JMLR: Workshop and Conference Proceedings vol 49:1–15, 2016

How to calculate partition functions using convex programming hierarchies: provable bounds for variational methods

Andrej Risteski[*] RISTESKI@CS.PRINCETON.EDU
Princeton University,
35 Olden Street, Princeton, NJ, 08540

Abstract

We consider the problem of approximating partition functions for Ising models. We make use of recent tools in combinatorial optimization: the Sherali-Adams and Lasserre convex programming hierarchies, in combination with variational methods to get algorithms for calculating partition functions in these families. These techniques give new, non-trivial approximation guarantees for the partition function *beyond the regime of correlation decay*. They also *generalize* some classical results from statistical physics about the Curie-Weiss ferromagnetic Ising model, as well as *provide a partition function counterpart* of classical results about max-cut on dense graphs (Arora et al., 1995). With this, we connect techniques from two apparently disparate research areas – optimization and counting/partition function approximations. (i.e. #-P type of problems).

Furthermore, we design to the best of our knowledge the first *provable, convex* variational methods. Though in the literature there are a host of convex versions of variational methods (Wainwright et al.; 2005; Heskes, 2006; Meshi et al., 2009), they come with no guarantees (apart from some extremely special cases, like e.g. the graph has a single cycle (Weiss, 2000)). We consider dense and low threshold rank graphs, and interestingly, the reason our approach works on these types of graphs is because local correlations *propagate* to global correlations – completely the opposite of algorithms based on *correlation decay*. In the process we design novel *entropy approximations* based on the low-order moments of a distribution.

Our proof techniques are very simple and generic, and likely to be applicable to many other settings other than Ising models.

Keywords: Ising models; threshold rank; partition function; Lasserre hierarchy; Sherali-Adams hierarchy; correlation decay; variational methods; variational inference;

1. Introduction

Calculating partition functions is a common task in machine learning: for a distribution p over a domain $\mathcal{D}$, specified up to normalization i.e. $p(\mathbf{x}) \propto f(\mathbf{x}), \mathbf{x} \in \mathcal{D}$ for some explicit function $f(\mathbf{x})$, we want to calculate the *partition function* (i.e. the normalization constant) $\sum_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x})$.[1] This task arises naturally in almost any problem involving learning, performing inference (i.e. calculating marginals) over graphical models, or estimating posterior distributions in latent variable models.

Broadly, two approaches are used for calculating partition functions: one is based on using Markov Chains to sample from the distribution p; the other is variational methods, which involve

[*] This work was supported in part by NSF grants CCF-0832797, CCF-1117309, CCF-1302518, DMS-1317308, Sanjeev Arora's Simons Investigator Award, and a Simons Collaboration Grant.

[1] $\mathcal{D}$ can also be continuous of course, in which case the sum becomes an integral, though in this paper we only will be concerned with discrete domains.

characterizing the partition function as the solution of a certain (intractable) optimization problem over the polytope of valid distributions over $\mathcal{D}$. In theory, the former are much better studied, the crowning achievements of which are probably (Jerrum et al., 2004) and (Jerrum and Sinclair, 1993), who proved certain Markov Chains mix rapidly in the case of permanent with non-negative entries and the ferromagnetic Ising model.

In practice however, variational methods are quite popular (Wainwright and Jordan, 2008; Blei et al., 2003; 2016). There are various reasons for this, the main being that they can be quite a bit faster than Markov Chain methods [2] and they tend to be easier to parallelize. With the exception of belief propagation (which can be viewed as a particular way to solve a certain *non-convex* relaxation of the optimization problem for calculating the partition function (Yedidia et al., 2003)) there is essentially no theoretical understanding. Additionally, the guarantees for belief propagation usually apply only in the regime of decay of correlations and locally tree-like graphs.

The contributions of our paper are two-fold.

First, we bring to bear recent tools in combinatorial optimization: the Sherali-Adams and Lasserre convex programming hierarchies, in combination with variational methods to get algorithms for calculating partition functions of Ising models. These techniques give new, non-trivial approximation guarantees for the partition function *beyond the regime of correlation decay*. They also *generalize* some classical results from statistical physics about the Curie-Weiss ferromagnetic Ising model, as well as *provide a partition function counterpart* of classical results about max-cut on dense graphs (Arora et al., 1995). With this, we connect techniques from two apparently disparate research areas – optimization and counting/partition function approximations. (i.e. #-P type of problems).

Second, we design to the best of our knowledge the first *provable, convex* variational methods. Though in the literature there are a myriad of convex versions of variational methods (Wainwright et al.; 2005; Heskes, 2006; Meshi et al., 2009), they come with no guarantees at all (except in some extremely special cases, like e.g. the graph has a single cycle (Weiss, 2000)). Our methods tackle dense and low threshold rank graphs, and interestingly, the reason our approach works on these types of graphs is because local correlations *propagate* to global correlations – which is completely the opposite of algorithms based on *correlation decay*. In the process we design novel *entropy approximations* based on the low-order moments of a distribution.

Our proof methods are extremely simple and generic and we believe they can be applied to many other families of partition functions.

Finally, one more important reason to study variational methods (albeit more theoretical in nature) is derandomization, since variational methods are usually deterministic. The gap between the state of the art in partition function calculation with and without randomization is huge. For instance, for the case of calculating permanents of non-negative matrices the algorithm due to (Jerrum et al., 2004) gets a factor $1+\epsilon$ approximation in time poly$(n, 1/\epsilon)$ with high probability (i.e. it's an FPRAS). In contrast, the best deterministic algorithm due to (Gurvits and Samorodnitsky, 2014) achieves only a factor 2^n approximation in time poly(n). (To make the situation even more drastic, the approach in (Gurvits and Samorodnitsky, 2014) can at best lead to a factor $\sqrt{2}^n$ approximation (Wigderson).)

