THE MINIMAX LEARNING RATE OF NORMAL AND ISING UNDIRECTED GRAPHICAL MODELS

LUC DEVROYE, ABBAS MEHRABIAN, AND TOMMY REDDAD

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. density estimation, distribution learning, graphical model, Markov random field, Ising model, multivariate normal, Fano's lemma. 2000 Math Subject Classification: 62G07, 82B20.

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 [30, 31, 32]. 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 [6].

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 $\sqrt{m/n}$.

Date: June 25, 2019.

2010 *Mathematics Subject Classification.* Primary: 62G07; secondary: 82B20.

Key words and phrases. Density estimation, distribution learning, graphical model, Markov random field, Ising model, multivariate normal, Fano's lemma.

Luc Devroye is supported by NSERC Grant A3456. Abbas Mehrabian is supported by a CRM-ISM postdoctoral fellowship and an IVADO-Apogee-CFREF postdoctoral fellowship. Tommy Reddad is supported by an NSERC PGS D scholarship 396164433.

Our techniques carry over to the continuous case and allow us prove a similar minimax rate 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 result for normal distributions. For precise definitions of all terms mentioned below, see Section 2.

Theorem 1.1 (Main result for learning normals)**.** *Let G be a given undirected graph with vertex set $\{1,\dots,d\}$ 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] in order to yield the optimal minimax rate of $\min\{1, d\sqrt{k/n}\}$ for learning mixtures of k independent d-dimensional multivariate normals, which was previously known only up to logarithmic factors.

We remark that for the class of mean-zero 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 $V = \{1,\dots,d\}$, edge set $E \subseteq \{\{i,j\} \colon i \neq j \in V\}$, 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

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

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 $\{1,\dots,d\}$ 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, 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 [9, 10, 19, 33]. It has also been studied in the learning theory community under the name *distribution learning*, starting from [21], 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.,* [15, page 100]). See [12] 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.,* [3, 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 spherical normals (*i.e.,* 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.

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? See [27, 29] for some lower bounds and [16, 22] for some upper bounds. Klivans and Meka [22] also give an efficient algorithm for learning the all of the parameters of an Ising model 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 [30, 31, 32].

Daskalakis, Dikkala, and Kamath [8] 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 (*e.g.,* [20]) 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 Σ and $\widetilde{\Sigma}$ are entrywise very close. However, Σ is non-singular and $\widetilde{\Sigma}$ is singular, and thus two mean-zero normal distributions with covariance matrices Σ 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.

2. Preliminaries

The goal of density estimation is to design an estimator $\hat{f}$ for an unknown function f taken from 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* [23], 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 μ 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 μ 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 [11, 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 $\mathcal{F}$ is the smallest risk over all possible estimators,

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

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

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

The following powerful result relates the minimax risk of a class of densities to an old well-studied combinatorial quantity called the Vapnik-Chervonenkis (VC) dimension [34]. Indeed, 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.,* [11, Chapter 4] for examples and applications.

Theorem 2.1 ([11, Section 8.1])**.** *There is a univeral 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 [4], Le Cam [25, 26], and Fano [17]. Each of these involve picking a finite subclass $\mathcal{G} \subseteq \mathcal{F}$, and using the fact that $\mathcal{R}_n(\mathcal{G}) \leq \mathcal{R}_n(\mathcal{F})$, developing a lower bound on the minimax risk of $\mathcal{G}$. See [9, 11, 37] for more details. We will use the following result, known as (generalized) Fano's lemma, originally due to Khas'minskii [17].

Lemma 2.2 (Fano's Lemma [37, 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 [24]. The underlying graph will always be undirected and without parallel edges or self-loops, so we will omit these qualifiers henceforth. Indeed, 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* or *Markov random field (MRF)* with respect to G if for every $\{i, j\} \notin E$, the variables X_i and X_j are conditionally independent given $\{X_k \colon k \neq i, j\}$.

