Universality of Spectral Independence with Applications to Fast Mixing in Spin Glasses

Nima Anari[*] Vishesh Jain[†] Frederic Koehler[‡] Huy Tuan Pham[§]
Thuy-Duong Vuong[¶]

Abstract

We study Glauber dynamics for sampling from discrete distributions μ on the hypercube $\{\pm 1\}^n$. Recently, techniques based on spectral independence have successfully yielded optimal $O(n)$ relaxation times for a host of different distributions μ. We show that spectral independence is universal: a relaxation time of $O(n)$ implies spectral independence.

We then study a notion of tractability for μ, defined in terms of smoothness of the multilinear extension of its Hamiltonian – $\log \mu$ – over $[-1, +1]^n$. We show that Glauber dynamics has relaxation time $O(n)$ for such μ, and using the universality of spectral independence, we conclude that these distributions are also fractionally log-concave and consequently satisfy modified log-Sobolev inequalities. We sharpen our estimates and obtain approximate tensorization of entropy and the optimal $\widetilde{O}(n)$ mixing time for random Hamiltonians, i.e. the classically studied mixed p-spin model at sufficiently high temperature. These results have significant downstream consequences for concentration of measure, statistical testing, and learning.

1 Introduction

In this paper, we study probability measures $\mu(\sigma)$ on the hypercube $\{\pm 1\}^n$ and the standard Markov chain for sampling from such distributions known as the Glauber dynamics. Any such distribution $\mu(x)$ with full support can be written in the form

$$\mu(\sigma) = \frac{1}{Z} \exp(H(\sigma))$$

for some function $H : \{\pm 1\}^n \to \mathbb{R}$, unique up to an additive constant, and corresponding normalizing constant Z. We say that such a distribution μ is the *Gibbs measure for Hamiltonian H*.

Directly sampling from such a distribution is not easy, because evaluating the normalizing constant Z ("partition function") can be computationally difficult. For this reason, samples are typically generated using a Markov chain approach. In particular, the *Glauber dynamics* or Gibbs sampler is a natural Markov chain with stationary distribution μ that in each step resamples a uniformly random coordinate i conditioned on the remaining coordinates $\sim i \coloneqq [n] - i$:

```
Let $\sigma^{(0)} \in \{\pm 1\}^n$ be the initial state.
for $t = 1, \ldots$ do
    Select $i$ uniformly at random from $[n] = \{1, \ldots, n\}$.
    Let $\sigma_j^{(t)} = \sigma_j^{(t-1)}$ for all $j \neq i$, and sample $\sigma_i^{(t)}$ from the conditional law $\mathbb{P}_\mu[\sigma = \cdot \mid \sigma_{\sim i} = \sigma_{\sim i}^{(t-1)}]$.
```

Each step of this chain is easy to implement given access to the Hamiltonian H, and in particular does not require knowledge of the normalizing constant Z. What is less obvious is how long the chain should be run. Understanding the *mixing time*, i.e. the number of steps which must be executed for the chain to approximately reach stationarity, for particular Hamiltonians H is a very important and mathematically difficult task which has

[*] Stanford University, anari@cs.stanford.edu
[†] University of Illinois Chicago, visheshj@uic.edu
[‡] Stanford University, fkoehler@stanford.edu
[§] Stanford University, huypham@stanford.edu
[¶] Stanford University, tdvuong@stanford.edu

been intensely studied for multiple decades (see e.g. [AF95; Mar99; GZ03; LP17] for background). It suffices to say that there are many approaches to prove rapid mixing with different strengths and weaknesses.

High-dimensional expansion and spectral independence. Recently, a fruitful approach to analyzing the Glauber dynamics has emerged based on connections to high-dimensional expansion and the geometry of polynomials. These ideas have been used to establish optimal mixing time bounds in many settings (see e.g. [Ana+19; CGM19; ALO20; Bla+21; CLV20; CLV21; Ana+21a; Ana+21b; Che+21; Ali+21; ALG22; CLV22; Che+22]).

One of the key concepts in this approach is *spectral independence*. A Gibbs measure μ is *η-spectrally independent* if $\lambda_{\max}(\Psi) \le \eta$ where $\Psi \in \mathbb{R}^{2n\times 2n}$ is the correlation matrix indexed by pairs in $[n] \times \{\pm 1\}$ with entries

$$\Psi_{(i,\tau_i),(j,\tau_j)} = \mathbb{P}_\mu[\sigma_i = \tau_i \mid \sigma_j = \tau_j] - \mathbb{P}_\mu[\sigma_i = \tau_i].$$

Proving that the Gibbs measure (and all of its conditionings) is $O(1)$-spectrally independent is the key first step for applying the high-dimensional expansion approach to mixing time (e.g. to apply Theorem 1.3 of [ALO20], which implies polynomial mixing time bounds).

Understanding this concept and finding ways to prove spectral independence is therefore very useful. Many authors have asked and studied how spectral independence relates to existing notions in the sampling literature. For example, Dobrushin's uniqueness criterion [Dob68] is a classical concept which considers the properties of a related-looking influence matrix. Recently, Liu [Liu21] and Blanca, Caputo, Chen, Parisi, Štefankovič, and Vigoda [Bla+21] proved that Dobrushin's criterion (and more generally, the existence of a contractive coupling) implies $O(1)$-spectral independence. A similar implication holds under other widely used criteria such as correlation decay [CLV20] or zero-freeness [Ali+21; CLV22].

1.1 Our Results

Universality of spectral independence. We prove a new result which directly connects spectral independence with the classical notion of spectral gap. The *spectral gap* or Poincaré constant of a Markov chain is the gap between the two largest eigenvalues of its transition matrix P. In other words, we say that P has *spectral gap* λ if $\lambda_2(P) = 1 - \lambda$. This is a key concept in the study of Markov chains — informally, the spectral gap controls the speed of mixing from a warm start. For this reason, the inverse spectral gap $1/\lambda$ is known as the *relaxation time* of the (lazy version of the) Markov chain [LP17].

We prove that *spectral independence is a relaxation of the spectral gap*, in other words, that $O(1)$-spectral independence necessarily holds if the Glauber dynamics has a large spectral gap.

Theorem 1.1. *If the Glauber dynamics of a distribution μ on the hypercube $\{\pm 1\}^n$ has spectral gap $\frac{1}{Cn}$, then μ is C-spectrally independent.*

We actually prove a stronger fact (Theorem 3.1), which is the natural generalization of this result to down-up walks. In particular, this means that the same relation holds for the Glauber dynamics with spins valued in arbitrary alphabets, not just binary spins.

Conceptually, this gives a simple explanation for the ubiquity of spectral independence: it is a *necessary* condition for the Glauber dynamics to have $O(n)$ relaxation time. It immediately implies that spectral independence holds in a large number of settings where mixing time analysis was performed using other methods (e.g., via stochastic localization [EKZ21; Ana+21a; CE22] or curvature arguments [CMT15; Erb+17]). This in turn has further nontrivial consequences: if the relaxation time is $O(n)$ under all external fields, we get fractional log-concavity of the generating polynomial, which in turn implies subadditivity of entropy and Brascamp-Lieb type inequalities [see Ana+21a; Ali+21; Bar+11; Bla+21].

Rapid mixing from smoothness of the multilinear extension. Building on the universality of spectral independence, we are able to prove new results concerning the mixing of the Glauber dynamics. We would like to understand what conditions on H naturally lead to rapid mixing of the Glauber dynamics. For inspiration, we know that for *continuous distributions* on $\mathbb{R}^n$ and the corresponding (continuous time) *Langevin dynamics*, strong log-concavity of the distribution implies rapid mixing via Bakry-Emery theory [BGL+14]. For a distribution with smooth density p, this just means that $\nabla^2 \log p \preceq -\epsilon I$ for some $\epsilon > 0$.

We identify a natural analogue of this fact on the discrete hypercube. First (as in e.g. [EG18]), we identify

H with its *multilinear extension* $H : \mathbb{R}^n \to \mathbb{R}$ defined by

$$H(x) = \sum_{S \subset [n]} \hat{H}(S) \prod_{i \in S} x_i$$

where $\hat{H}(S) := \frac{1}{2^n} \sum_{\sigma \in \{\pm 1\}^n} H(\sigma) \prod_{i \in S} \sigma_i$ is the Fourier transform of H viewed as a function on the hypercube [ODo14]. We prove that as long as $\nabla^2 H$ is spectrally small (i.e. H is sufficiently smooth in the usual sense), Glauber dynamics mixes rapidly:

Theorem 1.2. (Combined Theorem 4.1 and Theorem 4.2) *There exist absolute constants $A, B > 0$ for which the following holds. Suppose that μ is a probability measure with full support on the hypercube $\{\pm 1\}^n$, so $\mu(x) \propto \exp(H(x))$ for some function $H : \{\pm 1\}^n \to \mathbb{R}$. Suppose furthermore that*

$$\beta := \max_{\sigma \in \{\pm 1\}^n} \|\nabla^2 H(\sigma)\|_{\text{op}} \le A.$$

Then we have that:

1. *The spectral gap of the Glauber dynamics on μ is at least $\frac{1}{(1+B\beta)n}$.*
2. *μ^{hom}, the homogenization of μ (see Definition 2.14), is $\frac{1}{1+B\beta}$-fractionally log-concave.*
3. *μ satisfies approximate tensorization of entropy with constant $C = O(n^{B\beta})$.*
4. *The Glauber dynamics on μ satisfies the Modified Log-Sobolev Inequality (MLSI) with constant $\rho = \Omega(1/n^{1+B\beta})$.*
5. *The Glauber dynamics on μ satisfies*

$$\tau_{\text{mix}}(\epsilon) = O\left(n^{1+B\beta}[\log\log(1/\min_\sigma \mu(\sigma)) + \log(1/\epsilon)]\right).$$

We formally define the concepts of approximate tensorization, fractional log-concavity, etc. appearing here in the preliminaries (Section 2). Note that the first conclusion by itself only implies $\widetilde{O}(n^2)$ mixing time, whereas the final conclusion gives a much better bound (the constant A under which we can prove the spectral gap inequality is relatively small, and in particular $AB < 1$).

Remark 1.1. (Tightness) *It is not true that μ^{hom} (see Definition 2.14) is a log-concave distribution (in the sense of [Ana+19]) under the assumptions of this theorem — the relaxation to fractional log-concavity is required. Relatedly, the functional dependence of the Poincaré constant on β in conclusion (1) is optimal. By this we mean that β cannot be replaced by any function which is $o(\beta)$ as $\beta \to 0$, e.g. by β^2. This can be seen by examining one of various examples where the spectral gap is known exactly — in particular the Ising model on a cycle (Theorem 15.5 of [LP17]). Going back to the first point, if μ^{hom} were log-concave then this would imply a spectral gap of at least $1/n$ [Ana+19], which is not true. Finally, the result cannot be true for any value of $A > 1$, because then it would include the supercritical Curie-Weiss model for which mixing takes exponential time [LP17].*

Remark 1.2. (Need for a two-sided assumption) *Comparing to the continuous setting, we might guess that this result would be true under only a one-sided bound on $\nabla^2 H$, allowing for arbitrary large negative eigenvalues (somewhat in the spirit of Equation (4) of [ES22]). For example, if we define a Gibbs measure with respect to the standard Gaussian distribution with Radon-Nikodym derivative proportional to $\exp(H(x))$, then we only need $\nabla^2 H \prec (1-\epsilon)I$ to apply the Bakry-Emery criterion [BGL+14]. However, we know from [KLR22] that already in the case of quadratic/Ising interactions, large negative eigenvalues of H can make the spectral gap large and even make sampling computationally hard.*

Remark 1.3. *Bounds on β can be directly obtained from the tensor injective norm of the coefficients of H. We can apply existing results to control the coefficient tensor for natural random examples, like the mixed p-spin model discussed below, or when $H(x)$ counts the number of satisfied constraints in random k-XOR sat (in which case, the tensor will generally be sparse), see e.g. [ZZ21].*

Optimal mixing in the mixed p-spin model. One of the fundamental models in statistical physics corresponds to the case where the Hamiltonian H has random coefficients. In other words,

$$H(\sigma) = \sum_{p=2}^{\infty} \frac{\beta_p}{n^{(p-1)/2}} \sum_{1\leq i_1,\dots,\leq i_p\leq n} g_{i_1\cdots i_p}\sigma_{i_1}\cdots\sigma_{i_p} + \sum_i h_i\sigma_i$$

where the sum ranges over distinct $i_1,\dots,i_p$, each of the coefficients $g_{i_1\cdots i_p}$ is i.i.d. standard Gaussian, and we allow the external fields h_i to be arbitrary (in this work, they are allowed to depend on g). The corresponding (random) measure $\mu \propto \exp(H)$ on $\{\pm 1\}^n$ is known as the *mixed p-spin model* and it has been deeply studied in spin glass theory (see e.g. [Tal10; Pan13] for rigorous results).

In recent work (see below for more discussion of related work), Adhikari, Brennecke, Xu, and Yau [Adh+22] established an $\widetilde{O}(n^2)$ mixing time for this model at sufficiently high temperature (small β). Theorem 1.2 implies an improved mixing time bound of $\widetilde{O}(n^{1+O(\beta)})$ for this model, where $\beta := \sum_{p\geq 2}\sqrt{p^3\log p}\cdot\beta_p$. Our next result improves this to the optimal mixing time bound by proving that approximate tensorization of entropy (and hence, a modified Log-Sobolev inequality) holds with a dimension-free constant.

Theorem 1.3. *There exists an absolute constant $A > 0$ for which the following holds. Suppose $\beta_0 := \sum_{p\geq 2}\sqrt{p^3\log p}\cdot\beta_p \leq A$. Let $\beta := \sum_{p\geq 2}\sqrt{2^p p^3\log p}\cdot\beta_p$. Then, for the mixed p-spin model μ, with probability $\geq 1-\exp(-\Theta(N))$ over the randomness of the Gaussian interaction terms g,*

1. *μ (equivalently μ^{hom}) satisfies approximate tensorization of entropy with constant $C = O_\beta(1)$.*
2. *The discrete-time Glauber dynamics on μ satisfies the Modified Log-Sobolev Inequality (MLSI) with constant $\rho = \Omega_\beta(1/n)$.*
3. *The discrete-time Glauber dynamics on μ satisfies*

$$\tau_{\text{mix}}(\epsilon) = O_\beta\Big(n(\log\log(1/\min_\sigma \mu(\sigma)) + \log(1/\epsilon)\Big).$$

Downstream applications. The results we establish have a number of interesting downstream applications. We discuss a few in particular:

- **Consequences of the MLSI.** We obtain bounds on the Modified Log-Sobolev constant which have a number of useful consequences besides mixing time analysis. First of all, the MLSI gives more precise control on the dynamics of mixing in the form of *reverse hypercontractivity* — see [MOS13; Gro14]. Also, the MLSI implies subgaussian concentration of Lipschitz functions and transport-entropy inequalities [Van14; Gro14]. For example, we have established for the first time Lipschitz subguassian concentration for the high temperature mixed p-spin model.
- **Spectral independence from coupling.** Existence of a contractive coupling was previously known to imply spectral independence [Liu21; Bla+21]. Since contractive couplings imply a spectral gap, we recover this result from the universality of spectral independence (Theorem 3.1). The resulting proof is much shorter and has some notable quantitative advantages — we discuss this more in Section 5.1.
- **Learning graphical models.** It was recently observed that there is a useful connection between approximate tensorization of entropy and classical algorithms for learning distributions from data [KHR22]. Combining this connection with our results, we are able to dramatically improve the state of the art results for learning some types of graphical models from samples (Theorem 5.3). For example, we are able to learn the high-temperature SK model in total variation distance from $n^{3+O(\beta)}$ samples in polynomial time, which is close to the best information-theoretic guarantee known [DMR20]. In comparison, the best previous algorithmic results [KM17; Vuf+16; WSD19] could guarantee sample and time complexity of only $e^{O(\beta\sqrt{n})}$.
- **Identity testing.** Blanca, Chen, Štefankovič, and Vigoda [Bla+22] recently proved implications of approximate tensorization of entropy for identity testing of high-dimensional distributions in the *coordinate oracle* and *subcube oracle* query models. Combining our new results with their framework gives improved sample complexities for solving many testing problems (e.g. Corollary 5.1) – see Section 5.3 for more details.

