Proceedings of Machine Learning Research vol 178:1–44, 2022 35th Annual Conference on Learning Theory

Sampling Approximately Low-Rank Ising Models: MCMC meets Variational Methods

Frederic Koehler FKOEHLER@STANFORD.EDU
Stanford University

Holden Lee HOLDEN.LEE@DUKE.EDU
Duke University

Andrej Risteski ARISTESK@ANDREW.CMU.EDU
Carnegie Mellon University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We consider Ising models on the hypercube with a general interaction matrix J, and give a polynomial time sampling algorithm when all but $O(1)$ eigenvalues of J lie in an interval of length one, a situation which occurs in many models of interest. This was previously known for the Glauber dynamics when *all* eigenvalues fit in an interval of length one; however, a single outlier can force the Glauber dynamics to mix torpidly. Our general result implies the first polynomial time sampling algorithms for low-rank Ising models such as Hopfield networks with a fixed number of patterns and Bayesian clustering models with low-dimensional contexts, and greatly improves the polynomial time sampling regime for the antiferromagnetic/ferromagnetic Ising model with inconsistent field on expander graphs. It also improves on previous approximation algorithm results based on the naive mean-field approximation in variational methods and statistical physics.

Our approach is based on a new fusion of ideas from the MCMC and variational inference worlds. As part of our algorithm, we define a new nonconvex variational problem which allows us to sample from an exponential reweighting of a distribution by a negative definite quadratic form, and show how to make this procedure provably efficient using stochastic gradient descent. On top of this, we construct a new simulated tempering chain (on an extended state space arising from the Hubbard-Stratonovich transform) which overcomes the obstacle posed by large positive eigenvalues, and combine it with the SGD-based sampler to solve the full problem.

1. Introduction

An Ising model is a probability distribution on the hypercube $\{\pm 1\}^n$ of the form

$$p_{J,h}(\sigma) = \frac{1}{Z} \exp\left(\frac{1}{2}\langle \sigma, J\sigma\rangle + \langle h, \sigma\rangle\right)$$

where the normalizing constant Z is known as the *partition function*. The closely related problems of estimating the partition function Z and sampling from the Ising model are fundamental computational problems, both due to their central theoretical significance as well a plethora of applications—see for example Mezard and Montanari (2009); Talagrand (2010); Wainwright and Jordan (2008); Jerrum and Sinclair (1996); Hinton (2012); Murphy (2012). While computing the partition function Z exactly is #P-hard (Jerrum and Sinclair, 1993), and approximating it is NP-hard (see e.g., Sly and Sun (2012); Galanis et al. (2016)), a vast amount of work has been done to understand and characterize situations where this task is computationally tractable.

One of the dominant approaches in both theory and practice to sample from such models is the *Glauber dynamics* or *Gibbs sampler*. This is a Markov chain that at each step, resamples the spin of one coordinate from its conditional distribution. In general, this chain is expected to mix under appropriate assumptions on the weakness of the interactions in the model (e.g., presence of correlation decay, or uniqueness of the corresponding Gibbs measure on the tree). In certain special cases, the point at which the Glauber dynamics stops mixing rapidly is also exactly where sampling becomes hard: famously, this is the case for the antiferromagnetic Ising model on the worst-case d-regular graph (see e.g., Sly and Sun (2012); Chen et al. (2020)). However, this is not the case in general—there are many examples where Glauber dynamics fails to mix but other methods succeed to approximate the partition function and/or sample; see e.g., Jerrum and Sinclair (1993); Borgs et al. (2020); Risteski (2016); Guo and Jerrum (2017) for a few examples.

Variational methods are the main alternative to MCMC (Markov Chain Monte Carlo) methods in practice. In general, variational methods attempt to reduce to problem of computing the partition function to solving an optimization problem—see e.g., Wainwright and Jordan (2008); Mezard and Montanari (2009) for further background. Importantly, the strengths and limitations of variational methods are complementary to those of Glauber dynamics. Unlike Markov chain methods, variational methods are usually based on solving for an *approximation* of the true distribution, and hence may only achieve a comparatively crude approximation to the true distribution—a successful variational approximation may only output a distribution with KL divergence or Wasserstein distance $o(n)$ as opposed to $o(1)$ for the output of a rapidly mixing Markov chain. On the other hand, variational methods often work in both high and low-temperature settings and are closely related to textbook methods for solving low-temperature models, such as the Ising model on a high-dimensional lattice, the Curie-Weiss model, and the Sherrington-Kirkpatrick model (Talagrand, 2010; Mezard and Montanari, 2009; Parisi and Shankar, 1988).

To give a concrete example with strong theoretical guarantees, the *naive mean-field approximation*, which corresponds to approximating the Gibbs measure by a (small mixture of) product measure(s), is probably the most well-known variational method. It has been established that this approximation is in various senses accurate whenever the interaction matrix J has quantitatively low rank (more precisely, when $\|J\|_F^2 = \sum_i \lambda_i(J)^2 = o(n)$): see Basak and Mukherjee (2017); Eldan (2018); Eldan and Gross (2018); Eldan (2020); Augeri (2021) for a few of the works in this area. This condition essentially covers all of the main examples of Ising models where the mean-field approximation is known to be accurate, and for these models it covers both low and high temperature regimes (i.e., both strong and weak couplings). Correspondingly, there are approximation algorithms connected with the naive mean-field approximation (Risteski, 2016; Jain et al., 2018a,b, 2019) which approximate $\log Z$ within $o(n)$ additive error in subexponential time under this assumption (with improving runtime as the rank decreases, and with roughly matching computational lower bounds).

In this work, we seek to achieve the *best of both worlds* and combine the strengths of Glauber dynamics and variational inference. Recently, it was shown (Eldan et al., 2020; Anari et al., 2021) that the Glauber dynamics rapidly mix whenever the eigenvalues of J all lie within an interval of length 1, which is tight due to the example of the Curie-Weiss model (Levin and Peres, 2017). Our main result shows that by using a more sophisticated algorithm, we can sample in polynomial time from any Ising model with a constant number of eigenvalues outside of this interval, a situation which occurs in many examples of interest. To state our result, first note that without loss of generality, we can recenter the bulk of the eigenvalues to $[0, 1]$ by adding a multiple

of the identity to J. We provide an algorithm that samples from an Ising distribution with d_+ eigenvalues bigger than $1 - 1/c, c \in (1, \infty]$, and d_- negative eigenvalues $-\lambda_1, \ldots, -\lambda_{d_-}$ in time $(n \|J\|_{\text{op}})^{O(d_+)} e^{O(c(\lambda_1 + \cdots + \lambda_{d_-}))}$, as well as (multiplicatively) approximate the partition function.

In the special case of low-rank Ising models where the naive mean-field approximation is accurate, this gives a roughly comparable runtime to the previous approximation algorithms for estimating $\log Z$ (e.g., Jain et al. (2019)), while allowing us both to approximate Z much more accurately (within an arbitrary multiplicative factor) and also to sample; see Remark C.5 for further discussion. Our result also allows us to sample from models which are genuinely high-rank, for example the SK model with ferromagnetic interactions in the regime where the bulk has diameter at most 1 (see Section 3) in which case the naive mean-field approximation is known to be very inaccurate (see e.g., Thouless et al. (1977); Jain et al. (2019)). Our general result also continues a long tradition of seeking fixed-parameter tractable algorithms for optimization problems that are "approximately" low rank (Frieze and Kannan, 1996; Oveis Gharan and Trevisan, 2013).

Our techniques take inspiration from both variational and MCMC approaches. We describe them in detail later (see Section 2), but at a high-level our result is based on two key innovations: (1) for positive outlier eigenvalues, a rigorous version of the popular *simulated annealing* (Lovász and Vempala, 2006) and *tempering* heuristics (Marinari and Parisi, 1992), based in part on a decomposition of the measure into a mixture of high-temperature Ising models using the Hubbard-Stratonovich transform (Hubbard, 1959), and (2) for negative eigenvalues, a sampling approach based on importance sampling combined with the efficient solution of a related fixed point equation, which is done by constructing an appropriate (nonconvex) variational problem and running stochastic gradient descent. The key ideas behind both steps are clean and we believe the techniques may be useful for solving other sampling problems of interest.

In addition to this, we provide representative applications of our results to a diverse set of tasks: First, we give an algorithm to sample Ising models (antiferromagnetic or ferromagnetic, and potentially with inconsistent external fields) on expander graphs up to inverse temperature $\beta = O(1/\lambda)$ where λ is the second largest eigenvalue. This is outside the tree uniqueness regime; note that on general graphs, antiferromagnetic Ising is NP-hard past this threshold (Sly and Sun, 2012). Also, even when the model is ferromagnetic, inconsistent external fields make the sampling problem #BIS-hard in general[1]. Relatedly, we give the first results for sampling high-temperature Sherrington-Kirkpatrick models with strong ferromagnetic interactions.

We also show how to sample from a Hopfield network (Hopfield, 1982) with a fixed number of patterns in polynomial time. As an example Bayesian statistics application, we show how to sample from posteriors of mixtures of two Gaussians with symmetric means in fixed dimension. This provides complementary results to (Mou et al., 2019), who consider the same setting in an arbitrary dimension, but instead consider an easier task: sampling from the so-called power posterior of such a mixture—which is derived by weighing the prior substantially more in the Bayes formula for the posterior. More generally, we show how to sample from a regime of a more sophisticated clustering model (the Contextual Stochastic Block Model) with low-dimensional contexts.

1. Our results work in an expanded "high temperature" regime; in contrast algorithms for different #BIS-hard problems work in a low temperature regime by expanding around the ground states (Jenssen et al., 2020; Chen et al., 2021), so these approaches should be naturally complementary when they both apply.

1.1. Main results

Suppose that J is a symmetric matrix. We are interested in computing the partition function $Z_{J,h}$ and sampling from the distribution $P_{J,h}$ over $\{\pm 1\}^n$ given by

$$p_{J,h}(\sigma) = \frac{\exp\left(\frac{1}{2}\langle \sigma, J\sigma\rangle + \langle h, \sigma\rangle\right)}{Z_{J,h}}, \quad \text{where } Z_{J,h} = \sum_{\sigma\in\{\pm 1\}^n} \exp\left(\frac{1}{2}\langle \sigma, J\sigma\rangle + \langle h, \sigma\rangle\right). \quad (1)$$

Our main theorem is the following.

Theorem 1.1 *Let $c \in (1,\infty]$, $\varepsilon \in (0,1)$. Suppose that J is a symmetric matrix such that (1) J has d_+ eigenvalues that are greater than $1 - \frac{1}{c}$, and (2) its negative eigenvalues are $-\lambda_1, \dots, -\lambda_{d_-}$.*

1. *There is an algorithm (Algorithm 3) that with probability $\geq 1 - e^{-n}$, gives a e^{ε}-multiplicative approximation to $Z_{J,h}$ in time $O\left((\|J\|_{\text{op}}\, n)^{O(d_+ + 1)} e^{O(c(\lambda_1 + \cdots + \lambda_{d_-}))}/\varepsilon^2\right)$.*

2. *There is an algorithm (Algorithm 4) to sample from a distribution ε-close in TV-distance to $P_{J,h}$ in time $\left(\|J\|_{\text{op}}\, n \log\left(\frac{1}{\varepsilon}\right)\right)^{O(1+d_+)} e^{O(c(\lambda_1 + \cdots + \lambda_{d_-}))}$.*

Note that we can take $c = \infty$ in the theorem; in this case we assume that J has no negative eigenvalues, i.e., J is positive semi-definite, and we get the simpler bounds $O\left((\|J\|_{\text{op}}\, n)^{O(d_+ + 1)}/\varepsilon^2\right)$ and $\left(\|J\|_{\text{op}}\, n \log\left(\frac{1}{\varepsilon}\right)\right)^{O(1+d_+)}$. Excluding the dependence on $\|J\|_{\text{op}}$, for large positive eigenvalues the runtime only depends on the number of eigenvalues, but for negative eigenvalues, the runtime depends on their magnitude.

When there are n large eigenvalues, our runtime guarantee is similar to brute force[2]; see (Jain et al., 2019) for discussion of why this should be unavoidable under the Exponential Time Hypothesis (ETH). In the extreme case where there is just a single very large negative eigenvalue, it turns out the problem is also computationally hard. This arises from the discrete nature of the hypercube $\{\pm 1\}^n$ and stands in strong contrast to intuition from sampling continuous distributions, where very strong log-concavity is not an obstacle to efficient sampling. We prove the following negative result; see the full theorem (Theorem H.1) for a stronger runtime lower bound for estimating $\log Z$, conditional on the ETH.

Theorem 1.2 (Theorem H.1) *Let $\beta \geq 1$ be arbitrary and fixed. For any $a = (a_1, \dots, a_n) \in \mathbb{Z}^n$, define the Ising model with probability mass function $p_a : \{\pm 1\}^n \to [0,1]$ given by $p_a(\sigma) \propto \exp\left(-\beta n \langle a, \sigma\rangle^2\right)$. If there exists a polynomial time randomized algorithm to approximately sample within TV distance $1/2$ from Ising models of this form for any $a_1, \dots, a_n$, then* NP = RP.

2. Overview of techniques

This section has two parts: in the first, we recall some basic tools which we will use in our analysis. In the second, we give a full overview of our algorithm and the proof of our main result.

2. Note however, that Theorem 1.1 only gives nontrivial guarantees when $d_+ = o\left(\frac{n}{\log n}\right)$; it is an interesting question whether one can remove the $\log n$ factor.

2.1. Technical toolkit

Sampling from Ising models with bounded spectral diameter. As a basic ingredient, we use the following guarantee for Glauber dynamics on Ising models (see also Bauerschmidt and Bodineau (2019); Eldan et al. (2020)):

Theorem 2.1 ((Anari et al., 2021, Theorem 12)) *Let $J \in \mathbb{R}^{n\times n}$ be a symmetric matrix satisfying $0 \preceq J \prec I_n$, $h \in \mathbb{R}^n$ arbitrary. Then we have that:*

1. *The Poincaré and modified Log-Sobolev constants of $P_{J,h}$ are at most $n(1 - \|J\|_{\text{op}})^{-1}$.*
2. *For any $\epsilon > 0$, the discrete-time Glauber dynamics mixes to ϵ total variation distance of $P_{J,h}$ in $O(n \log(n/\epsilon)/(1 - \|J\|_{\text{op}}))$ steps.*

See Appendix A.2 for the definition of the Poincaré and modified log-Sobolev constant.

Hubbard-Stratonovich transform. The component of our algorithm which handles positive spike eigenvalues makes use of the multivariate version of the classical Hubbard-Stratonovich transform (Hubbard, 1959). This transform is commonly used in the analysis of quantum and statistical physics systems and in large deviation theory; for a few examples see (Talagrand, 2010; Bovier and Picco, 1998; Bauerschmidt and Bodineau, 2019; Hsu et al., 2012). The statement is given by Lemma 2.2 below; it is very useful despite its simplicity.

Lemma 2.2 *Let $X \in \mathbb{R}^{m\times n}$ be a matrix with d-dimensional column space V. Let $\sigma \in \mathbb{R}^n$. Then for any $\gamma > 0$,*

$$\exp\left(\frac{\gamma^2}{2}\|X\sigma\|^2\right) = \left(\frac{1}{2\pi\gamma^2}\right)^{d/2} \int_V \exp\left(\left\langle X^\top \mu, \sigma\right\rangle - \frac{1}{2\gamma^2}\|\mu\|^2\right) d\mu.$$

Proof We complete the square to find that

$$\left(\frac{1}{2\pi\gamma^2}\right)^{d/2} \int_V \exp\left(\left\langle X^\top \mu, \sigma\right\rangle - \frac{1}{2\gamma^2}\|\mu\|^2\right) d\mu$$
$$= \left(\frac{1}{2\pi\gamma^2}\right)^{d/2} \exp\left(\frac{\gamma^2}{2}\|X\sigma\|^2\right) \int_V \exp\left(-\frac{1}{2\gamma^2}\left\|\mu - \gamma^2 X\sigma\right\|^2\right) d\mu = \exp\left(\frac{\gamma^2}{2}\|X\sigma\|^2\right)$$

using the formula for the normalizing constant of a Gaussian distribution. ■

2.2. Proof overview

The proof of our main result, Theorem 1.1, combines two modular algorithmic ideas: a grid partitioning and simulated annealing/tempering strategy which handles the large positive eigenvalues, and an optimization and rejection sampling based strategy which handles the negative ones.

We briefly comment on the relation between our techniques and those used in the aforementioned literature on naive mean-field approximation, which do not seem as useful for sampling. In all of those works (algorithmic or non-algorithmic), the primary goal is to estimate $\log Z$ within an additive error which is small compared to n, but essentially always $\omega(1)$ as $n \to \infty$. The main reason for this is that the naive mean-field approximation is simply not accurate to $O(1)$ additive

error even in relatively basic examples (see e.g., Eldan (2020)). On the other hand, in almost all of those works (and also for Dense Max-CSP, e.g. Frieze and Kannan (1996)) the techniques used are general as far as the form of the distribution concerned: e.g., they can handle a log-likelihood which is not a quadratic function but a higher-order polynomial. Our analysis is based on decomposing the spectrum of the interaction matrix, which only seems to makes sense in the Ising case.

2.2.1. LARGE POSITIVE EIGENVALUES: DECOMPOSITION AND SIMULATED TEMPERING

Here we describe our method for sampling from Ising models with large positive eigenvalues. For simplicity, we describe the algorithm when the interaction matrix J is positive semidefinite and return to the general case later.

Warmup: Curie-Weiss model and generalizations. To motivate our approach, we start with a special case: sampling from a rank-one Ising model of the form $p_{ww^\top,0}(\sigma) \propto e^{\langle w,\sigma\rangle^2/2}$. This means the interaction matrix is simply $ww^\top$. A classical example of such a distribution is the *Curie-Weiss* model, in which case $w = \beta\mathbf{1}/\sqrt{n}$ where $\beta \geq 0$ is referred to as the inverse temperature. It is well known (Ellis, 2006; Talagrand, 2010) that the Curie-Weiss model exhibits symmetry breaking in its low temperature phase $\beta > 1$: the distribution becomes close to supported on two clusters of points, one with $\frac{1}{n}\sum_i \sigma_i \approx y$ and an opposite one with $\frac{1}{n}\sum_i \sigma_i \approx -y$ where y is a nontrivial (i.e., nonzero) solution of the fixed point equation $y = \tanh(\beta y)$. Because Glauber dynamics becomes trapped in one of the clusters, it will not mix (Levin and Peres, 2017).

There are many alternative algorithms to sample from the Curie-Weiss model. For example, the random variable $\sum_i \sigma_i$ is an integer between $-n$ and n and it is straightforward to write down its distribution under the Curie-Weiss model explicitly, letting us sample it; this can also be used with a Markov chain decomposition theorem to show mixing up to phase (Madras and Zheng, 2003). *However,* this approach which works well for the Curie-Weiss model does not generalize nicely — for a typical vector w, $\langle w, \sigma\rangle$ will take on 2^n many different values! There are multiple ways to provably sample from ferromagnetic Ising models which apply to Curie-Weiss (Jerrum and Sinclair, 1993; Guo and Jerrum, 2017), but we need to also sample from non-ferromagnetic ones.

We now explain an approach that *will* generalize nicely to rank-one models and beyond. We first describe this as a method to compute the partition function Z, and explain sampling at the end of this section. By applying the Hubbard-Stratonovich transform (Lemma 2.2), we have

$$Z = \sum_{\sigma\in\{\pm1\}^n} e^{\langle w,\sigma\rangle^2/2} = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-y^2/2} \sum_{\sigma\in\{\pm1\}^n} e^{y\langle w,\sigma\rangle}\,dy = \frac{2^n}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-y^2/2}\prod_{i=1}^n \cosh(yw_i)\,dy.$$

This is a one-dimensional integral: it's over an infinite domain, but the term $e^{-y^2/2}$ ensures that larger values of y contribute only a negligible amount to the integral. Hence, we only need to perform an integral over a bounded region which can be done using Riemann summation.

The general case: decomposition and integration. We now consider the much more general case of a positive semidefinite matrix J. We do not want to restrict ourselves to low-rank J, but rather J which have a smaller large number of eigenvalues greater than 1. For this reason, we only apply the Hubbard-Stratonovich transform over the large eigenspaces of J.

