Parallelising Glauber Dynamics

Holden Lee
Department of Applied Mathematics and Statistics, The Johns Hopkins University, Baltimore, MD, USA

Abstract

For distributions over discrete product spaces $\prod_{i=1}^n \Omega_i'$, Glauber dynamics is a Markov chain that at each step, resamples a random coordinate conditioned on the other coordinates. We show that k-Glauber dynamics, which resamples a random subset of k coordinates, mixes k times faster in χ^2-divergence, and assuming approximate tensorization of entropy, mixes k times faster in KL-divergence. We apply this to obtain parallel algorithms in two settings: (1) For the Ising model $\mu_{J,h}(x) \propto \exp(\frac{1}{2}\langle x, Jx\rangle + \langle h, x\rangle)$ with $\|J\| < 1 - c$ (the regime where fast mixing is known), we show that we can implement each step of $\widetilde{\Theta}(n/\|J\|_F)$-Glauber dynamics efficiently with a parallel algorithm, resulting in a parallel algorithm with running time $\widetilde{O}(\|J\|_F) = \widetilde{O}(\sqrt{n})$. (2) For the mixed p-spin model at high enough temperature, we show that with high probability we can implement each step of $\widetilde{\Theta}(\sqrt{n})$-Glauber dynamics efficiently and obtain running time $\widetilde{O}(\sqrt{n})$.

2012 ACM Subject Classification Theory of computation → Random walks and Markov chains

Keywords and phrases sampling, Ising model, parallel algorithm, Markov chain, Glauber dynamics

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2024.49

Category RANDOM

Related Version *Full Version*: https://arxiv.org/abs/2307.07131

Acknowledgements I want to thank Leo Du, Frederic Koehler, and Nicolas Loizou for helpful discussions.

1 Introduction

A key problem in computer science and statistics is to sample from a probability distribution given its probability mass function up to a constant of proportionality. The problem has been studied both over discrete spaces (such as Ω^n for a finite set Ω) and continuous spaces (such as $\mathbb{R}^n$); the goal is to give efficient algorithms for general classes of distributions, and in particular, to obtain optimal scaling in the dimension n. In this work we focus on minimizing the parallel running time, assuming a polynomial number of processors. In $\mathbb{R}^n$, it is natural to change multiple coordinates at a time using gradient-based algorithms such as Langevin dynamics and Hamiltonian Monte Carlo; many results have given algorithms that require a sublinear number of steps for log-concave distributions in various settings.

However, on discrete product spaces Ω^n, the canonical algorithm, Glauber dynamics, involves resampling coordinates one at a time, and hence requires at least n steps in general. A natural attempt to speed up Glauber dynamics with parallel computation is to resample k coordinates at a time. We establish that under general conditions, this simple idea does indeed speed up Glauber dynamics by a factor of approximately k.

To obtain a parallel algorithm, the task remains to give a fast parallel method of resampling k coordinates. We show that this can be done in the case of the Ising model $\mu_{J,h}$ over $\{\pm 1\}^n$ when the interaction matrix J is bounded away from 1 in operator norm, $\|J\| < 1 - c$, and in the case of the mixed p-spin model at high enough temperature, both of which are known to enjoy rapid mixing of standard Glauber dynamics.

The Ising model is a classical model from statistical physics which has probability mass function on $\{\pm 1\}^n$ given by

$$\mu_{J,h}(x) = \frac{1}{Z_{J,h}} \exp\left(\frac{1}{2}\langle x, Jx\rangle + \langle h, x\rangle\right), \quad \text{where } Z_{J,h} = \sum_{x\in\{\pm1\}^n} \exp\left(\frac{1}{2}\langle x, Jx\rangle + \langle h, x\rangle\right).$$

The regime $\|J\| < 1$ is exactly where (based on information of the operator norm alone) Glauber dynamics is known to have fast mixing [24, 8]. To sample k coordinates, we use approximate rejection sampling with a product distribution and a further recursion for certain "bad" sets. By taking $k = \widetilde{\Theta}(n/\|J\|_F)$, we obtain an algorithm with parallel running time $\widetilde{O}(\|J\|_F) = \widetilde{O}(\sqrt{n})$.

The **mixed p-spin model** with coefficients $\beta_2, \beta_3, \ldots$ and external field $h \in \mathbb{R}^n$ is the random measure on $\{\pm1\}^n$ given by[1]

$$\begin{aligned} &\mu_{\beta,g,h}(x) \propto \exp(H_{\beta,g,h}(x)), \\ &\text{where } H_{\beta,g,h}(x) = \sum_{p=2}^{\infty} \frac{\beta_p\sqrt{p!}}{n^{\frac{p-1}{2}}} \sum_{1\le i_1<\cdots<i_p\le n} g_{i_1,\cdots,i_p} x_{i_1}\cdots x_{i_p} + \sum_{i=1}^{n} h_i x_i \end{aligned} \tag{1}$$

and $g_{i_1,\ldots,i_p} \sim N(0,1)$. By taking $k = \Theta(\sqrt{n})$, we obtain an algorithm with parallel running time $\widetilde{O}(\sqrt{n})$.

1.1 Main results

Let μ be a distribution on $\prod_{i=1}^n \Omega_i'$. We define k-Glauber dynamics as the Markov chain which given a sample $x \in \prod_{i=1}^n \Omega_i'$, chooses a subset $S \subseteq [n]$ uniformly at random among subsets of size k, and resamples the coordinates in S conditional on coordinates in S^c, according to the distribution of μ. Let $P_{\mu,k}$ denote its Markov kernel.

We show that under general conditions, k-Glauber dynamics mixes k times faster in both χ^2 and KL-divergence. We say that a Markov kernel P with stationary distribution μ satisfies ρ-contraction in χ^2-divergence if $\mathcal{D}_{\chi^2}(\nu P\|\mu) \le \rho\mathcal{D}_{\chi^2}(\nu\|\mu)$ and similarly for $\mathcal{D}_{\mathrm{KL}}$; this can be iterated to give a mixing time bound. See Section 2.2 for background on functional inequalities (Poincaré inequality and approximate tensorization of entropy).

▶ **Theorem 1.1** (k-Glauber mixes k times faster). *Let μ be a distribution on $\Omega = \prod_{i=1}^n \Omega_i'$, and let $1 \le k \le n$. Below, let $C \ge 1$.*

1. *If μ satisfies a Poincaré inequality with constant Cn, then $P_{\mu,k}$ satisfies a Poincaré inequality with constant $O\left(\frac{Cn}{k}\right)$, and satisfies $(1-\Omega\left(\frac{k}{Cn}\right))$-contraction in χ^2-divergence.*
2. *If μ satisfies C-approximate tensorization of entropy (so that P_μ satisfies $(1-\Omega\left(\frac{1}{Cn}\right))$-contraction in KL-divergence), then $P_{\mu,k}$ satisfies $(1-\Omega\left(\frac{k}{Cn}\right))$-contraction in KL-divergence.*

Here, the $O(\cdot)$ and $\Omega(\cdot)$ hide only universal constants. The Poincaré inequality is equivalent to contraction in χ^2-divergence, so part (1) gives a $\Omega(k)$-factor speedup to mixing in χ^2. The analogue of the Poincaré inequality for KL is a modified log-Sobolev inequality. Although we need the slightly stronger notion of approximate tensorization of entropy to prove a speedup to mixing in KL, we note that many works that establish a modified log-Sobolev inequality do so using tensorization of entropy [12, 8]. See Section 2 for relevant background on mixing for Markov chains.

[1] The factor $\sqrt{p!}$ arises as we index only over increasing sequences.

We prove Theorem 1.1 as Corollary 3.6 of the more general Theorem 3.5. We view k-Glauber dynamics as randomly erasing k coordinates one by one, and then adding them back one by one according to the right conditional distributions. This realizes k-Glauber dynamics as a composition of down and up operators $D_{n\to n-1}\cdots D_{n-k+1\to n-k}U_{n-k\to n-k+1}\cdots U_{n-1\to n}$. The assumptions give contraction of $D_{n\to n-1}$, and our general theorem shows that the contraction of each $D_{j\to j-1}$ is at least as good as $D_{n\to n-1}$ (except for an additive factor). To do this, we realize $D_{j\to j-1}$ as $D_{n\to n-1}$ tensorized with erasure "noise" and projected, and bound how the factor of contraction changes under these operations. We make an analogy to bounding the Poincaré and log-Sobolev constants of a distribution μ on $\mathbb{R}^n$ convolved with Gaussian noise, and the proximal sampler based on iteratively adding and removing Gaussian noise (more specifically, sampling from the posterior distribution given a noisy Gaussian observation of the sample from μ).

Algorithmically, the challenge with implementing k-Glauber dynamics is that naive enumeration for the transition kernel takes 2^k time, and hence we must find a way to use the structure of the distribution to implement each step more efficiently. We show that in the case of the Ising model, we can efficiently simulate k-Glauber dynamics for $k = \widetilde{O}\left(n/\left\|J^{\diamond}\right\|_F\right)$, to obtain a parallel algorithm running in time $\widetilde{O}\left(\left\|J^{\diamond}\right\|_F\right)$, where $J^{\diamond}$ denotes J with diagonal entries set to 0[2]. Under the assumption that $\|J\| < 1$, this is always at most $\widetilde{O}(\sqrt{n})$.

▶ **Theorem 1.2.** *Let $c > 0$. With appropriate choice of constants depending only on c, if J is symmetric positive semi-definite with $\|J\| \le 1 - c$, then* ParallelIsingSampler *(Algorithm 1) with appropriate constants outputs a sample ε-close in TV distance from the Ising model $\mu_{J,h}$ and, with probability at least $1-\varepsilon$, runs in time $O\left(\max\left\{\left\|J^{\diamond}\right\|_F, 1\right\}\operatorname{poly}\log\left(\frac{n}{\varepsilon}\right)\right)$ on a parallel machine with $\operatorname{poly}(n)$ processors.*

We note that our algorithm is a high-accuracy sampler: the only dependence on ε is a poly-logarithmic dependence in the running time. Notably, the number of processors does not depend on ε. We rely on the result [8] that gives optimal ($O(n\ln n)$) mixing times for the Ising model for $\|J\| < 1$ based on the theory of entropic independence.

The first attempt to implement k-Glauber dynamics is to approximate the conditional distribution of k coordinates using a carefully chosen product distribution and use rejection sampling. Using concentration results (the Hanson-Wright inequality), if $\left\|J^{\diamond}_{S\times S}\right\|_F$ is small for the randomly chosen set S, then this succeeds with high probability. The complication is that $\left\|J^{\diamond}_{S\times S}\right\|_F$ can sometimes be large. If this is the case, then we recurse on $J_{S\times S}$. By controlling the expected size of $\left\|J^{\diamond}_{S\times S}\right\|_F$, we show that the recursive calls form a subcritical branching process and with high probability, add at most a polylogarithmic overhead to the running time.

▶ **Theorem 1.3.** *Consider the mixed p-spin model (1). There exists an absolute constant $\delta > 0$ such that if $\sum_{p\ge 2}\sqrt{p^3\ln p}\cdot\beta_p < \delta$ and $D(\beta) = \sum_{p\ge 2}\sqrt{2^p p^3\ln p}\cdot\beta_p < \infty$, then with probability $1-\exp(-\Omega(n))$ over g, given query access to $H_{\beta,g,h}$, there is an algorithm which outputs a sample ε-close in TV distance from $\mu_{\beta,g,h}$ and, with probability at least $1-\varepsilon$, runs in time $O_{D(\beta)}\left(\sqrt{n}\operatorname{poly}\log\left(\frac{n}{\varepsilon}\right)\right)$ on a parallel machine with $\operatorname{poly}\left(\frac{n}{\varepsilon}\right)$ processors.*

[2] While changing the diagonal entries of J does not change the Ising model, we need to allow J to have nonzero diagonal entries in order to be positive semi-definite.

Algorithm 1 Parallel Ising Sampler (ParallelIsingSampler).

Input: Interaction matrix $J \in \mathbb{R}^{n\times n}$, subset R of size m, external field $h \in \mathbb{R}^R$, error parameter $\varepsilon \in (0, \frac{1}{2})$.
Let $\varepsilon_{\text{step}} = \frac{\varepsilon}{2n^{C_4}}$.
if $\left\|J^{\otimes}_{R\times R}\right\|_F \le \frac{c_3}{\ln\left(\frac{2}{\varepsilon_{\text{step}}}\right)+1}$ ($J^{\otimes}$ denotes J with diagonal entries set to 0) **then**
    $y \leftarrow$ QuadraticApproxRejectionSampler $\left(H(x)=\frac{1}{2}\langle x, J_{R\times R}x\rangle+\langle h,x\rangle\,,\ \frac{c_3}{\ln\left(\frac{2}{\varepsilon_{\text{step}}}\right)+1},\varepsilon_{\text{step}}\right)$. (See Algorithm 2.)
else
    Let $s = \left\lceil \frac{c_1 m}{\left(\ln\left(\frac{2}{\varepsilon_{\text{step}}}\right)+1\right)\ln\left(\frac{n}{\varepsilon}\right)\left\|J^{\otimes}_{R\times R}\right\|_F} \right\rceil$.
    Let $T = \left\lfloor C_2 \ln\left(\frac{n}{\varepsilon}\right)\frac{m}{s}\right\rfloor$.
    Draw y from the product distribution $\nu_0(x) \propto e^{\langle h,x\rangle}$.
    for t from 1 to T **do**
        Choose $S \subseteq R$ a random subset of size s.
        $z \leftarrow$ ParallelIsingSampler$(J, S, J_{S\times R\setminus S}y_{R\setminus S} + h_S, \varepsilon)$
        Set $y_S = z$.
    end for
end if
Output: y (Approximate sample from $\mu_{J_{R\times R},h}$).

