Proceedings of Machine Learning Research vol 75:1–22, 2018 31st Annual Conference on Learning Theory

# The Mean-Field Approximation: Information Inequalities, Algorithms, and Complexity

**Vishesh Jain** VISHESHJ@MIT.EDU
**Frederic Koehler** FKOEHLER@MIT.EDU
*Massachusetts Institute of Technology. Department of Mathematics.*

**Elchanan Mossel** * ELMOS@MIT.EDU
*Massachusetts Institute of Technology. Department of Mathematics and IDSS.*

**Editors:** Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

## Abstract

The mean field approximation to the Ising model is a canonical variational tool that is used for analysis and inference in Ising models. We provide a simple and optimal bound for the KL error of the mean field approximation for Ising models on general graphs, and extend it to higher order Markov random fields. Our bound improves on previous bounds obtained in work in the graph limit literature by Borgs, Chayes, Lovász, Sós, and Vesztergombi and recent works by Basak and Mukherjee, and Eldan. Our bound is tight up to lower order terms.

Building on the methods used to prove the bound, along with techniques from combinatorics and optimization, we study the algorithmic problem of estimating the (variational) free energy for Ising models and general Markov random fields. For a graph $G$ on $n$ vertices and interaction matrix $J$ with Frobenius norm $\|J\|_F$, we provide algorithms that approximate the free energy within an additive error of $\epsilon n\|J\|_F$ in time $\exp(poly(1/\epsilon))$. We also show that approximation within $(n\|J\|_F)^{1-\delta}$ is NP-hard for every $\delta > 0$. Finally, we provide more efficient approximation algorithms, which find the optimal mean field approximation, for ferromagnetic Ising models and for Ising models satisfying Dobrushin's condition.

## 1. Introduction

One of the most widely studied models in statistical physics is the Ising model. An *Ising model* is specified by a probability distribution on the discrete cube $\{\pm 1\}^n$ of the form

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

where the collection $\{J_{i,j}\}_{i,j\in\{1,\dots,n\}}$ are the entries of an arbitrary real, symmetric matrix with zeros on the diagonal. The distribution $P$ is referred to as the *Boltzmann distribution*. The normalizing constant $Z = \sum_{x\in\{\pm1\}^n}\exp(\sum_{i,j=1}^n J_{i,j}x_i x_j)$ is called the *partition function* of the Ising model and the quantity $\mathcal{F} := \log Z$ is known as the *free energy*.

The free energy is a key physical quantity. It provides important information about the structure of the Boltzmann distribution. Given a naturally growing family of (possibly weighted) graphs with adjacency matrices $M_n$, one of the main problems of interest in statistical physics is to compute the asymptotics of the (suitably renormalized) free energy of the sequence of Ising models $J_n(\beta) =$

* Supported by ONR grant N00014-16-1-2227 and NSF CCF-1665252 and DMS-1737944.

$\beta M_n$ in the $n \to \infty$ limit for all values of $\beta$, where $\beta > 0$ is a parameter referred to as the *inverse temperature*. This is because understanding the behavior of the free energy reveals a wealth of information about the underlying Ising model. For instance, points of non-smoothness in the limiting free energy (as a function of $\beta$) reveal the location of *phase transitions*, which typically correspond to significant changes in the behavior of the underlying Boltzmann distribution e.g. the emergence of long-range correlations. In addition, many other quantities of interest (such as net magnetization) can be computed in terms of free energies.

Although originally introduced in statistical physics, Ising models and their generalizations have also found a wide range of applications in many different areas like statistics, computer science, combinatorics, social networks, and biology (see, e.g., the references in Basak and Mukherjee (2017)). Studying the free energy is of great interest in many of these applications as well.

In fact for every finite $\beta$, the free energy corresponds to the objective value of a natural optimization problem of its own. More precisely, the free energy is characterized by the following *variational principle* (dating back to Gibbs, see the references in Ellis (2007)):

$$\mathcal{F} = \max_{\mu} \left[ \sum_{i,j} J_{ij} \mathbb{E}_{\mu}[X_i X_j] + H(\mu) \right], \tag{1}$$

where $\mu$ ranges over all probability distributions on the boolean hypercube $\{\pm 1\}^n$. This can be seen by noting that

$$\mathbf{KL}(\mu||P) = \mathcal{F} - \sum_{i,j} J_{ij} \mathbb{E}_{\mu}[X_i X_j] - H(\mu), \tag{2}$$

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

By substituting $J = \beta M$ in equation Eq. (1), we see that the Boltzmann distribution is simply the maximum entropy distribution $\mu$ for a fixed value of the expected energy $\mathbb{E}_{\mu}[x^T M x]$. Thus, studying the free energy for different values of $\beta$ provides much richer information about the optimization landscape of $x \mapsto x^T M x$ over the hypercube than just the maximum value, e.g., in the max-cut case, the free energies encode information about non-maximal cuts as well (see e.g. Borgs et al. (2012) for related discussion).

Apart from the applications mentioned above, it is clear by definition that knowledge of the free energy (or equivalently, the partition function) allows one to perform fundamental inference tasks like computing marginals and posteriors in Ising models and their generalizations. Unfortunately, the partition function, which is defined as a sum of exponentially many terms, turns out to be both theoretically and computationally intractable. Closed form expressions for the partition function are extremely hard to come by; in fact, providing such an expression even for the Ising model on the standard 3-dimensional lattice remains one of the most outstanding problems in statistical physics. From a computational perspective, it is known that exactly computing the partition function of an Ising model with $J$ the adjacency matrix of a nonplanar graph is NP-hard (Istrail, 2000), and that approximate sampling/approximating the partition function is still NP-hard, even e.g. in the case of graphs with degree bounded by a small constant (Sly and Sun, 2012).

### 1.1. The mean-field approximation: structural results

Since exact computations, either analytic or otherwise, are typically infeasible, it is natural to look at schemes for approximating the partition function or the free energy. The *naive mean-field approximation* provides one of the simplest and most common methods for doing this.

The mean-field approximation to the free energy (also referred to as the *variational free energy*) is obtained by restricting the distributions $\mu$ in the variational characterization of the free energy (Eq. (1)) to be product distributions. Accordingly, we define the *variational free energy* by

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

Indeed, if $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)$ is the optimizer in the above definition, then the product distribution $\nu$ on the boolean hypercube, with the $i^{th}$ coordinate having expected value $\bar{x}_i$, minimizes $\mathbf{KL}(\mu||P)$ among all product distributions $\mu$. Moreover, it is immediately seen from Eq. (2) that the value of this minimum KL is exactly $\mathcal{F} - \mathcal{F}^*$. Thus, the quantity $\mathcal{F} - \mathcal{F}^*$, which measures the quality of the mean-field approximation, may be interpreted information theoretically as the divergence between the closest product distribution to the Boltzmann distribution and the Boltzmann distribution itself.

Owing to its simplicity, the mean field approximation has long been used in statistical physics (see, e.g., Parisi (1988) for a textbook treatment) and also in Bayesian statistics (Anderson and Peterson, 1987; Jordan et al., 1999; Wainwright and Jordan, 2008), where it is one of the prototypical examples of a *variational method*. As a variational method, the mean field approximation has the attractive property that it always gives a valid lower bound for the free energy. It is well known (Ellis and Newman, 1978) that the mean field approximation is very accurate for the Curie-Weiss model, which is the Ising model on the complete graph (see Example 1 for a complete description of the model). On the other hand, it is also known (Dembo and Montanari, 2010) that for very sparse graphs like trees of bounded arity, this is not the case. In recent years, considerable effort has gone into bounding the error of the mean-field approximation on more general graphs; we will give a detailed comparison of our results with recent work in Section 1.4. Our main structural result is the following inequality, which gives an explicit bound on the error of the mean field approximation for general graphs:

**Theorem 1** *Fix an Ising model $J$ on $n$ vertices. Let $\nu := \arg\min_\nu \mathbf{KL}(\nu||P)$, where $P$ is the Boltzmann distribution and the minimum ranges over all product distributions. Then,*

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

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

This result is *tight up to logarithmic factors*, not just for product distributions, but also for a large class of variational methods. In particular, this class includes approximation by bounded mixtures of product distributions (as in Jaakkola and Jordan (1998)), as well as (mixtures of) restricted classes of Ising models, e.g. Ising models on acyclic graphs (see the discussion of tractable families in Wainwright and Jordan (2008)). Some other methods for estimating the free energy, such as the Bethe approximation and the method of Risteski (2016), optimize over *pseudo-distributions* and so the theorem itself cannot be directly applied, but essentially the same obstruction should still apply.

**Theorem 2** *Let $(\mathcal{Q}_n)_{n=0}^\infty$ be a sequence of families of probability distributions on $\{\pm 1\}^n$ which are closed under the following two operations:*

1. *Conditioning on variables: if $Q \in \mathcal{Q}_n$, $i \in [n]$, and $x_i \in \{\pm 1\}$, then the conditional distribution of $X_{\sim i}$ under $Q$ given $X_i = x_i$, which is a probability distribution on $\{\pm 1\}^{n-1}$, is in $\mathcal{Q}_{n-1}$.*

2. *Taking products: if* $Q_1 \in \mathcal{Q}_m$ *and* $Q_2 \in \mathcal{Q}_n$*, then* $Q_1 \times Q_2 \in \mathcal{Q}_m \times \mathcal{Q}_n$.