[2] Markov Chain methods always produce the right answer in the end, but might take longer to converge; variational methods are based on solving an optimization problem, for which it is potentially possible to get stuck in a local optimum, but generally convergence is faster

2. Overview of results

We focus on *dense* Ising models first: an Ising model $p(\mathbf{x}) \propto \exp\left(\sum_{i,j} J_{i,j}\mathbf{x}_i\mathbf{x}_j\right), \mathbf{x} \in \{-1,1\}^n$ is Δ-*dense* if it satisfies $\Delta|J_{i,j}| \leq \frac{J_T}{n^2}, \forall i,j \in [n]$, where $J_T = \sum_{i,j}|J_{i,j}|$.

This is a natural generalization of the typical way to define density for combinatorial optimization problems (see e.g. (Yoshida and Zhou, 2014)). To see this consider a graph $G = (V,E)$ with $|E| = cn^2$. For optimization problems like max-cut or more generally CSPs, we care about objectives that look like

$$\mathbb{E}_{e\in E} f(e) = \sum_{e\in E} \frac{1}{|E|} f(e)$$

for some function f. Hence, the "weight" in front of each pair (i,j) in the objective is 0 if there is no edge or $\frac{1}{|E|}$. This corresponds to $\Delta = \frac{1}{c}$ in our definition. For partition function problems, however, scale matters (i.e. we cannot assume $\sum_{i,j} J_{i,j} = 1$), so the above generalization appears very organic.

Theorem 1 *For Δ-dense Ising models, there is an algorithm based on Sherali-Adams hierarchies which achieves an additive approximation of ϵJ_T to $\log \mathcal{Z}$, where $\mathcal{Z} = \sum_{\mathbf{x}\in\{-1,1\}^n} \exp\left(\sum_{i,j} J_{i,j}\mathbf{x}_i\mathbf{x}_j\right)$ and runs in time $n^{O\left(\frac{1}{\Delta\epsilon^2}\right)}$.*

Our second contribution are analogous claims for Ising models whose potentials look like low rank matrices. (More precisely, adjacency matrices of *low threshold rank* graphs, a concept introduced by (Arora et al., 2010) in the context of their algorithm for Unique Games.)

Concretely, an Ising model $p(\mathbf{x}) \propto \exp\left(\sum_{i,j} J_{i,j}\mathbf{x}_i\mathbf{x}_j\right), \mathbf{x} \in \{-1,1\}^n$ is *regular* if $\sum_j |J_{i,j}| = J', \forall i$. The adjacency matrix of a regular Ising model is the matrix $A_{i,j} = |J_{i,j}|/J'$. Then, we show:

Theorem 2 *There is an algorithm based on Lasserre hierarchies which achieves an additive aproximation of $\epsilon n J'$ to $\log \mathcal{Z}$, where $\mathcal{Z} = \sum_{\mathbf{x}\in\{-1,1\}^n} \exp\left(\sum_{i,j} J_{i,j}\mathbf{x}_i\mathbf{x}_j\right)$, and runs in time $n^{rank(\Omega(\epsilon^2))/\Omega(\epsilon^2)}$, where rank$(\tau)$ is the number of eigenvalues of the adjacency matrix A greater than or equal to τ.*

It's interesting that this property of the graph, previously introduced for purposes of combinatorial optimization problems like small-set expansion, Unique Games (Steurer, 2010; Arora et al., 2010), also helps with counting type problems.

Note that since we prove additive factor guarantees to $\log Z$, using the fact the $e^\epsilon \leq 1 + 2\epsilon$ for small enough ϵ, we can easily turn them to multiplicative factor guarantees on Z. While these guarantees are not as strong as one usually gets in the correlation decay regime (i.e. $1+\epsilon$ multiplicative factor approximations to $\mathcal{Z}$ in time poly$(n, \frac{1}{\epsilon})$), to the best of our knowledge, these are the first approximations guarantees for $\mathcal{Z}$ when correlation decay does not hold. We discuss interesting regimes of the potentials $J_{i,j}$ in Section 5.

2.1. Outline of the techniques

Our approach can be summarized as follows. We first express the value of the log-partition function as the solution of a certain (intractable) optimization problem, by using a variational characterization of the log-partition function dating all the way back to Gibbs. (See Lemma 3.) To be more precise, we express it as $\log \mathcal{Z} = \max_{\mu \in \mathcal{M}} \{E(\mu) + H(\mu)\}$, where $\mathcal{M}$ is the polytope of distributions over $\{-1,1\}^n$, $E(\mu)$ is an *average energy* term, which dependes on pairwise marginals of μ only, and $H(\mu)$ is the Shannon entropy of μ.

The source of intractability comes from the fact that we cannot optimize over the polytope $\mathcal{M}$: we will instead optimize over a larger polytope $\mathcal{M}'$, which will come by considering *pseudo-distributions*, derived from either Sherali-Adams or Lasserre hierarchies. Additionally, we need to design a relaxation of $H(\mu)$, since in general we cannot hope to express the entropy of a distribution as a function its low-order marginals only.

The entropy relaxation $\tilde{H}(\mu)$ needs to satisfy $\tilde{H}(\mu) \geq H(\mu)$ for $\mu \in \mathcal{M}$ and needs to be concave in the variables used in the Sherali-Adams and Lasserre relaxations. The relaxation we use (See Section 4) will be based upon the chain rule for entropy, so it will be easy to prove that it upper bounds $H(\mu)$ (Proposition 7).

The analysis of the quality of the relaxation proceeds by *rounding* the pseudo-distributions to an actual distribution. This is slightly different from the roundings in combinatorial optimization, as there we only care about producing a *single* good $\{-1,1\}^n$ solution. Here, because of the entropy term, we must crucially produce a *distribution* over $\{-1,1\}^n$. The observation then is that we can view *correlation rounding*, a rounding previously used in works on combinatorial optimization (Barak et al., 2011; Yoshida and Zhou, 2014) as producing a distribution over $\{-1,1\}^n$ which has the same entropy as the $\tilde{H}(\mu)$ we defined. (Theorems 11, 13).