1.2 Techniques

Universality of spectral independence. Interestingly, the proof of this result is very short once the correct definitions are in place, so we refer the reader to Section 3.

Rapid mixing for smooth Hamiltonians. The proof of this result combines three key ingredients: (1) universality of spectral independence, (2) a recursive spectral gap estimate of Adhikari, Brennecke, Xu, and Yau [Adh+22], and (3) a large body of existing tools from the high-dimensional expanders framework.

First, we want to prove that under the smoothness condition, the Glauber dynamics has an $\Omega(1/n)$ spectral gap. In the work [Adh+22], the authors developed a recursive method to prove spectral gap bounds for high temperature systems, *provided* that the spectral gap of systems with n spins is not much worse than that with $n-1$ spins. To make their argument work, they needed to prove a corresponding "continuity estimate", but their argument used facts specific to the mixed p-spin model (random H). Our key insight is that combining the universality of spectral independence with Oppenheim's trickle-down theorem [Opp18] and the local-to-global argument [KO18; AL20] proves that such continuity estimates *automatically* hold in a very general setting (Theorem 3.2). This lets us prove the spectral gap inequality for all models satisfying the smoothness condition.

By itself, the spectral gap bound only implies $\widetilde{O}(n^2)$ time mixing. To improve the mixing time bound we appeal to the universality of spectral independence again, which lets us derive fractional log-concavity of the generating polynomial [Ali+21]. This implies entropic independence which in turn implies the approximate tensorization of entropy estimate via the result of [Ana+21a]. The MLSI and mixing time bound are immediate consequences (see Section 2).

Improved bound for random interactions (mixed p-spin model). First, we discuss the special case of the Sherrington-Kirkpatrick (SK) model (pure 2-spin model). For sufficiently high temperature i.e. $1/\beta \gg 1/\epsilon$, the interaction matrix has operator norm at most $1+\epsilon$, thus the discussion above allows us to directly lower bound the modified log Sobolev constant by $1/n^{1+\epsilon}$. However, this is still a factor of n^{ϵ} away from the conjectured modified log Sobolev constant of $1/n$, when the interaction terms are i.i.d. Gaussians. How can we close the gap?

In the case of the SK model, after pinning $N-k$ spins, the resulting Hamiltonian $H^{(k)}$ is a $k \times k$ matrix with i.i.d. Gaussian entries of variance $1/N$, thus the corresponding operator norm scales approximately like $\sqrt{k/N}$, so that the local to global argument gives

$$\text{MLSI constant} \geq \exp\left(-\sum_{k} \frac{(1+\|H^{(k)}\|_{\text{op}})}{k}\right) \approx \exp\left(-\log n + \sum_{k} \frac{1}{\sqrt{kN}}\right) = n^{-1}\exp(-O(1)),$$

which is the desired lower bound for the modified log-Sobolev constant.

Unfortunately, when higher degree terms are present, pinning might still result in a subsystem with essentially the same operator norm. To see this, consider a pure 3-spin model, i.e. the Hamiltonian only contains terms of degree 3:

$$H(\sigma) = \frac{\beta}{N}\sum_{ijk} g_{ijk}\sigma_i\sigma_j\sigma_k,$$

where g_{ijk} are standard i.i.d. Gaussians and $\beta > 0$ is a small constant. It is well known that $\|H\| = \Theta(1)$ with high probability. Now, for any $\sigma \in \{\pm 1\}^n$, observe that pinning $\sigma_{[n]\setminus\{1,2\}}$, results in a subsystem on two spins with interaction matrix $H^{(2)}$ given by $H^{(2)}_{1,2} = \frac{1}{N}\sum_{k\notin\{1,2\}} g_{12k}\sigma_k$, and note that with the choice of pinning $\sigma_k := \text{sign}(g_{12k})$,

$$H^{(2)}_{1,2} = \frac{1}{N}\sum_{k\notin\{1,2\}} |g_{12k}| = \Theta(1) \text{ w.h.p.}$$

To circumvent the existence of these "bad pinnings", we switch to an "average-case" version of the local to global argument [Ali+21; ALV22]. Roughly speaking, for each $i \in [N]$ and $\sigma \in \{\pm 1\}^{N\setminus\{i\}}$, we want to establish that when pinning a *random* subset of $N-k$ spins according to σ, the operator norm of the conditional subsystem is sufficiently small, e.g. it scales like $\sqrt{k/N}$ with high probability. We emphasize that while the average case argument allows us to consider subsystems resulting from pinning a random subset of spins according to σ, we must still consider *all* possible σ; in particular, if we are to use the union bound over σ, we need to avoid bad events with exponentially (in n) small probability, even though the quantities involved (norms of subsystems on k spins) have order governed by k. Moreover, unlike in [ALV22], there are choices of i and σ for which we cannot expect the "good event" of avoiding a bad link to hold with a very high probability, e.g. $1 - 1/N^{10}$. Indeed, going back to the pure 3-spin example above, for the choice of σ there and for $i = 1$, $H^{(2)}_{1,2} = \Theta(1)$, so that with probability $\Omega(k/N)$, a random pinning of $N - k$ spins includes vertex 2 among the free spins and thus results in a subsystem with operator norm $\Theta(1)$ (instead of $O(\sqrt{k/N})$, as we might hope for). To summarize, we have to address two challenges: (i) prove a "norm-decay" statement for random $k \times k$ sub-matrices which holds with probability exponentially small in n (as opposed to k), and (ii) overcome the very heavy-tailed nature of the norm of random $k \times k$ sub-matrices for use in the average case local-to-global argument [Ali+21; ALV22]. We note that while the expected norm-decay of random submatrices of a matrix has been extensively studied in the random matrix theory literature (see, e.g. [RV07]), our requirement that the probability bounds hold with exponentially small probability in n (as opposed to k) makes existing techniques ineffective and needs a completely different argument.

We conclude by briefly discussing the ideas needed to overcome these challenges. As later detailed in Section 4, the average case local-to-global argument requires us to show that for any σ, in a random ordering of pinnings according to σ, a quantity roughly of the form $\mathbb{E}[\exp(\sum_k \|\nabla^2 H_k\|_{\text{op}}/k)]$ is bounded above by an absolute constant; here, H_k is the induced Hamiltonian on the k unpinned spins and the randomness is induced by the random ordering. Using standard concentration estimates for the norm of random sub-Gaussian matrices, we can bound the contribution of terms of degree at least 3 in the induced Hamiltonian by $(k/N)^{\Omega(1)}$, so that the main challenge lies in bounding the contribution of quadratic term in the induced Hamiltonian. As explained above, this term can have operator norm 1 with very large probability (at least (k/N)) and hence, we need a significantly more careful analysis of the average-case local to global iteration than in [ALV22]. In particular, we show that except with exponentially small (in n) probability over the randomness of the Hamiltonian, for all σ, over the randomness of the random ordering of pinnings, $\|\nabla^2 H_k\|_{\text{op}}$ is bounded by a small constant c with probability 1 (this is straightforward, given existing results), and crucially, it is bounded by $(k/N)^{1/2-\alpha}$ with probability at least $1 - (k/N)^{\alpha}$, for some $\alpha \in [c, 1/2)$. Even then, we have to control $\exp(\sum_k \|\nabla^2 H_k\|_{\text{op}}/k)$, where the challenge is in the complex dependencies among H_k (arising from one random permutation of the pinning order). We achieve this by breaking terms based on the typical and tail behavior, and combine these heavy-tailed random variables in a careful way with an appropriately weighted Hölder's inequality.

1.3 Related Work Many works have studied techniques for analyzing discrete Markov chains which were inspired by Bakry-Emery theory and related notions, see e.g. [Erb+17; Oll09; ELL17; CDP09] and references within for a few examples. These are quite different in nature from our results, and in particular they have not been used to obtain results for systems with spin glass interactions like the Sherrington-Kirkpatrick and mixed p-spin models from statistical physics (see e.g. [Tal10] for more background on these models).

There have been many recent works studying sampling problems in spin glass models. For spherical spin glasses, Gheissari and Jagannath [GJ19] showed that the Bakry-Emery criterion can be applied to prove mixing of the Langevin dynamics at high temperature. For the Sherrington-Kirkpatrick (SK) model on the hypercube, after a line of recent works [BB19; EKZ21; Ana+21a] the optimal $O(n \log n)$ time mixing bound is known up to inverse temperature $\beta < 0.25$ (see also [CE22] for an alternative proof). All of these works only need that the interaction matrix has sufficiently small operator norm, which is exactly the same as the assumption of Theorem 1.2 specialized to the case of Ising models. Interestingly, although the proof in [Ana+21a] is partially based on the high-dimensional expansion approach, none of these works could prove that approximate tensorization of entropy holds (with a dimension-free constant). The present work finally proves that approximate tensorization of entropy holds for sufficiently small β.

Using a different algorithmic approach, the works [AMS22; Cel22] constructed a sampler which sample up to the conjectured sharp threshold $\beta < 1$ in the SK model, albeit in a weaker metric (sublinear Wasserstein). Finally, as discussed above, the recent work of Adhikari, Brennecke, Xu, and Yau [Adh+22] which we build upon proved spectral gap for the mixed p-spin model for sufficiently small β (i.e. at sufficiently high temperature).

The works [ES22; Ali+21] and references within studied a few related analogues of (semi, fractional, etc.) "log-concavity" on the discrete hypercube. In this work, we have explicitly showed one such notion (fractional log-concavity of the generating polynomial) follows from a condition on the Hessian of the log-likelihood in the spirit of strong log-concavity.

2 Preliminaries

In this paper, we use the perspective on sampling arising from the theory of generating polynomials and high-dimensional expansion. This means, for example, that the Glauber dynamics on $\{\pm 1\}^n$ is interpreted via homogenization as the $n \leftrightarrow n-1$ down-up walk on $\binom{[2n]}{n}$ (see Section 2.4 below). We discuss this and other important background in this preliminaries section.

2.1 Down-Up Walk, Links, and Trickle-Down

Definition 2.1. (Down operator) *For a ground set Ω, and $|\Omega| \geq k \geq \ell$, the down operator $D_{k\to\ell} \in \mathbb{R}^{\binom{\Omega}{k}\times\binom{\Omega}{\ell}}$ is defined to be*

$$D_{k\to\ell}(S,T) = \begin{cases} \frac{1}{\binom{k}{\ell}} & \textit{if } T \subseteq S, \\ 0 & \textit{otherwise.} \end{cases}$$

Note that $D_{k\to\ell}D_{\ell\to m} = D_{k\to m}$.

Definition 2.2. (Up operator) *For a ground set Ω, $|\Omega| \geq k \geq \ell$, and density $\mu : \binom{\Omega}{k} \to \mathbb{R}_{\geq 0}$, the up operator $U_{\ell\to k} \in \mathbb{R}^{\binom{\Omega}{\ell}\times\binom{\Omega}{k}}$ is defined to be*

$$U_{\ell\to k}(T,S) = \begin{cases} \frac{\mu(S)}{\sum_{S'\supseteq T}\mu(S')} & \textit{if } T \subseteq S, \\ 0 & \textit{otherwise.} \end{cases}$$

Definition 2.3. (Down-up walk) *For a ground set Ω, $|\Omega| \geq k \geq \ell$, and density $\mu : \binom{\Omega}{k} \to \mathbb{R}_{\geq 0}$, the $k \leftrightarrow \ell$ down-up walk is defined by the row-stochastic matrix $D_{k\to\ell}U_{\ell\to k}$. Similarly, the up-down walk is defined by $U_{\ell\to k}D_{k\to\ell}$.*

Proposition 2.1. ([see, e.g., KO18; AL20; ALO20]) *The operators $D_{k\to\ell}U_{\ell\to k}$ and $U_{\ell\to k}D_{k\to\ell}$ both define Markov chains that are time-reversible and have nonnegative eigenvalues. Moreover μ and $\mu D_{k\to\ell}$ are respectively their stationary distributions.*

Definition 2.4. (Link) *Given μ a distribution on $\binom{\Omega}{k}$, and a set $T \subset \Omega$ with $|T| \leq k$, we define μ_T, the* induced distribution on the link of T, *to be the probability distribution over $\Omega \setminus \binom{T}{k-|T|}$ given by the conditional law $\mu_T(S') = \mathbb{P}_{S\sim\mu}[S = S' \cup T \mid T \subset S]$.*

We will slightly abuse terminology and often refer to the induced distribution of μ on the link of T simply as the link of μ at T.

Oppenheim's Trickle-Down. Oppenheim's trickle-down theorem inductively bounds the high-dimensional expansion of simplicial complexes, i.e. the spectral gap of certain up-down walks.

Theorem 2.1. ([Opp18]) *Let μ be a distribution on $\binom{\Omega}{k}$ with $k \geq 3$, let P denote the corresponding $1 \leftrightarrow k$ up-down walk, and suppose that $\lambda_2(P) < 1$. For $i \in \Omega$, let μ_i denote the link of μ at i, and let P_i denote the corresponding $1 \leftrightarrow k-1$ up-down walk. Suppose that for all $i \in \Omega$, $\lambda_2(P_i) \leq \lambda$. Then*

$$\lambda_2(P) \leq \frac{(1-2/k)\lambda}{1-\lambda}.$$

Remark 2.1. (Lazy vs active walk) *With our definition, the $1 \leftrightarrow k$ up-down walk has a probability of $1/k$ of staying at the same vertex. If P denotes this "lazy" walk and P' denotes the "active" walk which always moves to a new vertex, we have $P = \frac{1}{k}I + \frac{k-1}{k}P'$ so $\lambda_2(P) = \frac{1}{k} + \frac{k-1}{k}\lambda_2(P')$. The more common "active" form of trickle-down states that if $\lambda_2(P') < 1$ and $\lambda_2(P'_i) \leq \lambda'$ for all i, then $\lambda_2(P') \leq \frac{\lambda'}{1-\lambda'}$. Noting that $\lambda_2(P) < 1$ if and only if $\lambda_2(P') < 1$ and $\lambda_2(P_i) \leq \lambda$ if and only if $\lambda_2(P'_i) \leq \lambda' := (\lambda - (1/k-1)) \cdot (k-1)/(k-2)$, we have*

$$\begin{aligned}\lambda_2(P) \leq \frac{1}{k} + \frac{k-1}{k}\frac{\lambda'}{1-\lambda'} &= \frac{1}{k} + \frac{k-1}{k}\frac{\frac{k-2}{k-1}\lambda'}{\frac{k-2}{k-1} - \frac{k-2}{k-1}\lambda'} \\ &= \frac{1}{k} + \frac{k-1}{k}\frac{\lambda - 1/(k-1)}{1-\lambda} = \frac{1}{k} + \frac{(k-1)\lambda - 1}{k-k\lambda} = \frac{(1-2/k)\lambda}{1-\lambda},\end{aligned}$$

matching our statement of the trickle-down theorem.

