# The minimax learning rates of normal and Ising undirected graphical models

**Luc Devroye**[*]   and **Abbas Mehrabian**[†]   and **Tommy Reddad**[‡]

*School of Computer Science*
*McGill University*
*Montréal, Québec, Canada*
*e-mail:* lucdevroye@gmail.com*; * abbas.mehrabian@gmail.com*; * tommy.reddad@gmail.com

**Abstract:** Let $G$ be an undirected graph with $m$ edges and $d$ vertices. We show that $d$-dimensional Ising models on $G$ can be learned from $n$ i.i.d. samples within expected total variation distance some constant factor of $\min\{1, \sqrt{(m+d)/n}\}$, and that this rate is optimal. We show that the same rate holds for the class of $d$-dimensional multivariate normal undirected graphical models with respect to $G$. We also identify the optimal rate of $\min\{1, \sqrt{m/n}\}$ for Ising models with no external magnetic field.

## Contents

## 1. Introduction

The Ising model is a popular mathematical model inspired by ferromagnetism in statistical mechanics. The model consists of discrete $\{-1, 1\}$ random variables representing magnetic dipole moments of atomic spins. The spins are arranged in a graph—originally a lattice, but other graphs have also been considered—allowing each spin to interact with its graph neighbors. Sometimes, the spins are also subject to an external magnetic field.

The Ising model is one of many possible mean field models for spin glasses. Its probabilistic properties have caught the attention of many researchers—see, e.g., the monographs of Talagrand [36, 37, 38]. The analysis of social networks has brought computer scientists into the fray, as precisely the same model appears there in the context of community detection [8].

In this work we view an Ising model as a probability distribution on $\{-1, 1\}^d$, and consider the following statistical inference and learning problem, known as *density estimation* or *distribution learning*: given i.i.d. samples from an unknown Ising model $I$ on a known graph $G$, can we create a probability distribution on $\{-1, 1\}^d$ that is close to $I$ in total variation distance? If we have $n$ samples, then how small can we make the expected value of this distance? We prove that if $G$ has $m$ edges, the answer to this question is bounded from above and below by constant factors of $\sqrt{(m + d)/n}$. In the case when there is no external magnetic field, the answer is instead $\sqrt{m/n}$.

Our techniques carry over to the continuous case and allow us to prove a similar minimax risk for learning the class of $d$-dimensional normal undirected graphical models on $G$. It is surprising that the minimax rate for this class was not known, even when $G$ is the complete graph, corresponding to the class of all $d$-dimensional normal distributions.

### 1.1. Main results

We start by stating our results for normal distributions. For formal definitions of all terms mentioned below, see Section 2. We express minimax risks as functions of the number of samples $n$.

**Theorem 1.1** (Main result for learning normals)**.** *Let $G$ be a given undirected graph with vertex set $\{1, \ldots, d\}$ and $m$ edges. Let $\mathcal{F}_G$ be the class of $d$-dimensional multivariate normal undirected graphical models with respect to $G$. Then, the minimax rate for learning $\mathcal{F}_G$ in total variation distance is bounded from above and below by constant factors of $\min\{1, \sqrt{(m + d)/n}\}$.*

The upper bound follows from standard techniques (see Section 3.1) and a lower bound of $\min\{1, \sqrt{d/n}\}$ is known (see Section 1.2); our main technical contribution is to show a lower bound of $\min\{1, \sqrt{m/n}\}$, from which Theorem 1.1 follows. This theorem immediately implies a tight result on the minimax rate for learning the class of all $d$-dimensional normals, if we take the graph $G$ to be complete. In this specific case, the upper bound is already known, so our contribution is the matching lower bound.

**Corollary 1.2.** *The minimax rate for learning the class of all d-dimensional multivariate normal distributions in total variation distance is bounded from above and below by constant factors of* $\min\{1, d/\sqrt{n}\}$.

In fact, this result can be extended using the techniques of [2] to yield the lower bound of $\Omega(\min\{1, d\sqrt{k/n}\})$ for the minimax learning rate of *mixtures* of $k$ many $d$-dimensional multivariate normals, which was previously known only up to logarithmic factors.

We remark that for the class of *zero-mean* normal undirected graphical models, we prove a lower bound of $\min\{1, \sqrt{m/n}\}$ while the best known upper bound is $\min\{1, \sqrt{(m+d)/n}\}$. In practice, the underlying graph is typically connected, which means that $m \geq d - 1$, so these bounds match.

We prove similar rates as in Theorem 1.1 for the class of Ising models, which resemble discrete versions of multivariate normal distributions. An Ising model in dimension $d$ is supported on $\{-1, 1\}^d$ and comes with an undirected graph $G = (V, E)$ with vertex set $V = \{1, \ldots, d\}$, edge set $E \subseteq \{\{i, j\} : i \neq j \in V\}$, interactions $w_{ij} \in \mathbb{R}$ for each $\{i, j\} \in E$, and external magnetic field $h_i \in \mathbb{R}$ for $1 \leq i \leq d$ such that $x \in \{-1, 1\}^d$ appears with probability proportional to

$$\exp\left\{\sum_{\{i,j\}\in E} w_{ij}x_ix_j + \sum_{i=1}^{d} h_ix_i\right\}.$$

Note that our definition has no temperature parameter; we have absorbed it into the weights.

**Theorem 1.3** (Main result for learning Ising models). *Let $G$ be a given undirected graph with vertex set $\{1, \ldots, d\}$ and $m$ edges. Let $\mathcal{I}_G$ be the class of d-dimensional Ising models with underlying graph $G$.*

(i) *The minimax rate for learning $\mathcal{I}_G$ in total variation distance is bounded from above and below by constant factors of* $\min\{1, \sqrt{(m+d)/n}\}$.
(ii) *Let $\mathcal{I}'_G$ be the subclass $\mathcal{I}_G$ of Ising models with no external magnetic field. The minimax rate for learning $\mathcal{I}'_G$ in total variation distance is bounded from above and below by constant factors of* $\min\{1, \sqrt{m/n}\}$.

In all of the above cases, it is assumed that the full structure and labeling of the underlying graph $G$ is known in advance. We next consider the case in which it is only known that the underlying graph has $d$ vertices and $m$ edges.

**Theorem 1.4.** *Let $\mathcal{F}_{d,m}$ and $\mathcal{I}_{d,m}$ be the class of all normal and Ising undirected graphical models with respect to some unknown graph with d vertices and m edges. The minimax learning rates for $\mathcal{F}_{d,m}$ and $\mathcal{I}_{d,m}$ are both bounded from above by a constant factor of* $\min\{1, \sqrt{(m+d)\log(d)/n}\}$, *and bounded from below by a constant factor of* $\min\{1, \sqrt{(m+d)/n}\}$.

The lower bound in this theorem follows immediately from our lower bounds for the case in which the graph is known.

In the next section we review related work. In Section 2 we discuss preliminaries. Theorem 1.1, Theorem 1.3, and Theorem 1.4 are proved in Section 3, Section 4, and Section 5, respectively. We conclude with some open problems in Section 6.

### 1.2. Related work

Density estimation is a central problem in statistics and has a long history [12, 13, 25, 39]. It has also been studied in the learning theory community under the name *distribution learning*, starting from [27], whose focus is on the computational complexity of the learning problem. Recently, it has gained a lot of attention in the machine learning community, as one of the important tasks in unsupervised learning is to understand the distribution of the data, which is known to significantly improve the efficiency of learning algorithms (e.g., [20, page 100]). See [16] for a recent survey from this perspective.

An upper bound on the order of $d/\sqrt{n}$ for estimating $d$-dimensional normals can be obtained via empirical mean and covariance estimation (e.g., [2, Theorem B.1]) or via Yatracos' techniques based on VC-dimension (e.g., [4, Theorem 13]). Regarding lower bounds, Acharya, Jafarpour, Orlitsky, and Suresh [1, Theorem 2] proved a lower bound on the order of $\sqrt{d/n}$ for normals with identity covariance matrix, which implies the same lower bound for general normals. The lower bound for general normals was recently improved to a constant factor of $\frac{d}{\sqrt{n}\log n}$ by Asthiani, Ben-David, Harvey, Liaw, Mehrabian, and Plan [2]. In comparison, our result shaves off the logarithmic factor. Moreover, their result is nonconstructive and relies on the probabilistic method, while our argument is fully deterministic. (Recently, Ashtiani et al. [3] have also shaved off the logarithmic factor and have shown that their probabilistic construction achieves the optimal lower bound of $\frac{d}{\sqrt{n}}$ as well.)

For the Ising model, the main focus in the literature has been on learning the structure of the underlying graph rather than learning the distribution itself, i.e., how many samples are needed to reconstruct the underlying graph with high probability? Bresler [10] and Vuffray, Misra, Lokhov, and Chertkov [43] provided efficient algorithms for this task on bounded-degree graphs. For general graphs, see [33, 35] for some lower bounds and [21, 28] for some upper bounds. Klivans and Meka [28] also give an efficient algorithm for learning all of the parameters of an Ising model up to an additive error given a natural parametric constraint. Otherwise, the Ising model itself has been studied by physicists in other settings for nearly a century. See the books of Talagrand for a comprehensive look at the mathematics of the Ising model [36, 37, 38].

Daskalakis, Dikkala, and Kamath [11] were the first to study Ising models from a statistical point of view. However, their goal is to test whether an Ising model has certain properties, rather than estimating the model, which is our goal. Moreover, their focus is on designing efficient testing algorithms. They prove polynomial sample complexities and running times for testing various properties of the model.