To do this, let $c > 0$ be an arbitrary small constant. Using the spectral decomposition of J, we can decompose $J = J^\perp + J^\parallel$ so that $J^\perp$ and $J^\parallel$ are both positive semidefinite, $\|J^\perp\|_{\text{op}} \leq 1 - c$,

and $J^{\|}$ spans the eigenspaces of J above $1-c$, which we denote as $V^{\|}$ with dimension d. Let $J^{\|} = X^\top X$ be an arbitrary factorization; then by an analogous application of the Hubbard-Stratonovich transform (Lemma 2.2) we have

$$\begin{aligned} Z &= \sum_{\sigma\in\{\pm1\}^n} \exp\left(\frac{1}{2}\left\langle \sigma, J^{\perp}\sigma\right\rangle + \langle h,\sigma\rangle\right)\exp\left(\frac{1}{2}\left\|X\sigma\right\|^2\right) \\ &= \left(\frac{1}{2\pi}\right)^{d/2}\sum_{\sigma\in\{\pm1\}^n} \exp\left(\frac{1}{2}\left\langle \sigma, J^{\perp}\sigma\right\rangle + \langle h,\sigma\rangle\right)\int_{V^{\|}}\exp\left(\left\langle X^\top\mu^{\|},\sigma\right\rangle - \frac{1}{2}\left\|\mu^{\|}\right\|^2\right)d\mu^{\|} \\ &= \left(\frac{1}{2\pi}\right)^{d/2}\int_{V^{\|}}\exp\left(-\frac{1}{2}\left\|\mu^{\|}\right\|^2\right)\sum_{\sigma\in\{\pm1\}^n}\exp\left(\frac{1}{2}\left\langle \sigma, J^{\perp}\sigma\right\rangle + \left\langle h+X^\top\mu^{\|},\sigma\right\rangle\right)d\mu^{\|}. \quad (2) \end{aligned}$$

We see the resulting integral is now over a d-dimensional subspace; just like the example, the integrand has a damping term $\exp\left(-\frac{1}{2}\left\|\mu^{\|}\right\|^2\right)$ which allows us to truncate it to a bounded domain while changing the integral by only a small amount. Each of the integrands involves a sum over exponentially many $\sigma\in\{\pm1\}^n$, but we can recognize this sum as the partition function of an Ising model with interaction matrix $J^{\perp}$. Since $J^{\perp}$ has no large eigenvalues, and we can sample from this class of models using Glauber dynamics (Theorem 2.1), we can approximate the corresponding partition function using a relatively standard reduction from sampling to integration (see e.g., Bezáková et al. (2008); this reduction is via a form of simulated annealing, not to be confused with the related but different concept of simulated tempering described later). Finally, using Riemann summation to actually compute the integral gives the estimate of Z.

The simulated tempering chain: sampling with exponentially small error. In principle, given the previous result for approximating the partition function, we could apply standard reductions from approximate counting to sampling in order to approximately sample from the Ising model. This would be quite suboptimal, because the running time of such an algorithm would depend polynomially on the error parameter ϵ (desired total variation distance to the true distribution). In comparison, MCMC methods, when they work, generally depend logarithmically on the error parameter ϵ and we would like our algorithm to have this property too.

To achieve the desired logarithmic dependence on $1/\varepsilon$, we construct a new Markov chain. The first step is to observe that the formula (2) we derived comes with a simple probabilistic interpretation: it can be understood as a decomposition of the original Ising model into a mixture of high-temperature Ising models with additional external field $X^\top\mu^{\|}$. The associated joint distribution over the pair $(\sigma,\mu^{\|})$ is

$$p(\sigma,\mu^{\|}) \propto \exp\left(\frac{1}{2}\left\langle \sigma, J^{\perp}\sigma\right\rangle + \left\langle h+X^\top\mu^{\|},\sigma\right\rangle - \frac{1}{2}\left\|\mu^{\|}\right\|^2\right). \quad (3)$$

With this understanding, all we need to do is construct a Markov chain which can sample quickly on the joint $(\sigma,\mu^{\|})$ space. However, a standard Metropolis-Hastings sampler has the same issue as the original Glauber dynamics: the joint distribution in $(\sigma,\mu^{\|})$ space is multimodal just like the original distribution.

The key to solving this problem is to use a faster chain based on *simulated tempering* (Marinari and Parisi, 1992). We actually define the Markov chain on a further expanded state space of

$(\ell, \sigma, \mu^{\|})$ where ℓ is an additional *temperature* variable, so that the chain mixes to a distribution which conditional on the temperature ℓ being at its "coldest" setting is the desired distribution. The point is that the chain mixes rapidly at the "hottest" temperature, which combined with a choice of temperature schedule where distributions at adjacent distributions have constant overlap, provides a bridge between the different modes at the colder temperature. We actually consider a variant of simulated tempering where we approximately equalize the probability for each grid cell so that they will all be visited—this can be thought of as a Markov chain analogue of grid search—with a final step of importance sampling to attain the right probabilities.

Simulated tempering is a beautiful idea, but it isn't always guaranteed to work: indeed, Marinari and Parisi (1992) proposed their original simulated tempering chain exactly for the purpose of sampling from Ising models, but it does not come with a mixing time guarantee (and obviously, no sampling method will work for Ising models which are computationally hard to sample (Sly and Sun, 2012)). In our setting, we can establish a Poincaré inequality and prove rapid mixing by using a *Markov chain decomposition theorem* (Madras and Randall, 2002; Ge et al., 2018). Such a decomposition theorem allows us to conclude fast mixing once we show mixing within each grid cell as well as a "coarse-grained" chain where each grid cell is considered as a single state. Mixing within each grid cell is immediate from the fact that for fixed $\mu^{\|}$, Glauber dynamics for $p(\sigma, \mu^{\|})$ mixes rapidly, and mixing of the coarse-grained chain follows from equalization of the probabilities of grid cells and overlap of distributions at adjacent temperatures.

2.2.2. LARGE NEGATIVE EIGENVALUES: NONCONVEX VARIATIONAL PROBLEM AND IMPORTANCE SAMPLING.

Warmup example. To explain our method of handling large negative eigenvalues, it helps to start with a much easier special case of the argument. Consider $p_{-ww^\top,0}(\sigma) \propto e^{-\langle w,\sigma\rangle^2/2}$ for $\sigma \in \{\pm 1\}^n$, i.e., a rank one Ising model with interaction matrix $-ww^\top$. We claim that we can sample from $P_{-ww^\top,0}$ using *rejection sampling*: (1) first, sample $\sigma_0 \sim \mathsf{Uniform}(\{\pm 1\}^n)$, and then (2) with probability $e^{-\langle w,\sigma_0\rangle^2/2}$ output $\sigma = \sigma_0$, and otherwise restart with step (1). From the definition, it's clear that this process draws a sample from P; the only concern is how long it takes. The runtime is a geometric random variable with parameter $p = \mathbb{E}_{\sigma_0} e^{-\langle w,\sigma_0\rangle^2/2}$ and using Jensen's inequality we have $p \geq e^{-\mathbb{E}_{\sigma_0}\langle w,\sigma_0\rangle^2/2} = e^{-\|w\|^2/2}$. Hence, the expected runtime is $1/p = \exp(\|w\|^2/2)$ (constant time provided $\|w\| = O(1)$).

This is an artificially simple example because: (1) the Ising model we considered had no positive eigenvalues, and (2) there was no external field. In all of the cases of serious interest, rejection sampling from the uniform distribution has extremely bad runtime (exponential in dimension n). However, generalizing this example leads us naturally to a more sophisticated algorithm which works more generally.

The general importance sampling argument and fixed point equation. The actual problem we need to solve is this: sample from an Ising model with external field h and interaction matrix J with the following structure: $J = J_+ - J_-$ with $0 \preceq J_+ \preceq 1 - \frac{1}{c}$ and $0 \preceq J_-$ with small trace. (We use the previous annealing argument to eliminate any larger positive eigenvalues.) We will let $Q(\sigma) \propto e^{\frac{1}{2}\langle\sigma,J\sigma\rangle + \langle h,\sigma\rangle}$ denote the Ising model we ultimately want to sample from.

To have any hope of succeeding with the rejection sampling approach, we need a smart proposal distribution. Since we have a sampler for the Ising model $P_{J_+,h}(\sigma) \propto e^{\frac{1}{2}\langle\sigma,J_+\sigma\rangle + \langle h,\sigma\rangle}$, this would be an obvious choice of proposal distribution. However, this is a bad idea: the distribution $p_{J_+,h}$

and the target distribution many be concentrated around different regions[3], in which case rejection sampling will perform poorly. A smarter choice is to consider a tilted proposal distribution with additional external field $\lambda \in \mathbb{R}^n$, i.e., an Ising model of the form $P_{J_+,h+\lambda}(\sigma) \propto e^{\frac{1}{2}\langle \sigma, J_+\sigma\rangle + \langle h+\lambda,\sigma\rangle}$. Then the relative density satisfies $\frac{dQ}{dP_{J_+,h+\lambda}}(\sigma) \propto e^{-\frac{1}{2}\langle \sigma, J_-\sigma\rangle - \langle \lambda,\sigma\rangle}$ and if we specifically consider tilts of the form $\lambda = -J_-\mu$, we can complete the square to write

$$\frac{dQ}{dP_{J_+,h-J_-\mu}}(\sigma) = \frac{1}{Z(\mu)} e^{-\frac{1}{2}\langle \sigma-\mu, J_-(\sigma-\mu)\rangle}$$

where $Z(\mu) := \mathbb{E}_{P_{J_+,h-J_-\mu}}[e^{-\frac{1}{2}\langle \sigma-\mu, J_-(\sigma-\mu)\rangle}]$ is the normalizing constant. Note that $Z(\mu) \le 1$ since J_- is positive semidefinite. To lower bound $Z(\mu)$, analogous to the "warmup example," we can apply Jensen's inequality, which gives

$$\log Z(\mu) \ge -\mathbb{E}_{P_{J_+,h-J_-\mu}}[\langle \sigma-\mu, J_-(\sigma-\mu)\rangle/2] = -\langle J_-, \mathbb{E}_{P_{J_+,h-J_-\mu}}[(\sigma-\mu)(\sigma-\mu)^\top]\rangle. \quad (4)$$

For arbitrary μ, the right hand side of this inequality does not seem particularly tractable. *However*, if were fortunate enough to choose μ which is a solution of the fixed point equation

$$\mu = \mathbb{E}_{P_{J_+,h-J_-\mu}}[\sigma] \quad (5)$$

then on the right hand side of (4), the term $\mathbb{E}_{P_{J_+,h-J_-\mu}}[(\sigma-\mu)(\sigma-\mu)^\top]$ (4) is simply a covariance matrix. Because $P_{J_+,h-J_-\mu}$ is an Ising model with all eigenvalues lying in an interval of length $1-\frac{1}{c}$, its covariance matrix is bounded in operator norm by c (Eldan et al., 2020). Hence by the matrix Hölder inequality, we have $\log Z(\mu) \ge -\langle J_-, \mathbb{E}_{P_{J_+,h-J_-\mu}}[(\sigma-\mu)(\sigma-\mu)^\top]\rangle \ge -c\operatorname{Tr}(J_-)$. Provided such a μ exists, this lets us perform importance sampling with expected running time $e^{c\operatorname{Tr}(J_-)}$, by using $P_{J_+,h-J_-\mu}$ as the proposal distribution, which we can sample from using Glauber dynamics by Theorem 2.1.

Solving the fixed point equation: variational argument and nonconvex SGD. There is only one problem remaining: how do we find a solution of the fixed point equation (5), or even know that one exists? To show existence, we use what is known as a *variational argument*: we construct a functional $G(\mu)$ and prove that (1) any critical point of G solves our desired equation (5), and (2) G has at least one global minima, hence at least one critical point. This strategy is quite familiar in the context of variational inference (e.g., constructing BP fixed points (Mezard and Montanari, 2009)), as well as in other fields in mathematics like classical mechanics and PDEs (Evans, 2010).

In our case, we can first assume J_- is strictly positive definite without loss of generality (by adding a small copy of the identity to J_-, which preserves the distribution and only slightly increases the trace). Then we consider the functional

$$G(\mu) := \log \mathbb{E}_{P_{J_+,h}}[e^{\langle \mu, -J_-\sigma\rangle}] + \frac{1}{2}\langle \mu, J_-\mu\rangle. \quad (6)$$

Differentiating, we obtain

$$\nabla G(\mu) = -J_-\mathbb{E}_{P_{J_+,h-J_-\mu}}[\sigma] + J_-\mu \quad (7)$$

3. For a concrete example, suppose $J_+ = 0$, $J_- = \mathbf{1}\mathbf{1}^\top/n$ and $h = \mathbf{1}$. Then by explicit calculation, it can be shown that mean without the J_- term is much further from zero than with the J_- term included.

and because J_- is invertible, this means that $\nabla G(\mu) = 0$ iff μ solves the fixed point equation (5).

To show there exists a global minimizer of $G(\mu)$, we observe that $G(0) = 0$ and by Hölder's inequality that $G(\mu) \geq -\|J_-\mu\|_1 + \langle \mu, J_-\mu\rangle/2 \geq -\sqrt{n}\|J_-\|_{\text{op}}\|\mu\| + \langle \mu, J_-\mu\rangle/2$. The first negative term grows at most linearly in $\|\mu\|$, whereas the second positive term grows quadratically in $\|\mu\|$ because J_- is positive definite. Thus, for all μ with $\|\mu\|$ sufficiently large, we must have that $G(\mu) > 0$. Hence the infimum of G must be achieved within a compact ball around 0, and so G has at least one global minima and at least one critical point.

Now that we have shown that a fixed point exists, there is a clear way to make this argument constructive: run stochastic gradient descent to try to minimize $G(\mu)$, starting from zero. Based on (7), we can indeed compute a stochastic gradient of G provided we can sample from $P_{J_+, h - J_-\mu}$, which we do via Glauber dynamics (Theorem 2.1). While SGD is not guaranteed to find the global minimum, we can use the result of Ghadimi and Lan (2013) to guarantee that SGD at least finds an approximate critical point, which is sufficient.

The general case: Positive and negative eigenvalues. We now describe how to combine the techniques to deal with general case when $J = J^+ - J^-$ can have both positive and negative eigenvalues. In the PSD case, we computed the partition function for (3) over a grid of $\mu^{\|}$'s. We cannot include the negative definite part in $J^{\perp}$, but we know from our variational argument that we can approximate $p_{J^{\perp} - J_-, h + X^\top \mu^{\|}}$ with $p_{J^{\perp}, h + X^\top \mu^{\|} + f(\mu^{\|})}$ for some $f(\mu^{\|})$ we can compute; hence we run the annealing and tempering argument on these distributions instead, with a final step of importance/rejection sampling to bring us back to $p_{J^{\perp} - J_-, h + X^\top \mu^{\|}}$.

3. Applications

Our results specialize to give new sampling guarantees for a many models of interest. All of these are Ising models, so in each application we will describe the particular interaction matrix which arises and the resulting runtime guarantee. In all of the applications, the behavior in the presence of an external field $h \in \mathbb{R}^n$ is of interest (for example, in the Hopfield network to preferentially weight the distribution towards a particular memory) and we automatically handle this case.

Hopfield Network with a fixed number of patterns. The Hopfield network is a neural model of associative memory (Hopfield (1982), see also Pastur and Figotin (1977, 1978); Little (1974)) which has been hugely influential and extensively studied. In particular, for rigorous mathematical results see the textbooks by Bovier and Picco (1998); Talagrand (2010). Formally, given patterns $\eta_1, \dots, \eta_m \in \{\pm 1\}^n$ the Hopfield network at inverse temperature β is the Ising model with interaction matrix $J = \frac{\beta}{2n} \sum_{v=1}^m \eta_v \eta_v^\top$. This is thought of as a "Hebbian" learning rule because for each memory η_v and neurons (coordinates) i and j, the term $(\eta_v)_i (\eta_v)_j$ is positive if $(\eta_v)_i = (\eta_v)_j$ and negative otherwise. Therefore if J is thought of as the "wiring" of the neurons, then for each pattern all of the neurons which "fire together," i.e., have the same spin, are "wired together".

Most of the interest in this model has been in the case of low/zero-temperature, which means the parameter β is large. Glauber dynamics (Gibbs sampling) has long been considered as a natural dynamics for the Hopfield network. Informally, the patterns stored in the network serve as "attractors" which trap the dynamics. This is interesting as in a sense it means the network exhibits memory; however, from the sampling perspective this means that the vanilla Glauber dynamics are not expected to mix in the most interesting regime of this model.

When the number of patterns m is fixed (a regime which has been rigorously studied in e.g., Gentz and Löwe (1999); Bovier and Picco (1998); Talagrand (2010)), we obtain the first polynomial time sampling algorithm for the Gibbs measure of this model that works for any fixed $\beta > 0$. Based on the rigorous results in this model (see Bovier and Picco (1998); Talagrand (2010)), when each pattern is independently sampled $\eta_i \sim \mathsf{Uniform}(\{\pm 1\}^n)$ and $\beta > 1$ the distribution will be almost entirely supported on $2m$ clusters corresponding to each of the patterns $\{\pm \eta_i\}_{i=1}^m$ and so ordinary Glauber dynamics will not mix rapidly. (This should not be too difficult to formally prove given their results, though we did not do this.) Note that our sampling results apply to arbitrary patterns η_i, not just the commonly studied case where the patterns are uniformly random from the hypercube.

Antiferromagnetic and Ferromagnetic Ising Model on expanders and random graphs. Suppose that A is the adjacency matrix of a graph; then the *antiferromagnetic Ising model* at inverse temperature β has interaction matrix $J = -\beta A$. It is known that for worst-case graphs of maximum degree d, that polynomial time sampling is only possible for $\beta = O(1/d)$ (Sly and Sun (2012), in fact the precise threshold is known as a function of d). However, this should be far from tight in other cases of interest, such as on a uniformly random d-regular graph: in this model, it is known that the symmetry breaking phase transition is at scaling $\beta = \Theta(1/\sqrt{d})$ (see Coja-Oghlan et al. (2020) and references within) and we would expect the sampling regime of the model to be similar.

Based on our main result, we can indeed recover the correct scaling in the random d-regular graph setting, as a special case of a much more generic result about spectral expanders. Let $\lambda = \max\{|\lambda_2(A)|, |\lambda_n(A)|\}$; then our results give a polynomial time sampler whenever $\beta d = O(\log n)$ (so that our algorithm is polynomial time) and provided $\beta\lambda < 1$. For example, in the case of a *Ramanujan graph* of degree d we have $\lambda \leq 2\sqrt{d-1}$ and so we can sample in polynomial time whenever $\beta < \frac{1}{2\sqrt{d-1}}$, which is a dramatic improvement over $O(1/d)$. Because of Friedman's Theorem, we know the same result holds for the a uniformly random d-regular graph since it will be almost-Ramanujan (Friedman, 2008). Note that it is the presence of the "trivial" eigenvalue λ_1 which prevents the result from being deduced from the pre-existing works (e.g., Eldan et al. (2020)) which can handle related models (diluted d-regular SK model) without outlier eigenvalues. Our result also applies analogously if there are a couple of outlier eigenvalues, e.g., on bipartite expanders.

A completely analogous consequence of our theory is for the case of *ferromagnetic* Ising models on expanders, where we have $J = \beta A$. In this case, the famous result of Jerrum and Sinclair (Jerrum and Sinclair, 1993) proves that sampling is possible when the external field h is *consistent* i.e., $h_i \geq 0$ for all i. However, when the signs of the external fields h_i are allowed to disagree, sampling from the ferromagnetic Ising model is #BIS-Hard (Goldberg and Jerrum, 2007). So our result also implies sampling algorithms for the ferromagnetic Ising model with inconsistent external field on expanders up to larger inverse temperatures than were previously known.

Sherrington-Kirkpatrick Model with Ferromagnetic Interaction. The Sherrington-Kirkpatrick model is one of the most famous spin glass models, and the SK model with ferromagnetic interactions is a natural variant which exhibits a combination of ferromagnetic and spin glass behaviors—see e.g., Chen (2014); Comets et al. (1999); Talagrand (2010) for rigorous probabilistic analysis of this model. The interaction matrix J is given by $J_{ij} = \frac{\beta_1}{n} + \beta_2 W_{ij}$ where W is a matrix sampled from the Gaussian Orthogonal Ensemble (so $W_{ij} \sim N(0, 1/n)$). Since $\|W\|_{\text{op}} \leq 2(1 + o(1))$ with high probability by classical results in random matrix theory (Anderson et al., 2010), we are able to sample in polynomial time from this model for any fixed β_1, as long as $\beta_2 < 1/4$.