*Furthermore, suppose that* $(\mathcal{Q}_n)_{i=1}^{\infty}$ *does not contain the class of all probability distributions induced by Ising models. Then, there exists a sequence* $(J_i)_{i=1}^{\infty}$ *of Ising models of increasing size* $n_i$ *and with Boltzmann distributions* $P_{J_i}$ *such that*

$$\mathbf{KL}(Q_{n_i}||P_{J_i}) = \Omega(n_i^{2/3}\|J_{n_i}\|_F^{2/3}),$$

*where* $Q_{n_i} := \arg\min_{Q\in\mathcal{Q}_{n_i}} \mathbf{KL}(Q, P_{J_i})$.

**Remark 3** *The above theorem shows that our bound is essentially optimal in the natural univariate quantity* $n\|J\|_F$*. However, the construction we use does not contradict any upper bound of the form* $O(n^{1-\alpha}\|J\|_F^{2\alpha})$ *for* $\alpha \in [0,1]$*. As there is always a trivial bound* $O(n)$ *for the mean-field approximation (consider the optimal point-mass distribution), we may assume that* $\|J\|_F = o(n^{1/2})$ *and ask about the supremum of all* $\alpha$ *such that an upper bound of this form holds. We conjecture that the supremum is, in fact,* $\alpha = 1/3$*, so that our bound is optimal in this stronger sense as well.*

Our methods extend in a straightforward manner not just to Ising models with external fields, but indeed to general higher order Markov random fields, as long as we assume a bound $r$ on the order of the highest interaction (i.e. size of the largest hyper-edge). The results also generalize naturally to the case of non-binary alphabets but for simplicity, we only discuss the binary case.

**Definition 4** *Let* $J$ *be an arbitrary function on the hypercube* $\{\pm 1\}^n$ *and suppose that the degree of* $J$ *is* $r$ *i.e. the Fourier decomposition of* $J$ *is* $J(x) = \sum_{\alpha\subset[n]} J_\alpha x^\alpha$ *with* $r = \max_{J_\alpha\neq 0}|\alpha|$*. The corresponding* order $r$ (binary) Markov random field *is the probability distribution on* $\{\pm 1\}^n$ *given by*

$$P(X = x) = \frac{1}{Z}\exp(J(x))$$

*where the normalizing constant* $Z$ *is referred to as the* partition function*. For any polynomial* $J$ *we define* $J_{=d}$ *to be its* $d$*-homogeneous part and* $\|J\|_F$ *to be the square root of the total Fourier energy of* $J$ *i.e.* $\|J\|_F^2 := \sum_\alpha |J_\alpha|^2$.

**Theorem 5** *Fix an order* $r$ *Markov random field* $J$ *on* $n$ *vertices. Let* $\nu := \arg\min_\nu \mathbf{KL}(\nu||P)$*, where* $P$ *is the Boltzmann distribution and the minimum ranges over all product distributions. Then,*

$$\mathbf{KL}(\nu||P) = \mathcal{F} - \mathcal{F}^* \leq 2000r \max_{1\leq d\leq r} d^{1/3}n^{d/3}\|J_{=d}\|_F^{2/3}\log^{1/3}(d^{1/3}n^{d/3}\|J_{=d}\|_F^{2/3} + e).$$

### 1.2. Examples

We give a few examples of natural families of Ising models in order to illustrate the consequences of our bounds.

**Example 1 (Curie-Weiss)** *As our first example, we show how our bounds imply classical results about the Curie-Weiss model (see Ellis and Newman (1978)), in which* $J_{ij} = (\beta/2n)$ *for* $i \neq j$ *and there is a uniform external field* $h$*. In this case, we can explicitly solve the variational problem; indeed, by checking the first-order optimality condition (Eq. (3)), we see that an optimal product distribution with marginals* $\mathbb{E}[X_i] = x_i$ *must have* $x_i = \tanh(\sum_{j:j\neq i}\beta x_j/n + h)$*. Furthermore,*

*since $x_i < x_j$ implies that $\tanh(\sum_{k:k\neq i} \beta x_k/n + h) > \tanh(\sum_{k:k\neq j} \beta x_k/n + h)$, it follows that we cannot have $x_i < x_j$ for any pair $(i,j)$. Therefore, the optimal product distribution has all marginals equal to $x$, where $x$ is a solution of*

$$x = \tanh((1-1/n)\beta x + h).$$

*Taking $n \to \infty$ and $h = 0$, this correctly predicts a phase transition at $\beta = 1$; the mean field equations go from having just one solution ($x = 0$) to two additional "symmetry-breaking" solutions with $x \neq 0$. By Theorem 5, we see that for any constant $\beta, h$, the normalized free energy $\mathcal{F}/n$ agrees with $\mathcal{F}^*/n$ in the $n \to \infty$ limit with error decaying at least as fast as $\tilde{O}(n^{-1/3})$.*

**Example 2 (Uniform edge weights on graphs of increasing degree)** *Fix $\beta \in \mathbb{R}$ and a sequence of graphs $(G_{n_i})_{i=1}^{\infty}$ with the number of vertices $n_i$ going to infinity, and let $m_i$ be the corresponding number of edges. Then, it is natural to look at the model with uniform edge weights equal to $\beta n_i/m_i$, since this makes the maximum value of $x^T Jx$ on the order of $\Theta(n_i)$, which is the same scale as the entropy term in the variational definition of the free energy (Eq. (1)). We say the model is* ferromagnetic *if $\beta > 0$ and* anti-ferromagnetic *if $\beta < 0$. Observe that $\|J\|_F = \beta n_i/\sqrt{m_i}$, so that by Theorem 1, we have $|\mathcal{F}/n_i - \mathcal{F}^*/n_i| = O(n_i^{1/3}\log^{1/3} n_i/m_i^{1/3})$. In particular, this goes to $0$ as long as $m_i = \omega(n_i \log n_i)$.*

**Example 3 (Uniform edge weights on $r$-uniform hypergraphs)** *Fix $\beta \in \mathbb{R}$ and let $(G_{n_i})_{i=1}^{\infty}$ be a sequence of $r$-uniform hypergraphs with $n_i$ vertices and $m_i$ hyperedges. Analogous to the graph case, we let $J(x) = \frac{\beta n_i}{m_i}\sum_{S\in E(G_{n_i})} x_S$, so that the maximum of $J$ is on the same order as the entropy term in the free energy. We still have $\|J\|_F = \beta n_i/\sqrt{m_i}$, and see by Theorem 5 that $|\mathcal{F}/n_i - \mathcal{F}^*/n_i| = O(n_i^{(r-1)/3}\log n_i/m_i^{1/3})$. This converges to $0$ as long as $m_i = \omega(n_i^{r-1}\log n_i)$.*

### 1.3. Algorithmic results

Next, we study the algorithmic aspects of the mean field approximation and variational methods. We begin by showing that in a certain *high-temperature regime* (specifically, the range of parameters satisfying the *Dobrushin uniqueness criterion* Dobrushin (1968)), the minimization problem defining the variational free energy is convex.

**Theorem 6** *Suppose $J$ is the interaction matrix of an Ising model with arbitrary external field $h_i$ at vertex $i$, and suppose that for every row $i$ of $J$, we have $\sum_j 2|J_{ij}| \leq 1$. Then, the maximization problem defining the variational free energy is concave, and hence can be solved to additive $\epsilon$-error in time $poly(n, \log(1/\epsilon))$.*

**Remark 7** *Note that in the literature (e.g. Dobrushin (1968)), the Dobrushin uniqueness criterion is stated as $\sum_j |J_{i,j}| \leq 1$. This corresponds to the above condition $\sum_j 2|J_{i,j}| \leq 1$ in our normalization, since we do not insert a factor of $1/2$ in front of the quadratic term in the definition of (variational) free energy.*

A well known heuristic for finding the optimal mean-field approximation (see, e.g., the discussion in Wainwright and Jordan (2008)) consists of iterating the *mean-field equations* to search for a fixed point. The mean field equations are just the first-order optimality conditions for $\mathcal{F}^*$:

$$x^* = \tanh^{\otimes n}(2Jx^* + h). \tag{3}$$

In the Dobrushin uniqueness regime, we prove that this message passing algorithm in fact converges exponentially fast to the optimum of the variational free energy.

**Theorem 8** *Suppose $J$ is the interaction matrix of an Ising model with arbitrary external field $h_i$ at vertex $i$, and suppose that for every row $i$ of $J$, we have $\sum_j 2|J_{ij}| \leq 1 - \eta$ for some uniform $\eta > 0$. Let $x^*$ be the optimizer of the optimization problem given by $\mathcal{F}^*$. Let $x_0$ be an arbitrary point in $[-1, 1]^n$ and iteratively define $x_n := \tanh^{\otimes n}(2Jx_{n-1} + h)$. Then,*

$$\|x_n - x^*\|_\infty \leq (1 - \eta)^n \|x_0 - x^*\|_\infty \leq 2(1 - \eta)^n.$$

**Remark 9** *The high-temperature assumption is necessary for this algorithm to converge quickly to the optimum. In the super-critical Curie Weiss model without external field (Example 1 with $\beta > 1$ and $h = 0$), we see that $x = (0, \ldots, 0)$ is a critical point for the variational free energy (fixed point of the mean-field equations) but not the global optimum. Furthermore, even if we start from the point $(\epsilon, \ldots, \epsilon)$ for $\epsilon$ a small positive number, we see that for $\beta$ large, iterating the mean field equations converges exponentially slowly in $\beta$ as $\tanh'(\beta)$ is exponentially small in $\beta$.*

Even though there exist such situations where the optimization problem defining the variational free energy is *non-convex* and the message passing algorithm may fail or converge exponentially slowly (see Remark 9), there is a way to solve the optimization problem in polynomial time as long as the model is ferromagnetic.

