important classes of GGMs like *pairwise normalizable* and *non-frustrated* models [131]. A number of equivalent definitions are known for walk-summability. The following definition is perhaps the easiest to work with: Θ is walk-summable if making all off-diagonal entries of Θ negative preserves the fact that Θ is positive definite. Perhaps less well known, walk-summable models are exactly those GGMs with Symmetric Diagonally Dominant (SDD, see Preliminaries) precision matrices under a rescaling of the coordinates —

see e.g. [27, 160]. In the linear algebra literature [27], a walk-summable matrix Θ is referred to as a symmetric H-matrix with nonnegative diagonal.

Analysis for learning Walk-Summable GGMs. The analysis of learning walk-summable models is considerably different from the attractive case, because supermodularity (and even *weak supermodularity* [48]) of the conditional variance fails to hold – see Section 5.10.1. Regardless, we are still able to prove that GreedyAndPrune learns all walk-summable models with sample complexity that scales logarithmically with n. We also propose a variant HybridMB that achieves better sample complexity.

The key idea in this analysis is that a single greedy step can serve as a kind of sparse *weak preconditioner*, roughly in terms of the $\ell_\infty \to \ell_\infty$ geometry considered in CLIME. More precisely, we show that after a single step of greedy, the unknown sparse regression vector has small ℓ_1-norm (independent of n and scaling correctly with the noise level). This is shown in the proof of Theorem 49, based on effective resistance arguments relatd to Lemma 26. The ℓ_1-norm bound not only implies that greedy works, but also that appropriate invovations of ℓ_1-based methods (like the Lasso) can now be guaranteed to work. We emphasize that such bounds do not hold without our "weak preconditioning" step.

Concretely, we propose an algorithm called HybridMB based on this idea and show that it learns walk-summable GGMs without any condition number dependence. This algorithm does the following to learn the neighborhood of node i, where some technical details are left to the full algorithm description given in the Appendix:

1. (Greedy step) Set j to be the minimizer of $\widehat{\mathrm{Var}}(X_i|X_j)$.

2. (Lasso with implicit weak preconditioning) Solve for w, a in

$$\min_{w,a:\|w\|_1\le\lambda} \hat{\mathbb{E}}\left[\left(X_i - \sum_{k\notin\{i,j\}} w_k \frac{X_k}{\sqrt{\widehat{\mathrm{Var}}(X_k|X_j)}} - aX_j\right)^2\right].$$

We detail the selection of λ in the full version of the algorithm — see the

Appendix.

3. (Pruning step) We perform a pruning step similar to GreedyAndPrune to zero out some of the entries of w, and to test if j is an actual neighbor.

4. Return j (if it passed the test) and the remaining support of w as the neighborhood of i.

The analysis of HybridMB uses the aforementioned structural results for walk-summable models and a statistical analysis for the regression problem arising after the greedy step. The regression analysis is similar in spirit to the usual generalization bounds for ℓ_1-constrained regression but slightly more subtle. The key insight is that the output of the algorithm is the same if we replace X_k by $X_k - \mathbb{E}[X_k|X_j]$; this change of basis is unknown to the algorithm, but the analysis is much easier because X_j becomes independent of the other regressors.

Theorem 36 (Informal version of Theorem 50)**.** *Fix a walk-summable, κ-nondegenerate GGM. Algorithm* HybridMB *runs in polynomial time and returns the true neighborhood of every node i with high probability given $m \geq C(d/\kappa^4)\log(n)$ samples, where C is a universal constant.*

We can also prove a similar (but slightly weaker) guarantee for Algorithm GreedyAndPrune — see Theorem 51. For context, we note that prior to our work, Anandkumar, Tan, Huang and Willsky [4] gave an inefficient $n^{O(d)}$ time algorithm for learning walk-summable models with similar guarantees and requiring some additional assumptions.

The above structure learning result requires κ-nondegeneracy and sparsity of the entire model. However, it is proved using the following general result for sparse linear regression, which requires only a joint walk-summability assumption:

Theorem 37 (Informal version of Theorem 49)**.** *Suppose that $Y = w \cdot X + \xi$ where w is d-sparse, $\xi \sim N(0, \sigma^2)$ is independent of multivariate Gaussian r.v. $X \sim N(0, \Sigma)$, and suppose that the joint distribution of $(X_1, \ldots, X_n, Y)$ is a walk-summable GGM.*

Given m samples from this model, WS-Regression *runs in polynomial time and returns $\hat{w}$ such that*

$$\mathbb{E}[(w \cdot X - \hat{w} \cdot X)^2] = O(\sigma^2 \sqrt{d \log(n)/m})$$

with high probability.

Although this result gives a "slow rate" of $\sqrt{1/m}$, it is quite different from the usual slow rate guarantee for the Lasso. The latter guarantees an upper bound on the prediction error of the form $O(\sigma RW\sqrt{\log(n)/m} + RW\log(n)/m)$ where R is an ℓ_1 norm bound on w and W is an ℓ_∞ bound on X, see e.g. [158, 174]. To interpret this, we can rescale the problem so that $R, W = \Theta(1)$. Then Theorem 49 guarantees error on the order of the noise level σ^2 using $O(d\log(n))$ samples – in comparison, the standard slow rate result only guarantees error on the order of σ plus an additional term. This difference is the key to achieving structure recovery from $O(\log n)$ samples: σ can be orders of magnitude smaller compared to RW in our applications. Compared to ℓ_0-constrained least squares, which requires runtime $O(n^d)$, the above result is computationally efficient and still has the correct dependence on d and σ^2.

General Models. There do exist some well-conditioned GGMs which are not walk-summable. However, our analysis actually shows that our methods (GreedyAndPrune, HybridMB) also recover similar sample complexity bounds to [38] under their assumptions (the aforementioned $\ell_\infty \to \ell_\infty$ condition number bound) — see Theorem 52. Therefore, our results are a strict extension of the situation considered in prior work.

Non-Gaussian Models. It's well-known that many results for Gaussian Graphical Models can be generalized to other distributions in the following sense: if we can learn a GGM with precision matrix $\Theta = \Sigma^{-1}$, then the result will generally extend to estimating $\Theta = \Sigma^{-1}$ for X with sufficiently strong concentration assumptions. The reason is that for any result which depends only on the first two moments of X (i.e. any quantity definable in terms of Σ, μ), we can generalize it to such an X

by considering the Gaussian with matching first and second moments, and higher moments are generally needed only for concentration purposes.

We briefly note that the guarantees for our algorithms will also extend in this sense if, for any $w \in \mathbb{R}^n$, the sub-Gaussian constant of $w \cdot X$ is upper bounded by $C\text{Var}(w \cdot X)$ for a fixed constant C, as the needed concentration estimates generalize [188]. On the other hand, for non-Gaussian distributions the connection between Θ and conditional independence will not generally hold.

5.2.1 Further Discussion

GGMs vs Ising Models. There exist parallels but also surprisingly significant differences between learning GGMs and Ising models. For Ising models, Bresler [30] gave a greedy algorithm that builds a superset of the neighborhood around each node and then prunes to learn the true graph structure using $O(f(d)\log n)$ samples and under some relatively mild assumptions. This greedy algorithm is able to perform structure learning in Ising models even when they exhibit long range correlations, which was previously considered a difficult case to analyze. However in our setting, and unlike the previously described situation for Ising models, variables have real values and can have arbitrarily small or large variance. It turns out this changes the problem dramatically, as it means that the inter-node fluctuations in the random field (which contribute to $\text{Var}(X_i)$) may be orders of magnitude larger than the per-node fluctuations (corresponding to $\text{Var}(X_i|X_{\sim i})$). This is exactly the setting $\sigma \ll RW$ discussed in the context of sparse linear regression. Related problems can also arise in Ising models when the ℓ_1 norm is large, as in (for example) the Sherrington-Kirkpatrick model [107].

As a result of this difference, greedy methods fail to learn general GGMs from $O(\text{polylog}(n))$ samples (see Section 5.12), so any analysis of greedy methods must rely on structural results for a subclass of models. The same issue comes up when learning the model directly from ℓ_1-constrained regression guarantees as in [190, 109] — in fact, we will see in Section 5.11 that natural methods based only on ℓ_1 regularization fail even in some relatively simple attractive GGMs (where greedy works).

Sparse Linear Regression and Submodularity. As previously mentioned, Das and Kempe [48] studied the problem of sparse regression without assuming the restricted eigenvalue condition. While in sparse regression, in order to learn the parameters accurately (in additive error) some bound on the condition number is needed, they studied the problem of selecting a subset of columns that maximizes squared multiple correlation (a.k.a. minimizes mean squared error). They then gave approximation guarantees for greedy algorithms under an approximate submodularity condition; however, they did not address the natural *random design* setting where submodularity is assumed in the infinite-sample limit, but we need to analyze the behavior in a finite-sample setting (essentially, they studied this as a purely algorithmic problem).

Our algorithm for attractive models follows the same supermodularity-based strategy, but has no knowledge of the true model satisfying weak supermodularity except for the samples it sees. Therefore it requires a careful analysis of the interaction between the greedy iteration and noise. In the more general setting of walk-summable GGMs, we show the conditional variance does not satisfy an approximate supermodularity condition with any constant submodularity ratio. (See Remark 13.)

Some other Related Work on Sparse Linear Regression. In the literature on sparse regression, it is well known that the analyses of the Lasso which work well in a compresssed sensing style setting (i.e. with restricted eigenvalues, incoherent columns, etc.) is not always the correct tool to use when the coordinates of X (columns of the design matrix) are highly correlated — see e.g. [185, 88, 41]. For example, the work of Koltchinskii and Minsker [114] discusses this issue in the context of Brownian motion and other situations and develops general new guarantees for ℓ_1-penalized regression which apply under correlated design (as well as infinite dimensional settings). They consider the case where the response is a linear combination of well-separated measurements in time, which is incomparable to the situation we analyze. It would be interesting to see if the ideas used in Algorithms HybridMB and GreedyAndPrune can be used in some of these other settings.

5.3 Organization

Here we briefly outline the structure of the remainder of this section, which contains the proofs of the main Theorems as well as some simple simulations and experiments validating the theory. Each item below corresponds to a single section below.

1. Preliminaries: we explain some fundamental facts about GGMs and fix the notation we use throughout the rest of the paper.

2. Structural results for walk-summable models: In this section, we use the connection between walk-summability, SDD matrices, and electrical circuits to establish a number of new structural results about walk-summable GGMs that will be useful for learning them. As mentioned earlier, the fundamental fact we establish in this section which is needed in all of our algorithms is that a single step of a greedy method (Orthogonal Matching Pursuit) can serve as a "weak preconditioner" for sparse linear regression, in terms of ℓ_1/ℓ_∞ geometry. In particular, we establish the key Lemma 26 stated above.

3. Estimating changes in conditional variance: In this section, we recall the various facts we will need about ordinary least squares regression and prove a useful quantitative estimate for estimating changes in conditional variance.

4. Learning all attractive GGMs efficiently: we use further structural results about supermodularity in attractive GGMs and the results developed in the previous two sections to prove Theorem 34.

5. Information-theoretic optimal learning of attractive GGMs: In this section, we show how the result of the previous section can be improved as far as sample complexity if we are willing to sacrifice runtime, by giving a very precise analysis of a natural algorithm using ℓ_0-constrained squares, proving Theorem 35.

6. Hybrid ℓ_1-regression guarantees: In this section, in preparation for proving our results about learning general walk-summable models, we develop the needed statistical guarantees for a variant of the LASSO where a single coordinate

in the regression is left unregularized and also give an analysis of Orthogonal Matching Pursuit in essentially the same setup.

7. Regression and structure learning in walk-summable models: In this section, we first show that supermodularity fails in walk-summable models, even if we ask for supermodularity to only hold approximately. We then proceed to establish Theorem 37 for sparse linear regression in general walk-summable models and use this result to derive Theorem 36 for structure recovery in κ-nondegenerate walk-summable models.

8. Simulations and Experiments: In this section, we compare the methods proposed in this paper to those in a number of previous works on both simulated and real data. The simulations show that all previous methods indeed fail to achieve competitive sample complexity in simple settings where the precision matrix is not well-conditioned.

9. Some difficult examples: In this short final section, we give some examples which are not walk-summable and which break both the algorithms proposed previous to this paper and in this paper as well. We show that these examples are, however, not computationally hard to learn.

5.4 Preliminaries

In this section we set out some notation and basic facts about GGMs which will be used throughout.

Notations. Given a GGM with precision matrix Θ, d will always denote the maximum degree of the underlying graph. Thus, Θ has at most $d+1$ nonzero entries in each row. For a vector x and index i, $X_{\sim i} = ((X_j) : j \neq i)$. For a square matix $S \in \mathbb{R}^{k \times k}$ and $I \subseteq [k]$, S_I denotes the $I \times I$ principal submatrix of S. We will say a symmetric matrix M is SDD (Symmetric Diagonally Dominant) if its diagonal is nonnegative and for every row i, $M_{ii} \geq \sum_{j \neq i} |M_{ij}|$. We often use the notation $\hat{\mathbb{E}}$ to

denote the *empirical expectation*, i.e. expectation taken over the sample of data given to the algorithm.

We recall that conditioning on $X_i = x_i$ for any x_i yields a new GGM with the precision matrix having row i and column i deleted. In particular, the conditional precision matrix does not depend on the value of x_i chosen. Similarly, the value of the mean μ does not affect the covariance structure at all — so μ does not play an interesting role in the structure learning problem and we assume $\mu = 0$ without loss of generality in what follows; handling $\mu \neq 0$ just requires adding a constant term to every regression problem. We summarize the facts that we use the most below.

Fact 1 ([119])**.** *Let X be drawn from a mean 0 GGM with precision matrix Θ. Then, for any i, $X_i | X_{\sim i} = x_{\sim i}$ is distributed as $N(\langle w^{(i)}, x_{\sim i} \rangle, 1/\Theta_{ii})$ where $w^{(i)}$ is the vector with $w^{(i)}_j = -\Theta_{ij}/\Theta_{ii}$.*

Thus, if we fix an index i, then samples X from the GGM can be interpreted as a linear regression problem as $(X_{\sim i}, X_i)$ where $X_i = \langle w^{(i)}, X_i \rangle + N(0, 1/\Theta_{ii})$. This establishes the basic connection between learning GGMs and linear regression: if we can solve the above regression problem well, perhaps we can recover the non-zero entries of Θ from the coefficients. But as is well known in the literature, just fitting the coefficients using ordinary least squares is not sufficient (or necessarily possible) as we have very few samples.

By positive definiteness, we have $\Theta_{i,i} \geq 0$ and $\Theta_{i,i}\Theta_{j,j} - \Theta^2_{i,j} \geq 0$, or equivalently $0 \leq \frac{|\Theta_{i,j}|}{\sqrt{\Theta_{i,i}\Theta_{j,j}}} \leq 1$. To identify the graph we need the present edges to not be too weak. So it makes sense to assume (following the notation of [4, 140]) there is a $\kappa > 0$ such that

$$\kappa \leq \frac{|\Theta_{i,j}|}{\sqrt{\Theta_{i,i}\Theta_{j,j}}} \leq 1 \tag{5.1}$$

Definition 14 ([4, 140])**.** *We say a GGM is κ-nondegenerate if it satisfies* (5.1) *for all i, j such that $\Theta_{ij} \neq 0$.*

Conditional Variance. Conditional variances of the form $\mathrm{Var}(X_i | X_S)$ play a central role in all our algorithms. We first review the basic definition and some of their

properties.

Definition 15 (Conditional Variance)**.** *For X an arbitrary real-valued random variable and Y an arbitrary random variable or collection of random variables on the same probability space, let*[6]

$$\mathrm{Var}(X|Y) := \mathbb{E}[(X - \mathbb{E}[X|Y])^2].$$

By the Pythagorean Theorem, conditional variance obeys the *law of total variance* [23]:

$$\mathrm{Var}(X) = \mathrm{Var}(X|Y) + \mathrm{Var}(\mathbb{E}[X|Y]).$$

and more generally, $\mathrm{Var}(X|Y) = \mathrm{Var}(X|Y,Z) + \mathrm{Var}(\mathbb{E}[X|Y,Z]|Y)$. The last identity is also sometimes referred to as the law of total conditional variance.

The κ-nondegeneracy assumption implies a quantitative lower bound on conditional variances $\mathrm{Var}(X_i|X_S)$ when the conditioning set does not include all of i's neighbors.

Lemma 27. *Fix a node i in a κ-nondegenerate GGM, and let S be set of nodes not containing all neighbors of i. Then*

$$\mathrm{Var}(X_i|X_S) \geq \frac{1+\kappa^2}{\Theta_{ii}}$$

Proof. Let $j \notin S$ be a neighbor of i. By the law of total conditional variance, we have

$$\mathrm{Var}(X_i|X_S) = \mathrm{Var}(X_i|X_{\sim i}) + \mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}]|X_S) = \frac{1}{\Theta_{ii}} + \mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}]|X_S),$$

where in the last equality we used Fact 1. Thus, as $\mathbb{E}[f^2] \geq \mathrm{Var}(f)$, and the definition of κ-nondegeneracy

$$\mathrm{Var}(X_i|X_S) - \frac{1}{\Theta_{ii}} = \mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}]|X_S)$$

[6] In an alternate convention which we do not use, $\mathrm{Var}(X|Y)$ is defined to be the random variable $\mathbb{E}[(X - \mathbb{E}[X|Y])^2|Y]$ and our definition is the same as $\mathbb{E}\mathrm{Var}(X|Y)$.

$$= \mathbb{E}[(\mathbb{E}[X_i|X_{\sim i}] - \mathbb{E}[X_i|X_S])^2]$$
$$\geq \mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}] - \mathbb{E}[X_i|X_S]|X_{\sim j}) = \frac{\Theta_{ij}^2}{\Theta_{ii}^2\Theta_{jj}} \geq \frac{\kappa^2}{\Theta_{ii}}$$

where the last equality follows from Fact 1 and the last inequality is by the definition of κ. The Lemma follows by rearranging. □

The following basic fact about Gaussians will be useful:

Lemma 28. *If X and Y are jointly Gaussian random variables then $\mathbb{E}[X|Y] = \mathbb{E}[X] + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}(Y - \mathbb{E}[Y])$ and $\mathrm{Var}(X) - \mathrm{Var}(X|Y) = \frac{\mathrm{Cov}(X,Y)^2}{\mathrm{Var}(Y)}$.*

Proof. Because the random variables are jointly Gaussian, we know that $\mathbb{E}[X|Y]$ must be an affine function of Y. From $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$ and $\mathrm{Cov}(\mathbb{E}[X|Y], Y) = \mathrm{Cov}(X, Y)$ the coefficients are determined, proving the first formula. Then the second formula follows from the law of total variance, $\mathrm{Var}(X) - \mathrm{Var}(X|Y) = \mathrm{Var}(\mathbb{E}[X|Y])$. □

We will also use the following concentration inequality often. Recall that a χ^2-random variable with D degrees of freedom is a random variable with the law of $\sum_{i=1}^{D} Z_i^2$, where $Z_i \sim N(0, 1)$ are independent standard Gaussians.

Lemma 29 (Lemma 1, [116])**.** *Suppose U is χ^2-distributed with D degrees of freedom. Then $\Pr(U - D \geq 2\sqrt{D\log(1/\delta)} + 2\log(1/\delta)) \leq \delta$ and $\Pr(D - U \geq 2\sqrt{D\log(1/\delta)}) \leq \delta$. In particular, $U \leq 2D$ with probability at least $1 - \delta$ as long as $D \geq 8\log(1/\delta)$.*

5.5 Structural results for walk-summable models

5.5.1 Background: Walk-Summable Models are SDD after rescaling

Definition 16 ([131])**.** *A Gaussian Graphical Model with invertible precision matrix $\Theta \succ 0$ is* walk-summable *if $D - \overline{A} \succ 0$ where $\Theta = D - A$ decomposes Θ into diagonal and off-diagonal components, and $\overline{A}$ is the matrix with $\overline{A}_{ij} = |A_{ij}|$.*

It is well-known (and immediate) that the class of walk-summable matrices includes the class of SDD matrices. Indeed, the motivation for introducing walk-summable matrices was to generalize the notion of SDD matrices.

Definition 17. *A matrix M is* symmetric diagonally dominant *(SDD) if it is symmetric and $M_{ii} \geq \sum_{j:j\neq i} |M_{ij}|$ for every i.*

Perhaps less well-known, a natural converse holds: all walk-summable matrices are simply rescaled SDD matrices, where the rescaling is in the natural sense for a bilinear form. Furthermore, this rescaling is easy to find algorithmically (if we have access to Θ), requiring just a top eigenvector computation. This result can be found explicitly in [160]; it also appears in [27] and closely related results for M-matrices appear in [64].

Theorem 38 (Theorem 4.2 of [160])**.** *Suppose Θ is walk-summable. Then there exists a diagonal matrix D with positive entries such that $D\Theta D$ is an SDD matrix.*

Proof. We include the proof for completeness — it is the same as in [160].

First, we observe that we can reduce to the case $\mathrm{diag}(\Theta) = \vec{1}$ by replacing Θ by $D_1\Theta D_1$ where D_1 is the diagonal matrix with $(D_1)_{ii} = 1/\sqrt{\Theta_{ii}}$. Next, let $\overline{\Theta} = I - \overline{A}$ and note that when we write the decomposition $0 \prec \overline{\Theta} = I - \overline{A}$ that $\overline{A}$ has all nonnegative entries, so we can apply the Perron-Frobenius Theorem to find an eigenvector v with positive entries and eigenvalue $\lambda = \|\overline{A}\| < 1$. Now define $D_2 = \mathrm{diag}(v)$, and we claim that $D_2\Theta D_2$ is an SDD matrix. It suffices to check that $0 \leq D_2\overline{\Theta}D_2\vec{1} = D_2\overline{\Theta}v$ entry-wise, and because D_2 is diagonal with nonnegative entries it suffices to check that $\overline{\Theta}v \geq 0$. This follows as

$$\overline{\Theta}v = (I - \overline{A})v = (1 - \lambda)v \geq 0$$

entrywise. □

Example 6. *In Example 1 of [131] it was observed that the matrix*

$$\begin{bmatrix} 1 & -r & r & r \\ -r & 1 & r & 0 \\ r & r & 1 & r \\ r & 0 & r & 1 \end{bmatrix}$$

itself stops being SDD when $r > 1/3$, but remains walk-summable until a little past $r = 0.39$. When $r = 0.39$, the corresponding Perron-Frobenius eigenvector for $\overline{A}$ is roughly $(0.557, 0.435, 0.557, 0.435)$ and applying the rescaling from Theorem 38 we get

$$\begin{bmatrix} 0.310634 & -0.0945889 & 0.121147 & 0.0945889 \\ -0.0945889 & 0.189366 & 0.0945889 & 0. \\ 0.121147 & 0.0945889 & 0.310634 & 0.0945889 \\ 0.0945889 & 0. & 0.0945889 & 0.189366 \end{bmatrix}$$

which is an SDD matrix.

The SDD rescaling given by Theorem 38 will play a key role in our analysis. Conceptually, converting a walk-summable matrix to its SDD form is a way to take the extra degrees of freedom in the model specification (arbitraryness in the scaling of the X_i) and fix them in a way that is useful in the analysis – i.e. a gauge fixing. In particular, under the SDD rescaling there are meaningful relations between the different rows of Θ which fail to hold in general.

5.5.2 Background: SDD systems, Laplacians, and electrical flows

Definition 18. *A matrix L is a* generalized Laplacian *if it is SDD and for every $i \neq j$, $L_{ij} \leq 0$. We think of this graph theoretically as the Laplacian of the weighted graph with edge weights $-L_{ij}$ between distinct i and j and self loops of weight $L_{ii} - \sum_{j \neq i} |L_{ij}|$ at vertex i.*

We review the standard reduction between solving SDD systems and Laplacian systems. Suppose Θ is an SDD matrix. Then we can write $\Theta = L - P$ where L is a (generalized) Laplacian having positive entries on the diagonal and nonnegative entries off the diagonal, and P has negative off-diagonal entries and corresponds to the positive off-diagonal entries of Θ. Now we observe that

$$\begin{bmatrix} L & P \\ P & L \end{bmatrix} \begin{bmatrix} x \\ -x \end{bmatrix} = \begin{bmatrix} \Theta x \\ -\Theta x \end{bmatrix} \tag{5.2}$$

and the left matrix is itself a (generalized) Laplacian matrix on a weighted graph which we will refer to as the "lifted graph".

The inverse of a Laplacian has a natural interpretation in terms of electrical flows, where the edge weights are interpreted as conductances of resistors. In this case the self loops can be thought of as resistors connected directly to electrial ground. In the next Lemma we summarize the relevant facts about this interpretation, as can be found in e.g. [26]

Lemma 30. *Suppose that L is a (generalized) Laplacian matrix. Then if L^+ is the pseudo-inverse of L, and we define the* effective resistance $R_{\mathit{eff}}(i,j) := (e_i - e_j)^T L^+ (e_i - e_j)$ *then R_{eff} satisfies:*

- *(Nonnegativity) $R_{\mathit{eff}}(i,j) \geq 0$.*
- *(Monotonicity) $R_{\mathit{eff}}(i,j) \leq \frac{1}{|L_{ij}|}$, and more generally R_{eff} decreases when adding edges to the original adjacency matrix.*
- *(Triangle inequality) $R_{\mathit{eff}}(i,k) \leq R_{\mathit{eff}}(i,j) + R_{\mathit{eff}}(j,k)$ for any i,j,k.*

In the generalized Laplacian case, we can think of $\mathrm{Var}(X_i|X_S)$ as being the effective resistance from node i to ground when all of the nodes in S are connected by wires (without resistance) to ground.

5.5.3 Key structural results for Walk-Summable GGM

First we prove a fundamental fact about κ-nondegeneracy in walk-summable models, mentioned earlier: the maximum degree d always satisfies $d = O(1/\kappa^2)$ in κ-nondegenerate walk-summable models. This result is tight for star graphs.

Lemma 31. *In a κ-nondegenerate walk-summable GGM, the maximum degree of any node is at most $1/\kappa^2$.*

Proof. Rescale the coordinates so that the diagonal of Θ is all-1s, and reorder them so that X_1 corresponds to the node of maximum degree d with neighbors $2, \ldots, d+1$. Define $\overline{\Theta}$ to be the sign-flipped version of Θ such that all off-diagonal entries are negative; by the definition of walk-summability we know $\overline{\Theta}$ is still PSD. Let $v = (1, \kappa, \ldots, \kappa) \in \mathbb{R}^{d+1}$ and $S = \{1, \ldots, d+1\}$; then using that the off-diagonals are negative, κ-nondegeneracy we find that $\Theta_{d+1,d+1} v \leq (1 - d\kappa^2, 0, \ldots, 0)$ coordinate-wise, hence using $\overline{\Theta} \succeq 0$ we find

$$0 \leq v^T \Theta_{d+1,d+1} v \leq v^T (1 - d\kappa^2, 0, \ldots, 0) = 1 - d\kappa^2.$$

Rearranging we see that $d \leq 1/\kappa^2$. □

In the remainder of this subsection we prove some key structural results about walk-summable/SDD GGM using the SDD to Laplacian reduction and the electrical interpretation of the inverse Laplacian; these results will be crucial for analyzing the algorithms for both attractive and general walk-summable GGMs.

The following key Lemma, which shows that the variance between two adjacent random variables in the SDD GFF cannot differ by too much, will be crucial in the analysis of our algorithm in non-attractive models. Why is this useful? Informally, this is because for the greedy method to significantly reduce the variance of node i, at least one neighbor of i needs to provide a good "signal-to-noise ratio" for estimating X_i, and under the SDD scaling, this inequality shows that the neighbors do not have too much extra noise (compared to $|\Theta_{ij}|$ which roughly corresponds to the level of signal between nodes i and j).

Lemma 32. *Suppose that Θ is an invertible SDD matrix. Let $\Sigma = \Theta^{-1}$. If $\Theta_{ij} \neq 0$, then*

$$\Sigma_{ii} \leq 1/|\Theta_{ij}| + \Sigma_{jj}.$$

Proof. Let M be the generalized Laplacian matrix resulting from applying the SDD to Laplacian reduction from Σ, i.e. M is the left hand-side of (5.2). Let the standard basis for $\mathbb{R}^{2n}$ be denoted $e_1, \ldots, e_n, e'_1, \ldots, e'_n$. Observe from (5.2) that

$$\Sigma_{ii} = e_i^T \Theta^{-1} e_i = e_i^T M^+ (e_i - e'_i) = \frac{1}{2}(e_i - e'_i)^T M^+ (e_i - e'_i).$$

Let node label i be the node corresponding to e_i in the graph corresponding to M, and label i' be that corresponding to e'_i. Observe that in the graph corresponding to M, either i is adjacent to j and i' is adjacent to j', or i is adjacent to j' and i' is adjacent to j. Let $r = R_{\text{eff}}(i, j)$ in the first case and $r = R_{\text{eff}}(i, j')$ in the second case. By the triangle inequality (Lemma 30) and monotonicity of effective resistance (Lemma 30),

$$2\Sigma_{ii} = R_{\text{eff}}(i, i') \leq 2r + R_{\text{eff}}(j, j') \leq 2/|\Theta_{ij}| + 2\Sigma_{jj}$$

which proves the result. □

Remark 7. *Note that the above Lemma is for Θ under the true SDD scaling. It would not make sense for general Θ, because the left hand and right hand sides do not behave the same way when we rescale X_i and X_j.*

The following two lemmas show that in a SDD GGM, the variance of a single node can be bounded as long as we condition on any of its neighbors. In comparison, if we don't condition on anything then the variance can be arbitrarily large: consider the Laplacian of any graph plus a small multiple of the identity.

Lemma 33. *Suppose that i is a non-isolated node in an SDD GGM. Then for any neighbor j it holds that*

$$\text{Var}(X_i | X_j) \leq \frac{1}{|\Theta_{ij}|}$$

Proof. This result can be obtained from the previous Lemma 32 by taking an appropriate limit which sends $\Sigma_{jj} \to 0$. We give an alternate and direct proof below.

Apply the SDD to Laplacian reduction to the precision matrix (with row and column j eliminated) as in Lemma 32 to get a generalized Laplacian L, and then form the standard Laplacian M by adding an additional row and column $n+1$ with $M_{i,n+1} = L_{ii} - \sum_{j=1}^{n} L_{ij}$ and $M_{n+1,n+1} = \sum_{j=1}^{n} M_{j,n}$. Then $u = Lv$ iff there exists z s.t. $(u, z) = M(v, 0)$ where $(v, 0)$ denotes the vector in $\mathbb{R}^{n+1}$ given by adding final coordinate 0. Furthermore it must be that $\sum_i u_i + z = 0$ because (u, z) lies in the span of M. Using the relation between L and M and the triangle inequality and monotonicity (Lemma 30) through the added node $n+1$ we observe

$$\begin{aligned}
\mathrm{Var}(X_i|X_j) &= \frac{1}{2}(e_i - e_i')^T L^{-1}(e_i - e_i') \\
&= \frac{1}{2}(e_i - e_i')^T M^{+}(e_i - e_i') \\
&\leq \frac{1}{2}(e_i - e_{n+1})^T M^{+}(e_i - e_{n+1}) + \frac{1}{2}(e_i' - e_{n+1})^T M^{+}(e_i' - e_{n+1}) \\
&\leq \frac{1}{2}\frac{1}{M_{i,n+1}} + \frac{1}{2}\frac{1}{M_{i',n+1}} \leq \frac{1}{|\Theta_{ij}|}.
\end{aligned}$$

□

Lemma 34. *Suppose that i is a non-isolated node with d neighbors in an SDD GGM. Then for at least one neighbor j it holds that*

$$\mathrm{Var}(X_i|X_j) \leq \frac{4d}{\Theta_{ii}}$$

Proof. We establish the following dichotomy: either $\mathrm{Var}(X_i)$ is already small, or if it is large then there is a j s.t. $1/|\Theta_{ij}|$ is small so $\mathrm{Var}(X_i|X_j)$ is small. Observe by Cauchy-Schwartz that

$$\begin{aligned}
\Theta_{ii}\mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}]) = \Theta_{ii}\mathrm{Cov}(\mathbb{E}[X_i|X_{\sim i}], \mathbb{E}[X_i|X_{\sim i}]) &= \sum_j -\Theta_{ij}\mathrm{Cov}(\mathbb{E}[X_i|X_{\sim i}], X_j) \\
&\leq \sum_j |\Theta_{ij}|\sqrt{\mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}])\mathrm{Var}(X_j)}
\end{aligned}$$

so

$$\begin{aligned}\Theta_{ii}\sqrt{\mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}])} \le \sum_j |\Theta_{ij}|\sqrt{\mathrm{Var}(X_j)} &\le \sum_j |\Theta_{ij}|\sqrt{\mathrm{Var}(X_i) + 1/|\Theta_{ij}|} \\ &\le \sqrt{\mathrm{Var}(X_i)}\sum_j |\Theta_{ij}| + \sum_j \sqrt{|\Theta_{ij}|} \\ &\le \sqrt{\mathrm{Var}(X_i)}\sum_j |\Theta_{ij}| + \sqrt{d\Theta_{ii}}\end{aligned}$$

where in the second inequality we used Lemma 32, in the third inequality we used $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, and in the fourth inequality we used Cauchy-Schwartz and the SDD assumption.

Suppose that $\mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}]) > 4d/\Theta_{ii}$. Then by subtracting $d\sqrt{\Theta_{ii}}$ from both sides we see

$$\frac{1}{2}\Theta_{ii}\sqrt{\mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}])} \le \sqrt{\mathrm{Var}(X_i)}\sum_j |\Theta_{ij}| \le \sqrt{\mathrm{Var}(X_i)}d\max_j |\Theta_{ij}|$$

so using that $\mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}]) = \mathrm{Var}(X_i) - 1/\Theta_{ii} \ge \mathrm{Var}(X_i)/2$ under our assumption, we find

$$\frac{\Theta_{ii}}{4d} \le \frac{\Theta_{ii}}{2d}\sqrt{\frac{\mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}])}{\mathrm{Var}(X_i)}} \le \max_j |\Theta_{ij}|.$$

Let j be the maximizer, then from Lemma 33 we find $\mathrm{Var}(X_i|X_j) \le \frac{1}{|\Theta_{ij}|} \le \frac{4d}{\Theta_{ii}}$, assuming that $\mathrm{Var}(X_i) > 4d/\Theta_{ii}$. Otherwise, by the law of total variance we know $\mathrm{Var}(X_i|X_j) \le \mathrm{Var}(X_i) \le 4d/\Theta_{ii}$. □

Remark 8 (Electrical intuition for Lemma 34)**.** *We explain the electrical intuition behind Lemma 34 in the case of attractive GGMs. First w.l.o.g. we rescale Θ to be a generalized Laplacian (Theorem 38). By the electrical interpretation, we think of the edges of the graph are a collection of resistors connecting the nodes, and we imagine connecting the plus end of a 1-volt battery to node i, so the effective resistance between the plus and minus end of the battery is the reciprocal of the total current which flows. Then $1/\Theta_{ii}$ is the effective resistance when we connect all of the neighbors of node i directly to the negative end of the battery.*

When we do this experiment, we know that a lot of the current is either (1) going directly from node i to ground or (2) going from node i to one of its neighbors j. In case (1), $\mathrm{Var}(X_i)$ *is already small. Otherwise, we are in case (2). In this case, we would expect that if we only grounded node j, then the resulting effective resistance* $\mathrm{Var}(X_i|X_j)$ *should already be quite small; more precisely, within a $O(d)$ factor of grounding all of them, and this is exactly what Lemma 34 says.*

The following example shows that the assumption that the matrix is SDD (or walk-summable) is necessary for the previous Lemmas to be true:

Example 7 (Failure of Lemma 33 in Non-SDD GGM)**.** *Consider for κ fixed and C large*

$$\Theta := \begin{bmatrix} 1 & C & -C \\ C & C^2/\kappa^2 & -C^2/\kappa^2 + 1 \\ -C & -C^2/\kappa^2 + 1 & C^2/\kappa^2 \end{bmatrix}$$

We can verify that as $C \to \infty$ that the variances (i.e. diagonal of Θ^{-1}) remain $\Theta(1)$ *and the matrix is positive definite; furthermore this model is κ-nondegenerate. However, even after conditioning out the first node, the variance of the second (and third) node remains* $\Omega(1) \gg 1/C$.

5.6 Estimating changes in conditional variance

As alluded to before, our algorithms rely on estimating (differences of) conditional variances $\mathrm{Var}(X_i|X_S)$. The classical approach for estimating them is to solve a linear regression problem trying to predict X_i from X_S. As we are working in a sample-starved regime and deal with ill-conditioned matrices, we require very fine grained results about such estimates. We collect such results in this section.

For the analysis of Algorithm HybridMB we only need the basic facts from Section 4.1; for the analysis of Algorithm GreedyAndPrune the key additional fact we need is encapsulated as Lemma 39 in Section 4.3 below; finally, for the analysis of the Algorithm SearchAndValidate we will also directly use the results stated in Section 4.2.