2.2 Generating Polynomial, Tilts, and Fractional Log-Concavity

Definition 2.5. *The multivariate generating polynomial $g_\mu \in \mathbb{R}[z_1, \dots, z_n]$ associated to a density $\mu\colon 2^{[n]} \to \mathbb{R}_{\geq 0}$ is given by*

$$g_\mu(z_1, \dots, z_n) := \sum_S \mu(S) \prod_{i \in S} z_i = \sum_S \mu(S) z^S,$$

Here we have used the standard notation that for $S \subseteq [n]$, $z^S = \prod_{i \in S} z_i$.

Definition 2.6. (Measure tilted by external field) *For a distribution μ on $2^{[n]}$ and vector $\lambda = (\lambda_1, \dots, \lambda_n) \in \mathbb{R}^n_{>0}$, which we refer to as the* external field, *we denote the measure μ tilted by external field λ by the notation $\lambda * \mu$, formally defined as*

$$\mathbb{P}_{\lambda * \mu}[S] = \frac{1}{Z_\lambda} \mu(S) \cdot \prod_{i \in S} \lambda_i,$$

where the normalizing constant Z_λ is defined so that $\lambda * \mu$ is a probability measure. Note that for any $(z_1, \dots, z_n) \in \mathbb{R}^n_{\geq 0}$,

$$g_{\lambda * \mu}(z_1, \dots, z_n) \propto g_\mu(\lambda_1 z_1, \dots, \lambda_n z_n).$$

In [Ali+21], the notion of fractional log-concavity of the multivariate generating polynomial was developed, generalizing the concept of log-concave polynomials (see e.g. [Ana+19]).

Definition 2.7. (Fractional log-concavity) *Consider a homogeneous distribution $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ and let $g_\mu(z_1, \dots, z_n)$ be its multivariate generating polynomial. For $\alpha \in [0, 1]$, we say that μ is α-fractionally log-concave (α-FLC) if $\log g_\mu(z_1^\alpha, \dots, z_n^\alpha)$ is concave, viewed as a function over $\mathbb{R}^n_{\geq 0}$.*

2.3 Spectral and Entropic Independence

Definition 2.8. (Correlation matrix) *Let μ be a probability distribution over $2^{[n]}$. Its correlation matrix $\Psi^{\text{cor}}_\mu \in \mathbb{R}^{n \times n}$ is defined by*

$$\Psi^{\text{cor}}_\mu(i, j) = \begin{cases} 1 - \mathbb{P}_\mu[i] & \text{if } j = i, \\ \mathbb{P}_\mu[j \mid i] - \mathbb{P}_\mu[j] & \text{otherwise.} \end{cases}$$

Definition 2.9. (Spectral independence) *For $\eta \geq 0$, a distribution $\mu : 2^{[n]} \to \mathbb{R}_{\geq 0}$ is said to be η-spectrally independent (at the link $\emptyset$) if*

$$\lambda_{\max}(\Psi^{\text{cor}}_\mu) \leq \eta.$$

Remark 2.2. *The original definition of spectral independence in [ALO20] imposes such a requirement on μ as well as all of its links. Here, we follow the convention in [Ana+21a] and use the term spectral independence to refer only to a spectral norm bound on the correlation matrix of μ, with the understanding that in applications, we will require spectral independence of all links of μ as well.*

Fact 2.1. (Remark 70 of [Ali+21]) *A distribution μ on $\binom{[n]}{k}$ is η-spectrally independent iff*

$$\nabla^2 \log g_\mu(z_1^{1/\eta}, \dots, z_n^{1/\eta})\Big|_{z=\vec{1}} = (1/\eta)^2 D \Psi^{\text{cor}}_\mu - (1/\eta) D \preceq 0$$

where D is the diagonal matrix with entries $D_{ii} = \mathbb{P}_\mu[i]$.

Moreover, μ is $1/\eta$-FLC iff $\lambda * \mu$ is η-spectrally independent for all external fields $\lambda \in \mathbb{R}^n_{\geq 0}$.

Lemma 2.1. ([ALO20]) *Suppose $P = U_{1 \to k} D_{k \to 1}$ is the transition operator for the $1 \leftrightarrow k$ up-down walk for a distribution μ on $\binom{[n]}{k}$. Then*

$$\lambda_2(P) = \frac{\lambda_{\max}(\Psi^{\text{cor}}_\mu)}{k}.$$

Proof. We include the proof of this result for completeness. From the definitions we see that for the vector d with $d_i = \mathbb{P}[i]$, we have that

$$\Psi_\mu^{\mathrm{cor}} = kP - \vec{1}d^T = k\left(P - \frac{1}{k}\vec{1}d^T\right),$$

i.e. $P = \frac{1}{k}\vec{1}d^T + \frac{1}{k}\Psi_\mu^{\mathrm{cor}}$. Observe that $\frac{1}{k}d$ is the stationary distribution of P, so P is self-adjoint [LP17] with respect to the inner product $\langle \cdot, \Pi \cdot \rangle$ where $\Pi := \frac{1}{k}\mathrm{diag}(d)$ and $\langle \cdot, \cdot \rangle$ denotes the Euclidean dot product. We see by using the variational characterization of eigenvalues that

$$\lambda_2(P) = \sup_v \frac{\langle v, \Pi(P - \frac{1}{k}\vec{1}d^T)v\rangle}{\langle v, \Pi v\rangle} = \sup_v \frac{\langle v, \frac{1}{k}\Pi\Psi_\mu^{\mathrm{cor}} v\rangle}{\langle v, \Pi v\rangle} = \frac{\lambda_{\max}(\Psi_\mu^{\mathrm{cor}})}{k}.$$

□

Definition 2.10. (Entropic independence) *A probability distribution μ on $\binom{[n]}{k}$ is said to be C-entropically independent, for $C \geq 1$, if for all probability distributions ν on $\binom{[n]}{k}$,*

$$\mathcal{D}_{\mathrm{KL}}(\nu D_{k\to 1} \parallel \mu D_{k\to 1}) \leq \frac{C}{k}\mathcal{D}_{\mathrm{KL}}(\nu \parallel \mu).$$

This is an exact analogue of spectral independence, replacing variance by entropy. It is also a generalization of subadditivity of entropy which itself is equivalent to a generalized Brascamp-Lieb inequality, see e.g. [Ana+21a; Bar+11; Che+22; Bla+21] for discussion.

Theorem 2.2. ([Ana+21a]) *A distribution μ on $\binom{[n]}{k}$ is $(1/C)$-FLC if and only if $\lambda * \mu$ is C-entropically independent for all external fields $\lambda \in \mathbb{R}^n_{>0}$ (in particular, all links of μ are C-entropically independent).*

2.4 Functional Inequalities

Let P be the transition matrix of an ergodic, reversible Markov chain on a finite set Ω, with (unique) stationary distribution μ. The Dirichlet form of P is defined, for $f, g \colon \Omega \to \mathbb{R}$, by

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

For later use, we record an equivalent expression for the Dirichlet form of down-up walks.

Lemma 2.2. *Let μ be a distribution on $\Omega := \binom{[n]}{k}$ and let P denote the transition matrix of the $k \leftrightarrow k-1$ down-up walk. Then, for any $f, g : \Omega \to \mathbb{R}$,*

$$\mathcal{E}_P(f,g) = \mathbb{E}_{S_{k-1}\sim \mu D_{k\to k-1}}[\mathrm{Cov}(f(S), g(S) \mid S_{k-1})].$$

Proof. For notational convenience, let $\mu_{k-1} := \mu D_{k\to k-1}$ and $\Omega_{k-1} := \binom{[n]}{k-1}$. We have,

$$\begin{aligned}
\mathcal{E}_P(f,g) &= \frac{1}{2}\sum_{x,y\in\Omega}\sum_{z\in\Omega_{k-1}} \mu(x)D_{k\to k-1}(x,z)U_{k-1\to k}(z,y)(f(x)-f(y))(g(x)-g(y)) \\
&= \frac{1}{2}\sum_{z\in\Omega_{k-1}}\sum_{x,y\in\Omega} \mu_{k-1}(z)U_{k-1\to k}(z,x)U_{k-1\to k}(z,y)(f(x)-f(y))(g(x)-g(y)) \\
&= \sum_{z\in\Omega_{k-1}} \mu_{k-1}(z)\frac{\mathbb{E}_{X\sim\mu|z, Y\sim\mu|z}[(f(X)-f(Y))(g(X)-g(Y))]}{2} \\
&= \mathbb{E}_{S_{k-1}\sim\mu_{k-1}}[\mathrm{Cov}(f(S), g(S) \mid S_{k-1})].
\end{aligned}$$

□

Definition 2.11. *The spectral gap or Poincaré constant of P is defined to be γ, where γ is the largest value such that for every $f\colon \Omega \to \mathbb{R}$,*

$$\gamma \operatorname{Var}_\mu[f] \le \mathcal{E}_P(f, f).$$

The modified log-Sobolev constant of P is defined to be the the largest value ρ such that for every $f\colon \Omega \to \mathbb{R}_{\ge 0}$,

$$\rho \operatorname{Ent}_\mu[f] \le \mathcal{E}_P(f, \log f),$$

where $\operatorname{Ent}_\mu[f] = \mathbb{E}_\mu[f \log f] - \mathbb{E}_\mu[f] \log(\mathbb{E}_\mu[f])$.

Definition 2.12. *Let P be an ergodic Markov chain on a finite state space Ω and let μ denote its (unique) stationary distribution.For any $\varepsilon \in (0, 1)$, we define the ε-*total variation mixing time *to be*

$$\tau_{mix}(\varepsilon) = \max\bigl\{\min\{t \ge 0 \mid d_{TV}(\mathbb{1}_x P^t, \mu) \le \varepsilon\} \bigm| x \in \Omega\bigr\},$$

where $\mathbb{1}_x$ is the point mass supported at x and d_{TV} is the total variation distance [CT12].

The following relationships between the ε-(total variation) mixing time of P, $\tau_{\mathrm{mix}}(\varepsilon)$, and its Poincaré and modified log-Sobolev constants is standard (see, e.g., [BT06; LP17]):

$$(\gamma^{-1} - 1) \log\left(\frac{1}{2\varepsilon}\right) \le \tau_{\mathrm{mix}}(\varepsilon) \le \gamma^{-1} \log\left(\frac{1}{\varepsilon} \cdot \frac{1}{\min_{x\in\Omega} \mu(x)}\right), \tag{2.1}$$

$$\tau_{\mathrm{mix}}(\varepsilon) \le \rho^{-1}\left(\log\log\left(\frac{1}{\min_{x\in\Omega}\mu(x)}\right) + \log\left(\frac{1}{2\varepsilon^2}\right)\right).$$

Definition 2.13. (Approximate tensorization of entropy) *A distribution μ on $\binom{\Omega}{n}$ satisfies approximate tensorization of entropy with constant C if for all probability distributions ν on $\binom{\Omega}{n}$,*

$$\mathcal{D}_{\mathrm{KL}}(\nu D_{n\to n-1} \parallel \mu D_{n\to n-1}) \le (1 - 1/(Cn))\, \mathcal{D}_{\mathrm{KL}}(\nu \parallel \mu).$$

This is a generalization of the classical definition of approximate tensorization of entropy (see [Mar15; CMT15]), as we observe in the following remark. Explaining this also requires introducing the important concept of *homogenization* which we use throughout this paper:

Definition 2.14. (Homogenization) *Let μ be a distribution on a product space $\Omega' = \Omega'_1 \times \cdots \times \Omega'_n$. Then μ can naturally be viewed as a distribution μ^{hom} over $\binom{\Omega}{n}$, where $\Omega = \cup_{i=1}^n \Omega'_i \times \{i\}$ by identifying $\sigma \in \Omega'$ with the set $\{(\sigma_1, 1), (\sigma_2, 2), \ldots, (\sigma_n, n)\}$. Note that under this identification, the Glauber dynamics corresponds to the $n \leftrightarrow n-1$ down-up walk.*

Remark 2.3. *Let μ be a distribution on a product space and define its homogenization μ^{hom} as above. In this case, approximate tensorization of entropy with constant C for μ^{hom} is equivalent to the assertion that for any positive measurable function f,*

$$\operatorname{Ent}_\mu[f] \le C \sum_{v=1}^{n} \mathbb{E}_\mu[\operatorname{Ent}_v[f]]$$

where

$$\operatorname{Ent}_k[f] := \operatorname{Ent}_{\mu(\sigma_k = \cdot \mid \sigma_{\sim k})}[f]$$

is the entropy functional with respect to the conditional measure of $\sigma_k \in \Omega'_k$ given $\sigma_j \in \Omega'_j$ for all $j \ne k$ and for $\sigma \sim \mu$.

It is an immediate consequence of the data processing inequality (see, e.g., [Ana+21a]) that if μ satisfies approximate tensorization of entropy with constant C, then the $n \leftrightarrow n-1$ down-up walk for μ has modified log-Sobolev constant at least $1/Cn$.

3 Universality of Spectral Independence

In this section, we present our key result on the universality of spectral independence, Theorem 3.1, in the general setting of down-up walks (on pure simplicial complexes).