Note that a recursion is not necessary in this algorithm. Intuitively, the mean-field nature of the p-spin model ensures that with high probability all marginal distributions of $O(\sqrt{n})$ coordinates are well-approximated by a product distribution. Though we do not investigate this further, a recursive algorithm could potentially eliminate the poly$(1/\varepsilon)$ dependence on the number of processors as in Theorem 1.2. The proof of Theorem 1.3 is in Section 5 in the full version.

We view our result on the Ising model and the p-spin model as proofs of concept for parallelisation using k-Glauber dynamics, and hope it serves as a useful framework for constructing parallel algorithms for other families of discrete distributions. As discussed in the next section, using a different parallel algorithm, the work [35] obtains Theorem 1.2 but not Theorem 1.3.

1.2 Related work

We note that our Theorem 1.1 can be viewed as a complement of "local-to-global" results for mixing of the down-up walks [41, 1, 19], and is not implied by those results. Those results aim to establish mixing of Glauber dynamics (or the down-up walk) from mixing of simpler chains, while we start by assuming mixing of Glauber dynamics. In particular, [19] apply the reverse strategy: for the spin systems on graphs they consider, they show that mixing of θn-Glauber dynamics (for appropriate θ) implies mixing of Glauber dynamics.

When contraction of Glauber dynamics is derived directly from either spectral or entropic independence using local-to-global arguments, then the same arguments can be used to establish mixing of the k-Glauber (e.g., using k-uniform block factorization of entropy [19], the analogue of approximate tensorization of entropy). However, this does not apply for distributions for which mixing is established through other methods. The recent work [9]

shows that a Poincaré inequality implies spectral independence, but the bound obtained for k-Glauber through spectral independence is lossy (resulting in a power of n). Our work can be seen as giving a general conceptual reason why mixing for Glauber must imply mixing for k-Glauber.

1.2.1 Continuous sampling

For log-concave distributions on $\mathbb{R}^n$, a long line of works on the underdamped Langevin algorithm and Metropolis-adjusted Langevin algorithm have led to high-accuracy sampling using $\widetilde{O}(n^{1/2})$ steps [3]. The randomized midpoint method for underdamped Langevin dynamics allows sampling in the weaker Wasserstein metric in $\widetilde{O}(n^{1/3})$ steps [45], and can furthermore be fully parallelised to obtain ε error with $\operatorname{poly}\left(\frac{n}{\varepsilon}\right)$ processors. These dependencies are assuming the condition number is $O(1)$.

We note that the Ising model for $\|J\| < 1$ can be decomposed as a log-concave mixture of product distributions [30, 10, 32], so these algorithms give an alternative approach to parallel algorithms for the Ising model. However, this decomposition is highly specific to the Ising model. Moreover, the Wasserstein guarantee is incompatible with a TV guarantee, and the complexity of our approach scales with $\|J\|_F$.

1.2.2 Parallel algorithms for discrete sampling

Recent work [6, 4, 7] has investigated the question of obtaining fast parallel algorithms for approximate sampling in settings where fast parallel algorithms for approximate *counting* (or computing a partition function) exist. In particular, for distributions satisfying transport stability and where the *log-Laplace* transform can be efficiently calculated (e.g., using the efficient algorithm for computing partition functions), [7] gives a $\operatorname{poly}\log(n/\varepsilon)$-time algorithm with $\operatorname{poly}(n/\varepsilon)$ many processors (i.e., a RNC algorithm). This includes problems such as determinantal point processes and Eulerian tours. Notably they use the continuous algorithm (randomized midpoint method, discussed above) even though the problem is discrete.

In the setting of Ising models, however, we do not have a fast parallel algorithm for counting. Several works [26, 35] have studied the problem assuming the associated Dobrushin influence matrix has bounded norm. By using simultaneous updates, [35] obtains a factor-$\frac{n}{C}$ speedup for distributions whose Dobrushin influence matrix has norm bounded by C, in particular giving RNC algorithms when $C = O(1)$ and the mixing time is $O(n \ln n)$. The result of [35] can also give Theorem 1.2 with a different algorithm, but cannot be used to derive Theorem 1.3. See Appendix A in the full version for details.

On the practical side, designers of Markov chain Monte Carlo algorithms in discrete spaces have taken inspiration from continuous algorithms, for example, by using gradient information to inform the proposal distribution and allow updating multiple coordinates at once [29, 50, 42]. Theoretical guarantees for these algorithms remain to be understood.

1.2.3 Diffusion models and the proximal sampler

Stochastic localization [23] is a measure-valued stochastic process that converges to a point mass, which is distributed according to a desired distribution μ. As a technique, it gives a way of decomposing probability distributions that has been useful in proving functional inequalities and mixing time [17, 18], and more recently, in constructing new, time-*inhomogeneous* algorithms for sampling [22, 39].

Diffusion models [46, 47, 48] are a successful paradigm for generative modeling in machine learning, where the task is to learn and then generate samples from a distribution where only samples are given. Though the details may differ, they consist of a forward process which

adds noise to the data; reversing the process can then generate a sample from random noise. It has been observed [38] that a stochastic localization process can be viewed as the reverse process of a diffusion model.

Our analysis of k-Glauber dynamics is inspired by the analysis of the proximal sampler [34, 16, 25], which does alternating Gibbs sampling by adding Gaussian noise to the current sample, and then "de-noising" by sampling from the posterior distribution; this fits in the framework discussed above. In their analysis, [16] show that proximal sampler mixes at least as fast as Langevin in terms of χ^2 and KL-divergence. [25] show a $\widetilde{O}(n^{1/2})$ dimension dependence using a carefully chosen Gaussian proposal distribution to implement the posterior sampling step. We view the k-Glauber dynamics as a discrete analogue of the proximal Langevin algorithm, where the noise consists of erasing k coordinates, and our proof follows this analogy. In our application, we also require a careful choice of product distribution for the proposal.

2 Preliminaries

While many of the notions are generalizable, we will restrict ourselves to finite state spaces, and identify all measures with their probability mass functions. For more background on Markov chains, see [40].

2.1 Markov kernels

For finite sets A and B, a Markov kernel K from A to B is a function $A \times B \to \mathbb{R}_{\geq 0}$ or equivalently, a matrix $\mathbb{R}_{\geq 0}^{A \times B}$, where the rows sum to 1. If μ is a measure on A, then μK is a measure on B; if f is a function $B \to \mathbb{R}$, then Kf is a function $A \to \mathbb{R}$; these correspond to matrix-vector multiplication. Composition of kernels K_1 from A to B and K_2 from B to C gives a kernel $K_1 K_2$ from A to C, which corresponds to matrix multiplication. For f, g functions on A and μ a measure on A, let $\langle f, g \rangle_\mu = \sum_{x \in A} \mu(x) f(x) g(x)$. For a kernel $K : A \times B \to \mathbb{R}_{\geq 0}$, given measures μ_1, μ_2 on A and B respectively, we think of K as a linear map $L^2(\mu_1) \to L^2(\mu_2)$; then its adjoint $K^* : B \times A \to \mathbb{R}_{\geq 0}$ is a linear map $L^2(\mu_2) \to L^2(\mu_1)$ satisfying $\langle f, Kg \rangle_{\mu_1} = \langle K^* f, g \rangle_{\mu_2}$ for any $f \in L^2(\mu_1)$, $g \in L^2(\mu_2)$.

▶ **Definition 2.1.** *k-**Glauber dynamics** with stationary distribution μ on Ω is the Markov chain where at each step, if the current sample is x, we choose a subset S uniformly at random in $\binom{\Omega}{k}$ (subsets of size k), and resample the coordinates in S according to $\mu(X_S | X_{S^c} = x_{S^c})$. Let $P_{\mu,k}$ denote the transition operator. For $k = 1$, we simply call it Glauber dynamics, and let P_μ denote the Markov kernel.*

▶ **Definition 2.2.** *Let $0 \leq \ell \leq k \leq n$. Let μ be a distribution on $\binom{[n]}{k}$. Define the **down operator** $D_{k \to \ell}$ and **up operator** $U_{\ell \to k}$ as Markov kernels $\binom{[n]}{k} \times \binom{[n]}{\ell} \to \mathbb{R}_{\geq 0}$ and $\binom{[n]}{\ell} \times \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$, respectively, with*

$$D_{k \to \ell}(A, B) = \mathbb{1}_{B \subseteq A} \frac{1}{\binom{k}{\ell}} \qquad U_{\ell \to k}(B, A) = \mathbb{1}_{B \subseteq A} \frac{\mu(A)}{\sum_{A' \supseteq B} \mu(A')}.$$

*Let $\mu_\ell = \mu D_{k \to \ell}$ for $0 \leq \ell \leq k$, and define the $k \leftrightarrow \ell$ **down-up walk** and $\ell \leftrightarrow k$ **up-down walk** by*

$$P^{\triangledown}_{k \leftrightarrow \ell} = D_{k \to \ell} U_{\ell \to k} \qquad P^{\triangle}_{\ell \leftrightarrow k} = U_{\ell \to k} D_{k \to \ell}.$$

Note that $D_{k\to\ell}$ does not depend on μ while $U_{\ell\to k}$ does; we suppress the dependency in the notation. Note that $D_{k\to\ell}D_{\ell\to m} = D_{k\to m}$ and $U_{m\to\ell}U_{\ell\to k} = U_{m\to k}$. As operators, $D_{k\to\ell} : L^2(\mu_\ell) \to L^2(\mu_k)$ and $U_{\ell\to k} : L^2(\mu_k) \to L^2(\mu_\ell)$ are adjoint.

▶ **Definition 2.3.** *Let μ be a measure on $\Omega' = \Omega'_1 \times \cdots \times \Omega'_n$. Define the* ***homogenization*** *of μ to be the measure μ^{hom} over $\binom{\Omega}{n}$, where $\Omega = \bigcup_{i=1}^n \Omega'_i \times \{i\}$ and $\sigma \in \Omega'$ is identified with $\{(\sigma_1, 1), \ldots, (\sigma_n, n)\}$. (For short, we will write $\Omega = \bigsqcup_{i=1}^n \Omega'_i$ in the following.) For any property $\mathcal{P}$ defined for measures $\binom{\Omega}{n}$, we say that μ satisfies $\mathcal{P}$ if μ^{hom} satisfies $\mathcal{P}$.*

Under this identification, k-Glauber dynamics corresponds to the $n \leftrightarrow n-k$ down-up walk, as the down step corresponds to erasing k coordinates and the up step corresponds to restoring them with the correct conditional probabilities.

2.2 Functional inequalities

▶ **Definition 2.4.** *Let $M = (\Omega, P)$ be an ergodic, reversible Markov chain with stationary distribution μ. Define the associated Dirichlet form as the inner product*

$$\mathcal{E}_P(f, g) = \langle f, (I - P)g\rangle_\mu = \frac{1}{2}\sum_{x,y\in\Omega} \mu(x)P(x, y)(f(x) - f(y))(g(x) - g(y))$$

When μ is a distribution on $\Omega = \prod_{i=1}^n \Omega'_i$, we write $\mathcal{E}_\mu = \mathcal{E}_{P_\mu}$; we will similarly make other such replacements without comment.

▶ **Definition 2.5.** *Keeping the assumptions above, we say that P satisfies a* ***Poincaré inequality*** *with constant C if for all $f : \Omega \to \mathbb{R}$,*

$$\mathrm{Var}_\mu(f) \le C\mathcal{E}_P(f, f).$$

We say μ satisfies a Poincaré inequality with constant C if the Glauber dynamics with stationary distribution μ, P_μ, satisfies a Poincaré inequality with constant C.

When P is self-adjoint (M is reversible), this is the same as saying that $\lambda_2(P) \le 1 - \frac{1}{C}$, where $\lambda_k(\cdot)$ denotes the kth largest eigenvalue.

▶ **Definition 2.6.** *Let $f : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ be a strictly convex function with $f(1) = 0$. For measures $\nu \ll \mu$ on Ω, define the* ***f-divergence*** *by*

$$\mathcal{D}_f(\nu\|\mu) = \mathbb{E}_{x\sim\mu} f\left(\frac{\nu(x)}{\mu(x)}\right).$$

In particular, define the χ^2 and KL-divergences by $\mathcal{D}_{\chi^2} = D_{(x-1)^2}$ and $\mathcal{D}_{\mathrm{KL}} = D_{x\ln x}$.