5.6.1 Background: Fixed Design Linear Regression

In this section we recall the standard model for linear regression with Gaussian noise and the usual ordinary least squares estimator and some classical facts about it. See Chapter 14 of [104] for a reference.

Definition 19 (Fixed design regression with Gaussian noise)**.** *The (well-specified)* fixed design regression *model is specified by an unknown parameter $w \in \mathbb{R}^k$, known* design matrix $\mathbb{X} : m \times k$ *with $m > k$ and observations*

$$\mathbb{Y} = \mathbb{X}w + \Xi$$

where $\Xi \sim N(0, \sigma^2 I)$. In other words, $\mathbb{Y} \sim N(\mathbb{X}w, \sigma^2 I)$.

Definition 20 (Ordinary Least Squares (OLS) Estimator)**.** *The OLS estimator for w in the fixed design regression model is the minimizer of*

$$\min_w \|\mathbb{Y} - \mathbb{X}w\|_2^2$$

explicitly given by

$$\hat{w} := (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

assuming that $\mathbb{X}$ has maximal column rank. The corresponding estimator for σ is given by

$$\hat{\sigma}^2 := \frac{1}{m-k}\|\mathbb{Y} - \mathbb{X}\hat{w}\|_2^2.$$

Fact 2 ([104])**.** *Under the fixed design regression model with Gaussian noise, $\hat{w} \sim N(w, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1})$ and $\frac{(m-k)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{m-k}$ where χ^2_{m-k} denotes a χ^2-distribution with $m-k$ degrees of freedom. Furthermore, $\hat{w}$ and $\hat{\sigma}$ are independent.*

Lemma 35. *For any $\delta \in (0, 1)$,*

$$\Pr\left(\left|\frac{\hat{\sigma}^2}{\sigma^2} - 1\right| > 2\sqrt{\frac{\log(2/\delta)}{m-k}} + 2\frac{\log(2/\delta)}{m-k}\right) \leq \delta.$$

Proof. Combine Fact 2 and and the concentration inequality from Lemma 29. □

We end with a geometric interpretation of the OLS coordinates which is analogous to Lemma 28. In statistics this is known as the equivalence of the regression t-test and the 1-variable regression F-test [104].

Lemma 36.

$$\min_w \|\mathbb{Y} - \mathbb{X}w\|_2^2 - \min_{w:w_i=0} \|\mathbb{Y} - \mathbb{X}w\|_2^2 = \frac{\hat{w}_i^2}{[(\mathbb{X}^T\mathbb{X})^{-1}]_{ii}}$$

Proof sketch. Let $\mathbb{X}_i$ be the i'th column of $\mathbb{X}$. By the definition of the OLS estimate $\hat{w}$ and the Pythagorean theorem, the left hand side is equal to $\min_{w:w_i=0} \|\mathbb{X}\hat{w} - \mathbb{X}w\|_2^2$. By another application of the Pythagorean theorem, this equals $\|\mathbb{X}_i\hat{w}_i - \text{Proj}_{V_i} \mathbb{X}_i\hat{w}_i\|_2^2 = \hat{w}_i^2\|\mathbb{X}_i - \text{Proj}_{V_i} \mathbb{X}_i\|_2^2$ where V_i is the subspace spanned by the columns of $\mathbb{X}$ except for i. Finally $\|\mathbb{X}_i - \text{Proj}_{V_i} \mathbb{X}_i\|_2^2 = \frac{1}{[(\mathbb{X}^T\mathbb{X})^{-1}]_{ii}}$ by applying Schur complement formulas. □

5.6.2 Background: Wishart Matrices

Under fixed design, the matrix $\mathbb{X}$ was considered to be a deterministic quantity. Random design (see e.g. [94] for references) corresponds to the case where the rows of $\mathbb{X}$ are i.i.d. samples from some distribution, which fits the usual setup in statistical learning theory.

Definition 21 (Random design linear regression with Gaussian covariates)**.** *The random design linear regression model with Gaussian covariates with m samples is given by a (typically unknown) covariance matrix $\Sigma : k \times k$, i.i.d. samples $X^{(1)}, \ldots, X^{(m)} \sim N(0, \Sigma)$ and corresponding observations*

$$Y^{(i)} = \langle X^{(i)}, w\rangle + \xi^{(i)}, \quad i = 1, \ldots, m \tag{5.3}$$

where each $\xi^{(i)} \sim N(0, \sigma^2)$ is independent noise. (The assumption that $\xi^{(i)}$ is independent is referred to as the model being well-specified*.)*

The OLS estimator is defined as before in Definition 20 where the rows of the design matrix $\mathbb{X}$ are the samples $X_1, \ldots, X_m$ and $\mathbb{Y} = (Y^{(i)})_{i=1}^m$. From (2) we still

have that for fixed $X_1, \ldots, X_m$ (i.e. considering only the randomness over $\xi_1, \ldots, \xi_m$)

$$\hat{w}_{OLS} \sim N(w, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}).$$

Therefore reasoning about the OLS estimator under random design can be reduced to understanding the random matrix $\mathbb{X}^T\mathbb{X}$, which is referred to as a *Wishart matrix* (with m degrees of freedom). We recall here a standard concentration inequality for Wishart matrices when $\Sigma = I$. (This inequality generalizes to the sub-Gaussian case and we have specialized it for simplicity.)

Theorem 39 (Theorem 4.6.1, [188])**.** *Suppose that $X^{(1)}, \ldots, X^{(m)} \sim N(0, I)$ are independent Gaussian random vectors in $\mathbb{R}^k$, then*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} X^{(i)}(X^{(i)})^T - Id \right\| \leq C_1 \left(\sqrt{\frac{k}{m}} + \sqrt{\frac{\log(2/\delta)}{m}} \right)$$

for some absolute constant $C_1 > 0$, with probability at least $1 - \delta$.

This leads to a multiplicative guarantee for general Wishart matrices:

Lemma 37. *Suppose $\epsilon \in (0, 1/2)$ and $\delta > 0$. Then for any m such that $\epsilon \leq C_1 \left(\sqrt{\frac{k}{m}} + \sqrt{\frac{\log(2/\delta)}{m}} \right)$ and $X^{(1)}, \ldots, X^{(m)} \sim N(0, I)$ we have that*

$$(1 - \epsilon)\Sigma \preceq \frac{1}{m} \sum_i X_i X_i^T \preceq (1 + \epsilon)\Sigma$$

with probability at least $1 - \delta$.

Proof. This is equivalent to showing that

$$(1 - \epsilon)I \preceq \frac{1}{m} \sum_i \Sigma^{-1/2} X^{(i)} (\Sigma^{-1/2} X^{(i)})^T \preceq (1 + \epsilon)I$$

since the PSD ordering is preserved under matrix congruence. The above follows from applying Theorem 39 to $\bar{X}^{(i)} = \Sigma^{-1/2} X^{(i)}$. □

Definition 22. *Given i.i.d. mean-zero random vectors* $X^{(1)}, \ldots, X^{(m)}$ *the* empirical covariance matrix *is*

$$\widehat{\Sigma} := \frac{1}{m} \sum_i X^{(i)} (X^{(i)})^T.$$

5.6.3 Estimating changes in conditional variance

We are now ready to state what we need for estimating changes in conditional variance. Recall the basic setup: Given samples from X from a GGM at various stages in our algorithm we use estimates for conditional variances of the form $\mathrm{Var}(X_i|X_S)$ by regressing X_i against X_S. What we really we need are not actual values of $\mathrm{Var}(X_i|X_S)$ but to find a variable $j \notin S$ that gives non-trivial (or even the*most*) advantage in predicting $X_i|X_{S\cup\{j\}}$. So we need to quantify the relative advantage of including an additional variable j on top of S.

We can abstract the above in the regression setting as follows: Given samples for regression (X, Y), and an index j check if $\mathrm{Var}(Y|X) = \mathrm{Var}(Y|X_{\sim j})$. That is, whether including feature x_j gives non-trivial advantage in regression. This is akin to the classical *regression t-test* in statistics (see [104]) used to test the null hypothesis that $w_i = 0$ in a linear regression problem.

In the greedy steps in our learning algorithm, we will need to not only find a feature which has a nonzero value for predicting Y, but in fact we want to find one of the most predictive features. We do so by exploiting what is known as a *non-central* F-statistic [104]. The following lemma quantifies the *usefulness* of a particular coordinate for estimating Y. Crucially, this Lemma shows we can estimate the (normalized) change in conditional variance much more accurately than we can actually estimate the individual conditional variances. Note that by Lemma 36 that the term which appears in the Lemma, $\frac{|\hat{w}_j|^2}{(\hat{\Sigma}^{-1})_{jj}}$, also equals the difference in squared loss over the data between the OLS estimator constrained to $w_j = 0$ and the unconstrained OLS estimator.

Lemma 38. *Consider the Gaussian random design regression setup* (5.3)*, fix* $j \in$

$\{1, \ldots, k\}$ *and let*

$$\gamma := \frac{\mathrm{Var}(Y|X_{\sim j}) - \mathrm{Var}(Y|X)}{\mathrm{Var}(Y|X)}$$

where $X_{\sim j} = (X_i)_{i \neq j}$. *We have*

$$\left| \frac{|\hat{w}_j|}{\hat{\sigma}\sqrt{(\hat{\Sigma}^{-1})_{jj}}} - \sqrt{\gamma} \right| \leq \sqrt{\frac{4\log(4/\delta)}{m}} + \sqrt{\frac{\gamma}{64}}$$

and

$$\left| \frac{|\hat{w}_j|}{\sigma\sqrt{(\hat{\Sigma}^{-1})_{jj}}} - \sqrt{\gamma} \right| \leq \sqrt{\frac{2\log(4/\delta)}{m}} + \sqrt{\frac{\gamma}{64}}$$

with probability at least $1 - \delta$ *as long as* $m \geq m_0 = O(k + \log(4/\delta))$.

Proof. We prove this result directly. Alternatively and essentially equivalently, one could derive a similar result by using classical results in the fixed design regression setting for non-central F-statistics (Theorem 14.11 of [104], see also Section 5.8 below) and then analyzing their behavior under random design using matrix concentration. A benefit of this direct analysis is that it generalizes more straightforwardly to sub-gaussian settings.

Recall from Lemma 28 (applied for fixed X_S and then taking expectations) that

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X_{\sim j}] + \frac{\mathrm{Cov}(Y, X_j|X_{\sim j})}{\mathrm{Var}(X_j|X_{\sim j})}(X_j - \mathbb{E}[X_j|X_{\sim j}])$$

and that

$$\mathrm{Var}(Y|X_{\sim j}) - \mathrm{Var}(Y|X) = \frac{\mathrm{Cov}(Y, X_j|X_{\sim j})^2}{\mathrm{Var}(X_j|X_{\sim j})}$$

so

$$w_j^2 \mathrm{Var}(X_j|X_{\sim j}) = \mathrm{Var}(Y|X_{\sim j}) - \mathrm{Var}(Y|X). \tag{5.4}$$

i.e. $\frac{w_j^2}{\sigma^2(\Sigma^{-1})_{jj}} = \gamma$. We know that for fixed X, over the randomness of ξ we have

$\hat{w}_{OLS} \sim N(w, \frac{\sigma^2}{m}\hat{\Sigma}^{-1})$ by Fact 2, so

$$\frac{\hat{w}_j}{\sigma\sqrt{(\hat{\Sigma}^{-1})_{jj}}} \sim N\left(\frac{w_j}{\sigma\sqrt{(\hat{\Sigma}^{-1})_{jj}}}, \frac{1}{m}\right).$$

Using that $(\Sigma^{-1})_{jj} = \frac{1}{\mathrm{Var}(X_j|X_S)}$, $\sigma = \sqrt{\mathrm{Var}(Y|X)}$, and $\gamma = \frac{\mathrm{Var}(Y|X_{\sim j}) - \mathrm{Var}(Y|X)}{\mathrm{Var}(Y|X)}$ and (5.4) we find

$$\frac{\hat{w}_j}{\sigma\sqrt{(\hat{\Sigma}^{-1})_{jj}}} \sim N\left(\pm\sqrt{\gamma\frac{(\Sigma^{-1})_{jj}}{(\hat{\Sigma}^{-1})_{jj}}}, \frac{1}{m}\right)$$

where the sign is the sign of w_j. Applying $||a| - |b|| \leq |a - b|$ and the Gaussian tail bound over the randomness of $\hat{w}$ we find

$$\Pr\left(\left|\frac{|\hat{w}_j|}{\sigma\sqrt{(\hat{\Sigma}^{-1})_{jj}}} - \sqrt{\gamma\frac{(\Sigma^{-1})_{jj}}{(\hat{\Sigma}^{-1})_{jj}}}\right| > t\right) \leq \Pr\left(\left|\frac{\hat{w}_j}{\sigma\sqrt{(\hat{\Sigma}^{-1})_{jj}}} \mp \sqrt{\gamma\frac{(\Sigma^{-1})_{jj}}{(\hat{\Sigma}^{-1})_{jj}}}\right| > t\right) \leq 2e^{-mt^2/2}.$$

Applying Lemma 35 gives

$$\left|\frac{\hat{\sigma}}{\sigma} - 1\right| \leq 2\sqrt{\frac{\log(4/\delta)}{m-k-1}} + 2\frac{\log(4/\delta)}{m-k-1}$$

with probability at least $1 - \delta/2$. Therefore as long as $m \geq m_1 = O(k + \log(4/\delta))$ we have $\frac{\hat{\sigma}}{\sigma} \in (7/8, 9/8)$. Taking $t = \sqrt{2\log(4/\delta)/m}$ we have

$$\begin{aligned}\left|\frac{|\hat{w}_j|}{\hat{\sigma}\sqrt{(\hat{\Sigma}^{-1})_{jj}}} - \sqrt{\gamma}\right| &\leq \sqrt{\frac{\sigma}{\hat{\sigma}}}\left|\frac{|\hat{w}_j|}{\sigma\sqrt{(\hat{\Sigma}^{-1})_{jj}}} - \sqrt{\gamma\frac{\hat{\sigma}}{\sigma}}\right| + \sqrt{\gamma}\left|1 - \sqrt{\frac{(\hat{\Sigma}^{-1})_{jj}}{(\Sigma^{-1})_{jj}}}\right| \\ &\leq \sqrt{\frac{4\log(4/\delta)}{m}} + \sqrt{\frac{\gamma}{64}}\end{aligned}$$

applying Lemma 37 and requiring $m \geq m_2 = O(k + \log(4/\delta))$, with probability at least $1 - \delta$. A simpler variant of this argument gives the result for $\frac{|\hat{w}_j|}{\sigma\sqrt{(\hat{\Sigma}^{-1})_{jj}}}$ as well. □

In our analysis we will often need to estimate multiplicative changes in a quantity of the form $\mathrm{Var}(Y|X_{\sim j}) - V$ (where e.g. $V = \mathrm{Var}(Y|X, X')$ for some X') so we will

use the following variant of the previous Lemma:

Lemma 39. *Consider the Gaussian random design regression setup* (5.3)*, fix* $j \in \{1, \dots, k\}$*, let* $V > 0$ *be arbitrary s.t.* $V < \mathrm{Var}(Y|X)$ *and let*

$$\gamma' := \frac{\mathrm{Var}(Y|X_{\sim j}) - \mathrm{Var}(Y|X)}{\mathrm{Var}(Y|X_{\sim j}) - V}$$

where $X_{\sim j} = (X_i)_{i \neq j}$*. We have*

$$\left| \sqrt{\frac{1}{\mathrm{Var}(Y|X_{\sim j}) - V}} \frac{|\hat{w}_j|}{\sqrt{(\hat{\Sigma}^{-1})_{jj}}} - \sqrt{\gamma'} \right| \leq \sqrt{\frac{\mathrm{Var}(Y|X)}{\mathrm{Var}(Y|X_{\sim j}) - V} \cdot \frac{2\log(4/\delta)}{m}} + \sqrt{\frac{\gamma'}{64}}$$

with probability at least $1 - \delta$ *as long as* $m \geq m_0 = O(k + \log(4/\delta))$*.*

Proof. This follows from Lemma 38 after multiplying through in the guarantee by $\sqrt{\gamma'/\gamma}$, using that $\sigma = \sqrt{\mathrm{Var}(Y|X)}$. □

5.7 Learning all attractive GGMs efficiently

Definition 23. *We say that a GGM is* attractive *(or* ferromagnetic*) if* $\Theta_{ij} \leq 0$ *for all* $i \neq j$*. (This is the same as requiring that* Θ *is an M-matrix.)*

Lemma 40. *If* Θ *is the precision matrix of an attractive GGM, then there exists an invertible diagonal matrix* D *with nonnegative entries such that* $D\Theta D$ *is a generalized Laplacian.*

Proof. This follows immediately from Theorem 38. □

A particularly important example of an attractive GGM is the *discrete Gaussian free field* — see [167] for a reference to this and the closely related literature on the *continuum Gaussian free field.*

Definition 24. *The* discrete Gaussian free field *on a weighted graph* G *with zero boundary conditions on* S *is the GGM with* Θ *the Laplacian of* G*, after eliminating the rows and columns corresponding to the nodes in* S*.*

Without boundary conditions, the GFF should be translation invariant and so it does not exist as a probability distribution. One can approach the free boundary situation by taking the Laplacian and adding ϵI to make it invertible, which gives a learnable model that is arbitrarily poorly conditioned.

Example 8 (Gaussian simple random walk)**.** *Consider the discrete Gaussian free field on a path of length n with zero boundary condition on the first node. This process is the same as a simple random walk with $N(0,1)$ increments. That is the resulting distribution is of the form $(X_1, \ldots, X_n)$ where $X_i = \sum_{j \le i} \eta_j$ for independent and identical $\eta_j \sim N(0,1)$. From the GFF perspective, we can think of this as a discretization of Brownian motion (the one-dimensional (continuum) Gaussian free field).*

Remark 9. *Every attractive GGM can be realized from a Gaussian Free Field on a weighted graph in the following way: given an attractive GGM, first rescale the coordinates using the above Lemma so that it is a generalized Laplacian. Then, by adding one node to the model we can make the precision matrix into a standard Laplacian on some weighted graph, and conditioning out the added node recovers the original precision matrix.*

Our main theorem of this section is a sample-efficient algorithm for learning attractive GGMs:

Theorem 40. *Fix a κ-nondegenerate attractive GGM. Algorithm* GreedyAndPrune *returns the true neighborhood of every node i with probability at least $1 - \delta$ for $\nu = \kappa^2/\sqrt{32}, K = 64d\log(4/\kappa^2) + 1$ as long as the number of samples $m \ge m_1$ for $m_1 = O((1/\kappa^2)(K\log(n) + \log(4/\delta)))$. The combined run-time (over all nodes) of the algorithm is $O(K^3 m n^2)$.*

Note that the above immediately implies Theorem 34.

As mentioned in the introduction, Algorithm GreedyAndPrune learns the neighborhood of a node by doing greedy forward selection to minimize the conditioned variance, and then doing pruning to remove non-neighbors from the candidate

Algorithm OrthogonalMatchingPursuit$(T, \mathbb{X}, \mathbb{Y})$:

1. Set $S_0 := \{\}$.
2. For t from 1 to T:
   (a) Choose j which minimizes
   $$\min_{w \,:\, \mathrm{supp}(w) \subset S_{t-1} \cup \{j\}} \|\mathbb{Y} - \mathbb{X}w\|_2^2$$
   (b) Set $S_t := S_{t-1} \cup \{j\}$
3. Return S_T.

Algorithm GreedyAndPrune(i, ν, T):

1. Run OrthogonalMatchingPursuit for T steps to predict $\mathbb{X}_i$ from the other columns of $\mathbb{X}$, i.e. setting $\mathbb{Y} = \mathbb{X}_i$ and $\mathbb{X}' = \mathbb{X}_{\sim i}$.
2. Define $\hat{\Theta}_{ii}$ by $1/\hat{\Theta}_{ii} = \widehat{\mathrm{Var}}(X_i|X_S)$.
3. For $j \in S$:
   (a) Let $S' := S \setminus \{j\}$ and $\hat{w} := \hat{w}(i, S')$.
   (b) If $\widehat{\mathrm{Var}}(X_i|X_{S'}) - \widehat{\mathrm{Var}}(X_i|X_S) < \nu/\hat{\Theta}_{ii}$, set $S := S'$.
4. Return S.

neighborhood. The greedy forward selection step is known in the compressed sensing literature as *Orthogonal Matching Pursuit*[7] (OMP) (see e.g. [183]). We give a description of the OMP algorithm in the general setting of Section 5.6.1 below, along with pseudocode for GreedyAndPrune.

Remark 10 (Implementation: Merging neighborhoods)**.** *In order to return an actual estimate for the inverse precision matrix, we add in our implementation of* GreedyAndPrune *a merging step which includes an edge (i, j) iff it is in the computed neighborhood of node i and in the computed neighborhood of node j. Then to estimate the entries, we use OLS to predict node X_i from its neighbors and estimate*

[7]The terms forward selection and OMP are not always used for the exact same algorithm. In the language of [48], the algorithm we consider would be called forward selection and not OMP.

the conditional variance of X_i. We define $\hat{\Theta}_{ii}$ to be the inverse of the estimated conditional variance, and $-\hat{\Theta}_{ij}/\hat{\Theta}_{ii}$ to be the OLS coefficient. Finally, we symmetrize $\hat{\Theta}$ by picking the smaller of absolute norm between $\hat{\Theta}_{ij}$ and $\hat{\Theta}_{ji}$; the same step is used in CLIME [39].

5.7.1 Combinatorial proof of supermodularity

As a first step toward proving Theorem 40, we first show that the conditional variance function is supermodular.

Definition 25. *Given a universe U, a function $f : 2^U \to \mathbb{R}$ is* supermodular *if for any $S \subset T$,*

$$f(S) - f(S \cup \{j\}) \geq f(T) - f(T \cup \{j\}).$$

(This is the same as saying $-f$ is submodular*.)*

Supermodularity of the conditional variance of a node in the GFF (and hence, by using the reduction from Remark 9, all attractive GGMs) was previously shown independently in [128, 129] using two different methods. The proof in [128] is algebraic using the Schur complement formula, whereas the proof in [129] converts the problem into one about electrical flows and argues via Thomson's principle. We give a third different proof which has the benefit of being transparent and combinatorial in nature.

Theorem 41. *For any node i in a ferromagnetic GGM, $\mathrm{Var}(X_i|X_S)$ is a monotonically decreasing, supermodular function of S.*

Proof. By rescaling we may assume w.l.o.g. that $\Theta_{ii} = 1$ for all i. Define Θ_S to be the precision matrix corresponding to conditioning S out (i.e. Θ with the rows and columns of S removed), and $\Sigma_S = \Theta_S^{-1}$. Then, if we write $\Theta_S = I - A_S$, by Neumann series formula (as $\Theta_S \succ 0$, $\|A_S\| < 1$ using Perron-Frobenius), we see

$$\Sigma_S = (I - A_S)^{-1} = \sum_{k=0}^{\infty} A_S^k. \tag{5.5}$$

Writing this out explicitly for $(\Sigma_S)_{i,i}$ gives

$$\mathrm{Var}(X_i|X_S) = \sum_{k=0}^{\infty} \sum_{v_1,\ldots,v_k \notin S} (-\Theta_{iv_1}) \cdots (-\Theta_{v_k i}), \tag{5.6}$$

where the $k = 0$ term in the sum is interpreted to be 1, so $\mathrm{Var}(X_i|X_S)$ is a nonnegative weighted sum over walks avoiding S and returning to i in the final step. The above expression is clearly monotonically increasing in S as all off-diagonal entries of Θ are negative (and also follows from law of total variance); to verify supermodularity, we just need to check that

$$\mathrm{Var}(X_i|X_S) - \mathrm{Var}(X_i|X_{S\cup\{j\}}) = \sum_{k=0}^{\infty} \sum_{\substack{v_1,\ldots,v_k \notin S, \\ j \in \{v_1,\ldots,v_k\}}} (-\Theta_{iv_1}) \cdots (-\Theta_{v_k i})$$

is a monotonically decreasing function of $S \subseteq [n] \setminus \{i, j\}$, but this is clear once we apply (5.6) as the set of cycles that are eliminated from the sum by adding j only shrinks as we increase S. □

Supermodularity of the conditional variance has the following consequence which will later be useful in showing that the greedy algorithm makes non-trivial progress in each step.

Lemma 41. *For any node i in a ferromagnetic GGM, if S is a set of nodes that does not contain i or all neighbors of i, and T is the set of neighbors of i not in S, then there exists some node $j \in T$ such that*

$$\mathrm{Var}(X_i|X_S) - \mathrm{Var}(X_i|X_{S\cup\{j\}}) \geq \frac{\mathrm{Var}(X_i|X_S) - 1/\Theta_{ii}}{|T|}.$$

Proof. This is a standard consequence of supermodularity – we include the proof for completeness.

Consider adjoining the elements of T to S one at a time, and then apply super-

modularity to show

$$\begin{aligned}\mathrm{Var}(X_i|X_S) - \mathrm{Var}(X_i|X_{S\cup T}) &\leq \sum_{j\in T}(\mathrm{Var}(X_i|X_S) - \mathrm{Var}(X_i|X_{S\cup\{j\}})) \\ &\leq |T| \max_{j\in T}(\mathrm{Var}(X_i|X_S) - \mathrm{Var}(X_i|X_{S\cup\{j\}})).\end{aligned}$$

Rearranging and using $\mathrm{Var}(X_i|X_{S\cup T}) = 1/\Theta_{ii}$ (by the Markov property) gives the result. □

From (5.5) we see immediately that the entries of the covariance Σ of an attractive GGM are always nonnegative (this is why they are called attractive/ferromagnetic); we record this fact for future use.

Lemma 42 (Griffith's inequality)**.** *In an attractive GGM,* $\mathrm{Cov}(X_i, X_j) \geq 0$ *for any* i, j*.*

This fact is very well-known, holds for arbitrary ferromagnetic graphical models (i.e. not just Gaussian) and is referred to as *Griffith's inequality*. See [82] for a more general proof.

5.7.2 Greedy Subset Selection in Attractive Models

In this section we give a guarantee for *subset selection* using OMP, by showing that after a small number of rounds OMP finds a set S such that $\mathrm{Var}(X_i|X_S)$ is close to minimal. The sample complexity analysis is complicated by the fact that supermodularity holds at the level of the population loss (i.e. for an infinite amount of data) whereas it would be more convenient if it held for the empirical conditional variance, so we have to deal with both the regression noise and the randomness of the regressors. First we prove the following lemma which gives a stronger version of Lemma 27 for ferromagnetic GGMs:

Lemma 43. *Fix* i *a node in a* κ*-nondegenerate ferromagnetic GGM, and let* S *be set*

of nodes and let T be the set of neighbors of i not in S. Then

$$\mathrm{Var}(X_i|X_S) \geq \frac{1+|T|\kappa^2}{\Theta_{ii}}$$

Proof. By the law of total variance, Griffith's inequality (Lemma 42), and the law of total variance again

$$\begin{aligned}\mathrm{Var}(X_i|X_S) - \frac{1}{\Theta_{ii}} &= \mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}]|X_S) \\ &= \mathrm{Var}(\sum_{j\in T} \frac{-\Theta_{ij}}{\Theta_{ii}} X_j|X_S) \\ &\geq \sum_{j\in T} \frac{\Theta_{ij}^2}{\Theta_{ii}^2}\mathrm{Var}(X_j|X_S) \geq \frac{1}{\Theta_{ii}}\sum_{j\in T}\frac{\Theta_{ij}^2}{\Theta_{ii}\Theta_{jj}} \geq \frac{|T|\kappa^2}{\Theta_{ii}}.\end{aligned}$$

□

Lemma 44. *Suppose that X is distributed according to an κ-nondegenerate ferromagnetic GGM and i is a node of degree at most d. Let $\sigma^2 := \frac{1}{\Theta_{ii}}$ and $w_j^* = \frac{-\Theta_{ij}}{\Theta_{ii}}$ for all $j \neq i$. Then using T rounds of OMP to predict X_i given $X_{\sim i}$ from m i.i.d. samples, we have that $\mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}]|X_S) \leq (1-1/2d)^{T-1}\frac{8d}{\Theta_{ii}}$ with probability at least $1-\delta$ provided that $m = \Omega((d+1/\kappa^2)(T\log(n)+\log(2/\delta)))$.*

Proof. We prove by induction that for every $1 \leq t \leq T$ that

$$\mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}]|X_{S_t}) \leq (1-1/2d)^{t-1}\frac{8d}{\Theta_{ii}}.$$

Note that by Lemma 34 there exists a node j such that $\mathrm{Var}(X_i|X_j) \leq \frac{4d}{\Theta_{ii}}$. By taking a union bound, we may assume that:

1. $\mathrm{Var}(X_i|X_{S_1}) \leq \frac{8d}{\Theta_{ii}}$ using the above fact combined with Lemma 35 assuming that $m = \Omega(\log(n/\delta))$ to guarantee that the estimated conditional variances have small multiplicative error.

2. For all subsets U of $[n]$ of size at most T and $j \in [n]$, applying Lemma 39 we

have

$$\left|\frac{1}{\sqrt{\mathrm{Var}(X_i|X_{U\setminus\{j\}}) - 1/\Theta_{ii}}}\hat{R}(U,j) - \sqrt{\gamma'}\right| \leq \sqrt{\frac{\mathrm{Var}(X_i|X_U)}{\mathrm{Var}(X_i|X_{U\setminus\{j\}}) - 1/\Theta_{ii}}}\sqrt{\frac{4(T\log(n) + \log(12/\delta))}{m}} + \sqrt{\frac{\gamma'}{64}}$$

where

$$\gamma' = \gamma'(U,j) := \frac{\mathrm{Var}(X_i|X_{U\setminus\{j\}}) - \mathrm{Var}(X_i|X_U)}{\mathrm{Var}(X_i|X_{U\setminus\{j\}}) - 1/\Theta_{ii}}$$

and

$$\hat{R}(U,j) := \frac{(\hat{w}_U)_j}{((\hat{\Sigma}_{U,U})^{-1})_{jj}} = \sqrt{\|\mathbb{X}_i - \mathbb{X}\hat{w}_U\|_2^2 - \|\mathbb{X}_i - \mathbb{X}\hat{w}_{U\setminus\{j\}}\|_2^2}$$

using Lemma 36 in the last equality where $\hat{w}_U$ is the OLS estimate using only the coordinates in U. This holds assuming that $m = \Omega(T\log(4n) + \log(1/\delta))$.

Before proceeding, we observe that

$$\sqrt{\frac{\mathrm{Var}(X_i|X_U)}{\mathrm{Var}(X_i|X_{U\setminus\{j\}}) - 1/\Theta_{ii}}} \leq \sqrt{\frac{\mathrm{Var}(X_i|X_{U\setminus\{j\}})}{\mathrm{Var}(X_i|X_{U\setminus\{j\}}) - 1/\Theta_{ii}}} \leq \max(\sqrt{2}, \sqrt{2/d'\kappa^2}) \quad (5.7)$$

where d' is the degree of node i in the graph with the nodes in $U \setminus \{j\}$ removed, by the law of total variance (first inequality) and the following case analysis: either $\mathrm{Var}(X_i|X_{U\setminus\{j\}}) \geq 2/\Theta_{ii}$, in which case $\frac{\mathrm{Var}(X_i|X_{U\setminus j\}})}{\mathrm{Var}(X_i|X_{U\setminus\{j\}})-1/\Theta_{ii}} \leq 2$, or $\mathrm{Var}(X_i|X_{U\setminus\{j\}}) \leq 2/\Theta_{ii}$ in which case $\frac{\mathrm{Var}(X_i|X_{U\setminus\{j\}})}{\mathrm{Var}(X_i|X_{U\setminus\{j\}})-1/\Theta_{ii}} \leq 2/d'\kappa^2$ by Lemma 43.

The first point above gives the base case for the induction. By Lemma 41, if $\mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}]|S_t) \neq 0$ then there exists a k such that

$$\gamma'(S_t \cup \{k\}, k) = \frac{\mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}]|X_{S_t}) - \mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}]|X_{S_t\cup\{k\}})}{\mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}]|X_{S_t\cup\{k\}})} \geq \frac{1}{d'}$$

where (as above) d' is the degree of i in the set of non-neighbors of S_t. Combined with (5.7) and $d' \leq d$ we now see that the second guarantee above ensures that at every time t, the j selected by OMP (i.e. j where $S_{t+1} = S_t \cup \{j\}$) satisfies $\gamma'(S_t \cup \{j\}, j) \geq 1/2d$

as long as $m = \Omega((d + 1/\kappa^2)(T\log(n) + \log(12/\delta)))$. We therefore have that

$$\mathrm{Var}(X_i|X_{S_t}) - 1/\Theta_{ii} \leq (1 - 1/2d)(\mathrm{Var}(X_i|X_{S_{t-1}}) - 1/\Theta_{ii})$$

for all $1 < t \leq T$, which combined with the induction hypothesis gives the result (using that $\mathrm{Var}(X_i|X_{S_t}) - 1/\Theta_{ii} = \mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}]|X_{S_t})$ by law of total variance). □

5.7.3 Structure Recovery for Attractive GGMs

To give a final result for structure recovery, we show how to combine the previous analysis of greedy forward selection with a simple analysis of pruning (backward selection).

Lemma 45. *Let i be a node of degree at most d in a κ-nondegenerate attractive GGM. Fix $\delta > 0$ and suppose that $m = \Omega((d + 1/\kappa^2)(T\log(n) + \log(2/\delta)))$ where $T = \Theta(d\log(2d/\kappa^2))$. Then with probability at least $1 - \delta$, the neighborhod of node i is correctly recovered by Algorithm* GreedyAndPrune *with $\nu = \Theta(\kappa^2)$.*

Proof. By Lemma 44 with $T = 1 + 2d\log(16d/\kappa^2)$, with probability at least $1 - \delta/2$ we have that $\mathrm{Var}(\mathbb{E}[X_i|X_{\sim i}] \mid X_S) \leq \kappa^2/2$ where S is the set returned by OMP as long as $m = \Omega((d+1/\kappa^2)(T\log(n)+\log(24/\delta)))$. From Lemma 27 we see this implies that S contains the true neighborhood of node i.

We now analyze the pruning step for any S which is a superset of the true neighborhood of size at most T. By Lemma 27 and the Markov property, we know that if j is a true neighbor then $\gamma(S, j) \geq \kappa^2$, and otherwise $\gamma(S, j) = 0$. Applying Lemma 38 and taking the union bound over the at most n^T possible sets, we find that exactly the true edges are kept with probability at least $1 - \delta/2$ as long as $m = \Omega((T\log(n) + \log(8/\delta))/\kappa^2)$. Therefore the entire neighborhood recovery succeeds with probability at least $1 - \delta$. □

Theorem 42. *Let X be distributed according to a κ-nondegenerate GGM on n nodes with maximum degree d. Fix $\delta > 0$, then with probability at least $1 - \delta$ Algorithm* GreedyAndPrune *run at every node with $T = \Theta(d\log(2d/\kappa^2))$ and*

$\nu = \Theta(\kappa^2)$ *successfully recovers the true graph as long as* $m = \Omega((1/\kappa^2)(d\log(2d/\kappa^2) + \log(2/\delta))\log(n))$.

Proof. This follows from Lemma 45 by taking the union bound over the n nodes and recalling from Lemma 31 the bound $d \leq 1/\kappa^2$. □

Remark 11 (Input specification)**.** *In the description of the algorithms throughout this paper, we assume we have access to i.i.d. samples from the distribution. However, it is straightforward to verify that the algorithms only depend on the empirical covariance matrix, and can be run given only the empirical covariance matrix in polynomial time.*

5.8 Information-theoretic optimal learning of attractive GGMs

In this section we give an $O(n^d)$ time algorithm for recovering attractive GGMs which matches the information-theoretic lower bounds up to constants, improving the result of the previous section at the cost of computational efficiency.

5.8.1 Background: Noncentral F-statistics

In the analysis of the $O(n^d)$ time algorithm, we will need to compare empirical variances between predictors supported on very different sets of variables. In comparison, in the analysis of greedy methods we only needed to consider adding or removing a single variable at a time. In order to handle the new setting, we recall the definition of noncentral F-statistics and their connection to fixed design regression.

