Mean-field approximation and correlation rounding

Alexander Cai, Oliver Cheng, Tianze Jiang*

Abstract

The following abstract is from [RT12; JKM18; JKR19] and related works. Computing the *free energy* (logarithm of the partition function) of an Ising model is a key problem in the studies of statistical physics. While the exact problem is computationally intractable, a common scheme for approximating this key quantity is through the mean-field approximation, which bounds the free energy from below. In this paper, we will show this approximation scheme's generally tight error bound via a fundamental technique called correlation rounding.

Expanding on correlation rounding, we will then use it to approximate MaxCut on an expander-like graph via convex relaxation. Observing that MaxCut is equivalent to Max-A-Posteriori estimations for Ising models (and related to the variational estimate of the free energy), our proof shows how this simple technique can be powerful in analyzing complex graphical models.

1 Introduction

The *Ising model* is the de facto standard model in statistical physics used to test the strength of proposed techniques. It describes a graph of n atoms, each with either positive or negative magnetic spin. These spins may correlate with one another, and hence we can characterize an Ising model by a symmetric matrix J, where $(J)_{ij}$ represents the coupling strength between spins i and j. In addition, this coupling matrix is scaled by the inverse temperature $\beta = \frac{1}{k_B T}$, where the Boltzmann constant k_B is fixed and T is the temperature ([JKR19]). The Hamiltonian, or energy of the system, is then given by

$$E_\beta(x) = \sum_{i,j} (\beta J_{ij}) x_i x_j = x^\top J_\beta x \tag{1.1}$$

where $J_\beta = \beta J$. At equilibrium, the probability distribution of states is given by the Boltzmann distribution, which is our main object of interest:

$$\mathbb{P}_\beta[X = x] \propto \mathsf{exp}(-E_\beta(x)) \tag{1.2}$$

The normalizing constant $Z_\beta = \sum_x \mathsf{exp}(-E_\beta(x))$ is related to the (Gibbs) free energy $\mathcal{F}_\beta$ of the system by:

$$\mathcal{F}_\beta = \beta^{-1} \log Z_\beta \tag{1.3}$$

For our purposes, we will set $\beta = 1$, though in each section we will briefly mention its effect on our approximations and algorithms. A remarkable note from physics is that, as the size of the system as $n \to \infty$, properties of the system often undergo *phase transitions* as β varies ([JKR19]).

Computing (or even approximating) the free energy $\mathcal{F}$ (or partition function Z) is a central task in analyzing Ising models as it can easily be turned into efficient sampling algorithms or computing

*Course project in completion of the class CS224 (Algorithms for Data Science) Fall 2023. We thank Prof. Chen for providing valuable feedbacks and extra references.

marginals/conditionals. However, the exact computation of $\mathcal{F}$ for a general J is #-P complete (very hard) ([SJ89]), and even approximation remains NP-hard ([SS12]). Instead of summing naively, one can formulate the free energy in terms of an optimization over distributions.

Proposition 1 (Gibbs Variational Principle)**.** *For $\mathcal{F} = \log Z$, the free energy, one has that:*

$$\mathcal{F} = \max_{\mu \in \Delta(\mathbb{R}^n)} \left[\sum_{i<j} J_{ij} \mathbb{E}_\mu \left[X_i X_j \right] + H(\mu) \right]. \tag{1.4}$$

where $\Delta(\mathbb{R}^n)$ denotes the probability simplex over $\mathbb{R}^n$ and $H(\mu) = \mathbb{E}_\mu[-\log \mu]$ denotes the entropy of the distribution μ. Moreover, the minimizer of (1.4) is exactly the Boltzmann distribution $\mu^* = P$.

Proof. The proof relies on the fact that $\mathbf{KL}(\mu \| P) \geq 0$ for P being our Ising model. In fact, recall that $E(X) = X^\top J X$ is the Hamiltonian and $\log Z = -\log P(X) - E(X)$ for all X, hence:

$$\begin{aligned} \log Z - H(\mu) &= \mathbb{E}_\mu[-\log P - E(X) + \log \mu] \\ &= \mathbb{E}_\mu[-\log P + \log \mu] - \mathbb{E}_\mu[X^\top J X] \\ &= \mathbf{KL}(\mu \| P) + \sum J_{ij} \mathbb{E}_\mu[X_i X_j] \end{aligned}$$

We conclude the proof as $\mathbf{KL}$ is always non-negative. □

Unfortunately, for large n the space of possible distributions grows exponentially, and finding such a μ from (1.4) is intractable. Hence one considers reducing the space of possible distributions to contain only product distributions, where μ can be represented by a n-dimensional vector x. This technique is called *mean-field approximation*. Hence we can define the mean-field variational free energy as

$$\mathcal{F}^* = \max_{x \in [-1,1]^n} \left[\sum_{i<j} J_{ij} x_i x_j + \sum_i H\left(\frac{x_i + 1}{2} \right) \right] \tag{1.5}$$

where $H(p) = -p \log p - (1-p) \log(1-p)$ denotes the entropy of the Bern(p) distribution. It is clear that $\mathcal{F}^* \leq \mathcal{F}$ since the space of distributions being optimized over in (1.5) is a strict subset of that in (1.4). Theorem 1 will provide bounds as well as an idea of how to well approximate the marginals of the true solution to the mean-field equations [JKR19]. Furthermore, the critical points to (1.5) correspond to the fixed points of the map $x \mapsto \tanh(Jx)$, but this is only known to converge for high temperature (and hence low β) regimes [JKM18]. In general, it is not clear how to optimize over (1.5).