Often, the problem of density estimation is framed slightly differently than we have presented it: given $\varepsilon \in (0, 1)$, we can 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 $\delta \in (0, 1)$, we 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

$$m_{\mathcal{F}}(\mathcal{R}_n(\mathcal{F})) = n,$$

so that determining one also determines the other. Moreover, δ is often fixed to be some small constant like $1/3$ when studying $m_{\mathcal{F}}(\varepsilon, \delta)$, since it can be shown that all other values of $m_{\mathcal{F}}$ for smaller δ are within a $\log(1/\delta)$ factor of $m_{\mathcal{F}}(\varepsilon, 1/3)$. 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 which will be used throughout (see Horn and Johnson [18] for the proofs). 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 $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \colon \|x\|_2 = 1\}$ is the unit $(d-1)$-sphere. Recall also the *Frobenius norm* of A, sometimes also called the *Hilbert-Schmidt norm*, $\|A\|_F = \sqrt{\sum_{i,j=1}^d A_{ij}^2} = \sqrt{\mathrm{tr}(A^\intercal A)}$. When A has all real eigenvalues, we write $\lambda_i(A)$ for the i-th largest eigenvalue of A. In general, we write $\sigma_i(A) = \sqrt{\lambda_i(A^\intercal A)}$ 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 $\|A\| \leq \|A\|_F$. For any matrix $B \in \mathbb{R}^{d \times d}$, we have $\|AB\|_F \leq \min\{\|A\| \|B\|_F, \|B\| \|A\|_F\}$. Finally, when A is invertible, $\sigma_i(A^{-1}) = \sigma_{d-i}(A)^{-1}$ for every $1 \leq i \leq d$.

Throughout this paper, we let $c_1, c_2, \ldots \in \mathbb{R}$ denote positive universal constants. We liberally reuse these symbols, *i.e.,* every c_i may differ between proofs and statements of different results. From now on, 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,

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

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 [24, 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 [3, Theorem 13], which is for an upper bound on the minimax risk of all multivariate normals, corresponding to the case in which G is complete. Let $\mathcal{A}$ be the Yatracos class of $\mathcal{F}_G$,

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

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

$$\mathcal{A}' = \Big\{\{x \in \mathbb{R}^d \colon x^\mathsf{T}Ax + b^\mathsf{T}x > c\} \colon A \in \mathbb{R}^{d\times d},\, b \in \mathbb{R}^d,\, c \in \mathbb{R}\Big\}.$$

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 $\mathcal{B} = \{\{x \colon f(x) > 0\} \colon f \in \mathcal{G}\}$. Dudley [13, Theorem 7.2] proved that $\mathrm{VC}(\mathcal{B}) \leq \dim(\mathcal{G})$. (See [11, 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 spherical normals was 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 $c_1, c_2 > 0$ such that for any graph $G = ([d], E)$ with m edges, where $n \geq c_1 m$,*

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

Note that if $n < c_1 m$, then $\mathcal{R}_n(\mathcal{F}_G) \geq \mathcal{R}_{c_1 m}(\mathcal{F}_G) \geq c_2\sqrt{1/c_1}$, which implies the lower bound in Theorem 1.1 in this regime for 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 mean-zero, 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 [14] and independently Varshamov [35] 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 $2^{m/5}$. □

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

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

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.*

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

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

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

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_Σ for $f_{0,\Sigma}$ from now on. We define the set $\mathcal{W} = \{\Sigma(s) \colon s \in \mathcal{S}\}$ of covariance matrices, and let

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

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. *There exist $c_1, c_2 > 0$ such that for any $\Sigma, \widetilde{\Sigma} \in \mathcal{P}_d$ satisfying $\max\{\|\Sigma^{-1} - I\|_F, \|\widetilde{\Sigma}^{-1} - I\|_F\} \leq c_1$,*

$$\mathrm{KL}(f_\Sigma \parallel f_{\widetilde{\Sigma}}) \leq c_2\|\widetilde{\Sigma}^{-1} - \Sigma^{-1}\|_F^2.$$

Proof. We consider a symmetrized KL-divergence, often called the *Jeffreys divergence* [23],

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

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

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

e.g., by [23, Section 9.1]. By the Cauchy-Schwarz inequality for the inner product $\langle A, B\rangle = \mathrm{tr}(A^\mathsf{T} B)$,