3.1 $k \leftrightarrow k-1$ Spectral Gap Implies Spectral Independence

THEOREM 3.1. *Let μ be a distribution on $\binom{[n]}{k}$. If the $k \leftrightarrow k-1$ down-up walk has spectral gap $\frac{1}{Ck}$, then μ is C-spectrally independent.*

Proof. From Fact 2.1, C-spectral independence is equivalent to the inequality $D\Psi_\mu^{\text{cor}} \preceq CD$, where D is the diagonal matrix with entries $D_{ii} = \mathbb{P}_\mu[i]$. This is equivalent to showing that for all vectors $v \in \mathbb{R}^n$,

$$\mathrm{Var}_{S\sim\mu}\left[\sum_{i\in S} v_i\right] \le C\,\mathbb{E}_{S\sim\mu}\left[\sum_{i\in S} v_i^2\right].$$

From the definition of the spectral gap and Lemma 2.2, we have for any function $f : \binom{[n]}{k} \to \mathbb{R}$ that

$$\mathrm{Var}_{S\sim\mu}[f(S)] \le Ck \cdot \mathbb{E}_{S_{k-1}\sim\mu D_{k\to k-1}}[\mathrm{Var}[f(S) \mid S_{k-1}]].$$

Applying this inequality to the function $f(S) = \sum_{i\in S} v_i$, we observe that

$$\begin{aligned}\frac{1}{Ck}\mathrm{Var}_{S\sim\mu}\left[\sum_{i\in S} v_i\right] &\le \mathbb{E}_{S_{k-1}\sim\mu D_{k\to k-1}}\left[\mathrm{Var}\left[\sum_{i\in S} v_i \mid S_{k-1}\right]\right]\\ &= \mathbb{E}_{S_{k-1}\sim\mu D_{k\to k-1}}\left[\mathrm{Var}\left[\sum_{i\in S\setminus S_{k-1}} v_i \mid S_{k-1}\right]\right]\\ &\le \mathbb{E}_{S_{k-1}\sim\mu D_{k\to k-1}}\left[\mathbb{E}\left[\sum_{i\in S\setminus S_{k-1}} v_i^2 \mid S_{k-1}\right]\right] = \frac{1}{k}\mathbb{E}_{S\sim\mu}\left[\sum_{i\in S} v_i^2\right],\end{aligned}$$

as desired. Here, in the second inequality we have used that $\mathrm{Var}[f(X)] \le \mathbb{E}[f(x)^2]$ and the fact that the sum is over the set $S \setminus S_{k-1}$ which has size exactly one, and in the last equality we used symmetry. ☐

REMARK 3.1. *It is well-known that a bounded Poincaré constant implies that the largest eigenvalue of the covariance matrix is also bounded. In contrast, Theorem 3.1 shows spectral independence, which is a much stronger property. Spectral independence exactly controls the largest eigenvalue of the* influence *matrix rather than the covariance; reinterpreted in terms of the covariance matrix, it asserts a PSD upper bound not just by a multiple of the identity, but by a multiple of the diagonal matrix D of marginals, which is often much smaller. For example, for the uniform distribution on $\binom{[n]}{k}$, spectral independence tells us that the largest eigenvalue of the covariance matrix is $O(k/n)$, rather than just $O(1)$.*

3.2 Trickle-Down of $k \leftrightarrow k-1$ Spectral Gap As a consequence of the result established in the previous section, we can prove an analogue of the trickle-down theorem (which bounds spectral gaps of $k \leftrightarrow 1$ walks inductively) for the $k \leftrightarrow k-1$ walk. This follows by combining Theorem 3.1 with Oppenheim's trickle-down theorem [Opp18] and the "local-to-global" argument [AL20; KO18].

THEOREM 3.2. *Let μ be a distribution on $\binom{[n]}{k}$ with $k \ge 3$ so that for every $i \in [n]$, the $k-1 \leftrightarrow k-2$ down-up walk on the link μ_i has spectral gap at least $1/C(k-1)$, and such that the $k \leftrightarrow k-1$ down-up walk on μ is ergodic. Then the $k \leftrightarrow k-1$ down-up walk on μ has spectral gap at least $1/C''k$, where $C'' := C\frac{k-1-C}{k-2C}$.*

Proof. Theorem 3.1 implies that each link μ_i is C-spectrally independent for every i. We claim that μ is C'-spectrally independent for $C' := \frac{C(k-2)/(k-1)}{1-C/(k-1)}$. Indeed, using Lemma 2.1 and Theorem 2.1:

$$\lambda_{\max}(\Psi_\mu^{\text{cor}}) = k\lambda_2(P) \le k\frac{(1-2/k)C/(k-1)}{1-C/(k-1)} = \frac{C(k-2)/(k-1)}{1-C/(k-1)} = C'.$$

Next, we use spectral independence to perform a step of the local-to-global argument. By the law of total variance, for any function f

$$\begin{aligned}
\operatorname{Var}_\mu[f] &= \operatorname{Var}_{i\sim\mu D_{k\to1}}[\mathbb{E}_\mu[f \mid i]] + \mathbb{E}_{i\sim\mu D_{k\to1}}[\operatorname{Var}_\mu[f \mid i]] \\
&= \operatorname{Var}_{\mu D_{k\to1}}[U_{1\to k} f] + \mathbb{E}_{i\sim\mu D_{k\to1}}[\operatorname{Var}_\mu[f \mid i]] \\
&= \mathbb{E}[(U_{1\to k}(f - \mathbb{E}[f]))^2] + \mathbb{E}_{i\sim\mu D_{k\to1}}[\operatorname{Var}_\mu[f \mid i]] \\
&= \mathbb{E}[(f - \mathbb{E}[f])(D_{k\to1}U_{1\to k}(f - \mathbb{E}[f]))] + \mathbb{E}_{i\sim\mu D_{k\to1}}[\operatorname{Var}_\mu[f \mid i]] \\
&\le \lambda_2(D_{k\to1}U_{1\to k}) \operatorname{Var}_\mu[f] + \mathbb{E}_{i\sim\mu D_{k\to1}}[\operatorname{Var}_\mu[f \mid i]] \\
&\le \frac{C'}{k} \operatorname{Var}_\mu[f] + \mathbb{E}_{i\sim\mu D_{k\to1}}[\operatorname{Var}_\mu[f \mid i]]\,,
\end{aligned}$$

where in the last step we used $\lambda_2(D_{k\to1}U_{1\to k}) = \lambda_2(U_{1\to k}D_{k\to1}) \le C'/k$ via Lemma 2.1 and spectral independence.

Hence, rearranging and applying the spectral gap bound for the $k-1 \leftrightarrow k-2$ walks of the links, we have

$$\begin{aligned}
\operatorname{Var}_\mu[f] &\le \frac{1}{1 - C'/k}\, \mathbb{E}_{i\sim\mu D_{k\to1}}[\operatorname{Var}_\mu[f \mid i]] \\
&\le \frac{C(k-1)}{1 - C'/k}\, \mathbb{E}_{S_{k-1}\sim\mu D_{k\to k-1}}[\operatorname{Var}_\mu[f \mid S_{k-1}]]\,,
\end{aligned}$$

which bounds the inverse spectral gap of the $k \leftrightarrow k-1$ walk by $1/C''k$ for

$$C'' := \frac{C(1-1/k)}{1 - C'/k} = \frac{C(1-1/k)}{1 - \frac{C(k-2)/(k-1)}{k - Ck/(k-1)}} = \frac{C(k-1)(1 - C/(k-1))}{k - Ck/(k-1) - C(k-2)/(k-1)} = \frac{C(k-1-C)}{k-2C}.$$

□

4 Results for Gibbs Measures on the Hypercube

Notation. In this section, we consider distributions of the form $\mu(\sigma) \propto \exp(H(\sigma))$ where $H : \{\pm1\}^n \to \mathbb{R}$. Following [EG18], we identify H with its *multilinear extension* $H : \mathbb{R}^n \to \mathbb{R}$ defined by

$$H(x) = \sum_{S\subset[n]} \hat{H}(S) \prod_{i\in S} x_i$$

where $\hat{H}(S) := \frac{1}{2^n}\sum_{\sigma\in\{\pm1\}^n} H(\sigma)\prod_{i\in S}\sigma_i$ is the Fourier transform of H viewed as a function on the hypercube [ODo14]. This is also known as the *harmonic extension* since the Laplacian of H vanishes, and for $x \in [-1,1]^n$ it admits an equivalent expression

$$H(x) = \mathbb{E}_{\sigma\sim\otimes_i \operatorname{Ber}_\pm(x_i)}[H(\sigma)] \tag{4.2}$$

where $\operatorname{Ber}_\pm(x)$ denotes the distribution of a random variable valued in $\{\pm1\}$ with mean x. For $\sigma \in \{\pm1\}^n$, define

$$B_j(\sigma) := \partial_j H(\sigma)$$

to be the *cavity field* at site j, where $\partial_j H$ is the usual partial derivative applied to the multilinear extension of H. Because H is multilinear, the cavity field B_j does not depend on σ_j.

Note that in our notational convention, B_j refers to the same object as in [Adh+22] but ∂_j can differ in sign from their definition. More generally, following [Adh+22] we define versions of H and B_j for reduced versions of the original system which appear when performing induction. The generalization of H is parameterized by disjoint sets $A, B \subset [n]$ and $\sigma_A \in \{\pm1\}^A$ and given by

$$H^{[A,B]}_{\sigma_A}(\sigma_{(A\cup B)^c}) := \sum_{S\subset B^C} \hat{H}(S) \prod_{i\in S} \sigma_i.$$

In the other words, this is the multilinear extension of H evaluated at the vector $(\sigma_A, \sigma_{(A\cup B)^C}, 0_B)$. Similarly, we define

$$B^{[A,B]}_j(\sigma_{(A\cup B)^c}) := \partial_j H^{[A,B]}_{\sigma_A}(\sigma_{(A\cup B)^c})$$

where on the left hand side, the dependence on σ_A is omitted from the notation for convenience.

4.1 Results under the Smoothness Condition In the main result of this section, we prove that smallness of the Hessian of H implies that the Gibbs measure $\mu(x) \propto \exp(H(x))$ is fractionally log-concave and the Glauber dynamics has $\Omega(1/n)$ spectral gap (i.e. relaxation time $O(n)$). Afterwards, in Theorem 4.2 we prove that by combining our fractional log-concavity estimate with the entropic independence framework [Ana+21a], we obtain strong bounds on the mixing time, MLSI constant, and approximate tensorization constant of μ.

Theorem 4.1. *There exist absolute constants $A, B > 0$ for which the following holds. Suppose that μ is a probability measure with full support on the hypercube $\{\pm 1\}^n$, so $\mu(x) \propto \exp(H(x))$ for some function $H : \{\pm 1\}^n \to \mathbb{R}$. Suppose furthermore that*

$$\beta := \max_{\sigma \in \{\pm 1\}^n} \|\nabla^2 H(\sigma)\|_{\mathrm{op}} \le A.$$

Then we have that:

1. *The Poincaré constant (spectral gap) for the discrete-time Glauber dynamics on μ is at least $\frac{1}{(1+B\beta)n}$.*
2. *μ^{hom} is $\frac{1}{1+B\beta}$-fractionally log-concave.*

This result will be proved by combining the results we established in Section 3 with an important estimate established in [Adh+22], which we now recall.

We start with some notation. Let

$$T := \sup_{A,B: A\cap B = \emptyset} \sup_{\sigma_A \in \{\pm 1\}^{|A|}} \sup_{\sigma \in \{\pm 1\}^{N-|A\cup B|}} \left\| \left(\partial_i B_j^{[A,B]}(\sigma) \right)_{1 \le i,j \le N - |A\cup B|} \right\|_{\mathrm{op}}$$

and for any $0 \le k \le N$, let a_{N-k} be the worst-case dimension-free Poincaré constant among all subsystems with $|A \cup B| = k$. Precisely, a_{N-k} is defined to be the smallest positive number so that for all A, B disjoint with $|A \cup B| = k$ and all $\sigma_A \in \{\pm 1\}^A$, the Glauber dynamics for every measure of the form $\mu_{\sigma_A}^{[A,B]}(\sigma) \propto \exp(H_{\sigma_A}^{[A,B]}(\sigma))$ on $\{\pm 1\}^{(A\cup B)^c}$ has spectral gap at least $\frac{1}{a_{N-k}(N-k)}$.

The following key result from [Adh+22] relates the values of a_{N-k} between different values of k:

Lemma 4.1. (Proposition 4.1 of [Adh+22]) *There exist absolute constants $C, \beta_0 > 0$ for which the following holds. Let $C_r := (Cr)^2 e^{Cr} = \Theta_C(r^2)$ for $r \le 1/C$. Suppose that $\beta < \beta_0$ and $\epsilon \in (0, 10^{-2})$ are such that $T \le 5\beta$ and $a_{N-k-1} < \epsilon / C_\beta$. Then*

$$\left(1 - \frac{C\beta^2 e^{C\beta} \max(1, a_{N-k})^2}{\epsilon(N-k)}\right) a_{N-k} \le \left(1 - \frac{1}{N-k}\right) a_{N-k-1} + \frac{(1+4\epsilon)^5}{N-k}.$$

The key difficulty in using this relation to inductively bound the spectral gap is the presence of the term $\max(1, a_{N-k})^2$ which, if large, will make the bound trivial. In [Adh+22], this was overcome using a "continuity argument" (Section 3 there) which uses properties specific to the p-spin model (random H). It turns out we can eliminate the need for a specialized continuity argument completely using the general results established in Section 3.

We first check that the assumption on T is satisfied. Actually, it ends up to be equivalent to our definition of β.

Lemma 4.2. *With the notation above, $T = \beta$.*

Proof. Observe that for any point $x \in [-1, 1]^n$, by linearity and (4.2) we have

$$\nabla^2 H(x) = \mathbb{E}_{\sigma \sim \otimes_i \mathrm{Ber}_\pm(x_i)}[\nabla^2 H(\sigma)].$$

So by the triangle inequality $\|\nabla^2 H(x)\|_{\mathrm{op}} \le \max_{\sigma \in \{\pm 1\}^n} \|\nabla^2 H(\sigma)\|_{\mathrm{op}}$, and hence

$$\sup_{x \in [-1,1]^n} \|\nabla^2 H(x)\|_{\mathrm{op}} = \max_{\sigma \in \{\pm 1\}^n} \|\nabla^2 H(\sigma)\|_{\mathrm{op}}.$$

From the definition, it's clear that $T \ge \beta$ and since $H^{[A,B]}(\sigma_{(A\cup B)^c})$ is the multilinear extension of H evaluated at the vector $(\sigma_A, \sigma_{(A\cup B)^C}, 0_B)$, it follows from the above argument that $T \le \beta$. □

Proof. [Proof of Theorem 4.1] First, we observe that if we establish conclusion (1) concerning the spectral gap, it will automatically imply conclusion (2) regarding fractional log-concavity via Theorem 3.1. Indeed, by Fact 2.1, we know that establishing μ^{hom} is $\frac{1}{1+B\beta}$-FLC is equivalent to showing that for all external fields λ, the tilted measure $\lambda * \mu^{\text{hom}}$ is $(1+B\beta)$-spectrally independent. Note that tilting μ^{hom} is equivalent to first changing the degree-one part of H, and then homogenizing the resulting measure. Furthermore, changing the degree-one part of H does not change $\nabla^2 H$ and hence the assumption of this theorem is invariant to arbitrary tilts by external fields. Since the Glauber dynamics is the $n \leftrightarrow n-1$ down-up walk on μ^{hom}, it therefore suffices to prove that the spectral gap of the Glauber dynamics is at least $1/(1+B\beta)n$, which is precisely what conclusion (1) says.

It remains to prove conclusion (1). We bound a_{N-k} by induction on $(N-k)$. The induction hypothesis is $a_{N-k} \le 1+\delta$ with $\delta \le 1/2$ to be chosen later. For base cases, we have the bound when $N-k \le 2$ because in this case the measure is an Ising model (on one or two sites), so the desired bound follows from Theorem 1 of [EKZ21].

Let $k \le N-3$ and assume that $a_{N-k-1} \le 1+\delta$. We will prove $a_{N-k} \le 1+\delta$.

By Theorem 3.2,

$$a_{N-k} \le a_{N-k-1}\frac{N-k-1-a_{N-k-1}}{N-k-2a_{N-k-1}} = a_{N-k-1}T_{N-k-1}$$

with

$$T_{N-k-1} := \frac{1}{2} + \frac{(N-k)/2-1}{N-k-1-a_{N-k-1}} \le \frac{1}{2} + \frac{3/2}{2-3/2} = 7/2$$

for $k \le N-3$.