Posterior in Low-Dimensional Gaussian Mixture Model. A basic clustering problem in Bayesian statistics is posterior inference in the two-component (symmetric) Gaussian mixture model. More specifically, we will consider that we have data points $b_1, \dots, b_n \in \mathbb{R}^p$ and we want to sample from the posterior under the following Bayesian model: $u \sim N(0, I_p/p)$, $v \sim \mathsf{Uniform}(\{\pm 1\}^n)$ are the latent cluster assignments and independently $b_i \sim N(v_i \sqrt{\mu/n}\, u, I_p/p)$. In other words, we posit that the data points were generated by a balanced mixture of two spherical Gaussians with means $\pm\sqrt{\mu/n}\, u$ and u itself is sampled from a Gaussian distribution. (For simplicity, we assumed that the data is scaled and centered so that the variance of the components is I_p/p; the scalings here are chosen in part to maintain consistency with the next example.) In this case, the posterior on the cluster assignments v is given by $p(v \mid b) \propto \exp\left(\frac{p\mu}{2n(1+\mu)} \langle vv^\top, BB^\top \rangle\right)$ where $B \in \mathbb{R}^{n \times p}$ is the matrix with rows b_i. (See Appendix G for the derivation.) Note that this is an Ising model with $J = \frac{p\mu}{2n(1+\mu)} BB^\top$ and the rank of J is at most p. Hence, our main result lets us sample from this distribution (posterior in the Gaussian Mixture Model) in polynomial time in fixed dimension p. In the case of a balanced mixture, the posterior will always be bimodal due to the symmetry of swapping the two cluster assignments, and so Glauber dynamics would not be expected to mix. (Also, our algorithms works for general data points $b_1, \dots, b_n$ in which case the posterior can be an arbitrary positive semidefinite Ising model of rank p — in particular, it could be a Hopfield network and have even more than two modes.) In fact, the Hubbard-Stratonovich transform and our algorithm as a whole has a natural interpretation in terms of searching over the latent vector u in this case (see Appendix E). Finally, we note that this example can be easily generalized to asymmetric mixture (mixing weights not 50/50); this just changes the prior, which results in an external field in the (Ising model) posterior.

Remark 3.1 *Importantly, the posterior sampling result we establish does not rely on the data being a typical sample from the posited Bayesian model. This is useful because in many machine learning and statistics applications the data is not exactly generated from the posited model, and nevertheless sampling from the posterior is very useful. On the other hand, if the data is indeed generated from the model (i.e., well-specified) then posterior sampling lets us compute the Bayes-optimal estimator of quantities of interest, e.g., compute* $\Pr(v_i = v_j \mid B)$ *in the GMM example which is the Bayes-optimal estimate of* $1(v_i = v_j)$*, the indicator that* i *and* j *are from the same component.*

Posterior in Low-Dimensional Contextual SBM. The contextual stochastic block model (Deshpande et al., 2018) is a more complex version of the previous GMM model in which the cluster structure is also reflected in the community structure of a graph. We consider the low-dimensional version of this model where the dimension of the contexts p is small—this is morally related to, but different from, the spiked Wishart model with side information, see e.g., Montanari and Venkataramanan (2021). For simplicity, we describe the Gaussianized version of this model below, though our results also apply analogously to the original SBM version.

The generative model is $v \sim \mathsf{Uniform}(\{\pm 1\}^n)$, $u \sim N(0, I_p/p)$, W is a GOE matrix, i.e., a symmetric matrix where independently $W_{ij} \sim N(0, 1/n)$ for $i < j$ and $W_{ii} \sim N(0, 2/n)$, and $Z \in \mathbb{R}^{n \times p}$ is a matrix with iid $N(0, 1/p)$ entries. Then we observe

$$A = \frac{\lambda}{n} vv^\top + W, \qquad B = \sqrt{\frac{\mu}{n}} vu^\top + Z, \qquad u \sim N(0, I_p/p).$$

Informally, words A_{ij} is some indication of whether v_i and v_j are likely to agree, and rows of B are context/feature vectors in $\mathbb{R}^p$ from a mixture of two spherical gaussians with means $\pm\sqrt{\mu/n}\, u$,

where each gaussian corresponds to one community assignment. In this model, the posterior (see Appendix G for the derivation) is $p(v \mid A, B) \propto \exp\left(\frac{\lambda}{2}\langle vv^\top, A\rangle + \frac{p\mu}{2n(1+\mu)}\langle vv^\top, BB^\top\rangle\right)$, so it is an Ising model where the interaction matrix is the weighted sum of A and $BB^\top$. We can sample from this using our result as long as the dimension p is fixed (since $BB^\top$ is rank at most p) and provided $\lambda\|A\|_{\text{op}} < 1/2$. Note that if A is actually generated from the model, then $\|A\|_{\text{op}} \leq 2(1 + o_{n\to\infty}(1))$ due to well-known results on spiked Wigner matrices (see Perry et al. (2018) and references within) in which case we would have mixing for $\lambda < 1/4$. Like our previous application, the sampler works fine with any context matrix B.

Acknowledgements

This work was done in part while the authors were visiting the Simons Institute for the Theory of Computing. F.K. was supported in part by NSF award CCF-1704417, NSF Award IIS-1908774, and N. Anari's Sloan Research Fellowship.

References

Amir Abboud, Karl Bringmann, Danny Hermelin, and Dvir Shabtay. Seth-based lower bounds for subset sum and bicriteria path. *ACM Transactions on Algorithms (TALG)*, 18(1):1–22, 2022.

Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and non-convex SGD. *Advances in Neural Information Processing Systems*, 31, 2018.

Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Entropic independence in high-dimensional expanders: Modified Log-Sobolev inequalities for fractionally log-concave polynomials and the Ising model. *arXiv preprint arXiv:2106.04105*, 2021.

Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*. Number 118. Cambridge university press, 2010.

Fanny Augeri. A transportation approach to the mean-field approximation. *Probability Theory and Related Fields*, 180(1):1–32, 2021.

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

Roland Bauerschmidt and Thierry Bodineau. A very simple proof of the LSI for high temperature spin systems. *Journal of Functional Analysis*, 276(8):2582–2588, 2019.

Ivona Bezáková, Daniel Štefankovič, Vijay V Vazirani, and Eric Vigoda. Accelerating simulated annealing for the permanent and combinatorial counting problems. *SIAM Journal on Computing*, 37(5):1429–1454, 2008.

Christian Borgs, Jennifer Chayes, and Boris Pittel. Phase transition and finite-size scaling for the integer partitioning problem. *Random Structures & Algorithms*, 19(3-4):247–288, 2001.

Christian Borgs, Jennifer Chayes, Tyler Helmuth, Will Perkins, and Prasad Tetali. Efficient sampling and counting algorithms for the Potts model on $\mathbb{Z}^d$ at all temperatures. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 738–751, 2020.

Anton Bovier and Pierre Picco. *Mathematical aspects of spin glasses and neural networks*, volume 41 of *Progress in Probability*. Birkhauser, 1998.

Wei-Kuo Chen. On the mixed even-spin Sherrington-Kirkpatrick model with ferromagnetic interaction. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pages 63–83, 2014.

Zongchen Chen, Kuikui Liu, and Eric Vigoda. Rapid mixing of Glauber dynamics up to uniqueness via contraction. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1307–1318. IEEE, 2020.

Zongchen Chen, Andreas Galanis, Leslie A Goldberg, Will Perkins, James Stewart, and Eric Vigoda. Fast algorithms at low temperatures via Markov chains. *Random Structures & Algorithms*, 58(2):294–321, 2021.

Amin Coja-Oghlan, Philipp Loick, Balázs F Mezei, and Gregory B Sorkin. The Ising antiferromagnet and max cut on random regular graphs. *arXiv preprint arXiv:2009.10483*, 2020.

Francis Comets, Giambattista Giacomin, and Joel L Lebowitz. The Sherrington-Kirkpatrick model with short range ferromagnetic interactions. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 328(1):57–62, 1999.

Yash Deshpande, Andrea Montanari, Elchanan Mossel, and Subhabrata Sen. Contextual stochastic block models. *arXiv preprint arXiv:1807.09596*, 2018.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin monte carlo via convex optimization. *The Journal of Machine Learning Research*, 20(1):2666–2711, 2019.

Martin Dyer and Alan Frieze. Computing the volume of convex bodies: a case where randomness provably helps. *Probabilistic combinatorics and its applications*, 44(123-170):0754–68052, 1991.

Ronen Eldan. Gaussian-width gradient complexity, reverse log-Sobolev inequalities and nonlinear large deviations. *Geometric and Functional Analysis*, 28(6):1548–1596, 2018.

Ronen Eldan. Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probability Theory and Related Fields*, 176(3):737–755, 2020.

Ronen Eldan and Renan Gross. Decomposition of mean-field Gibbs distributions into product measures. *Electronic Journal of Probability*, 23:1–24, 2018.

Ronen Eldan, Frederic Koehler, and Ofer Zeitouni. A spectral condition for spectral gap: Fast mixing in high-temperature Ising models. *arXiv preprint arXiv:2007.08200*, 2020.

Richard S Ellis. *Entropy, large deviations, and statistical mechanics*, volume 1431. Taylor & Francis, 2006.

Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Soc., 2010.

Joel Friedman. *A proof of Alon's second eigenvalue conjecture and related problems*. American Mathematical Soc., 2008.

Alan Frieze and Ravi Kannan. The regularity lemma and approximation schemes for dense problems. In *Proceedings of 37th Conference on Foundations of Computer Science*, pages 12–20. IEEE, 1996.

Andreas Galanis, Daniel Štefankovič, and Eric Vigoda. Inapproximability of the partition function for the antiferromagnetic Ising and hard-core models. *Combinatorics, Probability and Computing*, 25(4):500–559, 2016.

David Gamarnik and Eren C Kızıldağ. Algorithmic obstructions in the random number partitioning problem. *arXiv preprint arXiv:2103.01369*, 2021.

Rong Ge, Holden Lee, and Andrej Risteski. Simulated tempering Langevin monte carlo ii: An improved proof using soft Markov chain decomposition, 2018.

Rong Ge, Holden Lee, and Jianfeng Lu. Estimating normalizing constants for log-concave distributions: Algorithms and lower bounds. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 579–586, 2020.

Barbara Gentz and Matthias Löwe. The fluctuations of the overlap in the Hopfield model with finitely many patterns at the critical temperature. *Probability theory and related fields*, 115(3): 357–381, 1999.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Leslie Ann Goldberg and Mark Jerrum. The complexity of ferromagnetic Ising with local fields. *Combinatorics, Probability and Computing*, 16(1):43–61, 2007.

Heng Guo and Mark Jerrum. Random cluster dynamics for the Ising model is rapidly mixing. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1818–1827. SIAM, 2017.

Geoffrey E Hinton. A practical guide to training restricted Boltzmann machines. In *Neural networks: Tricks of the trade*, pages 599–619. Springer, 2012.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1–6, 2012.

John Hubbard. Calculation of partition functions. *Physical Review Letters*, 3(2):77, 1959.

Vishesh Jain, Frederic Koehler, and Elchanan Mossel. The mean-field approximation: Information inequalities, algorithms, and complexity. In *Conference On Learning Theory*, pages 1326–1347. PMLR, 2018a.

Vishesh Jain, Frederic Koehler, and Elchanan Mossel. The vertex sample complexity of free energy is polynomial. In *Conference On Learning Theory*, pages 1395–1419. PMLR, 2018b.

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*, pages 1226–1236, 2019.

Matthew Jenssen, Peter Keevash, and Will Perkins. Algorithms for #BIS-hard problems on expander graphs. *SIAM Journal on Computing*, 49(4):681–710, 2020.

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 Alistair Sinclair. The Markov chain monte carlo method: an approach to approximate counting and integration. *Approximation Algorithms for NP-hard problems, PWS Publishing*, 1996.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

William A Little. The existence of persistent states in the brain. *Mathematical biosciences*, 19(1-2): 101–120, 1974.

László Lovász and Santosh Vempala. Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithm. *Journal of Computer and System Sciences*, 72(2):392–417, 2006.

Neal Madras and Dana Randall. Markov chain decomposition for convergence rate analysis. *Annals of Applied Probability*, pages 581–606, 2002.

Neal Madras and Zhongrong Zheng. On the swapping algorithm. *Random Structures & Algorithms*, 22(1):66–97, 2003.

Enzo Marinari and Giorgio Parisi. Simulated tempering: a new Monte Carlo scheme. *EPL (Europhysics Letters)*, 19(6):451, 1992.

Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

Andrea Montanari and Ramji Venkataramanan. Estimation of low-rank matrices via approximate message passing. *The Annals of Statistics*, 49(1):321–345, 2021.

Wenlong Mou, Nhat Ho, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. Sampling for Bayesian mixture models: MCMC with polynomial-time mixing. *arXiv preprint arXiv:1912.05153*, 2019.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

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.

Giorgio Parisi and Ramamurti Shankar. Statistical field theory. *Physics Today*, 41(12):110, 1988.

Leonid A Pastur and Alexander L Figotin. Exactly soluble model of a spin glass. *Sov. J. Low Temp. Phys*, 3(6):378–383, 1977.

Leonid Andreevich Pastur and AL Figotin. Theory of disordered spin systems. *Theoretical and Mathematical Physics*, 35(2):403–414, 1978.

Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Optimality and sub-optimality of pca i: Spiked random matrix models. *The Annals of Statistics*, 46(5):2416–2451, 2018.

Andrej Risteski. How to calculate partition functions using convex programming hierarchies: provable bounds for variational methods. In *Conference on Learning Theory*, pages 1402–1416. PMLR, 2016.

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*, pages 361–369. IEEE, 2012.

Daniel Štefankovič, Santosh Vempala, and Eric Vigoda. Adaptive simulated annealing: A near-optimal connection between sampling and counting. *Journal of the ACM (JACM)*, 56(3):1–36, 2009.

Michel Talagrand. *Mean field models for spin glasses: Volume I: Basic examples*, volume 54. Springer Science & Business Media, 2010.

David J Thouless, Philip W Anderson, and Robert G Palmer. Solution of 'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977.

Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

Overview of Appendix

The Appendix includes complete proofs of all of the main results. We set out notations and definitions in Appendix A. Appendix B formalizes the argument for handling negative outlier eigenvalues. Appendix C gives the proof of the part of Theorem 1.1 for estimating the partition function, and Appendix D gives the proof for sampling. Appendix E provides a re-interpretation of the Hubbard-Stratonovich transform in terms of Gaussian mixture posteriors, and Appendix F contains supporting technical lemmas for the previous sections. Appendix G contains additional calculations related to the examples. Finally, we prove the computational hardness results in Appendix H.

Appendix A. Notation and definitions

A.1. Notation

For a set $I \subseteq A$, we let $I \cdot c := \{cx : x \in I\}$; for instance, $\widehat{Z} \in Z \cdot [\frac{1}{2}, 2]$ means $\frac{1}{2}Z \leq \widehat{Z} \leq 2Z$.

We will often omit subscripts and superscripts for probability distributions; when we need to be precise, we will indicate the variables as superscripts (for example, $p^{\sigma,\mu}$, $p^{\sigma|\mu}$). We use a lowercase

letter p to denote the probability density functions and an uppercase letter P to denote the corresponding probability measure. All probability densities are with respect to the uniform measure on the hypercube and Lebesgue measure on $\mathbb{R}^n$. When we write $\propto$, the constants of proportionality do not depend on the variables to the left of the conditioning.

We collect here some notation used in the paper for easy reference.

Probability distributions and partition functions.

$$\begin{aligned}
p_{J,h}(\sigma) &= \frac{1}{Z_{J,h}} \exp\left(\frac{1}{2}\langle \sigma, J\sigma\rangle + \langle h, \sigma\rangle\right) \\
Z_{J,h} &= \sum_{\sigma\in\{\pm 1\}^n} \exp\left(\frac{1}{2}\langle \sigma, J\sigma\rangle + \langle h, \sigma\rangle\right) \\
p^{\sigma,\mu^{\|}}_{J^{\|},J^{\perp},h} &\propto p_{J^{\perp},h+X^\top\mu^{\|}}(\sigma)\exp\left(-\frac{n}{2}\left\|\mu^{\|}\right\|^2\right) \\
&= \exp\left(\frac{1}{2}\left\langle \sigma, J^{\perp}\sigma\right\rangle + \left\langle h + X^\top\mu^{\|}, \sigma\right\rangle - \frac{n}{2}\left\|\mu^{\|}\right\|^2\right) \\
p^{\sigma,y}_{J^{\|},J^{\perp},h}(\sigma, y) &= p^{\sigma,\mu^{\|}}_{J^{\|},J^{\perp},h}(\sigma, Qy) \\
Z_{J^{\|},J^{\perp},h}(\mu^{\|}) &= Z_{J^{\perp},h+X^\top\mu^{\|}}\exp\left(-\frac{n}{2}\left\|\mu^{\|}\right\|^2\right) \\
Z_{J^{\|},J^{\perp},h} &= \int_{V^{\|}} Z_{J^{\|},J^{\perp},h}(\mu^{\|})\, d\mu^{\|}
\end{aligned}$$

Decomposing J.

$$\begin{aligned}
J &= J_+ - J_- \\
J_+ &= \frac{1}{n}XX^\top \\
J^{\|} &= \frac{1}{n}X^\top P^{\|}X \\
J^{\perp} &= \frac{1}{n}X^\top P^{\perp}X = J_+ - J^{\|} \\
J^{\perp}_{\text{all}} &= J^{\perp} - J_- \\
V &= \text{ subspace of } \mathbb{R}^n \text{ spanned by eigenvectors of } J_+ \text{ with eigenvalues } > 1 - \frac{1}{c} \\
Q &= \ n \times d \text{ matrix whose columns are an orthogonal basis for } V
\end{aligned}$$

Probability distributions, partition functions, and partition function estimates from annealing/tempering.

$$\begin{aligned}
\text{Grid}^d_{L,\eta} &= \left\{-L + \frac{1}{2}\eta, -L + \frac{3}{2}\eta, \ldots, L - \frac{1}{2}\eta\right\}^d \\
\mu(y^*) &= \ \text{approximate critical point of } G(u) = \log \mathbb{E}_{\sigma\sim P_{J^{\perp},X^\top Qy^*}}[e^{-\langle u, J_-\sigma\rangle}] + \frac{1}{2}\langle u, J_-u\rangle \\
h(y^*) &= \mu(y^*) + X^\top Qy^* + h
\end{aligned}$$

$$
\begin{aligned}
B(y^*) &= \text{ hypercube with sides parallel to the standard axes, centered at } y^* \text{ with side length } \eta \\
p_{\ell,y^*} &= p_{\beta_\ell J^\perp, h(y^*)} \text{ where } \beta_\ell = \frac{\ell-1}{n} \\
p_{M+1}(\sigma, y^*) &= \frac{\int_{B(y^*)} \exp\left(\frac{1}{2}\left\langle \sigma, J_{\text{all}}^\perp \sigma\right\rangle + \left\langle X^\top Q y + h, \sigma\right\rangle - \frac{n}{2}\|y\|^2\right) dy}{\int_{[-L,L]^d} \sum_{\sigma\in\{\pm1\}^d} \exp\left(\frac{1}{2}\left\langle \sigma, J_{\text{all}}^\perp \sigma\right\rangle + \left\langle X^\top Q y + h, \sigma\right\rangle - \frac{n}{2}\|y\|^2\right) dy} \\
g_\ell(\sigma) &= \exp\left(\frac{1}{2}\left(\beta_{\ell+1} - \beta_\ell\right)\left\langle \sigma, J^\perp \sigma\right\rangle\right) = \exp\left(\frac{1}{2n}\left\langle \sigma, J^\perp \sigma\right\rangle\right), \quad 1 \le \ell \le M-1 \\
g_M(\sigma) = g_{M,y^*}(\sigma) &= \frac{\exp\left(-\frac{1}{2}\left\langle \sigma, J_- \sigma\right\rangle\right)}{\exp\left(\left\langle \mu(y^*), \sigma\right\rangle\right)} \int_{B(y^*)} \exp\left(\left\langle X^\top Q(y-y^*), \sigma\right\rangle - \frac{n}{2}\|y\|^2\right) dy \\
\widehat{Z}_\ell(y^*) &= \text{ estimate for } Z_\ell(y^*) \\
\widehat{Z}(y^*) &= \widehat{Z}_{M+1}(y^*) \\
Z_\ell(y^*) &= Z_{\beta_\ell J^\perp, h(y^*)} \\
R_\ell(y^*) &= \frac{Z_\ell(y^*)}{\widehat{Z}_\ell(y^*)} \\
q_{\ell,y^*} &= \frac{Z_\ell(y^*)}{\widehat{Z}_\ell(y^*)} p_{\ell,y^*} \\
p_\ell^{\text{st}}(\sigma, y) &= \left(\widehat{Z}_\ell(y^*) \sum_{y\in \text{Grid}_{L,\eta}^d} R_\ell(y)\right)^{-1} \exp\left(\frac{1}{2}\left\langle \sigma, J^\perp \sigma\right\rangle + \left\langle h(y^*), \sigma\right\rangle\right).
\end{aligned}
$$

A.2. Background on Markov chains

Let P be a measure on some space Ω and T be the transition kernel of the "natural" Markov chain associated with P, e.g., Glauber dynamics (Algorithm 1) when P is defined on the hypercube $\Omega = \{\pm1\}^n$. The Poincaré and modified log-Sobolev constants of P are defined as

$$
\begin{aligned}
C_{\text{P}}(P) &= \sup\left\{\frac{\text{Var}_P(f)}{\mathcal{E}_P(f,f)} : f : \{\pm1\}^n \to \mathbb{R}, \text{Var}_P(f) \neq 0\right\} \\
C_{\text{MLS}}(P) &= \sup\left\{\frac{2\,\text{Ent}_P(f)}{\mathcal{E}_P(f, \log f)} : f : \{\pm1\}^n \to \mathbb{R}_{\ge 0}, \text{Ent}_P(f) \neq 0\right\}
\end{aligned}
$$

where $\text{Ent}_P(f) = \mathbb{E}_P[f \log f] - \mathbb{E}_P[f] \log \mathbb{E}_P[f]$, and

$$
\begin{aligned}
\mathcal{E}_P(f,g) &= \mathbb{E}_P[f \cdot \mathcal{L}_P g] \\
\text{where} \quad \mathcal{L}_P f &= (\text{id} - T) f.
\end{aligned}
$$

In particular, for Glauber dynamics on $\{\pm1\}^n$,

$$
(\mathcal{L}_P f)(\sigma) = \frac{1}{n}\sum_{i=1}^n \left(\mathbb{E}_P[f(x) | x_{-i} = \sigma_{-i}] - f(\sigma)\right).
$$

Here, for $\sigma \in \{\pm1\}^n$, $\sigma_{-i} \in \{\pm1\}^{n-1}$ denotes all coordinates except the ith one. Note that some texts use instead the reciprocal of $C_{\text{P}}, C_{\text{MLS}}$, or do not include the $\frac{1}{n}$.

