Parallel Sampling from the Ising p-Spin Model

Nima Anari
Stanford University, CA, USA

Aniket Das
Stanford University, CA, USA

Alireza Haqi
Stanford University, CA, USA

Abstract

We study the parallel complexity of sampling from the high-temperature Ising mixed p-spin Gibbs measure, a canonical instance of a mean-field spin glass on the hypercube $\{\pm 1\}^n$. We propose two different algorithms for this problem, corresponding to two different regimes of accuracy.

Our first algorithm is a parallel implementation of a Markov chain known as block dynamics, combined with an approximate rejection sampling step that uses an Ising model in a novel way as a proposal distribution to approximate the quadratic interaction terms of the p-spin Hamiltonian. For any $\varepsilon > 0$, this algorithm runs in $n^{1/3}\mathrm{polylog}(n/\varepsilon)$ parallel time with $\mathrm{poly}(n/\varepsilon)$ work, and outputs a sample whose law is ε-close to the p-spin measure in total variation distance.

Our second algorithm uses Picard iterations to parallelize the Algorithmic Stochastic Localization (ASL) process of El Alaoui, Montanari, and Sellke (2025), and for any $\varepsilon > \varepsilon_n$, takes $\mathrm{polylog}(n/\varepsilon)$ parallel time and $\mathrm{poly}(n/\varepsilon)$ work to produce a sample that is ε-close to the p-spin measure in the normalized 2-Wasserstein metric. Here, $\varepsilon_n > 0$ is a threshold that goes to 0 as $n \to \infty$. Our result constitutes a doubly exponential improvement in the ε dependence of the runtime and an exponential improvement in the ε dependence of the total work when compared to naïve ASL, whose runtime scales as $\exp(\mathrm{poly}(1/\varepsilon))$.

2012 ACM Subject Classification Theory of computation → Parallel algorithms; Theory of computation → Random walks and Markov chains; Mathematics of computing → Probabilistic algorithms

Keywords and phrases spin glasses, parallel sampling, Glauber dynamics, stochastic localization

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2026.71

Category RANDOM

Related Version *Full Version*: https://arxiv.org/abs/2607.12348

Funding This work is supported by funds from NSF-2125511, NSF CCF-2045354, NSF MAI-2501597.

1 Introduction

Efficient sampling from a probability distribution which is specified only up to an unknown normalizing constant is a central algorithmic problem across computer science, statistical physics, and machine learning. Generally, given a finite state space Ω and a Hamiltonian $H : \Omega \to \mathbb{R}$, the target distribution is a Gibbs measure of the following form:

$$\mu(x) = \frac{1}{Z}\exp(H(x)), \qquad Z = \sum_{x\in\Omega}\exp(H(x)) \tag{1}$$

where the normalizing constant Z, also known as the partition function, is unknown and typically hard to compute. This setting includes classical problems in approximate counting [32] and randomized algorithms such as approximating the volume of convex bodies [15] and approximating matrix permanents [31], Markov-chain Monte Carlo methods in statistical physics [41], and posterior sampling in high-dimensional statistical models.

Our work studies the problem of sampling from the Gibbs measure of the Ising mixed p-spin model, which is a Gibbs measure on the hypercube $\{\pm 1\}^n$ whose Hamiltonian is a random polynomial with Gaussian coefficients.

$$H(x) = \sum_{p=2}^{P} \frac{\beta_p \sqrt{p!}}{n^{\frac{p-1}{2}}} \cdot \sum_{J \subseteq [n], |J|=p} G_J \prod_{k \in J} x_k + \langle x, h \rangle, \qquad \mu(x) = \frac{1}{Z} \exp(H(x)) \tag{2}$$

Here, $P \geq 2$ is an integer, the disorder $G_J \overset{i.i.d}{\sim} \mathcal{N}(0,1)$ for every choice of J, and h is an arbitrary external field independent of G. The p-spin Hamiltonian can be equivalently characterized as a Gaussian process on $\{\pm 1\}^n$ with $\mathbb{E}_G[H(x)] = \langle h, x \rangle$ and $\mathrm{Cov}[H(x)H(y)] = N\xi(\langle x,y \rangle / N)$, where $\xi(t) = \sum_{p=2}^{P} \beta_p^2 t^p$ denotes the mixture function. The mixed p-spin model is a canonical instance of a mean-field spin glass, which plays a central role in several areas of probability, mathematical physics, and average-case complexity [37, 46, 40, 38]. Notable special cases include the pure p-spin model where only one β_p is nonzero, and the Sherrington-Kirkpatrick (SK) model where $P = 2$.

We consider the problem of approximate sampling from the p-spin Gibbs measure in the quenched setting, where the disorder G is sampled once and then fixed, and the algorithm is tasked with approximately sampling from μ given random access to the disorder coefficients (which can be stored with $\mathrm{poly}(n)$ memory). Given an error parameter $\varepsilon > 0$, the goal is to design an algorithm that runs in time $\mathrm{poly}(n, 1/\varepsilon)$ for any realization of G, and produces an output distributed as $\hat{\mu}$ such that $\mathrm{dist}(\hat{\mu}, \mu) \leq \varepsilon$ for typical realizations of G. Here, dist denotes any appropriate metric between probability distributions.

Under this setting, [1] and [7] analyze Glauber dynamics [27], a canonical Markov chain for sampling from discrete distributions, which, at every step, resamples a randomly chosen coordinate according to its conditional marginal, while fixing the remaining coordinates. The authors identify a high-temperature regime for the p-spin model under which Glauber dynamics efficiently samples from μ. Specifically, they show that when $\sum_{p=2}^{P} \beta_p \sqrt{p^3 \ln(p)} \leq C$ for some small enough $C = \Theta(1)$, Glauber dynamics takes $O(n \ln(n/\varepsilon))$ time to output a sample distributed as $\hat{\mu}$ satisfying $d_{\mathrm{TV}}(\hat{\mu}, \mu) \leq \varepsilon$ with high probability over the randomness of G. Since the error parameter $\varepsilon > 0$ can be made as small as desired, we refer to such a guarantee as a *High Accuracy Guarantee*.

In another line of work, [19] propose a different approximate sampler for the p-spin model with a zero external field (i.e., $h = 0$ in Equation (2)) based on an approximate discretization of an Itô diffusion process known as Eldan's stochastic localization [20]. Their algorithm, named Algorithmic Stochastic Localization (ASL), offers a weaker sampling guarantee, albeit supporting a broader range of temperatures than that covered by the high-temperature condition defined above. Concretely, there exists an error threshold $\varepsilon_n = o_n(1)$ such that for any $\varepsilon > \varepsilon_n$, ASL runs in time $O(\mathrm{poly}(n) \exp(\mathrm{poly}(1/\varepsilon)))$ and with high probability over G, approximately samples from μ up to error ε in the normalized 2-Wasserstein metric. In other words, the output $\hat{x}$ of their algorithm can be coupled with some $x^\star \sim \mu$ such that $n^{-1}\,\mathbb{E}[\|\hat{x} - x^\star\|] \leq \varepsilon^2$. We call such a guarantee a *Low Accuracy Guarantee* for two key reasons. Firstly, for a fixed problem instance, the accuracy of ASL can only be certified above a threshold $\varepsilon > \varepsilon_n$. Secondly, their guarantees are presented in the normalized 2-Wasserstein metric, which is strictly weaker than the total variation metric for the range of ε tolerated by their algorithm.

A common drawback of both algorithms considered above is that they are inherently sequential. Indeed, Glauber dynamics sequentially updates one coordinate after another in a random order, while Algorithmic Stochastic Localization discretizes the trajectory of a continuous Markov process, which naturally evolves in a sequential fashion. Consequently, these algorithms fail to utilize the full potential of modern hardware, which is generally capable of massive parallelism, both in terms of parallel computation and parallel memory access. This motivates the central problem of our work:

Can we design an efficient approximate sampling algorithm for the Ising p-spin model that can take advantage of parallel hardware?

Specifically, we consider the problem of sampling from the p-spin Gibbs measure in the PRAM model of computation, where the algorithm is allowed to query the disorder G in multiple rounds, with each round comprising polynomially many synchronous queries that are executed in parallel. Given an error tolerance $\varepsilon > 0$, our goal is to design an algorithm that approximately samples from μ up to error ε in some suitable metric for typical realizations of G, while also optimizing for two criteria: *Parallel Runtime* or *Round Complexity*, which denotes the number of rounds of parallel queries made to G, and *Work*, which denotes the total number of queries made to G. Ideally, for every realization of G, we want the parallel runtime to scale at most sublinearly in n and ε^{-1}, and the work to scale at most polynomially in n and ε^{-1}.