Let $\epsilon = \delta/10$. Below, we will choose $\delta = \omega_{\beta\to 0}(\beta^2)$. In particular, for β_0 sufficiently small, we have

$$a_{N-k-1} \le 1+\delta \le \frac{\epsilon}{C_\beta}.$$

Hence, by Lemma 4.1 we have

$$\left(1 - \frac{T_{N-k-1}^2 C\beta^2 e^{C\beta}(1+\delta)^2}{\epsilon(N-k)}\right) a_{N-k} \le \left(1-\frac{1}{N-k}\right) a_{N-k-1} + \frac{(1+4\epsilon)^5}{N-k}$$

thus

$$a_{N-k} \le \left(1-\frac{1}{N-k}\right) a_{N-k-1} + a_{N-k}\frac{T_{N-k-1}^2 C\beta^2 e^{C\beta}(1+\delta)^2}{\epsilon(N-k)} + \frac{(1+4\epsilon)^5}{N-k}$$

Since $a_{N-k} \le a_{N-k-1}T_{N-k-1}$, we can bound the second term by $\frac{T_{N-k-1}^3 C\beta^2 e^{C\beta}(1+\delta)^2}{\epsilon(N-k)} a_{N-k-1}$. Assuming $\epsilon \ge 3\delta^{-1}T_{N-k-1}^3 C\beta^2 e^{C\beta}$ we have

$$\begin{aligned} a_{N-k} &\le \left(1-\frac{1}{N-k}+\frac{\delta}{3(N-k)}\right) a_{N-k-1} + \frac{(1+4\epsilon)^5}{N-k} \\ &\le \left(1+\frac{\delta/3-1}{N-k}\right)(1+\delta) + \frac{(1+4\epsilon)^5}{N-k} \\ &\le (1+\delta) + \frac{1}{N-k}\left((\delta/3-1)(\delta+1) + (1+4\epsilon)^5\right). \end{aligned}$$

Recall that $\epsilon = \delta/10$. Substituting this in the second term, we can verify that for all $\epsilon = [0, 0.01]$, $g(\epsilon) = (10\epsilon/3-1)(10\epsilon+1) + (1+4\epsilon)^5 < 0$, so that the induction holds provided that $\epsilon \ge 3\delta^{-1}T_{N-k-1}^3 C\beta^2 e^{C\beta}$ (i.e. $10\epsilon^2 > 3T_{N-k-1}^3 C\beta^2 e^{C\beta}$) and $\delta = \omega_{\beta\to 0}(\beta^2)$. Since $T_{N-k-1} \le 7/2$ under the inductive hypothesis, it is readily seen that taking $\epsilon = \Omega_C(\beta)$ suffices for the induction to hold.

Finally, the bound on the Poincaré constant is $a_N \le 1+\delta = 1+\Theta_C(\beta)$.

□

THEOREM 4.2. *Suppose μ and β satisfy the same assumptions as Theorem 4.1. Then, with the notation there, we have:*

1. *μ (equivalently μ^{hom}) satisfies approximate tensorization of entropy with constant $C = O(n^{B\beta})$.*

2. *The discrete-time Glauber dynamics on μ satisfy the Modified Log-Sobolev Inequality (MLSI) with constant $\rho = \Omega(1/n^{1+B\beta})$.*

3. *The discrete-time Glauber dynamics on μ satisfy*

$$\tau_{mix}(\epsilon) = O\Big(n^{1+B\beta}(\log\log(1/\min_{\sigma}\mu(\sigma)) + \log(1/\epsilon)\Big).$$

The proof of this theorem requires the following intermediate lemma showing approximate entropy tensorization for a system satisfying the conditions of Theorem 4.2 under pinning with $N-k = O(1)$ free (unpinned) spins. [CMT14, Lemma 2.2] showed this result only for 2-spin systems, but the proof applies more generally. We repeat the proof for the sake of completeness. The details are in Section B.

Lemma 4.3. *Let μ be a distribution satisfying the assumptions of Theorem 4.2. Any pinning of μ where the number of free spins $N-k$ is constant ($N-k = O(1)$) has approximate entropy tensorization with constant $C = \exp(O(\beta))$.*

Proof. [Proof of Theorem 4.2] The second and third conclusion follow directly from approximate tensorization of entropy (see Section 2).

Let $\alpha = 1/(1+B\beta)$. By Theorem 5 of [Ana+21a], we have that if the measure μ^{hom} is α-FLC then for any measure ν absolute continuous with respect to μ and $k_0 = \lceil 1/\alpha \rceil$,

$$\mathcal{D}_{\text{KL}}(\nu D_{n\to(n-k_0)} \parallel \mu^{\text{hom}} D_{n\to(n-k_0)}) \le (1-\kappa)\,\mathcal{D}_{\text{KL}}(\nu \parallel \mu^{\text{hom}})$$

where

$$\kappa = \frac{(3-1/\alpha)^{1/\alpha - \lceil 1/\alpha\rceil}\prod_{i=0}^{\lceil 1/\alpha\rceil - 1}(2-i)}{(n+1)^{1/\alpha}} \ge n^{1/\alpha - 1} = n^{B\beta}$$

This is equivalent to block approximate entropy tensorization with block size k_0 [Bla+21, Eq.(1.5)] i.e.

$$\frac{k_0}{n}\operatorname{Ent}_\mu[f] \le n^{B\beta}\frac{1}{\binom{n}{k_0}}\sum_{S\in\binom{[n]}{k_0}}\mathbb{E}_\mu[\operatorname{Ent}_S[f]].$$

Since $|S| = k_0 = O(1+B\beta) = O(1)$ and $\beta = O(1)$, combining this with Lemma 4.3 gives

$$\frac{1}{n}\operatorname{Ent}_\mu[f] \le n^{B\beta}\frac{\exp(O(\beta))}{k_0\binom{n}{k_0}}\sum_{S\in\binom{[n]}{k_0}}\sum_{v\in S}\mathbb{E}_\mu[\operatorname{Ent}_v[f]] = n^{B\beta}\frac{\exp(O(\beta))}{n}\sum_{v\in[n]}\mathbb{E}_\mu[\operatorname{Ent}_v[f]],$$

as required. □

4.2 Results for the p-Spin Model In this subsection, we prove Theorem 1.3. As mentioned in the introduction, we will need to rely on an average case local to global argument. Specifically, we will need the following theorem, which is a slight modification of [ALV22, Theorem 20] and follows from the same proof.

Theorem 4.3. *Consider a distribution $\mu : \binom{[n]}{k} \to \mathbb{R}_{\ge 0}$. Suppose that for every set T of size $\le k-2$, μ_T is $(k-|T|)(1-\rho(T))$-entropically independent i.e.*

$$\mathcal{D}_{\text{KL}}(\nu D_{k-|T|\to 1} \parallel \mu_T D_{k-|T|\to 1}) \le (1-\rho(T))\,\mathcal{D}_{\text{KL}}(\nu \parallel \mu_T).$$

Suppose also that there exist constants $k_0 \ge 2$ and $C(k_0)$ such that for all T with $|T| \ge k-k_0$, μ_T satisfies approximate entropy tensorization with constant $C(k_0)$.

Finally, for a set T of size $\ge k-1$, define the harmonic mean

$$\gamma_T := \mathbb{E}_{e_1,\ldots,e_{|T|}\sim\text{uniformly random permutation of } T}\Big[\big(\rho(\emptyset)\rho(\{e_1\})\rho(\{e_1,e_2\})\cdots\rho(\{e_1,\ldots,e_{k-k_0}\})\big)^{-1}\Big]^{-1}.$$

Then the operator $D_{k\to(k-1)}$ satisfies

$$\mathcal{D}_{\mathrm{KL}}(\nu D_{k\to(k-1)} \parallel \mu D_{k\to(k-1)}) \le (1-\kappa)\,\mathcal{D}_{\mathrm{KL}}(\nu \parallel \mu),$$

with

$$\kappa := C(k_0)^{-1} \min\left\{ \gamma_T \;\middle|\; T \in \binom{[n]}{k-1} \right\}.$$

Notation. For each k, let A_k be a uniformly random set of k spins[1]. We will bound the norm of the tensor associated to A_k when A_k^c is fixed according to σ. It will be more convenient to work with the normalization $\sigma \in \{\pm1\}^N/\sqrt{N}$. For a sequence S of coordinates and a spin configuration σ on a set of coordinates containing S, we denote $\sigma_S = \prod_{i\in S}\sigma_i$. We write $|S|$ for the length of S, and write $S \subset A$ if all entries of S are in A. For sequences S, T of coordinates, we denote by $g_{(S,T)}$ the entry of the disorder indexed by (S,T); and $g_{\{S,T\}}$ the sum of all entries whose index is the union of S and T (with the same order of elements within S and T). For a given realisation of A_k, disjoint subsets A', B' of A_k and $\sigma' \in \{\pm1\}^{A_k\setminus B'}/\sqrt{N}$, we define the matrix $\Delta^{[A',B']}$, with rows and columns indexed by $A_k \setminus (A' \cup B')$ (here, we are suppressing dependence on σ, σ', A_k):

$$\Delta^{[A',B']}_{i,j} = \sum_{p\ge2} \frac{1}{\sqrt{N}}\beta_p \sum_{s+s'=p-2} \sum_{\substack{S\subset A_k^c, |S|=s,\\ S'\subset A_k\setminus B', |S'|=s'}} g_{\{i,j,S,S'\}}\sigma_S\sigma'_{S'}, \tag{4.3}$$

i.e. $\partial_i B_j^{[A',B']}(\sigma') = \Delta^{[A',B']}_{i,j}$.

The key additional ingredient we need to prove Theorem 1.3 using Theorem 4.3 is the following estimate.

PROPOSITION 4.1. *There exists a constant $C > 0$ such that, with notation as in the proof of Theorem 1.3, with probability at least $1-\exp(-\Theta(N))$ over the choice of the Hamiltonian $H = H(g)$,*

$$\sup_{I\in[N],\sigma\in\{\pm1\}^{[N]\setminus\{I\}}} \mathbb{E}_{\pi\in \mathrm{Sym}(N-\{I\})}\left[\exp\left(\frac{B\sum_{N\ge k>k_0}\sup_{A',B'\subset A_k}\sup_{\sigma'\in\{\pm1\}^{A_k\setminus B'}}\|\Delta^{[A',B']}\|_{\mathrm{op}}}{k}\right)\right] \le C. \tag{4.4}$$

Here A_k is the union of $\{I\}$ and the last $k-1$ elements according to the permutation π on $[N]\setminus\{I\}$.

The proof of Proposition 4.1 is presented at the end of this section.

Given Theorem 4.2, Theorem 4.3 and Proposition 4.1, the proof of Theorem 1.3 is immediate.

Proof. [Proof of Theorem 1.3]

Let B be the constant appearing in Theorem 4.1. For $A \subseteq [n]$ and $\sigma_A \in \{\pm1\}^A$, as in Section 4.1, let H_{σ_A} be the Hamiltonian of the subsystem where the spins in A are pinned according σ_A.

We want to establish approximate tensorization of entropy, which is the same as entropy contraction of $D_{N\to(N-1)}$ for μ^{hom}. The set T of size $N-1$ in Theorem 4.3 corresponds to a tuple $(i,\sigma) \in [N] \times \{\pm1\}^{[N]\setminus\{i\}}$. Each permutation $e_1, \ldots, e_{|T|}$ of T corresponds to a permutation π of $[N]\setminus\{i\}$. For each k, $\rho(\{e_j \mid j \le N-k\})$ corresponds to the KL-divergence contraction of the $D_{k\to1}$ operator wrt $\mu_{\sigma_{A^c_{\pi,k}}}$ where $A^c_{\pi,k} = \{e_j \mid j \le N-k\}$.

By Theorem 4.1, $\mu_{\sigma_{A^c_{\pi,k}}}$ is $\frac{1}{1+BT_{\sigma_{A^c_{\pi,k}}}}$-fractionally log concave so by Theorem 2.2 the $D_{k\to1}$ operator contracts KL-divergence by

$$\rho(\sigma_{A^c_{\pi,k}}) = 1 - \frac{1+BT_{\sigma_{A^c_{\pi,k}}}}{k}$$

with $T_{\sigma_{A^c_{\pi,k}}} = \sup_{A',B'\subseteq A_{\pi,k}} \|\Delta^{[A',B']}\|_{\mathrm{op}}$.

[1] Note that for this subsection, k is the number of free (unpinned) spins, unlike in Section 4.1.

Let k_0 be a constant to be chosen later. Then, by Lemma 4.3 and Theorem 4.3

$$
\begin{aligned}
&\max \gamma_T^{-1} \equiv \sup_{\sigma,i} \gamma_{\sigma,i}^{-1}\\
&= \exp(\beta k_0) \sup_{\sigma,i} \mathbb{E}_\pi \left[\prod_{N \geq k > k_0} \left(1 - \frac{1 + BT_{\sigma_{A^c_{\pi,k}}}}{k} \right)^{-1} \right]\\
&\leq \exp(\beta k_0) \sup_{\sigma,i} \mathbb{E}_\pi \left[\exp \left(\sum_{N \geq k > k_0} \frac{1}{k} + \frac{B \sum_{N\geq k>k_0} \sup_{A',B' \subset A_k} \sup_{\sigma' \in \{\pm 1\}^{A_k \setminus B'}} \|\Delta^{[A',B']}\|_{\mathrm{op}}}{k} \right) \right]\\
&\leq N \exp(\beta k_0 + C)
\end{aligned}
$$

where the last inequality holds with probability $\geq 1 - \exp(\Theta(N))$ by Proposition 4.1, and k_0 and C are chosen according to Proposition 4.1. ∎

Finally, we prove Proposition 4.1.

Proof. [Proof of Proposition 4.1] Throughout we fix an index $I \in [N]$ according to Proposition 4.1 and let A_k denote a random set of size k which has the distribution of I together with a uniformly random subset of $[N] \setminus \{I\}$ of size $k-1$. We also fix $\sigma \in \{\pm 1\}^N/\sqrt{N}$.

Recall the definition of the matrix $\Delta^{[A',B']}$ from Eq. (4.3). Note that $g_{\{i,j,S,S'\}}$ is a sum of $O\left(\binom{p}{2}\binom{p-2}{s'}\right)$ independent standard Gaussians (the union $S \cup S' \cup \{i,j\}$ has size at most p; given this union, we can uniquely recover S as the part of the union belonging to A_k^c; we can then choose the indices i, j from the remaining elements, and together with size constraints, this determines S' up to constantly many choices. Finally, there are at most $\binom{p-2}{s'}$ ways to choose the positions of the indices corresponding to S'). We write (suppressing the dependence on σ, σ'):

$$\Delta^{[A',B']} = \Delta_0^{[A',B']} + \Delta_1^{[A',B']},$$

where $\Delta_0^{[A',B']}$ includes the terms with $s' = 0$ and $\Delta_1^{[A',B']}$ consists of the remaining terms (i.e. those with $s' \geq 1$). We also denote $\Delta_{0,p}^{[A',B']}$ and $\Delta_{1,p}^{[A',B']}$ to denote the parts of these matrices stratified by $p \geq 2$.

For notational convenience, given $p \geq 2$, $\sigma \in \{\pm 1\}^N/\sqrt{N}$, $S' \subset A_k \setminus (A' \cup B')$ of size $s' \geq 0$ and $i, j \in A_k \setminus B'$, we define

$$X_{\sigma,p}^{i,j,S'} := \sum_{S \subset A_k^c} \frac{1}{\sqrt{N}} g_{\{i,j,S',S\}} \sigma_S.$$

Note that these are independent (over the choice of S', i, j) Gaussians with variance $O\left(p^2\binom{p-2}{s'}N^{-1}\right)$.