1.1 Summary and Organization

This expository paper will review the results from Jain et al. in [JKM18; JKR19] and connect these results to the large body of MaxCut results by Raghavendra et al. in [RT12]. In particular we go over the improved bounds [JKM18; JKR19] manages to obtain on the error of the mean-field estimation. Their primary tool to improve these bounds is called correlation rounding. Our two major topics are as follows:

1. We introduce correlation rounding and prove the following error bound on the mean-field approximation:

Theorem 1 (Estimation error of the optimal mean-field)**.** *Consider an Ising model with coupling matrix J. Let $\mathcal{F}$ be the free energy and $\mathcal{F}^* \leq \mathcal{F}$ be the optimal mean-field variational estimation, then $\mathcal{F} - \mathcal{F}^* \lesssim (n\|J\|_F)^{2/3}$.*

This bound turns out to be generically tight even beyond the space of product distributions which the mean-field approximation optimizes over when $\|J\|_F = \Theta(n)$. We discuss the class of distributions for which this holds true and the result in Section 2.3.

2. We discuss results obtained by applying correlation rounding to MaxCut. In particular, we introduce an algorithm using Sherali-Adams convex relaxation to obtain an approximate cut with $O(\delta)$ error in $n^{O(\delta^{-2})}$-time for an expander-like graph.

The papers we survey contain interesting results we will not have space to cover such as generalizations to k-Markov Random Fields [JKR19], runtime bounds [JKR19], and other classes of constraint satisfaction-type problems [RT12]. We refer the interested reader to the cited papers.

1.2 Preliminaries and notations

We present fundamental notation and definitions we will use throughout the paper. For shorthand for any $S \subset [n]$, we will notate $X_S = (X_{i_1}, \ldots, X_{i_{|S|}})$. Furthermore, for any (pseudo)distribution μ we will use μ_S to denote the marginal distribution restricted on indices S.

We use TV as the total variation distance between two distributions, and $\mathbf{KL}$ as the Kullback-Leiber divergence. Pinsker's inequality tells us that

$$2\,\mathsf{TV}^2(\mu,\nu) \leq \mathbf{KL}(\mu\|\nu), \tag{1.6}$$

which we will use throughout (see e.g. [PW23]). The *information entropy* of a distribution μ is defined as $H(\mu) \triangleq \mathbb{E}_{x\sim\mu}[-\log \mu(x)]$. The *mutual information* between two distributions X, Y with marginals P_X, P_Y respectively and a joint of $P_{X,Y}$ is defined as $I(X,Y) = \mathbf{KL}(P_{X,Y}\|P_X \times P_Y)$. Intuitively, having lower mutual information means our joint distribution is close to the independent product of its marginals. Finally, the conditional entropy $H(X \mid Y)$ is defined as the information entropy of $\mathbb{P}(X \mid Y = y)$ taken in expectation over $y \sim Y$, and hence one can verify that $I(X;Y) = H(X) - H(X \mid Y)$.

2 Mean-field error bound via correlation rounding

2.1 Correlation rounding

The fundamental tool we will use to prove 1 is correlation rounding. We seek to answer the following question: Under what criteria can one approximate a general distribution using a product distribution?

At a high level, for any collection of random variables $X_1, X_2, \ldots, X_n$, one would intuitively expect one of the following two cases to hold:

1. The average covariance between pairs is small and hence the collection is close to independent.

2. The average covariance is not small, but some coordinates are "bad" in the sense that they contribute to a large dependency in the collection and that conditioning on (even an average configuration of them) removes the pair-wise dependency on the rest of the variables.

This intuition is indeed true and has been formalized rigorously in many different settings in the literature, most notably as the "pinning lemma" in the studies of statistical physics ([IV00]).