Compared to the complexity of sequential sampling, the theory of parallel sampling for mean-field spin glasses is far less developed. To our knowledge, the only prior parallel algorithm for sampling from the p-spin Gibbs measure is the work of [34], which considers the p-spin model in the high-temperature regime and achieves a high-accuracy guarantee with a parallel runtime of $O(\sqrt{n} \cdot \mathrm{polylog}(n/\varepsilon))$ and $O(\mathrm{poly}(n/\varepsilon))$ work. Beyond this, for the special case of the high-temperature SK model (i.e., $P = 2$ and $\beta_2 < 1/4$), [11] show that the Restricted Gaussian Dynamics algorithm of [12] is a high-accuracy approximate parallel sampler with parallel runtime $O(\mathrm{polylog}(n/\varepsilon))$ and $O(\mathrm{poly}(n/\varepsilon))$ work.

1.1 Our Contributions

In this work, we develop two different algorithms for parallel sampling from the Ising mixed p-spin Gibbs measure. The first is a high-accuracy algorithm which samples from the p-spin model up to ε total variation error in $O(n^{1/3}\mathrm{polylog}(n/\varepsilon))$ parallel time and $O(\mathrm{poly}(n, \log(n/\varepsilon)))$ work. To our knowledge, this constitutes an improvement both in terms of parallel runtime and work compared to the previous best known result of [34], which exhibits a parallel runtime of $O(\sqrt{n}\mathrm{polylog}(n/\varepsilon))$ with $O(\mathrm{poly}(n/\varepsilon))$ work.

▶ **Theorem 1** (High-Accuracy Parallel p-spin Sampler, Informal Version of Theorem 7)**.** *Let μ denote the Ising p-spin Gibbs measure as defined in Equation (2). Assume the model is in the high-temperature regime, i.e., $\sum_{p=2}^{P} \beta_p \sqrt{p^3 \ln(p)} \le C$ for some $C = \Theta(1)$. Then, there exists a parallel sampling algorithm (Algorithm 1) satisfying the following for any $\varepsilon > 0$.*

- *The algorithm runs in parallel time $\widetilde{O}(n^{1/3} \ln(1/\varepsilon))$ with $O(\mathrm{poly}(n, \log(n/\varepsilon)))$ work.*
- *With probability at least $1 - \exp(-cn)$ over the randomness of the disorder G, the algorithm outputs $\hat{x} \sim \hat{\mu}$ such that $d_{\mathrm{TV}}(\hat{\mu}, \mu) \le \varepsilon$.*

Our second algorithm obtains a low-accuracy approximate sampling guarantee in the normalized 2-Wasserstein metric, and enjoys a parallel runtime of $O(\mathrm{polylog}(n/\varepsilon))$ with $O(\mathrm{poly}(n/\varepsilon))$ work. This significantly improves upon the $O(\mathrm{poly}(n) \cdot \exp(\mathrm{poly}(1/\varepsilon)))$ runtime of [19] and constitutes the first parallel algorithm for sampling from the p-spin model with $\mathrm{polylog}(n/\varepsilon)$ parallel runtime in any metric.

▶ **Theorem 2** (Low-Accuracy Parallel p-spin Sampler, Informal Version of Theorem 16)**.** *Let μ denote the Ising p-spin Gibbs measure as defined in Equation (2) with $h = 0$. Assume the model is in the high-temperature regime, i.e., $\sum_{p=2}^{P} \beta_p \sqrt{p^3 \ln(p)} \le C$ for some $C = \Theta(1)$. Then, there exists an $\varepsilon_n > 0$ such that $\varepsilon_n \to 0$ as $n \to \infty$, and a parallel sampling algorithm (Algorithm 4) which satisfies the following for any $\varepsilon > \varepsilon_n$:*

- *The algorithm runs in parallel time $O(\mathrm{polylog}(n/\varepsilon))$ with $O(\mathrm{poly}(n/\varepsilon))$ work.*
- *With probability at least $1 - o(1)$ over the randomness of the disorder G, the output $\hat{x}$ of the algorithm can be coupled to some $x^\star \sim \mu$ such that $n^{-1}\,\mathbb{E}[\|\hat{x} - x^\star\|^2] \le \varepsilon^2$.*

For the special case of the SK model, our low-accuracy algorithm satisfies the above guarantee for $\beta_2 \leq 0.375$. To the best of our knowledge, this constitutes the only parallel sampler for the SK model running in $O(\mathrm{polylog}(n/\varepsilon))$ time with $O(\mathrm{poly}(n/\varepsilon))$ work in this regime of inverse temperatures.

1.2 Related Work

1.2.1 Sampling from mean-field spin glasses

For the Sherrington–Kirkpatrick model ($P = 2$), classical mixing criteria such as Dobrushin's condition [14] imply $\widetilde{O}(n^2)$ mixing time of Glauber dynamics only up to the vanishing window $\beta_2 = O(n^{-1/2})$. This falls short of the dimension-free β_2 regime predicted in the statistical physics literature and leads to algorithmically unsatisfactory guarantees. The first major breakthrough was the work of [21] which used stochastic localization to prove that Glauber dynamics mixes in $\widetilde{O}(n^2)$ time up to $\beta_2 < 1/4$. The mixing time was subsequently sharpened to $\widetilde{O}(n)$ by [6] and [12], and the range of β_2 was improved to $\beta_2 < 0.295$ by [8]. For the Ising p-spin model, [1] proved an $\widetilde{O}(n^2)$ mixing time for Glauber dynamics in the high-temperature regime $\sum_{2\leq p\leq P} \beta_p \sqrt{p^3 \ln(p)} = O(1)$. The mixing time was later improved to $\widetilde{O}(n)$ by [7].

In a different direction, [17] (see also [10]) introduced Algorithmic Stochastic Localization (ASL) for the SK model with zero external field. For any $\varepsilon > \varepsilon_n$ for a fixed $\varepsilon_n = o_n(1)$, their algorithm runs in time $O(\mathrm{poly}(n)\exp(\mathrm{poly}(1/\varepsilon)))$ and obtains a low-accuracy guarantee: sampling up to ε-error in the normalized 2-Wasserstein distance. Their algorithm works up to $\beta_2 < 1$, which is conjectured to be the sharp threshold up to which the SK model admits a polynomial-time approximate sampler. Their framework was later extended to the Ising p-spin model in [19], obtaining a similar low-accuracy guarantee but covering a broader range of temperatures than that of [1] and [7].

There also exists a parallel line of work on the spherical p-spin model, where the Gibbs measure μ is supported on the sphere $\{x \in \mathbb{R}^n \mid \|x\|^2 = n\}$, a setting that is often more analytically tractable than the hypercube. In the high-temperature regime, [26] proved that a continuous-time diffusion process known as Langevin dynamics samples from the spherical p-spin model in total variation in $\widetilde{O}(1)$ time. Beyond the high-temperature regime considered in [26], the recent breakthrough of [30] designed an improved version of ASL for the spherical p-spin model with zero external field that samples up to $o(1)$ error in total variation in $O(\mathrm{poly}(n))$ time beyond the high-temperature regime considered by [26]. Building upon their techniques, [29] proved that continuous-time annealed Langevin dynamics samples up to total variation error $\exp(-\Omega(n^{-1/5}))$ in $O(\mathrm{poly}(n))$ time beyond the high-temperature regime.

1.2.2 Parallel sampling

Parallel sampling has been a longstanding topic of interest in theoretical computer science, and more recently, in machine learning, where diffusion-based generative models are often deployed on massively parallel hardware [33, 44, 28]. In theoretical computer science, the problem dates back to at least [39], who developed an RNC algorithm (i.e., running in $O(\mathrm{polylog}(n))$ parallel time on a PRAM with $O(\mathrm{poly}(n))$ work) for finding a perfect matching, and asked whether one can also sample a uniform random perfect matching in parallel; this remains a major open problem to date.

Recent works study parallel sampling assuming access to global counting oracles. One such model is the weighted counting oracle, where one has access to the log Laplace transform $L(w) = \ln \mathbb{E}_\mu[\exp(\langle w, x\rangle)]$. A weaker model is the unweighted counting oracle, which returns

conditional marginals $\mathbb{P}_{X\sim\mu}[X_i = a \mid X_S = \omega_S]$ for $S \subseteq [n]$, $i \notin S$, $a \in \{\pm 1\}$, and $\omega_S \in \{\pm 1\}^S$. Such models are of interest for several combinatorial distributions such as spanning trees, Eulerian tours, planar perfect matchings, and determinantal point processes, where such oracles can be implemented on a PRAM in $O(\text{polylog}(n))$ time and $O(\text{poly}(n))$ work [13]. Assuming access to weighted counting oracles, [5, 3] obtained RNC samplers for distributions satisfying $\|\nabla^2 L(w)\|_{\text{op}} = O(1)$. For more general distributions that do not satisfy such a regularity condition, [28] developed an algorithm running in $\widetilde{O}(n^{2/3})$ parallel time with polynomial work. The parallel runtime was later improved to $\widetilde{O}(\sqrt{n})$ by [2]. In the weaker unweighted counting model, [4] developed an $O(n^{2/3})$ time parallel sampler with $O(\text{poly}(n))$ work. The parallel runtime was later improved to $O(\sqrt{n})$ by [2].