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

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

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

so that

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

Write $\Sigma^{-1} = I + \Delta$ just as in the proof of Lemma 3.3. Then,

$$\|\Sigma\| = \sigma_1(\Sigma) = \frac{1}{\sigma_d(\Sigma^{-1})} \le \frac{1}{1 - \|\Delta\|} \le \frac{1}{1 - \|\Delta\|_F} \le \frac{1}{1 - c_1},$$

and the same bound holds for $\|\widetilde{\Sigma}\|$, whence $\mathrm{J}(f_\Sigma \parallel f_{\widetilde{\Sigma}}) \le c_2 \|\widetilde{\Sigma}^{-1} - \Sigma^{-1}\|_F^2$. □

Unfortunately, the L^1-distance between multivariate normals does not have such a nice expression as the Jeffreys divergence does. To control some of the quantities involved in the computation of the L^1-distance, we recall some properties of sub-gaussian random variables.

The *sub-gaussian norm* of a random variable X is defined to be

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

A random variable X is called *sub-gaussian* if $\|X\|_{\psi_2} < \infty$. Observe in particular that $\mathcal{N}(0,1)$ and any bounded random variable are sub-gaussian. Recall now the following well-known large deviation inequality for quadratic forms of sub-gaussian random vectors.

Theorem 3.5 (Hanson-Wright inequality [36, Theorem 6.2.1], see also [7, Example 2.12])**.** *Let $X = (X_1, \dots, X_d)$ be a random vector with independent mean-zero components satisfying $\max_{1\le i\le d} \|X_i\|_{\psi_2} \le K$, and let $A \in \mathbb{R}^{d\times d}$. Then, for every $t \ge 0$,*

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

for some universal constant $C > 0$.

A square matrix is called *zero-diagonal* if all its diagonal entries are zero.

Lemma 3.6. *Let $X = (X_1, \dots, X_d)$ 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 $A \in \mathbb{R}^{d\times d}$ be symmetric and zero-diagonal. Then,*

(i) $\mathbf{E}\{X^\mathsf{T} AX\} = 0$.
(ii) $\mathbf{E}\{(X^\mathsf{T} AX)^2\} = 2\|A\|_F^2$.

(iii) There exists $c_3 > 0$ such that for any integer k we have

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

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

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

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

$$\mathbf{E}\{X^{\mathsf{T}}AX\} = \sum_{i,j} A_{ij}\,\mathbf{E}\{X_iX_j\} = \sum_{i=1}^d A_{ii}\,\mathbf{E}\{X_i^2\} + \sum_{i\neq j} A_{ij}\,\mathbf{E}\{X_i\}\,\mathbf{E}\{X_j\} = 0.$$

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

$$\begin{aligned}\mathbf{E}\{(X^{\mathsf{T}}AX)^2\} &= \mathbf{E}\left\{\left(\sum_{i,j} A_{ij}X_iX_j\right)^2\right\} \\ &= \sum_{i=1}^d A_{ii}^2\,\mathbf{E}\{X_i^4\} + \sum_{i\neq j}\big(A_{ij}^2 + A_{ij}A_{ji}\big)\,\mathbf{E}\{X_i^2X_j^2\} \\ &= 2\sum_{i\neq j} A_{ij}^2 = 2\|A\|_F^2.\end{aligned}$$

For (iii), we integrate

$$\begin{aligned}\mathbf{E}\{(X^{\mathsf{T}}AX)^k\} &\le \int_0^\infty \mathbf{P}\{|X^{\mathsf{T}}AX|^k \ge t\}\,\mathrm{d}t \\ \text{(by Theorem 3.5)}\qquad &\le 2\int_0^\infty e^{-C\frac{t^{1/k}}{K^2\|A\|}}\,\mathrm{d}t + 2\int_0^\infty e^{-C\frac{t^{2/k}}{K^4\|A\|_F^2}}\,\mathrm{d}t \\ &= 2\Gamma(k+1)\left(\frac{K^2\|A\|}{C}\right)^k + 2\Gamma(k/2+1)\left(\frac{K^4\|A\|_F^2}{C}\right)^{k/2} \\ &\le c_3^k K^{2k} k! \|A\|_F^k,\end{aligned}$$

for some $c_3 > 0$.