We also define the Cheeger constant of the Markov chain by

$$\Phi = \min_{A \subseteq \Omega, P(A) \leq \frac{1}{2}} \frac{Q(A, A^c)}{P(A)}$$
$$\text{where} \quad Q(A, B) = \int_A T(x, B)\, P(dx).$$

Appendix B. Sampling with negative definite spikes using a variational argument

The proof of the following result gives a generic algorithm which, given sampling access to a distribution P and its tilts, samples from any distribution Q which is reweighted by a negative definite quadratic form with small trace. As stated, the result applies to any distribution supported on a $\sqrt{n}$-radius sphere, not just discrete distributions on the hypercube. In fact, when $-J$ is strictly negative definite, the exact same argument applies not just to distributions on the sphere, but supported on any compact set.

Theorem B.1 *Suppose we are given a sampling oracle for a distribution P supported on the sphere $\{x : \|x\| = \sqrt{n}\}$ and all of its tilts*

$$\frac{dP_\lambda}{dP}(x) \propto e^{\langle \lambda, x \rangle}.$$

Also, suppose that for any λ the covariance matrix of P_λ is upper bounded in spectral norm by M. Then for any $J \succeq 0$ and $\varepsilon > 0$, if we define the reweighted measure

$$\frac{dQ}{dP}(x) \propto e^{-\langle x, Jx \rangle / 2},$$

then there exists an algorithm which with probability at least $1 - \delta$, outputs $\lambda \in \mathbb{R}^n$ such that

$$\log \frac{dQ}{dP_\lambda}(x) \leq M \operatorname{Tr}(J) + \varepsilon$$

with runtime and oracle complexity polynomial in n, $1/\varepsilon$, M, $\log(1/\delta)$, and $\|J\|_{\text{op}}$.

Specializing this result to the case of Ising models gives the following algorithmic result.

Corollary B.2 *Suppose that J is an arbitrary symmetric matrix and decompose $J = J_+ - J_-$ where both J_+, J_- are positive semidefinite and suppose that $\|J_+\|_{\text{op}} \leq 1 - \frac{1}{c}$ for $c > 0$. Let $h \in \mathbb{R}^n$ be arbitrary, and define $Q(\sigma) \propto \exp(\frac{1}{2}\langle \sigma, J\sigma \rangle + \langle h, \sigma \rangle)$ and $P_\lambda(\sigma) \propto \exp(\frac{1}{2}\langle x, J_+ x \rangle + \langle h + \lambda, x \rangle)$. There exists an algorithm which with probability at least $1 - \delta$, outputs $\lambda \in \mathbb{R}^n$ such that*

$$\log \frac{dQ}{dP_\lambda}(\sigma) \leq c \operatorname{Tr}(J_-) + \varepsilon$$

with runtime and oracle complexity polynomial in n, $1/\varepsilon$, M, $\log(1/\delta)$, and $\|J_-\|_{\text{op}}$.

Proof This follows by applying Theorem B.1 with $\delta' = \delta/2$. First, we recall from Eldan et al. (2020) (as a consequence of the Poincaré inequality) that we can take $M = \frac{1}{1 - \|J_+\|_{\text{op}}} \leq c$ where M is the upper bound on the spectral norm of the covariance matrix of P_μ as defined in Theorem B.1. If we supposed we had access to an exact sampler from each of the distributions P_μ, this would imply

the result. Since we instead will implement each sampling call with a Markov chain (the Glauber dynamics) which can draw samples extremely close to the distribution P_μ, the actual result follows by coupling these outputs to a hypothetical process which has exact samples.

More precisely, from Theorem 2.1 we can draw a sample from any of the distributions P_λ in polynomial time in the sense that for any $\varepsilon > 0$, with $\mathrm{poly}(n, \log(1/\varepsilon))$ time we can generate a sample with total variation distance at most ε. If q is the maximum number of queries made by the algorithm from Theorem 2.1, then by taking $\varepsilon = \delta/2q$ and using the union bound, we can with probability at least $1 - \delta/2$ couple all of the outputs of the Markov chains invoked at every oracle call with samples from the true distribution P_λ. Therefore, with total probability at least $1 - \delta$, the algorithm which uses Markov chain samplers will output λ satisfying the guarantee of Theorem B.1. This proves the result. ■

We now proceed to the proof of Theorem B.1. In the algorithm and analysis, we will use the fact that stochastic gradient descent with an appropriate step size schedule is able to find approximate critical points of smooth functions (a stronger and more explicit result is given in the original statement in Ghadimi and Lan (2013), see also Allen-Zhu (2018)).

Theorem B.3 (Corollary 2.5 of Ghadimi and Lan (2013)) *Suppose that f is a differentiable function which is L-smooth with respect to the Euclidean norm $\|\cdot\|$ in the sense that for all x, y*

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|.$$

Let $f^* := \inf_x f(x)$ and define

$$D_f := \sqrt{\frac{2(f(x_1) - f^*)}{L}}.$$

Then there exists a polynomial time algorithm (2-RSG, the two-phase randomized stochastic gradient algorithm) which given oracle access to (identical, independent copies of) a stochastic gradient oracle g such that $\mathbb{E}[g(x_t) \mid x_t] = \nabla f$ and $\mathbb{E}[\exp(\|g(x_t)\|^2/\sigma^2) \mid x_t] \le 1$ and $\varepsilon > 0$, with probability at least $1 - \delta$ outputs x such that $\|\nabla f\| \le \varepsilon$ using $\mathrm{poly}(D_f, \log(1/\delta), \sigma, L, 1/\varepsilon)$ runtime and oracle calls.

Proof [Proof of Theorem B.1] First, we can assume $J \succeq \varepsilon/Mn$ without loss of generality by adding $(\varepsilon/Mn)I$ to J, which does not change the measure Q and increases the trace by just ε/M. (This only changes the final guarantee by an additional additive ε, which can be trivially corrected by dividing ε by 2.)

The key idea of the proof is a variational argument. Define the functional

$$G(\mu) := \log \mathbb{E}_P[e^{\langle \mu, -JX\rangle}] + \frac{1}{2}\langle \mu, J\mu\rangle$$

and observe that its derivative can be expressed in terms of the tilted measure $P_{-J\mu}$:

$$\nabla G(\mu) = -J\mathbb{E}_{P_{-J\mu}}[X] + J\mu.$$

Now observe that for any μ,

$$\frac{dQ}{dP_{-J\mu}}(x) \propto e^{-\langle x, Jx\rangle/2 + \langle J\mu, x\rangle} \propto e^{-\frac{1}{2}\langle x-\mu, J(x-\mu)\rangle}$$

and so

$$\frac{dQ}{dP_{-J\mu}}(x) = \frac{1}{Z} e^{-\frac{1}{2}\langle x-\mu, J(x-\mu)\rangle}$$

where

$$Z := \mathbb{E}_{P_{J\mu}}[e^{-\frac{1}{2}\langle X-\mu, J(X-\mu)\rangle}].$$

From the definition and the fact that J is psd, we have $Z \leq 1$. Also, by Jensen's inequality

$$Z = \mathbb{E}_{P_{J\mu}}[e^{-\frac{1}{2}\langle X-\mu, J(X-\mu)\rangle}] \geq \exp\left(-\mathbb{E}_{P_{J\mu}}\left[\frac{1}{2}\langle X-\mu, J(X-\mu)\rangle\right]\right).$$

Observe that if $\Sigma := \mathbb{E}_{P_{J\mu}}[XX^\top] - \mathbb{E}_{P_{J\mu}}[X]\mathbb{E}_{P_{J\mu}}[X]^\top$ then

$$\begin{aligned}
\mathbb{E}_{P_{J\mu}}[(X-\mu)(X-\mu)^\top] &= \mathbb{E}_{P_{J\mu}}[XX^\top] - \mathbb{E}_{P_{J\mu}}[X]\mu^\top - \mu\mathbb{E}_{P_{J\mu}}[X]^\top + \mu\mu^\top \\
&= \Sigma + \mathbb{E}_{P_{J\mu}}[X]\mathbb{E}_{P_{J\mu}}[X]^\top - \mathbb{E}_{P_{J\mu}}[X]\mu^\top - \mu\mathbb{E}_{P_{J\mu}}[X]^\top + \mu\mu^\top \\
&= \Sigma + (\mathbb{E}_{P_{J\mu}}[X] - \mu)(\mathbb{E}_{P_{J\mu}}[X] - \mu)^\top
\end{aligned}$$

so

$$\begin{aligned}
\log Z &\geq -\frac{1}{2}\langle \mathbb{E}_{P_{J\mu}}[(X-\mu)(X-\mu)^\top], J\rangle \\
&= -\frac{1}{2}\langle \Sigma + (\mathbb{E}_{P_{J\mu}}[X] - \mu)(\mathbb{E}_{P_{J\mu}}[X] - \mu)^\top, J\rangle \\
&\geq -\frac{1}{2}\|\Sigma\|_{\text{op}} \operatorname{Tr}(J) - \frac{1}{2}\|J\mathbb{E}_{P_{J\mu}}[X] - J\mu\|_2 \|\mathbb{E}_{P_{J\mu}}[X] - \mu\|_2 \\
&= -\frac{1}{2}\|\Sigma\|_{\text{op}} \operatorname{Tr}(J) - \frac{1}{2}\|\nabla G(\mu)\|_2 \|\mathbb{E}_{P_{J\mu}}[X] - \mu\|_2.
\end{aligned}$$

Note that the final lower bound can be maximized if we can find a critical point of G. We next argue that such a critical point exists.

Note that $G(0) = 0$ by definition and because we reduced to the case $J \succeq (\varepsilon/Mn)I$,

$$\begin{aligned}
G(\mu) &\geq \log \mathbb{E}_P[e^{\langle \mu, -JX\rangle}] + (\varepsilon/2Mn)\|\mu\|_2^2 \\
&\geq -\|J\mu\|_1 + (\varepsilon/2Mn)\|\mu\|_2^2 \geq -\|J\|_{\text{op}}\|\mu\|_2\sqrt{n} + (\varepsilon/2Mn)\|\mu\|_2^2, \qquad (8)
\end{aligned}$$

which is positive provided $\|\mu\|_2 > 2Mn^{3/2}\|J\|_{\text{op}}/\varepsilon$. Hence the global minimum of G must be attained somewhere on the compact set $\mathcal{K} = \{\mu : \|\mu\|_2 \leq 2Mn^{3/2}\|J\|_{\text{op}}/\varepsilon\}$. At this point, we have proved the existence of a critical point. We next show that one can be approximately found with stochastic gradient descent initialized at zero, by checking the assumptions of Theorem B.3.

By the invertibility of J, any solution of the equation $0 = \nabla G(\mu) = -J\mathbb{E}_{P_{J\mu}}[X] + J\mu$ satisfies $\mu = \mathbb{E}_{P_{J\mu}}[X]$ and hence $\mu \in [-1, 1]^n$ and $\|\mu\|_2 \leq \sqrt{n}$. In particular the global minimum satisfies this, so combined with (8) we have

$$\inf_\mu G(\mu) \geq \inf_{r \leq \sqrt{n}}[-r\sqrt{n}\|J\|_{\text{op}} + (\varepsilon/2Mn)r^2] > -\frac{Mn^2}{2\varepsilon}\|J\|_{\text{op}}^2$$

Since

$$\nabla^2 G(\mu) = -J\mathbb{E}_{P_{J\mu}}[XX^\top]J + J$$

we have that $\|\nabla^2 G(\mu)\|_{\text{op}} \le M\|J\|_{\text{op}}^2 + \|J\|_{\text{op}} =: L$ which means that $G(\mu)$ is L-smooth with respect to the Euclidean norm. Recalling that $\nabla G(\mu) = -J\mathbb{E}_{P_{J\mu}}[X] + J\mu$, we see that if $x \sim P_{J\mu}$, which we have a sampling oracle for by assumption, then $g(\mu) := -J(x-\mu)$ is a stochastic gradient oracle for $G(\mu)$ satisfying $\|g(\mu)\| \le \|J\|_{\text{op}}\|x-\mu\| \le 2\|J\|_{\text{op}}\sqrt{n}$. This means that all of the assumptions of Theorem B.3 are satisfied and we can find an ε-approximate critical point of G using $\text{poly}(n, \|J\|_{\text{op}}, M, 1/\varepsilon)$ runtime and calls to the sampling oracle. Outputting $\lambda := -J\mu$ gives the result.

■

Remark B.4 *The variational argument in the proof is partially inspired by, thought different from, some previous arguments in the variational methods literature; for example, the construction of Belief Propagation fixed points using the Bethe free energy, and variants of this argument which arise from the Thouless-Anderson-Palmer and naive mean-field free energy (see e.g., Mezard and Montanari (2009); Wainwright and Jordan (2008)). As with all such variational arguments, the key idea is to construct a solution to a fixed point equation by writing it as the gradient of a well-behaved functional. To make a more explicit connection with that literature, consider the special case where P is a product measure on the hypercube $\{\pm 1\}^n$, so $P(\sigma) \propto e^{\langle h_0, \sigma\rangle}$ for some $h_0 \in \mathbb{R}^n$ encoding the bias of each coordinate. Then the equation $\nabla G(\mu) = 0$ is equivalent to $-J\mu = -J\tanh(h_0 - J\mu)$ and because J is invertible, it simplifies to the fixed-point equation*

$$\mu = \tanh(h_0 - J\mu).$$

This is almost the same as the naive mean-field fixed point equation, except that in that case, the diagonal of J must be zeroed out whereas in our case they are not. Relatedly, $G(\mu)$ is not the same as the naive mean-field free energy corresponding to Q, and the positive definiteness of J is not needed to solve the naive mean-field equations but plays a key role in our variational argument.

Appendix C. Estimating the partition function

In this section, we develop and analyze an algorithm for computing the partition function Z.

Application of the Hubbard-Stratonovich transform. Based on the Hubbard-Stratonovich transform, we can easily prove the following Theorem. (We warn the reader that the notation has a couple minor cosmetic differences from the Technical Overview, with the goal of minimizing ambiguity.)

Theorem C.1 *Let $J \in \mathbb{R}^{n\times n}$ be a symmetric matrix, and write $J = \frac{1}{n}X^\top X - J_-$ for $X \in \mathbb{R}^{m\times n}$ and J_- negative semi-definite.*

Let $V \subseteq \mathbb{R}^m$ be a subspace. Let $P^\parallel$ and $P^\perp$ be the projections onto V and $V^\perp$. Let $J^\parallel = \frac{1}{n}X^\top P^\parallel X$ and $J^\perp = J - J^\parallel$. Then

$$Z_{J,h} = \left(\frac{n}{2\pi}\right)^{d/2} Z_{J^\parallel, J^\perp, h} \quad \text{where} \quad Z_{J^\parallel, J^\perp, h} = \int_{V^\parallel} Z_{J^\perp, h + X^\top \mu^\parallel} \exp\left(-\frac{n}{2}\left\|\mu^\parallel\right\|^2\right) d\mu^\parallel.$$

Note that in the special case that $V = \mathbb{R}^m$ and $J_- = O$, this gives a decomposition of the probability measure in terms of product distributions in a similar manner to (Bovier and Picco, 1998; Bauerschmidt and Bodineau, 2019).

Proof [Proof of Theorem C.1] We decompose $J = J^{\perp} + J^{\parallel}$ and apply Lemma 2.2 to $X \leftarrow P^{\parallel}X$ with $\gamma^2 = 1/n$:

$$\begin{aligned}
Z_{J,h} &= \sum_{\sigma\in\{\pm1\}^n} \exp\left(\frac{1}{2}\langle\sigma, J\sigma\rangle + \langle h,\sigma\rangle\right) \\
&= \sum_{\sigma\in\{\pm1\}^n} \exp\left(\frac{1}{2}\left\langle\sigma, J^{\perp}\sigma\right\rangle + \langle h,\sigma\rangle\right)\exp\left(\frac{1}{2n}\left\|P^{\parallel}X\sigma\right\|^2\right) \\
&= \left(\frac{n}{2\pi}\right)^{d/2}\sum_{\sigma\in\{\pm1\}^n} \exp\left(\frac{1}{2}\left\langle\sigma, J^{\perp}\sigma\right\rangle + \langle h,\sigma\rangle\right)\int_{V^{\parallel}}\exp\left(\left\langle X^{\top}P^{\parallel}\mu^{\parallel},\sigma\right\rangle - \frac{n}{2}\left\|\mu^{\parallel}\right\|^2\right)d\mu^{\parallel} \\
&= \left(\frac{n}{2\pi}\right)^{d/2}\int_{V^{\parallel}} Z_{J^{\perp},h+X^{\top}\mu^{\parallel}}\exp\left(-\frac{n}{2}\left\|\mu^{\parallel}\right\|^2\right)d\mu^{\parallel},
\end{aligned}$$

as desired. ■

We can define an associated probability distribution on $\{\pm1\}^n \times V$ with $Z_{J^{\parallel},J^{\perp},h}$ as its partition function:

$$p^{\sigma,\mu^{\parallel}}_{J^{\parallel},J^{\perp},h}(\sigma,\mu^{\parallel}) \propto \exp\left(\frac{1}{2}\left\langle\sigma, J^{\perp}\sigma\right\rangle + \left\langle h + X^{\top}\mu^{\parallel},\sigma\right\rangle - \frac{n}{2}\left\|\mu^{\parallel}\right\|^2\right).$$

Choosing an orthogonal linear transformation $Q : \mathbb{R}^d \to V$, we will also define the distribution $p^{\sigma,y}_{J^{\parallel},J^{\perp},h}(\sigma,y) = p^{\sigma,\mu^{\parallel}}_{J^{\parallel},J^{\perp},h}(\sigma,Qy)$. In Appendix E, we will interpret $p^{\sigma,y}_{J^{\parallel},J^{\perp},h}$ as the posterior of a Gaussian mixture model after seeing samples given by the columns of X.

Estimating the partition function. For a PSD matrix A, let $\mathrm{rank}_{\tau}(A)$ denote the number of eigenvalues of A that are $\geq \tau$. Note that $\mathrm{rank}_1(A) \leq \|A\|_F^2$. For ease of exposition, we first prove the theorem when in the case where J has no negative eigenvalues.