Contrary to the above, our work falls under the more challenging local oracle setting, where one has to design a parallel sampler for a Gibbs measure given access to its Hamiltonian. This constitutes a far weaker form of access compared to the global counting oracles described above. In fact, the problem of computing a counting oracle of a Gibbs measure given its Hamiltonian is NP-hard in the worst case [25], and highly nontrivial even for specialized Hamiltonians like the SK model. In this setting, [23, 36] develop RNC samplers for Ising models satisfying a Dobrushin-type condition. For mean-field spin glass models, this only covers a vanishing window of temperatures, e.g., $\beta_2 = O(n^{-1/2})$ for the SK model. More recently, [11] show that the Restricted Gaussian Dynamics algorithm of [12] is an RNC sampler for the SK model. For general Ising p-spin models, much less is known. To our knowledge, the only parallel sampler in this setting prior to our work is [34], which runs in $\widetilde{O}(\sqrt{n})$ parallel time with polynomial work.

1.3 Overview of Our Techniques

Algorithm 1 s-Glauber Dynamics with Parallel Rejection Sampling.

Sample $X^{(0)}$ from $\mu_0 \propto e^{\langle h,x\rangle}$
for $t = 0, \ldots, T-1$ **do**
  Sample $S \sim \text{Uniform}\binom{[n]}{s}$
  $X_S^{(t+1)} \leftarrow \textsc{ParallelRejectionSampler}(S, X_{\overline{S}}^{(t)}, \varepsilon/10T)$ *(see Algorithm 2)*
  $X_{\overline{S}}^{(t+1)} \leftarrow X_{\overline{S}}^{(t)}$
return $X^{(T)}$

1.3.1 High Accuracy Sampler

Motivated by the prior work of [34], our high-accuracy sampler for the p-spin model, stated in Algorithm 1, is a computationally efficient implementation of a Markov chain called s-Glauber dynamics (also known as s-block dynamics), whose kernel $P_s(X^{(t)}, \cdot)$ is defined as follows:

1. Sample S uniformly at random from all subsets of $[n]$ of size s.
2. Sample $X^{(t+1)}$ from the posterior $\mu\left(X^{(t+1)} \;\middle|\; X_S^{(t+1)} = X_S^{(t)}\right)$

The above Markov chain is reversible with respect to μ and, for $s = 1$, it corresponds to Glauber dynamics, a canonical sampling algorithm for discrete spaces. When μ is the p-spin Gibbs measure at high temperature, s-Glauber dynamics is known to exhibit a mixing time of $\widetilde{O}(n/s)$ [7, 34]. Although this mixing time improves upon increasing s, the computational complexity of implementing each step of s-Glauber dynamics worsens significantly as s grows.

Algorithm 2 Approximate Parallel Rejection Sampling with an Ising Proposal (ParallelRejectionSampler).

Input: S, pinning τ on $\overline{S}$, $\varepsilon_{\text{step}}$

Set $y = X_S$ and decompose the conditional Hamiltonian $H_{S,\tau}(y) = H(y, \tau)$ on S as follows

$$H(y) = \langle b_{S,\tau}, y\rangle + \frac{1}{2}\, y^\top A_{S,\tau} y + R_{S,\tau}(y).$$

Set the rejection parameter $\bar{c} = \Theta(1)$, $L = \Theta(\bar{c}\ln(1/\varepsilon_{\text{step}}))$, $\varepsilon_{\text{RGD}} = \Theta(\varepsilon_{\text{step}}/L)$, $\eta = \Theta(1)$, and $\delta = \Theta(\varepsilon_{\text{step}})$

$\hat{h} = \textsc{CenteringExternalField}(A_{S,\tau}, b_{S,\tau}, R_{S,\tau}, \eta, \delta)$ *(see Algorithm 6)*

for $\ell = 1$ **to** L **do In Parallel**

- $U_\ell, V_\ell \overset{\text{iid}}{\sim} \text{RGD}\Big(A_{S,\tau},\, b_{S,\tau} + \hat{h},\, \varepsilon_{\text{RGD}}\Big)$
- $W_\ell \sim \mathsf{Uniform}[0,1]$
- Accept U_ℓ if $W_\ell \le \min\big\{1, \bar{c}^{-1}\exp(\psi(U_\ell) - \psi(V_\ell))\big\}$ where $\psi(y) = R_{S,\tau}(y) - \Big\langle \hat{h}, y \Big\rangle$

return *the first accepted sample; declare FAILURE otherwise*

Indeed, a naïve implementation of the posterior sampling step (i.e., Step 2 above) takes $O(2^s)$ time, which is computationally prohibitive, and completely negates the potential benefits of any mixing time improvements for $s > 1$.

To circumvent this, our Algorithm 1 replaces the posterior sampling step in s-Glauber dynamics with an approximate rejection sampler (Algorithm 2), whose proposal distribution is a carefully chosen Ising model. The idea of using rejection sampling in the posterior sampling step also appears in the work of [34], which approximates the posterior sampling step with a product proposal. While this enables computational efficiency (as product measures can be sampled in $O(1)$ parallel time), their rejection sampling procedure fails to correctly approximate the posterior when $s = \omega(\sqrt{n})$, and obtains a parallel runtime of $\widetilde{O}(n^{1/2})$. In contrast, the choice of our Ising proposal, which serves as the key technical innovation underlying our parallel speedup, can be made to approximate the posterior to arbitrary accuracy as long as $s = O(n^{2/3})$, leading to an improved parallel runtime of $\widetilde{O}(n^{1/3})$.

To motivate the construction of our Ising proposal, we observe that the conditional posterior distribution $\mu_{S,\tau}(y) \coloneqq \mu(x_S = y \mid x_{\bar{S}} = \tau) \propto \exp(H_{S,\tau}(y))$ is also a p-spin Gibbs measure whose Hamiltonian, $H_{S,\tau}(y) = H(x)$ for $x_S = y$ and $x_{\bar{S}} = \tau$, is obtained by restricting H to the coordinates in S and pinning the coordinates in $\bar{S}$ according to τ. $H_{S,\tau}$ is then a polynomial of the underlying disorder G, and admits the following decomposition:

$$H_{S,\tau}(y) = \langle b_{S,\tau}, y\rangle + \frac{1}{2}\langle y, A_{S,\tau} y\rangle + R_{S,\tau}(y) \tag{3}$$

where $b_{S,\tau}$, $A_{S,\tau}$, and $R_{S,\tau}$ depend on G. Motivated by this decomposition, our rejection sampler uses a proposal distribution $\pi_{S,\tau}$ such that $\ln \pi_{S,\tau}$ serves as a quadratic approximation to $H_{S,\tau}$, i.e., $\pi_{S,\tau}(x) \propto \exp\Big(\frac{1}{2}\langle y, A_{S,\tau} y\rangle + \Big\langle b_{S,\tau} + \hat{h}, y\Big\rangle\Big)$[1]. The performance and accuracy of our algorithm are governed by two key problems: 1. How easily can we sample from the proposal? 2. How well does our rejection sampler approximate the posterior? Below, we discuss how we address each of these questions in our proof of Theorem 1.

[1] As explained in Section 3.2, the nonzero external field $\hat{h}$ *centers* our rejection sampling proposal, leading to improved accuracy.

Analyzing the parallel complexity of our algorithm is a delicate problem due to the choice of our proposal distribution. Indeed, sampling from an Ising model is in itself a nontrivial problem, which can even be NP-hard in the worst case [45, 24]. To resolve this, we use the concentration properties of the p-spin Hamiltonian to prove that sampling from $\pi_{S,\tau}$ is computationally tractable. In particular, the Hamiltonian of $\pi_{S,\tau}$ satisfies a spectral condition which ensures that a Markov chain known as Restricted Gaussian Dynamics (or RGD, see Algorithm 3) approximately samples from $\pi_{S,\tau}$ in total variation, with $\widetilde{O}(1)$ mixing time [12]. Moreover, each step of RGD can be implemented in $\widetilde{O}(1)$ parallel time with $O(\mathrm{poly}(n))$ work [11]. Thus, to ensure parallel efficiency, our rejection sampler uses approximate samples from the proposal distribution, drawn via $\widetilde{O}(1)$ parallel calls to RGD, thereby implying an $\widetilde{O}(1)$ parallel runtime for Algorithm 2, which in turn leads to an $\widetilde{O}(n/s)$ parallel runtime for Algorithm 1.

Algorithm 3 Restricted Gaussian Dynamics (RGD) [12, 11].