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

$$\mathbf{E}\{e^{tX^{\mathsf{T}}AX}\} - 1 = \sum_{k=1}^\infty \frac{\mathbf{E}\{(tX^{\mathsf{T}}AX)^k\}}{k!} \le \sum_{k=2}^\infty \frac{(c_3K^2t\|A\|_F)^k k!}{k!} \le 2c_3^2K^4t^2\|A\|_F^2,$$

by (i) and (iii), as long as $2c_3^2K^2t\|A\|_F \le 1$. □

Lemma 3.7. *There exist $c_1, c_2 > 0$ such that for any $\Sigma \in \mathcal{P}_d$ with $\operatorname{tr}(\Sigma^{-1} - I) = 0$ and $\|\Sigma^{-1} - I\|_F \le c_1$, we have*

$$1 \le \det(\Sigma) \le 1 + c_2\|\Sigma^{-1} - I\|_F^2.$$

Proof. Write $\Sigma^{-1} = I + \Delta$. Then,

$$\log\det\Sigma^{-1} = \sum_{i=1}^d \log(1 + \lambda_i(\Delta)) \le \sum_{i=1}^d \lambda_i(\Delta) = \operatorname{tr}(\Delta) = 0,$$

and the lower bound follows. Furthermore, observe that

$$|\lambda_i(\Delta)| \le \|\Delta\| = \|\Sigma^{-1} - I\| \le \|\Sigma^{-1} - I\|_F \le c_1.$$

If c_1 is sufficiently small, then

$$\log\det\Sigma^{-1} = \sum_{i=1}^{d}\log(1+\lambda_i(\Delta)) \geq \sum_{i=1}^{d}(\lambda_i(\Delta) - 2\lambda_i(\Delta)^2) = \operatorname{tr}(\Delta) - 2\operatorname{tr}(\Delta^2).$$

Then, since $\operatorname{tr}(\Delta) = 0$ and $\operatorname{tr}(\Delta^2) = \|\Delta\|_F^2$ by symmetry of Δ,

$$\log\det\Sigma = -\log\det\Sigma^{-1} \leq -\operatorname{tr}(\Delta) + 2\operatorname{tr}(\Delta^2) = 2\|\Sigma^{-1} - I\|_F^2,$$

and again for sufficiently small c_1,

$$\det(\Sigma) \leq e^{2\|\Sigma^{-1}-I\|_F^2} \leq 1 + 4\|\Sigma^{-1} - I\|_F^2. \qquad \square$$

Lemma 3.8. *There are* $c_1, c_2, c_3 > 0$ *such that for any* $\Sigma, \widetilde{\Sigma} \in \mathcal{P}_d$ *such that* $\Sigma^{-1} - I$ *and* $\widetilde{\Sigma}^{-1} - I$ *are zero-diagonal and* $\max\{\|\Sigma^{-1} - I\|_F, \|\widetilde{\Sigma}^{-1} - I\|_F\} \leq c_1$,

$$\|f_\Sigma - f_{\widetilde{\Sigma}}\|_1 \geq c_2\|\Sigma^{-1} - \widetilde{\Sigma}^{-1}\|_F - c_3(\|\Sigma^{-1} - I\|_F^2 + \|\widetilde{\Sigma}^{-1} - I\|_F^2).$$

Proof. By Lemma 3.7 and the triangle inequality,

$$\begin{aligned}
&\|f_\Sigma - f_{\widetilde{\Sigma}}\|_1 \\
&= (2\pi)^{-d/2}\int_{\mathbb{R}^d}\left|\frac{e^{-\frac{1}{2}x^\mathsf{T}\Sigma^{-1}x}}{\sqrt{\det(\Sigma)}} - \frac{e^{-\frac{1}{2}x^\mathsf{T}\widetilde{\Sigma}^{-1}x}}{\sqrt{\det(\widetilde{\Sigma})}}\right|\mathrm{d}x \\
&\geq \quad (2\pi)^{-d/2}\int_{\mathbb{R}^d} e^{-\frac{1}{2}x^\mathsf{T}x}\left|e^{-\frac{1}{2}x^\mathsf{T}(\Sigma^{-1}-I)x} - e^{-\frac{1}{2}x^\mathsf{T}(\widetilde{\Sigma}^{-1}-I)x}\right|\mathrm{d}x \\
&\quad - c_4(\|\Sigma^{-1} - I\|_F^2 + \|\widetilde{\Sigma}^{-1} - I\|_F^2) \\
&= \quad \mathbf{E}\Big\{\Big|e^{-\frac{1}{2}X^\mathsf{T}(\Sigma^{-1}-I)X} - e^{-\frac{1}{2}X^\mathsf{T}(\widetilde{\Sigma}^{-1}-I)X}\Big|\Big\} \\
&\quad - c_4(\|\Sigma^{-1} - I\|_F^2 + \|\widetilde{\Sigma}^{-1} - I\|_F^2),
\end{aligned}$$