▶ **Definition 2.7.** *We say that Markov kernel $P : \Omega_1 \times \Omega_2 \to \mathbb{R}$ satisfies* ***ρ-contraction in f-divergence*** *with respect to μ_1 if for all $\nu_1 \ll \mu_1$,*

$$\mathcal{D}_f(\nu_1 P\|\mu_1 P) \le \rho\mathcal{D}_f(\nu_1\|\mu_1).$$

Contraction in χ^2 and KL-divergence is also referred to as variance or entropy contraction, respectively.

▶ **Proposition 2.8.** *Let $P : \Omega_1 \times \Omega_2 \to \mathbb{R}_{\ge 0}$ be a Markov kernel. The following are equivalent, for $C \le 1$:*

1. *P satisfies $(1 - C)^2$-contraction in χ^2-divergence with respect to μ.*
2. *For all $f : \Omega_1 \to \mathbb{R}$,*

$$\mathrm{Var}_{\mu P}(Pf) \le (1 - C)^2 \mathrm{Var}_\mu(f).$$

3. *(For $\Omega_1 = \Omega_2$, P reversible) P satisfies a Poincaré inequality with constant $\frac{1}{C}$.*
4. *(For P of the form $P = DD^*$, e.g., $P^{\nabla}_{k\leftrightarrow k-1} = D_{k\to k-1}U_{k-1\to k}$) D satisfies $(1-C)$-contraction in χ^2-divergence.*
5. *(For $P = DD^*$) D^* satisfies $(1-C)$-contraction in χ^2-divergence.*

Here, the adjoint is with respect to the measures μ and μD.

Proof sketch. See full version. ◀

▶ **Definition 2.9.** *A measure μ on $\binom{[n]}{k}$ satisfies C-**approximate tensorization of entropy** if $D_{k\to k-1}$ satisfies $\left(1-\frac{1}{Ck}\right)$-contraction in KL-divergence, i.e., for any $\nu \ll \mu$,*

$$\mathcal{D}_{\mathrm{KL}}(\nu D_{k\to k-1}\|\mu D_{k\to k-1}) \le \left(1-\frac{1}{Ck}\right)\mathcal{D}_{\mathrm{KL}}(\nu\|\mu).$$

We have the following alternate characterization for a measure defined on a product space. Define the entropy of a function f on a probability space by $\mathrm{Ent}_\mu[f] = \mathbb{E}_\mu[f\ln f] - \mathbb{E}_\mu[f]\ln\mathbb{E}_\mu[f]$.

▶ **Proposition 2.10** ([19, Lemma 2.7])**.** *Let μ be a measure on $\Omega = \Omega'_1\times\cdots\times\Omega'_n$. Then μ satisfies C-approximate tensorization of entropy iff for all $f:\Omega\to\mathbb{R}_{\ge 0}$,*

$$\mathrm{Ent}_\mu[f] \le C\sum_{k=1}^{n}\mathbb{E}_\mu\left[\mathrm{Ent}_{\mu(X_k=\cdot|X_{\sim k}=x_{\sim k})}[f]\right],$$

where $\sim k$ denotes the coordinates besides k.

▶ **Remark 2.11.** *Proposition 2.8 shows that for contraction in χ^2-divergence, nothing is lost if we consider $P^{\nabla}_{k\leftrightarrow k-1}$ or $D_{k\to k-1}$, $U_{k-1\to k}$ separately. However, the distinction is important for KL, as contraction of $P^{\nabla}_{k\leftrightarrow k-1}$ may not imply contraction of $D_{k\to k-1}$ or $U_{k-1\to k}$ separately; hence the definition of approximate tensorization of entropy. Approximate tensorization of entropy is stronger than the modified log-Sobolev inequality (which implies mixing for $P^{\nabla}_{k\leftrightarrow k-1}$), but weaker than the log-Sobolev inequality.*

2.3 Additional notation

For $f:\prod_{i=1}^n\Omega'_i\to\mathbb{R}$, and $x\in\prod_{i\in S^c}\Omega'_i$, define the restriction $f_x:\prod_{i\in S}\Omega'_i$ by $f_x(y)=f(x,y)$ with (x,y) treated as an element of $\prod_{i=1}^n\Omega'_i$.

Let $x_{i\leftarrow b}$ denote x with x_i set to b. For $f:\{\pm1\}^n$, let $\mathrm{D}_i f(x) := \frac{1}{2}[f(x_{i\leftarrow 1}) - f(x_{i\leftarrow -1})]$ and define $\nabla f:\{\pm1\}^n\to\mathbb{R}^n$ by

$$\nabla f(x) = (\mathrm{D}_1 f(x),\dots,\mathrm{D}_n f(x))$$

and $\nabla^2 f:\{\pm1\}^n\to\mathbb{R}^{n\times n}$ by $(\nabla^2 f(x))_{i,j} = \mathrm{D}_i\mathrm{D}_j f(x)$ (note $(\nabla^2 f(x))_{i,i}=0$).

For $x\in\{\pm1\}^n$ and $S\subseteq[n]$, let x^S denote $\prod_{i\in S}x_i$. For a function $f:\{\pm1\}^n\to\mathbb{R}$, we denote the degree d part of f by $f^{(d)}$, and define $f^{\ge d}=\sum_{p\ge d}f^{(p)}$, etc., so that we have the decomposition

$$f(x) = \sum_{p=0}^{n} f^{(p)}(x) = \sum_{p=0}^{n}\sum_{|I|=p} a_I x^I$$

for some coefficients a_I. We take $f_x^{(d)}$ to mean that we take the restriction first and then the degree-d part.

For a scalar-valued function f and $x\in\mathbb{R}^n$, we let $f(x)$ denote coordinate-wise evaluation.

3 k-Glauber mixes k times as fast

To show that k-Glauber mixes k times more quickly than Glauber dynamics, write $P_\mu = D_{n\to n-1}U_{n-1\to n}$ and $P_{\mu,k} = D_{n\to n-1}\cdots D_{n-k+1\to n-k}U_{n-k\to n-k+1}\cdots U_{n-1\to n}$; the task is then to show that $D_{j\to j-1}$, $j \le n$ are roughly at least as contractive as $D_{n\to n-1}$. We note that this is like the reverse of the usual "local-to-global" argument for high-dimensional expanders [1] (and will be easier!). We also note this approach relates to inductive arguments in prior work (e.g., [20, Lemma 11]).

Viewing the problem in this way, we note the similarity to the proximal sampler [16], each step of which involves adding and removing Gaussian noise; they bound the contraction in χ^2 and KL-divergence of this process based on the Poincaré and log-Sobolev constants of the original distribution.

As inspiration, we first recall the following fact, which bounds the Poincaré or log-Sobolev constant of a convolution of two measures on $\mathbb{R}^n$. (The convolution $\mu_1 * \mu_2$ is defined as the distribution of $X+Y$ where $X \sim \mu_1$ and $Y \sim \mu_2$ are independent.) As we will not cover the theory of functional inequalities over $\mathbb{R}^n$, this is meant only as a suggestion of how we might proceed. For details and generalizations, see [14].

▶ **Lemma.** *Suppose that μ_1, μ_2 are distributions on $\mathbb{R}^n$ with Poincaré constants C_1, C_2, respectively. Then $\mu_1 * \mu_2$ has Poincaré constant bounded by $C_1 + C_2$. The same holds true for the log-Sobolev constant. In particular, this holds true for μ_2 being a Gaussian of variance C_1.*

Proof sketch. Let M_m denote multiplication by m. Then $\mu_i M_{m_i}^{-1}$ has Poincaré constant $m_i^2 C_i$. By tensorization, $\mu = \mu_1 M_{m_1}^{-1} \otimes \mu_2 M_{m_2}^{-1}$ has Poincaré constant $\max\left\{C_1 m_1^2, C_2 m_2^2\right\}$. Consider the projection $\pi(x_1, x_2) = \frac{x_1}{m_1} + \frac{x_2}{m_2}$. When $\frac{1}{m_1^2} + \frac{1}{m_2^2} = 1$, this can be realized as projection onto the vector $(\frac{1}{m_1}, \frac{1}{m_2})$, so the Poincaré constant does not increase: $C_{\mathrm{P}}(\mu\pi^{-1}) \le C_{\mathrm{P}}(\mu)$. Note that we exactly have $\mu\pi^{-1} = \mu_1 * \mu_2$. Choosing $\frac{1}{m_1^2} = \frac{C_1}{C_1+C_2}$ and $\frac{1}{m_2^2} = \frac{C_2}{C_1+C_2}$ gives the bound. The same argument works for the log-Sobolev constant. ◀

We will carry out the same steps as in this proof: define a tensorization operation which preserves the Poincaré or approximate tensorization of entropy constant, and a projection which can only improve it. As in the sketch above, it will help to weight the components appropriately in the tensorization step.

To obtain the $D_{k\to k-1}$ operator from the $D_{n\to n-1}$ operator, we will tensorize with the appropriate "noise" distribution, which in this case is that of erasing $n-k$ coordiates. We note that while bit-flip noise is the more natural analogue of gaussian noise, the "denoising" step is more difficult to implement, while erasure noise connects more nicely with existing notions.

3.1 Tensorization and projection

The following proposition is similar to classical results on preservation of functional inequalities under tensorization, which corresponds to taking a product of Markov chains. However, we need to work with operators between different spaces to obtain contraction for the down operator, so we need the result given in Proposition 3.1 . As in Remark 2.11, nothing would be lost for χ^2-divergence if we considered the down-up walk – so we could use existing results – but for KL we need to bound contraction for just the down operator. This is important because we aim to bound contraction for $P^{\nabla}_{n\leftrightarrow n-k} = D_{n\to n-1}\cdots D_{n-k+1\to n-k}U_{n-k\to n-k+1}\cdots U_{n-1\to n}$, and in the case of KL, we cannot obtain this by bounding contraction for just operators of the form $P^{\nabla}_{k\leftrightarrow k-1} = D_{k\to k-1}U_{k-1\to k}$.

▶ **Proposition 3.1** (Contraction under tensorization). *Suppose that P_i is a Markov kernel from Ω_i to Ω_i', for $i = 1, 2$. Let I_i denote the identity kernel on Ω_i and define $P = p(P_1 \otimes I_2) + (1-p)(I_1 \otimes P_2)$ as a Markov kernel from $\Omega_1 \times \Omega_2$ to $(\Omega_1' \times \Omega_2) \sqcup (\Omega_1 \times \Omega_2')$.*

Let $\mathcal{D}_f = \mathcal{D}_{\mathrm{KL}}$ or $\mathcal{D}_{\chi^2}$. If P_i satisfies $(1-\kappa_i)$-contraction in f-divergence with respect to μ_i, then P satisfies $(1-\kappa)$-contraction in f-divergence with respect to $\mu_1 \otimes \mu_2$, where $\kappa = \min\{p\kappa_1, (1-p)\kappa_2\}$. In particular, if $p = \frac{\kappa_2}{\kappa_1+\kappa_2}$, then $\kappa = \frac{\kappa_1\kappa_2}{\kappa_1+\kappa_2} = \frac{1}{\kappa_1^{-1}+\kappa_2^{-1}}$.