Input: symmetric interaction matrix $J \in \mathbb{R}^{m\times m}$, external field $h \in \mathbb{R}^m$, tolerance ε
Output: Sample X satisfying $d_{\mathsf{TV}}(\mathsf{Law}(X), \nu_{J,h}) \le \epsilon$ where
$\nu_{J,h}(x) \propto \exp(\frac{1}{2}\langle x, Jx\rangle + \langle h, x\rangle)$
Initialize $X^{(0)} \in \{\pm 1\}^m$ arbitrarily
Set $T = \Theta(\log(m/\epsilon))$
for $t = 1, \dots, T$ **do**
  Sample $y \sim \mathcal{N}(X^{(t)}, J^{-1})$
  Sample $X^{(t)}$ from the product measure $p(x) \propto \exp(\langle x, h + Jy\rangle)$
return $X^{(T)}$

Finally, the optimal choice of s is determined by the approximation error of our rejection sampler. There are two main sources of this error, namely, the approximate sampling error of RGD (which can be controlled to arbitrary precision without sacrificing algorithmic efficiency), and the error due to mismatch between the proposal $\pi_{S,\tau}$ and the conditional posterior $\mu_{S,\tau}$, which is governed by the log-density $\psi(y) = \ln \frac{\mathrm{d}\mu_{S,\tau}}{\mathrm{d}\pi_{S,\tau}}(y)$. We show that this error can be controlled by proving exponential tail bounds on ψ under $\pi_{S,\tau}$, which we establish via the concentration properties of the p-spin Hamiltonian and the isoperimetric properties of the proposal, and conclude that Algorithm 2 uniformly approximates the posterior sampling step of s-Glauber dynamics for $s \le O(n^{2/3})$, leading to an overall runtime of $\widetilde{O}(n^{1/3})$ for Algorithm 1.

1.3.2 Low Accuracy Sampler

Adapting the framework of [17, 19], our low-accuracy parallel sampler, presented in Algorithm 4, is based on Eldan's Stochastic Localization (SL) [20], which we briefly introduce. For any probability measure μ supported on $\{\pm 1\}^n$, the associated SL process $(y_t)_{t\ge 0}$ is defined by the following stochastic differential equation on $\mathbb{R}^n$.

$$\mathrm{d}y_t = m(y_t)\,\mathrm{d}t + \mathrm{d}B_t, \quad y_0 = 0, \qquad m(y) = \frac{\mathbb{E}_{x\sim\mu}[x\cdot\exp(\langle y, x\rangle)]}{\mathbb{E}_{x\sim\mu}[\exp(\langle y, x\rangle)]} \tag{4}$$

Here, $(B_t)_{t\ge 0}$ denotes the standard Brownian motion on $\mathbb{R}^n$, and $m(y)$ corresponds to the mean of the exponential tilt of the measure μ, which we call the tilted mean. A key property of SL is that $t^{-1}y_t$ is marginally distributed as $x^\star + \zeta_t$ where $x^\star \sim \mu$ and

$\zeta_t \sim \mathcal{N}(0, t^{-1}I)$. Intuitively, $t^{-1}y_t$ is approximately distributed as μ for large enough t, and thus, a suitable discretization of Equation (4), such as the Euler discretization below, can be used to approximately sample from μ.

$$\widetilde{y}_{(k+1)h} = \widetilde{y}_{kh} + hm(\widetilde{y}_{kh}) + (B_{(k+1)h} - B_{kh}),\ 0 \le k < K, \qquad y_0 = 0 \tag{5}$$

Being the Euler discretization of a continuous Markov process, the discrete dynamical system specified by Equation (5) naturally evolves in a sequential fashion, and thus, naively simulating K steps requires $\Theta(K)$ calls to an oracle for the tilted mean m, even on a PRAM. To circumvent this, we employ a parallelization technique based on the concept of Picard iteration, a classical technique for proving the existence and uniqueness of ODE solutions [42], which was later used algorithmically by [3] to develop parallel algorithms for sampling from continuous log-concave densities in $\mathbb{R}^n$. The key idea behind Picard iteration is to view the trajectory defined by Equation (5) as the solution to the following fixed point problem defined by an operator $\mathcal{T} : \mathbb{R}^{n\times K} \to \mathbb{R}^{n\times K}$ over discrete trajectories:

$$\mathcal{T}(z) \coloneqq z'; \qquad z'_{kh} = h\sum_{i=0}^{k-1} m(z_{ih}) \ + B_{kh}\ \forall\ k \in [K], \qquad z'_0 = 0 \tag{6}$$

By unrolling the recurrence, one can observe that the trajectory defined by Equation (5) is a fixed point of the operator $\mathcal{T}$, i.e., $\mathcal{T}(\widetilde{y}) = \widetilde{y}$. Consequently, the discretized SL process can be approximated via the following fixed point iteration over discrete trajectories, which we call Picard iteration.

$$\widetilde{y}^{(r)}_{kh} = h\sum_{i=0}^{k-1} m(\widetilde{y}^{(r-1)}_{ih}) + B_{kh},\ \forall\ k \in [K], \qquad \widetilde{y}^{(r)}_0 = 0 \tag{7}$$

Algorithm 4 Picard Algorithmic Stochastic Localization.

Input : Parallel depth R, steps K, step size h, parameters $(\eta, K_{\mathrm{AMP}}, K_{\mathrm{NGD}})$
$B_0 \leftarrow 0$
for $k \leftarrow 0$ **to** $K-1$ **do In Parallel**
  $B_{(k+1)h} \leftarrow B_{kh} + \varepsilon_k$, where $\varepsilon_k \sim \mathcal{N}(0, hI)$
  $\hat{y}^{(0)}_{kh} \leftarrow 0$
for $r \leftarrow 1$ **to** R **do**
  $\hat{y}^{(r)}_0 \leftarrow 0$
  for $k \leftarrow 0$ **to** $K-1$ **do In Parallel**
    $\hat{m}\big(\hat{y}^{(r-1)}_{kh}\big) \leftarrow \text{TAP-AMP}\big(\hat{y}^{(r-1)}_{kh}, \eta, q_\star(kh), K_{\mathrm{AMP}}, K_{\mathrm{NGD}}\big)$
      (see Algorithm 5) and Equation (18)
  for $k \leftarrow 1$ **to** K **do In Parallel**
    $\hat{y}^{(r)}_{kh} \leftarrow h\sum_{j=0}^{k-1} \hat{m}\big(\hat{y}^{(r-1)}_{jh}\big) + B_{kh}$
Output : $\hat{x} \leftarrow \mathrm{sign}\big(\hat{y}^{(R)}_{Kh}\big)$

The key algorithmic benefit of this reformulation lies in its inherent parallelizability, albeit at the cost of introducing some mild approximation error. Indeed, Equation (7) turns the problem of sequentially evaluating the recurrence in Equation (5) into a highly parallelizable fixed point computation over the whole trajectory: given access to a parallel oracle for the tilted mean m, each Picard iteration can be implemented efficiently in parallel across all time

points. In particular, at depth r, $(\widetilde{y}_{kh}^{(r-1)})_{k\in[K]}$ can be computed simultaneously in one parallel oracle call, after which the cumulative drift terms in Equation (7) can be computed via parallel prefix sums in $O(\log(k))$ time and $O(k)$ work. However, to guarantee the effectiveness of this approach as a parallel sampler, one needs to quantify how well the Picard iterate $\widetilde{y}^{(R)}$ approximates its limiting trajectory $\widetilde{y}$. To this end, one can show (see Lemma 6) that Lipschitzness of the tilted mean (which can be guaranteed in the high-temperature regime) suffices to ensure that the Picard iterates converge to the discretized SL process exponentially fast. In particular, $R = \widetilde{\Theta}(Kh)$ iterations suffice to approximate the discrete SL trajectory in Equation (5) to arbitrary accuracy.

The key challenge in algorithmically implementing the Picard iteration in Equation (7) arises from the approximate computation of the tilted mean m, which is highly nontrivial and is known to be NP-hard even for worst-case Hamiltonians [25]. To circumvent this, we make use of the TAP-AMP algorithm of [18] (Algorithm 5), which constructs an estimator $\hat{m}$ of the tilted mean for points along the continuous SL trajectory. This results in the following inexact fixed point iteration, which we call Picard Algorithmic Stochastic Localization (Algorithm 4).

$$\hat{y}_{kh}^{(r)} = h\sum_{i=0}^{k-1}\hat{m}(\hat{y}_{ih}^{(r-1)}) + B_{kh},\ \forall\ k\in[K],\quad \hat{y}_0^{(r)} = 0 \tag{8}$$

While standard perturbation-based analyses of inexact fixed point iterations rely on uniformly controlling the deviation between the ideal and the inexact iterates, such a strategy is not applicable to Equation (8) due to the absence of uniform guarantees on the estimation error of $\hat{m}$. In fact, the current best known analysis of the tilted mean estimator can certify only a very weak error bound of the form $\|\hat{m}(y_t) - m(y_t)\|^2 \le n\varepsilon_n^2$ for points along the continuous SL trajectory. Here, ε_n is a fixed error threshold which converges to 0 as $n\to\infty$. This subtle distinction constitutes a major obstacle in our analysis since the Picard trajectories at low depth (i.e., low values of r) are expected to deviate significantly from the continuous SL trajectory. Consequently, the currently available guarantees are insufficient for controlling $\|\hat{m}(\hat{y}_{kh}^{(r)}) - m(\hat{y}_{kh}^{(r)})\|$.