An alternative goal would be to estimate the parameters of the underlying model (see, e.g., [26]) rather than coming up with a model that is statistically close, which is our focus. We remark that these two goals are quantitatively different, although similar techniques may be used for both. In general, estimating the parameters of a model to within some accuracy does not necessarily result in a distribution that is close to the original distribution in a statistical sense. For instance, define

$$\Sigma = \begin{pmatrix} 1 & -0.99 \\ -0.99 & 1 \end{pmatrix} \qquad \text{and} \qquad \widetilde{\Sigma} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},$$

and observe that $\Sigma$ and $\widetilde{\Sigma}$ are entrywise very close. However, $\Sigma$ is non-singular and $\widetilde{\Sigma}$ is singular, and thus two zero-mean normal distributions with covariance matrices $\Sigma$ and $\widetilde{\Sigma}$ are at total variation distance 1 from one another. Conversely, if two distributions are close in total variation distance, their parameters are not necessarily close to within the same accuracy (see, e.g., [2, Section 1.3.1]).

## 2. Preliminaries

The goal of density estimation is to design an estimator $\hat{f}$ for an unknown function $f$ which belongs to a known class of functions $\mathcal{F}$. In the continuous case, $\mathcal{F}$ is a class of probability density functions with sample space $\mathcal{X} = \mathbb{R}^d$ for some $d \geq 1$; in the discrete case, $\mathcal{F}$ is a class of probability mass functions with a countable sample space $\mathcal{X}$. In either case, in order to create the estimator $\hat{f}$, we have access to samples $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f$. Our measure of closeness is the *total variation (TV) distance*: For functions $f, g : \mathcal{X} \to \mathbb{R}$, their TV-distance is defined as

$$\mathrm{TV}(f, g) = \|f - g\|_1 / 2,$$

where for any function $f$, the $L^1$-norm of $f$ is defined as

$$\|f\|_1 = \int_{\mathcal{X}} |f(x)| \, \mathrm{d}x \qquad \text{in the continuous case, and}$$

$$\|f\|_1 = \sum_{x \in \mathcal{X}} |f(x)| \qquad \text{in the discrete case.}$$

Further along, we will also need the *Kullback-Leibler (KL) divergence* or *relative entropy* [29], which is another measure of closeness of distributions defined by

$$\mathrm{KL}(f \parallel g) = \int_{\mathcal{X}} f(x) \log\left(\frac{f(x)}{g(x)}\right) \mathrm{d}x \qquad \text{in the continuous case, and}$$

$$\mathrm{KL}(f \parallel g) = \sum_{x \in \mathcal{X}} f(x) \log\left(\frac{f(x)}{g(x)}\right) \qquad \text{in the discrete case.}$$

Formally, in the continuous case, we can write $f = \frac{\mathrm{d}F}{\mathrm{d}\mu}$ for a probability measure $F$ and $\mu$ the Lebesgue measure on $\mathbb{R}^d$, and in the discrete case $f = \frac{\mathrm{d}F}{\mathrm{d}\mu}$ for a

probability measure $F$ and $\mu$ the counting measure on countable $\mathcal{X}$. In view of this unified framework, we say that $\mathcal{F}$ is a *class of densities* and that $\hat{f}$ is a *density estimate*, in both the continuous and the discrete settings.

The total variation distance has a natural probabilistic interpretation as $\mathrm{TV}(f,g) = \sup_{A \subseteq \mathcal{X}} |F(A) - G(A)|$, where $F$ and $G$ are probability measures corresponding to $f$ and $g$, respectively. So, the TV-distance lies in $[0,1]$. Also, it is well-known that the KL-divergence is nonnegative, and is zero if and only if the two densities are equal almost everywhere. However, it is not symmetric in general, and can become $+\infty$.

For density estimation there are various possible measures of distance between distributions. Here we focus on the TV-distance since it has several appealing properties, such as being a metric and having a natural probabilistic interpretation. For a detailed discussion on why TV is a natural choice, see [14, Chapter 5]. If $\hat{f}$ is a density estimate, we define the *risk* of the estimator $\hat{f}$ with respect to the class $\mathcal{F}$ as

$$\mathcal{R}_n(\hat{f}, \mathcal{F}) = \sup_{f \in \mathcal{F}} \mathbf{E}\{\mathrm{TV}(\hat{f}, f)\},$$

where the expectation is over the $n$ i.i.d. samples from $f$, and possible randomization of the estimator. The *minimax risk* or *minimax rate* for learning $\mathcal{F}$ is the smallest risk over all possible estimators,

$$\mathcal{R}_n(\mathcal{F}) = \inf_{\hat{f}:\, \mathcal{X}^n \to \mathbb{R}^{\mathcal{X}}} \mathcal{R}_n(\hat{f}, \mathcal{F}).$$

For a class of functions $\mathcal{F}$ defined on the same domain $\mathcal{X}$, its *Yatracos class* $\mathcal{A}$, first defined in [44], is the class of sets defined by

$$\mathcal{A} = \Big\{ \{x \in \mathcal{X} : f(x) > g(x)\} : f \neq g \in \mathcal{F} \Big\}.$$

The following powerful result relates the minimax risk of a class of densities to a well-studied combinatorial quantity called the Vapnik-Chervonenkis (VC) dimension [40], defined next. Let $\mathcal{A} \subseteq 2^{\mathcal{X}}$ be a family of subsets of $\mathcal{X}$. The *VC-dimension* of $\mathcal{A}$, denoted by $\mathrm{VC}(\mathcal{A})$, is the size of the largest set $X \subseteq \mathcal{X}$ such that for each $Y \subseteq X$ there exists $B \in \mathcal{A}$ such that $X \cap B = Y$. See, e.g., [14, Chapter 4] for examples and applications.

**Theorem 2.1** ([14, Section 8.1])**.** *There is a universal constant $c > 0$ such that for any class of densities $\mathcal{F}$ with Yatracos class $\mathcal{A}$,*

$$\mathcal{R}_n(\mathcal{F}) \leq c\sqrt{\mathrm{VC}(\mathcal{A})/n}.$$

On the other hand, there are several methods for obtaining lower bounds on minimax risk; we emphasize, in particular, the methods of Assouad [5], Le Cam [31, 32], and Fano [22]. Each of these involve picking a finite subclass $\mathcal{G} \subseteq \mathcal{F}$, and developing a lower bound on the minimax risk of $\mathcal{G}$ using the fact that $\mathcal{R}_n(\mathcal{G}) \leq \mathcal{R}_n(\mathcal{F})$—see [12, 14, 45] for examples. We will use the following result, known as (generalized) Fano's lemma, originally due to Has'minskiĭ [22], where $\log(\cdot)$ denotes the natural logarithm function.

**Lemma 2.2** (Fano's Lemma [45, Lemma 3]). *Let $\mathcal{F}$ be a finite class of densities such that*

$$\inf_{f \neq g \in \mathcal{F}} \|f - g\|_1 \geq \alpha, \qquad \sup_{f \neq g \in \mathcal{F}} \mathrm{KL}(f \parallel g) \leq \beta.$$

*Then,*

$$\mathcal{R}_n(\mathcal{F}) \geq \frac{\alpha}{4}\left(1 - \frac{n\beta + \log 2}{\log |\mathcal{F}|}\right).$$

In light of this lemma, to prove a minimax risk lower bound on a class of densities $\mathcal{F}$, we shall carefully pick a finite subclass of densities in $\mathcal{F}$, such that any two densities in this subclass are far apart in $L^1$-distance but close in KL-divergence.

Throughout this paper, we will be estimating densities from classes with a given graphical dependence structure, known as undirected graphical models [30]. The underlying graph will always be undirected and without parallel edges or self-loops, so we will omit these qualifiers henceforth. Let $G = (V, E)$ be a given graph with vertex set $V = \{1, \ldots, d\}$ and edge set $E$. A set of random variables $\{X_1, \ldots, X_d\}$ with everywhere strictly positive densities forms a *graphical model*, also called a *Markov random field*, with respect to $G$ if, for every $\{i, j\} \notin E$, the variables $X_i$ and $X_j$ are conditionally independent given $\{X_k : k \neq i, j\}$.

Often, the problem of density estimation is framed slightly differently than we have presented it: given $\varepsilon \in (0, 1)$, one might be interested in finding the smallest number of i.i.d. samples $m_{\mathcal{F}}(\varepsilon)$ for which there exists a density estimate $\hat{f}$ based on these samples satisfying $\sup_{f \in \mathcal{F}} \mathbf{E}\{\mathrm{TV}(\hat{f}, f)\} \leq \varepsilon$. Or, given $\delta \in (0, 1)$, one might want to find the minimum number of samples $m_{\mathcal{F}}(\varepsilon, \delta)$ for which there is a density estimate $\hat{f}$ satisfying $\sup_{f \in \mathcal{F}} \mathrm{TV}(\hat{f}, f) \leq \varepsilon$ with probability at least $1 - \delta$. The quantities $m_{\mathcal{F}}(\varepsilon)$ and $m_{\mathcal{F}}(\varepsilon, \delta)$ are known as *sample complexities* of the class $\mathcal{F}$. Note that $m_{\mathcal{F}}(\varepsilon)$ and $\mathcal{R}_n(\mathcal{F})$ are related through the equation

$$m_{\mathcal{F}}(\varepsilon) = \min\{n \geq 0 : \mathcal{R}_n(\mathcal{F}) \leq \varepsilon\},$$

so determining $\mathcal{R}_n(\mathcal{F})$ also determines $m_{\mathcal{F}}(\cdot)$. Moreover, $\delta$ is often fixed to be some small constant, say $1/3$, when studying $m_{\mathcal{F}}(\varepsilon, \delta)$. Then, there are versions of Theorem 2.1 and Lemma 2.2 for $m_{\mathcal{F}}(\varepsilon, 1/3)$, which introduce some extraneous $\log(1/\varepsilon)$ factors. In order to avoid such extraneous logarithmic factors, we focus on $\mathcal{R}_n(\mathcal{F})$—equivalently, $m_{\mathcal{F}}(\varepsilon)$—rather than $m_{\mathcal{F}}(\varepsilon, 1/3)$ or $m_{\mathcal{F}}(\varepsilon, \delta)$.