Proposition 2 (Correlation Rounding, [RT12]). *Let $X_1, \dots, X_n$ be a collection of $\{\pm 1\}$-valued random variables. Then, for any $\ell \in o(n)$, there exists some $S \subset [n]$ with $|S| \le \ell$ such that:*

$$\mathbb{E}_{X_S} \mathbb{E}_{\{u,v\} \in \binom{[n]}{2}} \left[\mathsf{cov}\left(X_u, X_v \mid X_S\right)^2 \right] \lesssim \frac{1}{\ell}.$$

Proof. Let μ denote a distribution over $\{\pm 1\}^n$ and consider the potential function $\Phi(\mu) = \frac{1}{n} \sum H(\mu_i) \ge 0$. For any $i \in [n]$, consider the average potential conditioned on X_i. The potential changes by:

$$\mathbb{E}_i[\mathbb{E}_{X_i} \Phi(\mu | X_i)] - \Phi(\mu) = \mathbb{E}_i \left[\frac{1}{n} \sum_j H(X_j | X_i) - H(X_i) \right] = -\mathbb{E}_{i,j} I(X_i, X_j) < 0.$$

One can thus consider the following process: in each step, we condition on one more variable X_i such that the potential drops by at least $\mathbb{E}_{i,j} I(X_i, X_j)$ (note that after conditioning on a variable, the terms involving that variable vanish, but we keep them for notational simplicity). Now suppose $\mathbb{E}_{i,j} I(X_i, X_j) \ge 1/\ell$ holds throughout the first t steps of conditioning, this means that the potential at the start has to be at least t/ℓ. However, it is clear that $\Phi(\mu) \lesssim 1$ since the binary entropy is bounded above, the process cannot go above $t = O(\ell)$ rounds before reaching a set S conditioned on which the sum of mutual information is small!

We need another lemma to conclude, which claims that

$$|\,\mathsf{cov}(X_i, X_j)| = 2\,\mathsf{TV}(P_{X_i,X_j}, P_{X_i} \times P_{X_j})$$

for $\{\pm 1\}$-valued random variables (see e.g. [BRS11]), combined with Pinsker's inequality (1.6) to get:

$$\ell^{-1} \gtrsim \mathbb{E}_{i,j} I(X_i, X_j) \gtrsim \mathbb{E}_{i,j}\, \mathsf{TV}(P_{X_i,X_j}, P_{X_i} \times P_{X_j})^2 \asymp \mathbb{E}_{i,j}\, \mathsf{cov}(X_i, X_j)^2$$

when conditioned on an average $X_S \sim \mu_S$ for some $|S| \lesssim \ell$. □

Remark 2.1. Later in this note, we will see that even conditioning on an *average* S successfully removes the dependency (Lemma 5) using essentially the same proof.

Remark 2.2. Allen and O'Donell conjectured in [AO15] that the bound on covariance can be tightened to $\frac{1}{\ell^2}$. However, [JKR19; JKM18] refutes this conjecture using theory from Sherrington-Kirkpatrick Spin Glass models to show that the bound is indeed tight for this class of Ising models.

2.2 Proof of Theorem 1

Previous results by Jain, Koehler, and Mossel in [JKM18] show that the error bound for the mean-field approximation is $O(n^{2/3}\|J\|_F^{2/3}\,\mathsf{log}^{1/3}(n\|J\|_F))$. The results of [JKR19] take out this log factor and achieved a generally tight error bound (Section 2.3). Theorem 1 can be generalized to k Markov Random Fields (and hence relates to problems such as Max-k-CSP) (Section 5, [JKR19]).

Proof of Theorem 1. We are looking for a product distribution that is statistically close to the true distribution μ with error $\epsilon > 0$. Take $\ell = \frac{1}{\epsilon^2 \log 2}$. Recall from Prop. 2 there exists some subset S such that $|S| \le \ell$, where the average covariance between pairs is small upon conditioning by S. We take ν_{x_S} to be the product distribution that agrees with the true μ on first moments $\mathbb{E}_{\nu_{x_S}}[X_i] = \mathbb{E}_\mu[X_i | X_S = x_S]$. It suffices to show that in expectation over the randomness of x_S, our

mean field approximation has $O(n^{2/3}\|J\|_F^{2/3})$ error, as by an averaging argument this implies there exists such a x_S that satisfies the error, and hence there exists a product distribution ν_{x_S} as well.

Recall from the definition of entropy we have that $H_\mu(X) = H_\mu(X|X_S) + H_\mu(X_S)$. From our definition of variational free energy 1.4, we have

$$\begin{aligned}\mathcal{F} &= \sum_{i<j} J_{ij}\mathbb{E}_\mu[X_iX_j] + H_\mu(X|X_S) + H_\mu(X_S) \\ &= \mathbb{E}_{x_S}\left[\sum_{i<j} J_{ij}\mathbb{E}_\mu[X_iX_j|X_S = x_S] + H_\mu(X|X_S = x_S)\right] + H_\mu(X_S) \end{aligned} \tag{2.1}$$

Here, the second equality comes from the Law of Total Expectation. Furthermore, we have that $H_\mu(X_S) \le \log 2^{|S|} \le \log 2^\ell \le \frac{1}{\epsilon^2}$. In addition, from direct application of the chain rule of entropy and the fact $H(A|B) \le H(A)$, we have that the second term can be written in terms of our product distribution ν_{X_S},

$$H_\mu(X|X_S = x_S) = H_\mu((X^{(i)})_{i=1}^n|X_S = x_S) \le \sum_{i=1}^n H_\mu(X^{(i)}|X_S = x_S) = H_{\nu_{x_S}}(X|X_S = x_S)$$