To circumvent this obstacle, we depart from the standard perturbation-based analysis of inexact fixed point iteration, and instead analyze Equation (8) directly. In particular, we prove that in the high-temperature regime, the tilted mean estimator $\hat{m}(y)$ itself is uniformly $O(1)$-Lipschitz. Hence, the inexact Picard iteration in Equation (8) converges exponentially fast to the discrete process $\hat{y}_{(k+1)h} = \hat{y}_{kh} + h\hat{m}(y_{kh}) + (B_{(k+1)h} - B_{kh})$ in $R = \widetilde{O}(Kh)$ iterations. We then prove a sign rounding stability guarantee for the limiting discrete trajectory $\hat{y}$, and show that for any $\varepsilon \ge \varepsilon_n$, the distribution of $\mathrm{sign}(\hat{y}_{Kh}^{(R)}/Kh)$ approximates the p-spin measure μ up to ε error in normalized 2-Wasserstein distance whenever $Kh = O(\log(n/\varepsilon))$ and $h = O(\mathrm{poly}(1/\varepsilon))$. Since the K iterations are executed in parallel in each round, Algorithm 4 exhibits a parallel runtime of $O(\mathrm{polylog}(n/\varepsilon))$ with $O(\mathrm{poly}(n/\varepsilon))$ work. Compared to the $O(\mathrm{poly}(n)\exp(\mathrm{poly}(1/\varepsilon)))$ runtime of Algorithmic Stochastic Localization, this constitutes a doubly exponential improvement in runtime and an exponential improvement in work in terms of ε.

2 Preliminaries

2.1 Notation

For differentiable functions $g:\mathbb{R}^d\to\mathbb{R}$, we use ∇g and $\nabla^2 g$ to denote their Euclidean gradient and Hessian, respectively. For functions $f:\{\pm1\}^m\to\mathbb{R}$, we define the discrete partial derivative, and the associated discrete gradient and discrete Hessian as follows:

Algorithm 5 AMP Algorithm for Tilted Mean Estimation (TAP-AMP) [19, Algorithm 1].

Input : Iterations $K_{\mathrm{AMP}}, K_{\mathrm{NGD}}$, step size η, overlap parameter q

$\hat{m}^{-1} \leftarrow 0$, $z^0 \leftarrow 0$

for $k \leftarrow 0$ **to** $K_{\mathrm{AMP}} - 1$ **do**

$\hat{m}^k \leftarrow \tanh(z^k)$

$\hat{q}_k \leftarrow \frac{1}{n}\sum_{i=1}^n \tanh^2(z_{i,k})$

$b_k \leftarrow (1-\hat{q}_k)\,\xi''(\hat{q}_k)$

$z^{k+1} \leftarrow \beta \nabla H(\hat{m}^k) + y - b_k \hat{m}^{k-1}$

$u^0 \leftarrow z^{K_{\mathrm{AMP}}}$, $\hat{m}^{+,0} \leftarrow \hat{m}^{K_{\mathrm{AMP}}}$

for $k \leftarrow 0$ **to** $K_{\mathrm{NGD}} - 1$ **do**

$u^{k+1} \leftarrow u^k - \eta \nabla \hat{\mathcal{F}}_{\mathrm{TAP}}(\hat{m}^{+,k}; y, q)$ *(see Equation (22))*

$\hat{m}^{+,k+1} \leftarrow \tanh(u^{k+1})$

Output : $\hat{m}^{+,K_{\mathrm{NGD}}}$

$$\partial_i f(x) = \frac{1}{2}\cdot[f(x_{\neg i}\ ;\ x_i = +1) - f(x_{\neg i}\ ;\ x_i = -1)],$$
$$(\nabla f(x))_i = \partial_i f(x), \qquad (\nabla^2 f(x))_{ij} = \partial_i \partial_j f(x) \tag{9}$$

For any two distributions μ and ν we use $d_{\mathsf{TV}}(\mu,\nu)$ to denote their total variation distance. When μ and ν are distributions on $\mathbb{R}^d$, we use $W_2(\mu,\nu)$ to denote their Euclidean 2-Wasserstein distance and $W_{2,d}(\mu,\nu) = d^{-1/2}W_2(\mu,\nu)$ to denote its normalized variant. Throughout, we use H and μ to denote the Ising p-spin Hamiltonian and its associated Gibbs measure as defined in Equation (2). We use $\xi(t) = \sum_{p=2}^P \beta_p^2 t^p$ to denote its mixture function. We also define the coefficients $\mathfrak{C}(\beta)$ and $\mathfrak{D}(\beta)$ as follows:

$$\mathfrak{C}(\beta) \coloneqq \sum_{p=2}^P \beta_p \sqrt{p^3 \ln(p)}\ ; \qquad \mathfrak{D}(\beta) \coloneqq \sum_{p=2}^P \beta_p \sqrt{2^p p^3 \ln(p)} \tag{10}$$

2.2 Stochastic Localization

Let μ denote any arbitrary measure on $\mathbb{R}^n$ with a finite second moment. Stochastic localization, introduced by [20], denotes the following diffusion process:

$$\mathrm{d}y_t = m(y_t,t)\,\mathrm{d}t + \mathrm{d}B_t, \qquad y_0 = 0$$
$$m(y,t) = \mathbb{E}_{x\sim\mu,\zeta\sim\mathcal{N}(0,1)}\left[x \;\middle|\; tx + \sqrt{t}\zeta = y\right] \tag{11}$$

Stochastic localization exhibits several interesting properties that have found innumerable applications in areas such as convex geometry and the analysis of Markov chains [35, 21, 12]. Among these, the following property, which appears in [16], will be of particular importance to us.

▶ **Theorem 3** (Equivalent Characterization of SL [16]). *Let $(y_t)_{t\geq 0}$ denote the stochastic localization process associated with the measure μ, as defined in Equation (11). Then, there exists an $x^\star \sim \mu$ and a standard Brownian motion $(W_t)_{t\geq 0}$ such that for any $t \geq 0$, $y_t = tx^\star + W_t$. Consequently, the marginal law of $t^{-1}y_t$ is equal to that of $x^\star + \zeta_t$ where $\zeta_t \sim \mathcal{N}(0, t^{-1}I)$.*

The following lemmas, which follow from Theorem 3, are vital for analyzing the rounding scheme used in Algorithm 4.

▶ **Lemma 4** (Rounding SL for the hypercube). *Let μ be a measure supported on $\{\pm 1\}^n$, and let $(y_t)_{t\geq 0}$ denote the associated SL process. For any $T > 0$, define $\hat{x}_T = \mathrm{sign}(T^{-1}y_T)$. Then, $W_{2,n}^2(\mathsf{Law}(\hat{x}_T), \mu) \leq 8\exp(-T/2)$.*

▶ **Lemma 5** (Stability of SL Rounding). *Let μ be a distribution on $\{\pm 1\}^n$, and let $(y_t)_{t\geq 0}$ denote the associated SL process. For any $T > 0$, let $\hat{z}_T$ be a random variable satisfying $W_2^2(\hat{z}_T, T^{-1}y_T) \leq n\Delta^2$. Then,*

$$W_{2,n}^2\big(\mathrm{sign}(\hat{z}_T), \mathrm{sign}(T^{-1}y_T)\big) \leq 8e^{-T/8} + 16\Delta^2. \tag{12}$$

2.3 Picard Iteration

Consider a discrete-time dynamical system $(x_{kh})_{0\leq k\leq K}$ on $\mathbb{R}^n$ defined by the following iteration:

$$x_{(k+1)h} = x_{kh} + hf(kh, x_{kh}) + g(kh), \qquad k \in \{0, \ldots, K-1\} \tag{13}$$

where $h > 0$ denotes the step-size and $T = Kh$ denotes the time horizon, and the functions f and g can potentially be random. We observe that solving the difference equation above is an inherently sequential task. Concretely, given an initialization x_0 and access to an oracle for computing f and g, computing the trajectory $(x_{kh})_{0\leq k\leq K}$ requires $\Theta(K)$ oracle calls.

The key idea in Picard iteration is to view the entire trajectory $(x_{kh})_{0\leq k\leq K}$ as the fixed point of a trajectory-level iteration in $\mathbb{R}^{n\times K}$ such that, given access to a *parallel oracle* for computing f and g, each round of this trajectory-level iteration can be computed in $\Theta(1)$ oracle calls. To begin, we unroll the recurrence in Equation (13) and observe that the solution $(x_{kh})_{0\leq k\leq K}$ satisfies the following:

$$x_{kh} = x_0 + \sum_{j=0}^{k-1} \left(hf(jh, x_{jh}) + g(jh)\right), \qquad 0 \leq k \leq K \tag{14}$$

The Picard iteration associated with Equation (13) is the following sequence of trajectories $(x_{kh}^{(r)})_{0\leq k\leq K},\ r \in \mathbb{N}$.

$$x_{kh}^{(r+1)} = x_0 + \sum_{j=0}^{k-1} \left(hf(jh, x_{jh}^{(r)}) + g(jh)\right) \tag{15}$$

We make two key observations regarding the above trajectory sequence:

- The solution x_{kh} to Equation (13) is a fixed point of Equation (15). Indeed, setting $x_{kh}^{(r)} = x_{kh}$ for every k and using Equation (14), we conclude that $x_{kh}^{(r+1)} = x_{kh}$.
- Each round of the iteration in Equation (15) can be computed using $\Theta(1)$ calls to a parallel oracle for computing f and g. Indeed, the values of $g(jh),\ 0 \leq j < K$ can be precomputed using one parallel oracle call, and in each round, the values of $f(jh, x_{jh}^{(r)}),\ 0 \leq j < K$ can be computed via one call to a parallel oracle to f.