Proof. First consider KL-divergence. Let ν be a measure on $\Omega_1 \times \Omega_2$. Let Π_i and Π_i' denote the projection kernels to Ω_i and Ω_i' respectively, for $i = 1, 2$ (i.e., taking marginals). We calculate

$$\begin{aligned} B :&= \mathcal{D}_{\mathrm{KL}}(p\nu(P_1 \otimes I_2) + (1-p)\nu(I_1 \otimes P_2) \| p(\mu_1' \times \mu_2) + (1-p)(\mu_1 \times \mu_2')) \\ &= p\mathcal{D}_{\mathrm{KL}}(\nu(P_1 \otimes I_2) \| \mu_1' \times \mu_2) + (1-p)\mathcal{D}_{\mathrm{KL}}(\mu(I_1 \otimes P_2) \| \mu_1 \times \mu_2'), \end{aligned} \tag{2}$$

since the spaces $\Omega_1' \times \Omega_2$ and $\Omega_1 \times \Omega_2'$ are disjoint. Now, by the chain rule of KL-divergence and entropy contraction,

$$\begin{aligned} \mathcal{D}_{\mathrm{KL}}(\nu(P_1 \otimes I_2) \| \mu_1' \times \mu_2) &= \mathbb{E}_{x_2 \sim \mu_2} \mathcal{D}_{\mathrm{KL}}((\nu(P_1 \otimes I_2))(\cdot|x_2) \| \mu_1') + \mathcal{D}_{\mathrm{KL}}(\nu(P_1 \otimes I_2)\Pi_2 \| \mu_2) \\ &= \mathbb{E}_{x_2 \sim \mu_2} \mathcal{D}_{\mathrm{KL}}(\nu(\cdot|x_2)P_1 \| \mu_1') + \mathcal{D}_{\mathrm{KL}}(\nu\Pi_2 \| \mu_2) \\ &\le (1-\kappa_1)\mathbb{E}_{x_2 \sim \mu_2} \mathcal{D}_{\mathrm{KL}}(\nu(\cdot|x_2) \| \mu_1) + \mathcal{D}_{\mathrm{KL}}(\nu\Pi_2 \| \mu_2). \end{aligned}$$

By symmetry, the same calculation holds for the second term in (2), giving us

$$\begin{aligned} B \le\ & p\left[(1-\kappa_1)\mathbb{E}_{x_2 \sim \mu_2} \mathcal{D}_{\mathrm{KL}}(\nu(\cdot|x_2) \| \mu_1) + \mathcal{D}_{\mathrm{KL}}(\nu\Pi_2 \| \mu_2)\right] \\ &+ (1-p)\left[(1-\kappa_2)\mathbb{E}_{x_1 \sim \mu_1} \mathcal{D}_{\mathrm{KL}}(\nu(\cdot|x_1) \| \mu_2) + \mathcal{D}_{\mathrm{KL}}(\nu\Pi_1 \| \mu_1)\right]. \end{aligned}$$

We wish to compare this with

$$\begin{aligned} A :&= \mathcal{D}_{\mathrm{KL}}(\mu \| \mu_1 \times \mu_2) \\ &= \underbrace{\mathbb{E}_{x_2 \sim \mu_2} \mathcal{D}_{\mathrm{KL}}(\nu(\cdot|x_2) \| \mu_1)}_{C_{2,1}} + \underbrace{\mathcal{D}_{\mathrm{KL}}(\nu\Pi_2 \| \mu_2)}_{C_2} \\ &= \underbrace{\mathbb{E}_{x_1 \sim \mu_1} \mathcal{D}_{\mathrm{KL}}(\nu(\cdot|x_1) \| \mu_2)}_{C_{1,2}} + \underbrace{\mathcal{D}_{\mathrm{KL}}(\nu\Pi_1 \| \mu_1)}_{C_1}. \end{aligned}$$

By convexity of KL-divergence, $C_{1,2} \ge C_2$ and $C_{2,1} \ge C_1$. Hence

$$\begin{aligned} B &\le p((1-\kappa_1)C_{2,1} + C_2) + (1-p)((1-\kappa_2)C_{1,2} + C_1) \\ &\le \max\{(1-p)(1-\kappa_2) + p, p(1-\kappa_1) + (1-p)\}[p(C_{2,1} + C_2) + (1-p)(C_{1,2} + C_1)] \\ &= (1 - \min\{(1-p)\kappa_2, p\kappa_1\})\, A. \end{aligned}$$

Next consider χ^2-divergence. It suffices to prove the following equivalent statement on contraction of variance (Proposition 2.8): for any $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$, considering $P^* = p(P_1^* \otimes I_2) \oplus (1-p)(I_1 \otimes P_2^*)$, we have

$$\mathrm{Var}_{\mu P}(P^* f) \le (1-\kappa)\,\mathrm{Var}_{\mu}(f),$$

which is equivalent to $\sigma_2(P^*) \le 1-\kappa$ and hence to $\lambda_2(PP^*)^2 = \sigma_2(PP^*)^2 \le (1-\kappa)^2$. Now $PP^* = p(P_1P_1^* \otimes I_2) + (1-p)(I_1 \otimes P_2P_2^*)$ is exactly the transition matrix of the weighted product of two Markov chains with $\lambda_2(P_iP_i^*) \le 1-\kappa_i$, so it is well-known that

$$\lambda_2(PP^*) \le \max\{p + (1-p)(1-\kappa_2), (1-p) + p(1-\kappa_1)\} = 1 - \max\{(1-p)\kappa_2, p\kappa_1\}.$$

(A quick way to see this is as follows: if $\{f_i\}$ are the eigenvectors of $P_1P_1^*$ and $\{g_j\}$ are the eigenvectors of $P_2P_2^*$, then $\{f_ig_j\}$ are the eigenvectors of P^*P, and the second largest eigenvalue is when $f_i = 1$ or $g_i = 1$.) ◀

Though we do not do it here, the proofs can be put on the same footing and generalized using the notion of f-entropy [14].

▶ **Proposition 3.2** (Contraction under projection). *Suppose that P is a Markov kernel from Ω_1 to Ω_2, and μ_1, μ_2 are measures on Ω_1, Ω_2 such that $\mu_1 P = \mu_2$. Let $\pi_i : \Omega_i \to \Omega_i'$ be maps and $\mu_i' = \mu_i \pi_i^{-1}$. Define the projected Markov kernel $P' : \Omega_1' \times \Omega_2' \to \mathbb{R}_{\geq 0}$ by*

$$P'(x_1', x_2') = \sum_{\substack{x_1 \in \Omega_1 \\ \pi_1(x_1) = x_1'}} \sum_{\substack{x_2 \in \Omega_2 \\ \pi_2(x_2) = x_2'}} \mu_1(x_1 | \pi(x_1) = x_1') P(x_1, x_2).$$

If P satisfies ρ-contraction in f-divergence with respect to μ_1, then P' also satisfies ρ-contraction in f-divergence with respect to μ_1'.

In words, in the projected Markov chain, given x_1', we draw x_1 projecting to x_1' from the "prior", move according to P, and then project back down; then the projected kernel always has at least as much contraction as the original one. See e.g., [37] for the statement for the Poincaré constant.

Proof. Let $\nu_1' \ll \mu_1'$ be a measure on Ω_1'. Define $\nu_1(x) = \nu_1'(\pi_1(x))\mu_1(X = x | \pi_1(X) = \pi_1(x))$. Then $\mathbb{E}_{\mu_1(\cdot|\pi_1(X)=x_1')} f\left(\frac{\nu_1(x)}{\mu_1(x)}\right) = \frac{\nu_1'(x_1')}{\mu_1'(x_1')}$, so

$$\mathcal{D}_f(\nu_1' \| \mu_1') = \mathbb{E}_{x_1' \sim \mu_1'} f\left(\frac{\nu_1'(x_1')}{\mu_1'(x_1')}\right) = \mathcal{D}_f(\nu_1 \| \mu_1).$$

By contraction of P and the data processing inequality,

$$\mathcal{D}_f((\nu_1 P)\pi_2^{-1} \| \mu_2') \leq \mathcal{D}_f(\nu_1 P \| \mu_2) \leq \rho \mathcal{D}_f(\nu_1 \| \mu_1) = \rho \mathcal{D}_f(\nu_1' \| \mu_1').$$

Finally, note that $(\nu_1 P)\pi_2^{-1} = \nu_1' P'$ by definition of P'. ◀

3.2 Contraction improves going down

We now show that for $k < n$, contraction in KL and χ^2 for $D_{k \to k-1}$ will only be better than contraction of $D_{n \to n-1}$, except up to an additive constant.

▶ **Lemma 3.3.** *Let μ be the uniform distribution on $\binom{[n]}{k}$. Then $D_{k \to k-1}$ has $\left(1 - \frac{1}{k}\right)$-contraction in KL, and μ satisfies 1-approximate tensorization of entropy. Moreover, $D_{k \to k-1}$ and $U_{k-1 \to k}$ satisfy $\left(1 - \frac{n}{k(n-k+1)}\right)$-contraction in χ^2.*

We note that the down-up walk on the uniform distribution on $\binom{[n]}{k}$ is a rescaling of the Bernoulli-Laplace diffusion model, for which mixing and functional inequalities have been extensively studied [21, 33, 27, 13, 43], and we have not attempted to find the best result for KL.

Proof. Note that μ is log-concave, by the results of [5] and the fact that $\binom{[n]}{k}$ is a matroid. Then 1-approximate tensorization of entropy follows from [8, Theorem 5].

To note the improved bound for χ^2, note first that the probability of staying at the same set is $\frac{1}{n-k+1}$, so $P^{\nabla}_{k \leftrightarrow k-1} = \frac{1}{n-k+1} I + \frac{n-k}{n-k+1} P^{\vee}_{k \leftrightarrow k-1}$ where $P^{\vee}_{k \leftrightarrow k-1}$ is the "non-lazy" walk which swaps an occupied and non-occupied space at random. Hence $I - P^{\nabla}_{k \leftrightarrow k-1} = \frac{n-k}{n-k+1}(I - P^{\vee}_{k \leftrightarrow k-1})$. This in turn satisfies

$$I - P^{\vee}_{k \leftrightarrow k-1} = \frac{n(n-1)/2}{k(n-k)} \cdot \frac{2}{n-1} \cdot L$$

where L is the generator of the Bernoulli-Laplace diffusion model with parameters (n, k) (a random transposition occurs with rate 1). Here, the $\frac{2}{n-1}$ is the scaling factor to convert to a discrete-time walk and $\frac{n(n-1)/2}{k(n-k)}$ takes into account that we are randomly choosing from $k(n-k)$ transpositions that move a particle to an unoccupied space, rather than an arbitrary transposition. By [43], $\lambda_{\min}(L) = 1$. Hence the spectral gap for $P^{\nabla}_{k\leftrightarrow k-1}$ is

$$\frac{n-k}{n-k+1} \cdot \frac{n(n-1)/2}{k(n-k)} \cdot \frac{2}{n-1} = \frac{n}{k(n-k+1)}. \qquad \blacktriangleleft$$

▶ **Lemma 3.4.** *Let μ be a distribution on $\prod_{i=1}^{n} \Omega_i'$ and μ^{hom} its homogenization on $\binom{\Omega}{n}$, where $\Omega = \bigsqcup_{i=1}^{n} \Omega_i'$, and $\mu_m = \mu^{hom} D_{n\to m}$. Let $\mathcal{D}_f = \mathcal{D}_{\mathrm{KL}}$ or $\mathcal{D}_{\chi^2}$. For each k, let κ_k be the largest number such that $D_k := D_{k\to k-1}$ satisfies $(1-\kappa_k)$-contraction in f-divergence with respect to μ_k. Suppose that for the uniform distribution $\binom{[n]}{k}$ that $D_{k\to k-1}$ satisfies $(1-\kappa_{\mathrm{BL},k})$-contraction in f-divergence. Then*

$$\kappa_k \geq \frac{n\kappa_n\kappa_{\mathrm{BL},k}}{n\kappa_n + k\kappa_{\mathrm{BL},k}} \geq \begin{cases} \frac{n\kappa_n}{k((n-k+1)\kappa_n+1)}, & \mathcal{D}_f = \mathcal{D}_{\chi^2} \\ \frac{n\kappa_n}{k(n\kappa_n+1)}, & \mathcal{D}_f = \mathcal{D}_{\mathrm{KL}}. \end{cases}$$

Proof. Consider the kernel $P = p(D_n \otimes I_{\binom{[n]}{k}}) + (1-p)(I_{\binom{\Omega}{n}} \otimes D_k)$ from $\binom{\Omega}{n} \times \binom{[n]}{k}$ to $\binom{\Omega}{n-1} \times \binom{[n]}{k} \cup \binom{\Omega}{n} \times \binom{[n]}{k-1}$, where D_k denotes the down operator $\binom{[n]}{k} \times \binom{[n]}{k-1} \to \mathbb{R}_{\geq 0}$. By tensorization (Proposition 3.1), for $p = \frac{\kappa_{\mathrm{BL},k}}{\kappa_n+\kappa_{\mathrm{BL},k}}$ and $\kappa = \frac{1}{\kappa_n^{-1}+\kappa_{\mathrm{BL},k}^{-1}}$, P satisfies $(1-\kappa)$-contraction in f-divergence with respect to $\mu_n \otimes \mathsf{Uniform}\binom{[n]}{k}$. Define the projections $\pi_1 : \binom{\Omega}{n} \times \binom{[n]}{k} \to \binom{\Omega}{k}$ and $\pi_2 : \binom{\Omega}{n-1} \times \binom{[n]}{k} \cup \binom{\Omega}{n} \times \binom{[n]}{k-1} \to \binom{\Omega}{k-1} \cup \binom{\Omega}{k}$ both as

$$\pi(S, A) = S \cap \bigcup_{i\in A} \Omega_i,$$

i.e., if S corresponds to $x \in \prod_{i=1}^{n} \Omega_i'$, we keep only the coordinates in the set $A \in \binom{[n]}{k}$.

We claim the projected kernel as defined in Proposition 3.2 is

$$P' = \underbrace{\left(1 - \frac{p(n-k)}{n}\right) D_k}_{\text{(I)}} + \underbrace{\frac{p(n-k)}{n} I}_{\text{(II)}}$$

from $\binom{\Omega}{k}$ to $\binom{\Omega}{k-1} \cup \binom{\Omega}{k}$. To see this, we first identify $x_A \in \prod_{=1}^{i\in A} \Omega_i'$ with its homogenization in $\binom{\Omega}{|A|}$ (Definition 2.3). Then $\mu_1(\cdot|\pi_1(x_1) = x_A)$ is the distribution of (x, A) where x_{A^c} is distributed as $\mu(X_{A^c} = x_{A^c} | X_A = x_A)$. Under the transition P:

1. With probability p, we remove a coordinate i of x. In this case,
   a. with probability $\frac{k}{n}$, $i \in A$, and projection by π_2 then gives $x_{A\setminus\{i\}}$ for a random index i.
   b. with probability $\frac{n-k}{n}$, $i \notin A$, and projection by π_2.
2. With probability $1-p$, we remove an element i of A, and projection gives $x_{A\setminus\{i\}}$ for a random $i \in A$.