**Theorem 10** *Fix an Ising model $J$ on $n$ vertices which is ferromagnetic (i.e. $J_{ij} \geq 0$ for every $i, j$) and has uniform external field $h$ at every node. There is a randomized algorithm which runs in time $poly(n, 1/\epsilon, \log(1/\delta))$ and succeeds with probability at least $1 - \delta$ in solving the optimization problem defining $\mathcal{F}^*$ up to $\epsilon$-additive error.*

However, in the general case, we show that it is NP-hard to estimate the variational free energy. In fact, it is NP-hard to return an estimate to the free energy within additive error $n^{1-\delta}\|J\|_F^{1-\delta}$, whereas by Theorem 1 and Theorem 5, the true variational free energy is much closer than this.

**Theorem 11** *For any fixed $\delta > 0$, it is NP-hard to approximate the free energy $\mathcal{F}$ (or variational free energy $\mathcal{F}^*$) of an Ising model $J$ within an additive error of $n^{1-\delta}\|J\|_F^{1-\delta}$. More generally, for an $r$-uniform Markov random field, it is NP-hard to approximate $\mathcal{F}$ within an additive error of $(n^{r/2}\|J_{=r}\|_F)^{1-\delta}$.*

We now give an algorithm to approximate the free energy in the most general setting; in light of the NP-hardness result (Theorem 11) this approximation must be roughly on the scale of $n\|J\|_F$. In the general setting, the only previous algorithm which gives non-trivial guarantees for approximating the log-partition function is that of (Risteski, 2016), which requires time $n^{O(1/\epsilon^2)}$ as well as stronger density assumptions in order to provide a guarantee similar to Theorem 12. In comparison, the algorithm we give has the advantage that it runs in *constant-time* for fixed $\epsilon$.

**Theorem 12** *Fix $\epsilon > 0$. There is an algorithm which runs in time $2^{O(\log(1/\epsilon)/\epsilon^2)}$ and returns, with probability at least $0.99$, an implicit description of a product distribution $\mu$ such that*

$$\mathbf{KL}(\mu||P) \leq \epsilon n\|J\|_F + C\log(1/\epsilon)/\epsilon^2 + 0.5^{2^{1/\epsilon^2}} n$$

*and an estimate to the free energy $\hat{\mathcal{F}}$ such that*

$$|\mathcal{F} - \hat{\mathcal{F}}| \leq \epsilon n \|J\|_F + C' \log(1/\epsilon)/\epsilon^2 + 0.5^{2^{1/\epsilon^2}} n,$$

*where $C$ and $C'$ are absolute constants.*

**Remark 13** *Typically, the first term in the bound of Theorem 12 dominates. In particular, the last term $0.5^{2^{1/\epsilon^2}} n$ is dominated by the first term except in a very unusual regime where $\|J\|_F$ is extremely small, and even then it vanishes doubly-exponentially fast as we take $\epsilon \to 0$.*

Our algorithm extends to general order $r$ Markov random fields as well. In Theorem 22 we obtain an algorithm that runs in time $2^{O(\log(1/\epsilon)/\epsilon^{2r-2})}$ while in Theorem 23 the running time is $2^{O(\log(1/\epsilon)/\epsilon^2)} n^r$. See Appendix A for further details.

### 1.4. Comparison with previous work

As mentioned earlier, providing guarantees on the quality of the mean-field approximation for general graphs has attracted much interest in recent years. Notably, in the context of graph limits (Borgs et al., 2012), the following result (stated here in our notation[1]) was shown:

$$|\mathcal{F}^*/n - \mathcal{F}/n| \leq \frac{48}{n^{1/4}} + \frac{130 n \|J\|_\infty}{\sqrt{\log n}} + \frac{5|h|}{n^{1/2}}.$$

Here, $\|J\|_\infty$ denotes the absolute value of the largest entry of $J$. This result was sufficient for the application in (Borgs et al., 2012), i.e., proving convergence of the free energy density and the variational free energy density for sequences of dense graphs (i.e. those with $\Theta(n^2)$ many edges). In this case, it is natural to take $\|J\|_\infty = O(1/n)$ and thus, their error bound converges to 0 at rate $1/\sqrt{\log n}$. They used this bound to prove that defining the free energy density of a graphon in terms of the variational free energy density is asymptotically consistent with the combinatorial definition of the free energy in terms of sums over states (which cannot naively be made sense of in the graphon setting).

The bound in (Borgs et al., 2012) has two limitations: first, it does not provide any information about models where $\|J\|_\infty = \omega(\sqrt{\log n}/n)$ – a setting which includes essentially all natural models on graphs with $o(n^2)$ edges – and secondly, the convergence rate of $1/\sqrt{\log n}$ is very slow: in order to get $\epsilon$ error in the bound, we must look at graphs of size $2^{1/\epsilon^2}$, which raises the possibility that the approximation may perform badly even on very large graphs.

The papers of (Borgs et al., 2018) and (Basak and Mukherjee, 2017) resolve the first issue by giving bounds which extend to sparser graphs. In our context, the latter result is more relevant, and we refer the reader to the discussion in (Basak and Mukherjee, 2017) for the relationship to (Borgs et al., 2018). The main result of (Basak and Mukherjee, 2017) is that $|\mathcal{F}^*/n - \mathcal{F}/n| = o(1)$ whenever $\|J\|_F^2 = o(n^2)$. As noted by the authors, if we do not care about the rate of convergence, then this result is tight – there are simple examples of models with $\|J\|_F^2 = \Theta(n^2)$ where $|\mathcal{F}^*/n - \mathcal{F}/n| = \Omega(1)$. However, their result is focused on the asymptotic regime and does not give good

1. In their paper, the edge weights are normalized by $1/n$ so that on dense graphs, the limit as $n \to \infty$ will sensibly converge. Their bound is stated for the slightly more general setting of models over finite alphabets – to facilitate ease of comparison, we have stated it only in the simplest case of binary Ising models with uniform external field $h$.

control of the rate of convergence (though it is certainly possible to extract some effective bound from their argument). Most recently, a much better convergence rate was shown in the work of (Eldan, 2016) where, under a mild assumption on $\|J\|_\infty$, an explicit bound of $O(n^{5/6}\|J\|_F^{1/3})$ for the error of mean-field approximation was shown.

Our result improves on all of this previous work by giving an explicit bound (Theorem 2) of the form $\tilde{O}(n^{2/3}\|J\|_F^{2/3})$. Except for the log factor, our bound is always superior to that of (Eldan, 2016): see Remark 3. Concretely, in the setting of dense graphs with edge weights scaled by $1/n$, our bound shows a convergence rate of $\tilde{O}(n^{-1/3})$ for the rescaled error $|\mathcal{F}^*/n - \mathcal{F}/n|$ whereas the previously best known bound of (Eldan, 2016) showed only the slower rate $O(n^{-1/6})$.

It is interesting to note that both our result and (Borgs et al., 2012) use the Frieze-Kannan weak regularity lemma. However, our analysis introduces a number of new ideas that let us avoid the $2^{1/\epsilon^2}$ dependence which is typical in applications of the weak regularity lemma, thereby obtaining bounds with exponentially better dependence on $n$. Besides giving the best known convergence rate, our result is almost as strong as (Basak and Mukherjee, 2017; Eldan, 2016) asymptotically and has a much simpler proof which generalizes easily to higher-order Markov random fields.

As far as algorithmic results are concerned, there has been a very long line of work historically in understanding the performance of Markov Chain Monte Carlo methods (MCMC), especially the Glauber chain (Gibbs sampling). As mentioned earlier, it is known from (Dobrushin, 1968) that the Glauber dynamics mix rapidly in the Dobrushin uniqueness regime, where the entries of each row of $J$ are bounded by $(1-\eta)/2$. There has been a lot of work on improvements to this result, see for example (Mossel and Sly, 2013) for a tight result on bounded degree graphs. Although the Glauber dynamics typically cannot mix rapidly in the low-temperature regime (see e.g. Sly and Sun (2012)), in the special case where $J$ is ferromagnetic, there is a different Markov chain which can approximately sample from the Boltzmann distribution in polynomial time (Jerrum and Sinclair, 1990); see also recent work (Liu et al., 2017). With respect to Theorem 8, we note that some related ideas have been used in the convergence analysis of other algorithms like loopy belief propagation (see for example Tatikonda and Jordan. (2002); Mooij and Kappen (2007)).

Note that in situations where Markov chain methods do work, they allow for approximate sampling and approximation of the partition function to a higher precision than our results. However, in the general case where Markov chain methods typically have no guarantees, the previous best result is due to (Risteski, 2016), which gave a similar guarantee for approximating the free energy as our Theorem 12, but requiring stronger density assumptions as well as $n^{O(1/\epsilon^2)}$ time. It is interesting to note that this algorithm is also a variational method which *upper bounds* the free energy by relaxing Eq. (1) to *pseudo-distributions*, whereas mean-field approximation optimizes a *lower bound*.

### 1.5. Outline of the techniques

The proof of our main structural inequality is based on the weak regularity lemma of Frieze and Kannan (Theorem 16). Roughly speaking, this lemma allows us to (efficiently) partition the underlying weighted graph into a small number of blocks in a manner such that "cut-like" quantities associated to the graph approximately depend only on the *numbers* of edges between various blocks. It is well known (see, e.g., Borgs et al. (2012), and also Lemma 17) that the free energy and variational free energy fit into this framework. This observation shows that in order to prove Theorem 1, it is sufficient to prove the statement for such graphs composed of a small number of blocks.

In order to do this, we will first show that the free energy for such graphs is well approximated by an intermediate optimization problem (Eq. (5)) which is quite similar to the one defining the variational free energy. Next, we will use basic facts about entropy to show that the solution to this optimization problem is indeed close to the variational energy (Lemma 20). We now describe this intermediate optimization problem.