where the expectation is with respect to $X = (X_1, \ldots, X_d) \sim \mathcal{N}(0, I)$, a d-dimensional standard normal vector. Observe now the following chain of elementary inequalities,

$$1 + t \leq e^t \leq 1 + t + \frac{t^2}{2}\max\{e^t, 1\}, \tag{1}$$

which holds for all $t \in \mathbb{R}$. By the triangle inequality again,

$$\begin{aligned}
&\mathbf{E}\Big\{\Big|e^{-\frac{1}{2}X^\mathsf{T}(\Sigma^{-1}-I)X} - e^{-\frac{1}{2}X^\mathsf{T}(\widetilde{\Sigma}^{-1}-I)X}\Big|\Big\} \\
&\geq \quad (1/2)\,\mathbf{E}\Big\{|X^\mathsf{T}(\Sigma^{-1} - \widetilde{\Sigma}^{-1})X|\Big\} \qquad (2)\\
&\quad - (1/8)\,\mathbf{E}\Big\{(X^\mathsf{T}(\Sigma^{-1} - I)X)^2\max\{e^{-X^\mathsf{T}(\Sigma^{-1}-I)X/2}, 1\}\Big\} \qquad (3)\\
&\quad - (1/8)\,\mathbf{E}\Big\{(X^\mathsf{T}(\widetilde{\Sigma}^{-1} - I)X)^2\max\{e^{-X^\mathsf{T}(\widetilde{\Sigma}^{-1}-I)X/2}, 1\}\Big\}. \qquad (4)
\end{aligned}$$

We start with the term (3) in this expression. By Cauchy-Schwarz and Lemma 3.6 (iii), (iv), for some $c_5 > 0$

$$\begin{aligned}
&\mathbf{E}\Big\{(X^\mathsf{T}(\Sigma^{-1} - I)X)^2\max\{e^{X^\mathsf{T}(I-\Sigma^{-1})X/2}, 1\}\Big\} \\
&\leq \sqrt{\mathbf{E}\{(X^\mathsf{T}(\Sigma^{-1} - I)X)^4\}\big(\mathbf{E}\{e^{X^\mathsf{T}(I-\Sigma^{-1})X}\} + 1\big)} \\
&\leq c_5\sqrt{\|\Sigma^{-1} - I\|_F^4} = c_5\|\Sigma^{-1} - I\|_F^2.
\end{aligned}$$

A similar computation gives that (4) is also at most $c_5\|\widetilde{\Sigma}^{-1}-I\|_F^2$, so to complete the proof we need to bound (2) from below. By Hölder's inequality and Lemma 3.6 (ii), (iii), there exists some $c_6 > 0$ for which

$$\begin{aligned}\mathbf{E}\Big\{|X^{\mathsf{T}}(\Sigma^{-1} - \widetilde{\Sigma}^{-1})X|\Big\} &\geq \frac{\mathbf{E}\Big\{(X^{\mathsf{T}}(\Sigma^{-1} - \widetilde{\Sigma}^{-1})X)^2\Big\}^{3/2}}{\mathbf{E}\Big\{(X^{\mathsf{T}}(\Sigma^{-1} - \widetilde{\Sigma}^{-1})X)^4\Big\}^{1/2}} \\ &\geq \frac{(2\|\Sigma^{-1} - \widetilde{\Sigma}^{-1}\|_F^2)^{3/2}}{(c_6\|\Sigma^{-1} - \widetilde{\Sigma}^{-1}\|_F^4)^{1/2}} \\ &= (8/c_6)^{1/2}\|\Sigma^{-1} - \widetilde{\Sigma}^{-1}\|_F. \qquad \square\end{aligned}$$