Definition 26. *Suppose* $Z_1 \sim N(\delta, 1)$ *and for* $j > 1$, $Z_j \sim N(0,1)$ *with* $Z_1, \ldots, Z_m$ *independent. Then we write* $\sum_i Z_i \sim \chi^2_m(\delta^2)$ *where* $\chi^2_m(\delta^2)$ *is the* noncentral chi-square distribution *with noncentrality parameter* δ^2 *and* m *degrees of freedom.*

Definition 27. *If* $V \sim \chi^2_k(\delta^2)$ *and* $W \sim \chi^2_m$ *is independent of* V*, then we write*

$$\frac{V/k}{W/m} \sim F_{k,m}(\delta^2)$$

where $F_{k,m}(\delta^2)$ *is the* noncentral F-distribution *with degrees of freedom* k *and* m *and noncentrality parameter* δ^2*.*

Theorem 43 (Theorem 14.11 of [104])**.** *In the (Gaussian) fixed design regression model (Section 5.6.1), let* H *be a* q*-dimensional subspace of* $\mathbb{R}^k$*. Define*

$$T := \frac{m-k}{k-q} \frac{\|\mathbb{Y} - \mathbb{X}\hat{w}_0\|^2 - \|\mathbb{Y} - \mathbb{X}\hat{w}\|^2}{\|\mathbb{Y} - \mathbb{X}\hat{w}\|^2} = \frac{m-k}{k-q} \frac{\|\mathbb{X}\hat{w} - \mathbb{X}\hat{w}_0\|^2}{\|\mathbb{Y} - \mathbb{X}\hat{w}\|^2}$$

where $\hat{w}$ *is the unrestricted OLS estimator and* $\hat{w}_0$ *is the least squares estimator constrained to be inside of subspace* H*. (The second equality holds by the Pythagorean theorem.) Then* $T \sim F_{k-q,m-k}(\gamma)$ *where*

$$\gamma := \frac{\min_{w_0 \in H_0} \|\mathbb{X}(w - w_0)\|^2}{\sigma^2}.$$

More specifically, $\frac{1}{\sigma^2}\|\mathbb{Y} - \mathbb{X}\hat{w}\|^2 \sim \chi^2_{m-k}$ *and* $\frac{1}{\sigma^2}\|\mathbb{X}\hat{w} - \mathbb{X}\hat{w}_0\|^2 \sim \chi^2_{k-q}(\gamma)$ *and these random variables are independent.*

We also recall a convenient concentration inequality for noncentral χ^2-distributed random variables:

Lemma 46 (Lemma 8.1 of [21])**.** *Suppose that* $V \sim \chi^2_m(\delta^2)$*. Then*

$$\Pr(V \geq (m + \delta^2) + 2\sqrt{(m + 2\delta^2)t} + 2t) \leq e^{-t}$$

and

$$\Pr(V \leq (m + \delta^2) - 2\sqrt{(m + 2\delta^2)t}) \leq e^{-t}.$$

5.8.2 Structure learning by ℓ_0-constrained least squares

We perform structure recovery by, for every node i, performing several ℓ_0-constrained regressions and pruning the result. In the context of learning Gaussian graphical models, some algorithms in a similar spirit referred to as SLICE and DICE were proposed in [140] and they proved a sample complexity bound of $O(d/\kappa^2 \log(n))$ for the more sample-efficient method, DICE. We show our estimator SearchAndValidate

Algorithm SearchAndValidate(i,d,ν):

1. We assume the data has been split into two equally sized sample sets 1 and 2. Let $\hat{\mathbb{E}}_1$ and $\hat{\mathbb{E}}_2$ denote the empirical expectation over these two sets and define $\widehat{\mathrm{Var}}_2$ similarly.
2. For d' in 0 to d:
   (a) Find $w_{d'}$ minimizing
   $$\min_{w: w_i = 0, |\operatorname{supp}(w)| \le d'} \hat{\mathbb{E}}_1[(X_i - w_{d'} \cdot X)^2]$$
3. For d' in 0 to d (outer loop):
   (a) For d'' in 0 to d except d' (inner loop):
      i. Let $S_{d',d''} := \operatorname{supp}(w_{d'}) \cup \operatorname{supp}(w_{d''})$.
      ii. For j in $\operatorname{supp}(w_{d''}) \setminus \operatorname{supp}(w_{d'})$
         A. If $\widehat{\mathrm{Var}}_2(X_i | X_{S_{d',d''} \setminus \{j\}}) - \widehat{\mathrm{Var}}_2(X_i | X_{S_{d',d''}}) > \nu \widehat{\mathrm{Var}}_2(X_i | X_{S_{d',d''}})$, continue to next iteration of outer loop.
   (b) Return $\operatorname{supp}(w_{d'})$.

actually achieves optimal sample complexity $O((1/\kappa^2)\log(n))$ in the setting of attractive GGMs, and always achieves a sample complexity of $O((d/\kappa^2)\log(n))$ which gives a faster algorithm with the same sample complexity as DICE from [140], which has a slower runtime of $O(n^{2d+1})$. (It matches the runtime guarantee for SLICE in [140], which has a worse sample complexity guarantee.)

In Algorithm SearchAndValidate, the key step is performing ℓ_0-constrained regression to predict X_i; the loop in step 2 is required only because we do not know a priori the exact degree of node i, only an upper bound. With high probability, the support of one of the $w_{d'}$ will equal the exact neighborhood of node i, and then a straightforward validation procedure in step 3 (which uses a similar idea to Algorithm DICE in [140]) allows us to identify the correct $w_{d'}$ successfully. For the purposes of the analysis, for every pair of sets $S_0 \subset S$ not containing i define (as in

Theorem 43)

$$T(S_0, S) := \frac{n - |S|}{|S| - |S_0|} \frac{\|\mathbb{X}_i - \mathbb{X}\hat{w}_0\|^2 - \|\mathbb{X}_i - \mathbb{X}\hat{w}\|^2}{\|\mathbb{X}_i - \mathbb{X}\hat{w}\|^2} = \frac{n - |S|}{|S| - |S_0|} \frac{\|\mathbb{X}\hat{w} - \mathbb{X}\hat{w}_0\|^2}{\|\mathbb{X}_i - \mathbb{X}\hat{w}\|^2}$$

where $\hat{w}_0$ is the OLS estimator restricted to $\operatorname{supp}(w_0) \subset S_0$ and $\hat{w}$ is the OLS estimator restricted to $\operatorname{supp}(w) \subset S$.

The following Lemma analyzes the key step in the above algorithm; it shows that when d' equals the true degree of node i, the true support is returned. The crucial part which requires that the GGM is attractive is the application of Lemma 43, which guarantees that candidate supports which are far away from the true neighborhood perform much worse than the true neighborhood. This is crucial because there are many candidate neighborhoods far away from the true neighborhood, which means we need an improved bound to handle them and overcome the cost of taking the union bound.

Lemma 47. *In a κ-nondegenerate attractive GGM, if i is a node of degree d then ℓ_0 constrained regression over vectors with support size at most d returns the true neighborhood of node i with probability at least $1 - \delta$ as long as $m = \Omega(\log(n)/\kappa^2 + \log(2/\delta)/\kappa^2)$.*

Proof. First we consider the randomness over the samples of $X_{\sim i}$, i.e. over $\mathbb{X}$ with column i removed. By Lemma 37 and the union bound over all subsets S of $[n] \setminus \{i\}$ with $|S| \leq 2d$, it holds with probability at least $1 - \delta/2$ that for all w with $w_i = 0$ and $|\operatorname{supp}(w)| \leq 2d$,

$$\frac{1}{2}\mathbb{E}[(w^T X)^2] \leq \frac{1}{2} w^T \left(\frac{1}{m}\mathbb{X}^T\mathbb{X}\right) w \leq \mathbb{E}[(w^T X)^2] \tag{5.8}$$

as long as $m = \Omega(d \log(n) + \log(2/\delta))$. (Recall from Lemma 31 that $d \leq 1/\kappa^2$, so this holds under the hypothesis of the theorem.) We condition on this event and consider the remaining randomness over $\mathbb{X}_i$. Let S^* be the set of true neighbors of node i and let S_0 be any other subset of size at most d. Define $S := S^* \cup S_0$. Since the OLS estimators are defined by projection onto spans of the columns of $\mathbb{X}$, we can apply

the Pythagorean theorem to get

$$\|\mathbb{X}_i - \mathbb{X}\hat{w}_{S^*}\|^2 = \|\mathbb{X}_i - \mathbb{X}w_S\|^2 + \|\mathbb{X}\hat{w}_{S^*} - \mathbb{X}\hat{w}_S\|^2$$

and

$$\|\mathbb{X}_i - \mathbb{X}\hat{w}_{S_0}\|^2 = \|\mathbb{X}_i - \mathbb{X}w_S\|^2 + \|\mathbb{X}\hat{w}_{S_0} - \mathbb{X}\hat{w}_S\|^2.$$

Subtracting, we get that

$$\|\mathbb{X}_i - \mathbb{X}\hat{w}_{S_0}\|^2 - \|\mathbb{X}_i - \mathbb{X}\hat{w}_{S^*}\|^2 = \|\mathbb{X}\hat{w}_{S_0} - \mathbb{X}\hat{w}_S\|^2 - \|\mathbb{X}\hat{w}_{S^*} - \mathbb{X}\hat{w}_S\|^2.$$

To prove the result, it suffices to show with high probability that for any S_0 which does not contain S^* that the leftmost term is positive — then no such S_0 can be the minimizer of the ℓ_0-constrained regression, since S^* corresponds to a feasible point with smaller objective value. We achieve this by showing the right hand side is positive. Observe

$$\|\mathbb{X}\hat{w}_{S_0} - \mathbb{X}\hat{w}_S\|^2 - \|\mathbb{X}\hat{w}_{S^*} - \mathbb{X}\hat{w}_S\|^2 = \frac{d-q}{n-|S|}\|\mathbb{Y} - \mathbb{X}\hat{w}_S\|^2(T(S_0, S) - T(S^*, S)).$$

where $q = |S_0| = |S^*|$ so it suffices to show that $T(S_0, S) - T(S^*, S) \geq 0$. In fact, canceling out denominators, dividing by σ^2 and rearranging it suffices to show

$$\frac{1}{\sigma^2}\|\mathbb{X}\hat{w}_S - \mathbb{X}\hat{w}_{S_0}\|^2 \geq \frac{1}{\sigma^2}\|\mathbb{X}\hat{w}_S - \mathbb{X}\hat{w}_{S^*}\|^2$$

where by Theorem 43 the left hand side is according to $\chi^2_{d-q}(\gamma)$ with $\gamma := \frac{\min_{\text{supp}(w_0)\subset S}\|\mathbb{X}(w_0 - w^*)\|^2}{\sigma^2}$ and the right hand side is distributed according to χ^2_{d-q}, where $\sigma^2 := 1/\Theta_{ii}$. Observe by (5.8) that

$$\gamma \geq \frac{m}{2}\frac{\min_{\text{supp}(w_0)\subset S}\mathbb{E}[(X^T(w_0 - w^*))^2]}{\sigma^2} = \frac{m}{2}\frac{\min_{\text{supp}(w_0)\subset S}\text{Var}(X^T(w_0 - w^*))]}{\sigma^2} \geq \frac{m\kappa^2(d-q)}{2} \tag{5.9}$$

where the last inequality is by Lemma 43, since w_0 is supported on S_0 which is missing

$d-q$ of the neighbors of node i. Applying Lemma 46

$$\Pr(\frac{1}{\sigma^2}\|\mathbb{X}\hat{w}_S - \mathbb{X}\hat{w}_{S_0}\|^2 \leq (d-q+\gamma) - 2\sqrt{(d-q+2\gamma)t}) \leq e^{-t}$$

and applying Lemma 29

$$\Pr(\frac{1}{\sigma^2}\|\mathbb{X}\hat{w}_S - \mathbb{X}\hat{w}_{S^*}\|^2 \geq (d-q) + 2\sqrt{(d-q)t} + 2t) \leq e^{-t}.$$

Letting $t = \log(4dn^{d-q}/\delta)$, and taking the union bound over the at most n^{d-q} possible values of S_0 and then over the at most d possible values of q, we find that with probability at least $1-\delta/2$ for all possible S_0 and q that

$$\begin{aligned}\frac{1}{\sigma^2}\|\mathbb{X}\hat{w}_S - \mathbb{X}\hat{w}_{S_0}\|^2 - \frac{1}{\sigma^2}\|\mathbb{X}\hat{w}_S - \mathbb{X}\hat{w}_{S^*}\|^2 &\geq \gamma - 2\sqrt{(d-q+2\gamma)t} - 2\sqrt{(d-q)t} \\ &\geq \gamma - 4\sqrt{(d-q+2\gamma)t}.\end{aligned}$$

Finally, we see this is nonnegative as long as $\gamma = \Omega(t) = \Omega((d-q)\log(n) + \log(2/\delta))$, which by (5.9) holds as long as $m = \Omega(\frac{\log(n)+\log(2/\delta)}{\kappa^2})$. Therefore the desired result holds with total probability at least $1-\delta$, completing the proof. □

Theorem 44. *Fix $\delta > 0$. In a κ-nondegenerate attractive GGM, as long as $m = \Omega((1/\kappa^2)\log(n) + \log(2/\delta)/\kappa^2)$ it holds with probability at least $1-\delta$ that Algorithm* SearchAndValidate *with $\nu = \kappa^2/2$ returns the true neighborhood of every node i.*

Proof. By applying Lemma 47 and taking the union bound over nodes i, we know that as long as $m = \Omega((1/\kappa^2)\log(n) + \log(2/\delta)/\kappa^2)$ then with probability at least $1-\delta/2$ for every node i, for d' equal to the true degree of node i that $w_{d'}$ returned in step 2 of Algorithm SearchAndValidate is supported on exactly the true neighborhood of node i.

Furthermore, conditioned on the previous event (which only involves sample set 1), it holds with probability at least $1-\delta/2$ by taking the union bound over the possible values of d', d'' that (similar to the pruning argument used in analysis of

Algorithm GreedyAndPrune):

1. in step 3(a).ii, for every d' less than the true degree of node i and for d'' equal to the true degree of node i that the outer loop continues to the next step by applying Lemma 38, Lemma 36, and Lemma 27 and considering any j in the true neighborhood and missing from the support of $w_{d'}$.

2. In step 3 when d' equals the true degree of node i, step 3(b) is reached and the true support of node i is returned by applying Lemma 38 and Lemma 36.

as long as $m = \Omega((d+1/\kappa^2)\log(n)+\log(2/\delta)/\kappa^2)$. Using that $d \leq 1/\kappa^2$ by Lemma 31, we see the requirement on m holds and as desired, the algorithm succeeds with total probability at least $1-\delta$. □

A simplified argument in the general (non-attractive) case, using the weaker bound from Lemma 27 instead of Lemma 43, yields the following result in the general case.

Theorem 45. *Fix $\delta > 0$. In a κ-nondegenerate (not necessarily attractive) GGM with maximum degree d, as long as $m = \Omega((d/\kappa^2)\log(n) + \log(2/\delta)/\kappa^2)$ it holds with probability at least $1-\delta$ that Algorithm* SearchAndValidate *with $\nu = \kappa^2/2$ returns the true neighborhood of every node i.*

5.9 Hybrid ℓ_1 regression guarantees

In the next section, we will discuss algorithms for regression and structure learning in general walk-summable models. Since (as we will see) the conditional variance is not supermodular in these models, we need some fundamentally new tools to analyze this setting. It turns out that we will need to analyze a variant of ℓ_1-constrained least squares regression, which we do in this section as preparation.

Definition 28. *We define the* hybrid ℓ_1-regression model *to be given by*

$$Y = \langle w^*, X - \mathbb{E}[X|Z]\rangle + a^* Z + \xi$$

where $\|w\|_1 \leq W$ *and conditioned on* Z, $X - \mathbb{E}[X|Z] \sim N(0, \Sigma)$ *with* $\Sigma : n \times n$, $\Sigma_{ii} \leq R^2$ *for all* i, $\mathbb{E}Z^2 = 1$ *(w.l.o.g.), and* $\mathbb{E}\xi^2 = \sigma^2$ *with the noise* ξ *independent of* X, Z.

The corresponding function class is

$$\begin{aligned}\mathcal{F} &:= \{(x,z) \mapsto \langle w, x - \mathbb{E}[X|Z=z]\rangle + az : \|w\|_1 \leq W\} \\ &= \{(x,z) \mapsto \langle w, x\rangle + a'z : \|w\|_1 \leq W\}.\end{aligned}$$

and the *Empirical Risk Minimizer* (ERM) is given by taking the minimizer of

$$\min_{\|w\|_1 \leq W, a'} \hat{\mathbb{E}}[(Y - \langle w, X\rangle - a'Z)^2].$$

As mentioned in the introduction, it will be crucial in the analysis to look at the parameterization with a instead of a' even though algorithmically the ERM will be computed using the variable a' (as the change of basis given by subtracting off the conditional expectations is unknown and could only be approximated from data).

5.9.1 Guarantees for Empirical Risk Minimization (ERM)

There is a vast literature on generalization bounds for empirical risk minimization (and natural variants) using tools such as (local) Rademacher complexity, stability, etc. (see e.g. [14, 174, 165] and many related references). In the present context, many of these methods are not well-optimized because the noise and covariates are drawn from unbounded distributions and the squared loss is not uniformly Lipschitz (see e.g. the discussion in [136]). Fortunately, the framework developed in [136] avoids these issues and we are able to use it directly to give a good bound on the excess risk of the empirical risk minimizer.

Background: Learning without Concentration Framework

We recall the main result of [136]. In this framework, as with many results in statistical learning, the empirical process is controlled by a localized measure of complexity:

more precisely, fixpoints of local Rademacher averages defined below. See e.g. [14] for more context. In the present context, this is important both to get a better rate for the "realizable" part of the generalization bound (the term which doesn't depend on the noise level σ, which dominates in the realizable setting $\sigma = 0$), and also to handle the fact that the class $\mathcal{F}$ we consider is unbounded.

Let $\mathcal{F}$ be a class of (measurable) functions. Let X, Y be arbitrary random variables, suppose that f^* is a minimizer of $\mathbb{E}[(Y - f(X))^2]$ over $f \in \mathcal{F}$ (which we assume exists) and define $\xi := Y - f^*(X)$. Let $\|f\|_{L_2} = \sqrt{\mathbb{E}[f^2]}$ and let $D_2(f)$ be the L_2 ball of radius 1 around f, i.e. $D_2(f) = \{g : \mathbb{E}[(g(X) - f(X))^2] = 1\}$. The following two quantities, defined by fixed point equations, appear in the generalization bound: the intrinsic parameter (which does not depend on the noise model)

$$\beta_m^*(\gamma) = \inf \left\{ r > 0 : \mathbb{E} \sup_{f \in \mathcal{F} \cap r D_{f^*}} \left| \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \epsilon_i (f - f^*)(X_i) \right| \leq \gamma r \sqrt{m} \right\}$$

and the noise-sensitive/extrinsic parameter

$$\alpha_m^*(\gamma, \delta) = \inf \left\{ s > 0 : \Pr \left(\sup_{f \in \mathcal{F} \cap s D_{f^*}} \left| \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \epsilon_i \xi_i (f - f^*)(X_i) \right| \leq \gamma s^2 \sqrt{m} \right) \geq 1 - \delta \right\}.$$

Theorem 46 (Theorem 3.1, [136])**.** *Suppose $\mathcal{F}$ is a closed, convex class of functions and $f^*, X, Y, \alpha^*, \beta^*$ are defined as above. Let $\tau > 0$, define*

$$q := \inf_{f \in \mathcal{F} - \mathcal{F}} \Pr(|f| \geq 2\tau \|f\|_{L_2})$$

and assume that $q > 0$ (this is called the small-ball condition*). Then for any $\gamma < \tau^2 q/16$ and for every $\delta > 0$ it holds that for any $\hat{f}$ which is an empirical risk minimizer for i.i.d. samples $\{(X^{(i)}, Y^{(i)})\}_{i=1}^m$,*

$$\|\hat{f} - f^*\|_{L_2} \leq 2 \max \{\alpha_m^*(\gamma, \delta/4), \beta_m^*(\tau q/16)\}$$

with probability at least $1 - \delta - e^{-mq/2}$.

ERM Risk Bound

We return to the specific setting of hybrid ℓ_1-constrained regression and prove our desired bound.

Theorem 47. *As long as $m = \Omega(\log(n/\delta))$, if $\hat{w}, \hat{a}'$ is the empirical risk minimizer for hybrid L1 regression from m i.i.d. samples then*

$$\mathbb{E}[(\mathbb{E}[Y|X,Z] - \langle \hat{w}, X\rangle - \hat{a}'Z)^2] = O\left(RW\sigma\sqrt{\frac{\log(2n/\delta)}{m}} + \frac{\sigma^2\log(4/\delta)}{m} + \frac{R^2W^2\log(n)}{m}\right)$$

with probability at least $1-\delta$.

Proof. We first deal with the small-ball condition. Let $\tau = 1/2$. Observe that for any $f_1, f_2 \in \mathcal{F}$ that $f_1(X,Z) - f_2(X,Z)$ has a univariate Gaussian distribution, therefore

$$q := \Pr(|f| \geq 2\tau\|f\|_{L_2}) = 1 - \frac{1}{\sqrt{2\pi}}\int_{-2\tau}^{2\tau} e^{-x^2/2}dx \geq 1/4.$$

We take $\gamma = 1/300 < \tau^2 q/32$.

We now bound β^*. We have

$$\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{F}\cap rD_{f^*}}\left|\frac{1}{\sqrt{m}}\sum_{i=1}^m \epsilon_i(f-f^*)(X_i)\right| \\
&= \mathbb{E}\sup_{f\in\mathcal{F}\cap rD_{f^*}}\left|\frac{1}{\sqrt{m}}\sum_{i=1}^m \epsilon_i(\langle w-w^*, X_i - \mathbb{E}[X_i|Z_i]\rangle + (a-a^*)Z)\right| \\
&\leq 2RW\mathbb{E}\left\|\frac{1}{\sqrt{m}}\sum_{i=1}^n \epsilon_i\frac{X_i - \mathbb{E}[X_i|Z_i]}{W}\right\|_\infty + \sup_{f\in\mathcal{F}\cap rD_{f^*}}|a-a^*|\mathbb{E}|Z| \\
&\leq C(RW\sqrt{\log(n)} + \sup_{f\in\mathcal{F}\cap rD_{f^*}}|a-a^*|)
\end{aligned}$$

where the first inequality is by Holder's inequality and the triangle inequality, and the second is by the standard Gaussian tail bound combined with the union bound. To complete the bound observe that

$$\mathbb{E}[(\langle w-w^*, X - \mathbb{E}[X|Z]\rangle + (a-a^*)Z)^2] \geq (a-a^*)^2$$

so $a - a^* \leq r$ and

$$\mathbb{E} \sup_{f \in \mathcal{F} \cap r D_{f^*}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{m} \epsilon_i (f - f^*)(X_i) \right| \leq C(RW\sqrt{\log(n)} + r).$$

This is smaller than $\gamma r \sqrt{m}$ as long as $r = \Omega(\frac{RW}{\gamma}\sqrt{\frac{\log n}{m}})$ so $\beta_m^* = O(\frac{RW}{\gamma}\sqrt{\frac{\log n}{m}})$.

We proceed to bound α^* similarly.

$$\begin{aligned}
&\sup_{f \in \mathcal{F} \cap s D_{f^*}} \left| \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \epsilon_i \xi_i (f - f^*)(X_i) \right| \\
&\quad = \sup_{f \in \mathcal{F} \cap s D_{f^*}} \left| \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \epsilon_i \xi_i (\langle w - w^*, X_i - \mathbb{E}[X_i | Z_i] \rangle + (a - a^*) Z) \right| \\
&\quad \leq C(RW\sigma\sqrt{\log(2n/\delta)} + \sigma s \sqrt{\log(4/\delta)})
\end{aligned}$$

with probability at least $1-\delta$ as long as $m \geq m_1 = O(\log(n/\delta))$, where the inequality is by Holder's inequality and $|a - a^*| \leq s$ (as before), Bernstein's inequality (Theorem 2.8.2 of [188]) using that the product of sub-Gaussian r.v. (ξ_i and $X_i - \mathbb{E}[X_i|Z_i]$) is sub-exponential (Lemma 2.7.7 of [188]), and the union bound. The last quantity is upper bounded by $\gamma s^2 \sqrt{m}$ as long as $s^2 = \Omega(\frac{\sigma}{\gamma}\sqrt{\frac{\log(2n/\delta)}{m}})$ and $s = \Omega(\frac{\sigma}{\gamma}\sqrt{\frac{\log(4/\delta)}{m}})$. Therefore

$$(\alpha^*)^2 = O\left(\frac{RW\sigma}{\gamma} \sqrt{\frac{\log(2n/\delta)}{m}} + \frac{\sigma^2 \log(4/\delta)}{\gamma^2 m} \right).$$

Combining our estimates, it follows from Theorem 46 that

$$\mathbb{E}[(\hat{f} - f^*)^2] = O((\alpha_m^*)^2 + (\beta_m^*)^2) = O\left(\frac{RW\sigma}{\gamma} \sqrt{\frac{\log(2n/\delta)}{m}} + \frac{\sigma^2 \log(4/\delta)}{\gamma^2 m} + \frac{R^2 W^2 \log(n)}{\gamma m} \right)$$

with probability at least $1 - \delta - e^{-m/8} \geq 1 - 2\delta$ as long as $m = \Omega(\log(1/\delta) + m_1) = \Omega(\log(d/\delta))$. Since γ is just a constant, this gives the result. □

5.9.2 Guarantees for Greedy Methods

In this section we show that a simple greedy method can also solve this high-dimensional regression problem with the correct dependence on n, albeit with slightly

worse dependence on the other parameters. This is conceptually important as it shows that examples breaking greedy algorithms (in the sense of requiring $\omega(\log(n))$ sample complexity) also suffice to break analyses based on bounded ℓ_1-norm.

Lemma 48. *In the hybrid ℓ_1-regression model, there exists an input coordinate j such that*

$$\mathrm{Var}(\mathbb{E}[Y|X,Z] \mid Z, X_j) \leq \mathrm{Var}(\mathbb{E}[Y|X,Z] \mid Z)\left(1 - \frac{\mathrm{Var}(\mathbb{E}[Y|X,Z] \mid Z)}{R^2W^2}\right).$$

Proof. By expanding, applying Holder's inequality and using the assumption on R we have

$$\begin{aligned}
\mathrm{Var}(\mathbb{E}[Y|X,Z] \mid Z) &= \sum_j w_j \mathrm{Cov}(\mathbb{E}[Y|X,Z], X_j \mid Z) \\
&\leq W \max_j |\mathrm{Cov}(\mathbb{E}[Y|X,Z], X_j \mid Z)| \\
&\leq RW \max_j \left|\mathrm{Cov}\left(\mathbb{E}[Y|X,Z], \frac{X_j}{\sqrt{\mathrm{Var}(X_j|Z)}} \;\middle|\; Z\right)\right|.
\end{aligned}$$

Let j be the maximizer. Then by Lemma 28,

$$\mathrm{Var}(\mathbb{E}[Y|X,Z] \mid Z) - \mathrm{Var}(\mathbb{E}[Y|X,Z] \mid Z, X_j) = \frac{\mathrm{Cov}(\mathbb{E}[Y|X,Z], X_j \mid Z)^2}{\mathrm{Var}(X_j \mid Z)} \geq \frac{\mathrm{Var}(\mathbb{E}[Y|X,Z] \mid Z)^2}{R^2W^2}.$$

Rearranging gives that

$$\mathrm{Var}(\mathbb{E}[Y|X,Z] \mid Z, X_j) \leq \mathrm{Var}(\mathbb{E}[Y|X,Z] \mid Z)\left(1 - \frac{\mathrm{Var}(\mathbb{E}[Y|X,Z] \mid Z)}{R^2W^2}\right).$$

□

The above bound naturally leads to analyzing the recursion $x \mapsto x - cx^2$, which we do in the next Lemma.

Lemma 49. *Suppose that $x_1 \leq 1/2c$ and $x_{t+1} \leq (1 - cx_t)x_t$ for some $c < 1$. Then*

$$x_t \leq \frac{1}{c(t+1)}$$

Proof. We prove this by induction. Observe that $x(1-cx)$ is an increasing function in x for $x \leq \frac{1}{2c}$ since $1/2c$ corresponds to the vertex of the parabola, so using the assumption and the induction hypothesis,

$$x_t \leq x_{t-1}(1-cx_{t-1}) \leq 1/ct - 1/ct^2 = \frac{t-1}{ct^2} \leq \frac{t-1}{c(t^2-1)} \leq \frac{1}{c(t+1)}.$$

□

Lemma 50. *In the hybrid ℓ_1-regression model,*

$$\mathrm{Var}(\mathbb{E}[Y|X,Z] \mid Z) \leq R^2W^2.$$

Proof. By expanding, using Holder's inequality and Cauchy-Schwartz

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[Y|X,Z] \mid Z) &= \sum_j w_J \mathrm{Cov}(\mathbb{E}[Y|X,Z], X_j \mid Z) \\ &\leq W \max_j |\mathrm{Cov}(\mathbb{E}[Y|X,Z], X_j \mid Z) \\ &\leq W \max_j \sqrt{\mathrm{Var}(\mathbb{E}[Y|X,Z] \mid Z)\mathrm{Var}(X_j \mid Z)} \leq RW\sqrt{\mathrm{Var}(\mathbb{E}[Y|X,Z] \mid Z)}\end{aligned}$$

so $\mathrm{Var}(\mathbb{E}[Y|X,Z] \mid Z) \leq R^2W^2$. □

Remark 12 (Connection to Approximate Caratheodory)**.** *From the previous two lemmas, we can give a "matching pursuit" proof of the approximate Caratheodory theorem, which says that vectors of bounded ℓ_1-norm are well approximated in ℓ_2 by sparse vectors [188]. The standard proof of this result is probabilistic. Another proof, in a similar spirit, is given by using the guarantees of the Frank-Wolfe algorithm (see [36]).*

The remaining task is to analyze the behavior of the iteration under noise, which gives the main result:

Theorem 48. *For any $\epsilon \in (0,1)$, iterate t of OMP in the hybrid regression model satisfies*

$$\mathrm{Var}(\mathbb{E}[Y|X,Z]|Z, X_{S_t}) \leq \epsilon\sigma^2$$

as long as $t = \Omega(R^2W^2/\epsilon\sigma^2)$ *and* $m = \Omega(\frac{R^2W^2}{\epsilon^2\sigma^2}(t\log(4n) + \log(4/\delta)))$.

Proof. The argument is structured similarly to the proof of Lemma 44. Fix $\epsilon \in (0,1)$ to be optimized later: we bound the number of steps of OMP during which $\mathrm{Var}(\mathbb{E}[Y|X,Z]|Z,X_{S_t}) \geq \epsilon\sigma^2$. Note that once this bounds holds for some t, it holds for all larger t by the law of total variance. Fix an integer $T > 0$ to be optimized later.

First observe from Lemma 48 (applied after conditioning out X_{S_t}) that there exists a node j^* such that

$$\mathrm{Var}(\mathbb{E}[Y|X,Z]|Z,X_{j^*},X_{S_t}) \leq \mathrm{Var}(\mathbb{E}[Y|X,Z]|Z,X_{S_t})(1 - \frac{\mathrm{Var}(\mathbb{E}[Y|X,Z]|Z,X_{S_t})}{R^2W^2}).$$

From Lemma 39 and taking the union bound over all sets S of size $|S| \leq T$ we have

$$\begin{aligned}\left|\sqrt{\frac{1}{\mathrm{Var}(Y|X_{S\setminus j}) - \sigma^2}}\frac{|\hat{w}_j|}{\sqrt{(\hat{\Sigma}^{-1})_{jj}}} - \sqrt{\gamma'}\right| &\leq \sqrt{\frac{\mathrm{Var}(Y|X_S)}{\mathrm{Var}(Y|X_{S\setminus j}) - \sigma^2}\cdot\frac{2\log(n^T/\delta)}{m}} + \sqrt{\frac{\gamma'}{64}} \\ &\leq \sqrt{\frac{1+\epsilon}{\epsilon}\cdot\frac{2\log(n^T/\delta)}{m}} + \sqrt{\frac{\gamma'}{64}}\end{aligned}$$

using that $(1+x)/x = 1/x + 1$ is monotone decreasing, where

$$\gamma' = \gamma'(S,j) := \frac{\mathrm{Var}(X_i|Z,X_{S\setminus\{j\}}) - \mathrm{Var}(X_i|Z,X_S)}{\mathrm{Var}(X_i|Z,X_{S\setminus\{j\}}) - \sigma^2}.$$

Note that $\gamma'(S,j^*) \geq \epsilon\sigma^2/R^2W^2$. Therefore as long as $m = \Omega(\frac{R^2W^2}{\epsilon^2\sigma^2}(T\log(4n) + \log(4/\delta)))$ then OMP chooses a node j s.t.

$$\mathrm{Var}(\mathbb{E}[Y|X,Z]|Z,X_j,X_{S_t}) \leq \mathrm{Var}(\mathbb{E}[Y|X,Z]|Z,X_{S_t})(1 - \frac{\mathrm{Var}(\mathbb{E}[Y|X,Z]|Z,X_{S_t})}{2R^2W^2})$$

as long as $|S_t| \leq T$. Applying Lemma 50 and Lemma 49 we find that

$$\mathrm{Var}(\mathbb{E}[Y|X,Z]|Z,X_{S_t}) \leq \frac{2R^2W^2}{t+1}$$

for $t \leq T$. Therefore if $T \geq t \geq 2R^2W^2/\epsilon\sigma^2$ we are guaranteed that

$\mathrm{Var}(\mathbb{E}[Y|X,Z]|Z,X_{S_t}) \leq \epsilon\sigma^2$. Taking $\epsilon = 2R^2W^2/T\sigma^2$ gives the result. □

5.10 Regression and Structure Learning in Walk-Summable Models

5.10.1 Failure of (weak) supermodularity in SDD models

The following example shows that the conditional variance is not supermodular in the SDD case, unlike in the attractive/ferromagnetic case.

Example 9. *Consider the GGM given by SDD precision matrix*

$$\Theta = \begin{bmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & 1/2 \\ -1/2 & 1/2 & 1 \end{bmatrix}$$

and label the nodes (in order) by i, j, k. One can see (e.g. by computing effective resistances in the lifted graph) that $2\mathrm{Var}(X_i) = 3$, that $2\mathrm{Var}(X_i|X_j) = 2\mathrm{Var}(X_i|X_k) = 8/3$, and $2\mathrm{Var}(X_i|X_j, X_k) = 2$. Since $3 - 8/3 = 1/3 < 2/3 = 8/3 - 2$ this violates supermodularity.

The above example alone does not rule out the possibility that (negative) conditional variances in SDD models always have *submodularity ratio* introduced by [48] lower bounded by a constant. We recall the definition next:

Definition 29 ([48])**.** *The* submodularity ratio $\gamma(k)$ *of a function on subsets of a universe U, $f : 2^U \to \mathbb{R}_{\geq 0}$ is defined to be*

$$\gamma(k) := \min_{L \subset U, |S| \leq k, L \cap S = \emptyset} \frac{\sum_{x \in S} f(L \cup \{x\}) - f(L)}{f(L \cup S) - f(L)}$$

Note that $\gamma(k) \geq 1$ for a submodular function.

The significance of this ratio for a function f is that if the ratio is lower bounded by a constant then similar guarantees for submodular maximization follow ([48]); for

this reason such an f is sometimes called *weakly submodular* (as in e.g. [61]). Now, we give a counterexample showing that for general SDD matrices, this ratio can be arbitrarily small.

Example 10. *Fix $M > 0$ large. Let $\epsilon > 0$ be a parameter to be taken small, and consider the following precision matrix, which is SDD as long as $\epsilon < 1/2 < M$:*

$$\Theta = \begin{bmatrix} 1 & -\epsilon & \epsilon \\ -\epsilon & M & \epsilon - M \\ \epsilon & \epsilon - M & M \end{bmatrix}.$$

This has inverse Θ^{-1} equaling

$$\begin{bmatrix} (\epsilon - 2M)/(\epsilon + 2\epsilon^2 - 2M) & -(\epsilon/(\epsilon + 2\epsilon^2 - 2M)) & \epsilon/(\epsilon + 2\epsilon^2 - 2M) \\ -(\epsilon/(\epsilon + 2\epsilon^2 - 2M)) & (\epsilon^2 - M)/(\epsilon^2 + 2\epsilon^3 - 2\epsilon M) & (\epsilon + \epsilon^2 - M)/(\epsilon^2 + 2\epsilon^3 - 2\epsilon M) \\ \epsilon/(\epsilon + 2\epsilon^2 - 2M) & (\epsilon + \epsilon^2 - M)/(\epsilon^2 + 2\epsilon^3 - 2\epsilon M) & (\epsilon^2 - M)/(\epsilon^2 + 2\epsilon^3 - 2\epsilon M) \end{bmatrix}$$

so

$$\mathrm{Var}(X_1) - \frac{1}{\Theta_{11}} = \frac{-2\epsilon^2}{\epsilon + 2\epsilon^2 - 2M}$$

and (by computing the inverse of the top-left 2x2 submatrix of Θ) we find

$$\mathrm{Var}(X_1|X_3) - \frac{1}{\Theta_{11}} = \frac{M}{M - \epsilon^2} - 1 = \frac{\epsilon^2}{M - \epsilon^2}$$

and the difference is

$$\mathrm{Var}(X_1) - \mathrm{Var}(X_1|X_3) = \frac{\epsilon^3}{(M - \epsilon^2)(2M - 2\epsilon^2 - \epsilon)}$$

Therefore the submodularity ratio $\gamma = \gamma(2)$ *for $f(S) = \mathrm{Var}(X_1) - \mathrm{Var}(X_1|X_S)$ is upper bounded by (taking $L = \emptyset$)*

$$\gamma \leq \frac{f(\{2\}) + f(\{3\})}{f(\{2,3\})} = \Theta\left(\frac{\epsilon^3/M^2}{\epsilon^2/M}\right) = \Theta(\epsilon/M)$$

which is clearly arbitrarily small.