3. Preliminaries

We proceed with designing approximation algorithms for partition functions of Ising models first. Recall, an Ising model is a distribution $p : \{-1,1\}^n \to [0,1]$ that has the form $p(\mathbf{x}) \propto \exp\left(\sum_{i,j=1}^n J_{i,j}\mathbf{x}_i\mathbf{x}_j\right)$ and its partition function is $\mathcal{Z} = \sum_{\mathbf{x}\in\{-1,1\}^n} \left(\sum_{i,j=1}^n J_{i,j}\mathbf{x}_i\mathbf{x}_j\right)$.[3]

They are very commonly used in practical applications in machine learning because of their flexibility (and other appealing properties like being max-entropy distributions subject to moment constraints), and are extensively studied in theoretical computer science, statistical physics and probability theory.

A full survey is out of the scope of this paper, but we just mention that it can be shown that approximating Z within any polynomial factor is NP-hard for general potentials $J_{i,j}$ (Jerrum and Sinclair, 1993). When the potentials $J_{i,j}$ are all non-negative (also known as the *ferromagnetic* Ising model), (Jerrum and Sinclair, 1993) exhibit an FPRAS for computing $\mathcal{Z}$.

Let us set up the basic tools we will be using.

3.1. Variational methods

One of the main ideas all the algorithms will use is the following simple lemma, which characterizes Z as the solution of an optimization problem. It essentially dates back to Gibbs (Ellis, 2012), who used it in the context of statistical mechanics, though it has been rediscovered by machine learning

[3] There are many generalizations of this, allowing linear or higher order terms, as well as different domains than $\{-1,1\}^n$. Most results we prove can be generalized appropriately to these settings completely mechanically, so for clarity sake we focus on this case.

researchers (Wainwright and Jordan, 2008; Yedidia et al., 2003). For completeness, we reprove it here:

Lemma 3 (Variational characterization of $\log \mathcal{Z}$) *For any distribution $\mu : \{-1,1\}^n \to [0,1]$,*

$$\sum_{i,j} J_{i,j} \mathbb{E}_{\mu} [\mathbf{x}_i \mathbf{x}_j] + H(\mu) \leq \log \mathcal{Z}$$

with equality at $\mu = p$.

Proof For any distribution $\mu : \{-1,1\}^n \to [0,1]$, we can write the KL divergence between μ and p as

$$KL(\mu||p) = \mathbb{E}_{\mu} [\log \mu(\mathbf{x})] - \mathbb{E}_{\mu} [\log p(\mathbf{x})] = -H(\mu) - \sum_{i,j} J_{i,j} \mathbb{E}_{\mu} [\mathbf{x}_i \mathbf{x}_j] + \log \mathcal{Z}$$

Since the KL divergence is always non-negative, $-H(\mu) - \sum_{i,j} J_{i,j} \mathbb{E}_{\mu} [\mathbf{x}_i \mathbf{x}_j] + \log \mathcal{Z} \geq 0$. Hence, $\log \mathcal{Z} \geq H(\mu) + \sum_{i,j} J_{i,j} \mathbb{E}_{\mu} [\mathbf{x}_i \mathbf{x}_j]$ which proves the first claim of the lemma. However, equality is achieved whenever the KL divergence is 0, which happens when $\mu = p$. This finishes the second part of the lemma. ■

An immediate consequence of the above is the following:

Corollary 4 $\log \mathcal{Z} = \max_{\mu \in \mathcal{M}} \left\{ \sum_{i \sim j} J_{i,j} \mathbb{E}_{\mu} [\mathbf{x}_i \mathbf{x}_j] + H(\mu) \right\}$*, where $\mathcal{M}$ is the polytope of distributions over $\{-1,1\}^n$.*

We will use the above corollary as follows: instead of considering $\mu \in \mathcal{M}$, which is a polytope we cannot optimize over in polynomial time, we will consider $\mu \in \mathcal{M}'$, for a polytope $\mathcal{M}'$ satisfying $\mathcal{M} \subseteq \mathcal{M}'$, and feasible to optimize over. In fact, $\mathcal{M}'$ will be a polytope of *pseudo-distributions*, associated with either Sherali-Adams or Lasserre hierarchies. This idea is not new – it has appeared implicitly or explicitly in works on various types of belief propagation. (Wainwright and Jordan, 2008)

The novel thing is how we handle the entropy portion of the objective. Since $\mu \in \mathcal{M}'$ is no longer necessarily a distribution, we need to design surrogates for the entropy of μ. A popular choice in the literature is the so-called *Bethe* entropy, which roughly arises by taking the expression for the entropy of μ in terms of the pairwise marginals when the graph is a tree. (Of course, this expression is exact *only* if the graph is a tree. (Yedidia et al., 2003)) However, this approximation is not a relaxation of $\log \mathcal{Z}$ in the standard sense – the Bethe entropy is not an upper bound of the entropy, and the constructed approximation to $\log \mathcal{Z}$ is not concave in general, so the analysis proceeds by analyzing the belief propagation messages directly.[4]

We take a completely different approach. To get a proper relaxation for $\log \mathcal{Z}$, we design *functionals* $\tilde{H}(\mu)$ defined on $\mu \in \mathcal{M}'$, s.t. $\tilde{H}(\mu) \geq H(\mu)$ whenever $\mu \in \mathcal{M}$. In brief, we will use the following Corollary to 4:

[4] This approach usually works for graphs that are locally-tree-like (i.e. don't have short cycles), and for which some form of correlation decay holds.

Corollary 5 *If $\mathcal{M} \subseteq \mathcal{M}'$ and $H(\mu) \leq \tilde{H}(\mu)$ for $\mu \in \mathcal{M}$, then*

$$\log \mathcal{Z} \leq \max_{\mu \in \mathcal{M}'} \left\{ \sum_{i \sim j} J_{i,j} \mathbb{E}_{\mu} \left[\mathbf{x}_i \mathbf{x}_j \right] + \tilde{H}(\mu) \right\}$$

Subsequently, we will *round* the pseudo-distributions to actual distributions, in a manner that doesn't lose too much in terms of the value of the objective function.