The key point in the weak regularity lemma is that the number of blocks depends only on the desired quality of approximation, and *not* of the size of the underlying graph. Since we only care about the numbers of edges between the various blocks, this allows us to approximately rewrite the sum computing the partition function in terms of only *polynomially* many nonnegative summands, as opposed to the *exponentially* many nonnegative summands we started out with (Eq. (4)). Moreover, since none of the edge weights coming from the weak regularity lemma are too big, one can further group terms to reduce the number of summands to a polynomial in only the error parameter, independent of the number of vertices in the original graph (Lemma 19). This provides the desired intermediate optimization problem – the log of the largest summand of this much smaller sum approximates the free energy well (Eq. (6), Eq. (7)).

For the proof of Theorem 12, we show that solving (a slight variation of) this intermediate optimization problem amounts to solving a number of convex programs. However, since we want to provide algorithms which run in constant time (see Remark 27), we first need to rewrite these programs in a manner which uses only a constant number of variables and constraints. The proofs of the corresponding theorems for general order $r$ Markov random fields follow a similar outline, with the application of Theorem 16 replaced by Theorem 21 or Theorem 29.

## 2. Preliminaries

We will make essential use of the weak regularity lemma (Frieze and Kannan, 1999). Before stating it, we introduce some terminology. Throughout this section, we will deal with $m \times n$ matrices whose entries we will index by $[m] \times [n]$, where $[k] = \{1, \ldots, k\}$.

**Definition 14** *Given $S \subseteq [m]$, $T \subseteq [n]$ and $d \in \mathbb{R}$, we define the* $[m] \times [n]$ Cut Matrix $C = CUT(S, T, d)$ *by*

$$C(i,j) = \begin{cases} d & \text{if } (i,j) \in S \times T \\ 0 & \text{otherwise} \end{cases}$$

**Definition 15** *A* Cut Decomposition *expresses a matrix $J$ as*

$$J = D^{(1)} + \cdots + D^{(s)} + W$$

*where $D^{(i)} = CUT(R_i, C_i, d_i)$ for all $t = 1, \ldots, s$. We say that such a cut decomposition has* width $s$, coefficient length $(d_1^2 + \cdots + d_s^2)^{1/2}$ *and* error $\|W\|_{\infty \mapsto 1}$.

We are now ready to state the weak regularity lemma of Frieze and Kannan. The particular choice of constants can be found in (Alon et al., 2002).

**Theorem 16 (Frieze and Kannan (1999))** *Let $J$ be an arbitrary real matrix, and let $\epsilon > 0$. Then, we can find a cut decomposition of width at most $16/\epsilon^2$, coefficient length at most $4\|J\|_F/\sqrt{mn}$, error at most $4\epsilon\sqrt{mn}\|J\|_F$, and such that $\|W\|_F \leq \|J\|_F$.*

## 3. Proof of the main structural result

We begin by showing that both the free energy and the variational free energy are 1-Lipschitz with respect to the cut norm of the matrix of interaction strengths.

**Lemma 17** *Let $J$ and $D$ be the matrices of interaction strengths of Ising models with partition functions $Z$ and $Z_D$, and variational free energies $\mathcal{F}^*$ and $\mathcal{F}^*_D$. Then, with $W := J - D$, we have $|\log Z - \log Z_D| \leq \|W\|_{\infty\mapsto 1}$ and $|\mathcal{F}^* - \mathcal{F}^*_D| \leq \|W\|_{\infty\mapsto 1}$.*

**Proof** Note that for any $x \in [-1,1]^n$, we have

$$|\sum_{i,j} J_{i,j}x_ix_j - \sum_{i,j} D_{i,j}x_ix_j| = |\sum_i (\sum_j W_{i,j}x_j)x_i| \leq \sum_i |\sum_j W_{i,j}x_j| \leq \|W\|_{\infty\mapsto 1},$$

from which we immediately get that $|\mathcal{F}^* - \mathcal{F}^*_D| \leq \|W\|_{\infty\mapsto 1}$. Moreover, for any $x \in \{\pm 1\}^n$, we have

$$\exp\left(\sum_{i,j} J_{i,j}x_ix_j\right) \in \left[\exp\left(\sum_{i,j} D_{i,j}x_ix_j) \pm \|W\|_{\infty\mapsto 1}\right)\right].$$

Taking first the sum of these inequalities over all $x \in \{\pm 1\}^n$ and then the log, we get

$$\log Z \in \left[\log\left(\sum_{x\in\{\pm 1\}^n} \exp\left(x^T Dx\right)\right) \pm \|W\|_{\infty\mapsto 1}\right],$$

as desired. ■

**Remark 18** *For the remainder of this section, we take $D := D^{(1)} + \cdots + D^{(s)}$, where $D^1, \ldots, D^s$ are the cut matrices coming from applying Theorem 16 to $J$ with parameter $\epsilon/12$, so that $s \leq 2304/\epsilon^2$ and $\|J - D\|_{\infty\mapsto 1} \leq \|J\|_F/3$. By Lemma 17, it follows that $|\log Z - \log Z_D| \leq \epsilon n\|J\|_F/3$ and $|\mathcal{F}^* - \mathcal{F}^*_D| \leq \epsilon n\|J\|_F/3$. Thus, in order to show that $\mathcal{F} - \mathcal{F}^* \leq \epsilon n\|J\|_F$, it suffices to show that $\log Z_D - \mathcal{F}^*_D \leq \epsilon n\|J\|_F/3$.*

In order to show this, we begin by approximating $\log Z_D$ by the solution to an optimization problem. Let $R_i$ (resp. $C_i$) denote the rows (respectively columns) corresponding to the cut matrix $D^{(i)}$. Then, it follows by definition that

$$Z_D = \sum_{x\in\{\pm 1\}^n} \exp\left(\sum_{i=1}^s r_i(x)c_i(x)d_i\right),$$

where $r_i(x) = \sum_{a\in R_i} x_a$ and $c_i(x) = \sum_{b\in C_i} x_b$. By rewriting the sum in terms of the possible values that $r_i(x)$ and $c_i(x)$ can take, we get that

$$Z_D = \sum_{r,c} \exp\left(\sum_{i=1}^s r_ic_id_i\right)\left(\sum_{x\in\{\pm 1\}^n : r(x)=r, c(x)=c} 1\right), \tag{4}$$

where $r = (r_1, \ldots, r_s)$ ranges over all elements of $[-|R_1|, |R_1|] \times \cdots \times [-|R_s|, |R_s|]$ and similarly for $c$. The following lemma shows that for estimating the contribution of the term corresponding to some vector $x$, it suffices to know the components of $x$ up to some constant precision.

**Lemma 19** *Let $J, D^1, \ldots, D^s$ be as above. Then, given real numbers $r_i, r'_i, c_i, c'_i$ for each $i \in [s]$ and some $\upsilon \in (0,1)$ such that $|r_i|, |c_i|, |r'_i|, |c'_i| \leq n$, $|r_i - r'_i| \leq \upsilon n$ and $|c_i - c'_i| \leq \upsilon n$ for all $i \in [s]$, we get that $\sum_i d_i |r'_i c'_i - r_i c_i| \leq 8\|J\|_F \upsilon n s^{1/2}$.*

**Proof** Since $|r'_i c'_i - r_i c_i| \leq |c'_i||r'_i - r_i| + |r_i||c'_i - c_i| \leq 2\upsilon n^2$, it follows by Cauchy-Schwarz that

$$\sum_i d_i |r'_i c'_i - r_i c_i| \leq \left( \sum_i d_i^2 \right)^{1/2} 2s^{1/2} \upsilon n^2 \leq 8\|J\|_F \upsilon n s^{1/2}.$$

■

The previous lemma motivates grouping together configurations $x$ with similar values of $r_i(x), c_i(x)$. Accordingly, for any $r \in [-|R_1|, |R_1|] \times \cdots \times [-|R_s|, |R_s|]$, $c \in [-|C_1|, |C_1|] \times \cdots \times [-|C_s|, |C_s|]$ and $\upsilon > 0$, let

$$X_{r,c,\upsilon} := \{x \in \{\pm 1\}^n : |r_i(x) - r_i| \leq \upsilon n, |c_i(x) - c_i| \leq \upsilon n \text{ for all } i \in [s]\}.$$

Let $I_\upsilon := \{\pm \upsilon n, \pm 3\upsilon n, \pm 5\upsilon n, \ldots, \pm \ell \upsilon n\}$, where $\ell$ is the smallest odd integer satisfying $|\ell \upsilon n - n| \leq \upsilon n$, so $|I_\upsilon| \leq 1/\upsilon + 1$. Let

$$Z^*_{D,\upsilon} := \max_{r,c \in I_\upsilon^s} \exp\left( \sum_{i=1}^s r_i c_i d_i + \log |X_{r,c,\upsilon}| \right). \tag{5}$$

Then, it follows immediately from Lemma 19 that

$$Z^*_{D,\upsilon} \exp\left(-8\|J\|_F \upsilon n s^{1/2}\right) \leq Z_D \leq \sum_{r,c \in I_\upsilon^s} |X_{r,c,\upsilon}| \exp\left( \sum_{i=1}^s r_i c_i d_i \right) \exp\left(8\|J\|_F \upsilon n s^{1/2}\right).$$

In particular, since the outer sum is over $|I_\upsilon|^{2s}$ terms, it follows that

$$\log Z^*_{D,\upsilon} \geq \log Z_D - 8\|J\|_F \upsilon n s^{1/2} - 2s \log |I_\upsilon| \geq \log Z_D - 8\|J\|_F \upsilon n s^{1/2} - 2s \log(1/\upsilon + 1) \tag{6}$$

and

$$\log Z^*_{D,\upsilon} \leq \log Z_D + 8\|J\|_F \upsilon n s^{1/2}. \tag{7}$$

We can now prove Theorem 1: all we need to do is give an upper bound on $\mathcal{F} - \mathcal{F}^*$.