Thus by Prop. 2 and Cauchy-Schwarz we can modify the first term in (2.1) as follows

$$\begin{aligned}\mathbb{E}_{X_S}\left[\sum_{i<j} J_{ij}\mathbb{E}_\mu[X_iX_j|X_S]\right] &= \mathbb{E}_{x_S}\left[\sum_{i<j} J_{ij}\,\mathsf{cov}(X_i, X_j|X_S = x_S) + \mathbb{E}_\mu[X_i|X_S = x_S][X_j|X_S = x_S]\right] \\ &\le \sqrt{\sum_{i<j} J_{ij}^2}\sqrt{2\binom{n}{2}\mathbb{E}_{X_S}\mathbb{E}_{i,j}[\mathsf{cov}(X_i, X_j|X_S)^2]} \\ &\qquad + \mathbb{E}_{x_S}\mathbb{E}_\mu[X_i|X_S = x_S][X_j|X_S = x_S] \\ &\le O(\epsilon n\|J\|_F) + \mathbb{E}_{x_S}\left[\mathbb{E}_\mu[X_i|X_S = x_S]\mathbb{E}_\mu[X_j|X_S = x_S]\right].\end{aligned}$$

Here, the final inequality arises from Prop. 2. Note that for any $i, j \notin S$, we have

$$\mathbb{E}_{\nu_{x_S}}[X_iX_j] = \mathbb{E}_{\nu_{x_S}}[X_i]\mathbb{E}_{\nu_{x_S}}[X_j] = \mathbb{E}_\mu[X_i|X_S = x_S]\mathbb{E}_\mu[X_j|X_S = x_S].$$

Bringing all terms together, we have

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

We have the mean field approximation of free energy inside the expectation, hence upon choosing $\epsilon = \frac{1}{n^{1/3}\|J\|_F^{1/3}}$ we obtain

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

as desired. □

Remark 2.3. Jain et al. find tight subexponential time algorithms for achieving error to within $\|J\|_F^{2/3}n^{2/3}$ using Sherali-Adams relaxations when $\|J\|_F^2 = o(n)$ and showed corresponding hardness under gap-ETH. We refer the interested reader to Theorem 1.2 of [JKR19].

2.3 Theorem 1 is generally tight

Beyond the mean-field approximation with product distributions, we find that under fairly mild conditions on the types of distributions, we cannot obtain better bounds on the error of our free energy estimator. In this section, we will provide a construction of Ising models for which the error is $\Omega((n\|J\|_F)^{2/3})$ for a more general class of distributions given that $\|J\|_F = \Theta(\sqrt{n})$.

Theorem 2 ([JKM18]). *Let $(\mathcal{Q}_n)_{n=0}^{\infty}$ be a sequence of families of distributions each on $\{\pm 1\}^n$ which is 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 $X_{[n]\setminus i}$ under Q given $X_i = x_i$, is in $\mathcal{Q}_{n-1}$.*
2. *Taking products: if $Q_1 \in \mathcal{Q}_m$ and $Q_2 \in \mathcal{Q}_n$ then $Q_1 \times Q_2 \in \mathcal{Q}_m \times \mathcal{Q}_n = \mathcal{Q}_{m+n}$.*

If we take out the class of probability distributions induced by Ising models while maintaining the family's properties, there exists Ising models $(J_i)_{i=1}^{\infty}$ of increasing size n_i, with true distribution μ_{J_i} such that

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

where $Q_{n_i} \triangleq \mathsf{arg\,min}_{Q\in\mathcal{Q}_{n_i}} \mathbf{KL}(Q, \mu_{J_i})$.

Proof of Theorem 2. Consider the Ising model with coupling matrix $J \in \mathbb{R}^{k\times k}$ with true Boltzmann Gibbs distribution μ_J. Let $Q \triangleq \mathsf{arg\,min}_{Q\in\mathcal{Q}_k} \mathbf{KL}(Q|\mu_J)$. The premise of this construction will be to duplicate this model for all $m \in \mathbb{Z}^+$ to construct our sequence of Ising models. For any m, consider the Ising model on $n \triangleq mk$ nodes with a block diagonal coupling matrix consisting of copies of J.

$$J'_m \triangleq \begin{bmatrix} J & 0 & \cdots & 0 \\ 0 & J & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J \end{bmatrix}$$

We claim that $Q^{\otimes m} \in \mathcal{Q}_n$ is the closest distribution $\mu_{J'_m}$. Assume that there exists a closer distribution $Q_{J'_m} \in \mathcal{Q}_n$. Upon conditioning on the last $k(m-1)$ variables, chain rule for $\mathbf{KL}$ divergence immediately gives that there exists some distribution $\widetilde{Q}'_J$ on $\{\pm 1\}^k$, which must be closer to μ_J than Q_J. Since $\widetilde{Q}'_J \in \mathcal{Q}_k$ by the first property of our theorem, we have contradicted our original assumption.