3.2. Sherali-Adams and Lasserre hierarchies

We will be strongly using *hierarchies of convex relaxations*, capturing constraints on low-order moments and marginals of distributions. These are where our polytope $\mathcal{M}'$ will come from. While convex hierarchies have recently become relatively well-known in theoretical computer science, we still provide a (very) brief overview for completeness sake. For more details, the reader can consult (Barak et al., 2011; 2014; Laurent, 2009).

Recall, we are considering relaxations of the polytope of distributions over $\{-1,1\}^n$. The k-level *Sherali-Adams* hierarchy (henceforth SA(k)) has variables $\mu_S(\mathbf{x}_S), \mathbf{x}_S \in \{-1,1\}^{|S|}$ specifying local distributions over all subsets $S \subseteq [n]$, $|S| \leq k$. The distributions $\mu_S : \{-1,1\}^{|S|} \to [0,1]$ and $\mu_T : \{-1,1\}^{|T|} \to [0,1]$, for any S, T s.t. $|S \cup T| \leq k$ must be "consistent" on $S \cap T$. More precisely, it's the case that

$$\Pr_{\mathbf{x}_S \sim \mu_S} [\mathbf{x}_{S \cap T} = \alpha] = \Pr_{\mathbf{x}_T \sim \mu_T} [\mathbf{x}_{S \cap T} = \alpha], \forall S, T \subseteq [n], |S \cup T| \leq k$$

The fact that these constraints can be written as a linear program is well-known. (See e.g. (Barak et al., 2011))

We can also define a *conditioning* operation thanks to the existence of these local distributions. More precisely, for a vertex v, *conditioning* on v involves sampling v according to the local distribution $\mu_{\{v\}}$. This operation specifies a solution to the $k-1$-st level SA hierarchies: just define $\mu_S(\mathbf{x}_S) = \mu_{S \cup \{v\}}(\mathbf{x}_{S \cup v})$.

The additional power we get from the k-th level of the *Lasserre* hierarchy (henceforth LAS(k)) is that the semidefinite program provides vectors $v_{S,\alpha}$ for each subset S and possible assignment of values α to it, s.t. $\langle v_{S,\alpha}, v_{T,\beta} \rangle = \Pr_{\mu_{S \cup T}}(\mathbf{x}_S = \alpha, \mathbf{x}_T = \beta)$, if $|S \cup T| \leq k$.

4. Entropy respecting roundings

In this section we consider the functionals acting as surrogates for entropy. Recall, these need to be upper bounds on the entropy of a distribution μ on which we have essentially no handle other than having the first few moments. A clear candidate to do this is the *chain rule*.

Notice that for any set S of size at most k, where k is the number of levels of the Sherali-Adams or Lasserre hierarchy, $H(\mu_S)$ is a well-defined quantity: it's exactly

$$H(\mu_S) = \sum_{\mathbf{x}_S \in \{-1,1\}^{|S|}} \mu_S(\mathbf{x}_S) \log(\mu_s(\mathbf{x}_S))$$

Since these local quantities are essentially all the information about the joint distribution μ we have, our functional must involve such quantities only.

The simplest functional one can design surely is the following:

Definition *The* mean-field pseudo-entropy functional $H_{MF}(\mu)$ *is defined as* $H_{MF}(\mu) = \sum_{i=1}^{n} H(\mu_i)$.

Remark *Note, this is* not *the same as the usual mean-field approximation in statistical physics. The mathematical program analogue of that approximation would be to enforce that* $\mathbb{E}_\mu[\mathbf{x}_i\mathbf{x}_j] = \mathbb{E}_\mu[\mathbf{x}_i]\mathbb{E}_\mu[\mathbf{x}_j]$ *– which would result in a non-convex relaxation generally. We think the name is appropriate though, since the bound on the entropy is* mean-field*, i.e. results by treating* μ *as if it were a product distribution.*

Almost trivially for any $\mu \in \mathcal{M}$, the following proposition holds:

Proposition 6 *For any distribution* $\mu : \{-1,1\}^n \to [0,1]$, $H(\mu) \leq H_{MF}(\mu)$

Proof By the chain rule, $H(\mu) = \sum_{i=1}^{n} H(\mu_i|\mu_{[i-1]})$, where $[i-1]$ denotes the set $\{1, 2, \dots, i-1\}$ and $H(X|Y)$ is the conditional entropy of X given Y. However, since $H(\mu_i|\mu_{[i-1]}) \leq H(\mu_i)$ the claim trivially holds. ■

We will also consider generalizations of the above – where before applying the above ”mean-field” bound on the entropy, one can condition on a small subset first. Namely,

Definition *The* augmented mean-field pseudo-entropy functional *for subsets of size k,* $H_{aMF,k}(\mu)$ *is defined as* $H_{aMF,k}(\mu) = \min_{|S|\leq k} \left\{H(\mu_S) + \sum_{i\notin S} H(\mu_i|\mu_S)\right\}$.

The same proof as in Proposition 6 implies:

Proposition 7 $H(\mu) \leq H_{aMF,k}(\mu)$

Furthermore, it's quite easy to show that $H_{\text{aMF},k}(\mu)$, like $H_{\text{MF}}(\mu)$, is a concave function.

Lemma 8 *The pseudo-entropy functional* $H_{aMF,k}(\mu) = \min_{|S|\leq k} \left\{H(\mu_S) + \sum_{i\notin S} H(\mu_i|\mu_S)\right\}$ *is concave in the variables* $\{\mu_{S\cup\{i\}}\left(\mathbf{x}_{S\cup\{i\}}\right) | |S| \leq k, i \in [n]\}$.