Theorem C.2 *Let $\varepsilon, \delta \in (0,1)$. Suppose J is PSD. With probability $\geq 1-\delta$, Algorithm 3 outputs an e^{ε}-multiplicative approximation to $Z_{J,h}$,*

$$e^{-\varepsilon}Z_{J,h} \leq \widehat{Z}_{J,h} \leq e^{\varepsilon}Z_{J,h},$$

in time $(\|J\|_{\mathrm{op}}\, n)^{O(\mathrm{rank}_1(J)+1)}O\left(\log\left(\frac{1}{\delta}\right)/\varepsilon^2\right)$.

Given a probability distribution p on $\{\pm1\}^n$, we can define the Markov chain in Algorithm 1. For $\sigma \in \{\pm1\}^n$, we let $\sigma^{(i)} = (\sigma_1,\ldots,-\sigma_i,\ldots,\sigma_n)$ denote σ but with the ith coordinate flipped.

Algorithm 1: Glauber dynamics on $\{\pm1\}^n$

Input: Query access to probability distribution $p(\sigma)$ on $\{\pm1\}^n$, up to constant of proportionality; number of steps T.

for $1 \leq t \leq T$ **do**

  Choose a random coordinate i, and set $\sigma \leftarrow \sigma^{(i)}$ with probability $\frac{p(\sigma^{(i)})}{p(\sigma^{(i)})+p(\sigma)}$.

end

The following lemma gives fast mixing of Glauber dynamics for the Ising model, when the spectral norm of the interaction matrix is at most 1.

Lemma C.3 *Suppose $J \in \mathbb{R}^{n\times n}$ is symmetric and PSD with $\|J\|_{\text{op}} \le 1$. Then the modified log-Sobolev constant C_{MLS} for $P_{J,h}$ is at most $e^{1/2}n^2$, and the mixing time is bounded by $O(n^2 \log n)$.*

Proof For a symmetric matrix with diagonalization $A = UDU^\top$, let $D_{\le\tau}$ denote D with the entries $\ge \tau$ replaced by τ, and $A_{\le\tau} := UD_{\le\tau}U^\top$. By Theorem 2.1, the modified log-Sobolev constant for

$$p_{J_{\le 1-\frac{1}{n}},h}(\sigma) \propto \exp\left(\frac{1}{2}\left\langle \sigma, J_{\le 1-\frac{1}{n}}\sigma\right\rangle + \langle h, \sigma\rangle\right)$$

is bounded by $n\left(1 - \left\|J_{\le 1-\frac{1}{n}}\right\|_{\text{op}}\right)^{-1} = n^2$. Since

$$\log\left(\frac{p_{J,h}(\sigma)}{p_{J_{\le 1-\frac{1}{n}},h}(\sigma)}\right) - \log\left(\frac{Z_{J_{\le 1-\frac{1}{n}},h}}{Z_{J,h}}\right) = \frac{1}{2}\left\langle \sigma, (J - J_{\le 1-\frac{1}{n}})\sigma\right\rangle$$
$$\in \frac{1}{2}\left\|J - J_{\le 1-\frac{1}{n}}\right\|_2 n\cdot[0,1] \subseteq \left[0, \frac{1}{2}\right], \tag{9}$$

by the Holley-Stroock perturbation lemma, the modified log-Sobolev constant for $p_{J,h}$ is bounded by $e^{1/2}n^2$.

Finally, the exchange property holds for $p_{J,h}$ by (Anari et al., 2021, Lemma 37), so by (Anari et al., 2021, Lemma 36), the mixing time is bounded by $O((n + C_{\text{MLS}})\log n) = O(n^2 \log n)$. ■

Lemma C.3 implies that Glauber dynamics gives an efficient algorithm for sampling in our setting. To obtain an algorithm for partition function estimation, we use simulated annealing. Simulated annealing is a generic method to obtain an algorithm for estimating a partition function $\int_\Omega q\, d\omega$, given access to sampling oracles for a sequence of distributions $p_\ell \propto q_\ell$ such that (a) q_1 is known, (b) for each ℓ, p_ℓ and $p_{\ell+1}$ are "close," and (c) $p_{M+1} \propto q$.

Lemma C.4 *Let $0 < \varepsilon < 1$. Suppose that $p_\ell, 1 \le \ell \le M+1$ are distributions on Ω, and that in Algorithm 2 we are given sampling oracles for $\widetilde{p}_\ell$, $1 \le \ell \le M$ such that the following hold for each $1 \le \ell \le M$.*

1. *(Variance bound)* $\frac{\text{Var}_{P_\ell}(g_\ell(x))}{(\mathbb{E}_{P_\ell} g_\ell(x))^2} \le \sigma^2$.
2. *(Bias bound)* $\left|\mathbb{E}_{P_\ell} g_\ell(x) - \mathbb{E}_{\widetilde{P}_\ell} g_\ell(x)\right| \le \frac{\varepsilon}{4M}$.

Then taking $N \ge \frac{320\sigma^2 M}{\varepsilon^2}$ and $R \ge 32\log\left(\frac{1}{\delta}\right)$, with probability $1-\delta$, the output $\widehat{Z}$ satisfies $\widehat{Z} \in [e^{-\varepsilon}, e^{\varepsilon}]\cdot Z$.

The proof is standard and given in the appendix.

We can now give the algorithm and proof of Theorem C.2. We show that a non-adaptive temperature schedule of length $O(n)$ is sufficient for partition function estimation. Note that a shorter schedule of length $O(\sqrt{n}\log n \log\log n)$ is possible, and can be found in $n\,\text{polylog}(n)$ total queries to approximate sampling oracles at the different temperatures (Štefankovič et al., 2009), but we use a non-adaptive schedule for simplicity. Coordinate-wise sampling is also possible, but we will need a sequence of distributions at different temperatures for our sampling algorithm.

Algorithm 2: Simulated annealing for partition function estimation

Input: Sampling oracles for $\widetilde{p}_\ell$ (approximations to $p_\ell \propto q_\ell$) for $1 \le \ell \le M$ (distributions on Ω), for example, Glauber dynamics (Algorithm 1); $Z_1 = \int_\Omega q_1 \, d\omega$; number of samples N; number of trials R.

Output: Estimate of $\int_\Omega q_\ell \, d\omega$ for each $1 \le \ell \le M+1$.

Let $g_\ell(x) := \frac{q_{\ell+1}(x)}{q_\ell(x)}$.

for $1 \le r \le R$ **do**

  Let $\widehat{Z}_1^r = Z_1$.

  for $1 \le \ell \le M$ **do**

    Obtain samples $x_1, \dots, x_N \sim \widetilde{p}_\ell$.

    Let $\widehat{Y}_\ell = \frac{1}{N} \sum_{k=1}^N g_\ell(x_k)$.

    Let $\widehat{Z}_{\ell+1}^r = \widehat{Z}_\ell^r \widehat{Y}_\ell$.

  end

end

for $2 \le \ell \le M+1$ **do**

  Let $\widehat{Z}_\ell$ be the median of $\left\{ \widehat{Z}_\ell^r : 1 \le r \le R \right\}$.

end

Proof [Proof of Theorem C.2] We may assume $\varepsilon \ge 2^{-n}$. Set the temperature schedule as $\beta_\ell = \frac{\ell - 1}{n}$ for $1 \le \ell \le n+1$. Let $M = n+1$ be the length of the temperature schedule. We set parameters as suggested in Algorithm 3. Then the total time complexity of the algorithm is $O\left(\left(\frac{2L}{\eta}\right)^d MNRT\right)$ times the complexity of each Markov chain step, which gives complexity $O\left((\|J\|_{\text{op}} + 1)nd\right)^d \cdot O\left(\frac{\text{poly}(n) \log\left(\frac{1}{\varepsilon\delta}\right)}{\varepsilon^2}\right) = (\|J\|_{\text{op}} \, n)^{O(\text{rank}_1(J)+1)} O\left(\log\left(\frac{1}{\delta}\right)/\varepsilon^2\right)$.

Recall that we define the distribution $p^{\sigma,y}_{J^\parallel, J^\perp, h}(\sigma, y) = p^{\sigma, \mu^\parallel}_{J^\parallel, J^\perp, h}(\sigma, Qy)$. We now fix a particular y^*, and write for short $g_M = g_{M, y^*}$.

Choice of ratios g_ℓ. Define $Z_{J^\parallel, J^\perp, h}(\mu^\parallel) := Z_{J^\perp, h + X^\top P^\parallel \mu^\parallel} \exp\left(-\frac{n}{2} \left\|\mu^\parallel\right\|^2\right)$. We first compute

$$\begin{aligned}
\mathbb{E}_{p_M} g_M &= \frac{1}{Z_{J^\perp, h(y^*)}} \sum_{\sigma \in \{\pm 1\}^n} \exp\left(\frac{1}{2}\left\langle \sigma, J^\perp \sigma \right\rangle + \left\langle X^\top Q y^* + h, \sigma \right\rangle\right) \\
&\quad \cdot \int_{B(y^*)} \exp\left(\left\langle X^\top Q(y - y^*), \sigma \right\rangle - \frac{n}{2}\|y\|^2\right) \\
&= \frac{1}{Z_{J^\perp, h(y^*)}} \sum_{\sigma \in \{\pm 1\}^n} \int_{B(y^*)} \exp\left(\frac{1}{2}\left\langle \sigma, J^\perp \sigma \right\rangle + \left\langle X^\top Q y + h, \sigma \right\rangle - \frac{n}{2}\|y\|^2\right) dy \\
&= \frac{\int_{B(y^*)} Z_{J^\parallel, J^\perp, h}(Qy) \, dy}{Z_{J^\perp, h(y^*)}}.
\end{aligned}$$

Algorithm 3: Approximating partition function of Ising model. (Steps in italics are only needed in presence of a negative definite spike.)

Input: Ising model (J, h), cutoff L, discretization η dividing evenly into L, desired accuracy ε, failure probability δ, number of samples N, number of trials R, steps to run Markov chains T, threshold $c \in (1, \infty]$.

Output: Approximation of partition function $Z_{J,h}$.

Suggested parameters: $L = \Theta(\sqrt{\|J\|_{\text{op}}} + 1)$, $\eta \leq \frac{1}{ndL + 2n\sqrt{\|J\|_{\text{op}} d}}$, $N = \Theta\left(\frac{M}{\varepsilon^2}\right)$ where $M = n + 1$, $R = \Theta\left(\log\left(\frac{(L/\eta)^d}{\delta}\right)\right)$, and $T = \Theta\left(n^2 \log\left(\frac{n}{\varepsilon}\right)\right)$. Take $c = \infty$ if J is PSD.

If $\varepsilon \leq 2^{-n}$, calculate $Z_{J,h}$ by brute force.

Let $J = J_+ - J_-$ where J_+ and J_- are positive semi-definite and negative semi-definite, respectively, with column spaces intersecting only in 0.

Factor $J_+ = \frac{1}{n} X X^\top$ for $X \in \mathbb{R}^{n \times n}$.

Let V denote the subspace of $\mathbb{R}^n$ spanned by the eigenvectors of J_+ with eigenvalues $> 1 - \frac{1}{c}$. Let $P^{\|}$ and $P^{\perp}$ the projections to V and $V^{\perp}$. Let $Q \in \mathbb{R}^{n \times d}$ be the matrix with columns that are an orthonormal basis for V.

Let $J^{\|} = \frac{X^\top P^{\|} X}{n}$ and $J^{\perp} = \frac{X^\top P^{\perp} X}{n}$.

for $y^* \in \text{Grid}^d_{L,\eta} := \left\{-L + \frac{1}{2}\eta, -L + \frac{3}{2}\eta, \ldots, L - \frac{1}{2}\eta\right\}^d$ **do**

Let $\mu(y^*)$ be an approximate critical point of $G(u) = \log \mathbb{E}_{\sigma \sim P_{J^\perp, X^\top Q y^*}} [e^{-\langle u, J_- \sigma \rangle}] + \frac{1}{2} \langle u, J_- u \rangle$, found using stochastic gradient descent (Theorem B.2/B.3) with sampling oracle given by Glauber dynamics for $P_{J^\perp, X^\top Q y^* + h}$. (If $J_- = O$, let $\mu(y^*) = 0$.)

Let $B(y^*)$ denote the hypercube with sides parallel to the standard axes, centered at y^* with side length η.

Apply Algorithm 2 to the Ising model, with sampling algorithm given by running Glauber dynamics for T steps, for the following sequence of distributions ($1 \leq \ell \leq M = n + 1$):

$$p_\ell = p_{\frac{\ell - 1}{n} J^\perp, h(y^*)}$$

$$g_\ell(\sigma) = \exp\left(\frac{1}{2n} \left\langle \sigma, J^\perp \sigma \right\rangle\right), \quad 1 \leq \ell \leq n$$

$$g_{M, y^*}(\sigma) = \frac{\exp\left(-\frac{1}{2} \langle \sigma, J_- \sigma \rangle\right)}{\exp\left(\langle \mu(y^*), \sigma \rangle\right)} \int_{B(y^*)} \exp\left(\left\langle X^\top Q (y - y^*), \sigma \right\rangle - \frac{n}{2} \|y\|^2\right) dy$$

where $h(y) = \mu(y) + X^\top Q y + h$,

and initial partition function

$$Z_1 = Z_{O, h(y^*)} = 2^n \prod_{i=1}^{n} \cosh\left(\langle x_i, Q y^* \rangle + h_i\right)$$

to get estimates $\widehat{Z}_\ell(y^*)$ for $1 < \ell \leq M + 1$. Let $\widehat{Z}(y^*) := \widehat{Z}_{M+1}(y^*)$.

end

Return $\widehat{Z} = \left(\frac{n}{2\pi}\right)^{\frac{d}{2}} \sum_{y^* \in \text{Grid}^d_{L,\eta}} \widehat{Z}(y^*)$.

Hence

$$Z_1 \prod_{\ell=1}^{M} \mathbb{E}_{P_\ell} g_\ell = Z_{O,h(y^*)} \prod_{\ell=1}^{M-1} \frac{Z_{\beta_{\ell+1}J^\perp,h(y^*)}}{Z_{\beta_\ell J^\perp,h(y^*)}} \cdot \frac{\int_{B(y^*)} Z_{J^\parallel,J^\perp,h}(Qy)\,dy}{Z_{J^\perp,h(y^*)}} = \int_{B(y^*)} Z_{J^\parallel,J^\perp,h}(Qy)\,dy.$$

Variance of g_ℓ. With $g_\ell(\sigma) = \exp(\frac{1}{2}(\beta_{\ell+1}-\beta_\ell)\sigma^\top J^\perp \sigma) = \exp\left(\frac{1}{2n}\sigma^\top J^\perp \sigma\right)$, we bound

$$g_\ell(\sigma) \le \exp\left(\frac{1}{2n}\cdot n\right) = e^{1/2}, \qquad \frac{\mathbb{E}_{P_\ell} g_\ell^2}{(\mathbb{E}_{P_\ell} g_\ell)^2} \le \mathbb{E}_{P_\ell} g_\ell^2 \le e. \tag{10}$$

We also need to check the variance of

$$g_M(\sigma) = \exp\left(-\frac{n}{2}\|y^*\|^2\right) \int_{B(y^*)} \exp\left(\left\langle X^\top Q(y-y^*), \sigma\right\rangle + \frac{n}{2}\left(\|y^*\|^2 - \|y\|^2\right)\right) dy$$

Note that $\left\|\frac{X^\top P^\parallel X}{n}\right\|_{\text{op}} \le \|J\|_{\text{op}}$, so $\left\|P^\parallel X\right\|_{\text{op}} \le \sqrt{n\|J\|_{\text{op}}}$. We check how much the exponent can vary on $B(y^*)$:

$$\begin{aligned}\left|\left\langle X^\top Q(y-y^*), \sigma\right\rangle + \frac{n}{2}\left(\|y^*\|^2 - \|y\|^2\right)\right| &\le \|y-y^*\|\left(\left\|P^\parallel X\sigma\right\| + \frac{n}{2}\|y^*+y\|\right) \\ &\le \frac{\eta}{2}\sqrt{d}\left(\sqrt{n\|J\|_{\text{op}}}\sqrt{n} + nL\sqrt{d}\right) \le \frac{1}{2}\end{aligned} \tag{11}$$

when $\eta \le \frac{1}{ndL + 2n\sqrt{\|J\|_{\text{op}} d}}$. This makes $\frac{\mathbb{E}_{p_M}[g_M(\sigma)^2]}{\mathbb{E}_{p_M} g_M(\sigma)^2} \le e$ as well. We note g_M can be easily evaluated since it can be written as a product of integrals of a Gaussian on an interval.

Bias of $\mathbb{E}g_\ell$. For the approximate sampling oracle, we let $\widetilde{p}_\ell$ be the distribution after running Glauber dynamics for $\Theta\left(n^2 \log\left(\frac{n}{\varepsilon}\right)\right)$ steps (for an appropriate choice of constant). Then by Theorem C.3 and (10), $\left|\mathbb{E}_{P_\ell} g_\ell(\sigma) - \mathbb{E}_{\widetilde{P}_\ell} g_\ell(\sigma)\right| \le d_{\text{TV}}(P_\ell, \widetilde{P}_\ell)\cdot e^{1/2} \le \frac{\varepsilon}{4M}$.

Using Lemma C.4. By Lemma C.4 with δ replaced by $\frac{\delta}{(L/\eta)^d}$, using a union bound, we obtain that with probability $\ge 1-\delta$, for all $y^* \in \text{Grid}^d_{L,\eta}$, $\widehat{Z}(y^*) \in [e^{-\frac{\varepsilon}{2}}, e^{\frac{\varepsilon}{2}}]\cdot \int_{B(y^*)} Z_{J^\parallel,J^\perp,h}(Qy)\,dy$ and so

$$\sum_{y^*\in\text{Grid}^d_{L,\eta}} \widehat{Z}(y^*) \in [e^{-\frac{\varepsilon}{2}}, e^{\frac{\varepsilon}{2}}]\cdot \sum_{y^*\in\text{Grid}^d_{L,\eta}} \int_{B(y^*)} Z_{J^\parallel,J^\perp,h}(Qy)\,dy = [e^{-\frac{\varepsilon}{2}}, e^{\frac{\varepsilon}{2}}]\cdot \int_{\|y\|_\infty \le L} Z_{J^\parallel,J^\perp,h}(Qy)\,dy.$$

Error from cutoff. We would like to estimate $Z_{J^\parallel,J^\perp,h} = \int_{\mathbb{R}^d} Z_{J^\parallel,J^\perp,h}(Qy)\,dy$, so it remains to show that at least $e^{-\frac{\varepsilon}{2}}$ of the probability mass of $p(\sigma, y)$ is contained in $\{\pm 1\}^n \times [-L, L]^d$. For this, it suffices to fix σ, and show that $P(y \notin [-L,L]^d|\sigma) \le \frac{\varepsilon}{2}$. We have by Lemma E.2(3) that

$$p_{J^\parallel,J^\perp,h}(y|\sigma) = \left(\frac{n}{2\pi}\right)^{d/2} \exp\left(-\frac{n}{2}\left\|Qy - \frac{\sum_{i=1}^n \sigma_i P^\parallel x_i}{n}\right\|^2\right).$$

Using $\left\|X^\top P^\parallel\right\|_2 \le \sqrt{n\|J\|_{\text{op}}}$, we get

$$\left\|\frac{\sum_{i=1}^n \sigma_i P^\parallel x_i}{n}\right\| \le \left\|\frac{X^\top P^\parallel}{n}\right\|_2 \sqrt{n} \le \sqrt{\|J\|_{\text{op}}}.$$

Hence taking $L = \Omega\left(\sqrt{\|J\|_{\text{op}}} + 1\right) = \Omega\left(\sqrt{\|J\|_{\text{op}}} + \sqrt{\log\left(\frac{n}{\varepsilon}\right)/n}\right)$, we have

$$P(y \notin [-L, L]^d) \le \sum_{i=1}^{n} P(y_i \notin [-L, L]) = n \cdot \frac{\varepsilon}{2n} = \frac{\varepsilon}{2}. \tag{12}$$

Putting everything together and using Theorem C.1, we have with probability $\ge 1 - \delta$ that

$$\left(\frac{n}{2\pi}\right)^{\frac{d}{2}} \sum_{y^* \in \text{Grid}^d_{L,\eta}} \widehat{Z}(y^*) \in [e^{-\varepsilon}, e^{\frac{\varepsilon}{2}}] \cdot Z_{J^\parallel, J^\perp, h}.$$

■

C.1. Estimation with positive and negative spikes

We now analyze Algorithm 3 when there are negative spikes to prove Theorem 1.1(1).