The parallel efficiency of Picard iteration then depends on how fast the sequence of trajectories in Picard iteration converges to its fixed point trajectory. Below, we prove that Picard iteration exhibits an exponential convergence rate whenever $f(t, x)$ is uniformly L-Lipschitz in x.

▶ **Lemma 6** (Exponential convergence of Picard iteration under Lipschitzness). *Let $(x_{kh})_{0\le k\le K}$ denote the solution to Equation (13), where $T = Kh$, and let $(x_{kh}^{(r)})_{0\le k\le K,\ r\ge 0}$ denote the associated Picard iteration as defined in Equation (15). Suppose $f(t,\cdot)$ is uniformly L-Lipschitz in x and define $M = \max_{0\le k\le K}\|x_{kh}^{(0)} - x_{kh}\|$. Then, for any $r \ge 0$ and $0 \le k \le K$,*

$$\|x_{kh}^{(r)} - x_{kh}\| \le \frac{(khL)^r}{r!}\cdot M. \tag{16}$$

In particular, $R = \max\{e^2 LT, \ln(M/\varepsilon)\}$ Picard iterations suffice to ensure the following:

$$\max_{k\in\{0,\dots,K\}} \|x_{kh}^{(R)} - x_{kh}\| \le \varepsilon. \tag{17}$$

2.4 TAP Free Energy and the Tilted Mean Estimator

The construction of the tilted mean estimator (Algorithm 5) of El Alaoui, Montanari, and Sellke [19] is motivated by a deep structural property of the p-spin model: The mean of the tilted Gibbs measure at high temperature is an approximate stationary point of the Thouless–Anderson–Palmer (TAP) free energy [47, 37, 46], which we briefly introduce below.

Recall that the mixture function of the p-spin model is defined as $\xi(t) = \sum_{p=2}^{P}\beta_p^2 t^p$. For any $t \ge 0$, the asymptotic overlap $q^\star(t)$ is defined as follows:

$$q_{k+1}(t) = \mathbb{E}_{W\sim N(0,1)}\left[\tanh\left(t + \xi'(q_k) + W\sqrt{\xi'(q_k)+t}\right)\right], \quad q_k(t) = 0, \quad q^\star(t) = \lim_{k\to\infty} q_k(t) \tag{18}$$

Since $q_k(t)$ is independent of G, computable to high accuracy via a one-dimensional Gaussian integral, and approaches the limit $q^\star(t)$ exponentially fast, we assume, for simplicity, that $q^\star(t)$ is known exactly for $t = kh, k \in [K]$.

The TAP free energy associated with the p-spin Hamiltonian in Equation (2) is defined as follows:

$$\mathcal{F}_{\mathrm{TAP}}(m, y) = -H(m) - \langle y, m\rangle - \sum_{i=1}^{n} h(m_i) - \mathrm{ONS}(Q(m)) \tag{19}$$

$$\mathrm{ONS}(q) = \frac{n}{2}\left[\xi(1) - \xi(Q) - (1-Q)\xi'(Q)\right] \tag{20}$$

$$Q(m) = n^{-1}\|m\|^2, \qquad h(m) = -\frac{1+m}{2}\ln\left(\frac{1+m}{2}\right) - \frac{1-m}{2}\ln\left(\frac{1-m}{2}\right) \tag{21}$$

For the sake of tractability, El Alaoui, Montanari, and Sellke [19] also define an approximate version of the TAP free energy which replaces the $\mathrm{ONS}(Q(m))$ term by a quadratic approximation.

$$\hat{\mathcal{F}}_{\mathrm{TAP}}(m\,; y, q) = -H(m) - \langle y, m\rangle - \sum_{i=1}^{n} h(m_i) - \mathrm{ONS}(q) - \mathrm{ONS}'(q)(Q(m) - q) + \frac{n\Gamma}{8}(Q(m) - q)^2 \tag{22}$$

where q is set to be the asymptotic overlap $q^\star(t)$ for $t = kh$, and $\Gamma = \Theta(1)$ is a tunable parameter depending only on ξ. [19] show that the mean of the tilted Gibbs measure continues to be an approximate stationary point of $\hat{\mathcal{F}}_{\mathrm{TAP}}$. The tilted mean estimator in Algorithm 5 then works in two phases. The first phase runs a series of AMP updates to compute an approximate stationary point $\hat{m}^{K_{\mathrm{AMP}}}$ for $\hat{\mathcal{F}}_{\mathrm{TAP}}$ around which it is locally strongly convex. The second stage exploits this local strong convexity to refine the AMP estimate via natural gradient descent.

3 High Accuracy Parallel Sampler for the p-Spin Model

In this section, we analyze Algorithm 1 and prove the following Theorem.

▶ **Theorem 7** (Analysis of Algorithm 1). *For any $\varepsilon > 0$, the output $X^{(T)}$ of Algorithm 1 run with $s = \Theta(n^{2/3}\log(n/\varepsilon)^{-2/3})$ and $T = \Theta(n^{1/3}\log(n/\varepsilon)^{5/3})$ satisfies $d_{\mathsf{TV}}(\mathsf{Law}(X^{(T)}), \mu) \le \varepsilon$ with probability at least $1 - e^{-cn}$. Moreover, Algorithm 1 has a parallel runtime of $O(n^{1/3}\mathrm{polylog}(n/\varepsilon))$ with $O(\mathrm{poly}(n, \log(n/\varepsilon)))$ work.*

We describe the key technical ideas of our proof below, and refer the readers to the full version of our paper for complete proofs.

3.1 Bounds on the Hamiltonian of the Conditional Posterior

Consider any arbitrary $S \in \binom{[n]}{s}$. For any $x \in \{\pm 1\}^n$, let $x = (y, \tau)$ where $y = x_S$ denotes the block spins in S and $\tau = x_{\bar{S}}$ denotes the pinning outside S. Then the conditional distribution $\mu_{S,\tau}(y) = \mu(y \mid x_{\bar{S}} = \tau)$ of the spins in S satisfies $\mu_{S,\tau}(y) \propto \exp(H_{S,\tau}(y))$ where:

$$H_{S,\tau}(y) = \langle b_{S,\tau}, y\rangle + \tfrac{1}{2}\langle y, A_{S,\tau}y\rangle + R_{S,\tau}(y) \tag{23}$$

Here, $A_{S,\tau}$ is a symmetric matrix with zero diagonals with:

$$(A_{S,\tau})_{ij} = \sum_{p=2}^{P} \frac{\beta_p\sqrt{p!}}{n^{\frac{p-1}{2}}} \sum_{K\subseteq\bar{S},|K|=p-2} G_{K\cup\{i,j\}}\tau^K, \qquad i,j \in S,\ i \ne j \tag{24}$$

The remainder term is given by

$$R_{S,\tau}(y) = \sum_{p=3}^{P} \frac{\beta_p\sqrt{p!}}{n^{(p-1)/2}} \sum_{\substack{J\subseteq[n],|J|=p,\\ |J\cap S|\ge 3}} G_J\ y^{J\cap S}\ \tau^{J\cap\bar{S}} \tag{25}$$

As discussed earlier, each step of s-Glauber dynamics requires sampling from $\mu_{S,\tau}$, which can be computationally expensive. We address this via a parallelized implementation of rejection sampling in Algorithm 2 where $\mu_{S,\tau}(y)$ is approximated by an Ising model of the form $\pi_{S,\tau}(y) \propto \exp(\frac{1}{2}\langle y, A_{S,\tau}y\rangle + \left\langle b_{S,\tau} + \hat{h}, y\right\rangle)$, where $\hat{h}$ is an external field that satisfies $\hat{h} \approx \mathbb{E}_{y\sim\pi_{S,\tau}}[\nabla R_{S,\tau}(y)]$.

To ensure the parallel efficiency of Algorithm 2, we generate approximate samples from the Ising proposal $\pi_{S,\tau}$ using Restricted Gaussian Dynamics (Algorithm 3). To this end, the following theorem uses the concentration properties of the p-spin Hamiltonian to prove that $\|A_{S,\tau}\|_{\mathsf{op}} \le 1/4$ holds with high probability uniformly over all choices of S and τ whenever $s = O(n^{2/3})$. Consequently, Algorithm 3 can approximately sample from the proposal in $\widetilde{O}(1)$ parallel time.