Proof Since $H_{\text{aMF},k}(\mu) = \min_{|S|\leq k} \left\{H(\mu_S) + \sum_{i\notin S} H(\mu_i|\mu_S)\right\}$, and the minimum of concave functions is concave, all we need to show is that $H(\mu_S) + \sum_{i\notin S} H(\mu_i|\mu_S)$ is concave for all S. It's well known that entropy is a concave function, so $H(\mu_S)$ is concave. What remains to be shown is that $\sum_{i\notin S} H(\mu_i|\mu_S)$ is concave. But, since the sum of concave functions is concave, it suffices to prove $H(\mu_i|\mu_S)$ is concave.

The proof of this is essentially the same as the proof of concavity of entropy. Abusing notation a bit, we will denote as $\mu_A|\mathbf{x}_B$ the conditional distribution on the variables in A, conditioned on the variables in B having the value $\mathbf{x}_B$. We recall that

$$\begin{aligned}
H(\mu_i|\mu_S) &= \sum_{\mathbf{x}_S\in\{-1,1\}^{|S|}} \mu_s(\mathbf{x}_S) H(\mu_i|\mathbf{x}_s) \\
&= - \sum_{\mathbf{x}_S\in\{-1,1\}^{|S|}} \sum_{\mathbf{x}_i\in\{-1,1\}} \mu_s(\mathbf{x}_S)\mu_{i|\mathbf{x}_S}(\mathbf{x}_i) \log(\mu_{i|\mathbf{x}_S}(\mathbf{x}_i)) \\
&= - \sum_{\mathbf{x}_S\in\{-1,1\}^{|S|}} \sum_{\mathbf{x}_i\in\{-1,1\}} \mu_{S\cup\{i\}}(\mathbf{x}_{S\cup\{i\}}) \log(\mu_{i|\mathbf{x}_S}(\mathbf{x}_i)) \\
&= - \sum_{\mathbf{x}_S\in\{-1,1\}^{|S|}} \sum_{\mathbf{x}_i\in\{-1,1\}} \mu_{S\cup\{i\}}(\mathbf{x}_{S\cup\{i\}}) \log\left(\frac{\mu_{S\cup\{i\}}(\mathbf{x}_{S\cup\{i\}})}{\mu_s(\mathbf{x}_S)}\right)
\end{aligned}$$

We rewrite the last expression as a KL divergence as follows:

$$-\sum_{\mathbf{x}_S\in\{-1,1\}^{|S|}}\sum_{x_i\in\{-1,1\}}\mu_{S\cup\{i\}}(\mathbf{x}_{S\cup\{i\}})\log\left(\frac{\mu_{S\cup\{i\}}(\mathbf{x}_{S\cup\{i\}})}{\mu_s(\mathbf{x}_S)\frac{1}{2}}\right)+1=-KL(\mu_{S\cup\{i\}}||(\mu_S\times r))+1 \tag{4.1}$$

where r is a uniform distribution over $\{-1,1\}$.

Then, if $\mu^{\lambda}_{S\cup\{i\}}=\lambda\mu^1_{S\cup\{i\}}+(1-\lambda)\mu^2_{S\cup\{i\}}$, we want to show

$$H(\mu^{\lambda}_i|\mu^{\lambda}_S)\geq\lambda H(\mu^1_i|\mu^1_S)+(1-\lambda)H(\mu^2_i|\mu^2_S)$$

By (4.1) and the convexity of KL divergence,

$$\begin{aligned}H(\mu^{\lambda}_i|\mu^{\lambda}_S)&=-KL(\mu^{\lambda}_{S\cup\{i\}}||(\mu^{\lambda}_S\times r))+1\\&\geq-\lambda KL(\mu^1_{S\cup\{i\}}||(\mu^1_S\times r))-(1-\lambda)KL(\mu^2_{S\cup\{i\}}||(\mu^2_S\times r))+1\\&=\lambda H(\mu^1_i|\mu^1_S)+(1-\lambda)H(\mu^2_i|\mu^2_S)\end{aligned}$$

which is what we want.

■

4.1. Dense Ising models

We finally turn to designing an algorithm for "dense" Ising models.

There are multiple reasons to study this particular subclass: from the theoretical computer science point of view, we have various PTAS for constraint satisfaction problem when the constraint graph is dense (Yoshida and Zhou, 2014; Arora et al., 1995) so we might hope to get results better than the worst-case one ones for partition function calculation as well.

Another motivation comes from *mean-field* ferromagnetic Ising model (also known as the *Curie-Weiss* model (Ellis and Newman, 1978)), which is frequently studied as a very simplified model of ferromagnetism because one can get relatively easily results about global properties of the model like the partition function, magnetization, etc. In the mean-field model, each spin interacts (equally strongly) with every other spin.

We will, in this section, generalize the classical results about the ferromagnetic Curie-Weiss model, as well as provide the natural counterpart of the results in (Yoshida and Zhou, 2014; Arora et al., 1995) for partition functions.

Let us first review the standard results about Curie-Weiss. Recall, this model follows the distribution $p(\mathbf{x})\propto\exp\left(\sum_{i,j=1}^n\frac{J}{n}\mathbf{x}_i\mathbf{x}_j\right)$, $J>0$. It is easy to analyze because $p(\mathbf{x})$ factorizes and can be "reparametrized" in terms of the magnetization. Namely, since $\sum_{i,j=1}^n\frac{J}{n}\mathbf{x}_i\mathbf{x}_j=\frac{J}{n}(\sum_i\mathbf{x}_i)^2$, and $(\sum_i\mathbf{x}_i)^2\in[-n,n]$, one can show (Ellis and Newman, 1978):

Theorem 9 ((Ellis and Newman, 1978)) *For the Curie-Weiss model,*

$$\log\mathcal{Z}=(1\pm o(1))\left(n\max_{m\in[-1,1]}\left(Jm^2+\frac{1-m}{2}\log\frac{1-m}{2}+\frac{1+m}{2}\log\frac{1+m}{2}\right)\right)$$

The proof of this theorem involves rewriting the expression for $\mathcal{Z}$ as follows:

$$\mathcal{Z} = \sum_{\mathbf{x} \in \{-1,1\}^n} \exp\left(\sum_{i,j} \frac{J}{n} \mathbf{x}_i \mathbf{x}_j\right) = \sum_{l} \exp\left(\frac{J}{n} l^2\right) \cdot n_l$$

where n_l is the number of terms where $\sum_{i=1}^{n} \mathbf{x}_i = l$. Then, using Stirling's formula and some more algebraic manipulation, one can estimate the dominating term in the summation. The claim of the theorem then follows.