For $y^* \in \text{Grid}^d_{L,\eta}$, Algorithm 3 uses Corollary B.2 to find $\mu(y^*)$ such that

$$\log\left(\frac{dP_{J_+, X^\top Q y^* + h}}{dP_{J, X^\top Q y^* + h + \mu(y^*)}}\right) \le c \operatorname{Tr}(J_-) + 1. \tag{13}$$

Let $J^\perp_{\text{all}} = J^\perp - J_- = J - J^\parallel$. We first calculate

$$\begin{aligned}
\mathbb{E}_{P_M} g_M &= \frac{1}{Z_{J^\perp, h(y^*)}} \sum_{\sigma \in \{\pm 1\}^n} \exp\left(\frac{1}{2}\left\langle \sigma, J^\perp \sigma \right\rangle + \langle h(y^*), \sigma \rangle\right) \frac{\exp\left(-\frac{1}{2}\langle \sigma, J_- \sigma \rangle\right)}{\exp\left(\langle \mu(y^*), \sigma \rangle\right)} \\
&\quad \cdot \int_{B(y^*)} \exp\left(\left\langle X^\top Q(y - y^*), \sigma \right\rangle - \frac{n}{2}\|y\|^2\right) dy \\
&= \frac{1}{Z_{J^\perp, h(y^*)}} \sum_{\sigma \in \{\pm 1\}^n} \int_{B(y^*)} \exp\left(\frac{1}{2}\left\langle \sigma, (J^\perp_{\text{all}} - J_-)\sigma \right\rangle + \left\langle X^\top Q y + h, \sigma \right\rangle - \frac{n}{2}\|y\|^2\right) dy \\
&= \frac{\int_{B(y^*)} Z_{J^\parallel, J^\perp, h}(Qy)\, dy}{Z_{J^\perp, h(y^*)}}
\end{aligned} \tag{14}$$

as before.

We now bound $\frac{\mathbb{E} g_M^2}{(\mathbb{E} g_M)^2}$. First we bound

$$\begin{aligned}
g_M(\sigma) &= \frac{\exp\left(-\frac{1}{2}\langle \sigma, J_- \sigma \rangle\right)}{\exp\left(\langle \mu(y^*), \sigma \rangle\right)} \int_{B(y^*)} \exp\left(\left\langle X^\top Q(y - y^*), \sigma \right\rangle - \frac{n}{2}\|y\|^2\right) dy \\
&= \frac{\exp\left(\frac{1}{2}\langle \sigma, J^\perp_{\text{all}} \sigma \rangle + \langle X^\top Q y^* + h, \sigma \rangle\right)}{\exp\left(\frac{1}{2}\langle \sigma, J^\perp \sigma \rangle + \langle h(y^*), \sigma \rangle\right)} \cdot \int_{B(y^*)} \exp\left(\left\langle X^\top Q(y - y^*), \sigma \right\rangle - \frac{n}{2}\|y\|^2\right) dy \\
&\le \exp\left(c \operatorname{Tr}(J_-) + 1\right) \frac{Z_{J^\perp_{\text{all}}, X^\top Q y^* + h}}{Z_{J^\perp, h(y^*)}} \exp\left(-\frac{n}{2}\|y^*\|^2\right) \eta^d e^{1/2}
\end{aligned}$$

using (13) and (11). Next, again using (11), we bound

$$\begin{aligned}\mathbb{E}_{P_M} g_M(\sigma) &= \sum_{\sigma\in\{\pm 1\}^n} \frac{\exp\left(\frac{1}{2}\left\langle \sigma, J_{\text{all}}^{\perp}\sigma\right\rangle + \left\langle X^\top Q y^* + h, \sigma\right\rangle\right)}{Z_{J^\perp, h(y^*)}} \\ &\quad \cdot \int_{B(y^*)} \exp\left(\left\langle X^\top Q(y-y^*), \sigma\right\rangle - \frac{n}{2}\|y\|^2\right) dy \\ &\geq \frac{Z_{J_{\text{all}}^\perp, X^\top Q y^* + h}}{Z_{J^\perp, h(y^*)}} \exp\left(-\frac{n}{2}\|y^*\|^2\right) \eta^d e^{-1/2}\end{aligned}$$

Hence

$$\frac{\mathbb{E}_{P_M} g_M^2}{(\mathbb{E}_{P_M} g_M)^2} \leq \exp(2c\,\mathrm{Tr}(J_-) + 4),$$

and this is the extra multiplicative error we incur in estimation. The rest of the estimates in the proof are the same as before.

The above concludes the proof of our main result for computing the partition function. We now briefly discuss the performance of this algorithm under the "naive mean field" assumption $\|J\|_F^2 = o(n)$ referenced in the introduction and introduced in (Basak and Mukherjee, 2017).

Remark C.5 *Suppose we want to bound the performance of the algorithm from Theorem 1.1 in terms of Frobenius norms. This will be very wasteful compared to the original statement, but is useful for comparison.*

For simplicity, we can make the common assumption that the diagonal of J is zero, which means that the sum of the eigenvalues of J is zero. Then we can choose the interval $[-1/3, 1/3]$ as the interval of length at most one in the application of the Theorem. The runtime for estimating $\log Z$ to additive ε error will be at most

$$O\left(\frac{(\|J\|_{\text{op}}\, n)^{O(d_+ + 1)} e^{O(\lambda_1 + \cdots + \lambda_{d_-} - d_-/3)}}{\varepsilon^2}\right)$$

where $-\lambda_1, \ldots, -\lambda_{d_-}$ are the eigenvalues of J below $-1/3$. Now clearly we have $\sum_{i=1}^{d_-} \lambda_i \leq \sum_{i=1}^{d_-} 3\lambda_i^2 \leq 3\|J\|_F^2$ and $d_+ \leq 3\|J\|_F^2$. So we have a crude bound on the runtime as

$$O\left(\frac{(n\|J\|_{\text{op}})^{O(\|J\|_F^2)}}{\varepsilon^2}\right).$$

In particular, provided $\|J\|_F^2 = o(n/\log(n))$ we have that this is subexponential time. So the result works up to almost the same subexponential time regime as the algorithm in the work (Jain et al., 2019) when specialized to the setting of Ising models. Depending on the precise properties of J, the precise runtime of the new algorithm could be faster or slower than the algorithm of (Jain et al., 2019), but the approximation error for this one is much stronger (additive error ε to $\log Z$).

Appendix D. Sampling

We now turn to the problem of generating samples from the model; for the reader, note that this section builds on results and uses notation from the previous section on partition function estimation.

By choosing $y^* \in \text{Grid}_{L,\eta}^d = \left\{-L + \frac{1}{2}\eta, -L + \frac{3}{2}\eta, \ldots, L - \frac{1}{2}\eta\right\}^d$ with probability proportional to $\widehat{Z}(y^*)$ estimated by Algorithm 3 and then sampling from $P^{\sigma,y}_{J^\parallel,J^\perp,h}$ restricted to $\{\pm 1\}^n \times B(y^*)$, we can obtain an algorithm for sampling of the same order of complexity as in Theorem C.2. In this section, we give an algorithm that only has logarithmic dependence on ε and prove Theorem 1.1(2).

Let $Z_\ell(y^*) := Z_{\beta_\ell J^\perp, h(y^*)}$ for $1 \le \ell \le M$ and $Z_{M+1}(y^*) := \int_{B(y^*)} Z_{J^\parallel,J^\perp,h}(Qy)\,dy$. Denote the approximately normalized probabilities

$$q_{\ell,y^*} = \frac{Z_\ell(y^*)}{\widehat{Z}_\ell(y^*)} p_{\beta_\ell J^\perp, h(y^*)}$$

where $\beta_\ell = \frac{\ell-1}{n}$. Overloading notation, we will also write p_{ℓ,y^*} for $p_{\beta_\ell J^\perp,h(y^*)}$. Note that we can compute the ratios of different q_{ℓ,y^*}'s, as we have $q_{\ell,y^*}(\sigma) \propto \frac{1}{\widehat{Z}_\ell(y^*)} \exp\left(\frac{1}{2}\left\langle \sigma, \beta_\ell J^\perp \sigma\right\rangle + \left\langle h(y^*), \sigma\right\rangle\right)$.

We define a Markov chain on an expanded state space $\{1, \ldots, M\} \times \text{Grid}_{L,\eta}^d \times \{\pm 1\}^n$, where the first index denotes the "temperature" of the distribution. This is similar to a simulated tempering chain (Marinari and Parisi, 1992), with two types of moves: between temperatures and within temperatures. However, there are two differences with a standard simulated tempering chain:

1. We use a different normalizing constant $\widehat{Z}_\ell(y^*)$ for each value of y^*, in order to make sure the stationary distribution is roughly uniformly distributed over the $y^* \in \text{Grid}_{L,\eta}^d$.
2. Within any temperature other than the highest one, we do not allow moves that change y^*.

Finally, we do simulated tempering on the space $\text{Grid}_{L,\eta}^d \times \{\pm 1\}^n$ rather than $[-L, L]^d \times \{\pm 1\}^n$ for convenience; this adds an extra rejection sampling step at the end where we compare the distributions on $\{y^*\} \times \{\pm 1\}^n$ and on $B(y^*) \times \{\pm 1\}^n$, similar to the final ratio g_M in partition function estimation.

We need the modifications for technical reasons to make our proof work; it is an interesting question whether a more standard simulated tempering chain would work. Our proof strategy is based on a Markov chain decomposition theorem similar to Ge et al. (2018), which we will now introduce.

Given a Markov chain on Ω, we define two Markov chains associated with a partition of Ω.

Definition D.1 (Madras and Randall (2002)) *For a Markov chain $\mathcal{M} = (\Omega, T)$, and a set $A \subseteq \Omega$, define the **restriction of** $\mathcal{M}$ **to** A to be the Markov chain $\mathcal{M}|_A = (A, T|_A)$, where*

$$T|_A(x, B) = T(x, B) + \mathbb{1}_B(x) T(x, A^c).$$

(In words, $T(x, y)$ proposes a transition, and the transition is rejected if it would leave A.)

*Suppose the unique stationary measure of $\mathcal{M}$ is P. Given a partition $\mathcal{P} = \{A_j : j \in J\}$, define the **projected Markov chain with respect to** $\mathcal{P}$ to be $\overline{\mathcal{M}}^{\mathcal{P}} = (J, \overline{T}^{\mathcal{P}})$, where*

$$\overline{T}^{\mathcal{P}}(i, j) = \frac{1}{P(A_i)} \int_{A_i} \int_{A_j} T(x, dy)\, P(dx).$$

Algorithm 4: Simulated tempering on $\{1, \dots, M+1\} \times \mathrm{Grid}_{L,\eta}^d \times \{\pm 1\}^n$

Input: Ising model (J, h), steps to run Markov chain T (suggested $\Theta(n^4 d \log(n \|J\|_{\mathrm{op}} / \varepsilon))$.
Run Algorithm 3 to obtain partition function estimates $\widehat{Z}_\ell(y^*)$ for $1 < \ell \le M+1 = n+2$.
Let $\ell = 1$. Draw $y^* \in \mathrm{Grid}_{L,\eta}^d = \left\{-L + \frac{1}{2}\eta, -L + \frac{3}{2}\eta, \dots, L - \frac{1}{2}\eta\right\}^d$, and then draw $\sigma \sim P^{\sigma|y}_{O,h(y^*)}(\cdot|y^*)$.
for $1 \le t \le T$ **do**
  With probability $\frac{1}{4}$, if $\ell \ne M$, set $\ell \leftarrow \ell + 1$ with probability $\min\left\{\frac{q_{\ell+1,y^*}(\sigma)}{q_{\ell,y^*}(\sigma)}, 1\right\}$.
  With probability $\frac{1}{4}$, if $\ell \ne 1$, set $\ell \leftarrow \ell - 1$ with probability $\min\left\{\frac{q_{\ell-1,y^*}(\sigma)}{q_{\ell,y^*}(\sigma)}, 1\right\}$.
  With probability $\frac{1}{2}$, **begin**
    if $\ell = 1$ **then**
      With probability $\frac{1}{2}$, reselect a random $y^* \in \mathrm{Grid}_{L,\eta}^d$, and then draw $\sigma \sim P^{\sigma|y}_{O,h(y^*)}(\cdot|y^*)$.
    end
    Choose a random coordinate i, and set $\sigma \leftarrow \sigma^{(i)}$ with probability $\frac{q_{\ell,y^*}(\sigma^{(i)})}{q_{\ell,y^*}(\sigma^{(i)}) + q_{\ell,y^*}(\sigma)}$.
  end
end
if $\ell = M$ **then**
  Draw $U \sim \mathsf{Uniform}([0,1])$.
  if $U \le (4e \max \widehat{Z}_{M+1}(y^*) \exp(c \operatorname{Tr}(J_-) + 1))^{-1} \widehat{Z}_M(y^*) g_{n+1,y^*}(\sigma)$ **then**
    Return σ.
  end
end
If failed to return sample, re-run the procedure.

(In words, $\overline{T}(i,j)$ is the "total probability flow" from A_i to A_j.) We omit the superscript $\mathcal{P}$ when it is clear.

The following theorem lower-bounds the gap of the original chain in terms of the gap of the projected chain and the minimum gap of the restricted chains.

Theorem D.2 (Madras and Randall (2002)) *Let $\mathcal{M} = (\Omega, T)$ be a Markov chain with stationary measure P. Let $\mathcal{P} = \{A_j : j \in J\}$ be a partition of Ω such that $P(A_j) > 0$ for all $j \in J$. Then*

$$\frac{1}{2} \operatorname{Gap}(\overline{\mathcal{M}}^{\mathcal{P}}) \min_{j \in J} \operatorname{Gap}(\mathcal{M}|_{A_j}) \le \operatorname{Gap}(\mathcal{M}) \le \operatorname{Gap}(\overline{\mathcal{M}}^{\mathcal{P}}).$$

We can now prove our main theorem for sampling.

Proof [Proof of Theorem 1.1(2)] Let $\mathcal{M}$ be the simulated chain in Algorithm 4. Below, we condition on the event that all the $\widehat{Z}_\ell(y^*)$ are 2-multiplicative approximations of $Z_\ell(y^*)$, that is, $\widehat{Z}_\ell(y^*) \in [\frac{1}{2}, 2] \cdot Z_\ell(y^*)$. As in the proof of Theorem C.2, if we choose the failure probability to be $O\left(\frac{\varepsilon}{M}\left(\frac{\eta}{2L}\right)^d\right)$, by Lemma C.4 and a union bound—this time applied to the estimates at all levels $\widehat{Z}_\ell(y^*)$—this event happens with probability $1 - O(\varepsilon)$.

We let P^{st} denote the stationary measure for the simulated tempering chain, and P^{st}_ℓ denote the measure restricted to $\{\ell\} \times \text{Grid}^d_{L,\eta} \times \{\pm 1\}^n$.

We use Theorem D.2 with the partition given by $A_{\ell,y^*} = \{\ell\} \times \{y^*\} \times \{\pm 1\}^n$. The restriction $\mathcal{M}|_{A_{\ell,y^*}}$ is a lazy version of the Glauber dynamics chain for P_{ℓ,y^*} (that is, with all transition probabilities halved, or multiplied by $\frac{1}{4}$ in the case $\ell = 1$), which has Poincaré constant bounded by $O(n^2)$ by Lemma C.3.

First, note that by construction with the Metropolis-Hastings acceptance ratio, the stationary distribution satisfies

$$\overline{p}((\ell, y^*)) \propto R_\ell(y^*) := \frac{Z_\ell(y^*)}{\widehat{Z}_\ell(y^*)}. \tag{15}$$

For the projected chain, we use Lemma F.3. We check each of the conditions.

1. To bound the "bottleneck ratio", note that for $k < \ell$, letting $\overline{p}_j(y^*) = \overline{p}(y^*|j) = \frac{\overline{p}((j,y^*))}{\sum_{y \in \text{Grid}^d_{L,\eta}} \overline{p}((j,y))}$

$$\frac{\overline{p}_k(y^*)}{\overline{p}_\ell(y^*)} = \frac{R_k(y^*)/\sum_{y \in \text{Grid}^d_{L,\eta}} R_k(y)}{R_\ell(y^*)/\sum_{y \in \text{Grid}^d_{L,\eta}} R_\ell(y)} \geq \frac{1}{4}$$

   using the fact that the $\widehat{Z}_j(y)$ are 2-multiplicative approximations, so that $R_j(y^*) \in [\frac{1}{2}, 2]$ for each j, y^*.

2. From (15), we have $\frac{\overline{p}((\ell,y^*))}{\overline{p}((\ell,y^*))} \in [\frac{1}{4}, 4]$. Note that for $\ell, \ell \pm 1 \in [M]$,

$$\frac{p_{\ell\pm1,y^*}(\sigma)}{p_{\ell,y^*}(\sigma)} = \exp(\langle \sigma, (\beta_{\ell\pm1} - \beta_\ell) J\sigma \rangle) \frac{Z_\ell(y^*)}{Z_{\ell\pm1}(y^*)} = \Theta(1)$$

   because the ratio of individual terms in Z_{ℓ,y^*} and $Z_{\ell\pm1,y^*}$ is $\Theta(1)$. Hence

$$\begin{aligned}
\overline{T}((\ell, y^*), (\ell \pm 1, y^*)) &= \sum_{\sigma \in \{\pm 1\}^n} \min\left\{ \frac{\widehat{Z}_\ell(y^*)/Z_\ell(y^*)}{\widehat{Z}_{\ell\pm1}(y^*)/Z_{\ell\pm1}(y^*)} \cdot \frac{p_{\ell\pm1,y^*}(\sigma)}{p_{\ell,y^*}(\sigma)}, 1 \right\} p_{\ell,y^*}(\sigma) \\
&\geq \frac{1}{4} \sum_{\sigma \in \{\pm 1\}^n} \min\left\{ \frac{p_{\ell\pm1,y^*}(\sigma)}{p_{\ell,y^*}(\sigma)}, 1 \right\} p_{\ell,y^*}(\sigma) \\
&= \Omega(1) = \Omega\left(\frac{\overline{p}((\ell \pm 1, y^*))}{\overline{p}((\ell, y^*))} \right)
\end{aligned}$$

   where we used the fact that $\widehat{Z}_\ell(y^*)$ are 2-multiplicative approximations. We also note

$$\overline{T}((1, y^*), (1, z^*)) \geq \frac{1}{4} \left(\frac{\eta}{2L} \right)^d.$$

   Hence, condition 1 of Lemma F.3 holds with constant D_{high} and D_{adj}.

3. Finally, for any $1 \leq \ell \leq M$,

$$P\left(\{\ell\} \times \text{Grid}^d_{L,\eta} \times \{\pm 1\}^n \right) = \frac{\sum_{y \in \text{Grid}^d_{L,\eta}} R_\ell(y)}{\sum_{\ell=1}^M \sum_{y \in \text{Grid}^d_{L,\eta}} R_\ell(y)} \geq \frac{1}{4M}.$$

Hence by Lemma F.3, the Poincaré constant of $\mathcal{M}$ is $O(M^2) = O(n^2)$. Since $\mathcal{M}|_{A_{\ell,y^*}}$ have Poincaré constant bounded by $O(n^2)$ for each y^*, noting the spectral gap is the inverse of the Poincaré constant and using Lemma D.2, we get that the Poincaré constant of $\mathcal{M}$ is $C_{\mathrm{P}} = O\left(n^2 \cdot n^2\right) = O\left(n^4\right)$. For the mixing time, note that the starting distribution is the restriction of the stationary distribution to $\{1\} \times \mathrm{Grid}_{L,\eta}^d \times \{\pm 1\}^n$, which has at least $\frac{1}{4M}$ of the mass. Hence the time until the distribution is ε-close to the stationary distribution (and all restrictions to $\{\ell\} \times \mathrm{Grid}_{L,\eta}^d \times \{\pm 1\}^n$, $1 \le \ell \le M$ are ε-close) is $O\left(C_{\mathrm{P}} \log\left(\frac{M}{\varepsilon}\right)\right)$.