Claim 4.1. *With notation as above,*

$$\mathbb{P}_g \left[\forall p, \sup_{\sigma, A_k, A', B', S', \sigma'} \|\Delta_{1,p}^{[A',B']}\|_{\mathrm{op}} \geq C\beta_p \sqrt{\log p} \sum_{1 \leq s' \leq p-2} \binom{p}{s'}^{1/2} (k/N)^{s'/2} \right] \leq \exp(-1000N).$$

Proof. The proof of this claim is standard. Note that the $(i,j)^{th}$ entry of $\beta_p^{-1}\Delta_{1,p}^{[A',B']}$ is $\sum_{S' \subseteq A_k \setminus B'} X_\sigma^{i,j,S'} \sigma_{S'}$, which is a Gaussian of mean 0 and variance $O(p^2\binom{p-2}{s'}N^{-1}k^{s'}/N^{s'})$. Moreover, the entries of the matrix are independent. Therefore, by the concentration of the norm of random subgaussian matrices (e.g. [Ver18, Theorem 4.4.5]), we have that

$$\|\Delta_{1,p}^{[A',B']}\|_{\mathrm{op}} \gtrsim \beta_p p \sqrt{\log p} \sum_{1 \leq s' \leq p-2} \binom{p}{s'}^{1/2} (k/N)^{s'/2}$$

with probability at most $\exp(-100N \log p)$. We can then comfortably take the union bound over all choices appearing in the supremum. ∎

Furthermore, observe that

$$\sum_{p\geq 2}\sum_{k=k_0}^{N}\frac{1}{k}\beta_p p\sqrt{\log p}\sum_{1\leq s'\leq p-2}\binom{p}{s'}^{1/2}(k/N)^{s'/2}\leq \sum_{p\geq 2}\beta_p\sqrt{p^3\log p}\,2^{p/2},$$

which shows that on the event in Claim 4.1, for all choices of σ, I and for all permutations π,

$$\exp\left(\frac{B\sum_{N\geq k>k_0}\sup_{A',B'\subset A_k}\sup_{\sigma'\in\{\pm1\}^{A_k\setminus B'}}\|\Delta_1^{[A',B']}\|_{\mathrm{op}}}{k}\right)\lesssim 1. \tag{4.5}$$

It therefore remains to prove Proposition 4.1 with Δ replaced by Δ_0. Note that so far, in our analysis of $\Delta_1^{[A',B']}$, we did not need to use the randomness in the choice of A_k. This will be crucial in controlling $\|\Delta_0^{[A',B']}\|_{\mathrm{op}}$. Observe that $\Delta_0^{[A',B']}$ does not depend on σ'. Given σ and I, we say that A_k is *bad* if $\sup_p\sup_{A',B'\subseteq A_k}\|\Delta_{0,p}^{[A',B']}\|_{\mathrm{op}}/(\beta_p p\sqrt{\log p})\geq C(k/N)^{1/2-\alpha}$, where $\alpha<1/2$ is a constant (for instance, $\alpha=1/4$ is sufficient for us).

Claim 4.2. *For any σ and I, the probability that the number of bad A_k exceeds $\binom{N-1}{k-1}\cdot(k/N)^{\alpha}$ is at most $\exp(-1000N(N/k)^{\alpha})$. Hence, with probability at least $1-\exp(-900N(N/k)^{\alpha})$, simultaneously for all σ and I, the number of bad A_k is at most $\binom{N-1}{k-1}\cdot(k/N)^{\alpha}$.*

Proof. In proving this claim, we may assume without loss of generality that $k-1$ divides $N-1$. Then, to any permutation π of $[N-1]$, we naturally associate $(N-1)/(k-1)$ disjoint sets of size $k-1$. Together with I, these give $(N-1)/(k-1)$ sets of size k: $T_1,\dots,T_{(N-1)/(k-1)}$. By symmetry and double counting, it suffices to show that for any fixed permutation, with probability at least $1-\exp(-1000N\log(N/k)\log p)$, the fraction of $T_1,\dots,T_{(N-1)/(k-1)}$ which are bad is at most $(k/N)^{\alpha}$. By [Ver18, Theorem 4.4.5] and a union bound, the probability that a fixed T_i is bad is at most $\exp(-Ck(N/k)^{2\alpha})$. Since the Gaussians appearing in the matrices corresponding to distinct T_i are different (and in particular, independent), it follows that the probability that the number of bad T_i exceeds $(N/k)^{1-\alpha}$ is at most $\exp(-CN(N/k)^{\alpha})$. The final assertion follows by a union bound. □

The key claim in the proof of Proposition 4.1 is the following.

Claim 4.3. *With probability at least $1-\exp(-100N)$ over the choice of the Hamiltonian $H=H(g)$,*

$$\sup_{I,\sigma}\mathbb{E}_{\pi}\left[\underbrace{\exp\left(\frac{\sum_{k=k_0}^{N}\sup_{A',B'}\|\Delta_0^{[A',B']}\|_{\mathrm{op}}}{k}\right)}_{:=Z}\right]\lesssim 1.$$

Proof. We consider a 'good' realisation of the Hamiltonian i.e. a realisation satisfying (i) the conclusion of Claim 4.2 and (ii) $\sup_{I,\sigma,A_k,A',B'}\|\sum_{p\geq 2}\Delta_{0,p}^{[A',B']}\|\leq c$, for a sufficiently small constant c to be chosen later. By Claim 4.2 and [Adh+22, Lemma 6.1], this indeed holds with probability at least $1-\exp(-100N)$, provided that $\beta_0:=\sum_{p\geq 2}\sqrt{p^3\log p}\beta_p$ is sufficiently small depending on c. We show that for a good realisation, the conclusion of Claim 4.3 is satisfied.

To every permutation π, we naturally associate a sequence of sets $A_N,\dots,A_k,\dots,A_{k_0}$. Let N_k denote the operator norm of the submatrix corresponding to A_k. Let $\mathcal{B}_k$ denote the event that A_k is bad and $\mathcal{G}_k$ denote the complementary event that A_k is good. We write the random variable Z as

$$Z=\exp\left(\sum_{k=k_0}^{N}\frac{N_k\cdot 1_{\mathcal{G}_k}+N_k\cdot 1_{\mathcal{B}_k}}{k}\right).$$

Deterministically,

$$\sum_{k=k_0}^{N}\frac{N_k\cdot 1_{\mathcal{G}_k}}{k}\lesssim\left(\sum_{p\geq 2}\beta_p p\sqrt{\log p}\right)\sum_{k=k_0}^{N}k^{-1}(k/N)^{1/2-\alpha}\lesssim 1,$$

so it suffices to show that

$$\sup_{\sigma,I} \mathbb{E}_\pi \left[\exp \left(\sum_{k=k_0}^{N} \frac{N_k \cdot 1_{\mathcal{B}_k}}{k} \right) \right] \lesssim 1. \tag{4.6}$$

This follows from Hölder's inequality. Indeed, let $q_k = \gamma k^2$, where γ is an absolute constant ensuring that $\sum_{k=k_0}^{N} q_k^{-1} = 1$. By Claim 4.2, $\mathbb{E}_{A_k}[1_{\mathcal{B}_k}] \leq (k/N)^\alpha$ and we can always bound $N_k \leq c$ for a good realisation. Hence, we have

$$\mathbb{E}_{A_k} \left[\exp \left(q_k \cdot \frac{N_k \cdot 1_{\mathcal{B}_k}}{k} \right) \right] \leq \exp(cq_k/k)(k/N)^\alpha,$$

thus, using Hölder's inequality i.e. $\mathbb{E}[r_{k_0} \ldots r_N] \leq \mathbb{E}[r_{k_0}^{q_{k_0}}]^{\frac{1}{q_{k_0}}} \ldots \mathbb{E}[r_N^{q_N}]^{\frac{1}{q_N}}$ for $r_k = \exp(N_k 1_{\mathcal{B}_k}/k)$, we can bound the left hand side of Eq. (4.6) by

$$\prod_{k=k_0}^{N} \exp(c/k) \prod_{k=k_0}^{N} (k/N)^{\alpha/q_k} \lesssim N^c N^{-\alpha} \lesssim 1,$$

provided $\beta_0 = \sum_{p \geq 2} \sqrt{p^3 \log p}\beta_p$ is small enough so that $c \leq \alpha(= 1/4)$. □

Combining Eq. (4.5) and Claim 4.3 finishes the proof.
□

5 Applications

5.1 Spectral Independence from a Contractive Coupling As a quick application of Theorem 3.1, we provide a short and transparent alternate proof of recent results of Liu [Liu21] and Blanca et al. [Bla+21] showing, for instance, that for a measure μ on a product space, the existence of a contractive coupling for the Glauber dynamics implies that μ is spectrally independent.

More precisely, let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ be a product space, let d be a metric on Ω, and let μ be a measure on Ω. For $\kappa \in (0, 1)$, we say that μ satisfies property (†) with parameter κ with respect to the metric d if the following holds.

(†) For all valid configurations $(X_0, Y_0) \in \Omega \times \Omega$, there exists a coupling $(X_0, Y_0) \to (X_1, Y_1)$ of the Glauber dynamics P satisfying

$$\mathbb{E}[d(X_1, Y_1) \mid (X_0, Y_0)] \leq \kappa d(X_0, Y_0).$$

Theorem 5.1. (cf. Theorem 1.10 in [Bla+21]) *If μ satisfies property (†) with parameter $\kappa \leq (1 - \epsilon/n)$ with respect to any metric d, then μ is η-spectrally independent with constant $\eta = 1/\epsilon$.*

Proof. Since $\kappa \leq (1 - \epsilon/n)$, it follows by [LP17, Theorem 13.1] that

$$\gamma \geq 1 - \kappa \geq \frac{\epsilon}{n}.$$

and finally, applying Theorem 3.1, we conclude that μ is $1(/\epsilon)$-spectrally independent. □

Remark 5.1. *(1) There are two important advantages of the above theorem, compared to [Bla+21, Theorem 1.10]. First, our bound on η depends only on the contractive constant κ and not on the underlying metric (as in [Bla+21]). Second, even in the case of weighted Hamming metrics, our bound on η is a factor of 2 better than the corresponding bound in [Bla+21]. This factor of 2 is important in applications where $\epsilon \approx 1$, in which case our spectral independence bound can potentially be combined with local-to-global arguments to yield, e.g. modified log-Sobolev inequalities, with near-optimal dependence on n, whereas the bound of [Bla+21] will necessarily give a quadratically sub-optimal bound.*

(2) While the above theorem is stated only for the Glauber dynamics for simplicity, the same proof can be used to cover situations where (a) the contractive coupling is with respect to a different Markov chain and (b)

the spectral gaps of this Markov chain is within a constant factor of the spectral gap of the Glauber dynamics. Indeed, this is the situation for the so-called flip dynamics for the uniform distribution on q-colorings of graphs of maximum degree Δ (with $q > (11/6 - \epsilon_0)\Delta$ for a small absolute constant ϵ_0), and therefore, allows us to easily recover one of the main applications of [Liu21; Bla+21].

(3) While the approaches in [Bla+21; Liu21] generally lead to looser spectral independence guarantees than ours, their methods can still be useful as a way to bound the ∞-norm of the influence matrix.

5.2 Learning and Statistical Estimation There is a vast literature on learning exponential families and graphical models from data. In terms of learning these distributions (say in the total variation distance), there are easy to use guarantees known *information-theoretically* via computationally inefficient algorithms, e.g. [DMR20]. There are also guarantees for learning via polynomial time algorithms, e.g. the result of [KM17]. In some settings, the computationally efficient guarantees are extremely suboptimal in terms of their sample complexity. Is this an inherent limitation?

An example where existing results fail: spin glass inversion. Suppose we are given m i.i.d. samples from a Sherrington-Kirkpatrick model at inverse temperature $\beta > 0$ on n sites. As a reminder, this is the special case of the mixed p-spin model with quadratic interactions, so $\mu(x) \propto \exp(\frac{1}{2}\langle x, Jx\rangle)$ where J is a symmetric matrix with $J_{ij} \sim N(0, \beta^2/n)$, so J is proportional to a GOE random matrix. The computationally inefficient result of [DMR20] implies that obtaining total variation distance 0.01 can be done with high probability from $m = \tilde{O}(n^2)$ samples.

Applying a state of the art algorithmic result such as [WSD19; KM17; Vuf+16], the best we can get is a sample complexity of $e^{O(\beta\sqrt{n})}$ for obtaining guarantees for learning this distribution in total variation distance. The reason is that the ℓ_1 norm of a row of J is approximately $\beta\sqrt{n}$, and these results depend exponentially on this quantity (or worse, on the degree) — see e.g. Theorem 7.3 of [KM17]. This is a fundamental limitation of the analyses in all of these works.

Remark 5.2. *The SK model was prominently studied in a different context in statistical estimation [Cha07], where J is known up to normalization and the goal was to estimate a single parameter (the inverse temperature β) from a single sample. Here we are trying to estimate the entire distribution which requires many more samples from the distribution. Our problem has been considered under the name of* spin glass inversion *in the statistical physics literature [MV09; MM09], where heuristic message passing algorithms have been proposed.*

A different analysis via approximate tensorization. Many algorithms in the literature on learning discrete graphical models, including those referenced above, can be understood as variants of maximum likelihood (see e.g. [Van00]) or *pseudolikelihood* estimation [Bes75]. Pseudolikelihood estimation minimizes the loss

$$\hat{L}_p(q) := \sum_{j=1}^{n} \frac{1}{m} \sum_{i=1}^{m} \widehat{\mathbb{E}}_X[\log q((X_i)_j \mid (X_i)_{\sim j})]$$

where $X_1, \ldots, X_m$ are i.i.d. samples from the ground truth distribution p and $(X_i)_j$ denotes coordinate j of sample X_i. It was recently observed that approximate tensorization of entropy has immediate consequences for the sample complexity of pseudolikelihood estimation [KLR22].

Theorem 5.2. (Special case of Theorem 4 of [KHR22]) *Suppose that $\mathcal{P}$ is a class of probability distributions containing p and $C_{AT}(\mathcal{P}) := \sup_{q\in\mathcal{P}} C_{AT}(q)$ is the worst-case approximate tensorization constant in the class of distributions. Let*

$$\mathcal{R}_m := \mathbb{E}_{X_1,\ldots,X_m,\epsilon_1,\ldots,\epsilon_m}\left[\sup_{q\in\mathcal{P}} \frac{1}{m}\sum_{i=1}^{m} \epsilon_i \left[\sum_{j=1}^{n} \log q((X_i)_j \mid (X_i)_{\sim j})\right]\right]$$

be the expected Rademacher complexity of the class given m samples $X_1, \ldots, X_m \sim p$ i.i.d. and independent $\epsilon_1, \ldots, \epsilon_m \sim Uni\{\pm 1\}$ i.i.d. Rademacher random variables. Let $\hat{p}$ be the pseudolikelihood estimator from n i.i.d. samples from p, in other words let $\hat{p} = \arg\min_{q\in`\mathcal{P}} \hat{L}_p(q)$. Then

$$\mathbb{E}[\mathcal{D}_{\mathrm{KL}}(p \parallel \hat{p})] \le 2C_{AT}(\mathcal{P})\mathcal{R}_m.$$

Since in this work we established bounds on the approximate tensorization constant of a large class of distributions, they can be directly combined with this result to obtain new learning guarantees. The precise choice of the class $\mathcal{P}$ will depend on the application (larger classes $\mathcal{P}$ will require more sample complexity to learn).