We now recall some basic matrix analysis formulae that will be used throughout (see Horn and Johnson [24] for the proofs). We denote the identity matrix by $I$. For a matrix $A = (A_{ij}) \in \mathbb{R}^{d \times d}$, the *spectral norm* of $A$ is defined as $\|A\| = \sup_{x \in \mathbb{S}^{d-1}} \|Ax\|_2$, where $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ is the unit $(d-1)$-sphere. The trace and determinant of $A$ are denoted by $\mathrm{tr}(A)$ and $\det(A)$, respectively. Recall also the *Frobenius norm* of $A$, also called the *Hilbert-Schmidt norm* or the *Schur norm*, $\|A\|_F = \sqrt{\sum_{i,j=1}^d A_{ij}^2} = \sqrt{\mathrm{tr}(A^{\mathsf{T}}A)}$. When $A$ has only real eigenvalues, we write $\lambda_i(A)$ for the $i$th largest eigenvalue of $A$. In

general, we write $\sigma_i(A) = \sqrt{\lambda_i(A^\mathsf{T} A)}$ for the $i$th largest singular value of $A$. Then, $\det(A) = \prod_{i=1}^d \lambda_i(A)$, and for any $k \geq 1$, $\mathrm{tr}(A^k) = \sum_{i=1}^d \lambda_i^k(A)$. Furthermore, $\|A\| = \sigma_1(A)$, and $\|A\|_F = \sqrt{\sum_{i=1}^d \sigma_i(A)^2}$, so $\|A\| \leq \|A\|_F$. When $A$ is invertible, $\sigma_i(A^{-1}) = \sigma_{d-i}(A)^{-1}$ for every $1 \leq i \leq d$. Finally, for any matrix $B \in \mathbb{R}^{d \times d}$, we have (see, e.g., [23, Fact 7(c) in Section 24.4])

$$\max\{\sigma_d(A)\|B\|_F, \sigma_d(B)\|A\|_F\} \leq \|AB\|_F \leq \min\{\sigma_1(A)\|B\|_F, \sigma_1(B)\|A\|_F\}. \tag{1}$$

Throughout this paper, we let $c_0, c_1, c_2, \ldots \in \mathbb{R}$ denote positive absolute constants. We liberally reuse these symbols, i.e., every $c_i$ may differ between proofs and statements of different results. We denote the set $\{1, \ldots, d\}$ by $[d]$.

## 3. Learning normal graphical models

Let $d$ be a positive integer, $\mathcal{P}_d \subseteq \mathbb{R}^{d \times d}$ be the set of positive definite $d \times d$ matrices over $\mathbb{R}$, and $\mathcal{N}(\mu, \Sigma)$ denote the multivariate normal distribution with mean $\mu \in \mathbb{R}^d$, covariance matrix $\Sigma \in \mathcal{P}_d$, and corresponding probability density function $f_{\mu, \Sigma}$, where for $x \in \mathbb{R}^d$,

$$f_{\mu, \Sigma}(x) = \frac{\exp\{-\frac{1}{2}(x - \mu)^\mathsf{T} \Sigma^{-1}(x - \mu)\}}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}}.$$

Let $G = ([d], E)$ be a given graph with $m$ edges. Let $\mathcal{P}_G \subseteq \mathcal{P}_d$ be the following subset of all positive definite matrices,

$$\mathcal{P}_G = \Big\{\Sigma \in \mathcal{P}_d\colon \text{ if } \{i, j\} \notin E \text{ with } i \neq j \in [d], \text{ then } \Sigma_{ij}^{-1} = 0\Big\}.$$

The main result of this section is a characterization of the minimax risk of

$$\mathcal{F}_G = \big\{f_{\mu, \Sigma}\colon \mu \in \mathbb{R}^d,\, \Sigma \in \mathcal{P}_G\big\}.$$

It is known that $\mathcal{F}_G$ is precisely the class of $d$-dimensional multivariate normal graphical models with respect to $G$ [30, Proposition 5.2].

### 3.1. Proof of the upper bound in Theorem 1.1

We can already prove the upper bound in Theorem 1.1 without lifting a finger. The proof is similar to that of [4, Theorem 13], which gives the optimal upper bound on the minimax risk of all multivariate normals, corresponding to the special case in which $G$ is complete. Let $\mathcal{A}$ be the Yatracos class of $\mathcal{F}_G$,

$$\mathcal{A} = \Big\{\{x \in \mathbb{R}^d\colon f_{\mu, \Sigma}(x) > f_{\widetilde{\mu}, \widetilde{\Sigma}}(x)\}\colon (\mu, \Sigma) \neq (\widetilde{\mu}, \widetilde{\Sigma}) \in \mathbb{R}^d \times \mathcal{P}_G\Big\},$$

which, after taking logarithms and simplification, is easily seen to be contained in the larger class

$$\mathcal{A}' = \Big\{\{x \in \mathbb{R}^d\colon x^\mathsf{T} A x + b^\mathsf{T} x + c > 0\}\colon A^{-1} \in \mathcal{P}_G,\, b \in \mathbb{R}^d,\, c \in \mathbb{R}\Big\},$$

and thus $\mathrm{VC}(\mathcal{A}) \leq \mathrm{VC}(\mathcal{A}')$. It remains to upper-bound $\mathrm{VC}(\mathcal{A}')$.

In general, let $\mathcal{G}$ be a vector space of real-valued functions, and define $\mathcal{B} := \{\{x\colon f(x) > 0\}\colon f \in \mathcal{G}\}$. Dudley [17, Theorem 7.2] proved that $\mathrm{VC}(\mathcal{B}) \leq \dim(\mathcal{G})$. (See [14, Lemma 4.2] for a historical discussion.) In our case, the vector space $\mathcal{G}$ has a basis of monomials

$$\{1\} \cup \{x_i x_j \colon \{i,j\} \in E\} \cup \{x_i, x_i^2 \colon i \in [d]\},$$

so $\mathrm{VC}(\mathcal{A}') \leq m + 2d + 1$. By Theorem 2.1, there is a universal constant $c > 0$ such that

$$\mathcal{R}_n(\mathcal{F}_G) \leq c\sqrt{\frac{\mathrm{VC}(\mathcal{A}')}{n}} \leq c\sqrt{\frac{m + 2d + 1}{n}},$$

while the upper bound $\mathcal{R}_n(\mathcal{F}_G) \leq 1$ follows simply because the TV-distance is bounded by 1. □

### 3.2. Proof of the lower bound in Theorem 1.1

Since a lower bound on the order of $\min\{1, \sqrt{d/n}\}$ for normals is proved in [1, Theorem 2], the lower bound in Theorem 1.1 follows from subadditivity of the square root after the following proposition.

**Proposition 3.1.** *There exist $c_0, c_1 > 0$ such that for any graph $G = ([d], E)$ with $m$ edges, where $n \geq c_1 m$,*

$$\mathcal{R}_n(\mathcal{F}_G) \geq c_0\sqrt{m/n}.$$

Note that if $n < c_1 m$, then $\mathcal{R}_n(\mathcal{F}_G) \geq \mathcal{R}_{c_1 m}(\mathcal{F}_G) \geq c_0\sqrt{1/c_1}$, which implies the lower bound in Theorem 1.1 for such $n$. We prove Proposition 3.1 via Lemma 2.2. This involves choosing a finite subset of $\mathcal{F}_G$. Our normal densities will be zero-mean, but the covariance matrices will be chosen carefully. To make this choice, we use the next result which follows from an old theorem of Gilbert [19] and independently Varshamov [41] from coding theory.

**Theorem 3.2.** *There is a subset $Q \subseteq \{-1,1\}^m$ of size at least $2^{m/5}$ such that for any distinct $s, \widetilde{s} \in Q$ we have $\|s - \widetilde{s}\|_1 \geq m/3$.*

*Proof.* We give an iterative algorithm to build $Q$: choose a vertex from the hypercube, put it in $Q$, remove the hypercube points in the corresponding $L^1$-ball of radius $m/3$, and repeat. Since the intersection of this ball and the hypercube has size at most

$$\sum_{i=0}^{m/6} \binom{m}{i} \leq \left(\frac{em}{m/6}\right)^{m/6} = (6e)^{m/6} < 2^{4m/5},$$

the size of the final set $Q$ will be at least $2^{m/5}$. Note that the sum goes up to $m/6$ and not $m/3$ here, because we are working in the hypercube $\{-1,1\}^m$, hence two vertices with $m/6$ different coordinates have $L^1$ distance $m/3$. □

Let $\mathcal{S} \subseteq \{-1,1\}^m$ be as in Theorem 3.2, so that $|\mathcal{S}| \geq 2^{m/5}$ and for any distinct $s, \widetilde{s} \in \mathcal{S}$, $\|s - \widetilde{s}\|_1 \geq m/3$. Let $\delta > 0$ be a real number to be specified later. Enumerate the edges of $G$ from 1 to $m$, and for $s \in \mathcal{S}$, set $\Sigma(s)^{-1}$ to be the $d \times d$ matrix with entries

$$\left(\Sigma(s)^{-1}\right)_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j \text{ and } \{i,j\} \notin E, \\ \delta s_{\{i,j\}} & \text{if } i \neq j \text{ and } \{i,j\} \in E. \end{cases}$$

In other words, $\Sigma(s)^{-1}$ is symmetric with all ones on its diagonal, $\pm \delta$ everywhere along the nonzero entries of the adjacency matrix of $G$ according to the signs in $s$, and 0 elsewhere.