To conclude, note that

$$\inf_{Q\in\mathcal{Q}_n} \mathbf{KL}(Q\|\mu_{J'_m}) \geq m\mathbf{KL}(Q_J\|\mu_J)$$

Hence we have a linear relationship between the different levels of our Ising model. This implies that $\inf_{Q\in\mathcal{Q}_n} \mathbf{KL}(Q\|\mu_{J'_m}) = \Theta(n)$. Noting our assumption $\|J\|_F = \Theta(\sqrt{n})$ gives $\Omega(n) = n^{2/3}\|J\|_F^{2/3}$ as desired. □

3 More on correlation rounding: application in MaxCut

In this last section we digress from Ising models and present an interesting result from [RT12] concerning correlation rounding on MaxCut.

3.1 Mean Field Approximation and MaxCut

Let us establish the connection between MaxCut, a classical NP-hard (NP-complete in the decision variant) problem, and estimating the free energy of graphical models. The problem of MaxCut is defined as follows: given a graph G, split the nodes into two parts A and B such that the number of edges connecting nodes in A to nodes in B is maximized. One can reformulate finding a cut as assigning $x_i \in \{\pm 1\}$ to each node where $x_i x_j = -1$ when the edge ij is among the cut and $x_i x_j = 1$ otherwise. Hence we have

$$\mathsf{MaxCut} = \sum_{ij \in E} x_i^* x_j^*, \text{ where } x^* = \mathsf{arg\,min} \sum_{ij \in E} x_i x_j.$$

Using a similar idea, we can also formulate the objective for MaxCut as

$$\mathsf{MaxCut} = \sup_{\mu \in \mathcal{P}(\{\pm 1\}^n)} m \cdot \mathbb{E}_{\mu, ij \in E} \left[\frac{1}{4} (X_i - X_j)^2 \right].$$

To relate MaxCut back to the Ising model, set $J_{ij} = -\frac{\beta n}{m} \mathbb{1}_{ij \in E}$ where β is some inverse temperature. For any given graph G with n nodes and $m \in \omega(n)$ edges we have the following Gibbs distribution:

$$\mathbb{P}_G(x) \propto \mathsf{exp} \left(-\beta \frac{n}{m} \sum_{ij \in E} x_i x_j \right).$$

Firstly, it is easy to notice that the MaxCut assignment x^* is equivalent to the Max-A-Posterior assignment of $\mathbb{P}_G$. However, this property itself is not as useful to us. Instead, consider plugging in the delta distribution $\mu = \delta_{x^*}$ to (1.4), one has that (since $H(\delta) = 0$):

$$\mathcal{F} \geq -\frac{\beta n}{m} \sum_{E} x_i^* x_j^*.$$

But on the other hand, a simple bound on the entropy $H(\mu) \leq n$ leads to:

$$\mathcal{F} \leq \max_{\mu} \mathbb{E}_{\mu} \left[-\frac{\beta n}{m} \sum_{E} x_i x_j \right] + n.$$

Combining the above, one has that (let $M = \mathsf{max}_x(-\sum_E x_i x_j)$):

$$\frac{1}{\beta n} \mathsf{log}\, Z = \frac{1}{\beta n} \mathcal{F} \in \left[M, M + \frac{1}{\beta} \right].$$

This suggests that, when one can estimate the free energy with sufficiently low temperature, one can also approximate MaxCut with a small additive on the corresponding graph, and vice versa.

We will make an extra note on the parameter regime here. In the problem of MaxCut, the interesting regime to our studies lies in the dense case $m \in \omega(n)$ with a super-constant average degree. As a result, the norm of the coupling matrix can be written as $\|J_G\|_F^2 \asymp \beta^2 \frac{n^2}{m}$. Therefore, the condition that $\|J\|_F \in o(\sqrt{n})$ in our work, which may initially appear confusing, can be thus interpreted as MaxCut on dense graphs and constant optimality ratio. For a survey of solving MaxCut in the dense case, we refer to Section 2.2 of [JKR19].

3.2 Main result

We present a direct algorithm based on convex relaxation that achieves the following result:

Theorem 3. *Suppose a d-regular graph G with n nodes and $m \in \omega(n)$ edges such that the normalized adjacency matrix A has eigenvalues $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ where $|\lambda_2|, |\lambda_n| < \varepsilon < 0.1$. Then for any $\delta > 0$ there exists a $n^{O(\delta^{-2})}$-time algorithm which outputs Cut_δ such that*

$$\mathsf{Cut}_\delta \geq \mathsf{MaxCut} - (\varepsilon + \delta)m.$$

Remark 3.1. Here our normalized adjacency matrix A is defined as d^{-1} multiplied by the standard adjacency matrix so that A is doubly stochastic. Furthermore, the restriction that G is d-regular is not crucial to the argument and could be dropped with some extra work.

Let us first examine the result. Note that by randomly selecting the cut we get at least $\mathsf{MaxCut} \geq m/2$, so this indicates $\mathsf{Cut}_\delta \geq (1 - O(\delta))\,\mathsf{MaxCut}$ if $\delta \gtrsim \varepsilon$ is chosen. In practice, spectral concentration results typically bound ε to be $o_d(1)$ (for instance, in a random regular graph or an Erdős-Rényi graph, see [LLV17][TY19]). Therefore, in all practicality, our result can be interpreted as $n^{O(\delta^{-2})}$-time algorithm for $(1 - O(\delta))$-approximation for any constant $\delta > 0$.