Then (1a) and (2) give the term (I) and (1b) gives the term (II).

Because

$$\begin{aligned} &\mathcal{D}_f\left(\nu P' \| \left(1 - \frac{p(n-k)}{n}\right)\mu_{k-1} + \frac{p(n-k)}{n}\mu_k\right) \\ &= \left(1 - \frac{p(n-k)}{n}\right)\mathcal{D}_{\mathrm{KL}}(\nu D_k \| \mu_{k-1}) + \frac{p(n-k)}{n}\mathcal{D}_{\mathrm{KL}}(\nu\|\mu_k) \end{aligned}$$

(the component measures have disjoint support), we have that D_k satisfies $(1-\kappa_k)$-contraction in f-divergence iff P' satisfies $\left(1-\kappa_k\left(1-\frac{p(n-k)}{n}\right)\right)$-contraction in f-divergence. Proposition 3.2 then gives us

$$\kappa_k\left(1-\frac{\kappa_{\mathrm{BL},k}}{\kappa_n+\kappa_{\mathrm{BL},k}}\cdot\frac{n-k}{n}\right)\geq\frac{1}{\kappa_n^{-1}+\kappa_{\mathrm{BL},k}^{-1}}$$
$$\implies \kappa_k\geq\frac{n\kappa_n\kappa_{\mathrm{BL},k}}{n\kappa_n+k\kappa_{\mathrm{BL},k}}.$$

Plugging in the result of Lemma 3.3 then gives the result. ◀

▶ **Theorem 3.5.** *Let μ be a distribution on $\Omega=\prod_{i=1}^n\Omega_i'$, and consider the down and up operators defined with respect to its homogenization μ^{hom}.*

1. *If $D_{n\to n-1}$ satisfies $(1-\frac{1}{Cn})$-contraction in χ^2, then $D_{k\to\ell}$ satisfies $\prod_{j=\ell+1}^k\left(1-\frac{1}{j\left(C+\frac{n-j+1}{n}\right)}\right)$-contraction in χ^2.*
2. *If $D_{n\to n-1}$ satisfies $(1-\frac{1}{Cn})$-contraction in KL (i.e., μ satisfies C-approximate tensorization of entropy), then $D_{k\to\ell}$ and $P^{\nabla}_{k\leftrightarrow\ell}$ satisfy $\prod_{j=\ell+1}^k\left(1-\frac{1}{j(C+1)}\right)$-contraction in KL.*

We note that our tensorization construction in Lemma 3.1 currently depends on μ being a homogeneous distribution; it would be interesting to extend it beyond this case.

Proof. If $D_{n\to n-1}$ satisfies $\left(1-\frac{1}{Cn}\right)$-contraction in χ^2-divergence, then by Lemma 3.4, $D_{j\to j-1}$ satisfies $\left(1-\frac{1}{j\left(C+\frac{n-j+1}{n}\right)}\right)$-contraction in χ^2-divergence. If $D_{n\to n-1}$ satisfies $\left(1-\frac{1}{Cn}\right)$-contraction in χ^2-divergence, then we have $\left(1-\frac{1}{j(C+1)}\right)$-contraction in KL-divergence. Because $D_{k\to\ell}=D_{k\to k-1}\cdots D_{\ell+1\to\ell}$, taking the product from $\ell+1$ to k gives the result. ◀

From this, we can conclude that k-Glauber dynamics mixes at least $\Omega(k)$ times as fast in χ^2-divergence as Glauber dynamics, and assuming approximate tensorization of entropy, mixes at least $\Omega(k)$ times as fast in KL-divergence.

▶ **Corollary 3.6** (k-Glauber mixes k times as fast)**.** *Let μ be a distribution on $\Omega=\prod_{i=1}^n\Omega_i'$.*

1. *If μ satisfies a Poincaré inequality with constant Cn, then*

$$\lambda_2(P_{\mu,k})\leq\left(1-\frac{k}{n+1}\right)^{\frac{1}{C+1}}=1-\Omega\left(\max\left\{\frac{k}{(C+1)n},1\right\}\right)$$

   and $P_{\mu,k}$ satisfies a Poincaré inequality with constant $\Omega\left(\frac{(C+1)n}{k}\right)$.
2. *If μ satisfies C-approximate tensorization of entropy, then $P_{\mu,k}$ satisfies contraction in KL-divergence with constant*

$$\left(1-\frac{k}{n+1}\right)^{\frac{1}{C+1}}=1-\Omega\left(\max\left\{\frac{k}{(C+1)n},1\right\}\right).$$

Proof. Note that $P_{\mu,k}$ corresponds to $P^{\nabla}_{n\leftrightarrow n-k}$ after homogenization. For variance, we recall Proposition 2.8 which relates the Poincaré constant of P_μ and $D_{n\to n-1}$, and the Poincaré constant of $P_{\mu,k}$ and $D_{n\to n-k}$. The corollary then follows from Theorem 3.5 and the calculation

$$\prod_{j=n-k+1}^{n}\left(1-\frac{1}{j(C+1)}\right) \le e^{-\frac{1}{C+1}\sum_{j=n-k+1}^{n}\frac{1}{j}} \le e^{-\frac{1}{C+1}\ln\left(\frac{n+1}{n-k+1}\right)}$$
$$= \left(1-\frac{k}{n+1}\right)^{\frac{1}{C+1}} = 1-\Omega\left(\frac{k}{(C+1)n}\right). \quad \blacktriangleleft$$

Our theorem also implies contraction for $D_{2\to 1}$; this forms a kind of converse to local-to-global arguments that start with contraction of $D_{2\to 1}$ [2].

▶ **Corollary 3.7.** *Let μ be a distribution on $\Omega = \prod_{i=1}^{n}\Omega_i'$, and consider the down and up operators defined with respect to its homogenization μ^{hom}.*

1. *If $D_{n\to n-1}$ satisfies $(1-\frac{1}{Cn})$-contraction in χ^2, then $D_{2\to 1}$ satisfies $\left(1-\frac{1}{2(C+1)}\right)$-contraction in χ^2 and $\lambda_2(P^{\triangle}_{1\leftrightarrow 2}) \le 1-\frac{1}{2(C+1)}$.*
2. *If $D_{n\to n-1}$ satisfies $(1-\frac{1}{Cn})$-contraction in KL (i.e., μ satisfies C-approximate tensorization of entropy), then $D_{2\to 1}$ satisfies $\left(1-\frac{1}{2(C+1)}\right)$-contraction in KL.*

Proof. This follows from substituting $k=2$, $\ell=1$ into Theorem 3.5, appealing to Proposition 2.8 for the χ^2 result. ◀

4 Parallel sampling for Ising models

To apply Corollary 3.6 to the Ising model for $\|J\| < 1$, we use the fact that the Ising model satisfies approximate tensorization of entropy.

▶ **Theorem 4.1.** *Suppose $\|J\| < 1$. Then $\mu_{J,h}$ satisfies approximate tensorization of entropy with constant $\frac{1}{1-\|J\|}$.*

The proof is in Section 4.1. We first introduce a generic guarantee for approximate rejection sampling in Section 4.2, based on establishing concentration for the difference of the log-pdfs. In Section 4.3, we show that using an approximating product distribution – chosen as the solution to a variational problem – is sufficient as a proposal distribution for $\mu_{J,h}$ when $\|J\|_F$ is small enough, using the Hanson-Wright inequality. We put everything together in Section 4.4, combining the speedup of k-Glauber dynamics, known mixing for the Ising model, guarantee on the approximate rejection sampler, with a careful analysis of the recursion in the algorithm.

4.1 Approximate tensorization of entropy for the Ising model

For $\|J\| < 1$, we prove approximate tensorization of entropy for the Ising model $\mu_{J,h}$ (Theorem 4.1) by using the fact that it holds for rank-1 Ising models and using the needle decomposition in [24]. This is the same way that the modified log-Sobolev inequality is proved in [24].

▶ **Proposition 4.2** ([8, Proposition 32])**.** *Suppose $\|u\| < 1$. Then $\mu_{uu^\top,h}$ satisfies approximate tensorization of entropy with constant $\frac{1}{1-\|u\|^2}$.*

▶ **Theorem 4.3** (Needle decomposition of Ising measures [24])**.** *Consider an Ising model $\mu = \mu_{J,h}$ on $\{\pm 1\}^n$ with $J \succeq 0$. Let $f : \{\pm 1\}^n \to \mathbb{R}$ be any function. There exists a mixture decomposition (depending on f)*

$$\mu(x) = \int \mu_{u,v}(x)\, d\pi(u,v)$$

where π is a probability measure on $\mathbb{R}^{2n}$ such that:

1. *π-almost surely, $\mu_{u,v}$ is a probability measure of the form*

$$\mu_{u,v}(x) = \frac{1}{Z_{u,v}} \exp\left(\frac{1}{2}\langle u, x\rangle^2 + \langle v, x\rangle\right),$$

i.e., a rank one Ising model ("needle"). Furthermore $\|u\| \le \|J\|$.
2. *$\mathbb{E}_{\mu_{u,v}} f(X) = \mathbb{E}_{\mu} f(X)$ π-almost surely.*

Proof of Theorem 4.1. See full version. ◀

4.2 Approximate rejection sampling

Conditional distributions of the Ising model are again Ising models. We will show that with large probability, if we pick a random subset of not-too-large size, then we can approximate the distribution of those coordinates with a product distribution, and hence use the product distribution as a proposal for approximate rejection sampling. We discuss further the choice of product distribution (given in Algorithm 2) in Section 4.3.

First, we give a generic guarantee for the rejection sampling Algorithm 3, which can be used whenever the log-ratio between desired and proposal distributions $\ln \frac{dP}{dQ}$ has sufficiently decaying exponential tails. Note that because the normalizing constants are unknown, we draw two samples and use one as a reference to decide whether to accept the other. Lemma 4.4 appears as [25, Lemma 2] specialized to the distribution they consider, but holds more generally. We give the proof for completeness.

Algorithm 2 Quadratic approximate rejection sampler (QuadraticApproxRejectionSampler).

Input: Hamiltonian H in the form $H(x) = C + \langle h, x\rangle + \frac{1}{2}\langle x, Ax\rangle + H_{\ge 3}(x)$, δ such that $\left\|A^{\otimes}\right\|_F \le \delta < c_3$ and $0 \preceq A \preceq (1-c)I$, error ε.
Let $u_0 = \mathbf{0}$.
for t from 1 to $T = \Theta\left(\frac{1}{c}\ln\left(\frac{\sqrt{n}}{\delta}\right)\right)$ **do**
    Let $u_t = A^{\otimes}\tanh(h + u_{t-1})$.
end for
Let $\widehat{h} = u_T$.
Output: ApproxRejectionSampler$\left(q(x) \propto \exp\left(\langle h + \widehat{h}, x\rangle, g(x) = H(x) - \langle h + \widehat{h}, x\rangle, \left(\frac{2}{\varepsilon}\right)^{\frac{\delta}{c_3 - \delta}}\right)\right.$ (See Algorithm 3.)

Algorithm 3 Approximate rejection sampler (ApproxRejectionSampler).

Input: Oracle for sampling from Q, function g such that $\frac{dP}{dQ} \propto e^g$, error parameter c.
repeat (For parallel implementation, run $\lceil c \rceil$ times simultaneously and take the first success.)
    Draw $X, Z \sim Q$.
    Let $R = \exp(g(X) - g(Z))$.
    Draw $U \sim$ Uniform$([0,1])$.
until $U \le \frac{1}{c}R$
Output: X.