**Lemma 3.3.** *Suppose that $\delta^2 m \leq 1/8$. Then, for any $s \in \mathcal{S}$, the matrix $\Sigma(s)^{-1}$ is positive definite and its eigenvalues lie in $[1/2, 3/2]$.*

*Proof.* Since $\Sigma(s)^{-1}$ is symmetric and real, all its eigenvalues are real. Write $\Sigma(s)^{-1} = I + \Delta$, so that $\lambda_i(\Sigma(s)^{-1}) = 1 + \lambda_i(\Delta)$. Observe that

$$|\lambda_i(\Delta)| \leq \|\Delta\| \leq \|\Delta\|_F \leq \sqrt{2\delta^2 m} \leq 1/2. \tag{2}$$

Then, $1/2 \leq \lambda_i(\Sigma(s)^{-1}) \leq 3/2$ for every $1 \leq i \leq d$, and so $\Sigma(s)^{-1}$ is positive definite. $\square$

We will assume from now on that $\delta^2 m \leq 1/8$. In light of Lemma 3.3, $\Sigma(s)^{-1}$ is positive definite, so it is invertible, and we let $\Sigma(s)$ denote its inverse. Since we will always take the mean to be 0, we will write $f_\Sigma$ for $f_{0,\Sigma}$ from now on. We define the set $\mathcal{W} = \{\Sigma(s) : s \in \mathcal{S}\}$ of covariance matrices, and let

$$\mathcal{F} = \{f_\Sigma : \Sigma \in \mathcal{W}\}.$$

In order to prove Proposition 3.1 via Lemma 2.2, it suffices to exhibit upper bounds on the KL-divergence between any two densities in $\mathcal{F}$, and lower bounds on their $L^1$-distances.

**Lemma 3.4.** *For any $\Sigma, \widetilde{\Sigma} \in \mathcal{P}_d$ satisfying $\max\{\|\Sigma^{-1} - I\|_F, \|\widetilde{\Sigma}^{-1} - I\|_F\} \leq 1/2$,*

$$\mathrm{KL}(f_\Sigma \parallel f_{\widetilde{\Sigma}}) \leq 2\|\widetilde{\Sigma}^{-1} - \Sigma^{-1}\|_F^2.$$

*Proof.* We consider a symmetrized KL-divergence, often called the *Jeffreys divergence* [29],

$$\mathrm{J}(f_\Sigma \parallel f_{\widetilde{\Sigma}}) = \mathrm{KL}(f_\Sigma \parallel f_{\widetilde{\Sigma}}) + \mathrm{KL}(f_{\widetilde{\Sigma}} \parallel f_\Sigma),$$

which clearly serves as an upper bound on the quantity of interest. It is known that

$$\mathrm{J}(f_\Sigma \parallel f_{\widetilde{\Sigma}}) = \mathrm{tr}((\Sigma - \widetilde{\Sigma})(\widetilde{\Sigma}^{-1} - \Sigma^{-1}))/2,$$

e.g., by [29, Section 9.1]. Since $\Sigma - \widetilde{\Sigma}$ is symmetric, the inequality $\mathrm{tr}(A^\mathsf{T} B) \leq \|A\|_F \cdot \|B\|_F$, which is the Cauchy-Schwarz inequality for the inner product $\langle A, B \rangle = \mathrm{tr}(A^\mathsf{T} B)$, gives

$$\mathrm{J}(f_\Sigma \parallel f_{\widetilde{\Sigma}}) \leq \|\widetilde{\Sigma}^{-1} - \Sigma^{-1}\|_F \|\Sigma - \widetilde{\Sigma}\|_F/2.$$

Notice now that $\Sigma - \widetilde{\Sigma} = \Sigma(\widetilde{\Sigma}^{-1} - \Sigma^{-1})\widetilde{\Sigma}$, so by (1),

$$\|\Sigma - \widetilde{\Sigma}\|_F = \|\Sigma(\widetilde{\Sigma}^{-1} - \Sigma^{-1})\widetilde{\Sigma}\|_F \le \|\Sigma\| \cdot \|\widetilde{\Sigma}\| \cdot \|\widetilde{\Sigma}^{-1} - \Sigma^{-1}\|_F,$$

so that

$$\mathrm{J}(f_\Sigma \parallel f_{\widetilde{\Sigma}}) \le \|\Sigma\| \cdot \|\widetilde{\Sigma}\| \cdot \|\widetilde{\Sigma}^{-1} - \Sigma^{-1}\|_F^2/2.$$

Note that since $\Sigma$ is symmetric positive definite, we have $\|\Sigma\| = \lambda_1(\Sigma)$. Write $\Sigma^{-1} = I + \Delta$ as in the proof of Lemma 3.3, so the eigenvalues of $\Sigma^{-1}$ lie in $[1 - \|\Delta\|, 1 + \|\Delta\|]$. Therefore,

$$\|\Sigma\| = \lambda_1(\Sigma) = \frac{1}{\lambda_d(\Sigma^{-1})} \le \frac{1}{1 - \|\Delta\|} \le \frac{1}{1 - \|\Delta\|_F} \le \frac{1}{1 - 1/2} = 2,$$

and the same bound holds for $\|\widetilde{\Sigma}\|$, whence $\mathrm{J}(f_\Sigma \parallel f_{\widetilde{\Sigma}}) \le 2\|\widetilde{\Sigma}^{-1} - \Sigma^{-1}\|_F^2$. $\quad\square$

To use Lemma 2.2 we need a lower bound on $\mathrm{TV}(f_\Sigma, f_{\widetilde{\Sigma}})$. We will use the following bound, which is tight up to the constant factor.

**Lemma 3.5** ([15, Theorem 1.1], see also [7, Corollary 2]). *For any pair $\Sigma, \widetilde{\Sigma} \in \mathcal{P}_d$, we have*

$$\mathrm{TV}(f_\Sigma, f_{\widetilde{\Sigma}}) \ge \min\{1, \|\Sigma^{1/2}\widetilde{\Sigma}^{-1}\Sigma^{1/2} - I\|_F\}/100.$$

*Proof of Proposition 3.1.* We will use Lemma 2.2. Set $\delta = c_2/\sqrt{n}$ for a sufficiently small constant $c_2$. Since $n \ge c_1 m$ by assumption, by choosing $c_2 \le c_1/2\sqrt{2}$ we can be sure that $\delta\sqrt{2m} \le 1/2$.

Note that for any $\Sigma \in \mathcal{W}$ we have $\|\Sigma^{-1} - I\|_F = \delta\sqrt{2m} \le 1/2$. On the one hand, Lemma 3.4 gives that for any $\Sigma, \widetilde{\Sigma} \in \mathcal{W}$ we have

$$\mathrm{KL}(f_\Sigma \parallel f_{\widetilde{\Sigma}}) \le 2\|\widetilde{\Sigma}^{-1} - \Sigma^{-1}\|_F^2 \le 8m\delta^2,$$

where we have used the triangle inequality for the second inequality. On the other hand, applying the left inequality in (1) twice,

$$\|\Sigma^{1/2}\widetilde{\Sigma}^{-1}\Sigma^{1/2} - I\|_F = \|\Sigma^{1/2}(\widetilde{\Sigma}^{-1} - \Sigma^{-1})\Sigma^{1/2}\|_F \ge \sigma_d(\Sigma^{1/2})^2\|\widetilde{\Sigma}^{-1} - \Sigma^{-1}\|_F.$$

Recall that $\sigma_d(\Sigma^{1/2})$ denotes the smallest singular value of $\Sigma^{1/2}$. By Lemma 3.3, the eigenvalues of $\Sigma^{1/2}$ lie in $[\sqrt{2/3}, \sqrt{2}]$, and since $\Sigma^{1/2}$ is symmetric and positive definite, its singular values coincide with its eigenvalues, hence $\sigma_d(\Sigma^{1/2})^2 \ge 2/3$. Also note that $\|\widetilde{\Sigma}^{-1} - \Sigma^{-1}\|_F \ge \delta\sqrt{m/3}$ since $\widetilde{\Sigma}^{-1}$ and $\Sigma^{-1}$ differ in at least $m/3$ entries. Therefore, Lemma 3.5 gives

$$\mathrm{TV}(f_\Sigma, f_{\widetilde{\Sigma}}) \ge c_3 \min\{1, \delta\sqrt{m}\} = c_3\delta\sqrt{m}.$$

Thus, in the notation of Lemma 2.2, we have by Theorem 3.2 and the above discussion that for some $c_3 > 0$,

$$|\mathcal{F}| \ge 2^{m/5}, \quad \alpha \ge c_3\delta\sqrt{m}, \quad \beta \le 8\delta^2 m.$$

Recall that $\delta = c_2/\sqrt{n}$ for a sufficiently small constant $c_2$. By choosing $c_2$ small enough, we can be sure that

$$1 - \frac{n\beta + \log 2}{\log |\mathcal{F}|} \geq \frac{1}{2}.$$

Then, by Lemma 2.2, $\mathcal{R}_n(\mathcal{F}_G) \geq \alpha/8 \geq (c_2 c_3/8)\sqrt{m/n}$, completing the proof.
□