We significantly generalize the above claim using notions from theoretical computer science. The goal is to prove Theorem 1.

Let $J_T = \sum_{i,j} |J_{i,j}|$. As discussed in Section 2, we define the following notion of density inspired by the definition of a dense graph in combinatorial optimization (Yoshida and Zhou, 2014):

Definition *An Ising model is* Δ-dense *if* $\forall i \neq j, \Delta |J_{i,j}| \leq \frac{J_T}{n^2}$, $\Delta \in (0, 1]$.

We will consider the relaxation for $\log \mathcal{Z}$ given by the augmented pseudo-entropy functional and the level $k = O(1/(\Delta\epsilon^2))$ Sherali-Adams relaxation, namely:

$$\max_{\mu \in \mathrm{SA}(k), k = O(1/(\Delta\epsilon^2))} \left\{ \sum_{i,j} J_{i,j} \mathbb{E}_{\mu} \left[\mathbf{x}_i \mathbf{x}_j\right] + H_{\mathrm{aMF},k}(\mu) \right\} \tag{4.2}$$

We also recall *correlation rounding* as defined in (Barak et al., 2011). In correlation rounding, we pick a "seed set" of a certain size, condition on it, and round the rest of the variables independently. The usual thing to prove is that there is a good "seet set" of a small size to condition on. In particular, for the dense case, the following lemma was proven in (Yoshida and Zhou, 2014):

Lemma 10 ((Yoshida and Zhou, 2014)) *There exists a set S of size $k = O(1/(\Delta\epsilon^2))$, s.t.*

$$\left| \sum_{i,j} J_{i,j} \mathbb{E}_{\mu} \left[\mathbf{x}_i \mathbf{x}_j | \mathbf{x}_S\right] - \sum_{i,j} J_{i,j} \mathbb{E}_{\mu} \left[\mathbf{x}_i | \mathbf{x}_S\right] \mathbb{E}_{\mu} \left[\mathbf{x}_j | \mathbf{x}_S\right] \right| \leq \frac{100}{\Delta k} J_T$$

With this in hand, we proceed to the main theorem of this section:

Theorem 11 (Restatement of Theorem 1) *The output of 4.2 is an ϵJ_T additive approximation to* $\log Z$.

Proof The function 4.2 is optimizing is a sum of two terms: $\sum_{i \sim j} J_{i,j} \mathbb{E}_{\mu} \left[\mathbf{x}_i \mathbf{x}_j\right]$ and an entropy term. Following standard terminology in statistical physics, we will call the former term *average energy*.

We will analyze the quality of the convex relaxation by exhibiting a *rounding* of the pseudo-distribution to an actual distribution. There is a difference in what this means compared to the roundings we use in combinatorial optimization: there we only care about producing a *single* $\{+1, -1\}$ solution. Here, because of the entropy term, it's essential that we produce a *distribution* over $\{+1, -1\}$ solutions.

We use the fact that correlation rounding can be viewed as producing distributions with a fairly explicit expression for their entropy. Let S be the set of size $O(\frac{1}{\Delta\epsilon^2})$ that Lemma 10 gives. Consider the distribution $\tilde{\mu}(\mathbf{x}) = \mu(\mathbf{x}_S)\Pi_{i\notin S}\mu(\mathbf{x}_i|\mathbf{x}_S)$[5]. In other words, this is the distribution which rounds the variables in S according to their local distribution, and all other variables independently according to the conditional distribution on $\mathbf{x}_S$.

Consider the average energy first. By Lemma 10,

$$\left|\sum_{i,j} J_{i,j}\mathbb{E}_{\mu}\left[\mathbf{x}_i\mathbf{x}_j|\mathbf{x}_S\right] - \sum_{i,j} J_{i,j}\mathbb{E}_{\tilde{\mu}}\left[\mathbf{x}_i\mathbf{x}_j|\mathbf{x}_S\right]\right| \leq J_T\epsilon$$

Now consider the entropy term. The entropy of the distribution $\tilde{\mu}$ is $H(\tilde{\mu}) = H(\mu_S) + \sum_{i\notin S} H(\mu_i|\mu_S)$. But, since $H_{\text{aMF},k}(\mu) = \min_{|S|\leq k}\left\{H(\mu_S) + \sum_{i\notin S} H(\mu_i|\mu_S)\right\}$, $H_{\text{aMF},k}(\mu) \leq H(\tilde{\mu})$ follows. This immediately implies that

$$\left(\sum_{i,j} J_{i,j}\mathbb{E}_{\mu}\left[\mathbf{x}_i\mathbf{x}_j\right] + H_{\text{aMF},k}(\mu)\right) - \left(\sum_{i,j} J_{i,j}\mathbb{E}_{\tilde{\mu}}\left[\mathbf{x}_i\mathbf{x}_j\right] + H(\tilde{\mu})\right) =$$

$$\left(\sum_{i,j} J_{i,j}\mathbb{E}_{\mu}\left[\mathbf{x}_i\mathbf{x}_j\right] - \sum_{i,j} J_{i,j}\mathbb{E}_{\tilde{\mu}}\left[\mathbf{x}_i\mathbf{x}_j\right]\right) + \left(H_{\text{aMF},k}(\mu) - H(\tilde{\mu})\right) \leq J_T\epsilon$$

This exactly proves the claim we want.