**Lemma 20** *For any $\epsilon > 0$,*

$$\mathcal{F} - \mathcal{F}^* \leq \epsilon n \|J\|_F + 10^5 \log(e + 1/\epsilon)/\epsilon^2.$$

**Proof** Let $\gamma = \epsilon/48s^{1/2}$. Let $r = (r_1, \ldots, r_s), c = (c_1, \ldots, c_s) \in I_\gamma^s$ be such that $\log Z^*_{D,\gamma} = \sum_{i=1}^n r_i c_i d_i + \log |X_{r,c,\gamma}|$. Define

$$\bar{y}_j := \frac{1}{|X_{r,c,\gamma}|} \sum_{x \in X_{r,c,\gamma}} x_j,$$

and let $Y := (Y_1, \ldots, Y_n)$ be a random vector distributed uniformly in $X_{r,c,\gamma}$. Then by the chain rule for entropy,

$$\log |X_{r,c,\gamma}| = H(Y) \le \sum_{j=1}^{n} H(Y_j) = \sum_{j=1}^{n} H\left(\frac{1+\overline{y}_j}{2}\right).$$

Using this, we have

$$\begin{aligned}
\log Z^*_{D,\gamma} &= \sum_{i=1}^{n} r_i c_i d_i + \log |X_{r,c,\gamma}| \le \sum_{i=1}^{n} r_i c_i d_i + \sum_{j=1}^{n} H\left(\frac{1+\overline{y}_j}{2}\right) \\
&\le \left\{ \sum_{i=1}^{n} r_i(\bar{y}) c_i(\bar{y}) d_i + 8\|J\|_F \gamma n s^{1/2} \right\} + \sum_{j=1}^{n} H\left(\frac{1+\overline{y}_j}{2}\right) \\
&= \left\{ \sum_{i=1}^{n} r_i(\bar{y}) c_i(\bar{y}) d_i + \sum_{j=1}^{n} H\left(\frac{1+\overline{y}_j}{2}\right) \right\} + 8\|J\|_F \gamma n s^{1/2} \\
&\le \mathcal{F}^*_D + 8\|J\|_F \gamma n s^{1/2},
\end{aligned}$$

where the second line follows from $\overline{y}$ lying in the convex hull of $X_{r,c,\gamma}$ and Lemma 19, and the last line follows from the definition of $\mathcal{F}^*_D$. Thus, we get

$$\begin{aligned}
\mathcal{F}^*_D \ge \log Z^*_{D,\gamma} - 8\|J\|_F \gamma n s^{1/2} &\ge \log Z_D - 16\|J\|_F \gamma n s^{1/2} - 2s \log(1/\gamma + 1) \\
&\ge \log Z_D - \frac{\epsilon n \|J\|_F}{3} - 10^5 \log\left(\frac{1}{\epsilon} + e\right) \frac{1}{\epsilon^2},
\end{aligned}$$

where in the second inequality, we have used Eq. (6), and in the last line, we have used the values of $\gamma$ and $s$. Now, Remark 18 gives $\mathcal{F} - \mathcal{F}^* \le \epsilon n \|J\|_F + 10^5 \log(e + 1/\epsilon)/\epsilon^2$ as desired. ■

Finally, we use this bound to prove Theorem 1.

**Proof** [Proof of Theorem 1] Fix $M > e$ a constant to be optimized later. Observe that since $\mathbb{E}_\mu[\sum_{i,j} J_{i,j} X_i X_j] \le n\|J\|_F$ by Cauchy-Schwarz, and since $\mathcal{F}^* \ge n$, we always have $\mathcal{F} - \mathcal{F}^* \le n\|J\|_F$. Therefore, if $n\|J\|_F \le M$, we see that $\mathcal{F} - \mathcal{F}^* \le n\|J\|_F \le M^{1/3}(n\|J\|_F)^{2/3}$.

Next, we analyze the case when $n\|J\|_F > M$. Taking $\epsilon = \left(\frac{M \log(n\|J\|_F + e)}{n\|J\|_F \log M}\right)^{1/3}$ in Theorem 20 gives

$$\begin{aligned}
\mathcal{F} - \mathcal{F}^* &\le \frac{M^{1/3}}{\log^{1/3} M} n^{2/3} \|J\|_F^{2/3} \log^{1/3}(n\|J\|_F + e) + 10^5 \frac{\log^{2/3} M}{M^{2/3}} \frac{\log(n\|J\|_F + e)}{\log^{2/3}(n\|J\|_F + e)} n^{2/3} \|J\|_F^{2/3} \\
&\le \left( (M/\log M)^{1/3} + 10^5 (\log M / M)^{2/3} \right) n^{2/3} \|J\|_F^{2/3} \log^{1/3}(n\|J\|_F + e).
\end{aligned}$$

Finally, taking $(M/\log M) = 10^5$, we see that $\mathcal{F} - \mathcal{F}^* \le 200 n^{2/3} \|J\|_F^{2/3} \log^{1/3}(n\|J\|_F + e)$ for all values of $n\|J\|_F$. ■

## Acknowledgments

We thank David Gamarnik for insightful comments, Andrej Risteski for helpful discussions related to his work (Risteski, 2016), Yufei Zhao for introducing us to reference (Alon et al., 2002), and an anonymous reviewer for introducing us to the reference (Eldan, 2016).

## References

Nir Ailon and Noga Alon. Hardness of fully dense problems. *Information and Computation*, 205 (8):1117–1129, 2007.

Noga Alon, Fernandez de la Vega, Ravi Kannan, and Marek Karpinski. Random sampling and approximation of MAX-CSP problems. In *STOC*, 2002.

Noga Alon, Fernandez de la Vega, Ravi Kannan, and Marek Karpinski. Random sampling and approximation of MAX-CSPs. *J. Comput. System Sci.*, 67:212–243, 2003.

James Anderson and Carsten Peterson. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019, 1987.

Anirban Basak and Sumit Mukherjee. Universality of the mean-field for the potts model. *Probability Theory and Related Fields*, 168(3-4):557–600, 2017.

Christian Borgs, Jennifer T Chayes, László Lovász, Vera T Sós, and Katalin Vesztergombi. Convergent sequences of dense graphs ii. multiway cuts and statistical physics. *Annals of Mathematics*, 176(1):151–219, 2012.

Christian Borgs, Jennifer T Chayes, Henry Cohn, and Yufei Zhao. An $l^p$ theory of sparse graph convergence ii: Ld convergence, quotients, and right convergence. *Annals of Probability*, 46, 2018.

Amir Dembo and Andrea Montanari. Ising models on locally tree-like graphs. *Ann. Appl. Probab.*, 20(2):565–592, 2010. ISSN 1050-5164. doi: 10.1214/09-AAP627. URL http://dx.doi.org/10.1214/09-AAP627.

Roland Lvovich Dobrushin. The description of a random field by means of conditional probabilities and conditions of its regularity. *Theor. Prob. Appl.*, 13:197–224, 1968.

Ronen Eldan. Gaussian-width gradient complexity, reverse log-sobolev inequalities and nonlinear large deviations. *arXiv preprint arXiv:1612.04346*, 2016.

Richard S. Ellis. *Entropy, large deviations, and statistical mechanics*. Springer, 2007.

Richard S Ellis and Charles M Newman. The statistics of curie-weiss models. *Journal of Statistical Physics*, 19(2):149–161, 1978.

Alan Frieze and Ravi Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

Shayan Oveis Gharan and Luca Trevisan. A new regularity lemma and faster approximation algorithms for low threshold rank graphs. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 303–316. Springer, 2013.

Eslie Ann Goldberg and Mark Jerrum. The complexity of ferromagnetic ising with local fields. *Comb. Probab. Comput.*, 16(1):43–61, January 2007. ISSN 0963-5483. doi: 10.1017/S096354830600767X. URL http://dx.doi.org/10.1017/S096354830600767X.

Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media, 2012.

Piotr Indyk. Sublinear time algorithms for metric space problems. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 428–434. ACM, 1999.

Sorin Istrail. Statistical mechanics, three-dimensionality and np-completeness: I. universality of intracatability for the partition function of the ising model across non-planar surfaces (extended abstract). In *STOC*, 2000.

Tommi S Jaakkola and Michael I Jordan. Improving the mean field approximation via the use of mixture distributions. In *Learning in graphical models*, pages 163–173. Springer, 1998.

Vishesh Jain, Frederic Koehler, and Elchanan Mossel. The vertex sample complexity of free energy is polynomial. In *Conference on Learning Theory*, 2018.

M. Jerrum and A. Sinclair. Polynomial-time approximation algorithms for ising model (extended abstract). In *Automata, Languages and Programming*, pages 462–475, 1990.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Jingcheng Liu, Alistair Sinclair, and Piyush Srivastava. The ising partition function: Zeros and deterministic approximation. In *FOCS*, 2017.

J. M. Mooij and H. J. Kappen. Sufficient conditions for convergence of the sum-product algorithm. *IEEE Transactions on Information Theory*, 53(12):4422–4437, Dec 2007. ISSN 0018-9448. doi: 10.1109/TIT.2007.909166.

Elchanan Mossel and Allan Sly. Exact thresholds for ising–gibbs samplers on general graphs. *The Annals of Probability*, 41(1):294–328, 2013.

Giorgio Parisi. *Statistical field theory*. New York: Addison-Wesley, 1988.

Andrej Risteski. How to calculate partition functions using convex programming hierarchies: provable bounds for variational methods. In *COLT*, 2016.