## 4. Learning Ising graphical models

An *Ising model* describes a probability distribution on the binary hypercube $\{-1,1\}^d$ for some $d \geq 1$, where any particular vector $x \in \{-1,1\}^d$ is called a *configuration*. One such distribution is parameterized by a graph $G = ([d], E)$ with a set of edge weights $\{w_{ij}\}_{\{i,j\}\in E}$ called the *interactions*, and a set of weights $\{h_i\}_{i\in[d]}$ called the *external magnetic field*. These parameters define the *Hamiltonian* $H\colon \{-1,1\}^d \to \mathbb{R}$,

$$H(x) = \sum_{\{i,j\}\in E} w_{ij} x_i x_j + \sum_{i=1}^{d} h_i x_i.$$

Any configuration $x \in \{-1,1\}^d$ then appears with probability proportional to $\exp\{H(x)\}$. In fact, we can write $H(x) = H_{h,W}(x) = x^{\mathsf{T}} W x + h^{\mathsf{T}} x$ for a vector $h \in \mathbb{R}^d$ and a matrix $W \in \mathcal{M}_G$, where

$$\mathcal{M}_G = \Big\{ W \in \mathbb{R}^{d\times d}\colon \text{ if } \{i,j\} \notin E \text{ with } i \neq j \in [d], \text{ then } W_{ij} = 0 \Big\},$$

and in particular,

$$W_{ij} = \begin{cases} 0 & \text{if } \{i,j\} \notin E, \\ w_{ij}/2 & \text{if } \{i,j\} \in E. \end{cases}$$

The probability mass function of the Ising model with interactions $W$ and external magnetic field $h$ is denoted by $f_{h,W}$, where

$$f_{h,W}(x) = \frac{e^{H_{h,W}(x)}}{Z(h,W)}, \tag{3}$$

where the normalizing factor $Z(h,W)$, called the *partition function*, is defined by

$$Z(h,W) = \sum_{x\in\{-1,1\}^d} e^{H_{h,W}(x)}.$$

Probability distributions whose densities have the form (3) for general Hamiltonians are known as *Gibbs distributions* or *Boltzmann distributions*.

Given a graph $G$, let $\mathcal{I}_G$ be the class of all Ising models with interactions in $\mathcal{M}_G$, namely,

$$\mathcal{I}_G = \big\{ f_{h,W} \colon h \in \mathbb{R}^d, \, W \in \mathcal{M}_G \big\},$$

and let $\mathcal{I}'_G$ be the subclass with no external magnetic field, namely,

$$\mathcal{I}'_G = \{f_{0,W} \colon W \in \mathcal{M}_G\}.$$

As in Section 3, $\mathcal{I}_G$ is the class of all $d$-dimensional Ising models whose components form a graphical model with respect to $G$, and similarly for $\mathcal{I}'_G$.

We omit detailed proofs of the upper bounds in Theorem 1.3, since they are virtually identical to that of Theorem 1.1 as given in Section 3.1. For $\mathcal{I}_G$, the corresponding vector space has the basis

$$\{1\} \cup \{x_i x_j \colon \{i,j\} \in E\} \cup \{x_i \colon i \in [d]\},$$

with $m + d + 1$ elements, while for $\mathcal{I}'_G$, the corresponding vector space does not have the last $d$ basis vectors, so it has dimension $m + 1$. In the case that $m = 0$, the class $\mathcal{I}'_G$ contains only one distribution (the uniform distribution on $\{-1,1\}^d$), and thus $\mathcal{R}_n(\mathcal{I}'_G) = 0$. Thus for any $m \geq 0$ and any $G$ with $m$ edges, $\mathcal{R}_n(\mathcal{I}'_G) \leq c\sqrt{m/n}$ for some constant $c > 0$.

### 4.1. Proof of the lower bound in Theorem 1.3 (ii)

Since our Ising models in this section will have no external magnetic field, we write $f_W$ for $f_{0,W}$, $H_W$ for $H_{0,W}$, and $Z(W)$ for $Z(0,W)$. As in Section 3.2, the lower bound in Theorem 1.3 (ii) follows from the following proposition.

**Proposition 4.1.** *There exist $c_1, c_2 > 0$ such that for any graph $G = ([d], E)$ with $m$ edges, where $n \geq c_1 m$,*

$$\mathcal{R}_n(\mathcal{I}'_G) \geq c_2\sqrt{m/n}.$$

We appeal to Lemma 2.2 again. The construction and proof techniques are very similar to the previous section. Indeed, let $\mathcal{S} \subseteq \{-1,1\}^m$ be a set of sign vectors as in Theorem 3.2, satisfying $|\mathcal{S}| \geq 2^{m/5}$ and for any distinct $s, \widetilde{s} \in \mathcal{S}$, $\|s - \widetilde{s}\|_1 \geq m/3$. Enumerate the edges of $G$ from 1 to $m$, and for $s \in \mathcal{S}$, define the zero-diagonal symmetric matrix $W(s) \in \mathcal{M}_G$ with entries

$$W(s)_{ij} = \begin{cases} 0 & \text{if } i = j \text{ or } \{i,j\} \notin E, \\ \delta s_{\{i,j\}} & \text{if } \{i,j\} \in E. \end{cases}$$

Then let $\mathcal{W} = \{W(s) \colon s \in \mathcal{S}\}$ be a set of interactions, and $\mathcal{I} = \{f_W \colon W \in \mathcal{W}\}$ be the finite class of Ising models with interactions from $\mathcal{W}$.

Now, to control the $L^1$-distances and the KL-divergences between distributions in $\mathcal{I}$, a few intermediate computations are necessary. First, we recall some properties of sub-gaussian random variables. The *sub-gaussian norm* of a random variable $X$ is defined to be

$$\|X\|_{\psi_2} = \inf\Big\{t > 0 \colon \mathbf{E}\{e^{(X/t)^2}\} \leq 2\Big\}.$$

A random variable $X$ is called *sub-gaussian* if $\|X\|_{\psi_2} < \infty$ (see, e.g., [42, Section 2.5]). Observe in particular that any bounded random variable is sub-gaussian. Recall now the following concentration inequality for quadratic forms of sub-gaussian random vectors.

**Theorem 4.2** (Hanson-Wright inequality [42, Theorem 6.2.1], see also [9, Example 2.12])**.** *Let $X = (X_1, \ldots, X_d)$ be a random vector with independent zero-mean components satisfying $\max_{1 \le i \le d} \|X_i\|_{\psi_2} \le K$, and let $W \in \mathbb{R}^{d \times d}$. Then, for every $t \ge 0$,*

$$\mathbf{P}\left\{|X^\mathsf{T} W X - \mathbf{E}\, X^\mathsf{T} W X| > t\right\} \le 2 \exp\left\{-C \min\left\{\frac{t^2}{K^4 \|W\|_F^2}, \frac{t}{K^2 \|W\|}\right\}\right\},$$

*for some universal constant $C > 0$.*

A square matrix is called *zero-diagonal* if all its diagonal entries are zero. The following moment inequalities will come in handy.

**Lemma 4.3.** *Let $X = (X_1, \ldots, X_d)$ be a random vector with i.i.d. components where $\mathbf{E}\{X_1\} = 0$, $\mathbf{E}\{X_1^2\} = 1$, and $\|X_1\|_{\psi_2} \le K$. Let $W \in \mathbb{R}^{d \times d}$ be symmetric and zero-diagonal. Then,*

*(i)* $\mathbf{E}\{X^\mathsf{T} W X\} = 0$.
*(ii)* $\mathbf{E}\{(X^\mathsf{T} W X)^2\} = 2\|W\|_F^2$.
*(iii)* *There exists $c_3 > 0$ such that for any integer $k$ we have*

$$\mathbf{E}\{(X^\mathsf{T} W X)^k\} \le c_3^k K^{2k} k! \|W\|_F^k.$$

*(iv)* *There exist $c_1, c_2 > 0$ such that, for any $t > 0$, if $c_1 K^2 t \|W\|_F \le 1$ then*

$$\mathbf{E}\{e^{t X^\mathsf{T} W X}\} \le 1 + c_2 K^4 t^2 \|W\|_F^2.$$

*Proof.* Observation (i) follows simply by writing out the quadratic form,

$$\mathbf{E}\, X^\mathsf{T} W X = \sum_{i,j} W_{ij} \,\mathbf{E}\{X_i X_j\} = \sum_{i=1}^{d} W_{ii} \,\mathbf{E}\{X_i^2\} + \sum_{i \ne j} W_{ij} \,\mathbf{E}\{X_i\}\,\mathbf{E}\{X_j\} = 0.$$

To prove (ii), we expand the square, and notice that only the monomials of the form $\mathbf{E}\{X_i^4\}$ or $\mathbf{E}\{X_i^2 X_j^2\}$ are nonzero after taking expectations, so

$$\begin{aligned}
\mathbf{E}\{(X^\mathsf{T} W X)^2\} &= \mathbf{E}\left\{\left(\sum_{i,j} W_{ij} X_i X_j\right)^2\right\} \\
&= \sum_{i=1}^{d} W_{ii}^2 \,\mathbf{E}\{X_i^4\} + \sum_{i \ne j} (W_{ij}^2 + W_{ij} W_{ji} + W_{ii} W_{jj}) \,\mathbf{E}\{X_i^2 X_j^2\} \\
&= 2 \sum_{i \ne j} W_{ij}^2 = 2\|W\|_F^2.
\end{aligned}$$

For (iii), we integrate

$$
\mathbf{E}\{(X^\mathsf{T} W X)^k\} \leq \int_0^\infty \mathbf{P}\{|X^\mathsf{T} W X|^k \geq t\}\,\mathrm{d}t
$$

$$
\leq 2\int_0^\infty e^{-C\frac{t^{1/k}}{K^2\|W\|}}\,\mathrm{d}t + 2\int_0^\infty e^{-C\frac{t^{2/k}}{K^4\|W\|_F^2}}\,\mathrm{d}t
$$

$$
\text{(by Theorem 4.2)}
$$

$$
= 2\Gamma(k+1)\left(\frac{K^2\|W\|}{C}\right)^k + 2\Gamma(k/2+1)\left(\frac{K^4\|W\|_F^2}{C}\right)^{k/2}
$$

$$
\leq c_3^k K^{2k} k! \|W\|_F^k,
$$

for some $c_3 > 0$. Here, $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x}\,\mathrm{d}x$ is the gamma function, which is increasing on $[1, \infty)$ and satisfies $\Gamma(t+1) = t\Gamma(t)$ for any real $t$, and $\Gamma(k+1) = k!$ for any positive integer $k$.