■

Notice, in the case of the Curie-Weiss model, since $J > 0$, the value of the relaxation 4.2 is at least J_T, Theorem 11 gives a $1+\epsilon$ multiplicative factor approximation to $\log Z$ for any constant ϵ, so generalizes the statement of Theorem 9 to cases where the potentials $J_{i,j}$ might vary in magnitude and sign.

4.2. Low threshold rank Ising models

If we use the added power of the Lasserre hierarchy, we can also handle Ising models whose weights look like low rank matrices. We want to prove Theorem 2.

We will consider for simplicity in this section *regular* Ising models in the weighted sense, meaning $\sum_j |J_{i,j}| = J', \forall i$ [6]. The *adjacency matrix* of an Ising model will be the doubly-stochastic matrix with entries $|J_{i,j}|/J'$.

Let's recall the definition of threshold rank from (Arora et al., 2010):

Definition *The τ-threshold rank of a regular graph is the number of eigenvalues of the normalized adjacency matrix greater than or equal to τ.*

We will, in analogy, define the threshold rank of an Ising model.

[5] Notice this is an actual, well-defined distribution, and not only a pseudo-distribution anymore.

[6] Though we remind again, all of the claims can be appropriately generalized at the expense of more bothersome notation.

Definition *The τ-threshold rank of a regular Ising model is the number of eigenvalues of its adjacency matrix greater than or equal to τ.*

We will consider the following convex program:

$$\max_{\mu \in \text{LAS}(k)} \left\{ \sum_{i,j} J_{i,j} \mathbb{E}_{\mu} \left[\mathbf{x}_i \mathbf{x}_j \right] + H_{\text{aMF},k}(\mu) \right\} \tag{4.3}$$

Consider the vectors $v_i, i \in [n]$, s.t. $\langle v_j, v_j \rangle = \mathbb{E}_{\mu} \left[\mathbf{x}_i \mathbf{x}_j \right]$. Then, (Barak et al., 2011) prove that when the graph has low threshold rank, "local" correlations propagate to "global" correlations, and as a consequence of this, there is a set of size at most rank$(\Omega(\epsilon^2))/\Omega(\epsilon^2)$, such that conditioning on it causes the $\left| \sum_{i,j} J_{i,j} \mathbb{E}_{\mu} \left[\mathbf{x}_i \mathbf{x}_j | \mathbf{x}_S \right] - \sum_{i,j} J_{i,j} \mathbb{E}_{\tilde{\mu}} \left[\mathbf{x}_i \mathbf{x}_j | \mathbf{x}_S \right] \right|$ to drop below ϵJ_T. More precisely:

Lemma 12 ((Barak et al., 2011)) *There exists a set S of size $t \leq$ rank$(\Omega(\epsilon^2))/\Omega(\epsilon^2)$, where rank$(\tau)$ is the τ-threshold rank of the Ising model, s.t.* [7]

$$\left| \sum_{i,j} J_{i,j} \mathbb{E}_{\mu} \left[\mathbf{x}_i \mathbf{x}_j | \mathbf{x}_S \right] - \sum_{i,j} J_{i,j} \mathbb{E}_{\mu} \left[\mathbf{x}_i | \mathbf{x}_S \right] \mathbb{E}_{\mu} \left[\mathbf{x}_j | \mathbf{x}_S \right] \right| \leq \epsilon J_T$$

Hence, analogously as in Theorem 11, we get:

Theorem 13 (Restatement of Theorem 2) *The output of 4.3 is a ϵJ_T additive approximation to* $\log \mathcal{Z}$.

5. Discussion on interpreting the results

Since the above results are stated in terms of the additive approximation they provide for $\log \mathcal{Z}$, we discuss how one should interpret them in different "temperature regimes" i.e. different scales of the potentials $J_{i,j}$. Note that partition function approximation problems are not scale-invariant, and their hardness is sensitive to the size of the coefficients $J_{i,j}$.

For simplicity of the discussion, let's focus on the case where there is an underlying graph $G = (V, E)$, such that $J_{i,j} = \pm J$, for $(i, j) \in E(G)$, and 0 otherwise. Furthermore, let's assume the graph G is d-regular.

There are generically three regimes for the problem:

- "High temperature regime", i.e. when $|J| = O\left(\frac{1}{d}\right)$ for a sufficiently small constant in the $O(\cdot)$ notation. In this case, standard techniques like Dobrushin's uniqueness criterion show that there is correlation decay. This is the regime where generically Markov Chain methods work. Note that using such methods, generally one can get a $(1 + \epsilon)$-factor approximation for $\mathcal{Z}$ in time poly $\left(n, \frac{1}{\epsilon}\right)$, which is unfortunately much stronger than what our method gets in that regime. It would be extremely interesting to see if the methods in our paper can be modified to subsume this regime as well.

[7] Note, $J_T = nJ'$ in this case.

- "Around the transition threshold", i.e. when $|J| = \Theta(\frac{1}{d})$ for a sufficiently large constant in the Θ notation, such that there is no correlation decay. Generally, unless there is some special structure, Markov Chain methods will provide *no non-trivial* guarantee in this regime – however, we get an order ϵn additive factor approximation to $\log \mathcal{Z}$, which translates to a $(1+\epsilon)^n$ factor approximation of $\mathcal{Z}$. We do not, to the best of our knowledge, know how to get such results using *any other methods.*

- "Low temperature regime", i.e. when $|J| = \omega(1/d)$. In this case, in light of the variational characterization of $\log \mathcal{Z}$ and the fact that the entropy is upper bounded by n, the dominating term will typically be the energy term $\sum_{(i,j)\in E(G)} J_{i,j}\mathbb{E}_\mu[\mathbf{x}_i\mathbf{x}_j]$, so essentially the quality of approximation will be dictated by the hardness of the optimization problem corresponding to the energy term. (e.g., for the anti-ferromagnetic case, where all the potentials $J_{i,j}$ are negative, the optimization problem corresponding to the energy term is just max-cut, and we cannot hope for more than a constant factor approximation to $\log \mathcal{Z}$ for general (negative) potentials.)