Allan Sly and Nike Sun. The computational hardness of counting in two-spin models on d-regular graphs. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 361–369. IEEE, 2012.

S. Tatikonda and M. I. Jordan. Loopy belief propagation and gibbs measures. In *Uncertainty in Artificial Intelligence (UAI), Proceedings of the Eighteenth Conference*. 2002.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

## Appendix A. Hyper-graph Statements and proofs

**Proof** [Proof of Theorem 5] The proof is exactly the same as that of Theorem 1, except that for each $d$ from 1 to $r$, we use the following generalized weak regularity lemma to decompose $J_{=d}$:

**Theorem 21 (Alon et al. (2003))** *Let $J$ be an arbitrary $k$-dimensional matrix on $X_1 \times \cdots \times X_k$, where we assume that $k \geq 1$ is fixed. Let $N := |X_1| \times \cdots \times |X_k|$ and let $\epsilon > 0$. Then, in time $2^{O(1/\epsilon^2)}O(N)$ and with probability at least $0.99$, we can find a cut decomposition of width at most $4/\epsilon^2$, error at most $\epsilon\sqrt{N}\|J\|_F$, and the following modified bound on coefficient length: $\sum_i |d_i| \leq 2\|J\|_F/\epsilon\sqrt{N}$, where $(d_i)_{i=1}^s$ are the coefficients of the cut arrays.*

We omit further details. ■

The next two theorems are the analogs of Theorem 12 for general order $r$ Markov random fields. The first theorem runs in constant time for any fixed $\epsilon$, but the constant is of the form $2^{\tilde{O}((1/\epsilon)^{2r-2})}$. In contrast, the second algorithm runs in time $O_{\epsilon,r}(n^r)$ for any fixed $r$ and $\epsilon$, with the dependence of the constant on $\epsilon$ being the much improved $2^{\tilde{O}(1/\epsilon^2)}$. In (Jain et al., 2018), we build upon the latter result to provide a constant time algorithm, with similar guarantees as Theorem 22 but requiring time $2^{\tilde{O}(1/\epsilon^2)}$, for approximating the free energy.

**Theorem 22** *Fix $r \geq 3$. Then, there exists a constant $C = C(r)$ such that for any order $r$ Markov random field $J$ with Boltzmann distribution $P$ and free energy $\mathcal{F}$, and for any $\epsilon > 0$, there is an algorithm which runs in time $2^{O(\log(1/\epsilon)/\epsilon^{2r-2})}$ and returns, with probability at least $0.99$, an implicit description of a product distribution $\mu$ and estimate to the free energy $\hat{\mathcal{F}}$ such that*

$$\mathbf{KL}(\mu||P) \leq \max_{1\leq d\leq r} \epsilon n^{d/2}\|J_{=d}\|_F + C\log(1/\epsilon)/\epsilon^{2d-2} + 0.5^{2^{1/\epsilon^{2d-2}}} n$$

*and*

$$|\mathcal{F} - \hat{\mathcal{F}}| \leq \max_{1\leq d\leq r} \epsilon n^{d/2}\|J_{=d}\|_F + C\log(1/\epsilon)/\epsilon^{2d-2} + 0.5^{2^{1/\epsilon^{2d-2}}} n.$$

In the previous theorem, it is possible to improve the dependence on $\epsilon$ at the expense of introducing a factor of $n^r$ in the running time.

**Theorem 23** *Fix $r \geq 3$. Then, there exists a constant $C = C(r)$ such that for any order $r$ Markov random field $J$ with Boltzmann distribution $P$ and free energy $\mathcal{F}$, and for any $\epsilon > 0$, there is an algorithm which runs in time $2^{O(\log(1/\epsilon)/\epsilon^2)}n^r$ and returns, with probability at least $0.99$, an implicit description of a product distribution $\mu$ and estimate to the free energy $\hat{\mathcal{F}}$ such that*

$$\mathbf{KL}(\mu||P) \leq \epsilon \max_{1\leq d\leq r} n^{d/2}\|J_{=d}\|_F + C\log(1/\epsilon)/\epsilon^2 + 0.5^{2^{1/\epsilon^2}} n$$

*and*

$$|\mathcal{F} - \hat{\mathcal{F}}| \leq \epsilon \max_{1\leq d\leq r} n^{d/2}\|J_{=d}\|_F + C\log(1/\epsilon)/\epsilon^2 + 0.5^{2^{1/\epsilon^2}} n.$$

## Appendix B. An almost matching lower bound for a large class of variational methods

**Proof** [Proof of Theorem 2] Let $(\mathcal{Q}_n)_{n=0}^{\infty}$ be a sequence of families of probability distributions as in the theorem statement. By assumption, there exist $k$ and $J$ such that $\mathcal{Q}_k$ does not contain the probability distribution $P_J$ corresponding to the Ising model $J$ on $k$ nodes. We denote by $Q_J$ the probability distribution in $\mathcal{Q}_k$ which is closest to $P_J$. In particular, by the closure under products assumption, we have that $Q_J^{\otimes m} \in \mathcal{Q}_{mk}$ for all integers $m \geq 1$.

Consider the Ising model on $n := mk$ nodes whose matrix of interaction strengths $J'_n$ is the block diagonal matrix consisting of $m$ copies of $J$. Combinatorially, we can view $J'_n$ as $m$ vertex disjoint copies of $J$. We claim that $Q_J^{\otimes m}$ is the closest distribution in $\mathcal{Q}_{mk}$ to the Ising model $J'$. Suppose on the contrary that there is some other distribution $Q_{J'} \in \mathcal{Q}_{mk}$ which is strictly closer to $P_{J'}$ than $Q_J^{\otimes m}$. Then, the chain rule for KL divergence immediately implies that there exists some distribution $\tilde{Q}_J$ on $\{\pm 1\}^k$, obtained by conditioning $Q_{J'}$ on $k(m-1)$ variables, which is strictly closer to $P_J$ than $Q_J$. Since $\tilde{Q}_J \in \mathcal{Q}_k$ by assumption, and since $Q_J$ is the closest distribution to $P_J$ in this class, this gives a contradiction.

Therefore, we see that

$$\inf_{Q \in \mathcal{Q}_n} \mathbf{KL}(Q||P_{J'}) \geq m\, \mathbf{KL}(Q_J||P_J) = \Theta(n).$$

Furthermore, $\|J\|_F = \Theta(\sqrt{n})$ so that $n^{2/3}\|J\|_F^{2/3} = \Theta(n)$. Hence, we see that the variational method corresponding to $(\mathcal{Q}_n)_{n=0}^{\infty}$ must make an error of size $\Omega(n^{2/3}\|J\|_F^{2/3})$. ■

## Appendix C. The high-temperature regime

In this section, we show that in the high-temperature regime where Markov chain methods are guaranteed to mix quickly, the variational free energy functional is convex and furthermore, a simple message passing algorithm solves the corresponding optimization problem quickly.

**Lemma 24** *For $H(p) := H(Ber(p))$ and for any $p \in [0,1]$, we have*

$$H''(p) \leq -4.$$

**Proof** By definition,

$$H(p) = -p\log p - (1-p)\log(1-p).$$

Therefore,

$$H'(p) = -\log p - 1 + \log(1-p) + 1 = -\log p + \log(1-p),$$

and

$$H''(p) = -\frac{1}{p} - \frac{1}{1-p} \leq -4.$$

■

**Proof** [Proof of Theorem 6] Recall that $J$ is symmetric and has diagonal entries 0. Therefore, the assumption $\sum_j 2|J_{i,j}| \leq 1$ for all $i$, along with Gershgorin's disk theorem, shows that all the eigenvalues of $J$ lie in $[-1/2, 1/2]$. Observe that the Hessian of the corresponding quadratic form

is $2J$. Combining this with the strong concavity of entropy (Lemma 24) and the chain rule, which gives $\frac{d^2}{dx^2}H((1+x)/2) \le -1$, proves the concavity claim.

The runtime complexity follows from standard algorithms from convex optimization, e.g. standard guarantees for the ellipsoid method (Grötschel et al., 2012). ■

**Proof** [Proof of Theorem 8] Since $\tanh$ is 1-Lipschitz, we have for any $x^1, x^2 \in [-1,1]^n$ that

$$\|\tanh^{\otimes n}(2Jx^1 + h) - \tanh^{\otimes n}(2Jx^2 + h)\|_\infty \le 2\|Jx^1 - Jx^2\|_\infty \le (1-\eta)\|x^1 - x^2\|_\infty.$$

Since the optimum $x^*$ is a fixed point of the mean field equations, the above inequality shows that

$$\|\tanh^{\otimes n}(2Jx_{n+1} + h) - x^*\|_\infty \le (1-\eta)\|\tanh^{\otimes n}(2Jx_n + h) - x^*\|_\infty,$$

and iterating this inequality gives the desired conclusion. ■

## Appendix D. Computing the mean-field approximation in ferromagnetic models

**Proof** [Proof of Theorem 10] Consider the $m$-blow up of the Ising model, denoted by $J_m$, defined as follows: replace each vertex $i$ by $m$ vertices $(i,1), \dots, (i,m)$, add an edge of weight $J_{i,j}/m$ between vertices $(i,k)$ and $(j,\ell)$ for all $1 \le k, \ell \le m$, and assign a uniform external field $h$ at each vertex $(i,k)$.