Remark 13 (Submodularity ratio and κ)**.** *It's possible to show, based on Lemma 48 and the bounds in the proof of Theorem 49 to derive a partial lower bound for the submodularity ratio when we consider $S \subset T$ and restrict to j which are neighbors of i, by showing:*

$$f(S \cup \{j\}) - f(S) \geq \frac{\kappa^2}{4d}(f(U) - f(S)) \geq \frac{\kappa^2}{4d}(f(T \cup \{j\}) - f(T))$$

using the monotonicity of f (which follows from the law of total variance) in the last step, and under the assumption that the model is κ-nondegenerate and d-sparse. The above example shows that this dependence on κ is tight: by taking a fixed small ϵ and sending $M \to \infty$, the submodularity ratio can be as small as $O(\kappa^2)$ since $\kappa = \epsilon/\sqrt{M}$ in this model. It remains unclear if the submodularity ratio can be lower bounded in general in κ-nondegenerate models; even if such a bound did hold it could not be used to prove Theorem 49 since that result holds without a κ-nondegeneracy assumption.

5.10.2 Sparse regression

In this section we describe an algorithm to find a good predictor of node X_i with bounded degree d in a walk-summable GGM. To simplify the analysis, we assume the data has been split into 3 equally sized sample sets, each of size m; when there is no explicit mention, averages are taken over sample set 1.

The algorithm is conceptually straightforward: it does a single greedy step and then sets up an ℓ_1-constrained regression. The only complication is that we do not know $1/\Theta_{ii}$ a priori, but this appears in the ℓ_1-norm of the obvious regression we want to setup. Since we have multiplicative estimates for $1/\Theta_{ii}$, we can deal with this by searching over the possible values on a log scale.

We show this algorithm gives a result for sparse linear regression under the walk-summability assumption which (1) depends on sparsity only, not on norms (unlike the slow rate bound for LASSO) and (2) is computationally efficient (unlike brute force ℓ_0-constrained regression).

Algorithm WS-Regression(γ, d):

1. Choose j to minimize $\widehat{\mathrm{Var}}(X_i|X_j)$.
2. Let $s_0^2 := \exp(\lfloor \log(\widehat{\mathrm{Var}}(X_i|X_j)/8d) \rfloor - 1)$.
3. For ℓ in 0 to $\lceil \log(8d) + 3 \rceil$:
   (a) Let $s_\ell^2 := s_0 e^\ell$
   (b) Solve for w, a in

$$\min_{w,a:\|w\|_1 \le \lambda} \hat{\mathbb{E}}_2 \left[\left(X_i - \sum_{k \notin \{i,j\}} w_k \frac{X_k}{\sqrt{\widehat{\mathrm{Var}}(X_k|X_j)}} - aX_j \right)^2 \right]$$

   where $\lambda = \sqrt{2d} s_\ell$ and $\hat{\mathbb{E}}_2$ is empirical expectation over sample set 2.
   (c) Let $\hat{\sigma}^2 := \hat{\mathbb{E}}_3 \left[\left(X_i - \sum_{k \notin \{i,j\}} w_k \frac{X_k}{\sqrt{\widehat{\mathrm{Var}}(X_k|X_j)}} - aX_j \right)^2 \right]$ where $\hat{\mathbb{E}}_3$ is empirical expectation over sample set 3. If $\lambda^2 \ge 2d\gamma^2\hat{\sigma}^2$ (equivalently, $s_\ell^2 \ge \gamma^2\hat{\sigma}^2$), then exit the loop.
4. Return $w, a, j, \hat{\sigma}^2$.

Theorem 49. *Let i be a node of degree d in an SDD GGM and $\sigma^2 := 1/\Theta_{ii}$. Then WS-Regression($\gamma$) with $\gamma^2 = 2$ returns w, a such that*

$$\mathbb{E}\left[\left(\mathbb{E}[X_i|X_{\sim i}] - \sum_{k\notin\{i,j\}} w_k \frac{X_k}{\sqrt{\widehat{\mathrm{Var}}(X_k|X_j)}} - aX_j\right)^2\right] = O\left(\sigma^2\sqrt{\frac{d\log(2n/\delta)}{m}}\right)$$

and $\hat{\sigma}^2$ s.t. $1/2 \le \Theta_{ii}\hat{\sigma}^2 \le 2$ with probability at least $1-\delta$, as long as $m = \Omega(\log(n/\delta))$.

Proof. By Lemma 33, for any $k \sim i$ we have $\mathrm{Var}(X_i|X_j) \le 1/|\Theta_{ik}|$ therefore if we take j^* which minimizes $\mathrm{Var}(X_i|X_{j^*})$ then

$$\mathrm{Var}(X_i|X_{j^*}) \le 1/|\Theta_{ij}|$$

for all j. Similarly, applying Lemma 34 we know that

$$\mathrm{Var}(X_i|X_{j^*}) \le \frac{4d}{\Theta_{ii}}$$

By using Lemma 35 and taking the union bound over the randomness of sample set 1, we may assume that for every j, k, $\mathrm{Var}(X_k|X_j)/\sqrt{2} \le \widehat{\mathrm{Var}}(X_k|X_j) \le \sqrt{2}\mathrm{Var}(X_k|X_j)$, with probability at least $1 - \delta/3$ as long as $m = \Omega(\log(n/\delta))$. We condition on this event. Then for the j chosen in step 1 of the algorithm, we have that

$$\mathrm{Var}(X_i|X_j) \le \sqrt{2}\widehat{\mathrm{Var}}(X_i|X_j) \le \sqrt{2}\widehat{\mathrm{Var}}(X_i|X_{j^*}) \le 2\mathrm{Var}(X_i|X_{j^*}) \le 2/|\Theta_{ik}|$$

for all $i \sim k$, and similarly

$$\mathrm{Var}(X_i|X_j) \le \frac{8d}{\Theta_{ii}}. \tag{5.10}$$

Furthermore,

$$\mathrm{Var}\left(\frac{X_k}{\sqrt{\widehat{\mathrm{Var}}(X_k|X_j)}}\middle| X_j\right) \le \sqrt{2}$$

and

$$\sum_k \frac{|\Theta_{ik}|}{\Theta_{ii}}\sqrt{\widehat{\mathrm{Var}}(X_k|X_j)} \le \sum_k \frac{|\Theta_{ik}|}{\Theta_{ii}}\sqrt{2\mathrm{Var}(X_k|X_j)}$$
$$\le \sum_k \frac{|\Theta_{ik}|}{\Theta_{ii}}\sqrt{2(1/|\Theta_{ik}| + \mathrm{Var}(X_i|X_j))}$$
$$\le \sum_k \frac{|\Theta_{ik}|}{\Theta_{ii}}\sqrt{2(3/|\Theta_{ik}|)} = \frac{\sqrt{6}}{\Theta_{ii}}\sum_k \sqrt{|\Theta_{ik}|} \le \sqrt{6d/\Theta_{ii}}$$

using Lemma 32 in the second inequality and Cauchy-Schwartz and the SDD property in the final inequality. Given (5.10) we know that for one of the values of ℓ satisfies $e/\Theta_{ii} \le s_\ell^2 \le e^2/\Theta_{ii}$; call this ℓ^*. By Theorem 47 we have that with probability at least $1-\delta/3$ that for all of the loop iterations where $1/\Theta_{ii} \le s_\ell^2$ (so the global optimal w^*, a is in the constraint set) and $\ell \le \ell^*$

$$\mathbb{E}\left[\left(X_i - \sum_{k\notin\{i,j\}} w_k \frac{X_k}{\sqrt{\mathrm{Var}(X_k|X_j)}} - aX_j\right)^2\right] = O\left(\sqrt{1/\Theta_{ii}}\sqrt{24d/\Theta_{ii}}\sqrt{2}\sqrt{\frac{\log(n^2/\delta)}{m}}\right) \tag{5.11}$$

as long as $m = \Omega(\log(n/\delta))$, using that $d \le n$ in the union bound. Condition on this and consider only the randomness over sample set 3. By Bernstein's inequality and the union bound over the loop iterations, with probability at least $1-\delta/3$ as long as $m = \Omega(\log(n/\delta))$, for the above value of $\ell = \ell^*$ we have that the test in 3(c) succeeds and the loop exits, and that if the loop exited in a previous iteration then $\frac{1}{\Theta_{ii}} = \mathrm{Var}(X_i|X_{\sim i}) \le s_\ell^2$ so we can apply the above guarantee (5.11), giving the result. □

5.10.3 Structure learning

Theorem 50. *For an SDD, κ-nondegenerate GGM the following is true. Algorithm* HybridMB *with $\tau = \kappa^2/8, \gamma = 2$ returns the true neighborhood of every node i with probability at least $1-\delta$ as long as $m \ge m_1'$, where $m_1' = O((d/\kappa^4)\log(n/\delta))$ where d is the max degree in the graph.*

Proof. By Theorem 49 and the union bound, we may assume with probability at least

Algorithm HybridMB(τ, γ, d):

1. We suppose the samples are split into 3 equally sized sets as in the description of WS-Regression.
2. For every node i, apply WS-Regression which returns $w(i), a(i), j(i), \hat{\sigma}^2(i)$.
3. Define $u(i)_{j(i)} = a(i)$ and $u(i)_k = \frac{w(i)_k}{\sqrt{\widehat{\mathrm{Var}}(X_k|X_j)}}$.
4. Let $E = \{\}$.
5. For every pair of nodes a, b:
   (a) If $u(a)_b^2 \hat{\sigma}^2(b) \geq \tau \hat{\sigma}^2(a)$ and $u(b)_a^2 \hat{\sigma}^2(a) \geq \tau \hat{\sigma}^2(b)$: add (i, j) to E.
6. Return edge set E.

$1 - \delta$, as long as $m = \Omega((d/\kappa^4)\log(n/\delta))$ that for every node i we have $u(i)$ such that

$$\mathbb{E}\left[\left(\mathbb{E}[X_i|X_{\sim i}] - \sum_{k \neq i} u(k) X_k\right)^2\right] \leq \frac{\kappa^2}{16\Theta_{ii}}$$

and $\hat{\sigma}^2(i)$ which is within a factor of 2 of $1/\Theta_{ii}$. Applying the law of total variance and the conditional law of a single variable in the GGM, we find that

$$\left(\frac{u(k)}{\sqrt{\Theta_{kk}}} + \frac{\Theta_{ik}}{\Theta_{ii}\sqrt{\Theta_{kk}}}\right)^2 = \left(u(k) + \frac{\Theta_{ik}}{\Theta_{ii}}\right)^2 \mathrm{Var}(X_k|X_{\sim k}) \leq \frac{\kappa^2}{64\Theta_{ii}}$$

so if i and k are not neighbors, then $\Theta_{ik} = 0$ so

$$u(k)^2 \hat{\sigma}^2(k) \leq 2u(k)^2/\Theta_{kk} \leq \frac{\kappa^2 \hat{\sigma}_i^2}{16}$$

and if they are then $|\Theta_{ik}| \geq \kappa\sqrt{\Theta_{ii}\Theta_{kk}}$ so using the reverse triangle inequality

$$\begin{aligned} u(k)^2 \hat{\sigma}^2(k) \geq (1/2)u(k)^2/\Theta_{kk} \geq (1/2)(\kappa^2/\sqrt{\Theta_{ii}} - \kappa/8\sqrt{\Theta_{ii}}) &\geq (7/16)\kappa^2/\sqrt{\Theta_{ii}} \\ &\geq (7/32)\kappa^2 \hat{\sigma}^2(i). \end{aligned}$$

From these inequalities we see that in step 5 (a) exactly the correct edges are chosen.

□

Theorem 51. *For any SDD, κ-nondegenerate GGM the following is true. Algorithm* GreedyAndPrune *with $\tau = \kappa^2/8$ and $T = \Theta(d/\kappa^2)$ returns the true neighborhood of every node i with probability at least $1 - \delta$ as long as $m = \Omega((d^2/\kappa^6)\log(n/\delta))$ where d is the max degree in the graph.*

Proof. The proof is the same as for Theorem 50 except that we use Theorem 48 instead of Theorem 47, and use the slightly different pruning analysis from the proof of Theorem 40. □

Remark 14 (Implementation)**.** *In experiments, to reduce the number of free parameters in* HybridMB *we define $\gamma' = 2d\gamma^2$ and note that using γ' instead of γ actually allows d to be eliminated as a parameter. We also use a single sample set instead of sample splitting; we expect that the algorithm can still be proved correct without the splitting, at the cost of a more lengthy analysis.*

For completeness, we state a result for HybridMB under the ℓ_1-bounded assumption used in previous work like [39, 38]. The proof follows the proof of our main result, except that we can ignore the analysis of the first greedy step and simply use the a priori estimate for the ℓ_1 norm, which only shrinks under conditioning.

Theorem 52. *For any κ-nondegenerate GGM with precision matrix $\Theta : n \times n$ such that $\max_i \sum_{j=1}^{n} |\Theta_{ij}| \leq M$, Algorithm* HybridMB *with $\tau = O(\kappa^2)$ returns the true neighborhood of every node i with probability at least $1 - \delta$ as long as $m = \Omega(M^2 \log(n/\delta)/\kappa^4)$.*

This guarantee matches [38], which itself improves on the guarantee in [39]. In the same setting, GreedyAndPrune achieves a sample complexity of $O(\frac{M^4 \log(n/\delta)}{\kappa^6})$.

5.11 Simulations and Experiments

In this section, we will compare our proposed method (GreedyAndPrune) with popular methods previously introduced in the literature: the Graphical Lasso [67], the Meinhausen-Bühlmann estimator (based on the LASSO) [135], CLIME [39], and

ACLIME [38] (an adaptive version of CLIME). In the first subsection, we consider simple attractive GGMS and show that our method always performs well compared to previous methods and sometimes outperforms them considerably. In the second subsection, we compare the performance on a real dataset (from [37]) and show that our methods HybridMB and GreedyAndPrune again compare favorably. Our experiment also gives evidence that walk-summability is a reasonable assumption in practice.

5.11.1 Simple attractive GGMs where previous methods perform poorly

Three of the most popular methods for recovering a sparse precision matrix in practice are the Graphical Lasso (glasso) [67], the Meinhausen-Bülhmann estimator (MB) based on the Lasso [135], and the CLIME estimator [39]. The graphical lasso is the ℓ_1-penalized variant of the MLE (Maximum Likelihood Estimator) for the covariance matrix; CLIME minimizes the ℓ_1-norm of the recovered precision matrix $\hat{\Theta}$, given an ℓ_∞ constraint $|\Sigma\Omega - Id|_\infty \leq \lambda$ (where $|M|_\infty = \|M\|_{1\to\infty}$ is the entrywise max-norm). For Meinhausen-Bühlmann, we let the estimated $\hat{\Theta}$ have its rows be given by the appropriate lasso estimate, scaled appropriately by the corresponding estimate for the conditional variance. The current theoretical guarantees of these methods have very high sample complexity for general GFFs and we find simple examples in which the scaling of their sample complexity with n is poor. One example (which breaks the Meinhausen-Bühlmann estimator) is simply based off of a simple random walk observed at large times; the other examples we use are simple combinations of a path and cliques:

Example 11 (Path and cliques)**.** *Fix d and suppose $n/2$ is a multiple of d. Let B be a standard Brownian motion in 1 dimension, and let $X_1, \ldots, X_{n/2}$ be the values of the B at equally spaced points in the interval $[1/2, 3/2]$, i.e. $X_1 = B(1/2), X_2 = B((1/2) + 1/(n-1)), \ldots$ Equivalently, let the covariance matrix of this block be $\mathrm{Cov}(X_i, X_j) = 1/2 + \min(i,j)/n$, or take the Laplacian of the path and add the appropriate constant*

to the top-left entry.

Let the variables $X_{n/2+1,\ldots,X_n}$ be independent of the Brownian motion, and let their precision matrix be block-diagonal with $d \times d$ blocks of the form Θ_1 where Θ_1 is a rescaling of Θ_0 so that the coordinates have unit variance, and $\Theta_0 = I - (\rho/d)\vec{1}\vec{1}^T$ where $\rho \in (0, 1)$. In all experiments, we finally standardize the variables to have unit variance, following the usual recommendation (although the variances in this example are already bounded between 0.5 and 1.5).

The results of running all methods[8] on samples from this model are shown in Figure 5-1 for the Frobenius error with a fixed number of samples ($m = 150$) where the clique degree is $d = 4$ and the edge strength is $\rho = 0.95$. In Figure 5-2 we show the number of samples needed to recover the true edge structure for the same example with $d = 4$ in two cases, $\rho = 0.7$ and $\rho = 0.95$. We note that our definition of structure recovery is fairly generous — we apply a thresholding operation to the returned Θ matrix using the true value of $\kappa/2$, so the algorithms are not penalized for returning matrices with many small nonzero entries (which happens in practice at the optimal tuning of parameters, even though in the theory of e.g. [135] neighborhood estimates are made just from the support of the lasso estimate).

Note in particular that from Figure 5-2, we see the sample complexity of GreedyAndPrune scales like $O(\log(n))$, the information-theoretic optimal scaling which is in agreement with Theorem 40, while in the first example ($\rho = 0.7$) the sample complexity of the Graphical Lasso scales roughly like $\Theta(n)$ and in the second example ($\rho = 0.95$) the same is true for CLIME.

Recall that these examples are well-outside of the regime where the theoretical guarantees for methods like CLIME and Graphical Lasso can guarantee accurate reconstruction from $O(polylog(n))$ sammples, which is one reason we might expect them to be hard in practice. For example, the analysis of CLIME requires a bound on the entries of the inverse covariance (after rescaling the coordinates to have variance

[8]For the Graphical Lasso we used the standard R packages recommended in the original papers. For CLIME, we originally tested the standard R package but it was unable to reconstruct a path, presumably due to numerical issues. To fix this, we reimplemented CLIME using Gurobi and used a similar implementation for ACLIME.

$\Theta(1)$), but for the path Laplacian the entries of the precision matrix are of order $\Theta(n)$.

We describe one additional intuition as to why the Graphical Lasso should fails on this example: for the penalty $\lambda\|\hat{\Theta}\|_1$ to respect the structure of the path (where conditional variances are small) λ should be chosen small, but then the nodes in the cliques may gain spurious edges to the path and other cliques. With CLIME there is a similar concern that the ℓ_1 penalty for the two types of nodes does not scale properly. Different regularization parameters for the different types of edges could help in this particular example — however, it is typically difficult know beforehand which nodes have small and big conditional variances without effectively learning the GGM, as the way to show a node has low conditional variance almost always involves finding a good predictor of it from the other nodes. Concretely, in the case of ACLIME, it performed significantly worse than CLIME in most of our tests. On the other hand, the rescaling performed by our proposed algorithm HybridMB does resolve this issue in a principled way.

In the above two examples we tried, the (thresholded) Meinhausen-Bühlmann estimator successfully achieved similar sample complexity to our proposed methods, despite the fact that this example is again well outside of the regime where its theoretical guarantees are good. However, as we see in Figure 5-3 the sample complexity of this estimator is poor in another very simple example: a simple random walk with Gaussian steps run from times n to $2n$. (As before, this is the description of the model before standardizing coordinates to variance 1.) This is again not so surprising, as we know the Lasso (which the MB method is based upon) can only be guaranteed to obtain its "slow rate" guarantee when the coordinates of the input are highly dependent, and the slow rate guarantee for Lasso depends on norm parameters that are not sufficiently small in our example for good recovery guarantee.

5.11.2 Results for Riboflavin dataset

In this section we analyze the behavior of recovery algorithms on a popular dataset provided in [37]. This dataset has $m = 71$ samples and describes (log) expression

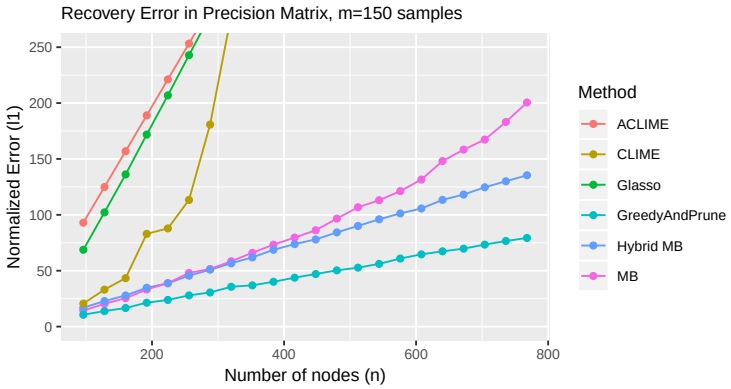

Figure 5-1: Normalized error (measured by $\|\hat{\Theta} - \Theta\|_1/n$ where $\|\cdot\|_1$ denotes the ℓ_1 norm viewing the matrix as a vector) in the precision matrix returned in Example 11 with $\rho = 0.95$. We note that this quantity should be expected to scale at least linearly, because some entries of Θ grow with n. Errors were averaged over 8 trials for each n and hyperparameters were chosen by grid search minimizing the recovery error in a separate trial, for each value of n. The tested parameters for λ in glasso were chosen from a log grid with 15 points from 0.0005 to 0.4, similarly for λ in MB, from 8 points from 1 to 32 for γ' in HybridMB (we set $\tau = 0$ for a more direct comparison to MB), for CLIME from a log grid with 15 points from 0.01 to 0.8, and for GreedyAndPrune k from a rounded log grid with 7 points from 3 to 24 and ν from a log grid with 8 points from 0.001 to 0.1.

levels for $n = 100$ genes in *B. subtilis*. We compared all of the methods listed above; our tables do not list the ACLIME results because it did not achieve nontrivial reconstruction (it's CV error as defined below was 0.98, which is essentially the same as the score for returning the identity matrix). We selected parameters using a 5-fold crossvalidation with the following least-squares style crossvalidation objective[9], after standardizing the coordinates to each have empirical variance 1 and mean 0:

$$E(\hat{\Theta}) := \frac{1}{nm_{holdout}} \sum_{i=1}^{n} \sum_{k=1}^{m_{holdout}} (X_i^{(k)} + \sum_{j \neq i} \frac{\hat{\Theta}_{ij} + \hat{\Theta}_{ji}}{2\hat{\Theta}_{ii}} X_i^{(k)})^2.$$

Note that the true Θ minimizes this objective as $m_{holdout} \to \infty$, making it equal to the sum of conditional variances; when the initial variances are set to 1, this objective simply measures the average amount of variance reduction achieved over

[9]An alternative which is sometimes used is the likelihood objective $\mathrm{Tr}(\hat{\Sigma}\hat{\Theta}) - \log\det(\hat{\Theta})$, but this objective is not very smooth due to the log det term and may equal ∞ even for entry-wise "good" reconstructions; since only glasso aims to return a positive definite matrix, we chose the simple least-squares objective instead.

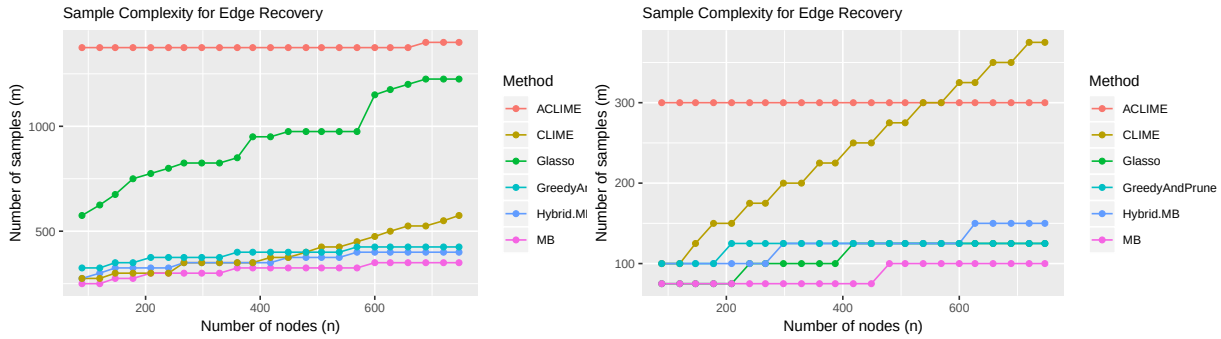

Figure 5-2: (a) $d = 4$ and $\rho = 0.7$, (b) $d = 4$ and $\rho = 0.95$. Number of samples needed to approximately recover true edge structure after thresholding using the test $\frac{|\hat{\Theta}_{ij}|}{\sqrt{\hat{\Theta}_{ii}\hat{\Theta}_{jj}}} > \kappa/2$, where κ is the κ for the true precision matrix from the information-theoretic assumption (5.1). Samples are drawn from the model in Example 11 with two different values for the edge strength ρ. Note that the sample complexity of GreedyAndPrune is consistent with the $O(\log(n))$ bound established in Theorem 40, whereas the graphical lasso and CLIME have sample complexity that appears to be roughly $\Theta(n)$ in the left and right examples respectively. The m shown is the minimal number of samples needed for the average number of incorrect edges per node (counting both insertions and deletions) to be at most 1. Trials and parameter selection was performed the same way as in the experiment for Figure 5-1, except that the parameters were chosen to minimize the number of incorrect edges, instead of error in the ℓ_1 norm.

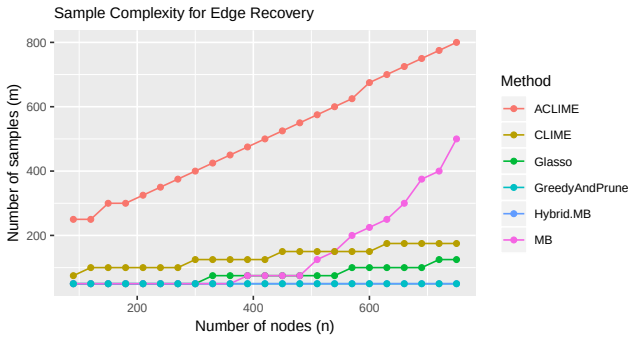

Figure 5-3: Large initial time simple random walk example: the setup is the same as in Figure 5-2, except that the ground truth model is a Gaussian simple random walk observed from times n to $2n$. We observe in this example that the sample complexity of ACLIME and the Lasso-based Meinhausen-Bühlmann estimator appear to scale roughly linearly in n, whereas the sample complexity of GreedyAndPrune and HybridMB is in fact constant over the observed values of n.

Method	CV Error	CV Parameters	# NNZ	Cond. No.	M	Δ_{WS}
Graphical Lasso	0.13	$\lambda = 0.01$	4378	968.6	54.8	8.7 %
CLIME	0.41	$\lambda = 0.21$	806	193.8	232.2	0.0 %
GreedyAndPrune	0.27	$k = 13, \nu = 0.01$	476	389.4	224	1.1 %
MB	0.17	$\lambda = 0.05$	1854	21439	156	1.1 %
HybridMB	0.19	$\gamma' = 21$	2758	1080843	324	2.2 %

Table 5.1: Results for precision matrix selected via 5-fold CV on Riboflavin dataset. The last 4 columns give summary statistics for the final recovered $\hat{\Theta}$ using the CV parameters on the entire dataset: M is the maximum ℓ_1 row norm for any row of Θ, the same as in the guarantee for CLIME cited earlier. The walk-summable relative error is $\Delta_{WS} := \frac{\|\tilde{\Theta} - \hat{\Theta}\|_F}{\|\hat{\Theta}\|_F}$ where $\tilde{\Theta}$ is the closest walk-summable matrix to $\hat{\Theta}$ in Frobenius norm. This shows that all of the estimated precision matrices are either walk-summable or close to walk-summable.

Method	Runtime (seconds)
Graphical Lasso	0.74
CLIME	2.12
GreedyAndPrune	0.19
MB	0.48
HybridMB	1.84

Table 5.2: Sequential runtime of methods on Riboflavin dataset with CV parameters, averaged over 10 runs. In all experiments, the graphical lasso implementation was from the glasso R package, CLIME was implemented by calling Gurobi from R (due to numerical limitations of the standard package), MB and HybridMB were implemented using the glmnet package, and for GreedyAndPrune we used a naive R implementation.

the coordinates.

The results of the cross-validation process[10] are shown in Table 5.1. As we see from the first 2 columns of the table, Graphical Lasso achieved the greatest amount of variance reduction but returned the densest estimate for Θ, MB and HybridMB had slightly less variance reduction, GreedyAndPrune had the sparsest estimate and achieved significantly more variance reduction that CLIME. We see that the chosen precision matrices have large condition number and row ℓ_1-norm M, comparable to the number of nodes n, which is significant in that known guarantees for Graphical Lasso,

[10]Essentially the same as before, parameters for Graphical Lasso were chosen from a log-scale grid from 0.001 to 0.5 with 15 points, for CLIME similarly from 0.01 to 0.8 with 20 points, and for GreedyAndPrune from a rounded log-scale grid from 3 to 26 with 7 points and from 0.001 to 0.1 with 8 points.

MB, CLIME and ACLIME are only interesting when these quantities are small (e.g. constant or $O(\log n)$). (Equivalently, the gap between variance and conditional variance is large; we note that the true gap may be even larger if we had access to more data, since we might be able to find even better estimators for each X_i given the other coordinates.) On the other hand, the recovered matrices are not far from walk-summable in Frobenius norm, suggesting that this is indeed a reasonable assumption.

In Table 5.2 we record the sequential runtimes of all of the methods on this dataset using the CV parameters. GreedyAndPrune was the fastest method. For larger datasets it is important to use parallelism, and we note we note that CLIME, MB, Hybrid.MB and GreedyAndPrune are "embarassingly parallelizable", as each node can be solved independently, but this is not the case for the Graphical Lasso. In practice, on our synthetic datasets and using 24 cores, CLIME becomes faster than the Graphical Lasso and GreedyAndPrune stays the fastest. In our experiment, we did not test our proposed method SearchAndValidate or the methods of [140], although they have good sample complexity guarantees, due to computational limitations; in [140], they report their methods requires on the order of days to run on this example.

We also performed a "semi-synthetic" experiment on this dataset, by taking the recovered (dense) Θ from Graphical Lasso, thresholding it to have $\kappa = 0.15$ and computing the sample complexity to recover the edges of the graphical model from sampled data (as in the synthetic experiments, with error of at most 0.25 incorrect edges per node, after thresholding at $\kappa/2$). All methods performed similarly on this test: the results are shown in Table 5.3.

Remark 15. *Several papers have been written on faster implementations of the graphical lasso, e.g. the Big & Quic estimator of [93]. However, these methods have mostly been developed/tested in the regime where λ is quite large: e.g. the documentation for the R package BigQuic implementing Big & Quic suggests using $\lambda \geq 0.4$ and that $\lambda = 0.1$ is too small to run in a reasonable time on large datasets. In practice, these methods may even fail to return the true optimum when given small λ; however, the*

Method	Number of Samples Needed	Optimal Parameters
Graphical Lasso	500	$\lambda = 0.005$
CLIME	550	$\lambda = 0.04$
GreedyAndPrune	550	$k = 6, \nu = 0.01$
MB	550	$\lambda = 0.01$
HybridMB	525	$\gamma' = 21$

Table 5.3: Number of samples needed to achieve error of at most 0.25 incorrect edges per node after thresholding in the semi-synthetic experiment: samples were drawn from a Θ given by thresholding the graphical lasso estimate from the Riboflavin dataset. The details of the thresholding, etc. are the same as in the synthetic experiment of Figure 5-2.

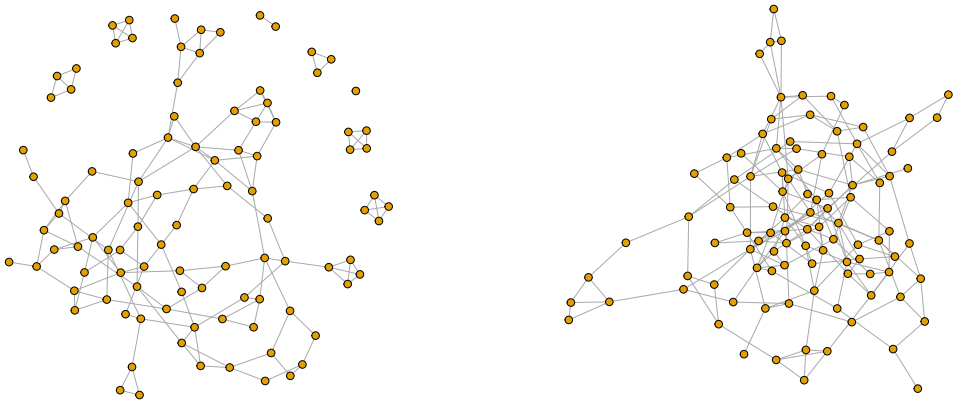

Figure 5-4: Left: thresholded graph from graphical lasso output on riboflavin data, used in semisynthetic experiment (see Table 5.3). Right: unthresholded graph output by GreedyAndPrune on Riboflavin data.

above experiment suggests this is an important regime in practice.

5.12 Some Difficult Examples

A natural question, given our results, is whether the GreedyAndPrune algorithm could possibly learn all sparse κ-nondegenerate GGMs with $O(\log n)$ sample complexity (without requiring walk-summability). Here we answer this question in the negative. Note by the analysis from Section 5.9.2 that if our GreedyAndPrune fails to succeed with $O(\log n)$ samples, then any analysis based on bounded ℓ_1-norm must also fail, since greedy methods always succeed if the ℓ_1-norm is small.

It is not too hard to find examples which break the greedy method when run once

from a single node, with the goal of recovering just that node's neighborhood. For example, if we take n pairs of near-duplicate variables (X_i, X_i') with $\mathrm{Var}(X_i) = \Theta(n)$ and $\mathrm{Var}(X_i - X_i') = \Theta(1)$ and define $Y = X_i - X_i'$ for some i, then using OMP to find a predictor of Y will fail to find the edge from X_i to Y with $O(\log n)$ samples. However, if we run a greedy method to find a predictor of X_i, then we actually will discover this edge. In the following example, we see there are edges which are not discovered from either direction:

Example 12 (Example breaking GreedyAndPrune)**.** *Fix $d > 2$ and let $Z_1, \ldots, Z_d$ be the result of taking d i.i.d. Gaussians and conditioning on $\sum_i Z_i = 0$. Define $X_i = Z_i + \delta W_i$ and $Y_i = Z_i + \delta W_i'$ where $W_i, W_i' \sim N(0,1)$ independently. Let Σ_0 be the covariance matrix of $X_1, \ldots, X_d, Y_1, \ldots, Y_d$ (so the Z are treated as latent variables).*

It can be checked that the GGM with covariance matrix Σ_0 remains κ nondegenerate for a fixed κ even as δ is taken arbitrarily small. Now consider the GGM which is block diagonal with first block Σ_0 and the second block the identity matrix, and suppose n is large. If we try to learn the neighbors of X_i, greedy will with high probability fail to find a superset of the correct neighborhood of node X_i, because after conditioning on Y_i, the angles between the residual of X_i and all of the other random variables are all near 90 degrees (going to 90 as $\delta \to 0$).

To summarize, this example is sparse and $\kappa = \Theta(1)$ nondegenerate but GreedyAndPrune fails to learn this GGM from $O(\log n)$ samples.

Remark 16. *Part of the motivation for the use of nearly-duplicated random variables is that one can prove (using essentially a modified version of Lemma 48)) that in a general sparse GGM there always exists at least one node i with at least one neighbor j such that $\mathrm{Var}(X_i|X_j)$ is noticeably smaller than $\mathrm{Var}(X_i)$. In this example, this is trivially true but is not useful for discovering connections between unpaired variables.*

Example 13 (Harder Example)**.** *The previous example, while it breaks* GreedyAndPrune*, cannot be a hard example in general because the edge structure is easy to determine from the covariance matrix. (The covariance matrix is roughly*

block diagonal and each block corresponds to a clique). The following variant seems significantly harder: start with Σ_0 from the previous example, and then Schur complement (i.e. condition) out $d/4$ many of the nodes to yield Σ_0'. Then the covariance matrix of the whole model is block diagonal with Σ_0' repeated $n/(d/4)$ times. Finally, we randomly permute the rows/columns.

Experimentally, it seems that Example 13 breaks the methods considered in our experiments in the high-dimensional regime where the number of samples is much less than the dimension n. However, this example itself cannot be computationally hard to learn: a simple algorithm to learn it thresholds the covariance matrix to find the sub-blocks made up of the paired nodes from a block, then picks a sub-block, conditions it out, and finds the remaining nodes from this block as the nodes whose conditional variance went down significantly.

The following important problem, first posed in [140], remains open: are κ-nondegenerate GGMs learnable from $O(\log n)$ samples with polynomial time algorithms?