3.3 The convex relaxation

Let us first express MaxCut as an optimization problem. Recall that we can write the objective as:

$$\mathcal{C}^* = \frac{1}{m}\mathsf{MaxCut} = \sup_{\mu \in \mathcal{P}} \mathbb{E}_{ij \in E}\left[\frac{1}{4}(X_i - X_j)^2\right]$$

where $\mathcal{P}$ is defined to be the class of singleton (delta) distributions where each X_i is fixed -1 or 1. As this optimization is computationally intractable in its worst form, it is natural to consider a convex relaxation to this objective. Specifically, we will consider:

$$\mathcal{C}^* \leq \mathcal{C}_k = \sup_{\widetilde{\mu} \in \mathcal{SA}_{k+1}} \widetilde{\mathbb{E}}_{ij \in E}\left[\frac{1}{4}(X_i - X_j)^2\right] \tag{3.1}$$

where $\mathcal{SA}_{k+1}$ represents the Sherali-Adams relaxation ([CT12]) to the $(k+1)$th level. We will omit the technical details of this relaxation technique except for the following crucial lemma:

Lemma 4 ([CT12]). *Any feasible solution to the t-th level Sherali-Adams relaxation is equivalent to a family of distributions $\{\mathcal{D}(S)\}_{|S| \leq t+1}$ such that they are locally consistent.*

Intuitively, the above says that we only need to solve for all the $(k+2)$-tuple joint marginals that are locally consistent by solving this program. Moreover, any joint $(k+2)$ indices make a real distribution. In the context of correlation rounding, however, (3.1) does not suffice via the original convex relaxation, as $\widetilde{\mu}$ itself does not give us a lot of useful information concerning the error bound. Instead, we will consider a slightly modified objective as follows:

$$\overline{\mathcal{C}}_k = \sup_{\widetilde{\mu} \in \mathcal{SA}_{k+1}} \mathbb{E}_{K \in [k]} \mathbb{E}_{|S|=K} \widetilde{\mathbb{E}}_{ij \in E, X_S \sim \widetilde{\mu}_S}\left[\frac{1}{4}(X_i - X_j)^2 \mid X_S\right] \tag{3.2}$$

Firstly, we note that the objective in (3.2) is still at least $\mathcal{C}^*$ as any delta measure is still incorporated in this class. Furthermore, we note that solving this relaxation can be turned into solving a convex program with $n^{O(k)}$ variables and $\mathsf{poly}(n^{O(k)})$ LP constraints, which can be done via standard

Algorithm: MaxCut via convex relaxation.

Input: graph G via A; parameter δ, ε.

1. Solve $\widetilde{\mu}$ optimizing (3.2) via e.g. the ellipsoid method. Let $k = \delta^{-2}$ and sample $K \sim [k]$, $S \sim \binom{[n]}{K}$ uniformly at random. Sample $X_S = x_S \sim \widetilde{\mu}_S$.
2. Sample $X_i = x_i \sim \widetilde{\mu}_{S \cup \{i\}} | X_S = x_S$ independently for each $i \notin S$.

Output: $\{x_i\}_i$, the cut assignment.

Figure 1: Algorithm for MaxCut on an expander-like graph.

ellipsoid methods in $\mathsf{poly}(n^{O(k)}, \log \eta^{-1})$-time within additive error η. We omit the dependence on η as it can be easily incorporated within the δ component in Theorem 3.

3.4 Average-case correlation rounding

Let us examine how correlation rounding can be employed in our application here. For our purposes here, we need a slightly stronger lemma than Proposition 2:

Lemma 5 (Average-case correlation rounding, [RT12]). *Let $X_1, \ldots, X_n$ be a collection of $\{\pm 1\}$ random variables. Then for any ℓ, one has that:*

$$\mathbb{E}_{L \in [\ell]} \mathbb{E}_{S \in \binom{[n]}{L}} \mathbb{E}_{X_S} \mathbb{E}_{u,v} \left[\mathsf{cov}(X_u, X_v \mid X_S)^2 \right] \lesssim \ell^{-1}$$

Proof Sketch. Consider the process on the potential Φ similar to the proof of Proposition 2, where we do the conditioning operation for ℓ steps choosing indices $i_1, i_2, \ldots, i_\ell$ along the way. Let $S_L = \{i_t\}_{t=1}^{L}$. For any sample path, take $L \in [\ell]$ uniformly at random one has that:

$$\ell^{-1} \gtrsim \mathbb{E}_L \Phi(\mu | X_{S_L}) - \Phi(\mu | X_{S_{L+1}})$$

because $\Phi(\mu | X_{S_L})$ strictly decreases with L from the data-processing inequality. And therefore, taking expectation over everything we get the desired claim. □

3.5 Proof of Theorem 3

We are now in place to complete the proof that the algorithm in Figure 1 does a good approximation in expectation. First we note this simple lemma.