Given a spin vector $X$ sampled from the Boltzmann distribution of $J_m$, define $Y_i \in [-m,m]$ to be the net spin of the vertices $(i,1), \dots, (i,m)$. Let $N_y$ denote the number of spin vectors which correspond to the net spin vector $y$ via the correspondence above. Then, we see that

$$\begin{aligned}\Pr(Y = y) &= \frac{1}{Z}\exp\left(\sum_{i,j}\frac{J_{ij}y_iy_j}{m} + h\sum_i y_i + \log N_y\right)\\ &= \frac{1}{Z}\exp\left(m\sum_{i,j}J_{i,j}(y_i/m)(y_j/m) + mh\sum_i(y_i/m) + m\sum_i H\left(\frac{1+y_i/m}{2}\right) \pm O(n\log m)\right).\end{aligned}$$

Let $Y_\epsilon$ be the set of $y$ such that

$$\sum_{i,j}J_{ij}(y_i/m)(y_j/m) + h\sum_i(y_i/m) + \sum_i H\left(\frac{1+y_i/m}{2}\right) < \mathcal{F}^* - \epsilon,$$

where $\mathcal{F}^*$ is the variational free energy of the original Ising model $J$. Note that $Z \ge e^{\mathcal{F}^* - O(n\log m)}$, as is readily seen by considering the net spin vector $y^*$ given by $y_i^* = mx_i^*$, where $x^* = (x_1^*, \dots, x_m^*)$ is the optimizer of the optimization problem defining $\mathcal{F}^*$. Then, the above inequality shows that for each $y \in Y_\epsilon$,

$$\Pr(Y = y) \le e^{-m\epsilon \pm O(n\log m)}.$$

Since $|Y_\epsilon| \le m^n$, the union bound shows that

$$\Pr(Y \in Y_\epsilon) \le e^{-m\epsilon \pm O(n\log m)} \le \frac{1}{3},$$

provided we take $m = \Omega(n \log(n)/\epsilon)$.

The preceding analysis shows the following: if we use the algorithm of (Jerrum and Sinclair, 1990) to draw $O(\log(1/\delta))$ independent (approximate) samples $X$ from the Boltzmann distribution of $J_m$, and use these (approximate) samples to obtain normalized net spin vectors $Y/m \in \{\pm 1\}^n$, then with probability $1 - \delta$, at least one of the sampled $Y/m$ solves the optimization problem defining $\mathcal{F}^*$ up to $\epsilon$-additive error in the objective. ■

**Remark 25** *It is known by the result of (Goldberg and Jerrum, 2007) that approximate sampling becomes #BIS-hard for ferromagnetic Ising models if we allow different (inconsistent) external fields for each node. Thus, our algorithm does not extend to this setting.*

## Appendix E. NP-hardness: Proof of Theorem 11

Our proof is an easy consequence of hardness of approximation results for dense CSPs. Specifically, we rely on a hardness result for *fully dense* MAX-r-LIN-2. In this problem, we are given $n$ free variables $x_1, \ldots, x_n$ to be assigned values in $\mathbb{F}_2^n$. Moreover, for each of the $\binom{n}{r}$ subsets $S$ of $[n]$ of size $r$, we are given a constraint $\sum x_S \equiv y_S \mod 2$, for $y_S$ fixed to be either 0 or 1. The goal is to find the maximum number of constraints which can be satisfied simultaneously by a single assignment of $x_1, \ldots, x_n$. For reasons of convenience, the objective value is defined to be $(1/2)(\#\text{ of satisfied constraints}) - (1/2)(\#\text{ of violated constraints})$.

**Theorem 26 (Ailon and Alon (2007))** *For $r \geq 2$ and any $\epsilon > 0$, it is NP-hard to approximate fully dense MAX-r-LIN-2 within an additive error of $n^{r-\epsilon}$.*

**Proof** [Proof of Theorem 11] We illustrate the reduction to our problem in the case $r = 2$. Given an instance of fully dense MAX-r-LIN-2 with constraints corresponding to fixed $(y_S)_{|S|=2}$, we consider the Ising model with matrix of interaction strengths $J$, where $J_{ij} = 1/2 - y_{\{i,j\}}$. It is readily seen that for any distribution $\mu$ on $\{\pm 1\}^n$,

$$\sum_{i,j} J_{ij} \mathbb{E}_\mu[X_i X_j] \leq \text{MAX-r-LIN-2}(y).$$

On the other hand, denoting by $x = (x_1, \ldots, x_n)$ the optimal assignment of the variables for MAX-r-LIN-2$(y)$, it is immediate that the deterministic distribution $\nu$ concentrated on $X_i = (-1)^{x_i}$ satisfies

$$\sum_{i,j} J_{ij} \mathbb{E}_\nu[X_i X_j] = \text{MAX-r-LIN-2}(y).$$

Thus, it follows that

$$\mathcal{F} = \max_\mu \left[ \sum_{i,j} J_{ij} \mathbb{E}_\mu[X_i X_j] + H(\mu) \right] = \text{MAX-r-LIN-2}(y) \pm n.$$

Finally, observe that $n\|J\|_F = \Theta(n^2)$. Therefore, approximating $\mathcal{F}$ within additive error $(n\|J\|_F)^{1-\delta}$ gives an $n^{2(1-\delta)}$ additive approximation to MAX-r-LIN-2$(y)$. ■

## Appendix F. A general algorithm for solving the variational problem

By Theorem 1, we have an upper bound on $\mathbf{KL}(\nu||P)$ (which is almost tight in the worst case) for the optimal product distribution $\nu$. Unfortunately, Theorem 11 shows that it is not always possible to efficiently find a product distribution which is as close to $P$ as $\nu$ is. In this section, we describe a provable algorithm which does essentially as well as possible without violating Theorem 11, with the additional benefit that it runs in *constant time* (independent of the size of the graph). Here we will use $\tilde{O}$ notation to hide logarithmic factors independent of $n$.

**Remark 27** *In order to provide a constant time guarantee on problems with unbounded input size, we will work under the usual assumptions on the computational model for sub-linear algorithms (as in, e.g., Alon et al. (2002); Frieze and Kannan (1999); Indyk (1999)). Thus, we can probe matrix entry $A(i,j)$ in $O(1)$ time. Note also that by the standard Chernoff bounds, it follows that for any set of vertices $V$ for which we can test membership in $O(1)$ time, we can also estimate $|V|/n$ to additive error $\epsilon$ w.h.p. in constant time using $\tilde{O}(1/\epsilon^2)$ samples. This approximation will always suffice for us and so, for the sake of exposition, we will henceforth ignore this technical detail and just assume that we have access to $|V|/n$ (as in, e.g., Frieze and Kannan (1999)).*

**Proof** [Proof of Theorem 12] Observe that it suffices to return the description of an $x \in [-1,1]^n$ such that $\mathcal{F}(x) := \sum_{i,j} J_{i,j}x_ix_j + \sum_i H((1+x_i)/2)$ satisfies

$$\mathcal{F}^* \leq \mathcal{F}(x) + \frac{2\epsilon n}{3}\|J\|_F + 0.5^{2^{1/\epsilon^2}} n.$$

Indeed, since

$$\mathcal{F} - \mathcal{F}^* \leq \frac{\epsilon n}{3}\|J\|_F + 10^6 \log(e+1/\epsilon)/\epsilon^2$$

by Lemma 20, the product distribution $\mu$ for which the $i^{th}$ coordinate has expected value $x_i$ will then satisfy the conclusions of the theorem.

Our strategy for finding such an $x$ will be to find an approximate maximizer of the problem defining $\mathcal{F}^*_D$, where $D$ is a sum of a small number of cut matrices which is close to $J$ in the $\|\cdot\|_{\infty\mapsto 1}$ norm. Specifically, we use the following algorithmic weak regularity lemma of Frieze and Kannan:

**Theorem 28 (Frieze and Kannan (1999))** *Let $J$ be an arbitrary real matrix, and let $\epsilon, \delta > 0$. Then, in time $2^{\tilde{O}(1/\epsilon^2)}/\delta^2$, we can, with probability at least $1-\delta$, find a cut decomposition of width $O(\epsilon^{-2})$, coefficient length at most $\sqrt{27}\|J\|_F/\sqrt{mn}$ and error at most $4\epsilon\sqrt{mn}\|J\|_F$.*

As in Remark 18, we will take $D := D^{(1)} + \cdots + D^{(s)}$, where the $D^{(i)}$ are obtained by applying Theorem 28 to $J$ with parameter $\epsilon/12$. In particular, Lemma 17 shows that $|\mathcal{F}^* - \mathcal{F}^*_D| \leq \frac{\epsilon n}{3}\|J\|_F$, so that any $x$ which satisfies $\mathcal{F}^*_D \leq \mathcal{F}(x) + \frac{\epsilon n}{3}\|J\|_F + 0.5^{2^{1/\epsilon^2}} n$ also satisfies the desired upper bound on $\mathcal{F}^* - \mathcal{F}(x)$.

For $r, c \in I_\gamma^s$, where $I_\gamma$ is as in the proof of Section 3, consider the following max-entropy program $\mathcal{C}_{r,c,\gamma}$:

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} H\left(\frac{1+x_i}{2}\right) \\ s.t. & \\ \forall i \in [n]: \quad & -1 \le x_i \le 1 \\ \forall t \in [s]: \quad & r_t - \gamma n \le \sum_{i \in R_t} x_i \le r_t + \gamma n \\ \forall t \in [s]: \quad & c_t - \gamma n \le \sum_{i \in C_t} x_i \le c_t + \gamma n \end{aligned}$$

Then, Lemma 19 shows that

$$\overline{\mathcal{F}}_D := \max_{r,c \in I_v^s} \sum_{i=1}^{s} r_i c_i d_i + \mathcal{C}_{r,c,\gamma}$$

satisfies $|\overline{\mathcal{F}}_D - \mathcal{F}_D^*| \le \frac{\epsilon n}{3}\|J\|_F$, provided we take $\gamma \le s^{-1/2}/24$. Let $(\overline{r}, \overline{c})$ denote the values of $(r, c)$ attaining $\overline{\mathcal{F}}_D$. It follows that if we can return an $x \in [-1,1]^n$ such that $x$ is feasible for $\mathcal{C}_{\overline{r},\overline{c},\gamma}$, and

$$\mathcal{C}_{\overline{r},\overline{c},\gamma} \le \sum_i H\left(\frac{1+x_i}{2}\right) + 2^{-2^{1/\epsilon^2}} n,$$

then we would be done.