Chapter 6

Learning RBMs with Bounded Weights

6.1 Introduction

In Chapter 2, we gave the first provable algorithms for learning RBMs, under the assumptions that the model is (1) sparse and (2) ferromagnetic. On the other hand, we also showed that learning general sparse RBMs is computationally intractable in general, because the conjecturally hard problem of learning a *sparse parity with noise* [184] can be embedded into a sparse RBM with a constant number of hidden units. The assumption of ferromagneticity (that variables are only positively correlated, not negatively correlated) rules out this example and plays a crucial role in the analysis of these works. Without ferromagneticity, viewing the marginal on X as a general Markov Random Field allows for using prior work [108] to give learning algorithms with runtime $n^{O(d_H)}$ where d_H is the maximum degree of a hidden node. This matches the lower bound of learning sparse parity with noise mentioned previously.

To summarize, the best previous results for learning RBMs either (1) make the assumption of ferromagneticity which makes building sparse parities impossible or (2) ignore all of the structure of the RBM except the max hidden degree, and pay the price of a $n^{\Theta(d_H)}$ runtime. This leaves open the question of developing algorithms whose runtime depends on some natural notion of a *complexity* measures of the RBM.

Our Results. In this chapter, we design an algorithm that is adaptive to a *norm* based complexity measure of the RBM, and often outperforms approach (2) above significantly, while not eliminating the possibility of negative correlation completely as in (1). This kind of complexity measure is often considered to be superior to alternatives like parameter-counting in practice, e.g. in the context of generalization bounds [13]. The key idea of our approach is to develop a novel connection between learning RBMs and their historical relative, feedforward neural networks. This connection allows us to establish new results for learning RBMs, by proving new results about learning feedforward neural networks (Section 6.2).

Our connection also validates the idea of so-called *supervised RBMs* as a natural distributional setting for classification with feedforward networks. Supervised RBMs, proposed by Hinton [90], treat one visible unit of the RBM as the label and the other visible units as the input to the classifier. This allows us to use the connection in the "reverse" direction — using natural structural assumptions on the RBM (like ferromagneticity) to give better results for solving supervised prediction tasks in an interesting distributional setting. Along these lines, we show that an assumption related to ferromagneticity, but allowing for some amount of negative correlation in the RBM, allows us to learn the induced feedforward network faster than would be possible without distributional assumptions (Section 6.3). Lastly, we present an experimental evaluation of our "supervised RBM" algorithm on MNIST and FashionMNIST to highlight the applicability of our techniques in practice (Section 6.5).

We note that in independent work, Bresler and Buhai [31] gave a new result for learning RBMs under a very different assumption: that there are few latent variables. Their algorithm uses a forward selection procedure which is more similar in spirit to the approach in Chapter 2 and [75, 30, 87] than the approach here.

6.2 Learning RBMs via New Results for Feedforward Networks

Relationship between RBMs and Feedforward Networks Our first result characterizes the relationship between RBMs and Feedforward networks. We show that there is a natural self-supervised prediction task in RBMs, of predicting the spin at node i given all other observed nodes, for which the Bayes-optimal predictor is *exactly given* by a two-layer feedforward network with a special family of tanh-like activations.

Theorem 53. *For any visible unit i in an arbitrary RBM,*

$$\mathbb{E}[X_i|X_{\sim i}] = \tanh\left(b_i^{(1)} + \sum_j \tanh(W_{ij}) f_{\beta_{ij}}\left(b_j^{(2)} + \sum_{k\neq i} W_{kj}X_k\right)\right) \tag{6.1}$$

where $\beta_{ij} = |\tanh(W_{ij})|$ and $f_\beta(x) := \frac{1}{\beta}\tanh^{-1}(\beta\tanh(x))$.

Proof. Observe that the conditional distribution of (X_i, H) given $X_{\sim i} = x_{\sim i}$ is given by

$$\Pr(X_i = x_i, H = h|X_{\sim i} = x_{\sim i}) \propto \exp\left(x_i(b_i^{(1)} + \sum_j W_{ij}h_j) + \langle W_{\sim i}^t x_{\sim i} + b^{(2)}, h\rangle\right) \tag{6.2}$$

where $W_{\sim i}$ denotes the $(n_1-1)\times n_2$ dimensional matrix given by deleting row i. Since the only quadratic terms left in the potential are between the remaining visible unit X_i and the hidden units h_j, this conditional distribution is exactly an Ising model on a star graph, i.e. a tree of depth 1 with root node corresponding to X_i. For all tree-structured graphical models, the conditional distribution of the root given the leaves can be computed exactly by Belief Propagation (see e.g. [138, 154]); in the case of Ising models it's known the general BP formula can be written with hyperbolic functions as above[1]. □

[1]For the readers convenience, we include a self-contained derivation of (6.1) from (6.2) in Appendix 6.7.1.

Remark 17. *An analogous result can be proved in the more general setting where the spins do not have to be binary; for example in a Potts model version of the RBM where each spin is valued in a set of size q, the conditional law of X_i given the others would be given again by a two-layer network where the last layer is a softmax. In this paper we focus on the binary case for simplicity.*

Remark 18. *The family of activation functions $f_\beta(x)$ naturally interpolates between the identity activation ($\beta = 1$ where $f_\beta(x) = x$) and* tanh *activation at $\beta = 0$, since*

$$\lim_{\beta \to 0} \frac{1}{\beta} \tanh^{-1}(\beta \tanh(x)) = \frac{\partial}{\partial \beta} \tanh^{-1}(\beta \tanh(x))\Big|_{\beta=0} = \tanh(x).$$

The exact structure of this prediction function is crucial in what follows and does not seem to have been known in the RBM literature, though some related ideas have been used to develop better heuristics for performing inference and training in RBMs (see discussion in Appendix 6.7).

Given this connection, we show that if we can solve the problem of learning such a neural network within sufficiently small error, then we can successfully learn the RBM. This reduces our RBM learning problem to that of learning feedforward neural networks in the setting that the input is bounded in ℓ_∞ norm.

Improved Results for Learning Feedforward Networks Subsequently, we give results for the feedforward network problem which are nearly optimal both in the terms of sample complexity (in the regime where λ is bounded) and in terms of computational complexity under the hardness of learning sparse parity with noise; some aspects of this result are new even for the well-studied case of learning neural networks with tanh activations (see Further Discussion).

Theorem 54 (Informal version of Corollary 6)**.** *Suppose that Y is a random variable valued in $\{\pm 1\}$, X is a random vector such that $\|X\|_\infty \leq 1$ almost surely and*

$$\mathbb{E}[Y|X] = \tanh\left(b^{(1)} + \sum_j w_j f_{\beta_j}\left(b_j^{(2)} + \sum_k W_{jk} X_k\right)\right)$$

where $b^{(1)} \in \mathbb{R}$, $\beta_j \in [0,1]$, w *is an arbitrary real vector and* W *is an arbitrary real matrix. Let* W_j *denote* column j *of* W *and suppose* $\|W_j\|_1 \leq \lambda$ *for every* j *and some* $\lambda \geq 2$. *Then if we run* ℓ_1*-constrained regression on the degree* D *monomial feature map* $\varphi_D(x) \mapsto \left(\prod_{i \in S} X_i\right)_{|S| \leq D}$ *with appropriate* ℓ_1 *constraint, the result* $\hat{w}$ *satisfies with high probability*

$$\mathbb{E}[\ell(\hat{w} \cdot \varphi_d(X), Y)] \leq OPT + \epsilon$$

where OPT *is the minimum logistic loss for any measurable function of* X, *as long as the number of samples* m *satisfies* $m = \Omega((|b^{(1)}|^2 \lambda^{O(D)}) \log(2n))$ *where* $D = O(\lambda \log(\|w\|_1 \lambda/\epsilon))$ *and the runtime of the algorithm is poly*(n^D).

We also show, under the standard assumption for hardness of learning sparse parity with noise, the following lower bound which shows that the runtime guarantee in our result is close to tight even in the usual setting of tanh neural networks ($\beta_j = 0$) — it is optimal up to log log factors in the exponent in its dependence on ϵ and $\|w\|_1$, and we also show that at least a subexponential dependence (essentially $2^{\sqrt{\lambda}}$) on λ is unavoidable (assuming the dependence on other parameters in the statement is fixed, since there are e.g. trivial algorithms that run in time 2^n).

Theorem 55 (Informal version of Theorem 63)**.** *There exists families of models (one with* ϵ *a constant, one with* $\|w\|_1$ *a constant) where a runtime of* $n^{\Omega\left(\frac{\log(\|w\|_1/\epsilon)}{\log\log(\|w\|_1/\epsilon)}\right)}$ *is needed for any algorithm to achieve* ϵ *error with high probability, regardless of its sample complexity. Even in the case of* tanh *activations (*$\beta_j = 0$ *for all* j*), there exists a sequence of models with* $\lambda = \Theta(n \log(n))$ *and* $\|w\|_1 = O(\sqrt{n})$ *which requires runtime* $n^{\Omega(\sqrt{\lambda/\log^2(\lambda)\log(n)}\log\|w\|_1)}$ *to achieve error* $\epsilon = 0.01$ *with high probability.*

To our knowledge, the fact that $n^{\log(\|w\|_1/\epsilon)/\log\log(\|w\|_1/\epsilon)}$ runtime is required to learn this class even for $\lambda = 1$, and by the above upper bound is tight up to the log log term, was not known before even for standard tanh networks. As far as the dependence on λ, a similar problem was studied in [166] where they proved the dependence cannot be polynomial using the result of [110] for intersection of halfspaces, based on a different assumption, though our lower bound seems to be somewhat stronger in the present context.

In particular the lower bounds on the runtime show that methods like the kernel trick cannot significantly improve the runtime compared to the simple method of writing out the feature map explicitly used in Theorem 54; however, writing out the feature map lets us use ℓ_1 regularization[2] instead of ℓ_2 which can give significant sample complexity advantages (e.g. $O(\log n)$ vs $O(n)$ for the usual sparse linear regression setups).

Structure Learning of RBMs As explained above, our reduction based on Theorem 53 lets us use the above feedforward network learning result to learn the structure of RBMs. By structure learning, we mean learning the *Markov blanket* of the each visible unit in the marginal distribution of the RBM over visible units, i.e. the minimal set of nodes S such that X_i is conditionally independent of all other X_j conditionally on X_S. We will also refer to the Markov blanket as the (two-hop) neighborhood of node i. This is a natural objective as other tasks such as distribution learning are straightforward in sparse models if the Markov blankets are known. As in the previous work on structure learning in other undirected graphical models (e., we will need some kind of quantitative nondegeneracy condition to guarantee nodes in the Markov blanket of node i are information-theoretically discoverable; it is not hard to see (e.g. using the bounds from [163]) that if two nodes are neighbors but their interaction is extremely weak then it becomes impossible to distinguish the model from the same model with the edge removed without a very large number of samples.

In Ising models and in ferromagnetic RBMs, there are simple conditions on the weight matrices which can ensure neighbors are information-theoretically discoverable. In a general RBM, there is no natural way to place constraints on the weights of the RBM to ensure this: the issue is that two nodes X_i and X_j can be independent even though they have two neighboring hidden units with non-negligible edge weights, since the effect of those hidden units can exactly cancel out so that X_i and X_j are independent or indistinguishably close to independent (a number of examples are given in the earlier Chapter). For this reason, we will instead make the following

[2]Interestingly, recent work [197] has shown in a special case connections between the implicit bias of gradient descent in feedforward networks and ℓ_1 regularization in function space.

assumption on the behavior of the model itself instead of on its weight matrix:

Definition 30. *We say that visible nodes i, j are η-nondegenerate two-hop neighbors if*

$$I(X_i; X_j | X_{\sim i,j}) = \mathbb{E}[\ell(\mathbb{E}[X_i | X_{\sim\{i,j\}}], X_i)] - \mathbb{E}[\ell(\mathbb{E}[X_i | X_{\sim i}], X_i)] \geq \eta$$

or if the same inequality holds with i and j interchanged. Here $I(X_i; X_j | X_{\sim i,j})$ is the conditional mutual information *between X_i and X_j conditional on $X_{\sim i,j}$, and the equality follows from Fact 3 in the Appendix and the definition of mutual information in terms of KL [47].*

Information-theoretically, this condition says that nontrivial information is gained about X_i by observing X_j, even after we have already observed $X_{\sim i,j}$. The fact that X_j is in the Markov blanket of node X_i exactly means that this quantity is nonzero. By Pinsker's inequality [47], η-nondegeneracy is also implied by a lower bound on the partial correlation $\mathrm{Cov}(X_i, X_j | X_{\sim i,j})$.

Example 14. *It is not hard to see that Ising models are equivalent to the marginal distribution of RBMs with maximum hidden node degree equal to 2. Consider an Ising model with minimum edge weight α and such that the maximum ℓ_1-norm into every node is upper bounded by λ and the external field is upper bounded by B, then $\eta \geq e^{-O(\lambda+B)}/\alpha$, see e.g. [29].*

Example 15. *In a ferromagnetic RBM with minimum edge weight α and maximum external field B, it can be shown that $\eta \geq e^{-O(\lambda_1+\lambda_2+B)}/\alpha^2$ (see Chapter 2 and [75]).*

In order for the RBM to be learnable with a reasonable number of samples (since general RBMs can represent arbitrary distributions with full support on the hypercube, which we saw in Chapter 2), we need to assume it has low complexity in the following sense:

Definition 31. *We say that an RBM is (λ_1, λ_2)-bounded if for any i, $\sum_j |\tanh(W_{ij})| + |b_i^{(1)}| \leq \lambda_1$ and the columns of W are bounded in ℓ_1 norm by λ_2.*

Note that λ_1 and λ_2 bound the ℓ_1 norm into the visible and hidden units, respectively. Based on our upper bounds and lower bounds for the learnability of feedforward networks, it should be less surprising that these parameters play a very different role in the computational learnability of RBMs.

Theorem 56 (Informal version of Theorem 64)**.** *Suppose all two-neighbors in a (λ_1, λ_2)-bounded RBM are η-nondegenerate. Given $m = \Omega(\lambda_2^{O(D)} \log(2n))$ i.i.d. samples from the RBM, where $D = O(\lambda_2 \log(\lambda_1\lambda_2/\eta))$, we can recover its structure with high probability in time poly(n^D).*

Based on this result we also give a result for learning the RBM in TV distance under the same assumption: see Theorem 65: the sample complexity of this method is essentially the above sample complexity plus $n^2(1-\tanh(\lambda_1))^{-d_2}$ where d_2 is the maximum 2-hop degree; the *poly*(n) dependence is required as even learning n bernoullis in TV requires $\Omega(n)$ sample complexity. Our algorithm encodes the distribution as a sparse Markov Random Field, but (if desired) this can easily be converted into a sparse RBM using an algorithm in Chapter 2. Therefore we learn the distribution properly, except that the learned RBM typically has more hidden units than the original RBM (i.e. it is overparameterized).

When interpreting these result, it is crucial not to confuse the ℓ_1 norm parameters λ_1, λ_2 of visible and hidden units with the maximum degrees of these units. Typically in Ising models, we should think of the weight of a typical edge as *shrinking* as d grows so that units stay near the sensitive region of their activation and the behavior of the model does not become trivial — this means that λ_1 and λ_2 may be much smaller than d. This is consistent with practical advice in the RBM literature, see e.g. [90]. Probably the most well known sufficient condition for being able to sample in an Ising model (or RBM) is *Dobrushin's uniqueness criterion* which is equivalent to the requirement that $\lambda_1, \lambda_2 \leq 1$ and this condition is actually tight for Glauber dynamics to mix quickly in the Ising model on the complete graph (Curie-Weiss Model) [122]. We discuss this further in Remark 21; in Dobrushin's uniqueness regime and under some mild nondegeneracy conditions we expect that $\eta = \Omega(1/d^2)$ so the

above algorithm has runtime $n^{\log(d)}$, which is an exponential improvement in the exponent compared to the best previously known result ($O(n^d)$ runtime by viewing the RBM as an MRF).

We also give lower bound results showing that the computational complexity of the above algorithm is essentially optimal in terms of λ_1 and η (based upon the hardness of learning sparse parity with noise) and nearly optimal in terms of λ_2 for an SQ (Statistical Query) algorithm, in the sense that any SQ algorithm needs at least sub-exponential dependence on λ_2 (given that the dependence on other parameters is not changed — e.g. obviously there is a 2^n time algorithm to learn this problem). In particular, this shows that our results for learning feedforward networks under ℓ_∞ are close to tight even in this application, where the input distribution is related to the label.

Theorem 57 (Informal version of Theorem 71)**.** *As before, λ_2 refers to the maximum ℓ_1-norm into any hidden unit and we choose parameters so that $\lambda_2 = poly(n)$ and $\lambda_1 = poly(n)$. There exists $\epsilon > 0$ so that no SQ algorithm with tolerance $n^{-\lambda_2^\epsilon}$ and access to $n^{\lambda_2^\epsilon}$ queries can structure learn an $\alpha = \Omega(1)$-nondegenerate (λ_1, λ_2)-bounded RBM.*

We also show (Theorem 68) that the η-nondegeneracy condition is required to achieve nontrivial guarantees even if we are only interested in distribution learning (i.e. in TV), assuming the hardness of learning sparse parity with noise.

6.3 Supervised RBMs

Since in many applications the input data to a classifier is clearly very structured (e.g. images, natural language corpuses, data on networks, etc.), it is interesting to consider the behavior of classification algorithms under structural assumptions on the data. RBMs are one (relatively simple) generative model which can generate interesting structured data. This suggests the idea of learning "supervised RBMs", as proposed by Hinton [90], where we assume the input and label are drawn from an RBM joint

distribution, so that predicting the label is a feedforward network by Theorem 53; in this model the label is just a special visible unit in the RBM. Based on the previous discussion about computational lower bounds, we know that assuming the input to a feedforward network comes from the corresponding RBM does not in general make learning easier, but we know that in RBMs there are very natural assumptions we can make to avoid these computational issues. Our final result is of exactly this flavor, showing how we can learn the supervised RBM under a ferromagneticity-related condition faster than is possible if we did not have a distributional assumption.

In order to emphasize the special role of the node which we want to predict, we will adopt a modified notation where the visible unit which we want to learn to predict is labeled Y and all other visible units are still labeled X. More precisely, we model the joint distribution over input features X valued in $\{\pm 1\}^{n_1}$, latent features H valued in $\{\pm 1\}^{n_2}$ and label $Y \in \{\pm 1\}$ as,

$$\Pr[X = x, H = h, Y = y] \propto \exp\left(\langle x, Wh\rangle + \langle h, w\rangle y + \langle b^{(1)}, x\rangle + \langle b^{(2)}, h\rangle + b^{(3)} y\right)$$

where the *weight matrix* W is a non-negative $n_1 \times n_2$ matrix, w is an arbitrary n_1 dimensional vector and $b^{(1)} \in \mathbb{R}^{n_2}$, $b^{(2)} \in \mathbb{R}^{n_2}$ and $b^{(3)} \in \mathbb{R}$ are arbitrary. Given the latent variables H, w can be seen as the linear predictor for Y.

Theorem 58 (Informal Version of Theorem 73)**.** *Suppose the interaction matrix W is ferromagnetic with minimum edge weight α. Further suppose one of the RBMs induced by conditioning on $Y = 1$ or $Y = -1$ is a (λ, λ)-RBM. Then there exists an algorithm that learns the predictor Y that minimizes logistic loss up to error ϵ. The algorithm has sample complexity $m = n_1^2 \exp(\lambda)^{\exp(O(\lambda))}(1/\alpha)^{O(1)} \log(n_1/\delta)/\epsilon^2$ and has runtime poly(m).*

Our main algorithm can be broken down into three main steps: (1) Use greedy maximization of conditional covariance $\mathrm{Cov}^{\mathsf{Avg}}$ to first learn the two-hop neighborhood $\mathcal{N}(i)$ of each observed variable i w.r.t. the hidden layer conditioned on the label (see Algorithm 2), (2) For each observed variable X_i, learn the conditional law of $X_i \mid X_{\mathcal{N}(i)}, Y$ using regression, and (3) Use the estimated distribution to compute

$\mathbb{E}[Y|X]$. Step (1) leverages tools from [75] but considers a setting where the RBM may in fact have some amount of negative correlation, as w has arbitrary signs and is allowed to have large norm. Step (2) can be achieved by simply looking at the conditional law under the empirical distribution; this is efficient as we learn small neighborhoods.

In step (3), we can make use of the following useful trick (a version of which can be found in [90]): we already have enough information to derive the law of $Y \mid X$ since we know the marginal law of Y (the fraction of $+$ and $-$ labels) and the law of $X \mid Y$. However, naively carrying out the Bayes law calculation is difficult because it involves partition functions (which are in general NP-hard to approximate, see e.g. [172]). We avoid computing the partition function by observing that if we define f_1, f_2 such that $\Pr(X, Y) \propto \exp(f_1(X)\mathbb{1}(Y = 1) + f_2(X)\mathbb{1}(Y = -1) + by)$, then the law of $Y \mid X$ follows a logistic regression model where

$$\mathbb{E}[Y \mid X] = \tanh\left(\frac{f_1(X) - f_2(X)}{2} + b\right)$$

for some constant $b \in \mathbb{R}$. Therefore if we know f_1, f_2 up to additive constants (which we can derive from the Fourier coefficients learned in (2)), we can simply fit a logistic regression model from data to learn h plus the missing constants, and we can prove this works using fundamental tools from generalization theory. We refer the reader to Appendix 6.10 for additional details.

Algorithm 2 LearnSupervisedRBMNbhd$(u, \tau, \mathcal{S})$ (Adapted from [75])

Set $S := \phi$
Set $i^* = \arg\max_v \widehat{\mathrm{Cov}}_{\mathcal{S}}^{\mathsf{Avg}}(u, v|S, Y)$, and $\eta^* = \max_v \widehat{\mathrm{Cov}}_{\mathcal{S}}^{\mathsf{Avg}}(u, v|S, Y)$
if $\eta^* \geq \tau$ **then**
    $S = S \cup \{i^*\}$
else
    Go to Step 8
Go to Step 2
For each $v \in S$, if $\widehat{\mathrm{Cov}}_{\mathcal{S}}^{\mathsf{Avg}}(u, v|S\backslash\{v\}, Y) < \tau$, remove v (*Pruning step*)
Return S

Observe that under the given distributional assumptions, our algorithm has run-

time complexity polynomial in the input dimension in contrast to Theorem 54 where the run time scales as $n^{\Omega(\lambda)}$. A simple example which shows the algorithm from this Theorem will outperform any algorithm without distributional assumptions (like Theorem 54) is given in Remark 24.

6.4 Discussion: Comparison to Prior work on Learning Neural Networks

In the neural network learning literature, various works prove positive results that either (1) work for any distribution with norm assumptions or (2) require strong distributional assumptions. The result of Theorem 54 falls into the category (1) and the result of Theorem 58 falls into category (2).

We first discuss the relation of Theorem 54 to other previous works of type (1). Perhaps the most closely related works are [166, 200, 76, 80]. All of these works assume the input is bounded in ℓ_2 norm and give learning results based on kernel methods; of course, these results could be applied under the assumption of ℓ_∞-bounded input, by using the inequality $\|x\|_2 \leq \sqrt{n}\|x\|_\infty$ and rescaling the input to have norm 1. For comparison, the best result in the ℓ_2 setting with tanh activation is given in [80], but this result (as is essentially necessary based on the known computational hardness results) has exponential dependence on the ℓ_2 norm of the weights in the hidden units, so doing such a reduction just using norm comparison bounds gives a runtime sub-exponential in dimension. Therefore it is indeed crucial for us to give a new analysis adapting to learning with input bounded in ℓ_∞. An interesting feature of this setting (as mentioned above) is that the kernel trick does not seem to be as useful for improving the runtime as the ℓ_2 setting, where it seems genuinely better than writing out the feature map [76, 80].

Due to the generality of direction (1), it is hard to design efficient algorithms. This further motivates direction (2), however, making the right distributional assumptions which allow for efficient learning while being well-motivated in context of real world

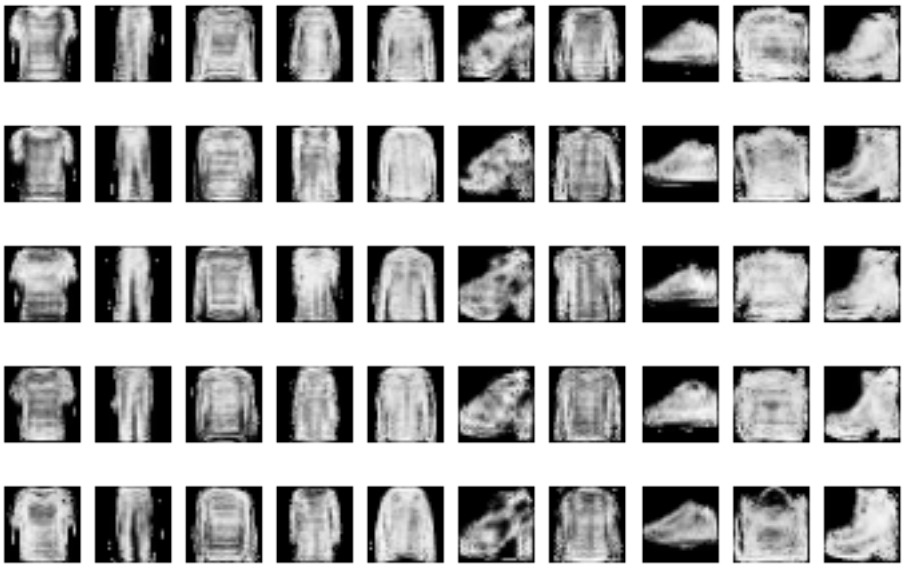

Figure 6-1: Five i.i.d. samples for each FashionMNIST class, drawn from the trained model by Gibbs sampling.

data can be very challenging. Most prior work has been limited to the Gaussian input [182, 173, 35, 202, 123, 56] or symmetric input [80, 72] assumptions which are not satisfied by real world data. The works of [142, 130] gave results for some simple tree-structured generative models. There has been some work in defining data based notions such as eigenvalue decay [77] and score function computability [71] to get efficient results. Our assumption for Theorem 58 in contrast exploits sparsity and nonnegative correlations among the input features conditional on the output label.

6.5 Experiments

In this section we present some simple experiments on MNIST and FashionMNIST to confirm that our method performs reasonably well in practice. In these experiments, we implemented the supervised RBM learning algorithm from Theorem 58 which makes use of the classification labels provided in the training data set. This algorithm outputs both a classifier (which predicts the label given the image) and also a generative model (which can sample images given a label).

For classification, we allowed the logistic regression (described as "step (3)" above)

to fit not just the bias term but also coefficients on the sum of Fourier coefficients for each pixel (an input of dimension $768 \times 10 = 7680$), since the runtime of the logistic regression step is almost negligible anyway. This is useful because it allows greater dynamic range in the influence of each pixel.

We observed a test accuracy of $97.22 \pm 0.16\%$ on MNIST; the training accuracy was 99.9% and we trained the logistic regression for 30 epochs (same as steps) of L-BFGS with line search enabled. For FashionMNIST, we obtained a test accuracy of $88.84 \pm 0.31\%$; the training accuracy was 92.19% and we trained the logistic regression for 45 epochs with L-BFGS as before. Overall training took a bit less than an hour each on a Kaggle notebook with a P100 GPU. Both datasets have $60,000$ training points and $10,000$ test; in both experiments we used a maximum neighborhood size of 12, and stopped adding neighbors if the conditional variance shrunk by less than 1%.

For context, we note that our accuracy on MNIST is better than what we would get using standard training methods for RBMs and logistic regression for classification; [69] reports accuracies of approximately 95% for CD and 96% using a more sophisticated TAP-based training method. The results are also around as good or better than what is achieved using many classical machine learning methods on these datasets [198]; for example, logistic regression achieves error 91.7% and 84.2% and polynomial kernel SVM achieves error 89.7% and 97.6% [198]. Of course, none of these results are as good as specialized deep convolutional networks (over 99% on MNIST). In contrast to other approaches using linear models such as kernel SVM, our approach also learns a generative model. Being able to sample from the generative model can give some insight into how the model classifies.

To evaluate the performance of the learned RBM as a generative model, we generated samples using Gibbs sampling starting from random initialization and run for 6000 steps. As is common practice, we output the probabilities generated in the last step instead of the sampled binary values, so that the result is a normal greyscale image. We display the resulting samples in Figures 6-1 and 6-2 (for reference, see randomly sampled training datapoints in Appendix 6.11): we note that the model

successfully generates samples with diversity, as in Figure 6-1 the model generates handbags both with and without handles, and in Figure 6-2 it renders both common styles for drawing the number 4.

It is clear that the model fails to generate as detailed of patterns exhibited in real FashionMNIST images since in our training algorithm, we represent a gray pixel as a random combination of black and white, so a checkerboard pattern of black and white and a patch of grey are not well-distinguished. We do this to ensure that our setup is comparable to classic RBM training [90]. It is potentially possible to fix this by adding spins over larger alphabets (e.g. real-valued) to the model.

6.6 Organization

Here we briefly outline the contents of each remaining section; each bold heading in the text below corresponds to a new section.

Section 6.7. Connections between Distribution Learning and Prediction in RBMs In this section we show that if you have learned the distribution of an RBM, then you have also in principle learned how to predict the output of corresponding feedforward networks. These feedforward networks are induced from a "self-supervised" prediction task: predicting the spin at node i given observations of all other spins. This connection leverages a classical observation in probabilistic inference: inference in all tree-structured graphical models has an exact solution known as Belief Propagation (see e.g. [154, 138]); perhaps surprisingly, this observation is useful even though the RBM itself is not tree structured. Conversely, in the next subsection we give quantitative bounds showing that sufficiently good predictors for this self-supervised objective for every node i allows us to recover the distribution of the corresponding RBM.

Section 6.8. Guarantees for Learning Feedforward Networks (with arbitrary distribution). In this section we prove upper and lower bounds for learning

one-layer feedforward networks with f_β activations in the hidden units and inputs X drawn from an arbitrary distribution such that $\|X\|_\infty \leq 1$.

In the first two subsections, we prove the needed approximation-theoretic results about our class of activations f_β, giving approximation results with uniform guarantees over the entire interval $\beta \in [0, 1]$. In the special case of $\beta = 0$, $f_\beta = \tanh$ and the needed result has essentially already been proved in the work of [166]. As explained in the first subsection, by a classical result of Bernstein (Theorem 59 below) it turns out that analyzing approximation theory for functions analytic on $[-1, 1]$ is equivalent to analyzing the function's extension into the complex plane. We develop the needed complex-analytic estimates (which crucially are uniform in β) in the following subsection. We note that the authors of [166] did not use Bernstein's result to prove their bound; their analysis of the $\beta = 0$ case is longer because they more or less reproduce the steps from the proof of the upper bound of Bernstein's Theorem.

After solving the approximation-theoretic question, we use them in an ℓ_1-regression based algorithm for learning feedforward networks, using an explicit polynomial feature map and the logistic version of the Lasso with its corresponding nonparametric generalization bounds. We derive the needed ℓ_1-norm bound in a clean way from the approximation-theoretic results using in part a Lemma of [168], previously used in [76]. This proves Theorem 54. In the last subsection, we prove that this result is nearly optimal under the hardness of sparse parity with noise, even in the case of tanh networks, using two different ways to construct a parity out of tanh units: one is a well-known construction from [86], the other is based on Taylor series expansion and is related to the MRF-to-RBM embedding result established in Chapter 2.

Section 6.9. Learning RBMs by Learning Feedforward Networks. In this section, we show how to derive structure recovery results (i.e. recovery of Markov blankets) for RBMs by using the feedforward network learning results developed in the previous section. Assuming η-nondegeneracy, we show how to learn the structure of the network by doing simple regression tests, e.g. comparing the minimal logistic

loss achieved predicting node i from all other nodes to the loss when node j is excluded from the input. This proves Theorem 56. We explain in more detail in Remark 21 how this result is a significant improvement over previous results in interesting regimes where we know that the RBM can actually be sampled from in polynomial time. Based on this, we prove a result for learning the distribution: by Theorem 56 this reduces to the case where the structure is known, so by proving a good estimate (Lemma 62) on the convergence of the natural predictor of X_i given its neighbors, the empirical conditional expectation and using the tools developed in Section 6.7.3 gives the result. A key point here is that the empirical conditional expectation converges at a much faster rate than e.g. relying on Theorem 62, which gives better sample complexity guarantees.

Finally, we again prove some computational hardness results. We establish that the algorithm's dependence is essentially optimal in terms of η and $\|w\|_1$ by using the Taylor-series based sparse parity construction from Chapter 2, related to the construction used above for tanh networks. For the dependence on λ_2, the hidden unit ℓ_1-norm, we use a third, different construction of parity from [133] for the RBM setting; this construction is not amenable to adding noise, but we are able to prove a lower bound on the runtime in terms of λ_2 for all SQ (Statistical Query) algorithms (see e.g. [24]).

Section 6.10. Learning a Feedforward Network by Learning RBMs. In this section, we prove Theorem 58, which lets us learn to predict in supervised RBMs under a natural conditional ferromagneticity condition in a provably more computationally efficient way than applying distribution-agnostic methods for learning feedforward networks like Theorem 54. In Remark 24 we give a simple example where the gap is provable and explain the (in this case) simple intuition as to how the approach of Theorem 58 uses the structure of the input data in a favorable way.

The idea of this learning algorithm is essentially to use Bayes rule to reduce computing the posterior on the label (i.e. $\Pr(Y|X)$) to computing the conditional likelihood of the observed X under the two possible values of the label. In some

situations where the conditional law of $Y|X$ is very simple, this approach may be overkill as it requires to model the law of X; however, we are interested in the setting where the label Y may have a large, complicated effect on X so this approach seems perfectly reasonable. An obvious issue with using Bayes rule in this way is that even if the the RBM is already known perfectly, computing the normalizing constant for the conditional distribution under $Y = +$ or $Y = -$ in such a model is #BIS-Hard [81]. Fortunately, for our application we show that we can estimate the needed ratio of normalizing constants from the data using a simple variant of logistic regression.

What remains is to learn how to estimate the conditional log-likelihoods i.e. $\Pr(X|Y)$. Fortunately, even though under our assumptions the original RBM was not ferromagnetic, the conditional models we get by applying Bayes rule are indeed ferromagnetic so we can apply the methods developed in [75] for learning such a model. Here we need the results of [75], which applies to a more general setting than the results in Chapter 2 (at the cost of higher sample complexity), as we expect the external fields in the resulting model to be inconsistent (have differing signs depending on the site). Once the structure is recovered, we can learn the coefficients of the log-likelihood using the results established in the previous section based on fast convergence of the empirical condition expectation, and using these coefficients we can accurately estimate $\Pr(X|Y)$ for the application of Bayes rule.

Section 6.11. Additional Experimental Data. In this section we include reference images from both datasets along with samples generated by our algorithm trained on MNIST.

6.7 Connections between Distribution Learning and Prediction in RBMs

To our knowledge, Theorem 53 has not been previously noted in the literature on RBMs. However, this is not the first time connections between RBMs and message passing algorithms for inference has been investigated: for example, the work of [195]

extensively studied the use of message passing algorithms (i.e. Belief Propagation and related algorithms) for estimating the mean and covariance matrix of nodes in an RBM, and the work of [69] used the related TAP approximation to derive better alternatives to constrastive divergence for training RBMs in practice. The key conceptual difference is that in these works, their goal is to solve a much harder problem (e.g. estimating marginals and $\log Z$) which is well-known to be NP-hard in general. In contrast, for our application to learning the relevant task ends up being predicting one node from the others, which it turns out is *not* computationally difficult if we know the model — conditioning on the other nodes breaks all cycles in the graph, which is the obstacle that makes inference difficult in general.