▶ **Lemma 4.4.** *Let $\widehat{P}$ be the distribution of the output of Algorithm 3. Then*

$$\frac{dP}{dQ}(X) = \frac{\mathbb{E}[R|X]}{\mathbb{E}R} \qquad \frac{d\widehat{P}}{dQ}(X) = \frac{\mathbb{E}[\min\{R, c\}|X]}{\mathbb{E}[\min\{R, c\}]}.$$

The acceptance probability is $p_{\text{accept}} = \frac{1}{c}\mathbb{E}\min\{R, c\} = \frac{1}{c}(\mathbb{E}R - \mathbb{E}[(R-c)\mathbb{1}_{R\geq c}])$ *and the TV distance is bounded by*

$$\mathcal{D}_{\text{TV}}(\widehat{P}, P) \leq \frac{\mathbb{E}[(R-c)\mathbb{1}_{R\geq c}]}{\mathbb{E}R}.$$

For $c \geq 1$*, the acceptance probability is at least* $\frac{1}{2c}$*.*

Proof. We calculate

$$\mathbb{E}[R|X] = \mathbb{E}[e^{g(X)-g(Z)}|X] = e^{g(X)}\mathbb{E}[e^{-g(Z)}]$$
$$\mathbb{E}[R] = \mathbb{E}[\mathbb{E}[R|X]] = \mathbb{E}[e^{g(X)}]\mathbb{E}[e^{-g(Z)}],$$

so $\frac{\mathbb{E}[R|X]}{\mathbb{E}[R]} = \frac{e^{g(X)}}{\mathbb{E}[e^{g(X)}]} = \frac{dP}{dQ}(X)$. Note $\frac{d\widehat{P}}{dQ}(X)$ is the probability of acceptance given X divided by the total probability of acceptance, which we calculate:

$$p_{\text{accept}}(X) := \mathbb{P}\left[U \leq \frac{1}{c}R|X\right] = \mathbb{E}\left[\min\left\{\frac{R}{c}, 1\right\}|X\right] = \frac{1}{c}\mathbb{E}[\min\{R, c\}|X]$$
$$p_{\text{accept}} = \mathbb{P}\left[U \leq \frac{1}{c}R\right] = \mathbb{E}\left[\mathbb{P}\left[U \leq \frac{1}{c}R|X\right]\right] = \frac{1}{c}\mathbb{E}[\mathbb{E}[\min\{R, c\}|X]] = \frac{1}{c}\mathbb{E}[\min\{R, c\}].$$

Dividing gives $\frac{d\widehat{P}}{dQ} = \frac{\mathbb{E}[\min\{R,c\}|X]}{\mathbb{E}[\min\{R,c\}]}$. Then

$$\begin{aligned}
\mathcal{D}_{\text{TV}}(\widehat{P}, P) &\leq \mathbb{E}_{X\sim Q}\max\left\{0, \frac{dP}{dQ}(X) - \frac{d\widehat{P}}{dQ}(X)\right\} \\
&\leq \mathbb{E}_{X\sim Q}\max\left\{0, \frac{\mathbb{E}[R|X]}{\mathbb{E}R} - \frac{\mathbb{E}[\min\{R,c\}|X]}{cp_{\text{accept}}}\right\} \\
&\leq \mathbb{E}_{X\sim Q}\max\left\{0, \frac{\mathbb{E}[R|X]}{\mathbb{E}R} - \frac{\mathbb{E}[\min\{R,c\}|X]}{\mathbb{E}R}\right\} \qquad \text{because } cp_{\text{accept}} \leq \mathbb{E}R \\
&\leq \mathbb{E}_{X\sim Q}\frac{\mathbb{E}[(R-c)\mathbb{1}_{R\geq c}|X]}{\mathbb{E}R} = \frac{\mathbb{E}[(R-c)\mathbb{1}_{R\geq c}]}{\mathbb{E}R}.
\end{aligned}$$

Finally, note that $\mathbb{P}(R \geq 1) \geq \frac{1}{2}$ by symmetry, so $p_{\text{accept}} \geq \frac{1}{c}\mathbb{P}(R \geq 1) \geq \frac{1}{2c}$. ◀

4.3 Concentration

To obtain concentration of the ratio in Lemma 4.4, we need a version of the Hanson-Wright inequality. We first state the classical inequality.

▶ **Theorem 4.5** (Hanson-Wright Inequality, [49, Theorem 6.2.1])**.** *There is a constant* c *such that the following holds. Let* $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ *be a random vector with independent, mean-zero,* K*-sub-gaussian coordinates. Let* $A \in \mathbb{R}^{n\times n}$ *be a matrix. Then for every* $t \geq 0$*,*

$$\mathbb{P}\left(|\langle X, AX\rangle - \mathbb{E}\langle X, AX\rangle| \geq t\right) \leq 2\exp\left[-c\min\left\{\frac{t^2}{K^4\|A\|_F^2}, \frac{t}{K^2\|A\|}\right\}\right].$$

We will use the following version, which is a consequence of [44, Corollary 2] (by taking $\Gamma(f) = \|\nabla f\|$ and noting that product distributions on $\{\pm 1\}^n$ satisfy a uniform modified log-Sobolev inequality) and allows a general function f.

▶ **Theorem 4.6** ([44])**.** *There is a constant* c *such that the following holds. Let* $X = (X_1, \ldots, X_n) \in \{\pm 1\}^n$ *be a random vector with independent coordinates. Let* $f : \{\pm 1\}^n \to \mathbb{R}$ *be a function. Then for every* $t \geq 0$*,*

$$\mathbb{P}\left(|f(X) - \mathbb{E}f(X)| \geq t\right) \leq 2\exp\left[-c\min\left\{\frac{t^2}{\mathbb{E}[\|\nabla f\|^2]}, \frac{t}{\max_{x\in\{\pm1\}^n}\|\nabla^2 f\|_F}\right\}\right].$$

Note the extra term $\mathbb{E}[\|\nabla f\|^2]$ compared to Theorem 4.5 which requires the random variables to be centered. This means that we cannot simply take $Q = \mu_h := \mu_{O,h}$ and $P = \mu_{J,h}$ for the reason that $\mathbb{E}[\|\nabla_x(\langle x, Ax\rangle)\|^2]$ can be $\Omega(n)$, while we need concentration to $O(1)$. Instead, in order to apply Theorem 4.6 for $f = \ln \frac{dP}{dQ}$, we would like to ∇f to be centered, that is, $\mathbb{E}_Q \nabla f = 0$. For this, we need to solve the variational problem

$$\mathbb{E}_{\mu_{h+h^*}} A^{\otimes} x = h^*. \tag{3}$$

We do this by fixed point iteration. Note this is a special case of "gradient" descent for Lipschitz and strongly monotone operators [15, 36].

▶ **Lemma 4.7** (Fixed point iteration). *Let (X, d) be a metric space, $c > 0$, and suppose $F : X \to X$ is $(1-c)$-Lipschitz (so it is a contraction mapping). Let $x_0 \in X$ and $x_t = F^{(t)}(x_0)$. Then*

$$d(F(x_t), x_t) \le (1-c)^t d(F(x_0), x_0)$$

and hence for $t = \Omega\left(\frac{1}{c} \ln\left(\frac{d(F(x_0),x_0)}{\varepsilon}\right)\right)$, we have $d(F(x_t), x_t) \le \varepsilon$.

Proof. We have $d(F(x_t), x_t) = d(F^{(t+1)}(x_0), F^{(t)}(x_0)) \le (1-c)^t d(F(x_0), x_0)$ by induction. Hence, it suffices to choose t such that $t \ln\left(\frac{1}{1-c}\right) \ge \ln\left(\frac{d(F(x_0),x_0)}{\varepsilon}\right)$, which gives the result. ◀

▶ **Lemma 4.8.** *Suppose that A is symmetric positive semi-definite with $A \preceq (1-c)I$. Let*

$$F(u) = A^{\otimes} \tanh(h + u).$$

Then for $t = \Omega\left(\frac{1}{c} \log\left(\frac{\sqrt{n}}{\varepsilon}\right)\right)$, we have that $\widehat{h} := F^{(t)}(\mathbf{0})$ satisfies

$$\left\|\mathbb{E}_{\mu_{h+\widehat{h}}} A^{\otimes} x - \widehat{h}\right\| \le \varepsilon. \tag{4}$$

Proof. Note that all diagonal entries of A are contained in $[0, 1-c]$, so $-(1-c)I \preceq A^{\otimes} \preceq (1-c)I$. Combining this with the fact that tanh is 1-Lipschitz, we obtain that F is $(1-c)$-Lipschitz. Note that $\|F(\mathbf{0}) - \mathbf{0}\| \le \sqrt{n}$. The result then follows from Lemma 4.7 and the fact that $F(u) = \mathbb{E}_{\mu_{h+u}} A^{\otimes} x$. ◀

Using this, for small enough $\|A\|_F$, we can obtain the exponential tails necessary to bound the TV-distance in Lemma 4.4. This fits in with the general fact that Ising models with small $\|J\|_F$ are well-approximated by product distributions [31], giving approximation guarantees for variational methods in this regime. We state the following lemma more generally with a higher-order term, so that we can also apply it for the p-spin model.

▶ **Lemma 4.9.** *There is a constant c_3 such that the following holds. Suppose $H(x) = \langle h, x\rangle + \frac{1}{2}\langle x, Ax\rangle + H_{\ge 3}(x)$ where A is symmetric and $H_{\ge 3}(x) = \sum_{|I|\ge 3} a_I x^I$ contains the terms of degree ≥ 3. If $\delta < c_3$,*

$$\max_{x\in\{\pm 1\}^n} \left\|\nabla^2 H(x)\right\|_F \le \delta \quad \text{and} \quad \max_{x\in\{\pm 1\}^n} \|\nabla H_{\ge 3}(x)\| \le \delta,$$

then the output of QuadraticApproxRejectionSampler *(Algorithm 2) is at most ε in TV distance from μ, and the acceptance probability in the call to* ApproxRejectionSampler *is at least $\frac{1}{2c} = \frac{1}{2}\left(\frac{\varepsilon}{2}\right)^{\frac{\delta}{c_3-\delta}}$.*

In the special case that $H_{\ge 3}(x) = 0$, the assumption simplifies to $\left\|A^{\otimes}\right\|_F \le \delta$.

Proof. Let

$$f(x) = H(x) - \left\langle h + \widehat{h}, x \right\rangle = \frac{1}{2}\langle x, Ax\rangle - \left\langle \widehat{h}, x\right\rangle + H_{\geq 3}(x).$$

To use Theorem 4.6, we calculate $\mathbb{E}[\|\nabla f\|^2]$. First note that $\mathbb{E}_{x\sim\mathsf{Uniform}(\{\pm1\}^n)}\nabla^2 H_{\geq 3}(x) = O$ (because for any $i, j \in [n]$ and $|I| \geq 3$, we have $\mathbb{E}_{x\sim\mathsf{Uniform}(\{\pm1\}^n)} x^{I\setminus\{i,j\}} = 0$) so by Jensen's inequality

$$\begin{aligned}\left\|A^{\diamondsuit}\right\|_F &= \left\|A^{\diamondsuit} + \mathbb{E}_{x\sim\mathsf{Uniform}(\{\pm1\}^n)}\nabla^2 H_{\geq 3}(x)\right\|_F \\ &\leq \mathbb{E}_{x\sim\mathsf{Uniform}(\{\pm1\}^n)}\left\|A^{\diamondsuit} + \nabla^2 H_{\geq 3}(x)\right\|_F \leq \max_{x\in\{\pm1\}^n}\left\|\nabla^2 H(x)\right\|_F \leq \delta.\end{aligned}$$

Let $\overline{x} = \mathbb{E}_{\mu_{h+\widehat{h}}} x$. From Lemma 4.8 we have that the output $\widehat{h}$ of fixed point iteration satisfies $\left\|A^{\diamondsuit}\overline{x} - \widehat{h}\right\| \leq \delta$. We then have

$$\begin{aligned}\mathbb{E}_{\mu_{h+\widehat{h}}}[\|\nabla f\|^2] &= \mathbb{E}_{\mu_{h+\widehat{h}}}\left[\left\|A^{\diamondsuit}x - \widehat{h} + \nabla H_{\geq 3}(x)\right\|^2\right] \\ &\leq 2\mathbb{E}_{\mu_{h+\widehat{h}}}\left[\left\|A^{\diamondsuit}(x-\overline{x}) + A^{\diamondsuit}\overline{x} - \widehat{h}\right\|^2\right] + 2\mathbb{E}_{\mu_{h+\widehat{h}}}\left[\|\nabla H_{\geq 3}(x)\|^2\right] \\ &= 2\mathbb{E}_{\mu_{h+\widehat{h}}}\left[\left\|A^{\diamondsuit}\overline{x} - \widehat{h}\right\|^2 + \left\|A^{\diamondsuit}(x-\overline{x})\right\|^2 + \|\nabla H_{\geq 3}(x)\|^2\right] \\ &\leq 2\left[\left\|A^{\diamondsuit}\overline{x} - \widehat{h}\right\|^2 + \left\|A^{\diamondsuit}\right\|_F^2 + \mathbb{E}_{\mu_{h+\widehat{h}}}\|\nabla H_{\geq 3}(x)\|^2\right] \leq 6\delta^2 \qquad (5)\end{aligned}$$

using the fact that the entries of $x - \overline{x}$ are independent, mean 0, with variance at most 1. Hence, by Theorem 4.6, $f(X) - \mathbb{E}f(X)$ is $O(\delta)$-sub-exponential, and so is $f(Z) - f(X)$, and there exists c_3 so that

$$\mathbb{P}\left(|f(Z) - f(X)| \geq t\right) \leq 2e^{-\frac{c_3 t}{\delta}}.$$

Then for $c = \left(\frac{2}{\varepsilon}\right)^{\frac{\delta}{c_3-\delta}}$,

$$\begin{aligned}\mathbb{E}[(R-c)\mathbb{1}_{R\geq c}] &\leq \int_{\ln c}^{\infty} e^t \cdot \mathbb{P}(f(Z) - f(X) \geq t)\, dt \\ &\leq \int_{\ln c}^{\infty} e^t 2e^{-\frac{c_3 t}{\delta}}\, dt \leq 2\int_{\ln c}^{\infty} e^{-(\frac{c_3}{\delta}-1)t}\, dt = 2c^{-\left(\frac{c_3}{\delta}-1\right)} = \varepsilon. \qquad (6)\end{aligned}$$

Moreover, by Jensen's inequality, $\mathbb{E}R \geq e^{\mathbb{E}[f(Z)-f(X)]} = 1$. Hence by Theorem 4.6, the output is at most ε in TV distance from μ and the acceptance probability is at least $\frac{1}{2c}$. ◀

4.4 Analysis of the Parallel Ising Sampler

▶ **Lemma 4.10.** *Let $S \subseteq [n]$, fix $x_{S^c} \in \{\pm1\}^{S^c}$, and let P be the distribution on $\{\pm1\}^S$ with mass function $p(x) = \mu_{J,h}(X_S = x | X_{S^c} = x_{S^c})$, and let Q be the product distribution in on $\{\pm1\}^S$ with mass function $q(x) \propto \exp\left(\langle J_{S\times S^c}x_{S^c} + h_S, x\rangle\right)$. Then the following hold.*
1. *$\frac{dP}{dQ}(x) \propto \exp\left(\frac{1}{2}\langle x, J_{S\times S}x\rangle\right)$.*
2. *$\mathcal{D}_{\mathrm{KL}}(P\|Q) \leq \|J_{S\times S}\| \cdot |S|$.*

Proof. Because x_{S^c} is constant, expanding the quadratic gives

$$\mu_{J,h}(X_S = x_S | X_{S^c} = x_{S^c}) \propto \exp\left(\frac{1}{2}(2\langle x_S, J_{S\times S^c}x_{S^c}\rangle + \langle x_S, J_{S\times S}x_S\rangle) + \langle h_S, x\rangle\right).$$

Dividing by $q(x_S)$ gives (1).