Since we want our algorithm to run in constant time, we rewrite this convex program in an equivalent way with only a constant number of variables and constraints. Let $(V_a)_{a=1}^A$ denote the common refinement of $\{R_i, C_i\}_{i=1}^s$. In particular, note that $A \le 2^{2s}$. Let $nv_a$ denote the number of vertices in $V_a$, and recall (Remark 27) that we can estimate $v_a$ to high precision in constant time by sampling. Then, by the concavity of entropy, it is readily seen that for the maximum entropy program $\mathcal{H}_{r,c,\gamma}$:

$$\begin{aligned} \max \quad & \sum_a v_a H\left(\frac{1+z_a/v_a}{2}\right) & & & \\ s.t. \quad & -v_a \le z_a & \le v_a & & \forall 1 \le a \le A \\ & r_t/n \le \sum_{a: V_a \subset R_t} z_a & \le r_t/n + \gamma & & \forall 1 \le t \le s \\ & c_t/n \le \sum_{a: V_a \subset C_t} z_a & \le c_t/n + \gamma & & \forall 1 \le t \le s, \end{aligned}$$

we have $n\mathcal{H}_{r,c,\gamma} = \mathcal{C}_{r,c,\gamma}$. Finally, each of these convex programs can be solved approximately using standard guarantees for the ellipsoid method (Grötschel et al., 2012) – in time $2^{O(1/\epsilon^2)}$, the returned $z_a$ is optimal up to an additive error of $2^{-2^{1/\epsilon^2}}$, and this completes the proof. ■

**Proof** [Proofs of Theorem 22 and Theorem 23] The proofs of these theorems are essentially the same as the proof of Theorem 12, and therefore we will omit details. We only note that for Theorem 22, we apply the following algorithmic regularity lemma of Frieze and Kannan generalizing Theorem 28:

**Theorem 29 (Frieze and Kannan (1999))** *Suppose $J$ is an arbitrary $k$-dimensional matrix on $X_1 \times \cdots \times X_k$, where we assume that $k \geq 3$ is fixed. Let $N := |X_1| \times \cdots \times |X_k|$ and let $\epsilon, \delta \in (0, 1]$. Then, in time $O(k^{O(1)}\epsilon^{-O(\log_2 k)}2^{\tilde{O}(1/\epsilon^2)}\delta^{-2})$, we can, with probability at least $1 - \delta$, find a cut decomposition of width $O(\epsilon^{2-2k})$, coefficient length at most $\sqrt{27}^k\|J\|_F/\sqrt{N}$ and error at most $\epsilon 2^k\sqrt{N}\|J\|_F$.*

For Theorem 23, we instead use Theorem 21. ■

## Appendix G. An improved result for graphs of low threshold rank

As in (Risteski, 2016), we are also able to handle graphs of low threshold rank using our methods; we refer to that paper for further discussion of low threshold rank graphs and their importance. The key additional ingredient which enables us to do so is the algorithmic regularity lemma of Gharan and Trevisan. We start with some preliminary definitions.

**Definition 30** *Let $J$ be the matrix of interaction strengths of an Ising model and define the* degree *of a vertex $u$ to be*

$$d(u) = \sum_v |J_{uv}|.$$

*Let $D = diag(d(u))$ be the matrix of degrees, then the* normalized adjacency matrix *$J_D$ is given by*

$$J_D := D^{-1/2}JD^{-1/2}$$

*Note that the eigenvalues of $J_D$ lie in the interval $[-1, 1]$.*

**Definition 31** *The* $\delta$-sum-of squares threshold rank *of $J$ is defined to be $t_\delta(J_D) := \sum_{i:|\lambda_i|>\delta} \lambda_i^2$, where $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of $J_D$.*

We can now state the algorithmic regularity lemma of Gharan and Trevisan.

**Theorem 32 (Gharan and Trevisan (2013))** *Let $J$ be the matrix of interaction strengths of an Ising model, let $\epsilon > 0$ and let $t := t_{\epsilon/2}(J_D)$. There exists a cut decomposition of $J$, $D = D^{(1)} + \cdots + D^{(s)}$, such that $s \leq 16t/\epsilon^2$,*

$$\|J - D\|_{\infty\mapsto 1} \leq 4\epsilon\|J\|_1$$

*and $|d_i| \leq \sqrt{t}/m$. Furthermore this decomposition can be computed in $poly(n, t, 1/\epsilon)$ time.*

Finally, we state and prove the main result of this section.

**Theorem 33** *Fix $\epsilon > 0$ and let $t = t_{\epsilon/2}(J_D)$ as in Theorem 32, then*

$$\mathcal{F} - \mathcal{F}^* \leq 3\epsilon\|J\|_1 + \frac{32t}{\epsilon^2}\log\left(\frac{2\sqrt{t}ns}{\epsilon\|J\|_1} + 1\right).$$

**Remark 34** *As in Theorems 12, 22, and 23, it is easily seen that the above result has a natural algorithmic counterpart. Unlike Theorems 12 and 22, this algorithm is not a constant time algorithm as Theorem 32 is itself polynomial time in the size of the graph.*

**Proof** We mimic the proof of Theorem 1. Apply Theorem 32 to get a matrix $D = D^{(1)} + \cdots + D^{(s)}$ and let $\mathcal{F}_D$ and ${\mathcal{F}^*}_D$ denote the free energy and variational energy of the Ising model with interaction matrix $D$; by Lemma 17 we know

$$|\mathcal{F}_D - \mathcal{F}| \leq \epsilon \|J\|_1, \|\mathcal{F}_D^* - \mathcal{F}| \leq \epsilon \|J\|_1.$$

Letting $Z_D$ denote the partition function of this Ising model, we see

$$Z_D = \sum_{r,c} \exp\left(\sum_{i=1}^{s} r_i c_i d_i\right) \left(\sum_{x \in \{\pm 1\}^n : r(x)=r, c(x)=c} 1\right),$$

where $r = (r_1, \ldots, r_s)$ ranges over all elements of $[-|R_1|, |R_1|] \times \cdots \times [-|R_s|, |R_s|]$ and similarly for $c$. Applying the argument from Lemma 19 now gives

**Lemma 35** *Let $J, D^1, \ldots, D^s$ be as above. Then, given real numbers $r_i, r_i', c_i, c_i'$ for each $i \in [s]$ and some $\upsilon \in (0,1)$ such that $r_i, c_i, r_i', c_i' \leq n$, $|r_i - r_i'| \leq \upsilon n$ and $|c_i - c_i'| \leq \upsilon n$ for all $i \in [s]$, we get that $\sum_i d_i |r_i' c_i' - r_i c_i| \leq 2\sqrt{t}\upsilon n s$.*

As before we use this lemma to group terms. Accordingly, for any $r \in [-|R_1|, |R_1|] \times \cdots \times [-|R_s|, |R_s|]$, $c \in [-|C_1|, |C_1|] \times \cdots \times [-|C_s|, |C_s|]$ and $\upsilon > 0$, let

$$X_{r,c,\upsilon} := \{x \in \{\pm 1\}^n : |r_i(x) - r_i| \leq \upsilon n, |c_i(x) - c_i| \leq \upsilon n \text{ for all } i \in [s]\}.$$

Let $I_\upsilon := \{\pm \upsilon n, \pm 3\upsilon n, \pm 5\upsilon n, \ldots, \pm \ell \upsilon n\}$, where $\ell$ is the smallest odd integer satisfying $|\ell \upsilon n - n| \leq \upsilon n$, so $|I_\upsilon| \leq 1/\upsilon + 1$. Let

$$Z^*_{D,\upsilon,\alpha} := \max_{r,c \in I_\upsilon^s} \exp\left(\sum_{i=1}^{s} r_i c_i d_i + \log |X_{r,c,\alpha\upsilon}|\right).$$

Then by following the argument from Theorem 1 we find

$$\log Z^*_{D,\upsilon,1} \geq \log Z_D - 2\sqrt{t}\upsilon n s - 2s \log |I_\upsilon| \geq \log Z_D - 2\sqrt{t}\upsilon n s - 2s \log(1/\upsilon + 1) \quad (8)$$

Finally, the argument from Lemma 20 now gives

$$\log Z^*_{D,\gamma,1} \leq \mathcal{F}_D^* + 2\sqrt{t}\upsilon n s$$

and so, letting $\upsilon = \frac{\epsilon \|J\|_1}{2\sqrt{t} n s}$ we find

$$\begin{aligned} \mathcal{F}_D^* &\geq \log Z_D - 2\sqrt{t}\upsilon n s - 2s \log(1/\upsilon + 1) \\ &\geq \log Z_D - \epsilon \|J\|_1 - \frac{32t}{\epsilon^2} \log\left(\frac{2\sqrt{t} n s}{\epsilon \|J\|_1} + 1\right) \end{aligned}$$

and finally

$$\mathcal{F} - \mathcal{F}^* \leq 3\epsilon \|J\|_1 + \frac{32t}{\epsilon^2} \log\left(\frac{2\sqrt{t} n s}{\epsilon \|J\|_1} + 1\right),$$

which completes the proof. ■