6.7.1 Conditional Law Derivation

In this Section we give, for the reader's convenience, a self-contained derivation of the conditional law (6.1) described in Theorem 53 for $\mathbb{E}[X_i|X_{\sim i}]$ from (6.2). As described in the proof of the Theorem, the result is obtained as a special case of the Belief Propagation algorithm as described in a number of references, including [138, 154], which is derived by performing a more general version of this calculation. First recall that the joint conditional law on X_i, H condiditioned on $X_{\sim i}$ is given by (6.2):

$$\Pr(X_i = x_i, H = h | X_{\sim i} = x_{\sim i}) \propto \exp\left(x_i(b_i^{(1)} + \sum_j W_{ij} h_j) + \langle W_{\sim i}^t x_{\sim i} + b^{(2)}, h \rangle \right).$$

The computation proceeds by rewriting this measure with respect to a "cavity" measure where all terms involving X_i are removed. For each hidden unit j, define a corresponding probability measure

$$\mu_{H_j \to X_i}(h_j) \propto \exp\left(\sum_{k \neq i} W_{kj} x_j h_j + b_j^{(2)} h_j \right)$$

under which $\sum_j h_j \mu_{H_j \to X_i}(h_j) = \tanh(\sum_k W_{kj} x_j + b_j^{(2)})$ and rewrite the joint probability over X, H as

$$\Pr(X_i = x, H = h | X_{\sim i} = x_{\sim i}) \propto \exp\left(x_i (b_i^{(1)} + \sum_j W_{ij} h_j) \right) \prod_j \mu_{H_j \to X_i}(h_j).$$

Now we compute that

$$\begin{aligned}
&\Pr[X_i = x_i | X_{\sim i} = x_{\sim i}] \\
&= \sum_h x_i \Pr(X_i = x_i, H = h | X_{\sim i} = x_{\sim i}) \\
&\propto \sum_h \exp\left(x_i (b_i^{(1)} + \sum_j W_{ij} h_j) \right) \mu_{H \to X_i}(h) \\
&= \exp(x_i b_i^{(1)}) \prod_{j=1}^{n_2} (\cosh(W_{ij}) + \sinh(x_i W_{ij}) \tanh(\sum_{k \neq i} W_{kj} x_j + b_j^{(2)})) \\
&\propto \exp(x_i b_i^{(1)}) \prod_{j=1}^{n_2} (1 + x_i \tanh(W_{ij}) \tanh(\sum_{k \neq i} W_{kj} x_j + b_j^{(2)})) \\
&= \exp\left(x_i b_i^{(1)} + \sum_{j=1}^{n_2} \log(1 + x_i \tanh(W_{ij}) \tanh(\sum_{k \neq i} W_{kj} x_j + b_j^{(2)})) \right)
\end{aligned}$$

where we used $\propto$ to ignore constants of proportionality independent of x_i and in the third line we used Lemma 51 below. Therefore if we use that

$$\log(1+\beta x_i) = \frac{1}{2} \log \frac{1+\beta x_i}{1-\beta x_i} + \frac{1}{2}(\log(1+\beta x_i) + \log(1-\beta x_i)) = \tanh^{-1}(\beta x_i) + \frac{1}{2}(\log(1+\beta) + \log(1-\beta))$$

where we see the last term does not depend on x, we can compute that

$$\begin{aligned}
&\mathbb{E}[X_i = x_i | X_{\sim i} = x_{\sim i}] \\
&= \frac{\sum_{x_i} x_i \exp\left(x_i b_i^{(1)} + \sum_{j=1}^{n_2} \log(1 + x_i \tanh(W_{ij}) \tanh(\sum_{k \neq i} W_{kj} x_j + b_j^{(2)})) \right)}{\sum_{x_i} \exp\left(x_i b_i^{(1)} + \sum_{j=1}^{n_2} \log(1 + x_i \tanh(W_{ij}) \tanh(\sum_{k \neq i} W_{kj} x_j + b_j^{(2)})) \right)} \\
&= \frac{\sum_{x_i} x_i \exp\left(x_i b_i^{(1)} + \sum_{j=1}^{n_2} x_i \tanh^{-1}(\tanh(W_{ij}) \tanh(\sum_{k \neq i} W_{kj} x_j + b_j^{(2)})) \right)}{\sum_{x_i} \exp\left(x_i b_i^{(1)} + \sum_{j=1}^{n_2} x_i \tanh^{-1}(\tanh(W_{ij}) \tanh(\sum_{k \neq i} W_{kj} x_j + b_j^{(2)})) \right)}
\end{aligned}$$

$$= \tanh\left(b_i^{(1)} + \sum_{j=1}^{n_2} \tanh^{-1}(\tanh(W_{ij})\tanh(\sum_{k\neq i} W_{kj}x_j + b_j^{(2)}))\right)$$

where in the final step we used that $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$. From this we get (6.1) by plugging in the definition of $f_{\beta_{ij}}$.

Lemma 51. *For any $z \in \mathbb{R}$ we have the formula for moment generating function of a recentered Bernoulli:*

$$\mathbb{E}_{X\sim Ber_\pm(\tanh(z))}[\exp(\lambda X)] = \cosh(\lambda) + \sinh(\lambda)\tanh(z)$$

where $Ber_\pm(\mu)$ denotes the distribution of a $\{\pm 1\}$-valued random variable with mean μ.

Proof. First recall that $\mathbb{E}_{X\sim Rad}[\exp(\lambda X)] = \cosh(\lambda)$ and $\mathbb{E}_{X\sim Rad}[X\exp(\lambda X)] = \tanh(\lambda)$. Therefore

$$\begin{aligned}
\mathbb{E}_{X\sim Ber_\pm(\tanh(z))}[\exp(\lambda X)] &= \mathbb{E}_{X\sim Rad}\left[e^{\lambda X}\frac{e^{zX}}{\cosh(z)}\right] \\
&= \frac{\cosh(z+\lambda)}{\cosh(z)} \\
&= \frac{\cosh(z)\cosh(\lambda) + \sinh(z)\sinh(\lambda)}{\cosh(z)} \\
&= \cosh(\lambda) + \sinh(\lambda)\tanh(z).
\end{aligned}$$

□

6.7.2 2-layer Tanh Neural Network as Bayes-Optimal Prediction in an RBM

In particular, (6.1) lets us realize *any* standard 2-layer tanh neural network as the Bayes-optimal predictor in an RBM in a natural limit where the number of hidden neurons goes to infinity, but the effect of each hidden neuron is very small, so that the ℓ_1 norm of the weights going into the top neuron stays bounded by a constant. Each

hidden unit in the neural network corresponds in a direct way to several duplicated hidden units in the RBM. The construction is given explicitly in the next Lemma; we will not use the statement explicitly but use it to develop intuition for (6.1).

Lemma 52. *Suppose that* $g(x) = \tanh\left(u_0 + \sum_{j=1}^{T} u_j \tanh\left(M_{j0} + \sum_k M_{jk} x_k\right)\right)$ *where* x *is* n*-dimensional, i.e.* g *is a 2-layer neural network with* tanh *activations. Then*

$$g(x) = \lim_{K\to\infty} \tanh\left(u_0 + \sum_{i=1}^{K}\sum_{j=1}^{T} \tanh(u_j/K) f_{|u_j/K|}\left(M_{j0} + \sum_k M_{jk} x_k\right)\right),$$

so by (6.1) *from Theorem 53 the restriction of* f *to* $\{\pm 1\}^n$ *is the Bayes-optimal predictor of a visible unit in an RBM with* $n+1$ *total visible units where the activations of the other visible units are known.*

Proof. This follows from the observation in Remark 18 and from Theorem 53 by building the corresponding RBM with KT hidden units. □

6.7.3 Distribution learning bounds from prediction bounds

In this section, we show how good estimates of the conditional prediction functions can be used in a direct way to recover the joint distribution of the RBM in total variation distance.

Algorithm 3 DistributionFromPredictors

1: For every i we suppose we are given $\hat{f}_i : \{\pm 1\}^n \to \mathbb{R}$ and set $\widehat{\mathcal{N}}(i)$ such that $\hat{f}_i$ is a predictor of node i from other nodes that depends only on those in the set $\widehat{\mathcal{N}}(i)$
2: Define $\mathcal{S} := \{S : \exists i, S \subset \widehat{\mathcal{N}}(i)\}$
3: **for** $S \in \mathcal{S}$ **do**
4: For all $i \in S$, define $\hat{w}_{S,i} := \mathbb{E}_{X\sim Uni(\{\pm 1\}^n)}[\tanh^{-1}(\hat{f}_i(X)) X_{S\setminus i}]$.
5: Define $\hat{w}_S := \frac{1}{|S|}\sum_{i\in S} \hat{w}_{S,i}$.
6: Return the MRF with unnormalized pmf $\exp\left(\sum_{S\in\mathcal{S}} \hat{w}_S X_S\right)$.

Lemma 53 ([163])**.** *Suppose* P, Q *are distributions over random variable* X *valued*

in $\{\pm 1\}^n$. *If* $P(x) \propto \exp(\sum_S p_S X_S)$ *and* $Q(x) \propto \exp(\sum_S q_S X_S)$ *then*

$$\mathbf{SKL}(P, Q) = \sum_S (p_S - q_S)(\mathbb{E}_P[X_S] - \mathbb{E}_Q[X_S]).$$

where $\mathbf{SKL}(P, Q) = \mathbf{KL}(P, Q) + \mathbf{KL}(Q, P)$ *is the symmetrized KL divergence.*

Proof. From the definition we see

$$\mathbf{SKL}(P, Q) = \mathbb{E}_P\left[\log \frac{P(x)}{Q(x)}\right] - \mathbb{E}_Q\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_P\left[\sum_S (p_S - q_S) X_S\right] - \mathbb{E}_Q\left[\sum_S (p_S - q_S) X_S\right]$$

so using linearity of expectation proves the result. □

The following definition captures the level of contiguity P has with the uniform measure when looking at small sets of coordinates.

Definition 32. *For any distribution* P *on* $\{\pm 1\}^n$ *and* $d \leq n$ *we define*

$$\delta_P(d) := \inf_{|S| \leq d} \inf_{x_S} 2^{|S|} P(X_S = x_S).$$

Lemma 54. *For any function* f *which depends on at most* d *coordinates,*

$$\mathbb{E}_P[f(X)^2] \geq \delta_P(d) \mathbb{E}_{X \sim \{\pm 1\}^n}[f(X)^2]$$

The following Lemma is a standard observation used in most previous works on learning Ising models including [29, 189, 108] and others.

Lemma 55. *A* (λ_1, λ_2)*-bounded RBM satisfies* $\delta_P(d) \geq (1 - \tanh(\lambda_1))^d$.

Proof. In the $d = 1$ case this follows from the law of total expectation as $\mathbb{E}[X_i | H, X_{\sim i}] = \tanh(b_i^{(1)} + \sum_j W_{ij} H_j)$ and the term inside the tanh has magnitude at most λ_1 by definition. For general d the result follows by induction, by using the above argument for a single spin and then applying the induction hypothesis to the model where than spin is plus and where that spin is minus, since these models are also (λ_1, λ_2)-bounded RBMs. □

Lemma 56. *Let $\hat{P}$ denote the distribution returned by Algorithm* DISTRIBUTION-FROMPREDICTORS *and let P be the true distribution. Let $\log P(x) = \sum_S w_S x_S$ and $\log \hat{P}(x) = \sum_S \hat{w}_S x_S$ be the Fourier expansions of the log-likelihoods. Then*

$$\begin{aligned}\mathbf{SKL}(\hat{P}, P) &\leq \sum_S |w_S - \hat{w}_S| \\ &\leq \sum_i \frac{2^{|\mathcal{N}(i)|/2+1}}{\sqrt{\delta_P(|\mathcal{N}(i) \cup \widehat{\mathcal{N}}(i)|)}} \sqrt{\mathbb{E}_{X'}[(\tanh^{-1}(\hat{f}_i(X')) - \tanh^{-1}(\mathbb{E}_P[X_i|X_{\sim i}]))^2]}\end{aligned}$$

where $X' \sim Uni(\{\pm 1\}^n)$.

Proof. Define w_S to be the true coefficient in the true MRF potential. By Lemma 53 and Holder's inequality we know $\mathbf{SKL}(P, \hat{P}) \leq 2\sum_S |\hat{w}_S - w_S|$. Then by Jensen's inequality and the Cauchy-Schwarz inequality,

$$\begin{aligned}\sum_S |\hat{w}_S - w_S| &\leq \sum_S \frac{1}{|S|} \sum_{i \in S} |\hat{w}_{S,i} - w_S| \\ &= \sum_i \sum_{S: i \in S} \frac{1}{|S|} |\hat{w}_{S,i} - w_S| \\ &\leq \sum_i 2^{|\mathcal{N}(i)|/2} \sqrt{\sum_{S: i \in S} (\hat{w}_{S,i} - w_S)^2}.\end{aligned}$$

Now using Plancherel's theorem [148], the fact that $f_i(x) = \tanh\left(\sum_{S: i \in S} w_S x_{S \setminus \{i\}}\right)$, and the definition of $\delta_P(d)$ gives the result. □

6.8 Guarantees for Learning Feedforward Networks (with Arbitrary Distribution)

In this section we prove upper and lower bounds for learning one-layer feedforward networks with f_β activations in the hidden units and inputs X drawn from an arbitrary distribution such that $\|X\|_\infty \leq 1$.

6.8.1 Preliminaries: Optimal Approximation of Analytic Functions

Identify $\mathbb{C}$ with $\mathbb{R}^2$ by taking x to be real and y to be the imaginary component of a complex number z. Define $\mathcal{E}_\rho$ to be the region bounded by the ellipse in $\mathbb{C} = \mathbb{R}^2$ centered at the origin with equation $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ with semi-axes $a = \frac{1}{2}(\rho + \rho^{-1})$ and $b = \frac{1}{2}|\rho - \rho^{-1}|$; the focii of the ellipse are ± 1. In the present context, this is sometimes referred to as a *Bernstein ellipse*. For an arbitrary function $f : [-1, 1] \to \mathbb{R}$, let $E_D(f)$ denote the error of the best polynomial approximation of degree D in infinity norm on the interval $[-1, 1]$ of f, i.e.

$$E_D(f) := \min_{P:\deg(P)\leq D} \max_{x\in[-1,1]} |f(x) - P(x)|. \tag{6.3}$$

The following theorem of Bernstein exactly characterizes the asymptotic rate at which $E_D(f)$ shrinks:

Theorem 59 (Theorem 7.8.1, [53]). *Let f be a function defined on $[-1, 1]$. Let ρ_0 be the supremum of all ρ such that f has an analytic extension on the interior of $\mathcal{E}_\rho$. Then*

$$\limsup_{D\to\infty} \sqrt[D]{E_D(f)} = \frac{1}{\rho_0}$$

where we interpret the rhs as ∞ when $\rho_0 = 0$.

For the definition of what it means for the function to be analytic on a region of the complex plane, we refer to a text on complex analysis such as [177]. For our application we need only the upper bound and we need a quantitative estimate for finite degree d. In the proof of the upper bound in [53], the following result is proved:

Theorem 60 (Quantitative Variant of Theorem 7.8.1, [53]). *Suppose f is analytic on the interior of $\mathcal{E}_{\rho_1}$ and $|f(z)| \leq M$ on the closure of $\mathcal{E}_{\rho_1}$. Then*

$$E_D(f) \leq \frac{2M}{\rho_1 - 1}\rho_1^{-D}.$$

This quantitative variant was previously used in [112] as part of a construction of

low-degree approximations to the ReLU activation with specific properties. Note that when applying this theorem, we should center f so that the constant M is small, since adding constants to f will obviously not change $E_d(f)$.

6.8.2 Approximation Guarantees for f_β Family of Activations

Recall that the activations f_β were defined in Theorem 53 to be $f_\beta(x) = \frac{1}{\beta}\tanh^{-1}(\beta\tanh(x))$. Recall that if $\beta = 1$ then $f_\beta(x) = x$ so the function is analytic everywhere on $\mathbb{C}$, and if $\beta = 0$ is is tanh so it is meromorphic. For the remaining values of $\beta \in (0,1)$, the function f_β is slightly more complicated (it has branch cuts), however we show it is still nicely behaved near the real line.

Lemma 57. *For $\beta \in [0,1]$ the function f_β is analytic on the strip $\{x+iy : |y| < \pi/2\}$.*

Proof. Observe that

$$f'_\beta(z) = \frac{1-\tanh^2(z)}{1-\beta^2\tanh^2(z)}..$$

Since tanh is analytic except at points of the form $z = \frac{\pi}{2}i + \pi k i$, the only other possible poles are solutions to $\beta^2\tanh^2(z) = 1$, i.e. solutions to $\tanh(z) = \pm 1/\beta$. Recalling that $\tanh^{-1}(z) = \frac{1}{2}(\log(1+z) - \log(1-z))$ and taking into account the branch cut from $(-\infty, 0]$ for the logarithm, we see that the solutions to $\tanh(z) = 1/\beta$ are of the form

$$z = \frac{1}{2}\log\frac{1+1/\beta}{1/\beta - 1} + \frac{\pi i}{2} + k\pi i$$

and for $\tanh(z) = -1/\beta$ of the form

$$z = \frac{1}{2}\log\frac{1/\beta - 1}{1+1/\beta} + \frac{\pi i}{2} + k\pi i$$

for $k \in \mathbb{Z}$. In particular we see that f'_β is analytic on the strip $\{x+iy : |y| < \pi/2\}$ so f_β is as well (since the region is simply connected, this can be proved by path integration [177]). □

To get a quantitative upper bound we will need to bound (the centered version of) f_β on the Bernstein ellipse, which will require us to back away from the singularities of

f'_β on the lines $y = \pm\pi/2$. The following Lemma proves that f'_β is uniformly bounded in a slightly smaller region:

Lemma 58. *For all $\beta \in [0,1]$, $|f'_\beta(z)| \le 2$ everywhere on the closed strip $\{x+iy : |y| \le \pi/4\}$.*

Proof. Observe that

$$f'_\beta(z) = \frac{1-\tanh^2(z)}{1-\beta^2\tanh^2(z)} = \frac{\cosh^2(z)-\sinh^2(z)}{\cosh^2(z)-\beta^2\sinh^2(z)}$$
$$= \frac{1}{1+(1-\beta^2)\sinh^2(z)} = \frac{1}{1+(1-\beta^2)\frac{\cosh(2z)-1}{2}}$$

using the identies $\cosh^2(x)-\sinh^2(x)=1$ and $\sinh^2(z)=\frac{\cosh(2z)-1}{2}$. Since $\cosh(2x+2iy) = \frac{e^{2x+2iy}+e^{-2x-2iy}}{2}$ we see that under the assumption $|y| \le \pi/4$ that $\cosh(2x+2iy)$ lies in the right half plane, therefore $|1+(1-\beta^2)\frac{\cosh(2z)-1}{2}| \ge |1-(1-\beta^2)/2| \ge 1/2$ which proves the result. □

Lemma 59. *For any $\beta \in [0,1]$, arbitrary $h \in \mathbb{R}$, and any $R \ge 0$,*

$$E_D(f_\beta(Rx+h)) \le \frac{4R(1+2R)}{(1+1/2R)^D}$$

Proof. Just for this proof define $g_{\beta,h}(x) := f_\beta(Rx+h) - f_\beta(h)$. We prove this bound by application of Bernstein's theorem. By Lemma 57 we know that f_β is analytic on the strip $\{x+iy : |y| < \pi/2\}$ so in particular it is analytic on the closed strip $\{x+iy : |y| \le \pi/4\}$, and by Lemma 58 we know that $|f'_\beta| \le 2$ on the closed strip.

We now compute ρ so that $R\mathcal{E}_\rho$ is contained in the latter strip. We solve

$$\frac{1}{2}(\rho - \rho^{-1}) = \frac{\pi}{4R}$$

which gives $\rho^2 - \frac{\pi}{2R}\rho - 1 = 0$ so $\rho = \frac{\pi/2R+\sqrt{\pi^2/4R^2+4}}{2} > 1+1/2R$. Since $|g'_{\beta,h}(z)| \le R|f'_\beta| \le 2R$ on the closure of the ellipse, it follows by the mean-value theorem that $|g_{\beta,h}| \le 2(1+1/2R)R \le 1+2R$ on $\mathcal{E}_{1+1/2R}$ and applying Theorem 60 gives the result. □

6.8.3 Learning Feedforward Networks under ℓ_∞ Bounded Input

Since the final activation in our network is tanh, we recall some useful facts about logistic regression and the logistic loss which we will use.

Definition 33. *The* logistic loss *is defined to be*

$$\ell(v, y) := \log(1 + e^{-2vy}).$$

We note that the factor of 2 in the exponent and the normalization differ depending on convention.

The following facts about the logistic loss which can be checked from the definition (or see a reference such as [165]):

Fact 3. *The following are true if $y \in \{\pm 1\}$ is fixed:*

1. *$\ell(v, y)$ is convex and 2-Lipschitz in v.*
2. *$\ell(v, y) = -\log \Pr(\hat{Y} = y)$ where $\hat{Y}$ is a $\{\pm 1\}$-valued random variable with expectation $\tanh(v)$.*
3. *$\frac{\partial}{\partial v}\ell(v, y) = \frac{-2ye^{-2vy}}{1+e^{-2vy}}$ and $\frac{\partial^2}{\partial v^2}\ell(v, y) = \frac{2}{1+\cosh(2v)}$.*

Furthermore if Y is a $\{\pm 1\}$-valued random variable (and v is deterministic) then

4. *$\mathbb{E}_Y \ell(v, Y) = \mathbf{KL}(\mathcal{L}(Y), \mathcal{L}(\hat{Y})) + H(Y)$ where $\hat{Y}$ is defined above, $\mathcal{L}(Y)$ denotes the law of random variable Y, $\mathbf{KL}$ denotes the Kullback-Liebler divergence and H denotes the Shannon entropy.*

We recall the following Theorem which states the agnostic learning guarantee for fitting ℓ_1-constrained predictors in logistic loss, i.e. the logistic version of the Lasso:

Theorem 61 (Consequence of Theorem 26.15 of [165])**.** *Suppose that X is a random vector in $\mathbb{R}^n$ such that $\|X\|_\infty \leq 1$ almost surely and Y is an arbitrary $\{\pm 1\}$-valued*

random variable. Then with probability at least $1-\delta$*, simultaneously for all* w *with* $\|w\|_1 \leq R$ *it holds that*

$$\hat{\mathbb{E}}[\ell(w \cdot X, Y)] \leq \mathbb{E}[\ell(w \cdot X, Y)] + 4R\sqrt{\frac{2\log(2n)}{m}} + 2R\sqrt{\frac{2\log(2/\delta)}{m}}$$

where $\hat{\mathbb{E}}$ *denotes the empirical expectation over* m *i.i.d. copies* $(X_1, Y_1), \ldots, (X_m, Y_m)$ *of* (X, Y)*.*

In order to bound the ℓ_1 norm of our predictor we will need the following Lemmas:

Lemma 60 ([168], Lemma 2.13 of [76])**.** *Suppose* $p(x) = \sum_{i=0}^{D} \beta_i x$ *and* $|p(x)| \leq M$ *for* $x \in [-1, 1]$*, then* $\sum_{i=0}^{D} \beta_i^2 \leq (D+1)(4e)^{2D} M^2$*.*

Lemma 61. *Suppose that* $p(x) = \sum_{i=0}^{D} a_i (w \cdot x)^i = \sum_{\alpha} u_\alpha x^\alpha$*. Then*

$$\sum_{\alpha} |u_\alpha| \leq \sqrt{\sum_i a_i^2}(1 + \|w\|_1)^D.$$

Proof. For any multi-index α let $w_\alpha := \prod_{i \in \alpha} w_i$ and observe by the multinomial theorem

$$p(w \cdot x) = \sum_i a_i (w \cdot x)^i = \sum_i a_i \sum_{|\alpha|=i} \binom{i}{\alpha} w_\alpha x^\alpha.$$

Therefore by the triangle inequality, multinomial theorem, and Cauchy-Schwarz inequality

$$\sum_{\alpha} |u_\alpha| \leq \sum_i |a_i| \sum_{|\alpha|=i} \binom{i}{\alpha} |w_\alpha| = \sum_i |a_i| \|w\|_1^i \leq \sqrt{\sum_i a_i^2 \sum_i \|w\|_1^{2i}} \leq \sqrt{\sum_i a_i^2}(1+\|w\|_1)^d$$

where in the last step we used $1 + x^2 + x^4 + \cdots + x^k \leq (1+x)^k$ for $x \geq 0$. □

Theorem 62. *Suppose that* Y *is a random variable valued in* $\{\pm 1\}$*,* X *is a random vector such that* $\|X\|_\infty \leq 1$ *almost surely and*

$$\mathbb{E}[Y|X] = \tanh\left(b^{(1)} + \sum_j w_j f_{\beta_j}\left(b_j^{(2)} + \sum_k W_{jk} X_k\right)\right)$$

where $b^{(1)} \in \mathbb{R}$, $\beta_j \in [0,1]$, w *is an arbitrary real vector and* W *is an arbitrary real matrix. Let* W_j *denote* column j *of* W. *Then* ℓ_1*-constrained regression on the degree* D *monomial feature map* $\varphi_D(x) \mapsto \left(\prod_{i\in S} X_i\right)_{|S|\le d}$ *with* ℓ_1 *constraint*

$$\|w\|_1 \le R := |b^{(1)}| + \sqrt{D+1}(4e)^{D+1} \sum_j |w_j|(1+\|W_j\|_1)^{D+1}$$

returns a predictor $\hat{w}$ *such that with probability at least* $1-\delta$,

$$\begin{aligned}&\mathbb{E}[\ell(\hat{w}\cdot\varphi_d(X),Y)] - \mathbb{E}[\ell(v^*(X),Y)]\\ &\le 8\sum_j |w_j| \frac{\|W_j\|_1 + 2\|W_j\|_1^2}{(1+2/\|W_j\|_1)^D} + 4R\sqrt{\frac{2D\log(2n)}{m}} + 2R\sqrt{\frac{2\log(2/\delta)}{m}}\end{aligned}$$

where $v^*(X) := \tanh^{-1}(\mathbb{E}[Y|X]) = b^{(1)} + \sum_j w_j f_{\beta_j}\left(b_j^{(2)} + \sum_k W_{jk}X_k\right)$ *is the minimizer of the expected logistic loss over all measurable functions of* X. *The runtime is* $poly(n^D)$.

Proof. The fact that $v^*(X)$ is the minimizer of the logistic loss $\mathbb{E}[\ell(h(X),Y)]$ over all X-measurable functions h can be seen from Fact 3. To derive the bound we combine the approximation-theoretic guarantees developed in the previous section with the ℓ_1 guarantee for logistic Lasso.

For the approximation step, define w^* so that $w^*\cdot\varphi_d(X)$ is given by replacing each activation f_{β_j} by its best polynomial approximation P_j on the interval $[b_j^{(2)} - \|W_j\|_1, b_j^{(2)} + \|W_j\|_1]$. By the triangle inequality and Lemma 59, for any $x \in \{\pm 1\}^n$,

$$|v^*(x) - w^*\cdot\varphi_d(x)| \le \sum_j |w_j||(f_{\beta_j} - P_j)(b_j^{(2)} + \sum_k W_{jk}x_k)| \le 4\sum_j \frac{|w_j|\left(\|W_j\|_1 + 2\|W_j\|_1^2\right)}{(1+2/\|W_j\|_1)^D}.$$

Since the logistic loss is 2-Lipschitz (Fact 3.1), this implies that

$$\mathbb{E}[\ell(w^*\cdot\varphi_D(X),Y)] \le \mathbb{E}[\ell(v^*(X),Y)] + 8\sum_j \frac{|w_j|\left(\|W_j\|_1 + 2\|W_j\|_1^2\right)}{(1+2/\|W_j\|_1)^D}. \tag{6.4}$$

Combining Lemma 58, Lemma 60 and Lemma 61 and using the triangle inequality

shows that $\|w^*\|_1 \leq R$ where R is as specified in the Theorem statement. Then applying Theorem 61 and combining it with (6.4) gives the desired inequality bounding the error of the predictor $\hat{w}$. □

To simplify usage of this Theorem, we give the following slightly less precise bound which will be used from now on:

Corollary 6. *In the same setting as Theorem 62, if we assume that* $\|W_j\|_1 \leq \lambda$ *for every* j *and* $\lambda \geq 2$*, then with probability at least* $1 - \delta$*,* $\mathbb{E}[\ell(\hat{w} \cdot \varphi_d(X), Y)] - \mathbb{E}[\ell(v^*(X), Y)] \leq \epsilon$ *as long as the number of samples* m *satisfies* $m = \Omega((|b^{(1)}|^2 \lambda^{O(D)}) \log(2n/\delta))$ *where* $D = O(\lambda \log(\|w\|_1 \lambda/\epsilon))$ *and the runtime of the algorithm is poly*(n^D)*.*

Proof. In order to make the first term of the bound on $\mathbb{E}[\ell(\hat{w} \cdot \varphi_d(X), Y)] - \mathbb{E}[\ell(v^*(X), Y)]$ at most $\epsilon/2$, we can upper bound it by $O(\|w\|_1 \lambda^2/(1 + 2/\lambda)^D)$ and see that it suffices to take $D = \Omega(\lambda \log(\|w\|_1 \lambda/\epsilon))$. Then $R = |b^{(1)}| + \exp(O(D))\|w\|_1 \lambda^{D+1} = |b^{(1)}| + \lambda^{O(D)}$ so it suffices to take $m = \Omega((|b^{(1)}|^2 + \lambda^{O(D)}) \log(2n/\delta))$ □

Remark 19. *In the analysis of Theorem 62 we did not concern ourselves with the exact constants in the runtime. However, if we are interested in optimizing the runtime it should be noted that instead of getting a precise estimate of the empirical risk minimizer when computing the logistic regression, one can achieve a similar statistical guarantee by using a single pass of stochastic mirror descent/exponentiated gradient (see reference text [36]), e.g. as used in [108] where the needed high-probability guarantees can be found.*

6.8.4 Nearly Matching computational lower bounds

In this section, we show that the runtime guarantee of Corollary 6 is close to optima: more precisely its runtime is optimal in $\|w\|_1$ and ϵ up to a $\log \log$ factor in the exponent, and also that at least sub-exponential dependence on λ is required. We first recall the definition of this problem and a standard hardness assumption for

learning sparse parity with noise. We phrase it in terms of a testing problem versus the uniform distribution, which is equivalent to a learning formulation (i.e. recovering S below), by boosting the probability of success and using a standard reduction of removing one coordinate at a time and testing (see e.g. [184]).

Definition 34. *The* k-sparse parity with noise *distribution is the following distribution on* (X, Y) *parameterized by* $\eta \in (0, 1/2)$ *and an unknown subset* S *of size* k*:*

1. *Sample* $X \sim \mathrm{Unif}(\{-1, +1\}^n)$.
2. *With probability* $1/2 + \eta$*, set* $Y = \prod_{s \in S} X_s$*, and with probability* $1/2 - \eta$*, set* $Y = (-1) \prod_{s \in S} X_s$.

The k-sparse parity with noise *problem is to test between the uniform and* k*-sparse parity with noise with sum of probability of Type I and Type II errors upper bounded by* 0.01*, given access to an oracle which generates samples from one of the two distributions.*

Assumption 1 (Hardness of learning sparse parity with noise)**.** *Suppose* k_n *is an arbitrary sequence of positive integers with* $k_n = o(n^{1-\epsilon})$ *for any* $\epsilon > 0$ *and* n *growing, any algorithm which solve the* k*-sparse parity with noise testing problem must have runtime* $n^{\Omega(k_n)}$.

The reason for the condition $k_n = o(n^{1-\epsilon})$ is simply because the number of sets of size n is 2^n, not n^n, so small correction factors in the exponent are needed when k is comparable to n. The best known algorithm for learning sparse parity with noise runs in time $n^{0.8k_n}$ [184].

Theorem 63. *In the setting of Corollary 6 and under Assumption 1, for* $\lambda \leq 2$ *there exists families of models (one with* ϵ *a constant, one with* $\|w\|_1$ *a constant) where a runtime of*

$$n^{\Omega\left(\frac{\log(\|w\|_1/\epsilon)}{\log\log(\|w\|_1/\epsilon)}\right)}$$

is needed for any algorithm to achieve ϵ error with high probability, regardless of its sample complexity and even in the case of tanh *activations ($\beta_j = 0$ for all j). There also exists a sequence of models with $\lambda = \Theta(n \log(n))$ and $\|w\|_1 = O(\sqrt{n})$ which requires runtime*

$$n^{\Omega(\sqrt{\lambda/\log^2(\lambda)\log(n)}\log\|w\|_1)}$$

to achieve error $\epsilon = 0.01$ with high probability.

Proof. We first show a lower bound of $n^{\Omega(\log(\|w\|_1/\epsilon))}$ for a family of models where $\lambda \leq 1$. Recall we are proving a lower bound in the $\beta_j = 0$ case where all activations are tanh. The lower bound is shown by building a parity function out of tanh functions exactly using a simple taylor series expansion argument, under the assumption that the input to the network is in the hypercube $\{\pm 1\}^n$. The construction proceeds in a similar fashion to the sparse parity with noise lower bound for learning RBMs of bounded hidden degree established in Chapter 2. We first describe the construction of a parity function on boolean inputs $x_1, \ldots, x_k$. It suffices to build this parity with a small (constant-size) coefficient, since we can repeat it to make the coefficient larger. We start from the fact that

$$\tanh(z) = 2\sum_k \frac{(-1)^k}{\pi^{2k+2}}(1 - 1/4^{k+1})\zeta(2k+2)z^{2k+1}$$

for $|z| < \pi/2$ and recall that the Riemann ζ function does not vanish on even integers [177], so every coefficient in this expansion is nonzero. Furthermore it is known that $\zeta(n) \to 1$ as $n \to \infty$, since this follows from the power series definition of $\zeta(s) = \sum \frac{1}{n^s}$, so we can write

$$\tanh(z) = \sum_k a_{2k+1} z^{2k+1}$$

where $a_{2k+1} \neq 0$ for any k and $|a_{2k+1}| = \Theta(1/\pi^{2k+2})$. From this we can see that for some constant $c \neq 0$,

$$x_1 \cdots x_{2k+1} = c\frac{(2k+1)^{2k+1}}{a_{2k+1}} \tanh\left(\frac{x_1 + \cdots + x_{2k+1}}{2k+1}\right) + p(x)$$

where $p(x)$ is of degree at most $k-1$, using that $x_i^2 = 1$ for all i on the hypercube; here the constant c (which is close to 1) is a fixed correction factor to handle the small effect of maximum-degree terms coming from expanding higher order terms in the tanh power series. We can inductively rewrite each of the highest-order coefficients of p in terms of tanh and lower order monomials: this ultimately gives us a way to write parity as a linear combination of tanh functions. Using this, we can rewrite $\tanh(\frac{1}{4}x_1 \cdots x_{2k+1})$ as a two-layer tanh network with $\|w\|_1 = k^{O(k)}$ and $\lambda \leq 1$. Taking $\epsilon = 1/16$ and using the hardness of k-sparse parity with noise, we get that the runtime for learning the corresponding network is at least $n^{\Omega(k)} = n^{\Omega(\log(\|w\|_1)/\log\log(\|w\|_1))}$.

We can similarly prove a lower bound of $n^{\Omega(\log(1/\epsilon)/\log\log(1/\epsilon))}$ for constant $\lambda, \|w\|_1$ by using the same method to convert $\tanh(\eta x_1 \cdots x_{2k+1})$ into a two-layer network and by taking $\eta = k^{-\Theta(k)}$ so that the ℓ_1 norm of the coefficients is shrunk to be at most 1. Taking $\epsilon = \Theta(\eta)$ and using the sparse parity with noise lower bound as above gives the result.

Finally, we give a lower bound showing exponential dependence on λ is necessary. We use the well-known fact that a parity can be written as a small sum of threshold functions [86]. For k even,

$$x_1 \cdots x_k = \mathbb{1}[x_1 + \cdots + x_k \geq -k] - 2(\mathbb{1}[x_1 + \cdots + x_k \geq -k+2] - \mathbb{1}[x_1 + \cdots + x_k \geq -k+4] + \cdots)$$

with a total of $k+1$ terms in the sum on the rhs. We now consider replacing each threshold function with the approximation $\mathbb{1}[a \geq b] \approx \frac{1+\tanh(\lambda'(a-b+1/2))}{2}$ for some $\lambda' > 0$. Note that the error of this approximation for a singe threshold unit and integers a, b is maximized when $a - b = 0$ where the error is $\frac{1-\tanh(\lambda'/2)}{2} = O(e^{-\lambda'})$. Therefore by Holder's inequality, the error in approximating $x_1 \cdots x_k$ by replacing all of the threshold functions is $O(ke^{-\lambda'}) = O(ke^{-\lambda/(k+1/2)})$, where we used that $\lambda = (k+1/2)\lambda'$ where λ is the hidden node ℓ_1 norm as used previously. By adding a tanh nonlinearity on top of the approximate parity, this gives an approximate construction of sparse parity with noise.

Taking $k = \sqrt{n}$ and $\lambda = \Theta(k^2 \log(n))$ we see that the resulting model is **TV**-

distance $n^{-\Theta(k)}$ from sparse parity with noise, so any algorithm with runtime $cn^{-\Theta(k)}$ cannot distinguish this model from sparse parity with noise with probability better than 75% for sufficiently small constant $c > 0$. From the assumed hardness of learning sparse parity with noise, any algorithm succeeding to distinguish this model from the uniform distribution with sufficiently small error probability requires runtime $n^{\Omega(k)} = n^{\sqrt{\lambda/\log^2(\lambda)\log(n)}\log\|w\|_1}$. □

Remark 20. *In the second construction in the proof of Theorem 63, based off of approximating threshold functions, the computational lower bound becomes stronger if we allow the algorithm access to less data (recall that for a fixed noise level, $\Theta(k\log n)$ samples suffice information-theoretically for sparse parity with noise). If we only allow to use $\Theta(k\log n)$ samples as information-theoretically required, then we can take $\lambda = \Theta(k(\log k + \log\log n))$ and the runtime required is $n^k = n^{\lambda/(\log\log n + \log(\lambda))}$.*

6.9 Learning RBMs by Learning Feedforward Networks

6.9.1 Structure and Distribution Learning Guarantees

In this section we discuss application of the prediction guarantees from the previous section to structure and distribution learning. As motivation, recall that in undirected graphical models the *Markov blanket* or *neighborhood* of a node i, the minimal set of nodes which separate node i from the rest of the model in the underlying graph, is one of the most interesting pieces of information to learn about a node. By the Markov property, node i interacts directly only with nodes in its Markov blanket, in the sense that X_i is conditionally independent of all other nodes X_k given the values of nodes X_j for all j in the markov blanket of i. Learning the markov blanket of all nodes, equivalently learning the underlying graph of the Markov Random Field, is referred to as *structure learning*. It is also well known (see e.g. [109]) that once we have performed structure learning, distribution learning (e.g. in total variation distance) becomes a conceptually straightforward task as it can typically be reduced

to solving low-dimensional regression problems.