To prove (iv), we use the power series representation of the exponential, so

$$
\mathbf{E}\{e^{tX^\mathsf{T} W X}\} - 1 = \sum_{k=1}^\infty \frac{\mathbf{E}\{(tX^\mathsf{T} W X)^k\}}{k!}
$$

$$
= \sum_{k=2}^\infty \frac{\mathbf{E}\{(tX^\mathsf{T} W X)^k\}}{k!} \qquad (\mathbf{E}\{tX^\mathsf{T} W X\} = 0)
$$

$$
\leq \sum_{k=2}^\infty \frac{(c_3 K^2 t\|W\|_F)^k k!}{k!} \qquad \text{(part (iii))}
$$

$$
\leq 2c_3^2 K^4 t^2 \|W\|_F^2 \qquad (\text{if } c_3 K^2 t\|W\|_F \leq 1/2),
$$

completing the proof. □

For the rest of this section, let $X = (X_1, \ldots, X_d)$ denote a uniformly random vector in $\{-1, 1\}^d$. All expectations will be with respect to this random variable.

**Lemma 4.4.** *There exist $c_1, c_2 > 0$ such that for any zero-diagonal symmetric $W \in \mathbb{R}^{d \times d}$ with $\|W\|_F \leq c_1$,*

$$
1 \leq 2^{-d} Z(W) \leq 1 + c_2 \|W\|_F^2.
$$

*Proof.* By the definition of $Z(W)$,

$$
2^{-d} Z(W) = \sum_{x \in \{-1,1\}^d} 2^{-d} e^{x^\mathsf{T} W x} = \mathbf{E}\{e^{X^\mathsf{T} W X}\}.
$$

On the one hand, by Lemma 4.3 (i),

$$
\mathbf{E}\{e^{X^\mathsf{T} W X}\} \geq \mathbf{E}\{1 + X^\mathsf{T} W X\} = 1,
$$

and on the other hand, by Lemma 4.3 (iv),

$$
\mathbf{E}\{e^{X^\mathsf{T} W X}\} \leq 1 + c_2 \|W\|_F^2,
$$

when $\|W\|_F \leq c_1$ for some sufficiently small positive constant $c_1$. □

**Lemma 4.5.** *There exist $c_1, c_2 > 0$ such that for any zero-diagonal symmetric matrices $W, \widetilde{W} \in \mathbb{R}^{d \times d}$ satisfying $\max\{\|W\|_F, \|\widetilde{W}\|_F\} \leq c_1$,*

$$\mathrm{KL}(f_W \parallel f_{\widetilde{W}}) \leq c_2(\|W\|_F^2 + \|\widetilde{W}\|_F^2).$$

*Proof.* We prove the inequality for $\mathrm{J}(f_W \parallel f_{\widetilde{W}}) = \mathrm{KL}(f_W \parallel f_{\widetilde{W}}) + \mathrm{KL}(f_{\widetilde{W}} \parallel f_W)$. By definition,

$$\mathrm{KL}(f_W \parallel f_{\widetilde{W}}) = 2^d \, \mathbf{E}\left\{ f_W(X) \log\left( \frac{f_W(X)}{f_{\widetilde{W}}(X)} \right) \right\}$$

$$= 2^d \, \mathbf{E}\left\{ \frac{e^{H_W(X)}}{Z(W)} \log\left( \frac{e^{H_W(X)} Z(\widetilde{W})}{e^{H_{\widetilde{W}}(X)} Z(W)} \right) \right\}.$$

From Lemma 4.4 we have $0 < 2^d / Z(W) \leq 1$, whence,

$$\mathrm{KL}(f_W \parallel f_{\widetilde{W}}) \leq \mathbf{E}\left\{ e^{H_W(X)}(H_W(X) - H_{\widetilde{W}}(X)) \right\} + \mathbf{E}\left\{ e^{H_W(X)} \log\left( \frac{Z(\widetilde{W})}{Z(W)} \right) \right\}.$$

$$(4)$$

We next bound the second term. By Lemma 4.3 (iv) and since $\|W\|_F \leq c_1$, $\mathbf{E}\{e^{H_W(X)}\} \leq 1 + c_3\|W\|_F^2 \leq 1 + c_3 c_1^2$. From Lemma 4.4 we have $\frac{Z(\widetilde{W})}{Z(W)} \leq 1 + c_4\|\widetilde{W}\|_F^2$, so $\log\left( \frac{Z(\widetilde{W})}{Z(W)} \right) \leq c_4\|\widetilde{W}\|_F^2$, thus

$$\mathbf{E}\left\{ e^{H_W(X)} \log\left( \frac{Z(\widetilde{W})}{Z(W)} \right) \right\} \leq (1 + c_3 c_1^2) c_4\|\widetilde{W}\|_F^2.$$

It can be shown similarly that

$$\mathrm{KL}(f_{\widetilde{W}} \parallel f_W) \leq \mathbf{E}\left\{ e^{H_{\widetilde{W}}(X)}(H_{\widetilde{W}}(X) - H_W(X)) \right\} + c_5\|W\|_F^2,$$

hence,

$$\mathrm{J}(f_W \parallel f_{\widetilde{W}}) \leq \mathbf{E}\left\{ \left( e^{H_W(X)} - e^{H_{\widetilde{W}}(X)} \right)(H_W(X) - H_{\widetilde{W}}(X)) \right\}$$
$$+ c_5(\|W\|_F^2 + \|\widetilde{W}\|_F^2).$$

For bounding the first term, using the elementary inequality

$$1 + t \leq e^t \leq 1 + t + t^2 \max\{e^t, 1\} \qquad \text{for any } t \in \mathbb{R}, \qquad (5)$$

we find that, for all $t, s \in \mathbb{R}$,

$$(e^t - e^s)(t - s) \leq (t - s)^2 + |t - s|\left( t^2 \max\{e^t, 1\} + s^2 \max\{e^s, 1\} \right).$$

Using this, we get the following upper bound,

$$\mathbf{E}\left\{\left(e^{H_W(X)} - e^{H_{\widetilde{W}}(X)}\right)(H_W(X) - H_{\widetilde{W}}(X))\right\}$$

$$\leq \quad \mathbf{E}\{(H_W(X) - H_{\widetilde{W}}(X))^2\} \tag{6}$$

$$+ \mathbf{E}\left\{|H_W(X) - H_{\widetilde{W}}(X)| H_W(X)^2 \max\{e^{H_W(X)}, 1\}\right\} \tag{7}$$

$$+ \mathbf{E}\left\{|H_W(X) - H_{\widetilde{W}}(X)| H_{\widetilde{W}}(X)^2 \max\{e^{H_{\widetilde{W}}(X)}, 1\}\right\}. \tag{8}$$

The term (6) is $2\|W - \widetilde{W}\|_F^2 \leq 2(\|W\|_F + \|\widetilde{W}\|_F)^2$ by Lemma 4.3 (ii) and the triangle inequality for the Frobenius norm. For bounding (7), we first observe that, clearly, $(\max\{e^{H_W(X)}, 1\})^2 \leq e^{2H_W(X)} + 1$. Then, using the inequality

$$\mathbf{E}\{AB\} \leq \sqrt{\mathbf{E}\{A^2\} \mathbf{E}\{B^2\}}, \tag{9}$$

which is the Cauchy-Schwarz inequality for the inner product $\langle A, B \rangle = \mathbf{E}\{AB\}$,

$$\mathbf{E}\left\{|H_W(X) - H_{\widetilde{W}}(X)| H_W(X)^2 \max\{e^{H_W(X)}, 1\}\right\}$$

$$\leq \sqrt{\mathbf{E}\{(H_W(X) - H_{\widetilde{W}}(X))^2 H_W(X)^4\}(\mathbf{E}\{e^{2H_W(X)}\} + 1)}.$$

By Lemma 4.3 (iv) and since $\|W\|_F \leq c_1$ by assumption, $\sqrt{\mathbf{E}\{e^{2H_W(X)}\} + 1} \leq c_6$. For the other factor, we apply (9) again:

$$\mathbf{E}\{(H_W(X) - H_{\widetilde{W}}(X))^2 H_W(X)^4\} \leq \sqrt{\mathbf{E}\{(H_W(X) - H_{\widetilde{W}}(X))^4\} \mathbf{E}\{H_W(X)^8\}}$$

$$\leq c_7\sqrt{\|W - \widetilde{W}\|_F^4 \|W\|_F^8} \leq (2c_1)^2 c_7 \|W\|_F^4,$$

where the second inequality follows from Lemma 4.3 (iii) applied to matrices $W - \widetilde{W}$ and $W$, and the third one follows from the triangle inequality, $\|W - \widetilde{W}\|_F \leq \|W\|_F + \|\widetilde{W}\|_F \leq 2c_1$. Putting everything together, we find that

$$\mathbf{E}\left\{|H_W(X) - H_{\widetilde{W}}(X)| H_W(X)^2 \max\{e^{H_W(X)}, 1\}\right\} \leq c_6 \sqrt{(2c_1)^2 c_7} \|W\|_F^2,$$

and a similar bound holds for (8), after which the result follows. $\qquad \square$