▶ **Theorem 8** (Uniform Operator Norm Bound for Ising Proposal). *There exists a numerical constant $C > 0$ such that for $\mathfrak{C}(\beta) \le C$ and $s = o(n)$, the following holds with probability at least $1 - \exp(-cn)$:*

$$\|A_{S,\tau}\|_{\mathsf{op}} \le \frac{1}{4}, \qquad \forall\ S \in \binom{[n]}{s},\ \tau \in \{\pm 1\}^{\bar{S}} \tag{26}$$

Naturally, the success probability of our rejection sampling procedure is governed by how well the Ising proposal $\pi_{S,\tau}$ approximates the conditional distribution $\mu_{S,\tau}$, or, in other words,

how large the influence of the remainder term $R_{S,\tau}$ is. We quantify this via a uniform bound on the discrete Hessian of $R_{S,\tau}$, which follows from the concentration properties of the p-spin Hamiltonian and a careful combinatorial bookkeeping.

▶ **Theorem 9** (Uniform Bound on Hessian of Remainder Term). *There exists a numerical constant $C > 0$ such that for $\mathfrak{C}(\beta) \le C$, the following holds with probability at least $1 - \exp(-cn)$ whenever $\Omega(\sqrt{n}) \le s \le o(n)$*

$$\|\nabla^2 R_{S,\tau}(y)\|_F^2 \le \frac{C_1 \mathfrak{C}(\beta)^2 s^3}{n^2}, \qquad \forall\ S \in \binom{[n]}{s},\ (y,\tau) \in \{\pm 1\}^n \tag{27}$$

3.2 Analysis of Parallel Rejection Sampling

In this section, we shall establish the accuracy and parallel efficiency of Algorithm 2. Henceforth, we condition on the event $\mathcal{E}$ that the bounds in Theorem 8 and Theorem 9 hold uniformly over all S and τ.

▶ **Theorem 10** (Analysis of Parallel Rejection Sampler). *For any $S \in \binom{[n]}{s}$ and $\tau \in \{\pm 1\}^{\overline{S}}$, let $\tilde{\mu}_{S,\tau}$ denote the law of the sample output by Algorithm 2. Conditioned on the event $\mathcal{E}$, the following holds whenever $s = \Theta(n^{2/3})$:*

$$d_{\mathsf{TV}}(\tilde{\mu}_{S,\tau}, \mu_{S,\tau}) \le \varepsilon_{\mathrm{step}} \qquad \forall S \in \binom{[n]}{s},\ \tau \in \{\pm 1\}^{\overline{S}}. \tag{28}$$

Moreover, Algorithm 2 has parallel runtime $O(\mathrm{polylog}(n/\varepsilon_{\mathrm{step}}))$ with $O(\mathrm{poly}(n, \log(1/\varepsilon_{\mathrm{step}})))$ work.

The proof of Theorem 10, whose complete proof is relegated to the full version, involves several technical components, the first of which is an accuracy guarantee for the approximate rejection sampling step in Algorithm 2. This result appears as Lemma 4.4 in Lee [34] and Lemma 2 in [22].

▶ **Lemma 11** (Accuracy of Rejection Sampling). *Let P and Q be probability measures satisfying $\frac{dP}{dQ}(y) \propto e^{\psi(y)}$. Let $Y, Z \overset{\text{iid}}{\sim} Q$, and define $R = \exp(\psi(Y) - \psi(Z))$. Consider the following rejection sampling procedure: Draw $U \sim \mathsf{Uniform}[0,1]$ and accept Y if $U \le \min\{1, R/c\}$ for some $c \ge 1$. Let $\tilde{P}$ denote the law of the accepted sample and assume $\psi(Y) - \psi(Z)$ exhibits subexponential tails of the form $\mathbb{P}[\psi(Y) - \psi(Z) \ge t] \le a\exp(-bt)$ for $b > 1$. Then, $d_{\mathsf{TV}}\left(P, \tilde{P}\right) \le \frac{a}{b-1} \cdot c^{-(b-1)}$*

In Algorithm 2, (X, Z) corresponds to (U_ℓ, V_ℓ), both of which are approximately distributed as $\pi_{S,\tau}$ (modulo the negligible sampling error of Algorithm 3, which we ignore for now). Since $\frac{d\mu_{S,\tau}}{d\pi_{S,\tau}}(y) \propto e^{\psi(y)}$, where $\psi(y) = R_{S,\tau}(y) - \langle \hat{h}, y\rangle$, Lemma 11 suggests that we can bound the TV error of Algorithm 2 by controlling the sampling error of Algorithm 3 and proving a subexponential tail bound for $|\psi(\widetilde{U}) - \psi(\widetilde{V})|$, where $\widetilde{U}, \widetilde{V} \overset{\text{iid}}{\sim} \pi_{S,\tau}$, or equivalently, for $|\psi(Y) - \mathbb{E}[\psi(Y)]|$ under $Y \sim \pi_{S,\tau}$. By Theorem 8 and [34, Theorem 4.1], $\pi_{S,\tau}$ satisfies approximate tensorization of entropy with high probability. The work of [43] then implies the following concentration bound:

$$\mathbb{P}_{Y \sim \pi_{S,\tau}}\left[|\psi(Y) - \mathbb{E}_{\pi_{S,\tau}}[\psi(Y)]|\right] \le 2\exp\left(-c\min\left\{\frac{t^2}{\mathbb{E}_{\pi_{S,\tau}}[\|\nabla\psi\|^2]}, \frac{t}{\max\limits_{y\in\{\pm1\}^S}\|\nabla^2\psi(y)\|_F}\right\}\right) \tag{29}$$

Algorithm 6 Centering External Field.

Input: $A_{S,\tau}$, $b_{S,\tau}$, $R_{S,\tau}$, accuracy $\eta > 0$, failure probability $\delta \in (0,1)$
Set iterations $T = \Theta(\log(n/\eta))$ and sample size $K = \Theta\big(s\eta^{-2}\log(sT/\delta)\big)$
Initialize $z_0 \leftarrow 0$
for $t = 0, \ldots, T-1$ **do**
  for $i = 1, \ldots, K$ **do** In parallel
    $U_i^{(t)} \sim \mathrm{RGD}(A_{S,\tau}, b_{S,\tau} + z_t, \delta/100KT)$
  $z_{t+1} = \frac{1}{K}\sum_{i=1}^{K} \nabla R_{S,\tau}(U_i^{(t)})$
return $\hat{h}_{S,\tau} = z_T$

Since $\|\nabla^2\psi(y)\|_F^2 = \|\nabla^2 R_{S,\tau}(y)\|_F \lesssim \mathfrak{C}(\beta)\cdot\frac{s^3}{n^2}$ holds whp uniformly for any S, τ and y by Theorem 9. Hence, setting $s = \Theta(n^{2/3})$ and $\mathfrak{C}(\beta)$ sufficiently small, $\max_{y\in\{\pm1\}^S} \|\nabla^2\psi(y)\|_F \le \delta$ for any $\delta = O(1)$. To control $\mathbb{E}_{\pi_{S,\tau}}[\|\nabla\psi\|^2]$, we note that:

$$\mathbb{E}_{\pi_{S,\tau}}[\|\nabla\psi\|^2] = \| \hat{h} - \mathbb{E}_{\pi_{S,\tau}}[\nabla R_{S,\tau}] \|^2 + \mathbb{E}_{\pi_{S,\tau}}\left[\| \nabla R_{S,\tau} - \mathbb{E}[\nabla R_{S,\tau}] \|^2\right] \tag{30}$$

Using the Poincare Inequality for $\pi_{S,\tau}$ and Theorem 9, the second term in Equation (30) can be made $O(\delta^2)$ for any $\delta = O(1)$ by setting $s = \Theta(n^{2/3})$ and $\mathfrak{C}(\beta)$ sufficiently small. The centering field $\hat{h}$ must be chosen so as to make the first term $\|\hat{h} - \mathbb{E}_{\pi_{S,\tau}}[\nabla R_{S,\tau}]\|^2$ as small as possible. Since $\pi_{S,\tau} = \nu_{A_{S,\tau}, b_{S,\tau}+\hat{h}}$ where $\nu_{J,v}(y) \propto \exp(\frac{1}{2}\langle y, Jy\rangle + \langle v, y\rangle)$, the optimal choice of $\hat{h}$ is given by the solution to the following fixed point problem:

$$h = \mathbb{E}_{\nu_{A_{S,\tau},\, h+b_{S,\tau}}}[\nabla R_{S,\tau}]. \tag{31}$$

As we shall demonstrate, Equation (31) admits a unique solution which can be efficiently approximated by an inexact fixed point iteration, namely Algorithm 6, which replaces the expectation in Equation (31) by an empirical average over approximate samples drawn via RGD. In particular, we prove the following guarantee.