As explained in the introduction, learning the structure requires a non-degeneracy condition on neighbors (recall the definition of η-nondegeneracy from above). In the introduction, we stated that if all edges are η-nondegenerate then we can learn the structure perfectly; in the next Theorem, we state a slightly more precise result giving the result we can successfully test between non-neighbors and η-nondegenerate neighbors, without requiring nondegeneracy on the entire model. Since our guarantee holds with high probability, using the union bound it immediately gives a result for structure recovery under η-nondegeneracy.

Theorem 64. *Let i and j be two visible nodes in a (λ_1, λ_2)-bounded RBM. Let H_0 be the hypothesis that nodes i and j are not two-hop neighbors and H_1 the hypothesis that nodes i and j are η-nondegenerate two-hop neighbors. Given $\delta > 0$ and $m = \Omega(\lambda_2^{O(D)} \log(2n/\delta))$ i.i.d. samples where $D = O(\lambda_2 \log(\lambda_1 \lambda_2/\eta))$, we can test in time $poly(n^D)$ between H_0 and H_1 with sum of Type I and Type II errors upper bounded by δ.*

Proof. We run the following testing procedure:

1. Run the ℓ_1 regression algorithm from Theorem 53 to predict X_i from $X_{\sim i}$ and from $X_{\sim i,j}$.

2. Repeat the previous step with i and j reversed.

3. If the decrease in prediction accuracy for removing i or j is at least $3\eta/4$ in either step 1 or step 2, reject H_0.

That this works follows by combining Theorem 53 and Corollary 6, by choosing $\epsilon = \eta/8$ under H_0 the difference in prediction error is at most 2ϵ whereas under H_1 it must be at least $\eta - 2\epsilon$. □

Assuming that all 2-hop neighbors in the RBM are η-nondegenerate, the above Theorem lets us recover the structure of the RBM (its 2-hop neighborhoods) in time $poly(n^D)$. In the following remark, we explain how large D is in the regimes where we know polynomial time sampling from the RBM is possible:

Remark 21 (Comparison to polynomial time sampling regimes)**.** *Dobrushin's uniqueness criterion is probably the most well-known sufficient condition for sampling to be possible in polynomial time in a general pairwise model. Dobrushin's condition is that for every node i, the total ℓ_1-norm of the edges touching node i is at most 1, where the mixing time guarantees for Glauber dynamics become worse as the maximum norm approaches 1 (see [122]). This condition is tight in the example of the Ising model on the complete graph (Curie-Weiss), or for the bipartite complete graph (i.e. dense RBM) with all edge weights positive and equal and an equal number of visible and hidden units.*

Under Dobrushin's uniqueness criterion on the RBM, we have that $\lambda_1, \lambda_2 \leq 1$ so $D = O(\log(1/\eta))$. As mentioned above, we cannot compute η in terms of just the edge weights for general models, but if we for example assume the model is d-regular and has all edge weights equal to $+1/d$ and no external field then it is not too hard to show that $\eta = \Omega(1/d^2)$ (see Chapter 2), so in this case the overall runtime is $n^{\log(d)}$. We expect that under Dobrushin's condition $\eta = \Omega(1/d^2)$ except in perhaps some rare degenerate situations. This means the runtime is improved by an exponential factor in the exponent compared to what one gets by just applying the RBM to MRF reduction, since learning d-wise MRFs is known to require n^d time in general [108].

In some other interesting contexts, it is also known that polynomial time sampling can only be guaranteed when $\lambda_1, \lambda_2 = O(1)$: for antiferromagnetic Ising models on bounded degree graphs with equal edge weights the sharp result is known for every d [169, 70, 172] and embedding these Ising models as RBMs with hidden nodes of degree 2 in a straightforward way gives models with $\lambda_1, \lambda_2 = O(1)$ and $\eta = \Omega(1/d^2)$ (see Example 14 above).

For distribution learning we will need the following technical Lemma, which is proved in Appendix 6.9.2 using the local Rademacher complexity framework [14]. Informally it says that if X is a random variable with a density with respect to the uniform measure on $\{\pm 1\}^n$ that is lower bounded by a constant, then given a number of samples m which is large with respect to the size of the domain the natural estimator of $\tanh^{-1}(\mathbb{E}[Y|X])$ has error which converges at a $1/m$ rate, which

generalizes the case of estimating the (exponential-family parameterization of) mean, the $n = 0$ case, in a natural way. Since the bound depends exponentially on n, we will only apply it in settings where we expect n is small. Similar bounds are used in previous works including [34, 29] and proved using different methods, though they are not quite as optimized (e.g. deriving this result from Lemma 3.2 of [29] would give a $1/\gamma^2$ dependence); this bound can be shown to be optimal up to constants.

Lemma 62. *Suppose that X is a random variable valued in $\{\pm 1\}^n$ with $\Pr(X = x) \geq \gamma/2^n$ for every x and Y is a random variable valued in $\{\pm 1\}$. Suppose that $|\mathbb{E}[Y|X]| \leq r$ for $r < 1$. Let $\hat{\mathbb{E}}[Y|X]$ be the empirical conditional expectation of Y given X based upon m i.i.d. samples of (X, Y) and define $h(X) := \min(\max(\mathbb{E}[Y|X], r), -r)$. Then with probability at least $1 - \delta$,*

$$\mathbb{E}[(\tanh^{-1}(h(X)) - \tanh^{-1}(\mathbb{E}[Y|X]))^2] \lesssim \frac{2^n/\gamma + \log(1/\delta)}{(1 - r^2)^2 m}$$

where $\lesssim$ denotes inequality up to an absolute constant.

We present the proof of this lemma in the subsequent subsection. From this Lemma we straightforwardly get the right result for learning a sparse RBM with known 2-hop neighborhoods.

Algorithm 4 DistributionFromStructure

1: We assume for every node i we are given a recovered neighborhood $\widehat{\mathcal{N}}(i)$. $\widehat{\mathcal{N}}(i)$
2: For every node i with neighborhood $\widehat{\mathcal{N}}(i)$, let $f_i(X) := \widehat{\mathbb{E}}[X_i|X_{\widehat{\mathcal{N}}(i)}]$ be the empirical conditional expectation of X_i given $X_{\widehat{\mathcal{N}}(i)}$.
3: Return the output of Algorithm DistributionFromPredictors run with these f_i.

Lemma 63. *For any (λ_1, λ_2)-bounded RBM where the maximum two-hop degree of any visible node is at most d_2 and where $\|b^{(1)}\|_\infty \leq B$, for $\delta > 0$ and $m = \Omega\left(n^2 \left(\frac{2}{(1-\tanh(\lambda_1))}\right)^{d_2+1} \log(n/\delta)/\epsilon^4\right)$ we have that with probability at least $1 - \delta$, Algorithm* DistributionFromStructure *given m samples and $\widehat{\mathcal{N}}(i) = \mathcal{N}(i)$ for every i returns a distribution $\hat{P}$ which is ϵ-TV close to the distribution of the RBM.*

Furthermore, if $w_S, \hat{w}_S$ are as defined as in Lemma 56 then

$$2\mathbf{TV}(P, \hat{P})^2 \leq \mathbf{SKL}(P, \hat{P}) \leq \sum_S |w_S - \hat{w}_S| \leq \epsilon^2.$$

Proof. By Lemma 56, Lemma 55 and Lemma 62 we have

$$\begin{aligned}
\mathbf{SKL}(\hat{P}, P) &\leq \sum_S |w_S - \hat{w}_S| \\
&\leq \sum_i \frac{2^{d_2/2+1}}{(1-\tanh(\lambda_1))^{d_2/2}} \sqrt{\mathbb{E}_{X \sim Uni(\{\pm 1\}^n)}[(\tanh^{-1}(h_i(X)) - \tanh^{-1}(\mathbb{E}_P[X_i|X_{\sim i}])^2]} \\
&\leq \sum_i \frac{2^{d_2/2+1}}{(1-\tanh(\lambda_1))^{d_2}} \sqrt{\mathbb{E}_{X_{\mathcal{N}(i)}}[(\tanh^{-1}(h_i(X)) - \tanh^{-1}(\mathbb{E}_P[X_i|X_{\sim i}])^2]} \\
&\leq \sum_i \frac{2^{d_2/2+1}}{(1-\tanh(\lambda_1))^{d_2}} \sqrt{\frac{2^{d_2}/(1-\tanh(\lambda_1))^{d_2} + \log(n/\delta)}{(1-\tanh(\lambda_1)^2)^2 m}}
\end{aligned}$$

and by Pinsker's inequality $\mathbf{TV}(\hat{P}, P)^2 \leq \mathbf{SKL}(\hat{P}, P)/2$ so the result follows. □

Theorem 65. *Suppose that all visible nodes in an RBM which are neighbors in the Markov blanket sense are η-nondegenerate neighbors, and that maximum 2-hop degree of any visible node is at most d_2. Then given $\delta > 0$ and $m = \Omega(\lambda_2^{O(D)} \log(2n/\delta) + n^2 \left(\frac{2}{(1-\tanh(\lambda_1))}\right)^{d_2+1} \log(n/\delta)/\epsilon^4)$ i.i.d. samples where $D = O(\lambda_2 \log(\lambda_1 \lambda_2/\eta))$ samples, Algorithm* DistributionFromStructure *run with the set of η-nondegenerate neighbors output by Theorem 64 returns with probability at least $1-\delta$ a distribution which is ϵ-TV close to the true distribution of the RBM.*

Proof. This follows by combining Theorem 64 and Lemma 63. □

Remark 22. *If we do not assume that all neighbors are η-nondegenerate, then by Theorem 68 it is impossible to get a nontrivial distribution learning guarantee assuming the hardness of learning sparse parity with noise, in the sense that the naive approach of forgetting the RBM structure entirely and using MRF learning results (e.g. [108]) cannot be improved.*

6.9.2 Proof of Lemma 62

We recall the statement of Lemma 62. Suppose that X is a random variable valued in $\{\pm 1\}^n$ with $\Pr(X = x) \geq \gamma/2^n$ for every x and Y is a random variable valued in $\{\pm 1\}$. Suppose that $|\mathbb{E}[Y|X]| \leq r$ for $r < 1$. Let $\hat{\mathbb{E}}[Y|X]$ be the empirical conditional expectation of Y given X based upon m i.i.d. samples of (X, Y) and define $h(X) := \min(\max(\mathbb{E}[Y|X], r), -r)$. Then with probability at least $1 - \delta$,

$$\mathbb{E}[(\tanh^{-1}(h(X)) - \tanh^{-1}(\mathbb{E}[Y|X]))^2] \lesssim \frac{2^n}{\gamma(1-r^2)^2 m} + \frac{\log(1/\delta)}{(1-r^2)^2 m}$$

We will prove the result by proving the analogous result without the $\tanh^{-1}$ first, as Lemma 64. The following general result reduces this to computing the local Rademacher complexity of the corresponding function class.

Theorem 66 (Corollary 5.3 of [14])**.** *Suppose that $\mathcal{F}$ is a class of functions from $\mathcal{X}$ to $[-1, 1]$ and $\ell(\hat{y}, y)$ is a loss which satisfies:*

1. *ℓ is L-Lipschitz in $\hat{y}$.*

2. *There is a constant $B \geq 1$ such that for any random variable X supported on $\mathcal{X}$ and random variable Y on $[-1, 1]$*

$$\mathbb{E}(f(X) - f^*(X))^2 \leq B\mathbb{E}[\ell(f(X), Y) - \ell(f^*(X), Y)]$$

where $f^*(X)$ is a minimizer of $\mathbb{E}[\ell(f(X), Y)]$ which we assume exists.

Then if $\psi(r)$ is a sub-root function (meaning a monotonically increasing non-negative function with $\psi(r)/\sqrt{r}$ monotonically decreasing) such that

$$\psi(r) \geq BL\mathbb{E} \sup_{f \in \mathcal{F}, L^2\mathbb{E}[(f-f^*)^2] \leq r} \frac{1}{m} \sum_{i=1}^{m} \sigma_i (f - f^*)(X_i) \tag{6.5}$$

where the σ_i are i.i.d. Rademacher random variables, then for any $r \geq \psi(r)$ with

probability at least $1-\delta$

$$\mathbb{E}[\ell(\hat{f}(X),Y)-\ell(f^*(X),Y)] \lesssim \frac{r}{B}+\frac{(L+B)\log(1/\delta)}{m}$$

where the notation $\lesssim$ *hides an absolute constant.*

Lemma 64. *Under the same setup as Lemma 62,*

$$\mathbb{E}[(h(X)-\mathbb{E}[Y|X])^2] \lesssim \frac{2^n}{\gamma m}+\frac{\log(1/\delta)}{m}.$$

Proof. We consider $\mathcal{F}$ the class of arbitrary functions from $\mathcal{X}$ to $[-r,r]$ and take $\ell(\hat{y},y) := (\hat{y}-y)^2$ to be the square loss so $L=2$ and $B=1$ satisfy the conditions above. It is clear from the definition of h that it is the empirical risk minimizer for this function class and loss. Since this class is convex we can take $\psi(r)$ to be defined by the rhs of (6.5) (Lemma 3.4 of [14]) and it remains to compute the fixed point of ψ. Thus if we write $g := f - f^*$

$$\psi(r) = 2\mathbb{E}\sup_{f:4\mathbb{E}[g^2]\le r}\frac{1}{m}\sum_{i=1}^m \sigma_i g(X_i)$$

and we observe by the assumption $\Pr(X=x)\ge \gamma/2^n$ that

$$\mathbb{E}_X[g^2] \ge \gamma \mathbb{E}_{X'\sim Uni(\{\pm 1\}^n)}[g(X')^2] = \gamma\sum_S \widehat{g}(S)^2$$

by Plancherel's Theorem [148] where $\widehat{g}(S)$ denotes the Fourier coefficient of g corresponding to set S, so that $g(x)=\sum_S \widehat{g}(S)x_S$ where $x_S=\prod_{s\in S}x_s$. Therefore by the above, the Cauchy-Schwarz inequality, and Jensen's inequality we have

$$\begin{aligned}\psi(r) &= 2\mathbb{E}\sup_{g:4\mathbb{E}[g^2]\le r}\frac{1}{m}\sum_{i=1}^m\sigma_i g(X_i)\\ &\le 2\mathbb{E}\sup_{g:\sum_S \hat{g}(S)^2\le r/4\gamma}\frac{1}{m}\sum_S \hat{g}(S)\frac{1}{m}\sum_{i=1}^m\sigma_i(X_i)_S\end{aligned}$$

$$\leq \sqrt{r/\gamma}\mathbb{E}\frac{1}{m}\sqrt{\sum_S \left(\sum_{i=1}^m \sigma_i(X_i)_S\right)^2}$$
$$\leq \frac{\sqrt{r}}{m\sqrt{\gamma}}\sqrt{\mathbb{E}\sum_S \left(\sum_{i=1}^m \sigma_i(X_i)_S\right)^2} = \frac{\sqrt{r}}{\sqrt{m\gamma}}2^{n/2}.$$

Solving for the fixed point of $r = \frac{\sqrt{r}}{\sqrt{m\gamma}}2^{n/2}$ gives $r^* = \frac{2^n}{\gamma m}$ so the result follows from Theorem 66. □

Proof of Lemma 62. Recall that the derivative of $\tanh^{-1}$ at x is $\frac{1}{1-x^2}$. Therefore on the domain $[-r, r]$ the function $\tanh^{-1}$ is $\frac{1}{1-r^2}$ Lipschitz. Therefore by the mean value theorem,

$$\mathbb{E}[(\tanh^{-1}(h(X)) - \tanh^{-1}(\mathbb{E}[Y|X]))^2] \leq \frac{1}{(1-r^2)^2}\mathbb{E}[(h(X) - \mathbb{E}[Y|X])^2]$$

and applying Lemma 64 gives the result. □

6.9.3 Matching Computational Lower Bounds

In the following sequence of theorems we show that our runtime guarantees for structure learning of RBMs cannot be significantly improved. The first result relies in part on the representation of sparse parity with noise given in Chapter 2; this embedding is constructed in a similar way to the first embedding used in Theorem 63. It shows the dependence on λ_1 and η is correct when asking for structure recovery.

Theorem 67. *In the same setup as Theorem 64 and under Assumption 1, there exists a family of instances parameterized by n going to infinity with $\lambda_2 \leq 2$ such that any algorithm which is able to achieve structure recovery for a model with all neighbors being η-nondegenerate requires runtime $n^{\Omega(\log(\lambda_1/\eta)/\log\log(\lambda_1/\eta))}$, regardless of its sample complexity.*

Proof. In Chapter 2, it was shown that for any fixed constant η (say $\eta = 1/8$), there exists an embedding of k-sparse parity with noise into an RBM where every hidden unit has incoming edges of total ℓ_1 norm upper bounded by 2 (i.e. satisfying $\lambda_1 \leq 2$)

and there are $2^{O(k)}$ hidden units; it can be checked straightforwardly that for $\eta = 1/8$ that $\lambda_2 = k^{O(k)}$. Therefore if we fix $\epsilon = \eta/2$ then when assuming the hardness of k-sparse parity with noise there is a $n^{\Omega(k)}$ runtime lower bound which matches since $\lambda_2 = e^{O(k)}$.

For the tightness in ϵ, by making the parity bias η exponentially small in $k \log(k)$, it's easy to check that by repeating the construction in Chapter 2 that we can make λ_2 a constant; then to find the parity with noise one needs ϵ exponentially small in $k \log k$ as well, and the hardness assumption implies the runtime must be $n^{\Omega(k)}$. □

By tensorizing this construction, we show that the η-nondegeneracy assumption is required, even if we only care about distribution learning. More precisely, we need it to learn in TV distance with runtime better than the pessimistic $n^{O(d_h)}$ result which follows from viewing the RBM as an unstructured MRF and using the result of [108].

Theorem 68. *There exists a family of RBMs with n nodes, maximum hidden node degree d_H, and $\lambda_1, \lambda_2 = O(1)$ such that any algorithm which can learn this family of RBMs within total variation distance at most $1/4$ requires $n^{\Omega(d_H)}$ time.*

Proof. The construction in Theorem 67 shows that there exists a family of RBMs given by embedding sparse parity with noise with the desired property, except that the total variation distance is only guaranteed to be $2^{-O(d_H \log(d_H))}$. By building a larger RBM consisting of $2^{d_H \log(d_H)}$ disjoint copies of the original RBM (note that the resulting increase in n is a multiplicative factor independent of the original n), we can boost the total variation distance to be arbitrarily close to 1. □

In order to give lower bounds with respect to λ_2 for fixed η, we need a significantly more involved argument. We first recall an approximate construction of parity (with low levels of noise) from [133]:

Theorem 69 (Theorem 7 of [133])**.** *There exists an RBM network with $n^2 + 1$ hidden units and weights poly$(n, \log(1/\epsilon))$ such that the marginal distribution P on the visible*

units satisfies $P(x) \propto e^{f(x)}$ *for some* f *satisfying*

$$\sup_{x \in \{\pm 1\}^n} |f(x)/C - x_1 \cdots x_n| \leq \epsilon$$

where $C > 0$ *satisfies* $C = poly(\log(n), \log(1/\epsilon))$.

This construction is for a dense parity, but obviously we can make the parity as sparse as we want by adding additional visible units not connected to anything else. More significantly, since the above theorem only constructs an ϵ-approximate instance of parity with noise $\eta = O(1/2 - 1/poly(n, 1/\epsilon))$, when n or $1/\epsilon$ is large it does not seem that the resulting distribution is computationally hard to distinguish from the uniform distribution, since Gaussian elimination over $\mathbb{F}_2$ has some chance of succeeding to find the parity. Since we need ϵ to be small for the model to be indistinguishable from sparse parity with noise, this appears to be a barrier to deriving a hardness result from the above Theorem. Instead, we will prove that our result cannot be significantly improved for SQ (Statistical Query) algorithms (for a reference, see [24]). In the Statistical Query model algorithms do not have access to data, but instead have access to an SQ oracle:

Definition 35. *An oracle for the statistical query model over distribution* $\mathcal{D}$ *over* X, Y *takes input* (g, τ) *where* g *is a function* $g : \{\pm 1\}^n \times \{\pm 1\} \to [-1, 1]$ *and* τ *is a tolerance, and gives output* v *with*

$$|\mathbb{E}_{X,Y \sim D}[g(X, Y)] - v| \leq \tau.$$

Standard arguments, i.e. implementing the needed regressions using standard gradient-based methods for convex optimization shows that our algorithm for learning RBMs can be implemented in the statistical query model (in this case, the separation of X and Y in the definition above is somewhat artificial but we will take Y to be a particular visible unit in the RBM). We will show that statistical query algorithms cannot do better than subexponential dependence on λ_2.

The following theorem statements a lower bound for learning concepts of large

SQ-dimension in the Statistical Query model. The definition of SQ-dimension can be found in [24], but for our purposes the only needed fact is that the class of k-parities over the uniform distribution $\{\pm 1\}^n$ has SQ-dimension $\binom{n}{k}$ [24].

Theorem 70 ([24])**.** *Let $\mathcal{F}$ be a class of functions over $\{\pm 1\}^n$ and D a distribution such that SQ-DIM$(\mathcal{F}, D) \geq d \geq 16$. Then if all queries are made with tolerance at least $1/d^{1/3}$, then at least $d^{1/3}/2$ queries are required to learn $\mathcal{F}$ with error less than $1/2 - 1/d^3$ in the statistical query model.*

Theorem 71. *Let S be an unknown subset of $[n]$ of size k and containing n and $\mathcal{D}$ is the distribution of the RBM produced by Theorem 69 on S where the other $n - |S|$ visible units are isolated and without external field. Let $\mathcal{F}$ be the class of parities on $[n-1]$. As before, λ_2 refers to the maximum ℓ_1-norm into any hidden unit and we choose parameters so that $\lambda_2 = poly(n)$ and $\|w\|_1 = poly(n)$. There exists $\epsilon > 0$ so that no SQ algorithm with tolerance $n^{-\lambda_2^\epsilon}$ and access to $n^{\lambda_2^\epsilon}$ queries can learn $\mathcal{F}$ with error less than $1/4$.*

Proof. In Theorem 69 we take $\epsilon = \exp(-n)$ which gives $\lambda_2 = poly(n)$. The resulting RBM is then within TV distance $\exp(-n)$ of the distribution of a parity over the uniform distribution with a small amount of label noise, so an SQ algorithm for the RBM setting implies an SQ algorithm for learning parity, and the result follows from the lower bound of Theorem 70. □

6.10 Learning a Feedforward Network under the RBM distributional assumption

In this section we reverse the connection between RBMs and Feedforward networks by using RBMs with certain structural assumptions as a useful *distributional assumption* for learning feedforward network. More formally, we assume our data is generated by the following Supervised RBM.

Definition 36. *A* Supervised Restricted Boltzmann Machine *is any joint distribution over random variables X valued in $\{\pm 1\}^{n_1}$, H valued in $\{\pm 1\}^{n_2}$ and label $Y \in \{\pm 1\}$ of the form*

$$\Pr[X = x, H = h, Y = y] \propto \exp\left(\langle x, Wh\rangle + \langle h, w\rangle y + \langle b^{(1)}, x\rangle + \langle b^{(2)}, h\rangle + b^{(3)} y\right)$$

where the weight matrix W *is an arbitrary $n_V \times n_H$ matrix and* external fields/biases $b^{(1)} \in \mathbb{R}^{n_1}$, $b^{(2)} \in \mathbb{R}^{n_2}$ *and $b^{(3)}$ are arbitrary, and X is referred to as the vector of* visible unit *activations and H the vector of* hidden unit *activations.*

We make the following additional assumptions on the parameters of the model.

Assumption 2 (Minimum Ferromagnetic Interaction)**.** *For all $i \in [n_1], j \in [n_2]$ either $W_{ij} = 0$ or $W_{ij} \geq \alpha$.*

We do not make any assumption on the weight w to the label. Therefore the model overall is not ferromagnetic.

Assumption 3 (Sparsity)**.** *For all $i \in [n_1]$, $\sum_{j=1}^{n_2} W_{ij} + |b_i^{(1)}| \leq \lambda$ and for either $y = -1$ or $y = 1$, for all $j \in [n_2]$ $\sum_{i=1}^{n_1} W_{ij} + |b_j^{(2)} + y w_j| \leq \lambda$.*

Here the sparsity assumption implies that under the conditioning of the label to either value, the sparsity parameter is bounded. This conditional sparsity can be exploited by an algorithm for learning the conditional distribution whereas a direct regression algorithm may be unable to gain from the same.

Remark 23. *Observe that the generative model of X itself is not sparse since Y is connected to all hidden nodes however conditioned on knowing the label Y, the model is now sparse. This assumption is more reasonable than assuming sparsity directly on the model of X which may not hold.*

Assumption 4 (Balanced Label)**.** *For $y \in \{\pm 1\}$, $\Pr[Y = y] \geq \beta$.*

The above assumption essentially rules out trivial constant learners. Using data, it is easy to check if this assumption is satisfied or not.

As before, we can compute the conditional mean function of the label as follows:

$$\mathbb{E}[Y|X=x] = \tanh\left(b^{(3)} + \sum_j \tanh^{-1}\left(\tanh(w_j)\nu_j\right)\right)$$

where $\nu_j := \tanh\left(b_j^{(2)} + \sum_i \tanh^{-1}\left(\tanh(W_{ij})X_i\right)\right) = \tanh\left(b_j^{(2)} + \sum_i W_{ij}X_i\right)$. This represents a 2-layer neural network and in the limit of infinite hidden nodes, it can represent all 2-layer tanh networks (see Lemma 52).

Assumption 5 (Boundedness)**.** *When* $\mathbb{E}[Y|X=x]$ *is re-expressed as* $\tanh(f^*(x)+b^*)$ *for some function* f^* *with no constant term and* $b^* \in \mathbb{R}$. $|b^*| \leq B$ *for some* $B > 0$.

The above assumption intuitively says that the effect on Y that does not depend on X is bounded. B can be bounded in terms of the network parameters.

Also observe that conditioned on a fixed label,

$$\Pr[X=x, H=h|Y=y] \propto \exp\left(\langle x, Wh\rangle + \langle b^{(1)}, x\rangle + \langle b^{(2)} + wy, h\rangle\right)$$

which is a sparse, ferromagnetic RBM with arbitrary external field. Thus, we capture a neural network problem with a conditional RBM distributional assumption on the input. This distributional assumption seems more natural than the Gaussian input distribution which is extensively used in prior work. Also, this assumption allows us to leverage prior known algorithms for structure learning of ferromagnetic RBMs to learn the prediction function.

6.10.1 Preliminaries: Structure Learning of RBMs with Ferromagnetic Interactions

Consider a RBM with the following additional assumptions:

Assumption 6 (Minimum Ferromagnetic Interaction)**.** *For all* $i \in [n_1], j \in [n_2]$ *either* $W_{ij} = 0$ *or* $W_{ij} \geq \alpha$.

Assumption 7 (Sparsity)**.** *For all* $i \in [n_1]$, $\sum_{j=1}^{n_2} W_{ij} + |b_i^{(1)}| \leq \lambda$ *and for all* $j \in [n_2]$, $\sum_{i=1}^{n_1} W_{ij} + |b_j^{(2)}| \leq \lambda$.

Under these assumptions, [75] has shown that a simple greedy algorithm based on covariance maximization suffices to learn the structure of the RBM. We use this result because the earlier analysis of Chapter 2 makes the further assumption of non-negative external fields which won't be true in our general situation. (It remains an interesting open problem to combine the good sample complexity guarantees of the analysis of Chapter 2 with the general setting of [75].)

The crucial structural property that [75] use is their algorithm is the following strengthening of the FKG inequality,

Lemma 65 (Lemma 2 of [75]). *For any observed nodes u, v and set $S \subseteq [n_1] \backslash \{u, v\}$,*

$$\mathsf{Cov}(u, v | X_S = x_S) := \mathbb{E}[X_u X_v | X_S = x_S] - \mathbb{E}[X_u | X_S = x_S] \, \mathbb{E}[X_v | X_S = x_S] \geq \alpha^2 \exp(-12\lambda).$$

Subsequently they define *average conditional covariance* $\mathrm{Cov}^{\mathsf{Avg}}(u, v|S) = \mathbb{E}_{x_S}[\mathrm{Cov}(u, v | X_S = x_S)]$ which straightforwardly is lower bounded by an application of the above lemma. Their final algorithm essentially greedily maximizes this average conditional covariance to build the neighborhood.

Theorem 72 (Theorem 2 of [75]). *Consider M samples $\mathcal{S}$ drawn from a RBM with arbitrary external field satisfying the given assumptions. For $\tau = \frac{\alpha^2}{2} \exp(-12\lambda)$ and $\delta = \exp(-2\lambda)/2$, with probability $1 - \zeta$,* LearnRBMNbhd$(u, \tau, \mathcal{S})$ *outputs* exactly *the two-hop neighborhood of observed variable u for*

$$M \geq \Omega\left((\log(1/\zeta) + T^* \log(n)) \frac{2^{2T^*}}{\tau^2 \delta^{2T^*}} \right) \text{ and } T^* = \frac{8}{\tau^2}.$$

Moreover, the algorithm runs in time $O(T^* M n)$.

6.10.2 Prediction from Distribution Learning

Here we will present our algorithm for learning the supervised RBM followed by a proof of its correctness. Instead of learning the label function directly, we will instead first learn the underlying generative model of X conditioned on a particular value of the label and use this knowledge to predict Y.

Theorem 73. *Given a supervised RBM satisfying Assumption 2, 3, 4 and 5, there exists an algorithm with sample complexity* $m = n^2 \exp(\lambda)^{\exp(O(\lambda))}(1/\alpha)^{O(1)}(1/\beta)^{O(1)} \log(n/\delta)/\epsilon^2$ *and runtime* $poly(m)$ *returns hypothesis* h *such that,*

$$\mathbb{E}[\ell(h(X), Y] - \mathbb{E}[\ell(h^*(X), Y] \leq \epsilon$$

where ℓ *is the logistic loss and* h^* *is the minimizer of the logistic loss.*

Remark 24. *For an example where this algorithm is better than if we have no distributional assumptions, observe that we can construct a ferromagnetic RBM where* $\mathbb{E}[Y|X]$ *is a sparse parity function by adapting in a straightforward way the reduction used in the proof of the part of Theorem 63 with bounded* λ *(the use of* tanh *as opposed to* f_β *in that construction is not fundamental, or we can use a finite version of Lemma 52), since the hidden units in that proof all have nonnegative weights. It's clear why Algorithm* LearnSupervisedRBMNBhd *is better than an algorithm which doesn't know the input distribution: under the true input distribution, the visible units involved in the parity are correlated so the algorithm can find them, which makes learning the sparse parity easy.*

Our main algorithm can be broken down into three main steps: 1) Use greedy maximization (similar to Algorithm 1 of [75]) to first learn the two-hop neighborhood $\mathcal{N}(i)$ of each observed variable i w.r.t. the hidden layer conditioned on the label, 2) For each observed variable X_i, learn the distribution for $X|Y = y$ for $y = \pm 1$, and 3) Use the estimated distribution to compute $\mathbb{E}[Y|X]$.

Structure Learning For notation simplicity, we will overload notation and represent $\text{Cov}^{\mathsf{Avg}}(u, v|S, Y) = \mathbb{E}_{x_S,y}[\text{Cov}(u, v|X_S = x_S, Y = y)]$ where $\text{Cov}(u, v|X_S = x_S, Y = y) = \mathbb{E}[X_u X_v|X_S = x_S, Y = y] - \mathbb{E}[X_u|X_S = x_S, Y = y]\ \mathbb{E}[X_v|X_S = x_S, Y = y]$. Then for structure learning, our algorithm essentially follows Algorithm 1 of [75] with the slight modification of conditioning w.r.t. Y.

Theorem 74. *Consider m samples $\mathcal{S}$ drawn from a supervised RBM satisfying Assumption 2, 3 and 4. For $\tau = \frac{\beta\alpha^2}{2}\exp(-12\lambda)$ and $\delta = \exp(-2\lambda)/2$, with probability $1-\zeta$,*

LearnSupervisedRBMNbhd$(u, \tau, \mathcal{S})$ *outputs* exactly *the two-hop neighbors of observed variable u w.r.t. the hidden layer, with*

$$m \geq \Omega\left((\log(1/\zeta) + T^* \log(n))\,\frac{2^{2T^*}}{\tau^2\beta\delta^{2T^*}}\right) \text{ and } T^* = \frac{8}{\tau^2}.$$

Moreover, the algorithm runs in time $O(T^*Mn)$.