For (2), we note that for $x \in \{\pm 1\}^S$, $\left|\frac{1}{2}\langle x, J_{S\times S}x\rangle\right| \le \frac{1}{2}\|J_{S\times S}\|\|x\|^2 \le \frac{1}{2}\|J_{S\times S}\||S|$. Hence

$$\frac{dP}{dQ}(x) = \frac{\exp\left(\frac{1}{2}\langle x, J_{S\times S}x\rangle\right)}{\int \exp\left(\frac{1}{2}\langle x, J_{S\times S}x\rangle\right) dQ(x)} \le \frac{e^{\frac{1}{2}\|J_{S\times S}\||S|}}{e^{-\frac{1}{2}\|J_{S\times S}\||S|}} = e^{\|J_{S\times S}\||S|}$$

and $\mathcal{D}_{\mathrm{KL}}(P\|Q) = \mathbb{E}_P \ln \frac{dP}{dQ} \le \|J_{S\times S}\||S|$. ◀

▶ **Lemma 4.11** (Bernstein's inequality for supermartingales [28, (1.6)]). *Let X_n be a martingale adapted to $\mathcal{F}_n$. Suppose that $|X_{n+1} - X_n| \le L$ and $\mathbb{E}[|X_{n+1} - X_n|^2|\mathcal{F}_n] \le \sigma^2$ with probability 1. Then*

$$\mathbb{P}(X_n - X_0 \ge t) \le \exp\left(-\frac{t^2}{2(tL + n\sigma^2)}\right).$$

We are now ready to prove our main theorem on the parallel Ising sampler.

Proof of Theorem 1.2. We first note that all lines in Algorithm 1 take logarithmic time with poly(n) processors (e.g., by a parallel implementation of matrix-vector multiplication). Note that a random subset of specified size s can be selected by generating a random number for each index, using a parallel sorting algorithm [11], and then selecting the smallest s elements. We will ignore logarithmic overhead for the rest of the proof.

Running time is bounded with high probability. We consider a tree associated with a run of the algorithm, where each node is labeled with a set, constructed as follows. Each node represents a time that ParallelIsingSampler is called, and each leaf node represents a time that ApproxRejectionSampler. Start with a root node v_1 labeled with $S_1 = [n]$. A node has T children, where T is the number calculated in line 7 of the algorithm. Each node is labeled with subset of indices marking out the submatrix $J_{S\times S}$ it is given.

Now consider exploring the tree in the following breadth-first manner. We will define a list B_t which will contain the vertices at the boundary of explored territory and a filtration $\mathcal{F}_t$. Let $B_0 = (v_1)$ and $\mathcal{F}_0$ be the trivial σ-algebra. Given B_t and $\mathcal{F}_t$, if B_t is non-empty, define B_{t+1} and $\mathcal{F}_{t+1}$ as follows. Let v_{t+1} be the first vertex in the list B_t, and let B_{t+1} be defined from B_t by removing v_{t+1} from B_t and adding its children. Let S_{t+1} denote the set of indices associated with v_{t+1}, considered as a set-valued random variable, and $\mathcal{F}_{t+1} = \sigma(\mathcal{F}_t, S_{t+1})$. Let $M_t = |B_t|$. We have $M_1 = \left\lfloor C_2 \ln\left(\frac{n}{\varepsilon}\right)\frac{n}{s}\right\rfloor$, and wish to bound the first time τ such that $M_\tau = 0$. We redefine $M_{\tau+k} = -k$ (for sake of making M_t a supermartingale, as we will show below).

For $t \ge 2$, consider $M_t - M_{t-1}|\mathcal{F}_{t-1}$. Let v denote the parent of v_t, and suppose v is associated with the set R, with $|R| = m$. Then $|S_t| = s := \left\lceil \frac{c_1 m}{\left(\ln\left(\frac{2}{\varepsilon_{\mathrm{step}}}\right)+1\right)\ln\left(\frac{n}{\varepsilon}\right)\left\|J^{\otimes}_{R\times R}\right\|_F} \right\rceil$. (We choose $c_1 \le \frac{1}{2}c_3$ to ensure that we always have $s \le m$.) Let D_t be the number of new children added. If $\left\|J^{\otimes}_{S_t\times S_t}\right\|_F \le c_3$ or $s = 1$, then v_t is a leaf and $D_t = 0$. Now consider $\left\|J^{\otimes}_{S_t\times S_t}\right\|_F > c_3$. In the current call to the algorithm, $s' = \left\lceil \frac{c_1 s}{\left(\ln\left(\frac{2}{\varepsilon_{\mathrm{step}}}\right)+1\right)\ln\left(\frac{n}{\varepsilon}\right)\left\|J^{\otimes}_{S_t\times S_t}\right\|_F} \right\rceil$. Then $D_t \le C_2 \ln\left(\frac{n}{\varepsilon}\right)\frac{s}{s'} \le \frac{C_2\left(\ln\left(\frac{2}{\varepsilon_{\mathrm{step}}}\right)+1\right)\ln\left(\frac{n}{\varepsilon}\right)^2\left\|J^{\otimes}_{S_t\times S_t}\right\|_F}{c_1}$. In either case, $M_t - M_{t-1} = D_t - 1$. We have that

$$D_t \le \frac{C_2\left(\ln\left(\frac{2}{\varepsilon_{\mathrm{step}}}\right)+1\right)\ln\left(\frac{n}{\varepsilon}\right)^2\left\|J^{\otimes}_{S_t\times S_t}\right\|_F \mathbb{1}\left[\left\|J^{\otimes}_{S_t\times S_t}\right\|_F > c_3\right]}{c_1}. \tag{7}$$

Hence, by Cauchy-Schwarz and Chebyshev's inequality,

$$\begin{aligned}\mathbb{E}[D_t|\mathcal{F}_{t-1}] &\le \frac{C_2\left(\ln\left(\frac{2}{\varepsilon_{\text{step}}}\right)+1\right)\ln\left(\frac{n}{\varepsilon}\right)^2}{c_1}\mathbb{E}\left[\left\|J^{\otimes}_{S_t\times S_t}\right\|_F^2|\mathcal{F}_{t-1}\right]^{1/2}\cdot \\ &\qquad \mathbb{P}\left[\left\|J^{\otimes}_{S_t\times S_t}\right\|_F > \frac{c_3}{\ln\left(\frac{2}{\varepsilon_{\text{step}}}\right)+1}\middle|\mathcal{F}_{t-1}\right]^{1/2} \\ &\le \frac{C_2\left(\ln\left(\frac{2}{\varepsilon_{\text{step}}}\right)+1\right)\ln\left(\frac{n}{\varepsilon}\right)^2}{c_1}\mathbb{E}\left[\left\|J^{\otimes}_{S_t\times S_t}\right\|_F^2|\mathcal{F}_{t-1}\right]^{1/2}\cdot\frac{\mathbb{E}\left[\left\|J^{\otimes}_{S_t\times S_t}\right\|_F^2|\mathcal{F}_{t-1}\right]^{1/2}}{c_3/\left(\ln\left(\frac{2}{\varepsilon_{\text{step}}}\right)+1\right)} \\ &= \frac{C_2\left(\ln\left(\frac{2}{\varepsilon_{\text{step}}}\right)+1\right)^2\ln\left(\frac{n}{\varepsilon}\right)^2}{c_1c_3}\mathbb{E}\left[\left\|J^{\otimes}_{S_t\times S_t}\right\|_F^2|\mathcal{F}_{t-1}\right]. \end{aligned} \tag{8}$$

Now because S_t is uniformly chosen at random from subsets of R of size s,

$$\begin{aligned}\mathbb{E}\left[\left\|J^{\otimes}_{S_t\times S_t}\right\|_F^2|\mathcal{F}_{t-1}\right] &= \mathbb{E}_{S\sim\mathsf{Uniform}\binom{R}{s}}\left[\left\|J^{\otimes}_{S\times S}\right\|_F^2\right] \\ &= \sum_{i,j\in R, i\neq j}\left(\frac{s}{m}\right)^2 J_{ij}^2 = \left(\frac{s}{m}\right)^2\left\|J^{\otimes}_{R\times R}\right\|_F^2 \le \frac{4c_1^2}{\left(\ln\left(\frac{2}{\varepsilon_{\text{step}}}\right)+1\right)^2\ln\left(\frac{n}{\varepsilon}\right)^2}, \end{aligned} \tag{9}$$

where we use the fact that $J^{\otimes}_{ii}=0$, all off-diagonal entries have probability $\left(\frac{s}{m}\right)^2$ of being included in S_t, and $s>1$. Combining (8) and (9) gives

$$\mathbb{E}[D_t|\mathcal{F}_{t-1}] \le \frac{4C_2c_1}{c_3}.$$

Choosing c_1 small enough (depending on C_2, c_3), we can ensure that $\mathbb{E}[M_t - M_{t-1}|\mathcal{F}_{t-1}] = \mathbb{E}[D_t - 1|\mathcal{F}_{t-1}] \le -\frac{1}{2}$, so that $M_t + \frac{t}{2}$ is a supermartingale for $t \ge 1$. By Doob's decomposition we can write $M_t = A_t + M'_t$ where $A_{t+1} \le A_1 - \frac{t}{2}$ is a predictable decreasing sequence and M'_t is a martingale.

We now bound the variance. Using (7),

$$\begin{aligned}\mathbb{E}[(M'_t - M'_{t-1})^2|\mathcal{F}_{t-1}] &\le \mathbb{E}[D_t^2|\mathcal{F}_{t-1}] \\ &\le \frac{C_2^2\left(\ln\left(\frac{2}{\varepsilon_{\text{step}}}\right)+1\right)^2\ln\left(\frac{n}{\varepsilon}\right)^4}{c_1^2}\mathbb{E}\left[\left\|J^{\otimes}_{S_t\times S_t}\right\|_F^2|\mathcal{F}_{t-1}\right] \\ &\le 4C_2^2\ln\left(\frac{n}{\varepsilon}\right)^2, \end{aligned}$$

where we use the bound (9). Finally, $|M'_{t+1} - M'_t| \le \frac{C_2\left(\ln\left(\frac{2}{\varepsilon_{\text{step}}}\right)+1\right)\ln\left(\frac{n}{\varepsilon}\right)^2}{c_1}\left\|J^{\otimes}\right\|_F$ with probability 1. Let T_0 be the T computed in line 7 in the first step of the algorithm. By Bernstein's inequality for martingales (Lemma 4.11), for $t \ge C\ln^4\left(\frac{n}{\varepsilon}\right)\max\left\{\left\|J^{\otimes}\right\|_F, 1\right\} \ge T_0$ for an appropriate constant C (depending on C_2, c_1, C_4),

$$\mathbb{P}(M_{t+1} > 0) = \mathbb{P}\left((M_{t+1} - M_1) > -T_0\right) \le \mathbb{P}\left(M'_{t+1} - M'_1 > \frac{t}{2} - T_0\right) \le \frac{\varepsilon}{2}.$$

This shows that with probability $\ge 1 - \frac{\varepsilon}{4}$, there are at most $t_{\max} = C\ln^4\left(\frac{n}{\varepsilon}\right)\max\{\|J\|_F, 1\}$ nodes.