**Lemma 4.6.** *There exist $c_1, c_2, c_3 > 0$ such that for any zero-diagonal symmetric matrices $W, \widetilde{W} \in \mathbb{R}^{d \times d}$ with $\max\{\|W\|_F, \|\widetilde{W}\|_F\} \leq c_1$,*

$$\|f_W - f_{\widetilde{W}}\|_1 \geq c_2 \|W - \widetilde{W}\|_F - c_3(\|W\|_F^2 + \|\widetilde{W}\|_F^2).$$

*Proof.* By Lemma 4.4 we have

$$\left|\frac{2^d}{Z(W)} - 1\right| \leq \frac{c_0 \|W\|_F^2}{1 + c_0 \|W\|_F^2} \leq c_0 \|W\|_F^2,$$

so by the triangle inequality and Lemma 4.3 (iv), there is $c_4 > 0$ for which

$$\|f_W - f_{\widetilde{W}}\|_1 = 2^d \, \mathbf{E}\left\{\left|\frac{e^{H_W(X)}}{Z(W)} - \frac{e^{H_{\widetilde{W}}(X)}}{Z(\widetilde{W})}\right|\right\}$$

$$\geq \mathbf{E}\left\{\left|e^{H_W(X)} - e^{H_{\widetilde{W}}(X)}\right|\right\} - c_4(\|W\|_F^2 + \|\widetilde{W}\|_F^2). \qquad (10)$$

By (5) and the triangle inequality again,

$$\mathbf{E}\left\{\left|e^{H_W(X)} - e^{H_{\widetilde{W}}(X)}\right|\right\} \geq \mathbf{E}\{|H_W(X) - H_{\widetilde{W}}(X)|\}$$

$$- \mathbf{E}\left\{H_W(X)^2 \max\{e^{H_W(X)}, 1\}\right\}$$

$$- \mathbf{E}\left\{H_{\widetilde{W}}(X)^2 \max\{e^{H_{\widetilde{W}}(X)}, 1\}\right\}.$$

We first bound the second term. By (9) and Lemma 4.3 (iii), (iv),

$$\mathbf{E}\left\{H_W(X)^2 \max\{e^{H_W(X)}, 1\}\right\} \leq \sqrt{\mathbf{E}\{H_W(X)^4\}(\mathbf{E}\{e^{2H_W(X)}\} + 1)} \leq c_5\|W\|_F^2,$$

and a similar analysis works for the third term. For the first term, by Hölder's inequality and Lemma 4.3 (ii), (iii) (applied to the symmetric zero-diagonal matrix $W - \widetilde{W}$), there is a $c_6 > 0$ for which

$$\mathbf{E}\{|H_W(X) - H_{\widetilde{W}}(X)|\} \geq \frac{\mathbf{E}\{(H_W(X) - H_{\widetilde{W}}(X))^2\}^{3/2}}{\mathbf{E}\{(H_W(X) - H_{\widetilde{W}}(X))^4\}^{1/2}} \geq c_6\|W - \widetilde{W}\|_F. \quad \square$$

The proof of Proposition 4.1 is now identical to that of Proposition 3.1.

### 4.2. Proof of the lower bound in Theorem 1.3 (i)

Let $\overline{\mathcal{I}_d}$ be the class of $d$-dimensional Ising models with no interactions. Note that in this case the problem is density estimation for a product distribution.

The lower bound in Theorem 1.3 (i) will follow from the next proposition along with Theorem 1.3 (ii) and subadditivity of the square root, just as in Section 3.2.

**Proposition 4.7.** *There exist $c_1, c_2 > 0$ such that if $n \geq c_1 d$,*

$$\mathcal{R}_n(\overline{\mathcal{I}_d}) \geq c_2\sqrt{d/n}.$$

*Proof sketch.* As in the above arguments, we pick a subclass of $2^{d/5}$ densities of $\overline{\mathcal{I}_d}$ and apply Lemma 2.2. The corresponding magnetic fields will have entries $\pm\delta$, with the signs specified by Theorem 3.2, so that any two of them differ in at least $d/6$ components. One can then show that the KL-divergence between any two of these densities is at most a constant factor of $\delta^2 d$, while the $L^1$-distances are at least some constant factor of $\delta\sqrt{d}$. The proofs are simpler than those in the previous section; for example, in this case, the partition functions can be computed exactly, and are equal for every density in the subclass. We omit the details. $\square$

## 5. Proof of the upper bound in Theorem 1.4

We give the proof for $\mathcal{F}_{d,m}$, and the proof for $\mathcal{I}_{d,m}$ is identical. Let $\mathcal{G}_{d,m}$ denote the set of all labeled graphs with vertex set $[d]$ and $m$ edges. Now, $\mathcal{F}_{d,m}$ has Yatracos class

$$\mathcal{A} = \bigcup_{(G,H)\in\mathcal{G}_{d,m}^2} \mathcal{A}_{G,H},$$

where

$$\mathcal{A}_{G,H} = \Big\{ \{x \in \mathbb{R}^d : g(x) > h(x)\} : g \in \mathcal{F}_G,\ h \in \mathcal{F}_H \Big\}.$$

Note that $|\mathcal{G}_{d,m}| \le \binom{\binom{d}{2}}{m} \le d^{2m}$, and $\mathrm{VC}(\mathcal{A}_{G,H}) \le 2m + 2d + 1$ for any $G, H \in \mathcal{G}_{d,m}$, as in the proof of the upper bound in Theorem 1.1. By properties of the VC-dimension of unions (see, e.g., [34, Exercise 6.11]),

$$\begin{aligned}
\mathrm{VC}(\mathcal{A}) = \mathrm{VC}\left( \bigcup_{(G,H)\in\mathcal{G}_{d,m}^2} \mathcal{A}_{G,H} \right) \\
\le c_1(m+d)\log(m+d) + c_2 \log d^{4m} \\
\le c_3(m+d)\log d,
\end{aligned}$$

so by Theorem 2.1,

$$\mathcal{R}_n(\mathcal{F}_{d,m}) \le c_4\sqrt{\frac{(m+d)\log d}{n}}.$$

## 6. Discussion

Our work raises several open problems.

1. *Higher order forms.* We have studied estimating densities that are proportional to the exponential of some quadratic form. One can ask for the minimax risk of the class of densities in which this form has a higher order. Namely, let $k, d \ge 1$ be given integers, and suppose that $\mathcal{F}$ is a class of densities supported on $\{-1, 1\}^d$, where each density $f \in \mathcal{F}$ is parameterized by weights $w_{i_1,\ldots,i_k} \in \mathbb{R}$ for each $1 \le i_1 < i_2 < \cdots < i_k \le d$, and when $x \in \{-1, 1\}^d$,

$$f(x) \propto \exp\left\{ \sum_{1 \le i_1 < \cdots < i_k \le d} w_{i_1,i_2,\ldots,i_k} x_{i_1} x_{i_2} \cdots x_{i_k} \right\}.$$

Then, just as in the proof of the upper bound of Theorem 1.3 (ii), we have that there is a universal constant $c_1 > 0$ for which

$$\mathcal{R}_n(\mathcal{F}) \le c_1 \min\left\{ 1, \sqrt{\frac{\binom{d}{k}}{n}} \right\}.$$

Can this be shown to be tight to within a constant factor? It is straightforward to see that the answer is *yes* for $k = 1$, and the results of this paper show that the answer is *yes* for $k = 2$. However, for $k \geq 3$, our techniques seem to fail. Auffinger and Ben Arous [6] noted that when the weights are $w_{i_1,\dots,i_k} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, the random $k$-th order form $g \colon \mathbb{S}^{d-1} \to \mathbb{R}$ defined by

$$g(x) = \sum_{i_1,\dots,i_k=1}^{d} w_{i_1,i_2,\dots,i_k} x_{i_1} x_{i_2} \dots x_{i_k}$$

blows up in complexity once $k \geq 3$. For example, they show that there is a $c_3 > 0$ for which $g$ has at least $e^{c_3 d}$ local minima on $\mathbb{S}^{d-1}$ in expectation, as long as $k \geq 3$. On the other hand, when $k \leq 2$, deterministically $g$ has only a constant number of local minima on $\mathbb{S}^{d-1}$. This gap in complexity may indicate that analyzing the case $k \geq 3$ for our purposes will require more sophisticated techniques.

2. *Tightness of the VC-dimension bound.* We proved that $\mathcal{R}_n(\mathcal{F})$ is bounded from above and below by constant factors of $\sqrt{\mathrm{VC}(\mathcal{A})/n}$, where $\mathcal{A}$ is the Yatracos class of $\mathcal{F}$, for $\mathcal{F} \in \{\mathcal{F}_G, \mathcal{I}_G, \mathcal{I}'_G\}$. The upper bound here holds for any class $\mathcal{F}$ by Theorem 2.1, and it can be easily seen that there are classes of densities for which this is not tight. Can we characterize the classes of densities $\mathcal{F}$ for which $\mathcal{R}_n(\mathcal{F})$ is in fact on the order of $\sqrt{\mathrm{VC}(\mathcal{A})/n}$?

3. *The minimax risk of unlabeled graphical models.* In our setting, the given graph $G$ is labeled, so we are given the specific pairs of coordinates which interact. What if only the structure of the graph $G$ is known, but its labeling is not? What if we know that $G$ is a tree? If only the number of edges of $G$ is known, Theorem 1.4 provides some bound that is tight up to a factor of $\sqrt{\log d}$. Can this gap be closed?