Proof. In order to apply Theorem 72 to our setting, the only two properties we need to show are 1) given the conditioning of Y, the average conditional covariance bound still holds, that is, $\mathrm{Cov}^{\mathsf{Avg}}(u, v|S \cup \{0\})$ is lower bounded for all $S \subseteq [n_2]\backslash\{u, v\}$ for v in the two-hop neighborhood of u, 2) $\Pr[X_S = x_S, Y = y]$ for all x_S and y. We have,

$$\mathrm{Cov}^{\mathsf{Avg}}(u, v|S, Y) = \sum_{y\in\pm 1}\sum_{x_S\in\{\pm 1\}^{|S|}} \Pr[X_S = x_S, Y = y]\mathrm{Cov}(u, v|X_S = x_S, Y = y)$$

By Assumption 3, we know that either for $y = 1$ or $y = -1$ (say $y = 1$ WLOG), the resulting RBM is sparse therefore we can apply Lemma 65 to the ones conditioned on $y = 1$. Also, we know that $\mathrm{Cov}(u, v|X_S = x_S, Y = y) \geq 0$ for all x_S and y due to FKG inequality for ferromagnetic RBMs. This implies that,

$$\begin{aligned}
\mathrm{Cov}^{\mathsf{Avg}}(u, v|S, Y) &\geq \sum_{x_S\in\{\pm 1\}^{|S|}} \Pr[X_S = x_S, Y = 1]\mathrm{Cov}(u, v|X_S = x_S, Y = 1) \\
&\geq \sum_{x_S\in\{\pm 1\}^{|S|}} \Pr[X_S = x_S, Y = 1]\alpha^2\exp(-12\lambda) \\
&\geq \Pr[Y = 1]\alpha^2\exp(-12\lambda) \geq \beta\alpha^2\exp(-12\lambda).
\end{aligned}$$

For the second part, let us order the elements of S of size k as $s_1, \ldots, s_k$, then we have

$$\Pr[X_S = x_S, Y = y] = \Pr[Y = y] \times \Pr[X_{s_1} = x_{s_1}|Y = y] \times \Pr[X_{s_2} = x_{s_2}|X_{s_1} = x_{s_1}, Y = y] \times \ldots$$

$$\times \Pr[X_{s_k} = x_{s_k} | X_{s_1} = x_{s_1}, \ldots, X_{s_{k-1}} = x_{s_{k-1}}, Y = y]$$

Since l_1-norm to the observed nodes is bounded by λ, by Bresler's property (see [29]) we have $\Pr[X_{s_r} = x_{s_r} | X_{s_1} = x_{s_1}, \ldots, X_{s_r} = x_{s_r}, Y = y] \geq \delta$. This implies that $\Pr[X_S = x_S, Y = y] \geq \beta\delta^{|S|}$ for all values of x_S and y. Now by applying Theorem 72 with the correct parameters, we get the required result. □

Distribution Learning Given the neighborhood of each observed node, we run Algorithm DistributionFromStructure and subsequently use Lemma 63 to guarantee that we obtin the weights of the unnormalized MRFs for distributions $X|Y = y$ for $y \in \{\pm 1\}$ up to epsilon accuracy. More formally,

Lemma 66. *Let the maximum two-hop degree of any visible node is at most d_2 and $\|b^{(1)}\|_\infty \leq B$. For $\delta > 0$ and $m = \Omega\left(n^2 \left(\frac{2}{(1-\tanh(\lambda))}\right)^{d_2+1} \log(n/\delta)/\epsilon^2\right)$ we have that with probability at least $1 - \delta$, Algorithm* DistributionFromStructure *given m samples and $\widehat{\mathcal{N}}(i) = \mathcal{N}(i)$ for every i returns unnormalized MRFs of $X|Y = y$ for $y \in \{\pm 1\}$ with coefficients $\hat{f}_S^{(y)}$ that are close to the coefficients of the true unnormalized MRFs $f_S^{(y)}$, that is,*

$$\sum_S |\hat{f}_S^{(y)} - f_S^{(y)}| \leq \epsilon.$$

Constructing the Predictor Observe that the joint distribution of X and Y can be represented as,

$$\Pr[X = x, Y = y] \propto \exp\left(\sum_S f_S^{(1)} x_S \mathbb{1}[y = 1] + \sum_S f_S^{(-1)} x_S \mathbb{1}[y = -1] + b^* y\right)$$

for some b^* and coefficients of the true unnormalized MRFs $f_S^{(y)}$ corresponding to conditioning of $Y = y$. This gives us,

$$\mathbb{E}[Y|X = x] = \tanh\left(\sum_S \frac{(f_S^{(1)} - f_S^{(-1)})}{2} x_S + b\right) \approx_\varepsilon \tanh\left(\sum_S \frac{(\hat{f}_S^{(1)} - \hat{f}_S^{(-1)})}{2} x_S + b\right)$$

Since we have estimates of $f_S^{(y)}$, to learn the predictor for Y we only need to find b^* which we can find by minimizing ℓ snce it is convex. Let $h_b = \sum_S \frac{(f_S^{(1)} - f_S^{(-1)})}{2} x_S + b$ and $\hat{h}_b = \sum_S \frac{(\hat{f}_S^{(1)} - \hat{f}_S^{(-1)})}{2} x_S + b$. We minimize $\hat{E}[\ell(h_b(X), Y)]$ over b and suppose the minimizer is $\hat{b}$. By Fact 3.3, $\ell(\hat{h}_b(X), Y) \leq \ell(h_b(X), Y) + 4\epsilon$. By Fact 3.4, h_{b^*} is the minimizer of the logistic loss. Then we have,

$$\hat{\mathbb{E}}[\ell(h_b(X), Y)] \leq \hat{\mathbb{E}}[\ell(\hat{h}_{b^*}(X), Y)] + 4\epsilon \leq \hat{\mathbb{E}}[\ell(h_{b^*}(X), Y)] + 8\epsilon.$$

Last we need a generalization bound that holds for our hypothesis class. For this we bound the Rademacher complexity (see [165] for more background) of the class of functions $\ell \circ \mathcal{H}$ where $\mathcal{H} := \{h_b || b| \leq B\}$.

$$\begin{aligned}
\mathcal{R}_m(\ell \circ \mathcal{H}) &\leq 2\mathcal{R}_m(\mathcal{H}) \\
&= \mathbb{E}_\sigma \left[\sum_{b||b| \leq B} \frac{1}{m} \sum_{i=1}^m \sigma_i h_b(x^{(i)}) \right] \\
&= \mathbb{E}_\sigma \left[\sum_{b||b| \leq B} \frac{1}{m} \sum_{i=1}^m \sigma_i \sum_S (f_S^{(1)} - f_S^{(-1)}) x_S + 2b \right] \\
&= 2\mathbb{E}_\sigma \left[\sum_{b||b| \leq B} \frac{1}{m} \sum_{i=1}^m \sigma_i b \right] \\
&= 2B\mathbb{E}_\sigma \left[\frac{1}{m} \left| \sum_{i=1}^m \sigma_i \right| \right] \\
&\leq \frac{2B}{\sqrt{m}}.
\end{aligned}$$

Here the first inequality follows from the contraction lemma (see [120]) and the last from standard properties of Radmeacher variables. Now applying Theorem 26.5 from [165] we get

$$|\mathbb{E}[\ell(h_b(X), Y)] - \mathbb{E}[\ell(\hat{h}_b(X), Y)]| \leq \frac{2B}{\sqrt{m}} + c\sqrt{\frac{\log(1/\delta)}{\sqrt{m}}}$$

where c is the maximum value of logistic loss by any hypothesis in the class. Observe that by Fact 3.4, logistic loss at h_{b^*} is bounded by a constant. Hence by Lipschitzness, we know that loss anywhere will be bounded by $O(\max(1, B))$. Therefore choosing $m \geq \Omega(B^2 \log(1/\delta)/\epsilon^2)$ suffices to get within ϵ. Combining this with before we get that the loss is within $O(\varepsilon)$ of the best loss.

Proof of Theorem 73 First, the algorithm runs LearnSupervisedRBMMbhd for each node to learn the structure of the induced RBM exactly with the given samples

$$m_1 = \exp(\lambda)^{\exp(O(\lambda))}(1/\alpha)^{O(1)}(1/\beta)^{O(1)} \log(n/\delta).$$

With the structure, we run DistributionFromStructure to learn both the induced RBMs for each conditioning of the label using $m_2 \geq \Omega\left(n^2 \left(\frac{2}{(1-\tanh(\lambda))}\right)^{d_2+1} \log(n/\delta)/\epsilon^2\right)$ samples where d_2 is the max 2-hop neighborhood size. Note that the dependence on λ is greater in m_1 than m_2. Subsequently, given the unnormalized mrfs, we run a simple optimization to find the bias term of the predictor using $m_3 \geq \Omega(B^2 \log(1/\delta)/\epsilon^2)$ samples. Combining the learnt mrf and the bias term, we get our hypothesis.

Remark 25. *If the model is not ferromagnetic, it is also possible and we expect it may be advantageous in some models to still use a similar indirect approach based on Bayes rule for learning a predictor of Y, but using the result of Theorem 53 instead of the greedy structure recovery method used in this section. The disadvantage of this approach is of course that its runtime for achieving structure recovery is slower.*

6.11 Additional Experimental Data

Figure 6-3 contains samples generated from the model trained on MNIST images. For reference, we also include samples from the true MNIST and FashionMNIST training sets in the same format as Figure 6-2 and Figure 6-1.

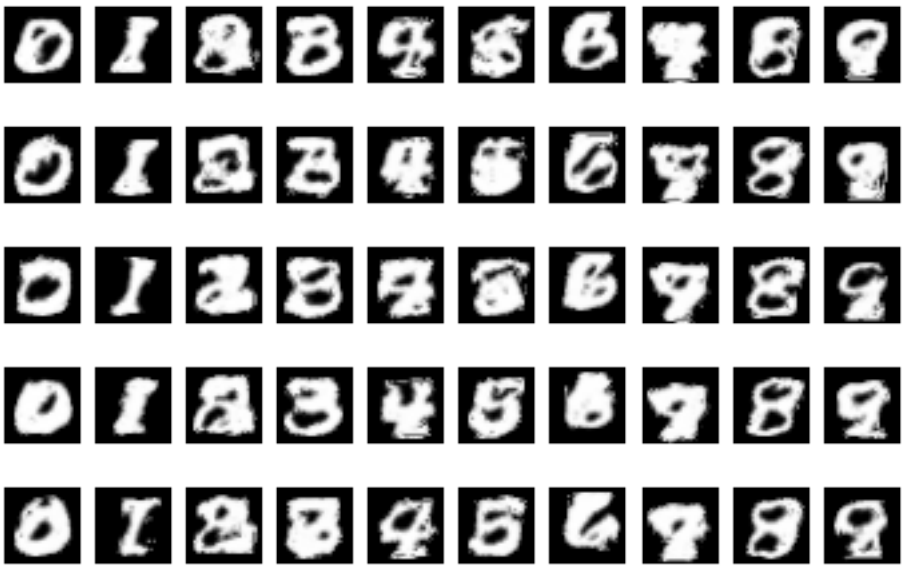

Figure 6-2: Five i.i.d. samples for each MNIST class, drawn from the trained model by Gibbs sampling.

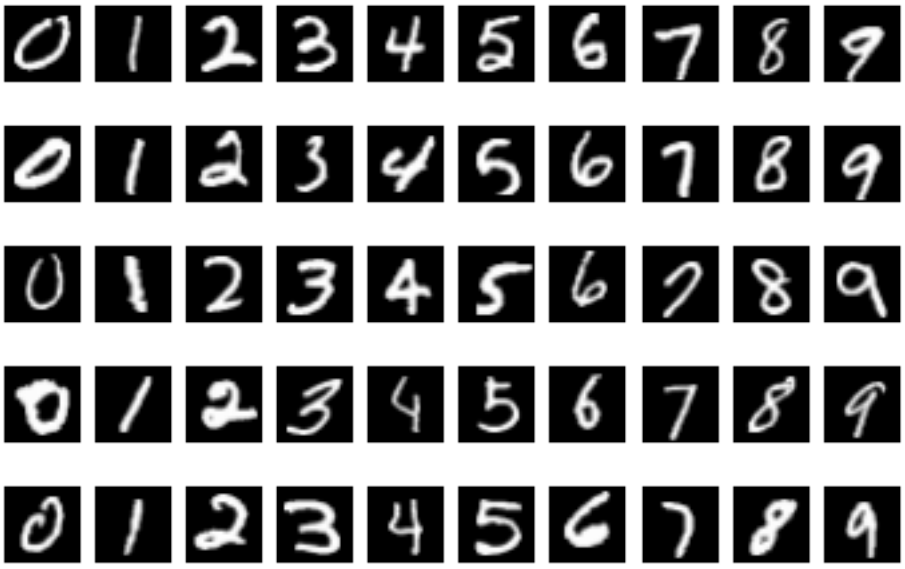

Figure 6-3: Reference MNIST images chosen randomly from training set.

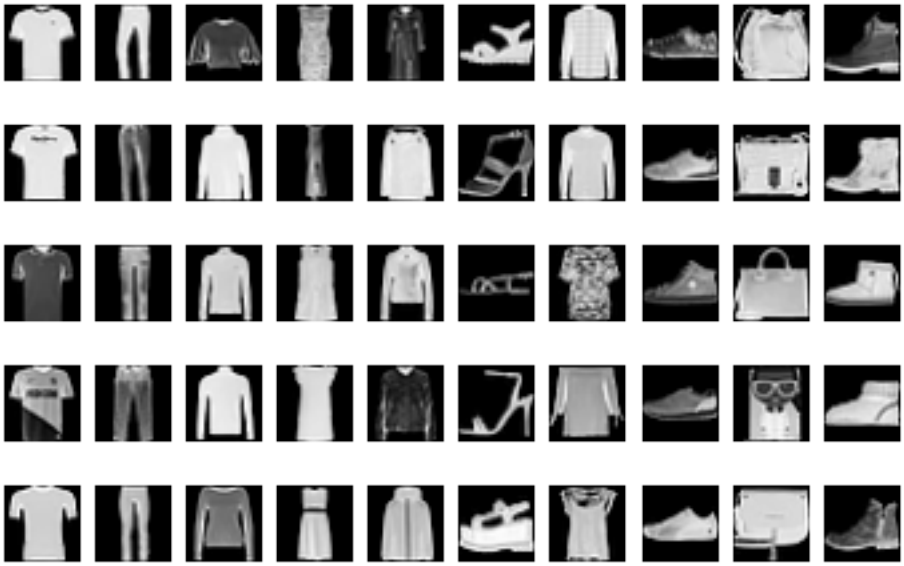

Figure 6-4: Reference FashionMNIST samples from training set.

Bibliography

[1] Nir Ailon and Noga Alon. Hardness of fully dense problems. *Information and Computation*, 205(8):1117–1129, 2007.

[2] Michael Aizenman, Joel L Lebowitz, and David Ruelle. Some rigorous results on the sherrington-kirkpatrick spin glass model. *Communications in mathematical physics*, 112(1):3–20, 1987.

[3] Sarah R Allen and Ryan O'Donnell. Conditioning and covariance on caterpillars. In *Information Theory Workshop (ITW), 2015 IEEE*, pages 1–5. IEEE, 2015.

[4] Animashree Anandkumar, Vincent YF Tan, Furong Huang, and Alan S Willsky. High-dimensional gaussian graphical model selection: Walk summability and local separation criterion. *Journal of Machine Learning Research*, 13(Aug):2293–2337, 2012.

[5] Animashree Anandkumar and Ragupathyraj Valluvan. Learning loopy graphical models with latent variables: Efficient methods and guarantees. *The Annals of Statistics*, pages 401–435, 2013.

[6] Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*. Number 118. Cambridge university press, 2010.

[7] James Anderson and Carsten Peterson. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019, 1987.

[8] Steen A Andersson, David Madigan, Michael D Perlman, et al. A characterization of markov equivalence classes for acyclic digraphs. *Annals of statistics*, 25(2):505–541, 1997.

[9] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. Provable bounds for learning some deep representations. In *International Conference on Machine Learning*, pages 584–592, 2014.

[10] Fanny Augeri. A transportation approach to the mean-field approximation. *arXiv preprint arXiv:1903.08021*, 2019.

[11] Afonso S Bandeira, Dmitriy Kunisky, and Alexander S Wein. Computational hardness of certifying bounds on constrained pca problems. *arXiv preprint arXiv:1902.07324*, 2019.

[12] Boaz Barak, Prasad Raghavendra, and David Steurer. Rounding semidefinite programming hierarchies via global correlation. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 472–481. IEEE, 2011.

[13] 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.

[14] Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[15] Alexander Barvinok. Computing the permanent of (some) complex matrices. *Foundations of Computational Mathematics*, 16(2):329–342, 2016.

[16] Alexander Barvinok. Combinatorics and complexity of partition functions. *Algorithms and Combinatorics*, 30, 2017.

[17] Anirban Basak and Sumit Mukherjee. Universality of the mean-field for the potts model. *Probability Theory and Related Fields*, 168(3-4):557–600, 2017.

[18] Katia Basso, Adam A Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. Reverse engineering of regulatory networks in human b cells. *Nature genetics*, 37(4):382, 2005.

[19] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

[20] William Bialek, Andrea Cavagna, Irene Giardina, Thierry Mora, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M Walczak. Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*, 109(13):4786–4791, 2012.

[21] Lucien Birgé et al. An alternative point of view on lepski's method. *Lecture Notes-Monograph Series*, 36:113–133, 2001.

[22] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[23] Joseph K Blitzstein and Jessica Hwang. *Introduction to probability*. Chapman and Hall/CRC, 2014.

[24] Avrim Blum, Merrick Furst, Jeffrey Jackson, Michael Kearns, Yishay Mansour, and Steven Rudich. Weakly learning dnf and characterizing statistical query learning using fourier analysis. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 253–262, 1994.

[25] Andrej Bogdanov, Elchanan Mossel, and Salil Vadhan. The complexity of distinguishing markov random fields. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 331–342. Springer, 2008.

[26] Béla Bollobás. *Modern graph theory*, volume 184. Springer Science & Business Media, 2013.

[27] Erik G Boman, Doron Chen, Ojas Parekh, and Sivan Toledo. On factor width and symmetric h-matrices. *Linear algebra and its applications*, 405:239–248, 2005.

[28] Christian Borgs, Jennifer T Chayes, László Lovász, Vera T Sós, and Katalin Vesztergombi. Convergent sequences of dense graphs ii. multiway cuts and statistical physics. *Annals of Mathematics*, 176(1):151–219, 2012.

[29] Guy Bresler. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 771–782. ACM, 2015.

[30] Guy Bresler. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 771–782. ACM, 2015.

[31] Guy Bresler and Rares-Darius Buhai. Learning restricted boltzmann machines with sparse latent variables. *Advances in Neural Information Processing Systems*, 33, 2020.

[32] Guy Bresler, David Gamarnik, and Devavrat Shah. Hardness of parameter estimation in graphical models. In *Advances in Neural Information Processing Systems*, pages 1062–1070, 2014.

[33] Guy Bresler, Frederic Koehler, and Ankur Moitra. Learning restricted boltzmann machines via influence maximization. In *Symposium on Theory of Computing (STOC)*, 2019.

[34] Guy Bresler, Elchanan Mossel, and Allan Sly. Reconstruction of markov random fields from samples: Some observations and algorithms. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 343–356. Springer, 2008.

[35] 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.

[36] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[37] Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278, 2014.

[38] T Tony Cai, Weidong Liu, Harrison H Zhou, et al. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 44(2):455–488, 2016.

[39] Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[40] Gruia Calinescu, Chandra Chekuri, Martin Pál, and Jan Vondrák. Maximizing a submodular set function subject to a matroid constraint. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 182–196. Springer, 2007.

[41] Emmanuel J Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on pure and applied Mathematics*, 67(6):906–956, 2014.

[42] Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, 2012.

[43] Wei-Kuo Chen and Si Tang. On convergence of bolthausen's tap iteration to the local magnetization. *arXiv preprint arXiv:2011.00495*, 2020.

[44] C Chow and Cong Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

[45] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223, 2011.

[46] Amin Coja-Oghlan and Will Perkins. Bethe states of random factor graphs. *arXiv preprint arXiv:1709.03827*, 2017.

[47] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[48] Abhimanyu Das and David Kempe. Submodular meets spectral: greedy algorithms for subset selection, sparse approximation and dictionary selection. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1057–1064. Omnipress, 2011.

[49] Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Optimal phylogenetic reconstruction. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 159–168, 2006.

[50] Fernandez de la Vega and Marek Karpinski. Approximation complexity of nondense instances of max-cut. *Electronic Colloquium on Computational Complexity*, 2006. TR06-101.

[51] Wenceslas Fernandez de la Vega and Claire Kenyon-Mathieu. Linear programming relaxations of maxcut. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 53–61. Society for Industrial and Applied Mathematics, 2007.

[52] Amir Dembo and Andrea Montanari. Ising models on locally tree-like graphs. *Ann. Appl. Probab.*, 20(2):565–592, 2010.

[53] Ronald A DeVore and George G Lorentz. *Constructive approximation*, volume 303. Springer Science & Business Media, 1993.

[54] Jian Ding, James R Lee, and Yuval Peres. Cover times, blanket times, and majorizing measures. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 61–70. ACM, 2011.

[55] Irit Dinur. The pcp theorem by gap amplification. *Journal of the ACM (JACM)*, 54(3):12, 2007.

[56] Simon Du, Jason Lee, Yuandong Tian, Aarti Singh, and Barnabas Poczos. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. In *International Conference on Machine Learning*, pages 1339–1348, 2018.

[57] Ronen Eldan. Gaussian-width gradient complexity, reverse log-sobolev inequalities and nonlinear large deviations. *arXiv preprint arXiv:1612.04346*, 2016.

[58] Ronen Eldan. Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probability Theory and Related Fields*, pages 1–19, 2019.

[59] Ronen Eldan and Renan Gross. Decomposition of mean-field gibbs distributions into product measures. *Electronic Journal of Probability*, 23, 2018.

[60] Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on Learning Theory*, pages 907–940, 2016.

[61] Ethan R Elenberg, Rajiv Khanna, Alexandros G Dimakis, and Sahand Negahban. Restricted strong convexity implies weak submodularity. *arXiv preprint arXiv:1612.00804*, 2016.

[62] Richard S Ellis and Charles M Newman. The statistics of curie-weiss models. *Journal of Statistical Physics*, 19(2):149–161, 1978.

[63] Joseph Felsenstein. *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA, 2004.

[64] Miroslav Fiedler and Vlastimil Ptak. On matrices with non-positive off-diagonal elements and positive principal minors. *Czechoslovak Mathematical Journal*, 12(3):382–400, 1962.

[65] Dimitris Fotakis, Michael Lampis, and Vangelis Th Paschos. Sub-exponential approximation schemes for csps: From dense to almost sparse. In *33rd Symposium on Theoretical Aspects of Computer Science*, 2016.

[66] Sacha Friedli and Yvan Velenik. *Statistical mechanics of lattice systems: a concrete mathematical introduction*. Cambridge University Press, 2017.

[67] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[68] Alan Frieze and Ravi Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

[69] Marylou Gabrié, Eric W Tramel, and Florent Krzakala. Training restricted boltzmann machine via the thouless-anderson-palmer free energy. In *Advances in neural information processing systems*, pages 640–648, 2015.

[70] Andreas Galanis, Daniel Štefankovič, and Eric Vigoda. Inapproximability of the partition function for the antiferromagnetic ising and hard-core models. *Combinatorics, Probability and Computing*, 25(4):500–559, 2016.

[71] Weihao Gao, Ashok V Makkuva, Sewoong Oh, and Pramod Viswanath. Learning one-hidden-layer neural networks under general input distributions. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1950–1959, 2019.

[72] Rong Ge, Rohith Kuditipudi, Zhize Li, and Xiang Wang. Learning two-layer neural networks with symmetric inputs. In *International Conference on Learning Representations*, 2019.

[73] Mrinalkanti Ghosh, Fernando Granha Jeronimo, Chris Jones, Aaron Potechin, and Goutham Rajendran. Sum-of-squares lower bounds for sherrington-kirkpatrick via planted affine planes. *arXiv preprint arXiv:2009.01874*, 2020.

[74] James Glimm and Arthur Jaffe. *Quantum physics: a functional integral point of view*. Springer Science & Business Media, 2012.

[75] Surbhi Goel. Learning ising and potts models with latent variables. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3557–3566, Online, 26–28 Aug 2020. PMLR.

[76] Surbhi Goel, Varun Kanade, Adam Klivans, and Justin Thaler. Reliably learning the relu in polynomial time. In *Conference on Learning Theory*, pages 1004–1042, 2017.

[77] Surbhi Goel and Adam Klivans. Eigenvalue decay implies polynomial-time learnability for neural networks. In *Advances in Neural Information Processing Systems*, pages 2192–2202, 2017.

[78] Surbhi Goel and Adam Klivans. Learning depth-three neural networks in polynomial time. *arXiv preprint arXiv:1709.06010*, 2017.

[79] Surbhi Goel, Adam Klivans, and Frederic Koehler. From boltzmann machines to neural networks and back again. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[80] Surbhi Goel, Adam R. Klivans, and Raghu Meka. Learning one convolutional layer with overlapping patches. In Jennifer G. Dy and Andreas Krause 0001, editors, *ICML*, volume 80 of *JMLR Workshop and Conference Proceedings*, pages 1778–1786. JMLR.org, 2018.

[81] Leslie Ann Goldberg and Mark Jerrum. The complexity of ferromagnetic ising with local fields. *Combinatorics, Probability and Computing*, 16(1):43–61, 2007.

[82] Robert B Griffiths. Rigorous results for ising ferromagnets of arbitrary spin. *Journal of Mathematical Physics*, 10(9):1559–1565, 1969.

[83] Robert B. Griffiths, C. A. Hurst, and S. Sherman. Concavity of magnetization of an ising ferromagnet in a positive external field. *Journal of Mathematical Physics*, 11(3):790–795, 1970.

[84] Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric algorithms and combinatorial optimization*. Springer Science & Business Media, 2012.

[85] Venkatesan Guruswami and Ali Kemal Sinop. Lasserre hierarchy, higher eigenvalues, and approximation schemes for graph partitioning and quadratic integer programming with PSD objectives. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 482–491. IEEE, 2011.

[86] András Hajnal, Wolfgang Maass, Pavel Pudlák, Mario Szegedy, and György Turán. Threshold circuits of bounded depth. *Journal of Computer and System Sciences*, 46(2):129–154, 1993.

[87] Linus Hamilton, Frederic Koehler, and Ankur Moitra. Information theoretic properties of markov random fields, and their algorithmic applications. In *Advances in Neural Information Processing Systems*, pages 2460–2469, 2017.

[88] Mohamed Hebiri and Johannes Lederer. How correlations influence lasso prediction. *IEEE Transactions on Information Theory*, 59(3):1846–1854, 2012.

[89] Geoffrey E Hinton. Deep belief networks. *Scholarpedia*, 4(5):5947, 2009.

[90] Geoffrey E Hinton. A practical guide to training restricted boltzmann machines. In *Neural networks: Tricks of the trade*, pages 599–619. Springer, 2012.

[91] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[92] Geoffrey E Hinton and Ruslan R Salakhutdinov. Replicated softmax: an undirected topic model. In *Advances in neural information processing systems*, pages 1607–1614, 2009.

[93] Cho-Jui Hsieh, Mátyás A Sustik, Inderjit S Dhillon, Pradeep K Ravikumar, and Russell Poldrack. Big & quic: Sparse inverse covariance estimation for a million variables. In *Advances in neural information processing systems*, pages 3165–3173, 2013.

[94] Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1, 2012.

[95] Shuai Huang, Jing Li, Liang Sun, Jieping Ye, Adam Fleisher, Teresa Wu, Kewei Chen, Eric Reiman, Alzheimer's Disease NeuroImaging Initiative, et al. Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation. *NeuroImage*, 50(3):935–949, 2010.

[96] Dmitry Ioffe and Yvan Velenik. A note on the decay of correlations under δ-pinning. *Probability theory and related fields*, 116(3):379–389, 2000.

[97] Vishesh Jain, Frederic Koehler, and Elchanan Mossel. The mean-field approximation: Information inequalities, algorithms, and complexity. In *Conference on Learning Theory (COLT)*, 2018.

[98] Vishesh Jain, Frederic Koehler, and Elchanan Mossel. The vertex sample complexity of free energy is polynomial. In *Conference on Learning Theory (COLT)*, 2018.

[99] 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.

[100] M. Jerrum and A. Sinclair. Polynomial-time approximation algorithms for ising model (extended abstract). In *Automata, Languages and Programming*, pages 462–475, 1990.

[101] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[102] Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636, 2007.

[103] David Karger, David Karger, and Nathan Srebro. Learning markov networks: Maximum bounded tree-width graphs. In *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, pages 392–401. Society for Industrial and Applied Mathematics, 2001.

[104] Robert W Keener. *Theoretical statistics: Topics for a core course*. Springer, 2011.

[105] Jonathan A. Kelner, Frederic Koehler, Raghu Meka, and Ankur Moitra. Learning some popular gaussian graphical models without condition number bounds. In *Neural Information Processing Systems (NeurIPS)*, 2020 (Spotlight Presentation).

[106] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[107] S Kirkpatrick and D Sherrington. Solvable model of a spin-glass. *Phys. Rev. Lett*, 35(26):1792–1796, 1975.

[108] Adam Klivans and Raghu Meka. Learning graphical models using multiplicative weights. In *FOCS*, 2017.

[109] Adam Klivans and Raghu Meka. Learning graphical models using multiplicative weights. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 343–354. IEEE, 2017.

[110] Adam R Klivans and Alexander A Sherstov. Cryptographic hardness for learning intersections of halfspaces. *Journal of Computer and System Sciences*, 75(1):2–12, 2009.

[111] Frederic Koehler. Fast convergence of belief propagation to global optima: Beyond correlation decay. In *Advances in Neural Information Processing Systems*, 2019.

[112] Frederic Koehler and Andrej Risteski. The comparative power of relu networks and polynomial kernels in the presence of sparse latent structure. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[113] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[114] Vladimir Koltchinskii and Stanislav Minsker. l_1-penalization in functional linear regression with subgaussian design. *Journal de l'École polytechnique-Mathématiques*, 1:269–330, 2014.

[115] Filip Korč, Vladimir Kolmogorov, Christoph H Lampert, et al. Approximating marginals using discrete energy minimization. In *ICML Workshop on Inferning: Interactions between Inference and Learning*. Citeseer, 2012.

[116] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[117] Steffen Lauritzen, Caroline Uhler, and Piotr Zwiernik. Maximum likelihood estimation in gaussian models under total positivity. *arXiv preprint arXiv:1702.04031*, 2017.

[118] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[119] Steffen L Lauritzen. Elements of graphical models. *Lectures from the XXXVIth International Probability Summer School in St-Flour, France*, 2011.

[120] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[121] Tsung-Dao Lee and Chen-Ning Yang. Statistical theory of equations of state and phase transitions. ii. lattice gas and ising model. *Physical Review*, 87(3):410, 1952.

[122] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[123] 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.

[124] Jingcheng Liu, Alistair Sinclair, and Piyush Srivastava. The ising partition function: Zeros and deterministic approximation. *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 986–997, 2017.

[125] Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

[126] Philip M Long and Rocco Servedio. Restricted boltzmann machines are hard to approximately evaluate or simulate. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 703–710, 2010.

[127] Christopher W Lynn and Daniel D Lee. Maximizing activity in ising networks via the tap approximation. *arXiv preprint arXiv:1803.00110*, 2018.

[128] Yifei Ma, Roman Garnett, and Jeff Schneider. Sigma-optimality for active learning on gaussian random fields. In *Advances in Neural Information Processing Systems*, pages 2751–2759, 2013.

[129] Satyaki Mahalanabis and Daniel Stefankovic. Subset selection for gaussian markov random fields. *arXiv preprint arXiv:1209.5991*, 2012.

[130] Eran Malach and Shai Shalev-Shwartz. A provably correct algorithm for deep learning that actually works. *arXiv preprint arXiv:1803.09522*, 2018.

[131] Dmitry M Malioutov, Jason K Johnson, and Alan S Willsky. Walk-sums and belief propagation in gaussian graphical models. *Journal of Machine Learning Research*, 7(Oct):2031–2064, 2006.

[132] Pasin Manurangsi and Prasad Raghavendra. A birthday repetition theorem and complexity of approximating dense csps. In *Proceedings of ICALP*, volume 80. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[133] James Martens, Arkadev Chattopadhya, Toni Pitassi, and Richard Zemel. On the representational efficiency of restricted boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 2877–2885, 2013.

[134] Claire Mathieu and Warren Schudy. Yet another algorithm for dense max cut: Go greedy. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '08, pages 176–182, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics.

[135] Nicolai Meinshausen, Peter Bühlmann, et al. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.

[136] Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

[137] Patricia Menéndez, Yiannis AI Kourmpetis, Cajo JF ter Braak, and Fred A van Eeuwijk. Gene regulatory networks from multifactorial perturbations using graphical lasso: application to the dream4 challenge. *PloS one*, 5(12):e14147, 2010.

[138] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[139] Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[140] Sidhant Misra, Marc Vuffray, and Andrey Y. Lokhov. Information theoretic optimal learning of gaussian graphical models. In *Proceedings of COLT 2020.*, 2020.

[141] Andrea Montanari. Estimating random variables from random sparse observations. *European Transactions on Telecommunications*, 19(4):385–403, 2008.

[142] Elchanan Mossel. Deep learning and hierarchal generative models. *arXiv preprint arXiv:1612.09057*, 2016.

[143] Elchanan Mossel, Ryan O'Donnell, and Rocco P Servedio. Learning juntas. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 206–212. ACM, 2003.

[144] Elchanan Mossel and Sébastien Roch. Learning nonsingular phylogenies and hidden markov models. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 366–375, 2005.

[145] Alfred Müller and Marco Scarsini. Archimedean copulae and positive dependence. *Journal of Multivariate Analysis*, 93(2):434–445, 2005.

[146] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[147] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294, 1978.

[148] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, New York, NY, USA, 2014.

[149] Dmitry Panchenko. *The Sherrington-Kirkpatrick model*. Springer Science & Business Media, 2013.

[150] Christos H Papadimitriou and Mihalis Yannakakis. Optimization, approximation, and complexity classes. *Journal of computer and system sciences*, 43(3):425–440, 1991.

[151] Giorgio Parisi. *Statistical field theory*. New York: Addison-Wesley, 1988.

[152] Viresh Patel and Guus Regts. Deterministic polynomial-time approximation algorithms for partition functions and graph polynomials. *SIAM Journal on Computing*, 46(6):1893–1919, 2017.

[153] Judea Pearl. *Causality*. Cambridge university press, 2009.

[154] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[155] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[156] Prasad Raghavendra and Ning Tan. Approximating csps with global cardinality constraints using sdp hierarchies. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 373–387. Society for Industrial and Applied Mathematics, 2012.

[157] Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, Bin Yu, et al. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[158] Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 2015.

[159] Andrej Risteski. How to calculate partition functions using convex programming hierarchies: provable bounds for variational methods. In *COLT*, 2016.

[160] Nicholas Ruozzi, Justin Thaler, and Sekhar Tatikonda. Graph covers and quadratic minimization. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1590–1596. IEEE, 2009.

[161] Itay Safran and Ohad Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *International Conference on Machine Learning*, pages 2979–2987, 2017.

[162] Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798. ACM, 2007.

[163] Narayana P Santhanam and Martin J Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012.

[164] Juliane Schäfer and Korbinian Strimmer. Learning large-scale graphical gaussian models from genomic data. In *AIP Conference Proceedings*, volume 776, pages 263–276. AIP, 2005.

[165] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[166] Shai Shalev-Shwartz, Ohad Shamir, and Karthik Sridharan. Learning kernel-based halfspaces with the 0-1 loss. *SIAM Journal on Computing*, 40(6):1623–1646, 2011.

[167] Scott Sheffield. Gaussian free fields for mathematicians. *Probability theory and related fields*, 139(3-4):521–541, 2007.

[168] Alexander A Sherstov. Making polynomials robust to noise. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 747–758, 2012.

[169] Alistair Sinclair, Piyush Srivastava, and Marc Thurley. Approximation algorithms for two-state anti-ferromagnetic spin systems on bounded degree graphs. *Journal of Statistical Physics*, 155(4):666–686, 2014.

[170] Martin Slawski and Matthias Hein. Estimation of positive definite m-matrices and structure learning for attractive gaussian markov random fields. *Linear Algebra and its Applications*, 473:145–179, 2015.

[171] Allan Sly and Nike Sun. The computational hardness of counting in two-spin models on d-regular graphs. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 361–369. IEEE, 2012.

[172] Allan Sly and Nike Sun. The computational hardness of counting in two-spin models on d-regular graphs. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 361–369. IEEE, 2012.

[173] Mahdi Soltanolkotabi. Learning relus via gradient descent. In *Advances in Neural Information Processing Systems*, pages 2007–2017, 2017.

[174] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in neural information processing systems*, pages 2199–2207, 2010.

[175] M. Steel. Recovering a tree from the leaf colourations it generatesunder a Markov model. *Appl. Math. Lett.*, 7(2):19–23, 1994.

[176] Mike Steel. *Phylogeny: Discrete and random processes in evolution*. SIAM, 2016.

[177] Elias M Stein and Rami Shakarchi. *Complex analysis*, volume 2. Princeton University Press, 2010.

[178] Michel Talagrand. *Mean field models for spin glasses. Volume I*, volume 54 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer-Verlag, Berlin, 2011. Basic examples.

[179] Michel Talagrand. *Mean Field Models for Spin Glasses. Volume II: Advanced Replica-Symmetry and Low Temperature*, volume 55 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer, Heidelberg (2011b). ISBN, 2011.

[180] Matus Telgarsky. Benefits of depth in neural networks. In *Conference on Learning Theory*, pages 1517–1539, 2016.

[181] David J Thouless, Philip W Anderson, and Robert G Palmer. Solution of'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977.

[182] 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.

[183] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.

[184] Gregory Valiant. Finding correlations in subquadratic time, with applications to learning parities and juntas. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 11–20. IEEE, 2012.

[185] Sara van de Geer, Johannes Lederer, et al. The lasso, correlated design, and improved oracle inequalities. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 303–316. Institute of Mathematical Statistics, 2013.

[186] Ramon van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

[187] Gaël Varoquaux, Alexandre Gramfort, Jean-Baptiste Poline, and Bertrand Thirion. Brain covariance selection: better individual functional connectivity models using population prior. In *Advances in neural information processing systems*, pages 2334–2342, 2010.

[188] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[189] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2016.

[190] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2016.

[191] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[192] Wei Wang, Martin J Wainwright, and Kannan Ramchandran. Information-theoretic bounds on model selection for gaussian markov random fields. In *2010 IEEE International Symposium on Information Theory*, pages 1373–1377. IEEE, 2010.

[193] Yuhao Wang, Uma Roy, and Caroline Uhler. Learning high-dimensional gaussian graphical models under total positivity without adjustment of tuning parameters. In *International Conference on Artificial Intelligence and Statistics*, pages 2698–2708, 2020.

[194] Yair Weiss and William T Freeman. Correctness of belief propagation in gaussian graphical models of arbitrary topology. In *Advances in neural information processing systems*, pages 673–679, 2000.

[195] Max Welling and Yee Whye Teh. Approximate inference in boltzmann machines. *Artificial Intelligence*, 143(1):19–50, 2003.

[196] Anja Wille, Philip Zimmermann, Eva Vranová, Andreas Fürholz, Oliver Laule, Stefan Bleuler, Lars Hennig, Amela Prelić, Peter von Rohr, Lothar Thiele, et al. Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome biology*, 5(11):R92, 2004.

[197] Blake Woodworth, Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Kernel and deep regimes in overparametrized models. *arXiv preprint arXiv:1906.05827*, 2019.

[198] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[199] Yuichi Yoshida and Yuan Zhou. Approximation schemes via sherali-adams hierarchy for dense constraint satisfaction problems and assignment problems. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 423–438. ACM, 2014.

[200] Yuchen Zhang, Jason D Lee, and Michael I Jordan. ℓ_1-regularized neural networks are improperly learnable in polynomial time. In *International Conference on Machine Learning*, pages 993–1001, 2016.

[201] Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948, 2014.

[202] 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.

[203] Shuheng Zhou, Sara van de Geer, and Peter Bühlmann. Adaptive lasso for high dimensional regression and gaussian graphical modeling. *arXiv preprint arXiv:0903.2515*, 2009.

[204] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.

[205] Xiaojin Zhu, John Lafferty, and Zoubin Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*, volume 3, 2003.

[206] Piotr Zwiernik. *Semialgebraic statistics and latent tree models*. CRC Press, 2015.