Proof of Proposition 3.1. In the notation of Lemma 2.2, we have by Theorem 3.2, Lemma 3.4, and Lemma 3.8, for some $c_1, c_2 > 0$,

$$|\mathcal{F}| \geq 2^{m/5}, \quad \alpha \geq c_1\delta\sqrt{m}, \quad \beta \leq c_2\delta^2 m,$$

as long as $\delta^2 m$ is smaller than some absolute constant. So, we may pick $\delta = c_3/\sqrt{n}$ for sufficiently small $c_3 > 0$ for which

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

Then, by Lemma 2.2, $\mathcal{R}_n(\mathcal{F}_G) \geq \alpha/8 \geq (c_1c_3/8)\sqrt{m/n}$ as long as $n \geq c_4 m$ for some $c_4 > 0$. □

4. Learning Ising Graphical Models

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

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

Any configuration $x \in \{-1,1\}^d$ then appears with probability proportional to $\exp\{H(x)\}$. In fact, we can write $H(x) = H_{h,W}(x) = x^{\mathsf{T}}Wx + h^{\mathsf{T}}x$ 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

$$f_{h,W}(x) = \frac{e^{H_{h,W}(x)}}{Z(h,W)}, \tag{5}$$

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

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

Probability distributions whose densities have the form (5) 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 $\mathcal{M}_G$,

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

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

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

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 in fact $\mathcal{R}_n(\mathcal{I}'_G) = 0$. Thus for any $m \ge 0$ and any G with m edges, $\mathcal{R}_n(\mathcal{I}'_G) \le c\sqrt{m/n}$ for some constant $c > 0$.

4.1. **Proof of the lower bound in Theorem 1.3 (ii).** Since all of 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 $n \ge c_1 m$,*

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

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 $|\mathcal{S}| \ge 2^{m/5}$ and for any distinct $s, \widetilde{s} \in \mathcal{S}$, $\|s - \widetilde{s}\|_1 \ge m/3$. 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-distance and KL-divergence between distributions in $\mathcal{I}$, a few intermediate computations are necessary. Throughout this section, let $X = (X_1, \dots, X_d)$ denote a uniformly random vector in $\{-1,1\}^d$. All expectations will be with respect to this random variable.

Lemma 4.2. *There exist $c_1, c_2 > 0$ such that for any zero-diagonal symmetric $W \in \mathbb{R}^{d\times d}$ with $\|W\|_F \le c_1$,*

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

Proof. We have

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

On the one hand, by Lemma 3.6 (i),

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

and on the other hand, by Lemma 3.6 (iv),

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

as long as $\|W\|_F \leq c_1$ for some sufficiently small constant $c_1 > 0$. □

Lemma 4.3. *There exist* $c_1, c_2 > 0$ *such that for any zero-diagonal symmetric* $W, \widetilde{W} \in \mathbb{R}^{d \times d}$ *satisfying* $\max\{\|W\|_F, \|\widetilde{W}\|_F\} \leq c_1$,

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

Proof. We again prove the result 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 Lemma 4.2,

$$\begin{aligned}
&\mathrm{J}(f_W \parallel f_{\widetilde{W}}) \\
&\quad = 2^d\, \mathbf{E}\left\{ (f_W(X) - f_{\widetilde{W}}(X)) \log\left(\frac{f_W(X)}{f_{\widetilde{W}}(X)} \right) \right\} \\
&\quad \leq 2^d\, \mathbf{E}\big\{ (f_W(X) - f_{\widetilde{W}}(X))(H_W(X) - H_{\widetilde{W}}(X)) \big\} + c_3 (\|W\|_F^2 + \|\widetilde{W}\|_F^2) \\
&\quad \leq \mathbf{E}\Big\{ \Big(e^{H_W(X)} - e^{H_{\widetilde{W}}(X)} \Big) (H_W(X) - H_{\widetilde{W}}(X)) \Big\} + c_4 (\|W\|_F^2 + \|\widetilde{W}\|_F^2).
\end{aligned}$$