Lemma 6. *The following is true for any $(X_1, X_2, \ldots, X_n)$ such that $\frac{1}{n^2} \sum_{i,j} \mathsf{cov}(X_i, X_j) \leq \delta$ and G satisfy the conditions of Theorem 3:*

$$\frac{1}{m} \sum_{ij \in E(G)} \mathsf{cov}(X_i, X_j) \leq \delta + \varepsilon.$$

Proof. Let $M \in \mathbb{R}^{n\times n}$ be the covariance matrix such that $M_{ij} = \mathsf{cov}(X_i, X_j)$, then one has that:

$$\begin{aligned}\frac{1}{m}\sum_{ij\in E(G)} \mathsf{cov}(X_i, X_j) &= \left\langle \frac{1}{n}A, M \right\rangle \\ &= \left\langle \frac{\mathbb{1}\mathbb{1}^\top}{n^2}, M \right\rangle + \left\langle \frac{1}{n}\left(A - \frac{\mathbb{1}\mathbb{1}^\top}{n}\right), M \right\rangle\end{aligned}$$

Note that the first term $\left\langle \frac{\mathbb{1}\mathbb{1}^\top}{n^2}, M \right\rangle \le \delta$ is by condition, and the second term can be bounded as:

$$\left\langle \frac{1}{n}\left(A - \frac{\mathbb{1}\mathbb{1}^\top}{n}\right), M \right\rangle \le \frac{1}{n}\lambda_{\mathsf{max}}\left(A - \frac{\mathbb{1}\mathbb{1}^\top}{n}\right)\cdot \mathsf{Tr}(M) \le \varepsilon$$

since $\mathsf{Tr}(M) = n$. This concludes the proof. □

To finish up the proof, we still need to show how to construct the cut given a $\widetilde{\mu}$ from optimizing (3.2) and Proposition 2. From Cauchy-Schwartz over Lemma 5, we see that:

$$\mathbb{E}_{K\in[k]}\mathbb{E}_{S\in\binom{[n]}{K}}\mathbb{E}_{X_S}\mathbb{E}_{u,v}\left|\mathsf{cov}(X_u, X_v \mid X_S)\right| \lesssim k^{-1/2}. \tag{3.3}$$

At a high level, we will consider Step 1 in Figure 1 as fixed for now and assume the global correlation rounding inequality (Lemma 5), which is true in expectation over Step 1. Then, we will show that sampling from Step 2 on average gives a good cut by bounding the difference between the relaxation ($\widetilde{E}[X_iX_j|X_S]$) and expected cut ($\widetilde{\mathbb{E}}[X_i|X_S]\widetilde{\mathbb{E}}[X_j|X_S]$), which happens to differ by exactly the covariance!

Formally, consider Step 2 in Figure 1: with expectation over randomness from sampling at Step 1, (3.3) can be re-written as:

$$\mathbb{E}_{\text{Step 1}}\left[\frac{1}{n^2}\sum_{i,j}\left|\mathsf{cov}(X_i, X_j|X_S)\right|\right] \lesssim k^{-1/2} \asymp \delta$$

since after fixing X_S the relevant terms disappear, and we may assume that $k \in o(n)$. Furthermore, the discrepancy in the expected cut size from this algorithm is conveniently:

$$\begin{aligned}\mathbb{E}[\overline{\mathcal{C}} - \mathcal{C}_{\mathsf{alg}}] &= \mathbb{E}_{\text{Step 1}}\left[\frac{-1}{m}\sum_{ij\in E}\widetilde{\mathbb{E}}[X_iX_j|X_S] - \widetilde{\mathbb{E}}[X_i|X_S]\widetilde{\mathbb{E}}[X_j|X_S]\right] \\ &= \mathbb{E}_{\text{Step 1}}\left[\frac{-1}{m}\sum_{ij\in E}\mathsf{cov}(X_i, X_j|X_S)\right] \le \delta + \varepsilon\end{aligned}$$

where the last inequality follows from Lemma 6 by treating the distribution on X_S as fixed from Step 1. Combined with the result following (3.2) that $\overline{\mathcal{C}} \ge \mathcal{C}^*$ the true MaxCut, our proof for Theorem 3 is complete (the algorithm performs well in expectation over Step 1 and Step 2).

4 Discussions

We conclude this note with some followup results as well as open directions for future work.

4.1 Followup works

We survey some results extending from those covered in this note. In [Eld20], the authors extend the Frobenius norms bounds to be based on the general Schatten-p norm $\|J\|_{S_p} := \left(\sum_{i\in[n]} |\lambda_i|^p\right)^{1/p}$. In particular, they derived a $O\left(\frac{1+p}{p}\left(n\|J\|_{S_p}\right)^{\frac{p}{1+p}}\right)$ bound on the mean-field approximation error. When $p = 2$, where the Schatten norm reduces to the Frobenius norm, this recovers Theorem 1. However, for optimal p one usually arrives at bounds that are almost dimension-free (see the examples therein). In a succeeding work [Aug21], the authors proved a bound of $\sqrt{n}\|J\|_F$, which also subsequently improves upon Theorem 1.