Finally, we note that in the call to ApproxRejectionSampler, the parameter needed to obtain error $\varepsilon_{\text{step}}$ is

$$c = \left(\frac{2}{\varepsilon_{\text{step}}}\right)^{\frac{\delta}{c_3-\delta}} = \left(\frac{2}{\varepsilon_{\text{step}}}\right)^{\frac{1}{\log(2/\varepsilon_{\text{step}})}} = e$$

and the acceptance probability is $\geq \frac{1}{2c} = \frac{1}{2e}$.

The number of tries until acceptance is a geometric random variable, which is subexponential, so standard concentration bounds show that the total number of tries is at most $O(\ln\left(\frac{1}{\varepsilon}\right))$ times the number of calls, with probability $\geq 1 - \frac{\varepsilon}{2}$. Putting everything together, we obtain $O(\max\{\|J\|_F, 1\} \operatorname{poly}\log\left(\frac{n}{\varepsilon}\right))$ running time with probability $\geq 1 - \varepsilon$.

Output is close in TV distance. Let A be a large constant to be determined.

Now consider coupling $y = y^{(0)}$ with a sequence of random variables $y^{(1)}, \ldots$, defined inductively as follows. Start with all vertices of the tree of recursive calls unmarked. Now given $y^{(t)}$, choose a node (in a fixed manner) all of whose children are marked, and mark it; then replace the output of that call to ParallelIsingSampler by a sample from the true distribution. We now choose constants so that $\mathcal{D}_{\text{TV}}(\mathcal{D}(y^{(t)}), \mathcal{D}(y^{(t+1)})) \leq \frac{\varepsilon}{n^A}$. There are two kinds of replacements to consider, a leaf node and a non-leaf node.

A leaf node corresponds to a call to ApproxRejectionSampler. If c_3 is small enough and $C_4 = A$, then by Lemma 4.9 and 4.4, the output of ApproxRejectionSampler is within $\frac{\varepsilon}{n^A}$ of the $\mu_{J_{R\times R},h}$.

A non-leaf node corresponds to T recursive calls to ParallelIsingSampler. Here we must appeal to mixing for the Ising model. By Theorem 4.1, approximate tensorization of entropy holds with constant $\frac{1}{c}$. Hence by Theorem 3.6, there is a constant C_0' such that if $C_0 = C_0' c$, then for any s, $t \cdot \frac{n}{s}$ steps of s-Glauber dynamics results in a distribution ν_t satisfying

$$\mathcal{D}_{\text{TV}}(\nu_t \| \mu_{J,h}) \leq \sqrt{\frac{1}{2}\mathcal{D}_{\text{KL}}(\nu_t \| \mu_{J,h})} \leq \sqrt{\frac{1}{2}\mathcal{D}_{\text{KL}}(\nu_0 \| \mu_{J,h}) e^{-C_0 t}}.$$

With the product initialization, we have by Lemma 4.10(2) (applied to the whole matrix) that $\mathcal{D}_{\text{KL}}(\nu_0 \| \mu_{J,h}) \leq \|J\| n \leq 2n$. Hence there exists a constant C_2' such that if $C_2 = C_2' A/c$, then with $T = C_2 \ln\left(\frac{n}{\varepsilon}\right) \frac{n}{s}$ steps, $\mathcal{D}_{\text{TV}}(\nu_T \| \mu_{J,h}) \leq \frac{\varepsilon}{n^A}$. By Lemma 4.10(1), for the Ising model $\mu_{J_{R\times R},h}$ the conditional distribution of X_S given $X_{R\setminus S} = y_{R\setminus S}$ is exactly the Ising model $\mu_{J_{S\times S}, J_{S\times R\setminus S} y_{R\setminus S} + h_S}$. Given that the conditional distributions are sampled exactly, then the only error is that from not having fully mixed, which we set to be $\frac{\varepsilon}{n^A}$.

This chain of coupled random variables establishes $\mathcal{D}_{\text{TV}}(\mathcal{D}(y), \mathcal{D}(y^{(t)})) \leq \frac{t\varepsilon}{n^A}$. Moreover, for $t > t_{\max}$, $\mathcal{D}_{\text{TV}}(\mathcal{D}(y^{(t)}), \mu_{J,h}) \leq \frac{\varepsilon}{2}$ by our high-probability bound, as the root node in $y^{(t)}$ will have been replaced with a perfect sample with probability $\geq 1 - \frac{\varepsilon}{2}$. It remains to note that $t_{\max} = C \ln^4\left(\frac{n}{\varepsilon}\right) \max\{\|J\|_F, 1\}$ with C depending polynomially on A. Hence we can choose A such that $\frac{t_{\max}\varepsilon}{n^A} \leq \frac{\varepsilon}{2}$, and this finishes the proof. ◀

References

1. Vedat Levi Alev and Lap Chi Lau. Improved analysis of higher order random walks and applications. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1198–1211, 2020. doi:10.1145/3357713.3384317.
2. Yeganeh Alimohammadi, Nima Anari, Kirankumar Shiragur, and Thuy-Duong Vuong. Fractionally log-concave and sector-stable polynomials: counting planar matchings and more. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 433–446, 2021. doi:10.1145/3406325.3451123.
3. Jason M Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. *arXiv preprint*, 2023. arXiv:2302.10249.

4 Nima Anari, Callum Burgess, Kevin Tian, and Thuy-Duong Vuong. Quadratic speedups in parallel sampling from determinantal distributions. In *Proceedings of the 35th ACM Symposium on Parallelism in Algorithms and Architectures*, pages 367–377, 2023. doi:10.1145/3558481.3591104.

5 Nima Anari, Shayan Oveis Gharan, and Cynthia Vinzant. Log-concave polynomials, entropy, and a deterministic approximation algorithm for counting bases of matroids. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 35–46. IEEE, 2018. doi:10.1109/FOCS.2018.00013.

6 Nima Anari, Nathan Hu, Amin Saberi, and Aaron Schild. Sampling arborescences in parallel. *arXiv preprint*, 2020. arXiv:2012.09502.

7 Nima Anari, Yizhi Huang, Tianyu Liu, Thuy-Duong Vuong, Brian Xu, and Katherine Yu. Parallel discrete sampling via continuous walks. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 103–116, 2023. doi:10.1145/3564246.3585207.

8 Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Entropic independence i: Modified log-sobolev inequalities for fractionally log-concave distributions and high-temperature ising models. *arXiv preprint*, 2021. arXiv:2106.04105.

9 Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Universality of spectral independence with applications to fast mixing in spin glasses. *arXiv preprint*, 2023. arXiv:2307.10466.

10 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.

11 Dina Bitton, David J DeWitt, David K Hsaio, and Jaishankar Menon. A taxonomy of parallel sorting. *ACM Computing Surveys (CSUR)*, 16(3):287–318, 1984. doi:10.1145/2514.2516.

12 Antonio Blanca, Pietro Caputo, Zongchen Chen, Daniel Parisi, Daniel Štefankovič, and Eric Vigoda. On mixing of markov chains: Coupling, spectral independence, and entropy factorization. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3670–3692. SIAM, 2022. doi:10.1137/1.9781611977073.145.

13 Alexandre Bristiel and Pietro Caputo. Entropy inequalities for random walks and permutations. *arXiv preprint*, 2021. arXiv:2109.06009.

14 Djalil Chafaï. Entropies, convexity, and functional inequalities, on ϕ-entropies and ϕ-sobolev inequalities. *Journal of Mathematics of Kyoto University*, 44(2):325–363, 2004.

15 George HG Chen and R Tyrrell Rockafellar. Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997. doi:10.1137/S1052623495290179.

16 Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pages 2984–3014. PMLR, 2022. URL: https://proceedings.mlr.press/v178/chen22c.html.

17 Yuansi Chen. An almost constant lower bound of the isoperimetric coefficient in the kls conjecture. *Geometric and Functional Analysis*, 31:34–61, 2021.

18 Yuansi Chen and Ronen Eldan. Localization schemes: A framework for proving mixing bounds for markov chains. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 110–122. IEEE, 2022.

19 Zongchen Chen, Kuikui Liu, and Eric Vigoda. Optimal mixing of glauber dynamics: Entropy factorization via high-dimensional expansion. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1537–1550, 2021. doi:10.1145/3406325.3451035.

20 Mary Cryan, Heng Guo, and Giorgos Mousa. Modified log-sobolev inequalities for strongly log-concave distributions. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1358–1370. IEEE, 2019. doi:10.1109/FOCS.2019.00083.

21 Persi Diaconis and Mehrdad Shahshahani. Time to reach stationarity in the bernoulli–laplace diffusion model. *SIAM Journal on Mathematical Analysis*, 18(1):208–218, 1987.

22 Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual*

Symposium on Foundations of Computer Science (FOCS), pages 323–334. IEEE, 2022. doi:10.1109/FOCS54457.2022.00038.

23 Ronen Eldan. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geometric and Functional Analysis*, 23(2):532–569, 2013.

24 Ronen Eldan, Frederic Koehler, and Ofer Zeitouni. A spectral condition for spectral gap: fast mixing in high-temperature ising models. *Probability theory and related fields*, 182(3-4):1035–1051, 2022.

25 Jiaojiao Fan, Bo Yuan, and Yongxin Chen. Improved dimension dependence of a proximal algorithm for sampling. *arXiv preprint*, 2023. arXiv:2302.10081.

26 Weiming Feng, Thomas P Hayes, and Yitong Yin. Distributed metropolis sampler with optimal parallelism. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2121–2140. SIAM, 2021. doi:10.1137/1.9781611976465.127.

27 Yuval Filmus, Ryan O'Donnell, and Xinyu Wu. Log-sobolev inequality for the multislice, with applications. *Electronic Journal of Probability*, 27:1–30, 2022.

28 David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

29 Will Grathwohl, Kevin Swersky, Milad Hashemi, David Duvenaud, and Chris Maddison. Oops i took a gradient: Scalable sampling for discrete distributions. In *International Conference on Machine Learning*, pages 3831–3841. PMLR, 2021. URL: http://proceedings.mlr.press/v139/grathwohl21a.html.

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

31 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. doi:10.1145/3313276.3316299.

32 Frederic Koehler, Holden Lee, and Andrej Risteski. Sampling approximately low-rank ising models: Mcmc meets variational methods. In *Conference on Learning Theory*, pages 4945–4988. PMLR, 2022. URL: https://proceedings.mlr.press/v178/koehler22a.html.

33 Tzong-Yow Lee and Horng-Tzer Yau. Logarithmic sobolev inequality for some models of random walks. *The Annals of Probability*, 26(4):1855–1873, 1998.

34 Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted gaussian oracle. In *Conference on Learning Theory*, pages 2993–3050. PMLR, 2021. URL: http://proceedings.mlr.press/v134/lee21a.html.

35 Hongyang Liu and Yitong Yin. Simple parallel algorithms for single-site dynamics. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1431–1444, 2022. doi:10.1145/3519935.3519999.

36 Nicolas Loizou, Hugo Berard, Gauthier Gidel, Ioannis Mitliagkas, and Simon Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Advances in Neural Information Processing Systems*, 34:19095–19108, 2021. URL: https://proceedings.neurips.cc/paper/2021/hash/9f96f36b7aae3b1ff847c26ac94c604e-Abstract.html.

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

38 Andrea Montanari. Sampling, diffusions, and stochastic localization. *arXiv preprint*, 2023. arXiv:2305.10690.

39 Andrea Montanari and Yuchen Wu. Posterior sampling from the spiked models via diffusion processes. *arXiv preprint*, 2023. arXiv:2304.11449.

40 Ravi Montenegro and Prasad Tetali. Mathematical aspects of mixing times in markov chains. *Foundations and Trends® in Theoretical Computer Science*, 1(3):237–354, 2006. doi:10.1561/0400000003.

41 Izhar Oppenheim. Local spectral expansion approach to high dimensional expanders part I: Descent of spectral gaps. *Discrete & Computational Geometry*, 59(2):293–330, 2018. doi:10.1007/s00454-017-9948-x.

42 Benjamin Rhodes and Michael Gutmann. Enhanced gradient-based mcmc in discrete spaces. *arXiv preprint*, 2022. arXiv:2208.00040.

43 Justin Salez. A sharp log-Sobolev inequality for the multislice. *Ann. H. Lebesgue*, 4:1143–1161, 2021. doi:10.5802/ahl.99.

44 Holger Sambale and Arthur Sinulis. Modified log-sobolev inequalities and two-level concentration. *arXiv preprint*, 2019. arXiv:1905.06137.

45 Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. *Advances in Neural Information Processing Systems*, 32, 2019.

46 Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. URL: http://proceedings.mlr.press/v37/sohl-dickstein15.html.

47 Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

48 Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint*, 2020. arXiv:2011.13456.

49 Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

50 Ruqi Zhang, Xingchao Liu, and Qiang Liu. A langevin-like sampler for discrete distributions. In *International Conference on Machine Learning*, pages 26375–26396. PMLR, 2022. URL: https://proceedings.mlr.press/v162/zhang22t.html.