It is not hard to see using (1) that for all $t, s \in \mathbb{R}$,

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

Using this, we get the following upper bound,

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

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

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

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

The term (6) is $2\|W - \widetilde{W}\|_F^2 \leq 2(\|W\|_F + \|\widetilde{W}\|_F)^2$ by Lemma 3.6 (ii) and the triangle inequality for the Frobenius norm. For (7), by two applications of the Cauchy-Schwarz inequality, and Lemma 3.6 (iii), (iv),

$$\begin{aligned}
&\mathbf{E}\Big\{ |H_W(X) - H_{\widetilde{W}}(X)| H_W(X)^2 \max\{e^{H_W(X)}, 1\} \Big\} \\
&\quad \leq \sqrt{ \mathbf{E}\{(H_W(X) - H_{\widetilde{W}}(X))^2 H_W(X)^4\} (\mathbf{E}\{e^{2H_W(X)}\} + 1) } \\
&\quad \leq c_5 \big(\mathbf{E}\{(H_W(X) - H_{\widetilde{W}}(X))^4\}\, \mathbf{E}\{H_W(X)^8\} \big)^{1/4} \\
&\quad \leq c_6 \|W - \widetilde{W}\|_F \|W\|_F^2 \\
&\quad \leq c_7 \|W\|_F^2.
\end{aligned}$$

A similar bound holds for (8), after which the result follows. □

Lemma 4.4. *There exist $c_1, c_2, c_3 > 0$ such that for any zero-diagonal symmetric $W, \widetilde{W} \in \mathbb{R}^{d\times d}$ with $\max\{\|W\|_F, \|\widetilde{W}\|_F\} < c_1$,*

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

Proof. By the triangle inequality, there is $c_4 > 0$ for which

$$\begin{aligned} \|f_W - f_{\widetilde{W}}\|_1 &= 2^d\,\mathbf{E}\left\{\left|\frac{e^{H_W(X)}}{Z(W)} - \frac{e^{H_{\widetilde{W}}(X)}}{Z(\widetilde{W})}\right|\right\} \\ \text{(by Lemma 4.2)} \qquad &\geq \mathbf{E}\Big\{\Big|e^{H_W(X)} - e^{H_{\widetilde{W}}(X)}\Big|\Big\} - c_4(\|W\|_F^2 + \|\widetilde{W}\|_F^2). \end{aligned}$$

By (1) and the triangle inequality again,

$$\begin{aligned} \mathbf{E}\Big\{\Big|e^{H_W(X)} - e^{H_{\widetilde{W}}(X)}\Big|\Big\} \geq \mathbf{E}\{|H_W(X) - H_{\widetilde{W}}(X)|\} \\ - (1/2)\,\mathbf{E}\Big\{H_W(X)^2\max\{e^{H_W(X)}, 1\}\Big\} \\ - (1/2)\,\mathbf{E}\Big\{H_{\widetilde{W}}(X)^2\max\{e^{H_{\widetilde{W}}(X)}, 1\}\Big\}. \end{aligned}$$

We bound the second term. By Cauchy-Schwarz and Lemma 3.6 (iii), (iv),

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

and a similar analysis works for the third term. For the first term, by Hölder's inequality and Lemma 3.6 (ii), (iii), there is a $c_6 > 0$ for which

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

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 $\mathcal{I}_d'$ be the class of d-dimensional Ising models with no interactions. 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 4.5. *There exist $c_1, c_2 > 0$ such that if $n \geq c_1 d$,*

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

Proof sketch. As in the above arguments, we pick a subclass of $2^{d/5}$ densities of $\mathcal{I}_d'$ 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. □

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

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

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.,* [28, 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,

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

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 rate 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 parametrized by weights $w_{i_1,\dots,i_k} \in \mathbb{R}$ for each $1 \le i_1 < i_2 < \cdots < i_k \le d$, and when $x \in \{-1,1\}^d$,

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

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 $k \ge 3$, our techniques seem to fail. Auffinger and Ben Arous [5] 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 some 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 rate 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 rate of Ising blockmodels.* For a given $S \subseteq [d]$ with $|S| = d/2$ and parameters $\alpha, \beta \in \mathbb{R}$, an *Ising blockmodel* 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. This model, as introduced by Berthet, Rigollet, and Srivastava [6], is motivated by social network analysis and the notion of communities in such networks. In their work [6], Berthet *et al.* are 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 α, β and S are unknown.