In [KLR22], the authors constructed an estimation algorithm (Algorithm 3 and Theorem 1.1 therein) by combining variational inference and Glauber dynamics. When restricted to Ising models like the setting of our expository, they give a $\frac{(n\|J\|_{\mathrm{op}})^{O(\|J\|_F^2)}}{\varepsilon^2}$ runtime bound for estimation up to additive ε error on $\log Z$, which is stronger than the convex-relaxation algorithm in [JKR19] (while we did not cover the algorithm here, it is in spirit similar to our MaxCut relaxation). See Remark C.5 therein for further discussions.

4.2 Open directions

We conclude by mentioning open directions related to the results in this expository.

1. Note in the final step of Theorem 1 we chose ϵ to obtain bounds of $O((n\|J\|_F)^{2/3})$. Furthermore, our lower bounds say that mean-field error is at least $\Omega(n)$ if $\|J\|_F^2 \in \Omega(n)$. The natural question here is whether one can improve the upper bound in the forms like $|\mathcal{F}^* - \mathcal{F}| \in O(n^{1-\alpha}\|J\|_F^{2\alpha})$ (see [Aug21] for $\alpha = 1/2$).

2. The runtime complexity of MaxCut convex relaxation $n^{O(\delta^{-2})}$ arises as an artifact from correlation rounding and the fact that covariance decays no quicker than $1/\sqrt{k}$. While correlation rounding may not be improved for spin glasses (Section 5 in [JKR19]), could it be that MaxCut can be approximately solved using $o(\delta^{-2})$ rounds of relaxations (Sherali-Adams, Sum-of-Squares)?

References

[AO15] Sarah R. Allen and Ryan O'Donnell. "Conditioning and covariance on caterpillars". In: *2015 IEEE Information Theory Workshop (ITW)*. 2015, pp. 1–5. DOI: 10.1109/ITW.2015.7133115.

[Aug21] Fanny Augeri. "A transportation approach to the mean-field approximation". In: *Probability Theory and Related Fields* 180.1-2 (2021), pp. 1–32.

[BRS11] Boaz Barak, Prasad Raghavendra, and David Steurer. "Rounding semidefinite programming hierarchies via global correlation". In: *2011 ieee 52nd annual symposium on foundations of computer science*. IEEE. 2011, pp. 472–481.

[CT12] Eden Chlamtac and Madhur Tulsiani. "Convex relaxations and integrality gaps". In: *Handbook on semidefinite, conic and polynomial optimization* (2012), pp. 139–169.

[Eld20] Ronen Eldan. "Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation". In: *Probability Theory and Related Fields* 176.3-4 (2020), pp. 737–755.

[IV00] Dmitry Ioffe and Yvan Velenik. "A note on the decay of correlations under δ-pinning". In: *Probability theory and related fields* 116.3 (2000), pp. 379–389.

[JKM18] Vishesh Jain, Frederic Koehler, and Elchanan Mossel. "The Mean-Field Approximation: Information Inequalities, Algorithms, and Complexity". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 1326–1347.

[JKR19] Vishesh Jain, Frederic Koehler, and Andrej Risteski. "Mean-field approximation, convex hierarchies, and the optimality of correlation rounding: a unified perspective". In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 2019, pp. 1226–1236.

[KLR22] Frederic Koehler, Holden Lee, and Andrej Risteski. "Sampling approximately low-rank Ising models: MCMC meets variational methods". In: *Conference on Learning Theory*. PMLR. 2022, pp. 4945–4988.

[LLV17] Can M Le, Elizaveta Levina, and Roman Vershynin. "Concentration and regularization of random graphs". In: *Random Structures & Algorithms* 51.3 (2017), pp. 538–561.

[PW23] Y Polyanskiy and Y Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2023+.

[RT12] Prasad Raghavendra and Ning Tan. "Approximating CSPs with global cardinality constraints using SDP hierarchies". In: *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*. SIAM. 2012, pp. 373–387.

[SJ89] Alistair Sinclair and Mark Jerrum. "Approximate counting, uniform generation and rapidly mixing Markov chains". In: *Information and Computation* 82.1 (1989), pp. 93–133. ISSN: 0890-5401. DOI: https://doi.org/10.1016/0890-5401(89)90067-9. URL: https://www.sciencedirect.com/science/article/pii/0890540189900679.

[SS12] Allan Sly and Nike Sun. "The computational hardness of counting in two-spin models on d-regular graphs". In: *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*. IEEE. 2012, pp. 361–369.

[TY19] Konstantin Tikhomirov and Pierre Youssef. "The spectral gap of dense random regular graphs". In: *The Annals of Probability* 47.1 (2019), pp. 362–419. DOI: 10.1214/18-AOP1263. URL: https://doi.org/10.1214/18-AOP1263.