Let P_{M+1} be the probability measure on $\{\pm 1\}^n \times \mathrm{Grid}_{L,\eta}^d$ with probability mass function given by

$$p_{M+1}(\sigma, y^*) = \frac{\int_{B(y^*)} \exp\left(\frac{1}{2}\left\langle \sigma, J_{\mathrm{all}}^{\perp}\sigma\right\rangle + \left\langle X^\top Q y + h, \sigma\right\rangle - \frac{n}{2}\|y\|^2\right) dy}{\int_{[-L,L]^d} \sum_{\sigma \in \{\pm 1\}^d} \exp\left(\frac{1}{2}\left\langle \sigma, J_{\mathrm{all}}^{\perp}\sigma\right\rangle + \left\langle X^\top Q y + h, \sigma\right\rangle - \frac{n}{2}\|y\|^2\right) dy};$$

that is, it is obtained from restricting $p^{\sigma,y}_{J_{\mathrm{all}}^{\perp}, X^\top Q y + h}(\sigma, y)$ to $\{\pm 1\}^n \times [-L, L]^d$ and then rounding y to the nearest grid point. Except for the fact that this measure is restricted to $[-L, L]^d$, this is the distribution we wish to sample from. We also know that

$$p_M^{\mathrm{st}}(\sigma, y) = \left(\widehat{Z}_M(y^*) \sum_{y \in \mathrm{Grid}_{L,\eta}^d} R_M(y)\right)^{-1} \exp\left(\frac{1}{2}\left\langle \sigma, J^{\perp}\sigma\right\rangle + \left\langle h(y^*), \sigma\right\rangle\right).$$

In terms of $\frac{p_{M+1}(\sigma,y)}{p_M^{\mathrm{st}}(\sigma,y)}$, the acceptance ratio in Algorithm 4 is given by

$$\begin{aligned}
&(4e \max_{y^* \in \mathrm{Grid}_{L,\eta}^d} \widehat{Z}_{M+1}(y^*) \exp(c \operatorname{Tr}(J_-) + 1))^{-1} \widehat{Z}_M(y^*) g_{n+1,y^*}(\sigma) \\
&= (4e \max_{y^* \in \mathrm{Grid}_{L,\eta}^d} \widehat{Z}_{M+1}(y^*) \exp(c \operatorname{Tr}(J_-) + 1))^{-1} \cdot \frac{\int_{[-L,L]^d} Z_{J^{\parallel}, J_{\mathrm{all}}^{\perp}, h}(Qy)\, dy}{\sum_{y \in \mathrm{Grid}_{L,\eta}^d} R_M(y)} \cdot \frac{p_{M+1}(\sigma, y^*)}{p_M^{\mathrm{st}}(\sigma, y^*)}
\end{aligned} \tag{16}$$

This is a constant times $\frac{p_{M+1}(\sigma,y^*)}{p_M^{\mathrm{st}}(\sigma,y^*)}$, so it is the correct rejection sampling ratio. We need to show that this is always at most 1, and give a lower bound for the coefficient of $\frac{p_{M+1}(\sigma,y)}{p_M^{\mathrm{st}}(\sigma,y)}$.

1. Ratio is at most 1: We first consider

$$\begin{aligned}
\frac{p_{M+1}(\sigma, y^*)}{p_M^{\mathrm{st}}(\sigma, y^*)} &= \frac{p_{J_{\mathrm{all}}^{\perp}, X^\top Q y^* + h}(\sigma)}{p_{J^{\perp}, h(y^*)}(\sigma) p_M^{\mathrm{st}}(y^*)} \cdot \frac{p_{M+1}(\sigma|y^*) p_{M+1}(y^*)}{p_{J_{\mathrm{all}}^{\perp}, X^\top Q y^* + h}(\sigma)} \\
&= \frac{p_{J_{\mathrm{all}}^{\perp}, X^\top Q y^* + h}(\sigma)}{p_{J^{\perp}, h(y^*)}(\sigma) \frac{R_M(y^*)}{\sum_{y \in \mathrm{Grid}_{L,\eta}^d} R_M(y)}} \cdot \frac{p_{M+1}(\sigma|y^*)}{p_{J_{\mathrm{all}}^{\perp}, X^\top Q y^* + h}(\sigma)} \frac{\int_{B(y^*)} Z_{J^{\parallel}, J^{\perp}, h}(Qy)\, dy}{\int_{[-L,L]^d} Z_{J^{\parallel}, J^{\perp}, h}(Qy)\, dy} \\
&\le 2 \cdot \left(\sum_{y \in \mathrm{Grid}_{L,\eta}^d} R_M(y)\right) \exp(c \operatorname{Tr}(J_-) + 1) \cdot \frac{p_{M+1}(\sigma|y^*)}{p_{J_{\mathrm{all}}^{\perp}, X^\top Q y^*}(\sigma)} \frac{\int_{B(y^*)} Z_{J^{\parallel}, J^{\perp}, h}(Qy)\, dy}{\int_{[-L,L]^d} Z_{J^{\parallel}, J^{\perp}, h}(Qy)\, dy}
\end{aligned}$$

where we used the guarantee obtained from Corollary B.2. We also note

$$\frac{p_{M+1}(\sigma|y^*)}{p_{J_{\text{all}}^\perp, X^\top Q y^* + h}(\sigma)} \propto \int_{B(y^*)} \exp\left(\left\langle X^\top Q(y-y^*)\right\rangle - \frac{n}{2}\|y\|^2\right) dy$$
$$\in \exp\left(-\frac{n}{2}\|y^*\|^2\right) \cdot [e^{-1/2}, e^{1/2}]$$

by (11); hence, because probabilities integrate to 1, the ratio is bounded by e. Combining with (16), we obtain that the acceptance ratio is bounded by

$$\frac{1}{2} \cdot \frac{\int_{B(y^*)} Z_{J^\parallel, J^\perp, h}(Qy)\, dy}{\max_{y^* \in \text{Grid}_{L,\eta}^d} \widehat{Z}_{M+1}(y^*)} \le \frac{1}{2} \cdot \frac{\int_{B(y^*)} Z_{J^\parallel, J^\perp, h}(Qy)\, dy}{\frac{1}{2}\max_{y^* \in \text{Grid}_{L,\eta}^d} Z_{M+1}(y^*)} \le 1.$$

2. Lower bound for coefficient: The reciprocal of the coefficient is

$$4e \frac{\max \widehat{Z}_{M+1}(y^*)}{\int_{[-L,L]^d} Z_{J^\parallel, J_{\text{all}}^\perp, h}(Qy)\, dy} \exp(c \operatorname{Tr}(J_-) + 1) \cdot \sum_{y \in \text{Grid}_{L,\eta}^d} R_M(y)$$
$$\le 4e \cdot \exp(c \operatorname{Tr}(J_-) + 1) \cdot 2 \max_{y^* \in \text{Grid}_{L,\eta}^d} \frac{\int_{B(y^*)} Z_{J^\parallel, J_{\text{all}}^\perp, h}(Qy)\, dy}{\int_{[-L,L]^d} Z_{J^\parallel, J_{\text{all}}^\perp, h}(Qy)\, dy} \cdot 2\left(\frac{2L}{\eta}\right)^d$$
$$\le 16 \exp(c \operatorname{Tr}(J_-) + 2)\left(\frac{2L}{\eta}\right)^d.$$

Thus we can apply Lemma F.2 with $C = 16\exp(c\operatorname{Tr}(J_-)+2)\left(\frac{2L}{\eta}\right)^d$. Replacing ε with $\frac{\varepsilon}{C}$, we get that the distribution restricted to $\{M\} \times \text{Grid}_{L,\eta}^d \times \{\pm 1\}^n$ after running for $\Omega\left(C_{\text{P}} \log\left(\frac{MC}{\varepsilon}\right)\right)$ steps is $\frac{\varepsilon}{4C}$ close to P_M^{st} in TV-distance. By Lemma F.2, an accepted sample will be $\frac{\varepsilon}{2}$ close to P_{M+1}. Finally, because L was chosen large enough so that $P(y \notin [-L,L]^d) \le \frac{\varepsilon}{4}$ as in (12), we conclude that the marginal distribution of σ is ε-close to $P_{J,h}$. The expected number of trials until acceptance will be $O(CM) = O(n\exp(c\operatorname{Tr}(J_-))(2L/\eta)^d)$. ■

Appendix E. Interpreting the Hubbard-Stratonovich transform as as Gaussian mixture posterior

In this Appendix, we discuss at length the properties of the Hubbard-Stratonovich transform and its possible interpretation as a Gaussian mixture model posterior. For the most part (and unlike all of the other appendices in this paper) this discussion is pedagogical, though some simple formulas stated here are used elsewhere in the paper.

Throughout this section, we consider the case when J is positive semi-definite (PSD). In this case, we can write $J = \frac{1}{n}X^\top X$ for $X \in \mathbb{R}^{d \times n}$, for $d = \operatorname{rank}(J) \le n$. Let $x_1, \ldots, x_n$ be the columns of X; we will re-interpret the Hubbard-Stratonovich transform as giving the posterior of a Gaussian mixture model after seeing samples $x_1, \ldots, x_n$. (The precise model is a very slight variant of the Gaussian mixture model described in the main text and applications sections.) We consider

the following augmented model, which is a density on $\{\pm 1\}^n \times \mathbb{R}^d$:

$$p_{X,h}(\sigma, \mu) = \frac{1}{Z^{\text{joint}}_{J,h}} \prod_{i=1}^{n} \exp\left(-\frac{1}{2}\left\|\sigma_i x_i - \mu\right\|^2 + h_i \sigma_i\right) \tag{17}$$

$$\text{where } Z^{\text{joint}}_{J,h} = \int_{\mathbb{R}^d} \sum_{\sigma \in \{\pm 1\}^n} \exp\left(-\frac{1}{2}\left\|\sigma_i x_i - \mu\right\|^2 + h_i \sigma_i\right) d\mu. \tag{18}$$

(As we will see below, $Z^{\text{joint}}_{J,h}$ does not depend on the choice of X.) Note this can be interpreted as the posterior distribution for a Gaussian mixture model (with two components, symmetric around 0 with identity covariance) $p(x|\mu) \propto \exp\left(-\frac{1}{2}\left\|x - \mu\right\|^2\right) + \exp\left(-\frac{1}{2}\left\|x + \mu\right\|^2\right)$ with uniform prior on μ and prior on σ given by $p_{\text{prior}}(\sigma) \propto e^{\langle h, \sigma \rangle}$, where σ represents the class assignments (to the Gaussian with mean μ or mean $-\mu$).

We summarize the connection in this lemma. We will drop the subscripts J, h when they are clear.

Lemma E.1 *Consider the distribution $p_{X,h}(\sigma, \mu)$ in (17) and let $J = \frac{1}{n} X^\top X$. The following hold:*

1. *The marginal distribution of σ is $p_{J,h}(\sigma)$ (in (1)).*
2. *The marginal distribution on μ is*

$$p(\mu) \propto e^{-\frac{n}{2}\|\mu\|^2} \prod_{i=1}^{n} \cosh(\langle x_i, \mu \rangle + h_i).$$

3. *The conditional distribution of σ given μ is a product distribution,*

$$p(\sigma|\mu) \propto \prod_{i=1}^{n} \exp\left(\sigma_i(\langle x_i, \mu \rangle + h_i)\right).$$

4. *The conditional distribution of μ given σ is a Gaussian distribution,*

$$p(\mu|\sigma) = \left(\frac{n}{2\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{n}{2}\left\|\mu - \frac{\sum_{i=1}^n \sigma_i x_i}{n}\right\|^2\right).$$

5. *The partition functions are related via*

$$Z^{\text{joint}}_{J,h} = \left(\frac{2\pi}{n}\right)^{n/2} \exp\left(-\frac{n}{2}\operatorname{Tr}(J)\right) Z_{J,h}.$$

As a consequence, to sample from $p(\sigma)$, it suffices to sample μ from the above distribution, and then sample μ conditional on μ (which is immediate).

We calculate the Hessian of $-\ln p(\mu)$:

$$-\nabla^2 \ln p(\mu) = nI - \sum_{i=1}^{n} x_i x_i^\top + \sum_{i=1}^{n} (1 - \operatorname{sech}^2(\langle x_i, \mu \rangle + h_i)) x_i x_i^\top.$$

Note that this is convex (and hence $p(\mu)$ is log-concave) when $J = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \preceq I$. This observation can be used to infer an efficient sampling algorithm for $p_{J,h}$ by first drawing a sample from $p(\mu)$ (using algorithms for log-concave sampling such as Langevin dynamics (Durmus et al., 2019)) and then drawing from $p(\sigma|\mu)$, as observed in Bauerschmidt and Bodineau (2019). This gives an alternative algorithm to the Glauber dynamics (which mix rapidly under the same assumption (Anari et al., 2021)), albeit one which is not as fast.

We note that our decomposition is similar, but slightly different from the decomposition in Bauerschmidt and Bodineau (2019). Both approaches decompose $p_{J,h}$ as a log-concave mixture of product distributions when $J \preceq I$. Our approach has the advantage that when J has a few large eigenvalues (eigenvalues greater than 1), the distribution on μ is still log-concave in the other directions. We note the log-concave decomposition technique was used extensively in analysis of the Hopfield model (Bovier and Picco, 1998; Talagrand, 2010).

Proof

1. The marginal distribution of σ is

$$\frac{1}{Z^{\text{joint}}_{J,h}} \int_{\mathbb{R}^d} \prod_{i=1}^n \exp\left(-\frac{1}{2}\left\|\sigma_i x_i - \mu\right\|^2 + h_i\sigma_i\right) d\mu = \frac{1}{Z^{\text{joint}}_{J,h}} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\sum_{i=1}^n \left\|\sigma_i x_i - \mu\right\|^2 + h_i\sigma_i\right) d\mu$$

$$= \frac{1}{Z^{\text{joint}}_{J,h}} \int_{\mathbb{R}^d} \exp\left(-\frac{n}{2}\left\|\mu - \frac{\sum_{i=1}^n \sigma_i x_i}{n}\right\|^2 + \frac{1}{2n}\sum_{i,j=1}^n \sigma_i x_i x_j^\top \sigma_j - \frac{1}{2}\sum_{i=1}^n \|x_i\|^2 + \langle h, \sigma\rangle\right) d\mu$$

$$= \frac{1}{Z^{\text{joint}}_{J,h}} \left(\frac{2\pi}{n}\right)^{n/2} \exp\left(-\frac{1}{2}\|X\|_F^2\right) \exp\left(\frac{1}{2}\sigma^\top\left(\frac{XX^\top}{n}\right)\sigma + \langle h, \sigma\rangle\right), \tag{19}$$

   where the last line uses the fact that the integral of $\exp\left(-\frac{n}{2}\|\mu - \mu_0\|^2\right)$ is a fixed normalizing constant, for any μ_0. Finally, we use $J = \frac{1}{n}XX^\top$.

2. This follows from factoring the product,

$$p(\mu) \propto \sum_{\sigma\in\{\pm1\}^n} \prod_{i=1}^n \exp\left(-\frac{1}{2}\left\|\sigma_i x_i - \mu\right\|^2 + h_i\sigma_i\right) \propto \prod_{i=1}^n \sum_{\sigma_i=\pm1} \exp\left(-\frac{1}{2}\left\|\sigma_i x_i - \mu\right\|^2 + h_i\sigma_i\right)$$

$$\propto e^{-\frac{1}{2}\|\mu\|^2} \prod_{i=1}^n \sum_{\sigma_i=\pm1} e^{\sigma_i(\langle x_i,\mu\rangle + h_i)} \propto e^{-\frac{n}{2}\|\mu\|^2} \prod_{i=1}^n \cosh(\langle x_i, \mu\rangle + h_i).$$

3–4. These follow directly by noting $p(\sigma|\mu) \propto p(\sigma, \mu)$ for fixed μ, and $p(\mu|\sigma) \propto p(\sigma, \mu)$ for fixed σ.

5. This follows from comparing normalizing constants in (19).

■

Lemma E.1 gives a decomposition of $p_{J,h}$ into a mixture of product distributions $p_{J,h}(\sigma) = \int_{\mathbb{R}^d} p(\sigma|\mu)p(\mu)\,d\mu$. We can instead only condition on the projection of μ to a rank-d subspace V and obtain a decomposition in terms of rank-$(n-d)$ Ising models. We will choose the rank-d subspace to contain the eigenvectors of J with large eigenvalue.

We define the distribution $p_{X,h,V}(\sigma, \mu^{\parallel}, \mu^{\perp})$ on $\{\pm 1\}^n \times V \times V^{\perp}$ by $p_{X,h,V}(\sigma, \mu^{\parallel}, \mu^{\perp}) = p_{X,h}(\sigma, \mu^{\parallel} + \mu^{\perp})$.

Lemma E.2 *Consider the distribution $p(\sigma, \mu^{\parallel}, \mu^{\perp})$. Let $P^{\parallel}$ and $P^{\perp}$ be the projections onto V and $V^{\perp}$, respectively and let $J^{\parallel} = \frac{1}{n} X^{\top} P^{\perp} X$, $J^{\perp} = J - J^{\parallel}$.*

1. *The joint distribution of $(\sigma, \mu^{\parallel})$ is given by*

$$p(\sigma, \mu^{\parallel}) = \frac{1}{Z^{\text{joint}}_{J,h}} \left(\frac{2\pi}{n}\right)^{(n-d)/2} \exp\left(-\frac{n}{2}\operatorname{Tr}(J)\right) \cdot \exp\left(\frac{1}{2}\left\langle \sigma, J^{\perp}\sigma\right\rangle + \left\langle X^{\top}P^{\parallel}\mu + h, \sigma\right\rangle - \frac{n}{2}\left\|\mu^{\parallel}\right\|^2\right).$$

2. *The distribution of σ given $\mu^{\parallel}$ is*

$$p(\sigma|\mu^{\parallel}) = p_{J^{\perp}, h + X^{\top}\mu^{\parallel}}(\sigma) \propto \exp\left(\frac{1}{2}\left\langle \sigma, J^{\perp}\sigma\right\rangle + \left\langle h + X^{\top}\mu^{\parallel}, \sigma\right\rangle\right).$$

3. *The distribution of $\mu^{\parallel}$ given σ is Gaussian,*

$$p(\mu^{\parallel}|\sigma) = \left(\frac{n}{2\pi}\right)^{\frac{d}{2}} \exp\left(-\frac{n}{2}\left\|\mu^{\parallel} - \frac{\sum_{i=1}^n \sigma_i P^{\parallel} x_i}{n}\right\|^2\right).$$

4. *Let $Z_{J^{\parallel}, J^{\perp}, h} := \int_{\mu^{\parallel} \in V} Z_{J^{\parallel}, J^{\perp}, h}(\mu^{\parallel})\, d\mu^{\parallel}$ where*

$$Z_{J^{\parallel}, J^{\perp}, h}(\mu^{\parallel}) := Z_{J^{\perp}, X^{\top}P^{\parallel}\mu + h} \exp\left(-\frac{n}{2}\left\|\mu^{\parallel}\right\|^2 d\mu^{\parallel}\right) \tag{20}$$

$$= \sum_{\sigma \in \{\pm 1\}^n} \exp\left(\frac{1}{2}\left\langle \sigma, J^{\perp}\sigma\right\rangle + \left\langle X^{\top}P^{\parallel}\mu + h, \sigma\right\rangle - \frac{n}{2}\left\|\mu^{\parallel}\right\|^2\right). \tag{21}$$

Then we have

$$Z^{\text{joint}}_{J,h} = \left(\frac{2\pi}{n}\right)^{d/2} \exp\left(-\frac{n}{2}\operatorname{Tr}(J^{\perp})\right) Z_{J^{\parallel}, J^{\perp}, h}. \tag{22}$$

Proof

1. We integrate $p(\sigma, \mu^{\parallel}, \mu^{\perp})$ along $V^{\perp}$ and complete the square in $\mu^{\perp}$; integrating gives a $\left(\frac{2\pi}{n}\right)^{(n-d)/2}$ normalizing constant:

$$\begin{aligned}
p(\sigma, \mu^{\parallel}) &= \int_{\mu^{\perp} \in V^{\perp}} p(\sigma, \mu^{\parallel}, \mu^{\perp})\, d\mu^{\perp} \\
&= \frac{1}{Z^{\text{joint}}_{J,h}} \int_{\mu^{\perp} \in V^{\perp}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \left(\left\| \sigma_i P^{\perp} x_i - \mu^{\perp} \right\|^2 + \left\| \sigma_i P^{\parallel} x_i - \mu^{\parallel} \right\|^2 \right) + \langle h, \sigma \rangle \right) d\mu_{\perp} \\
&= \frac{1}{Z^{\text{joint}}_{J,h}} \int_{\mu^{\perp} \in V^{\perp}} \exp\left(-\frac{1}{2} \left(n \left\| \mu^{\perp} \right\|^2 - \sum_{i=1}^{n} \sigma_i \left\langle \mu^{\perp}, P^{\perp} x_i \right\rangle + \sum_{i=1}^{n} \left\| P^{\perp} x_i \right\|^2 \right.\right. \right. \\
&\qquad \left.\left.\left. + \sum_{i=1}^{n} \left\| \sigma_i P^{\parallel} x_i - \mu^{\parallel} \right\|^2 \right) + \langle h, \sigma \rangle \right) d\mu_{\perp} \\
&= \frac{1}{Z^{\text{joint}}_{J,h}} \int_{\mu^{\perp} \in V^{\perp}} \exp\left(-\frac{n}{2} \left\| \mu^{\perp} - \frac{1}{n} \sum_{i=1}^{n} \sigma_i P^{\perp} x_i \right\|^2 + \frac{1}{2n} \left\| \sum_{i=1}^{n} \sigma_i P^{\perp} x_i \right\| \right. \\
&\qquad \left. - \frac{1}{2} \sum_{i=1}^{n} \left\| P^{\perp} x_i \right\|^2 - \frac{1}{2} \sum_{i=1}^{n} \left(\left\| P^{\parallel} x_i \right\|^2 - \left\langle X^{\top} P^{\parallel} \mu, \sigma \right\rangle + \left\| \mu^{\parallel} \right\|^2 \right) + \langle h, \sigma \rangle \right) d\mu_{\perp} \\
&= \frac{1}{Z^{\text{joint}}_{J,h}} \left(\frac{2\pi}{n} \right)^{(n-d)/2} \exp\left(\frac{1}{2} \left\langle \sigma, \frac{X^{\top} P^{\perp} X}{n} \sigma \right\rangle - \frac{1}{2} \left\| X \right\|_F^2 + \left\langle X^{\top} P^{\parallel} \mu + h, \sigma \right\rangle - \frac{n}{2} \left\| \mu^{\parallel} \right\|^2 \right)
\end{aligned}$$

   Finally, we rewrite in terms of $J^{\perp}$ by using $J^{\perp} = \frac{1}{n} X^{\top} P^{\perp} X$.