Exponential improvement in the example. We revisit the example of learning distributions like the SK model from data. We observe that when β is small enough so that approximate tensorization is satisfied, we get a dramatically improved guarantee for learning the distribution from samples:

THEOREM 5.3. *Suppose p is a distribution lying in $\mathcal{P}$, which is defined to be the class of distributions of the form*

$$p_{J,h}(x) \propto \exp\left(\frac{1}{2}\langle x, Jx\rangle + \langle h, x\rangle\right)$$

under the assumption that for some $R > 0$:

1. *$\|J\|_{OP} \le \alpha < A$ where $A > 0$ is the constant from Theorem 4.1.*
2. *$h_j \le R$ for every j, and $\|J_j\|_1 \le R$ for every row J_j.*

Let Θ denote the set of (J, h) pairs satisfying these conditions. Let $\hat{p}$ be the pseudolikelihood estimator from n i.i.d. samples from p, in other words let $\hat{p} = \arg\min_{q\in{}`\mathcal{P}} \hat{L}_p(q)$. (This is a convex program which can be efficiently optimized.) Then

$$\mathbb{E}[\mathcal{D}_{\mathrm{KL}}(p \parallel \hat{p})] = O\left(Rn^{1+B\alpha}\sqrt{\log(n)/m}\right)$$

where B is as defined in Theorem 4.1.

Proof. First, observe that the conditional density is

$$p_{J,h}(x_j \mid x_{\sim j}) = \frac{e^{\langle J_j, x\rangle x_j + h_j x_j}}{e^{\langle J_j, x\rangle x_j + h_j x_j} + e^{-\langle J_j, x\rangle x_j - h_j x_j}} = \frac{1}{1 + e^{-2\langle J_j, x\rangle x_j - 2h_j x_j}}$$

where J_j denotes row j of the matrix J. So

$$\log p_{J,h}(x_j \mid x_{\sim j}) = -\ell(-\langle J_j x\rangle x_j - h_j x_j)$$

where $\ell(z) = \log(1 + e^{2z})$ is the logistic loss, which is a 1-Lipschitz function. We can now bound the Rademacher complexity of this class:

$$\begin{aligned}
\mathcal{R}_n &= \mathbb{E}_{X_1,\dots,X_n,\epsilon_1,\dots,\epsilon_n}\left[\sup_{J,h\in\Theta} \frac{1}{m}\sum_{i=1}^m \epsilon_i \left[\sum_{j=1}^n \ell(-(X_i)_j(\langle J_j, X_i\rangle + h_j))\right]\right] \\
&\le \sum_{j=1}^n \mathbb{E}_{X_1,\dots,X_n,\epsilon_1,\dots,\epsilon_n}\left[\sup_{J,h\in\Theta} \frac{1}{m}\sum_{i=1}^m \epsilon_i \ell(-(X_i)_j(\langle J_j, X_i\rangle + h_j))\right] \\
&\le \sum_{j=1}^n \mathbb{E}_{X_1,\dots,X_n,\epsilon_1,\dots,\epsilon_n}\left[\sup_{J,h\in\Theta} \frac{1}{m}\sum_{i=1}^m \epsilon_i (X_i)_j(\langle J_j, X_i\rangle + h_j)\right] \\
&= \sum_{j=1}^n \mathbb{E}_{X_1,\dots,X_n,\epsilon_1,\dots,\epsilon_n}\left[\sup_{J,h\in\Theta} \frac{1}{m}\langle J_j, \sum_{i=1}^m \epsilon_i X_i (X_i)_j\rangle + \frac{h_j}{m}\sum_{i=1}^m \epsilon_i (X_i)_j\right] \\
&\le \sum_{j=1}^n \mathbb{E}_{X_1,\dots,X_n,\epsilon_1,\dots,\epsilon_n}\left[\frac{1}{m}R\left\|\sum_{i=1}^m \epsilon_i X_i (X_i)_j\right\|_\infty + \frac{R}{m}\left|\sum_{i=1}^m \epsilon_i (X_i)_j\right|\right]
\end{aligned}$$

where in the first inequality we moved the supremum inside the sum over j, in the second inequality we used Talagrand's contraction principle (Exercise 6.7.7 of [Ver18]), and in the last inequality we used Holder's inequality. Using Hoeffding's inequality and a standard tail bound for the maximum of subgaussian random variables (Exercise 2.5.10 of [Ver18]), we conclude that

$$\mathcal{R}_n \le nR\sqrt{\frac{2\log n}{m}} + nR\sqrt{1/m} \le 4Rn\sqrt{\frac{\log n}{m}}.$$

Appealing to Theorem 5.2 and Theorem 4.2 proves the result. ▯

Going back to the example of the SK model, we have $R = \tilde{O}(\beta\sqrt{n})$ so for β sufficiently small this implies we can learn the distribution to KL divergence $o(1)$, hence also TV distance $o(1)$ by Pinsker's inequality [CT12], with $m = \tilde{O}(n^{3+2B\beta})$ samples. This is not much worse than the sample complexity required for the inefficient method of [DMR20] and exponentially improves the previously mentioned algorithmic guarantees. We remark that this analysis can be adapted to the estimators proposed in [KM17; WSD19; Vuf+16] and other works, since their algorithms are based on similar principles to, though not identical to, the standard pseudolikelihood estimator. The key difference is our analysis of the estimator.

This results also works for many other models, e.g. diluted spin glasses, with similar improvements (exponential improvement in the dependence on the degree d). Finally, we note that while the guarantee of Theorem 5.2 is in expectation, it is straightforward to obtain a strong high probability guarantee by combining the same argument with standard tools from generalization theory (see [KHR22; SB14]).

Remark 5.3. *Obviously, this result (and the identity testing result in the next section) can be generalized to the case of higher-order interactions. In the special case of the Ising model, it is possible to tweak the above and obtain a version of the result for a larger value of $\|J\|_{OP}$ by using the result of [EKZ21] and appealing to comparison inequalities to bound the log-Sobolev or approximate tensorization constant; this results in a worse dependence on the dimension and on the size of the external field.*

5.3 Identity Testing In the identity testing problem, given an explicitly visible distribution μ and oracle access to samples from an unknown/hidden distribution π, the goal is to determine if these distributions are identical using as few samples from π as possible. [Bla+22] shows efficient identity testing algorithms for distributions μ satisfying approximate tensorization of entropy.

Theorem 5.4. ([Bla+22, Theorem 7.5]) *Consider distribution $\mu : \{\pm 1\}^{[n]} \to \mathbb{R}_{\geq 0}$ satisfying approximate tensorization of entropy with parameter C and b-marginally bounded assumption i.e. if for every $\Lambda \subseteq [n]$, every $x \in \{\pm 1\}^{\Lambda}$ with $\mu(X^{\Lambda} = x) > 0$, every $i \in [n] \setminus \Lambda$, and every $a \in \{\pm 1\}$, one has*

$$\text{either } \mu(X_i = a | X^{\Lambda} = x) = 0 \text{ or } \mu(X_i = a | X^{\Lambda} = x) \geq b$$

with $b := b(n)$ satisfying $\log\log(b) = O(\log n)$.

Suppose that there is FPRAS to estimate the marginals of μ i.e. estimating $\mu_i(\cdot | x_{-i})$ for any $x_{-i} \in \{\pm 1\}^{[n]\setminus\{i\}}$. Given query access to the Subcube oracle, there exist an identity testing algorithm for KL divergence with distance parameter ϵ with query complexity

$$O\left(\log\left(\frac{1}{b}\right) n \log\left(\frac{n}{\epsilon}\right)\right)$$

This immediately implies efficient identity testing algorithm for μ satisfying the preconditions Theorem 5.3, since μ is $b(n) = \exp(-\theta(\alpha\sqrt{n} + R))$-marginally bounded by the same argument as in proof of Theorem 5.3.

Corollary 5.1. *Let p be the distribution satisfying the assumptions in Theorem 5.3. There exists identity testing algorithm for KL divergence with distance parameter ϵ with query complexity $O(n^{3/2+B\beta+R}\log(n/\epsilon))$ with B being a constant, the same as in Theorem 4.2.*

Acknowledgements. Frederic Koehler was supported in part by NSF award CCF-1704417, NSF award IIS-1908774, and Nima Anari's Sloan Research Fellowship. Vishesh Jain was supported by NSF CAREER award DMS-2237646. Nima Anari was supported by NSF CAREER award CCF-2045354.

References

[Adh+22] Arka Adhikari, Christian Brennecke, Changji Xu, and Horng-Tzer Yau. *Spectral Gap Estimates for Mixed p-Spin Models at High Temperature.* 2022. DOI: 10.48550/ARXIV.2208.07844.

[AF95] David Aldous and James Fill. *Reversible Markov chains and random walks on graphs.* 1995.

[AL20] 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.* 2020, pp. 1198–1211.

[ALG22] Dorna Abdolazimi, Kuikui Liu, and Shayan Oveis Gharan. "A matrix trickle-down theorem on simplicial complexes and applications to sampling colorings". In: *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2022, pp. 161–172.

[Ali+21] 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*. STOC 2021. Virtual, Italy: Association for Computing Machinery, 2021, pp. 433–446. ISBN: 9781450380539. DOI: 10.1145/3406325.3451123.

[ALO20] Nima Anari, Kuikui Liu, and Shayan Oveis Gharan. "Spectral Independence in High-Dimensional Expanders and Applications to the Hardcore Model". In: *Proceedings of the 61st IEEE Annual Symposium on Foundations of Computer Science*. IEEE Computer Society, 2020.

[ALV22] Nima Anari, Yang P. Liu, and Thuy-Duong Vuong. *Optimal Sublinear Sampling of Spanning Trees and Determinantal Point Processes via Average-Case Entropic Independence*. 2022. DOI: 10.48550/ARXIV.2204.02570.

[AMS22] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. "Sampling from the Sherrington-Kirkpatrick Gibbs measure via algorithmic stochastic localization". In: *arXiv preprint arXiv:2203.05093* (2022).

[Ana+19] Nima Anari, Kuikui Liu, Shayan Oveis Gharan, and Cynthia Vinzant. "Log-concave polynomials II: high-dimensional walks and an FPRAS for counting bases of a matroid". In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 2019, pp. 1–12.

[Ana+21a] 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". In: *arXiv preprint arXiv:2106.04105* (2021).

[Ana+21b] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. "Entropic independence II: optimal sampling and concentration via restricted modified log-Sobolev inequalities". In: *arXiv preprint arXiv:2111.03247* (2021).

[Bar+11] Franck Barthe, Dario Cordero-Erausquin, Michel Ledoux, and Bernard Maurey. "Correlation and Brascamp–Lieb inequalities for Markov semigroups". In: *International Mathematics Research Notices* 2011.10 (2011), pp. 2177–2216.

[BB19] Roland Bauerschmidt and Thierry Bodineau. "A very simple proof of the LSI for high temperature spin systems". In: *Journal of Functional Analysis* 276.8 (2019), pp. 2582–2588.

[Bes75] Julian Besag. "Statistical analysis of non-lattice data". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 24.3 (1975), pp. 179–195.

[BGL+14] Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*. Vol. 103. Springer, 2014.

[Bla+21] 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: *arXiv preprint arXiv:2103.07459* (2021).

[Bla+22] Antonio Blanca, Zongchen Chen, Daniel Štefankovič, and Eric Vigoda. *Identity Testing for High-Dimensional Distributions via Entropy Tensorization*. 2022. DOI: 10.48550/ARXIV.2207.09102.

[BT06] Sergey G Bobkov and Prasad Tetali. "Modified logarithmic Sobolev inequalities in discrete settings". In: *Journal of Theoretical Probability* 19.2 (2006), pp. 289–336.

[CDP09] Pietro Caputo, Paolo Dai Pra, and Gustavo Posta. "Convex entropy decay via the Bochner-Bakry-Emery approach". In: *Annales de l'IHP Probabilités et statistiques*. Vol. 45. 3. 2009, pp. 734–753.

[CE22] Yuansi Chen and Ronen Eldan. "Localization schemes: A framework for proving mixing bounds for Markov chains". In: *arXiv preprint arXiv:2203.04163* (2022).

[Cel22] Michael Celentano. "Sudakov-Fernique post-AMP, and a new proof of the local convexity of the TAP free energy". In: *arXiv preprint arXiv:2208.09550* (2022).

[CGM19] 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)*. IEEE. 2019, pp. 1358–1370.

[Cha07] Sourav Chatterjee. "Estimation in spin glasses: A first step". In: *The Annals of Statistics* 35.5 (2007), pp. 1931–1946.

[Che+21] Xiaoyu Chen, Weiming Feng, Yitong Yin, and Xinyuan Zhang. *Rapid mixing of Glauber dynamics via spectral independence for all degrees.* 2021. arXiv: `2105.15005 [cs.DS]`.

[Che+22] Xiaoyu Chen, Weiming Feng, Yitong Yin, and Xinyuan Zhang. "Optimal mixing for two-state anti-ferromagnetic spin systems". In: *arXiv preprint arXiv:2203.07771* (2022).

[CLV20] 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)*. IEEE. 2020, pp. 1307–1318.

[CLV21] 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.* 2021, pp. 1537–1550.

[CLV22] Zongchen Chen, Kuikui Liu, and Eric Vigoda. "Spectral independence via stability and applications to holant-type problems". In: *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2022, pp. 149–160.

[CMT14] Pietro Caputo, Georg Menz, and Prasad Tetali. *Approximate tensorization of entropy at high temperature.* 2014. DOI: `10.48550/ARXIV.1405.0608`.

[CMT15] Pietro Caputo, Georg Menz, and Prasad Tetali. "Approximate tensorization of entropy at high temperature". In: *Annales de la Faculté des sciences de Toulouse: Mathématiques*. Vol. 24. 4. 2015, pp. 691–716.

[CT12] Thomas M Cover and Joy A Thomas. *Elements of Information Theory.* John Wiley & Sons, 2012.

[DMR20] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. "The minimax learning rates of normal and Ising undirected graphical models". In: *Electronic Journal of Statistics* 14.1 (2020), pp. 2338–2361.

[Dob68] Roland L'vovich Dobrushin. "The problem of uniqueness of a Gibbsian random field and the problem of phase transitions". In: *Functional Analysis and its Applications* 2.4 (1968), pp. 302–312.

[EG18] Ronen Eldan and Renan Gross. "Decomposition of mean-field Gibbs distributions into product measures". In: *Electronic Journal of Probability* 23 (2018), pp. 1–24.

[EKZ21] Ronen Eldan, Frederic Koehler, and Ofer Zeitouni. "A spectral condition for spectral gap: fast mixing in high-temperature Ising models". In: *Probability Theory and Related Fields* (2021), pp. 1–17.

[ELL17] Ronen Eldan, James R Lee, and Joseph Lehec. "Transport-entropy inequalities and curvature in discrete-space Markov chains". In: *A journey through discrete mathematics*. Springer, 2017, pp. 391–406.

[Erb+17] Matthias Erbar, Christopher Henderson, Georg Menz, and Prasad Tetali. "Ricci curvature bounds for weakly interacting Markov chains". In: *Electronic Journal of Probability* 22 (2017), pp. 1–23.