References

[1] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Near-optimal-sample estimators for spherical Gaussian mixtures. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1395–1403. Curran Associates, Inc., 2014. Available at http://papers.nips.cc/paper/5251-near-optimal-sample-estimators-for-spherical-gaussian-mixtures.pdf.

[2] H. Ashtiani, S. Ben-David, N. Harvey, C. Liaw, A. Mehrabian, and Y. Plan. Settling the sample complexity for learning mixtures of Gaussians. *ArXiv e-prints*, 2018. Available at https://arxiv.org/abs/1710.05209v2.

[3] H. Ashtiani, S. Ben-David, and A. Mehrabian. Sample-efficient learning of mixtures. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI '18, pages 2679–2686. AAAI Publications., 2018. Available at https://arxiv.org/abs/1706.01596.

[4] P. Assouad. Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris Sér. I Math.*, 296(23):1021–1024, 1983.

[5] A. Auffinger and G. Ben Arous. Complexity of random smooth functions on the high-dimensional sphere. *Ann. Probab.*, 41(6):4214–4247, 2013.

[6] Q. Berthet, P. Rigollet, and P. Srivastava. Exact recovery in the Ising blockmodel. *Ann. Statist.*, 2016. To appear, available at https://arxiv.org/abs/1612.03880.

[7] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

[8] C. Daskalakis, N. Dikkala, and G. Kamath. Testing Ising models. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1989–2007, 2018.

[9] L. Devroye. *A Course in Density Estimation*, volume 14 of *Progress in Probability and Statistics*. Birkhäuser Boston, Inc., Boston, MA, 1987.

[10] L. Devroye and L. Györfi. *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, 1985.

[11] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer Series in Statistics. Springer-Verlag, New York, 2001.

[12] I. Diakonikolas. Learning structured distributions. In *Handbook of Big Data*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 267–283. CRC Press, Boca Raton, FL, 2016.

[13] R. M. Dudley. Central limit theorems for empirical measures. *Ann. Probab.*, 6(6):899–929, 1978.

[14] E. N. Gilbert. A comparison of signalling alphabets. *Bell System Tech. J.*, 31(3):504–522, 1952.

[15] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2016.

[16] L. Hamilton, F. Koehler, and A. Moitra. Information theoretic properties of Markov random fields, and their algorithmic applications. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2463–2472. Curran Associates, Inc., 2017.

[17] R. Z. Has'minskiĭ. A lower bound for risks of nonparametric density estimates in the uniform metric. *Teor. Veroyatnost. i Primenen.*, 23(4):824–828, 1978.

[18] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, second edition, 2013.

[19] I. Ibragimov. Estimation of analytic functions. In *State of the Art in Probability and Statistics (Leiden, 1999)*, volume 36 of *IMS Lecture Notes Monogr. Ser.*, pages 359–383. Inst. Math. Statist., Beachwood, OH, 2001.

[20] A. Kalai, A. Moitra, and G. Valiant. Disentangling Gaussians. *Comm. ACM*, 55(2), 2012.

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

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

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

[24] S. L. Lauritzen. *Graphical Models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1996.

[25] L. Le Cam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53, 1973.

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

[27] N. P. Santhanam and M. J. Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Trans. Inform. Theory*, 58(7):4117–4134, 2012.

[28] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

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

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

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

[32] M. 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*. Springer, Heidelberg, 2011.

[33] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[34] V. N. Vapnik and A. J. Červonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Verojatnost. i Primenen.*, 16:264–279, 1971.

[35] R. R. Varšamov. The evaluation of signals in codes with correction of errors. *Dokl. Akad. Nauk*, 117(5):739–741, 1957.

[36] R. Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018. To appear, available at https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf.

[37] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, New York, 1997.

Email address: lucdevroye@gmail.com, abbas.mehrabian@gmail.com, tommy.reddad@gmail.com

School of Computer Science, McGill University, 3480 University Street, Montréal, Québec, Canada, H3A 2K6