2. This follows from fixing $\mu^{\parallel}$ in the joint probability density and expanding.

3. This follows from fixing σ in the joint density, expanding, and completing the square in $\mu^{\parallel}$.

4. This follows from setting the integral of the joint density equal to 1.

■

Finally, we note that although the interpretation as a Gaussian mixture posterior only makes sense when J is positive semi-definite, the decomposition still works for general symmetric J, as we can multiply the distribution by $\exp\left(-\frac{1}{2} \langle \sigma, J_{-} \sigma \rangle\right)$. We note that combining Lemma E.1, part 5, with Lemma E.2, part 4, gives us Theorem C.1 in the PSD case.

Appendix F. Technical lemmas for partition function estimation and sampling

In this section, we collect some technical lemmas we will need for analyzing our algorithms for partition function estimation and sampling.

F.1. Simulated annealing

For partition function estimation, we use the following lemma, which roughly says that when the variance of some random variables are close to 1, then the variance is additive under multiplication.

Lemma F.1 ((Ge et al., 2020, Lemma B.2), cf. Dyer and Frieze (1991)) *Let Y_ℓ, $\ell = 1, \dots, M$ be independent variables and let $\overline{Y}_\ell = \mathbb{E}Y_\ell$. Assume there exists $\eta > 0$ such that $\eta M \leq \frac{1}{5}$ and*

$$\mathbb{E}Y_\ell^2 \leq (1+\eta)\overline{Y}_\ell^2,$$

then for any $\varepsilon > 0$

$$\mathbb{P}\left(\frac{\left|Y_1 \cdots Y_M - \overline{Y}_1 \cdots \overline{Y}_M\right|}{\overline{Y}_1 \cdots \overline{Y}_M} \geq \frac{\varepsilon}{2}\right) \leq \frac{5\eta M}{\varepsilon^2}.$$

Proof [Proof of Lemma C.4] Let $Y_\ell = \mathbb{E}_{P_\ell}\frac{q_\ell}{p_\ell} = \frac{\int_\Omega q_{\ell+1}}{\int_\Omega q_\ell}$. By Lemma F.1 with $\eta = \frac{\sigma^2}{N}$,

$$\mathbb{P}\left(\prod_{\ell=1}^M Y_\ell \notin [e^{\varepsilon/2}, e^{\varepsilon/2}] \cdot \prod_{\ell=1}^M \overline{Y}_\ell\right) \leq \mathbb{P}\left(\frac{\left|\prod_{\ell=1}^M Y_\ell - \prod_{\ell=1}^M \overline{Y}_\ell\right|}{\prod_{\ell=1}^M \overline{Y}_\ell} \geq \frac{\varepsilon}{4}\right) \leq \frac{80\eta M}{\varepsilon^2} \leq \frac{1}{4}. \tag{23}$$

Now we consider the bias. We have

$$\mathbb{E}_{\widetilde{P}_\ell} g_\ell(x) \in \left[1 - \frac{\varepsilon}{4M}, 1 + \frac{\varepsilon}{4M}\right] \cdot \mathbb{E}_{P_\ell} g_\ell(x) \subseteq [e^{-\frac{\varepsilon}{2M}}, e^{\frac{\varepsilon}{2M}}] \cdot \mathbb{E}_{P_\ell} g_\ell(x).$$

Taking a product, we obtain

$$Z_1 \prod_{\ell=1}^M \overline{Y}_\ell \in \left[e^{-\frac{\varepsilon}{2}}, e^{\frac{\varepsilon}{2}}\right] \cdot Z. \tag{24}$$

Putting together (23) and (24), we obtain that for any r,

$$\mathbb{P}\left(\widehat{Z}^r \notin [e^{-\varepsilon}, e^{\varepsilon}]Z\right) \leq \frac{1}{4}.$$

The algorithm takes the median in order to boost this probability. As the median of R independent runs, $\widehat{Z}$ will fail to be contained in $[e^{-\varepsilon}, e^{\varepsilon}] \cdot Z$ only if at least half of the $\widehat{Z}^r$'s fail to be contained in $[e^{-\varepsilon}, e^{\varepsilon}] \cdot Z$. By the Chernoff-Hoeffding bound, this happens with probability at most δ when $R \geq 32 \log\left(\frac{1}{\delta}\right)$. ■

F.2. Rejection sampling

The following bounds the TV-error and expected running time for rejection sampling, given an inexact oracle for the proposal distribution.

Lemma F.2 *Suppose that P and Q are probability measures on Ω such that $\frac{dP}{dQ} \leq C$ everywhere. Suppose we have an oracle which gives samples from $\widetilde{Q}$, with $d_{\mathrm{TV}}(\widetilde{Q}, Q) \leq \frac{\varepsilon}{2C}$. Consider the following rejection sampling algorithm: draw $x \sim \widetilde{Q}$, and accept with probability $\frac{1}{C}\frac{dP}{dQ}(x)$; otherwise repeat the process. Let $\widetilde{P}$ be the resulting measure. Then $d_{\mathrm{TV}}(\widetilde{P}, P) \leq \varepsilon$, and the number of oracle calls is a geometric random variable with success probability at least $\frac{1}{2C}$ (and hence expected value at most $2C$).*

Proof Let $A \subseteq \Omega$ be measurable. First, we note that $\widetilde{P}(A) = \frac{\int_A \frac{dP}{dQ} d\widetilde{Q}}{\int_\Omega \frac{dP}{dQ} d\widetilde{Q}}$. To calculate $d_{\mathrm{TV}}(\widetilde{P}, P)$, we break up the difference as

$$\begin{aligned}\widetilde{P}(A) - P(A) &= \left(\frac{\int_A \frac{dP}{dQ}\, d\widetilde{Q}}{\int_\Omega \frac{dP}{dQ}\, d\widetilde{Q}} - \int_A \frac{dP}{dQ}\, d\widetilde{Q}\right) + \left(\int_A \frac{dP}{dQ}\, d\widetilde{Q} - \int_A \frac{dP}{dQ}\, dQ\right) \\ &\leq \left(\frac{\int_A \frac{dP}{dQ}\, d\widetilde{Q}}{\int_\Omega \frac{dP}{dQ}\, d\widetilde{Q}} \left(1 - \int_\Omega \frac{dP}{dQ}\, d\widetilde{Q}\right)\right) + \left(\int_A \frac{dP}{dQ}\, d\widetilde{Q} - \int_A \frac{dP}{dQ}\, dQ\right)\end{aligned}$$

Next note that

$$\left|\int_\Omega \frac{dP}{dQ}\, d\widetilde{Q} - 1\right| \leq \left|\int_\Omega \frac{dP}{dQ}\, d\widetilde{Q} - \int_\Omega \frac{dP}{dQ}\, dQ\right| \leq C d_{\mathrm{TV}}(Q, \widetilde{Q}) \leq \frac{\varepsilon}{2}.$$

Hence,

$$|\widetilde{P}(A) - P(A)| \leq \left|\int_\Omega \frac{dP}{dQ}\, d\widetilde{Q} - 1\right| + d_{\mathrm{TV}}(\widetilde{Q}, Q) \left\|\frac{dP}{dQ}\right\|_\infty \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2C} C = \varepsilon,$$

so $d_{\mathrm{TV}}(\widetilde{P}, P) \leq \varepsilon$. Finally, we check that the acceptance probability is

$$\int_\Omega \frac{1}{C} \frac{dP}{dQ} d\widetilde{Q} \geq \int_\Omega \frac{1}{C} \frac{dP}{dQ} dQ - \frac{1}{C} \cdot C d_{\mathrm{TV}}(Q, \widetilde{Q}) \geq \frac{1}{C} - \frac{\varepsilon}{2C} \geq \frac{1}{2C}.$$

■

F.3. Spectral gap of a projected chain

We use the following to bound the Poincaré constant of the projected Markov chain arising in the analysis of simulated tempering. A similar analysis appears in the proof in Ge et al. (2018).

Lemma F.3 *Let S be a countable set. Consider a reversible Markov chain on $[L] \times S$ with stationary distribution P and transition kernel T satisfying the following conditions. Let $P_\ell(j) = P((\ell, j))/P(\{\ell\} \times S)$.*

1. *(Bounded bottleneck ratio) For $k < \ell$, $\frac{P_k(j)}{P_\ell(j)} \geq \gamma$.*
2. *(Transitions at highest temperature and between adjacent temperatures) We have*

$$T((\ell_1, i_1), (\ell_2, i_2)) \geq \begin{cases} \frac{P_1(i_2)}{D_{\mathrm{high}}}, & \ell_1 = \ell_2 = 1, \quad i_1 \neq i_2 \\ \frac{1}{2D_{\mathrm{adj}}} \min\left\{\frac{P((\ell \pm 1, i_1))}{P((\ell, i_1))}, 1\right\}, & i_1 = i_2, \quad \ell_1 \neq L, \ell_2 = \ell_1 \pm 1 \end{cases}$$

3. *(Lower bound of probability for each level) For each ℓ, $P(\{\ell\} \times S) \geq \frac{r}{L}$.*

Then the following hold.

1. *(Cheeger constant) The Cheeger constant satisfies $\Phi \geq \frac{\gamma r}{2L \max\{D_{\mathrm{high}}, D_{\mathrm{adj}}\}}$.*

2. *(Poincaré constant) The associated Dirichlet form satisfies a Poincaré inequality with constant* $C_{\mathrm{P}} \leq \frac{8L^2 \max\{D_{\text{high}}, D_{\text{adj}}\}^2}{\gamma^2 r^2}$.

Proof Let $Q(x, B)$ denote $P(x)T(x, B)$ and $Q(A, B)$ denote $\sum_{x \in A} P(x)T(x, B)$. Note $Q(A, B) = Q(B, A)$ by reversibility. Let A_ℓ denote the sets such that $A = \bigcup_{\ell=1}^{L} \{\ell\} \times A_\ell$, i.e., A_ℓ is the ℓth layer of A.

To prove the bound on the Cheeger constant, for each A, it suffices to bound either $\frac{Q(A,A^c)}{P(A)}$ or $\frac{Q(A^c,A)}{P(A^c)}$. Without loss of generality, we suppose that $P_1(A_1) \leq \frac{1}{2}$. For each j, let ℓ_j denote the smallest ℓ such that $(\ell, j) \in A$. To lower bound $Q(A, A^c)$, we consider the contributions from n such that $\ell_j > 1$ and $\ell_j = 1$ separately.

1. $\ell_j > 1$: We have

$$\begin{aligned} Q((\ell_j, j), A^c) &\geq P((\ell_j, j))T((\ell_j, j), (\ell_j - 1, j)) \\ &\geq P((\ell_j, j)) \frac{1}{2D_{\text{adj}}} \min\left\{ \frac{P((\ell_j - 1, j))}{P((\ell_j, j))}, 1 \right\} \\ &= \frac{1}{2D_{\text{adj}}} \min\{P((\ell_j - 1, j)), P((\ell_j, j))\} \\ &\geq \frac{\gamma r}{2LD_{\text{adj}}} P([\ell_j, L] \times \{j\}). \end{aligned}$$

2. $\ell_j = 1$: Note $A_1 = \{j : \ell_j = 1\}$. We will bound $Q(\{1\} \times A_1, A^c)$ by looking at transitions within $\{1\} \times S$. We have

$$\begin{aligned} Q(\{1\} \times A_1, A^c) &\geq \sum_{j \in A_1} P((1, j))T((1, j), \{1\} \times A_1^c) \\ &\geq \sum_{j \in A_1} P((1, j)) \frac{P(\{1\} \times A_1^c)}{D_{\text{high}}} \\ &\geq \frac{1}{2D_{\text{high}}} \sum_{j \in A_1} P((1, j)) = \frac{1}{2D_{\text{high}}} P(\{1\} \times A_1) \\ &\geq \frac{\gamma r}{2LD_{\text{high}}} P([L] \times A_1). \end{aligned}$$

Adding the two parts,

$$\begin{aligned} Q(A, A^c) &\geq \frac{\gamma r}{2L \max\{D_{\text{adj}}, D_{\text{high}}\}} P\left(\left(\bigcup_{j:\ell_j > 1} [\ell_j, L] \times \{j\} \right) \cup ([L] \times A_1) \right) \\ &\geq \frac{\gamma r}{2L \max\{D_{\text{adj}}, D_{\text{high}}\}} P(A). \end{aligned}$$

The bound on the Poincaré constant follows immediately from Cheeger's inequality: the spectral gap of the chain is at least $\frac{1}{2}\Phi^2$, and the Poincaré constant is the inverse of the spectral gap. ■

Appendix G. Additional material related to examples

We give here the derivation of the posterior for the contextual SBM. Because we chose consistent notations between problems, the derivation of the posterior for the Gaussian mixture model is simply the special case of this argument where $\lambda = 0$ (so there is no graph/spiked Wigner information).

Posterior derivation in contextual SBM. Under the Gaussian contextual stochastic block model, we have

$$\begin{aligned} p(A, B \mid u, v) &\propto \exp\left(-\frac{n}{4}\left\|\frac{\lambda}{n}vv^\top - A\right\|_F^2 - \frac{p}{2}\left\|\sqrt{\frac{\mu}{n}}vu^\top - B\right\|_F^2\right) \\ &\propto \exp\left(\frac{\lambda}{2}\langle vv^\top, A\rangle - \frac{n}{4}\|A\|_F^2 + p\sqrt{\mu/n}\langle vu^\top, B\rangle - \frac{p}{2}\|B\|_F^2 - \frac{p}{2}\mu\|u\|^2\right) \end{aligned}$$

(note we dropped the term $\|vv^\top\|_F^2$ since it is a constant) and so

$$\begin{aligned} p(u, v \mid A, B) &= p(A, B \mid u, v)p(u, v)/p(A, B) \\ &\propto \exp\left(\frac{\lambda}{2}\langle vv^\top, A\rangle + p\sqrt{\mu/n}\langle B^\top v, u\rangle - \frac{p}{2}(1+\mu)\|u\|^2\right). \end{aligned}$$

Integrating over u, we have that the posterior distribution is

$$\begin{aligned} p(v \mid A, B) &\propto \int \exp\left(\frac{\lambda}{2}\langle vv^\top, A\rangle + p\sqrt{\mu/n}\langle B^\top v, u\rangle - \frac{p}{2}(1+\mu)\|u\|^2\right) du \\ &= \int \exp\left(\frac{\lambda}{2}\langle vv^\top, A\rangle - \frac{p}{2}(1+\mu)\left\|u - \frac{1}{1+\mu}\sqrt{\frac{\mu}{n}}B^\top v\right\|^2 + \frac{p\mu}{2n(1+\mu)}\|B^\top v\|^2\right) du \\ &\propto \exp\left(\frac{\lambda}{2}\langle vv^\top, A\rangle + \frac{p\mu}{2n(1+\mu)}\|B^\top v\|_2^2\right) \\ &\propto \exp\left(\frac{\lambda}{2}\langle vv^\top, A\rangle + \frac{p\mu}{2n(1+\mu)}\langle vv^\top, BB^\top\rangle\right) \end{aligned}$$

This is an Ising model without external field.

Appendix H. Computational hardness of sampling from rank-one models with large spike

Using the subset sum/number partitioning problem, we will show that sampling and (even crudely) approximating $\log Z$ from negative-definite rank-one models is NP-hard. The NP-hard problem we start with is given integers $a_1, \ldots, a_n$, determining whether there exists a partitioning into two sets such that the sum is equal. Equivalently, we seek to determine if there exists a sign vector $\sigma \in \{\pm 1\}^n$ such that

$$\sum_i a_i \sigma_i = 0.$$

This is not the first time this problem is connected to statistical physics—see e.g., discussion in Borgs et al. (2001); Gamarnik and Kızıldağ (2021).

Theorem H.1 *Let $\beta \geq 1$ be arbitrary and fixed. For any $a = (a_1, \ldots, a_n) \in \mathbb{Z}^n$, define the Ising model with probability mass function $P_a : \{\pm 1\}^n \to [0,1]$ given by*

$$P_a(\sigma) = \frac{1}{Z} \exp\left(-\beta n \langle a, \sigma \rangle^2\right)$$

If there exists a polynomial time randomized algorithm to approximately sample within TV distance $1/2$ from Ising models of this form for any $a_1, \ldots, a_n$, then $\mathsf{NP} = \mathsf{RP}$. Furthermore, for $\beta \geq 2\log(2)$, it is NP-hard to approximate the log partition function/free energy $\log Z$ of such a model within an additive error of $\frac{\beta n}{2}$, and under the Exponential Time Hypothesis (ETH), it is impossible to do so in subexponential time in the presence of an external field $b \in \mathbb{Z}^n$, i.e., for models of the form

$$P_{a,h}(\sigma) = \frac{1}{Z} \exp\left(-\beta n \langle a, \sigma \rangle^2 + \langle b, \sigma \rangle\right).$$

Proof Let $a_1, \ldots, a_n$ be an instance of the number partitioning problem. Consider the Ising model with probability mass function $P_a : \{\pm 1\}^n \to [0,1]$ given by for $\beta \geq 1$

$$P_a(\sigma) = \frac{1}{Z} \exp\left(-\beta n \langle a, \sigma \rangle^2\right),$$

where Z is the normalizing constant (partition function) so that the distribution has normalizing constant 1. Note that this is an Ising model with interaction matrix $-2\beta n a a^T$, which is negative definite and rank one as promised. If there exists at least one solution $\sum_i a_i \sigma_i = 0$ then

$$\Pr_{\sigma \sim P}\left(\sum_i a_i \sigma_i \neq 0\right) = \frac{\sum_{\sigma : \sum_i a_i \sigma_i \neq 0} e^{-\beta n \langle a, \sigma \rangle^2}}{\sum_{\sigma \in \{\pm 1\}^n} e^{-\beta n \langle a, \sigma \rangle^2}} \leq 2^n e^{-\beta n}$$

where we used that because the a_i are integers, if $\sum_i a_i \sigma_i \neq 0$ then $\langle a, \sigma \rangle^2 \geq 1$, and also that if there exists a solution $\sum_i a_i \sigma_i = 0$ then the denominator is at least 1. Thus, except with exponentially small probability in n, a sample from P will be a solution to the subset sum problem. In particular, it follows that a polynomial time (approximate) sampling algorithm implies $\mathsf{NP} = \mathsf{RP}$.

Similarly, observe that if there exists a solution to the subset sum instance then $\log Z \geq 0$ whereas if there does not exist a solution, then $\log Z \leq n[\log(2) - \beta] < -\frac{\beta n}{2}$, which establishes the NP-hardness of approximating $\log Z$. The last statement in the Theorem follows because solving subset sum in time $2^{o(n)}$ is known to be ETH-hard (see discussion in Abboud et al. (2022)), and the general subset problem (deciding if there exists σ so that $\sum_i a_i \sigma_i = b$) can be directly encoded as minimizing

$$\left(\langle a, \sigma \rangle - b\right)^2 = \langle a, \sigma \rangle^2 - 2b\langle a, \sigma \rangle + b^2,$$

which by the same argument as above implies that approximating $\log Z$ for the distribution $P_{a,h}$ with $h = 2b\beta n a$ is ETH-hard. ■