6. Conclusion

We presented simple new algorithms for calculating partition functions in Ising models based on variational methods and convex programming hierarchies. To the best of our knowledge, these techniques give new, non-trivial approximation guarantees for the partition function when correlation decay does not hold, and are the first provable, convex variational methods. Our guarantees are for dense or low threshold rank graphs, and in the process we design novel *entropy approximations* based on the low-order moments of a distribution.

We barely scratched the surface, and we leave many interesting directions open. Our methods are very generic, and are probably applicable to many other classes of partition functions apart from Ising models. One natural candidate is weighted matchings due to the connections to calculating non-negative permanents.

Another intriguing question is to determine if there is a similar approach that can subsume prior results on partition function calculation in the regime of correlation decay, as our guarantees are much weaker there. This would give a convex relaxation interpretation of these types of results.

References

Sanjeev Arora and Rong Ge. New tools for graph coloring. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 1–12. Springer, 2011.

Sanjeev Arora, David Karger, and Marek Karpinski. Polynomial time approximation schemes for dense instances of np-hard problems. In *Proceedings of the twenty-seventh annual ACM symposium on Theory of computing*, pages 284–293. ACM, 1995.

Sanjeev Arora, Boaz Barak, and David Steurer. Subexponential algorithms for unique games and related problems. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 563–572. IEEE, 2010.

Boaz Barak, Prasad Raghavendra, and David Steurer. Rounding semidefinite programming hierarchies via global correlation. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 472–481. IEEE, 2011.

Boaz Barak, Jonathan A Kelner, and David Steurer. Rounding sum-of-squares relaxations. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 31–40. ACM, 2014.

Alexander Barvinok. Polynomial time algorithms to approximate permanents and mixed discriminants within a simply exponential factor. *Ann Arbor*, 1001(48109):1109.

Mohsen Bayati, David Gamarnik, Dimitriy Katz, Chandra Nair, and Prasad Tetali. Simple deterministic approximation algorithms for counting matchings. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 122–127. ACM, 2007.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *arXiv preprint arXiv:1601.00670*, 2016.

Yiling Chen, Lance Fortnow, Nicolas Lambert, David M Pennock, and Jennifer Wortman. Complexity of combinatorial market makers. In *Proceedings of the 9th ACM conference on Electronic commerce*, pages 190–199. ACM, 2008.

Richard S Ellis. *Entropy, large deviations, and statistical mechanics*, volume 271. Springer Science & Business Media, 2012.

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

David Gamarnik and Dmitriy Katz. A deterministic approximation algorithm for computing the permanent of a 0, 1 matrix. *Journal of Computer and System Sciences*, 76(8):879–883, 2010.

Leonid Gurvits and Alex Samorodnitsky. Bounds on the permanent and some applications. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, pages 90–99, 2014. doi: 10.1109/FOCS.2014.18. URL http://dx.doi.org/10.1109/FOCS.2014.18.

David P Helmbold and Manfred K Warmuth. Learning permutations with exponential weights. *The Journal of Machine Learning Research*, 10:1705–1736, 2009.

Tom Heskes. Convexity arguments for efficient minimization of the bethe and kikuchi free energies. 2006.

Mark Jerrum and Alistair Sinclair. Polynomial-time approximation algorithms for the ising model. *SIAM Journal on computing*, 22(5):1087–1116, 1993.

Mark Jerrum and Umesh Vazirani. A mildly exponential approximation algorithm for the permanent. *Algorithmica*, 16(4):392–401, 1996.

Mark Jerrum, Alistair Sinclair, and Eric Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *Journal of the ACM (JACM)*, 51(4):671–697, 2004.

Monique Laurent. Sums of squares, moment matrices and optimization over polynomials. In *Emerging applications of algebraic geometry*, pages 157–270. Springer, 2009.

Liang Li, Pinyan Lu, and Yitong Yin. Approximate counting via correlation decay in spin systems. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 922–940. SIAM, 2012.

Liang Li, Pinyan Lu, and Yitong Yin. Correlation decay up to uniqueness in spin systems. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 67–84. SIAM, 2013.

Ofer Meshi, Ariel Jaimovich, Amir Globerson, and Nir Friedman. Convexifying the bethe free energy. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 402–410. AUAI Press, 2009.

Alistair Sinclair, Piyush Srivastava, and Marc Thurley. Approximation algorithms for two-state anti-ferromagnetic spin systems on bounded degree graphs. *Journal of Statistical Physics*, 155 (4):666–686, 2014.

Allan Sly. Computational transition at the uniqueness threshold. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 287–296. IEEE, 2010.

David Steurer. *On the complexity of unique games and graph expansion*. PhD thesis, Princeton University, 2010.

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.

Martin J Wainwright, Tommi S Jaakkola, and Alan S Willsky. Tree-reweighted belief propagation algorithms and approximate ml estimation by pseudo-moment matching.

Martin J Wainwright, Tommi S Jaakkola, and Alan S Willsky. A new class of upper bounds on the log partition function. *Information Theory, IEEE Transactions on*, 51(7):2313–2335, 2005.

Martin J Wainwright, Michael Jordan, et al. Log-determinant relaxation for approximate inference in discrete markov random fields. *Signal Processing, IEEE Transactions on*, 54(6):2099–2109, 2006.

Yair Weiss. Correctness of local probability propagation in graphical models with loops. *Neural computation*, 12(1):1–41, 2000.

Dror Weitz. Counting independent sets up to the tree threshold. In *Proceedings of the thirty-eighth annual ACM Symposium on Theory of Computing*, pages 140–149. ACM, 2006.

Avi Wigderson. personal communication.

Jonathan S Yedidia, William T Freeman, and Yair Weiss. Understanding belief propagation and its generalizations. 2003.

Yuichi Yoshida and Yuan Zhou. Approximation schemes via sherali-adams hierarchy for dense constraint satisfaction problems and assignment problems. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 423–438. ACM, 2014.