[ES22] Ronen Eldan and Omer Shamir. "Log concavity and concentration of Lipschitz functions on the Boolean hypercube". In: *Journal of Functional Analysis* 282.8 (2022), p. 109392.

[GJ19] Reza Gheissari and Aukosh Jagannath. "On the spectral gap of spherical spin glass dynamics". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*. Vol. 55. 2. Institut Henri Poincaré. 2019, pp. 756–776.

[Gro14] Leonard Gross. "Hypercontractivity, logarithmic Sobolev inequalities, and applications: a survey of surveys". In: *Diffusion, quantum theory, and radically elementary mathematics* 47 (2014), pp. 45–73.

[GZ03] Alice Guionnet and Bogusław Zegarlinksi. "Lectures on logarithmic Sobolev inequalities". In: *Séminaire de probabilités XXXVI*. Springer, 2003, pp. 1–134.

[KHR22] Frederic Koehler, Alexander Heckett, and Andrej Risteski. "Statistical Efficiency of Score Matching: The View from Isoperimetry". In: *arXiv preprint arXiv:2210.00726* (2022).

[KLR22] Frederic Koehler, Holden Lee, and Andrej Risteski. "Sampling Approximately Low-Rank Ising Models: MCMC meets Variational Methods". In: *arXiv preprint arXiv:2202.08907* (2022).

[KM17] Adam Klivans and Raghu Meka. "Learning graphical models using multiplicative weights". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 343–354.

[KO18] Tali Kaufman and Izhar Oppenheim. "High order random walks: Beyond spectral gap". In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2018.

[Liu21] Kuikui Liu. "From coupling to spectral independence and blackbox comparison with the down-up walk". In: *arXiv preprint arXiv:2103.11609* (2021).

[LP17] David A Levin and Yuval Peres. *Markov chains and mixing times*. Vol. 107. American Mathematical Soc., 2017.

[Mar15] Katalin Marton. "Logarithmic Sobolev inequalities in discrete product spaces: a proof by a transportation cost distance". In: *arXiv preprint arXiv:1507.02803* (2015).

[Mar99] Fabio Martinelli. "Lectures on Glauber dynamics for discrete spin models". In: *Lectures on probability theory and statistics*. Springer, 1999, pp. 93–191.

[MM09] Marc Mézard and Thierry Mora. "Constraint satisfaction problems and neural networks: A statistical physics perspective". In: *Journal of Physiology-Paris* 103.1-2 (2009), pp. 107–113.

[MOS13] Elchanan Mossel, Krzysztof Oleszkiewicz, and Arnab Sen. "On reverse hypercontractivity". In: *Geometric and Functional Analysis* 23.3 (2013), pp. 1062–1097.

[MV09] Enzo Marinari and Valery Van Kerrebroeck. "Intrinsic limitations of inverse inference in the pairwise Ising spin glass". In: *arXiv preprint arXiv:0911.1985* (2009).

[ODo14] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[Oll09] Yann Ollivier. "Ricci curvature of Markov chains on metric spaces". In: *Journal of Functional Analysis* 256.3 (2009), pp. 810–864.

[Opp18] Izhar Oppenheim. "Local spectral expansion approach to high dimensional expanders part I: Descent of spectral gaps". In: *Discrete & Computational Geometry* 59.2 (2018), pp. 293–330.

[Pan13] Dmitry Panchenko. *The sherrington-kirkpatrick model*. Springer Science & Business Media, 2013.

[RV07] Mark Rudelson and Roman Vershynin. "Sampling from large matrices: An approach through geometric functional analysis". In: *Journal of the ACM (JACM)* 54.4 (2007), 21–es.

[SB14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[Tal10] Michel Talagrand. *Mean field models for spin glasses: Volume I: Basic examples*. Vol. 54. Springer Science & Business Media, 2010.

[Van00] Aad W Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.

[Van14] Ramon Van Handel. *Probability in high dimension*. Tech. rep. PRINCETON UNIV NJ, 2014.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.

[Vuf+16] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. "Interaction screening: Efficient and sample-optimal learning of ising models". In: *Advances in neural information processing systems* 29 (2016).

[WSD19] Shanshan Wu, Sujay Sanghavi, and Alexandros G Dimakis. "Sparse logistic regression learns all discrete pairwise graphical models". In: *Advances in Neural Information Processing Systems* 32 (2019).

[ZZ21] Zhixin Zhou and Yizhe Zhu. “Sparse random tensors: Concentration, regularization and applications”. In: *Electronic Journal of Statistics* 15.1 (2021), pp. 2483–2516.

A Miscellaneous Facts

Equivalent expressions for the Dirichlet form. Let

$$\nu(\sigma) \propto \exp(H(\sigma))$$

be a binary spin system on the hypercube $\{\pm 1\}^n$. Note that the conditional law at site j is

$$\nu(\sigma_j \mid \sigma_{\sim j}) = \frac{\exp(\sigma_j \partial_j H(\sigma))}{\exp(\sigma_j \partial_j H(\sigma)) + \exp(-\sigma_j \partial_j H(\sigma))} = \frac{1}{1 + \exp(-2\sigma_j \partial_j H(\sigma))}$$

Recall that the Dirichlet form for Glauber dynamics is

$$\mathcal{E}(f,f) := \sum_{\sigma \sim \tau} \frac{\nu(\sigma)\nu(\tau)}{\nu(\sigma)+\nu(\tau)}(f(\sigma)-f(\tau))^2 = \frac{1}{2}\sum_{\sigma}\nu(\sigma)\sum_{j=1}^{n}\frac{\nu(\hat{\sigma}_j)}{\nu(\sigma_j)+\nu(\hat{\sigma}_j)}(f(\sigma)-f(\hat{\sigma}_j))^2$$

where $\hat{\sigma}_j$ denotes σ with the spin at site j flipped. Note that

$$\frac{\nu(\hat{\sigma}_j)}{\nu(\sigma_j)+\nu(\hat{\sigma}_j)} = \frac{1}{\nu(\sigma_j)/\nu(\hat{\sigma}_j)+1} = \frac{1}{1+\exp(2\sigma_j\partial_j H(\sigma))}.$$

This lets us establish the following fact which shows consistency with the notation in [Adh+22]:

Lemma A.1. (Standard, see e.g. [Adh+22]) *For all functions f, we have*

$$\mathcal{E}(f,f) = \frac{1}{4}\sum_{\sigma}\nu(\sigma)\sum_{j=1}^{n}\cosh^{-2}(\partial_j H(\sigma))(f(\sigma)-f(\hat{\sigma}_j))^2$$

Proof. We have (using that $\cosh(\partial_j H) = \cosh(\sigma_j \partial_j H)$ by evenness)

$$\begin{aligned}
&\frac{1}{4}\sum_{\sigma}\nu(\sigma)\sum_{j=1}^{n}\cosh^{-2}(\partial_j H(\sigma))(f(\sigma)-f(\hat{\sigma}_j))^2\\
&= \sum_{\sigma}\nu(\sigma)\sum_{j=1}^{n}\frac{1}{(\exp(\sigma_j\partial_j H(\sigma))+\exp(-\sigma_j\partial_j H(\sigma)))^2}(f(\sigma)-f(\hat{\sigma}_j))^2\\
&= \frac{1}{2}\sum_{\sigma}\nu(\sigma)\sum_{j=1}^{n}\frac{1}{1+(\exp(2\sigma_j\partial_j H(\sigma))+\exp(-2\sigma_j\partial_j H(\sigma)))/2}(f(\sigma)-f(\hat{\sigma}_j))^2\\
&= \frac{1}{2}\sum_{\sigma}\sum_{j=1}^{n}\nu(\sigma_{\sim j})\frac{\nu(\sigma_j \mid \sigma_{\sim j})}{1+(\exp(2\sigma_j\partial_j H(\sigma))+\exp(-2\sigma_j\partial_j H(\sigma)))/2}(f(\sigma)-f(\hat{\sigma}_j))^2\\
&= \frac{1}{2}\sum_{\sigma}\sum_{j=1}^{n}\nu(\sigma_{\sim j})\frac{1}{2+\exp(2\sigma_j\partial_j H(\sigma))+\exp(-2\sigma_j\partial_j H(\sigma))}(f(\sigma)-f(\hat{\sigma}_j))^2
\end{aligned}$$

where in the last step we averaged over pairs of σ agreeing on $\sigma_{\sim j}$ and used that

$$\frac{\nu(\sigma_j = +1 \mid \sigma_{\sim j}) + \nu(\sigma_j = -1 \mid \sigma_{\sim j})}{2} = \frac{1}{2}.$$

On the other hand, the Dirichlet form is

$$\begin{aligned}\mathcal{E}(f,f) &= \frac{1}{2}\sum_{\sigma}\sum_{j=1}^{n}\nu(\sigma_{\sim j})\frac{\nu(\sigma_j \mid \sigma_{\sim j})}{1+\exp(2\sigma_j\partial_j H(\sigma))}(f(\sigma)-f(\hat{\sigma}_j))^2\\ &= \frac{1}{2}\sum_{\sigma}\sum_{j=1}^{n}\nu(\sigma_{\sim j})\frac{1}{(1+\exp(-2\sigma_j\partial_j H(\sigma)))(1+\exp(2\sigma_j\partial_j H(\sigma)))}(f(\sigma)-f(\hat{\sigma}_j))^2\\ &= \frac{1}{2}\sum_{\sigma}\sum_{j=1}^{n}\nu(\sigma_{\sim j})\frac{1}{2+\exp(-2\sigma_j\partial_j H(\sigma))+\exp(2\sigma_j\partial_j H(\sigma))}(f(\sigma)-f(\hat{\sigma}_j))^2\end{aligned}$$

so we established the desired equality. □

B Deferred Proofs from Section 4

Proof. [Proof of Lemma 4.3] Consider $A \subseteq [n]$ and a pinning σ_A, then the pinned subsystem on $[n] \setminus A$ is of the form

$$\mu_{\sigma_A}^{[A,\emptyset]}(\sigma) \propto \exp(H_{\sigma_A}^{[A,\emptyset]}(\sigma))$$

Let $H_A := H_{\sigma_A}^{[A,\emptyset]}$ and consider the multilinear extension of H_A:

$$H_A(x) = \sum_{S\subset[n]\setminus A}\hat{H}(S)\prod_{i\in S}\sigma_i$$

Let $\mu_0(\sigma) \propto \exp(\sum_{i\in A}\hat{H}_A(\{i\})\sigma_i)$, so that μ_0 is a product distribution and hence satisfies approximate entropy tensorization with $C = 1$. Let

$$H_{A,\geq 2}(\sigma) = \sum_{S\subset[n]\setminus A,|S|\geq 2}\hat{H}(S)\prod_{i\in S}\sigma_i,$$

so that $\mu(\sigma) \propto \exp(H_{A,\geq 2}(\sigma))\mu_0(x)$. Then,

$$|H_{A,\geq 2}(\sigma)| = |\sum_{S\subseteq A}\hat{H}(S)\prod_{i\in S}\sigma_i| = |\sigma^{\intercal}\nabla^2 H_{A,\geq 2}\sigma| \leq \|\nabla^2 H_{A,\geq 2}\|_{\mathrm{OP}}\|\sigma\|_2^2 = (N-k)\|\nabla^2 H_{A,\geq 2}\|_{\mathrm{OP}} \leq O(\beta)$$

by Lemma 4.2. The required assertion then follows from Lemma B.1 below. □

Lemma B.1. (Comparison theorem for approximate entropy tensorization) *Consider distributions μ and μ' over Ω satisfying $\mu(x) \propto \mu'(x)\exp(W(x))$. Let $\|W\|_\infty = \sup_x|W(x)|$. Then, for any function $f : \Omega \to (0,\infty)$,*

$$\mathrm{Ent}_\mu[f] \leq \exp(2\|W\|_\infty)\,\mathrm{Ent}_{\mu'}[f].$$

Consequently, if μ' satisfies approximate entropy tensorization with constant C, then μ satisfies approximate entropy tensorization with constant $\exp(6\|W\|_\infty)C$.

Proof. Let $Z_\mu = \int d\mu$ and $Z_{\mu'} = \int d\mu'$ be normalization constants of μ and μ' respectively. It is easy to see that

$$Z_{\mu'} = \int \exp(-W(x))d\mu(x) \leq \exp(\|W\|_\infty)Z_\mu.$$

Thus $\exp(-\|W\|_\infty)\mu'(x) \leq \mu(x) \leq \exp(\|W\|_\infty)\mu'(x)$.

By the Donsker-Varadhan theorem,

$$\mathrm{Ent}_\mu[f] = \inf_{t>0}\int(f\log f - f\log t - f + t)d\mu$$

where$f \log f - f \log t - f + t = f(-\log(t/f) + (t/f - 1)) \geq 0$ sine $\log x \leq x - 1$ for $x \in (0, \infty)$. Thus

$$\begin{aligned}
\mathrm{Ent}_\mu[f] &= \inf_{t>0} \int (f \log f - f \log t - f + t) d\mu(x) \\
&= \inf_{t>0} \int (f(x) \log f(x) - f(x) \log t - f(x) + t) \exp(W(x)) Z_\mu^{-1} Z_{\mu'} d\mu'(x) \\
&\leq \inf_{t>0} \int (f(x) \log f(x) - f(x) \log t - f(x) + t) \exp(2\|W\|_\infty) d\mu'(x) \\
&= \exp(2\|W\|_\infty) \inf_{t>0} \int (f(x) \log f(x) - f(x) \log t - f(x) + t) d\mu'(x) \\
&= \exp(2\|W\|_\infty) \, \mathrm{Ent}_{\mu'}[f]
\end{aligned}$$

Next, since μ' satisfies approximate entropy tensorization with constant C

$$\begin{aligned}
\mathrm{Ent}_\mu[f] &\leq \exp(2\|W\|_\infty) \, \mathrm{Ent}_{\mu'}[f] \leq \exp(2\|W\|_\infty) C \sum_i \mathbb{E}_{x_{-i} \sim \mu'} [\mathrm{Ent}_{\mu'_{|x_{-i}}}[f_{|x_{-i}}]] \\
&\leq \exp(4\|W\|_\infty) C \sum_i \mathbb{E}_{x_{-i} \sim \mu'} [\mathrm{Ent}_{\mu_{|x_{-i}}}[f_{|x_{-i}}]] \\
&\leq \exp(6\|W\|_\infty) C \sum_i \mathbb{E}_{x_{-i} \sim \mu} [\mathrm{Ent}_{\mu_{|x_{-i}}}[f_{|x_{-i}}]]
\end{aligned}$$

where in the penultimate inequality we use the first statement and the fact that

$$\exp(-2\|W\|_\infty) \leq \mu'_{|x_{-i}}(x_i)/\mu_{|x_{-i}}(x_i) = \frac{\mu'(x)}{\mu(x)} \cdot \frac{\mu(x_{-i})}{\mu'(x_{-i})} = \frac{\mu'(x)}{\mu(x)} \cdot \frac{\int_{y : y_{-i} = x_{-i}} \mu(y)}{\int_{y : y_{-i} = x_{-i}} \mu'(y)} \leq \exp(2\|W\|_\infty),$$

and in the last inequality, we just use $\mu'(x_{-i}) \leq \exp(2\|W\|_\infty)\mu(x)$. $\square$