4. *The minimax risk of Ising blockmodels.* For a given $S \subseteq [d]$ with $|S| = d/2$ and parameters $\alpha, \beta \in \mathbb{R}$, a *bipartite mean-field model* [18], also called an *Ising blockmodel* [8], has density

$$f_{S,\alpha,\beta}(x) = \exp\left\{ \frac{\beta}{2d} \sum_{i \sim j} x_i x_j + \frac{\alpha}{2d} \sum_{i \not\sim j} x_i x_j \right\} \Big/ Z(\alpha, \beta)$$

for $x \in \{-1, 1\}^d$, where $i \sim j$ means that either $i, j \in S$ or $i, j \notin S$, and $i \not\sim j$ means that one of $i, j$ is in $S$ and one is not, and $Z(\alpha, \beta)$ is the normalizing factor. Motivated by social network analysis and the notion of communities in such networks, Berthet, Rigollet, and Srivastava [8] studied this model in a learning context. Their work is mainly concerned with the estimation or recovery of $S$ from $n$ independent samples of $f_{S,\alpha,\beta}$, but one can also ask for the minimax learning rate for this class of densities, if some or all of $\alpha, \beta$ and $S$ are unknown.

5. *Efficient algorithms for density estimation.* Here, we focused on statistical efficiency rather than computational efficiency. As the VC-based density estimate in Theorem 2.1 cannot be efficiently computed, a natural open

problem is to design efficient density estimators achieving the minimax risks for normal and Ising distributions. As far as we know, only for the class of multivariate normal distributions, which correspond to normal graphical models with respect to the complete graph, such an efficient algorithm exists. It is based on careful mean and covariance estimations (see [2, Appendix B]).

## Acknowledgements

## References

[1] ACHARYA, J., JAFARPOUR, A., ORLITSKY, A. and SURESH, A. T. (2014). Near-optimal-sample estimators for spherical Gaussian mixtures. In *Advances in Neural Information Processing Systems 27* (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger, eds.) 1395–1403. Curran Associates, Inc.

[2] ASHTIANI, H., BEN-DAVID, S., HARVEY, N., LIAW, C., MEHRABIAN, A. and PLAN, Y. (2018). Near-optimal sample complexity bounds for robust learning of Gaussians mixtures via compression schemes. *ArXiv e-prints*. Appeared in the proceedings of Neural Information Processing Systems 2018, available at https://arxiv.org/abs/1710.05209v3.

[3] ASHTIANI, H., BEN-DAVID, S., HARVEY, N., LIAW, C., MEHRABIAN, A. and PLAN, Y. (2019). Near-optimal sample complexity bounds for robust learning of Gaussians mixtures via compression schemes. *ArXiv e-prints*. To appear in the Journal of the ACM, available at https://arxiv.org/abs/1710.05209v4.

[4] ASHTIANI, H., BEN-DAVID, S. and MEHRABIAN, A. (2018). Sample-efficient learning of mixtures. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. AAAI'18* 2679–2686. AAAI Publications. Available at https://arxiv.org/abs/1706.01596.

[5] ASSOUAD, P. (1983). Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris Sér. I Math.* **296** 1021–1024. MR0777600

[6] AUFFINGER, A. and BEN AROUS, G. (2013). Complexity of random smooth functions on the high-dimensional sphere. *Ann. Probab.* **41** 4214–4247. MR3161473

[7] BARSOV, S. S. and UL'YANOV, V. V. (1987). Estimates of the proximity of Gaussian measures. *Sov. Math., Dokl.* **34** 462–466.

[8] BERTHET, Q., RIGOLLET, P. and SRIVASTAVA, P. (2019). Exact recovery in the Ising blockmodel. *Ann. Statist.* **47** 1805–1834. MR3953436

[9] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, Oxford. MR3185193

[10] BRESLER, G. (2015). Efficiently learning Ising models on arbitrary graphs. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing. STOC'15* 771–782. Association for Computing Machinery, New York, NY, USA. MR3388257

[11] DASKALAKIS, C., DIKKALA, N. and KAMATH, G. (2018). Testing Ising models. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms* 1989–2007. MR3775918

[12] DEVROYE, L. (1987). *A Course in Density Estimation. Progress in Probability and Statistics* **14**. Birkhäuser Boston, Inc., Boston, MA. MR0891874

[13] DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation: The $L_1$ View. Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics.* John Wiley & Sons, Inc., New York. MR0780746

[14] DEVROYE, L. and LUGOSI, G. (2001). *Combinatorial Methods in Density Estimation. Springer Series in Statistics.* Springer-Verlag, New York. MR1843146

[15] DEVROYE, L., MEHRABIAN, A. and REDDAD, T. (2019). The total variation distance between high-dimensional Gaussians. *ArXiv e-prints.* Available at https://arxiv.org/abs/1810.08693.

[16] DIAKONIKOLAS, I. (2016). Learning structured distributions. In *Handbook of Big Data. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 267–283. CRC Press, Boca Raton, FL. MR3674822

[17] DUDLEY, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6** 899–929. MR0512411

[18] FEDELE, M. and UNGUENDOLI, F. (2012). Rigorous results on the bipartite mean-field model. *Journal of Physics A: Mathematical and Theoretical* **45** 385001. MR2970551

[19] GILBERT, E. N. (1952). A comparison of signalling alphabets. *Bell System Tech. J.* **31** 504–522.

[20] GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep Learning. Adaptive Computation and Machine Learning.* MIT Press, Cambridge, MA. MR3617773

[21] HAMILTON, L., KOEHLER, F. and MOITRA, A. (2017). Information theoretic properties of Markov random fields, and their algorithmic applications. In *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.) 2463–2472. Curran Associates, Inc.

[22] HAS'MINSKIĬ, R. Z. (1978). A lower bound for risks of nonparametric density estimates in the uniform metric. *Teor. Veroyatnost. i Primenen.* **23** 824–828. MR0516279

[23] HOGBEN, L., ed. (2014). *Handbook of linear algebra*, second ed. *Discrete Mathematics and its Applications (Boca Raton).* CRC Press, Boca Raton, FL. MR3013937

[24] HORN, R. A. and JOHNSON, C. R. (2013). *Matrix Analysis*, Second ed. Cambridge University Press, Cambridge. MR2978290

[25] IBRAGIMOV, I. (2001). Estimation of analytic functions. In *State of the Art*

in Probability and Statistics (Leiden, 1999). *IMS Lecture Notes Monogr. Ser.* **36** 359–383. Inst. Math. Statist., Beachwood, OH. MR1836570

[26] Kalai, A., Moitra, A. and Valiant, G. (2012). Disentangling Gaussians. *Comm. ACM* **55**.

[27] Kearns, M., Mansour, Y., Ron, D., Rubinfeld, R., Schapire, R. E. and Sellie, L. (1994). On the learnability of discrete distributions. In *Proceedings of the Twenty-sixth Annual ACM Symposium on Theory of Computing. STOC'94* 273–282. ACM, New York, NY, USA.

[28] Klivans, A. R. and Meka, R. (2017). Learning graphical models using multiplicative weights. In *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017* 343–354. IEEE Computer Soc., Los Alamitos, CA. MR3734242

[29] Kullback, S. (1997). *Information Theory and Statistics.* Dover Publications, Inc., Mineola, NY. Reprint of the second (1968) edition. MR1461541

[30] Lauritzen, S. L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. The Clarendon Press, Oxford University Press, New York. MR1419991

[31] Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53. MR0334381

[32] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory. Springer Series in Statistics.* Springer-Verlag, New York. MR0856411

[33] Santhanam, N. P. and Wainwright, M. J. (2012). Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Trans. Inform. Theory* **58** 4117–4134. MR2943079

[34] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press Available at http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/copy.html. MR3277164

[35] Shanmugam, K., Tandon, R., Dimakis, A. G. and Ravikumar, P. (2014). On the information theoretic limits of learning Ising models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems. NIPS'14* **2** 2303–2311. MIT Press, Cambridge, MA, USA.

[36] Talagrand, M. (2003). *Spin Glasses: A Challenge for Mathematicians. Cavity and Mean Field Models. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics* **46**. Springer-Verlag, Berlin. MR1993891

[37] Talagrand, M. (2011). *Mean Field Models for Spin Glasses. Volume I. Basic Examples. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics* **54**. Springer-Verlag, Berlin. MR2731561

[38] Talagrand, M. (2011). *Mean Field Models for Spin Glasses. Volume II. Advanced Replica-Symmetry and Low Temperature. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics* **55**. Springer, Heidelberg. MR3024566

[39] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.*

*Springer Series in Statistics.* Springer, New York Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.  MR2724359

[40] VAPNIK, V. N. and ČERVONENKIS, A. J. (1971). The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Verojatnost. i Primenen.* **16** 264–279.  MR0288823

[41] VARŠAMOV, R. R. (1957). The evaluation of signals in codes with correction of errors. *Dokl. Akad. Nauk* **117** 739–741.  MR0095090

[42] VERSHYNIN, R. (2018). *High-dimensional probability. Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge University Press, Cambridge Available at `https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.html`.  MR3837109

[43] VUFFRAY, M., MISRA, S., LOKHOV, A. Y. and CHERTKOV, M. (2016). Interaction Screening: Efficient and Sample-Optimal Learning of Ising Models. In *Proceedings of the 30th International Conference on Neural Information Processing Systems. NIPS'16* 2603–2611. Curran Associates Inc., Red Hook, NY, USA.

[44] YATRACOS, Y. G. (1988). A note on $L_1$ consistent estimation. *Canad. J. Statist.* **16** 283–292.  MR0998220

[45] YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer, New York.  MR1462963