▶ **Theorem 12** (Analysis of Algorithm 6). *Conditioned on the event $\mathcal{E}$, the output $\hat{h}$ of Algorithm 6 satisfies $\|\hat{h} - \mathbb{E}_{\pi_{S,\tau}}[\nabla R_{S,\tau}]\| \le \eta$ with probability at least $1-\delta$. Moreover, Algorithm 6 runs in $O(\mathrm{polylog}(n, \eta^{-1}, \delta^{-1}))$ parallel time with $O\big(\mathrm{poly}(\frac{n}{\eta}, \ln(1/\delta))\big)$ work.*

The proof of Theorem 10, which is presented in our full version, then follows by combining Theorem 12 and Lemma 11

3.3 Proof of Theorem 7

Proof. Let $\mu_t = \mathsf{Law}(X^{(t)})$. Since $\mu_0(x) \propto \exp(\langle h, x\rangle)$, by Lemma 5.5 of [34],

$$\mathbb{P}\big[D_{\mathsf{KL}}(\mu_0 \,||\, \mu) \le Cn\big] \ge 1 - \exp(-cn). \tag{32}$$

Let $\mathcal{G} = \mathcal{E} \cap \big\{D_{\mathsf{KL}}(\mu_0 \,||\, \mu) \le Cn\big\}$. Then, $\mathbb{P}[\mathcal{G}] \ge 1 - e^{-cn}$, and henceforth, we condition on $\mathcal{G}$.

Let P_s and $\widetilde{P}_s$ denote the Markov kernel for s-Glauber dynamics and the Markov chain in Algorithm 1 respectively. By Theorem 10, for $s = \Theta((\frac{n}{\log(n/\varepsilon)})^{2/3})$,

$$\max_{x\in\{\pm1\}^n} d_{\mathsf{TV}}\big(P_s(x,\cdot), \widetilde{P}_s(x,\cdot)\big) \le \varepsilon_{\mathrm{step}} = \tfrac{\varepsilon}{10T}. \tag{33}$$

Hence,

$$d_{\mathsf{TV}}\left(\mu_T, \mu_0 P_s^T\right) \le T\varepsilon_{\text{step}} \le \tfrac{\varepsilon}{10}. \tag{34}$$

By [7, Theorem 6], the following holds with probability $1 - e^{-cn}$:

$$D_{\mathsf{KL}}\left(\mu_0 P_s^T \mid\mid \mu\right) \le \exp\left(-\tfrac{csT}{n}\right) Cn \le \exp\left(-\frac{cT}{n^{1/3}\log(n/\varepsilon)^{2/3}}\right) Cn. \tag{35}$$

Setting $T = \Theta(n^{1/3}\log(n/\varepsilon)^{5/3})$, applying Pinsker's inequality and taking appropriate union bounds, we conclude that the following holds with probability at least $1 - \exp(-cn)$:

$$d_{\mathsf{TV}}(\mu_T, \mu) \le d_{\mathsf{TV}}\left(\mu_T, \mu_0 P_s^T\right) + d_{\mathsf{TV}}\left(\mu_0 P_s^T, \mu\right) \le \tfrac{\varepsilon}{10} + \tfrac{\varepsilon}{2} \le \varepsilon. \tag{36}$$

To compute the parallel runtime, note that Algorithm 1 makes T calls to Algorithm 2 each of which has parallel runtime $O(\mathsf{polylog}(n/\varepsilon))$ with $O(\mathsf{poly}(n/\varepsilon))$ processors as per Theorem 10. Hence, Algorithm 1 exhibits a parallel runtime of $O(n^{1/3}\mathsf{polylog}(n/\varepsilon))$ with $O(\mathsf{poly}(n/\varepsilon))$ processors. ◀

4 Low Accuracy Parallel Sampler for the p-Spin Model

In this section, we key technical ideas of our proof below and refer the readers to the full version of our paper for complete proofs. We prove that Picard Algorithmic Stochastic Localization (Algorithm 4) samples from the mixed p-spin model with Wasserstein guarantees.

We first illustrate our results for the SK model. The TAP-AMP mean-estimation algorithm for the SK model (Algorithm 7) is slightly different from that for the mixed p-spin model, but it uses the same AMP iterations followed by NGD iterations. We prove that the TAP-AMP algorithms proposed in prior work for computing mean approximations in the SK and mixed p-spin models, Algorithm 7 and Algorithm 5, are Lipschitz with respect to the external field and $\widetilde{O}(1)$ parallel computable. This enables us to apply Picard iteration to the steps of algorithmic stochastic localization Equation (8), yielding a parallel speedup.

For the rest of the section, we first present an improved parameter dependence for algorithmic stochastic localization in Section 4.1. Then, using the improved parameters, we prove the parallelization result for algorithmic stochastic localization in Section 4.2, relying on the following two results:

- The TAP-AMP mean approximations are executable in $\widetilde{O}(1)$ parallel time (Section 4.3).
- The TAP-AMP mean approximations are Lipschitz (Section 4.4).

4.1 Sign Rounding for Algorithmic Stochastic Localization

El Alaoui, Montanari, and Sellke [17, 19] provide Wasserstein guarantees for algorithmic stochastic localization using TAP fixed points as mean approximations. Their analysis controls the propagation of the mean-estimation error along the discretized stochastic localization trajectory. They obtain an ε upper bound in normalized Wasserstein distance between the output of the sampling algorithm and the target distribution by using $\exp(\mathrm{poly}(1/\varepsilon))$ discrete steps to run the stochastic localization up to a time horizon $\mathrm{poly}(1/\varepsilon)$.

▶ **Proposition 13.** *For $\varepsilon_n < \varepsilon$, El Alaoui, Montanari, and Sellke [17, 19] provide algorithmic stochastic localization for a time horizon* $\mathrm{poly}(1/\varepsilon)$ *with* $\exp(\mathrm{poly}(1/\varepsilon))$ *discretization steps.*

In this work, in addition to the Picard parallelization and Lipschitz analysis, we improve the dependence of the discretization step size and the time horizon in stochastic localization on the parameter ε by analyzing sign rounding for algorithmic stochastic localization.

▶ **Theorem 14.** *For any n, there exists a value ε_n such that, for any $\varepsilon > \varepsilon_n$ and any high-temperature mixed p-spin model with Hamiltonian H_n and no external field, there is an algorithm based on algorithmic stochastic localization using TAP-AMP mean approximations (Algorithm 7 and Algorithm 5) whose output is within ε of the target distribution in normalized Wasserstein distance with probability $1 - o(1)$ over the disorder.*

Moreover, the algorithmic stochastic localization procedure uses a discretization step of size δ and time horizon t satisfying

$$\delta = \mathrm{poly}(\varepsilon), \qquad t = O(\log(1/\varepsilon)).$$

The output of the algorithm is generated by sign rounding, namely $\mathrm{sign}(\hat{y}_t)$, where $\hat{y}_t$ is the output of the discrete algorithmic stochastic localization process at time t. Furthermore, if y_t denotes the true stochastic localization process, then

$$W_{2,n}(y_t, \hat{y}_t) \le O(\varepsilon).$$

4.2 Parallel Time Analysis of Algorithm 4

In this section, we analyze the parallelization result for algorithmic stochastic localization using Picard iteration. The Picard iterations produce a trajectory that is close, in Wasserstein distance, to the true solution of the discrete stochastic localization process using the TAP-AMP mean approximation. This discrete process is itself close to the true continuous-time stochastic localization process by Theorem 14. Therefore, by the triangle inequality for Wasserstein distance, the input to the rounding step is close to the output of the true stochastic localization process.

We present our results for the SK model and the mixed p-spin model separately. For the SK model, the Gibbs measure is defined by

$$\mu_A(x) \propto \exp\left(\frac{\beta}{2}\langle x, Ax\rangle\right),$$

where $A \sim \mathrm{GOE}(n)$. Here $\mathrm{GOE}(n)$ denotes the distribution over symmetric matrices whose entries are Gaussian with variance of order $1/n$. Although the proof structure is nearly the same, we separate the two cases because, for the SK model, we obtain a parallel sampling algorithm in a temperature regime not covered by previous parallel algorithms. The analysis for the SK model yields the threshold $\beta_0 \approx 0.3753$ for Lipschitzness and parallelization, improving on the $\beta < 1/4$ threshold at which the previous $\mathrm{polylog}(n)$-depth algorithm for sampling from the SK model applies [11]. However, unlike the algorithm of Chen, Liu, Yin, and Zhang [11], our low-accuracy algorithm cannot achieve arbitrarily small error.

▶ **Theorem 15** (SK Model Parallelization)**.** *For any $\varepsilon \ge \varepsilon_n$ and inverse temperature $\beta < \beta_0 \approx 0.3753$, there exist parameters (Theorem 14)*

$$\eta, K_{\mathrm{AMP}}, K_{\mathrm{NGD}} = \mathrm{polylog}(n/\varepsilon), \quad K = \mathrm{poly}(1/\varepsilon), \quad \delta = \mathrm{poly}(\varepsilon), \quad t = K\delta = O(\log(1/\varepsilon))$$

such that the following holds. Let $R = O(\frac{\log^2(1/\varepsilon)}{\beta_0 - \beta})$, and execute Algorithm 4 on a Hamiltonian of the form $H_n(x) = x^\top Ax$ using TAP-AMP (Algorithm 7) as the approximate mean function $\hat{m}$ with parameters $(\eta, K_{\mathrm{AMP}}, K_{\mathrm{NGD}}, K, \delta)$.

Then Algorithm 4 outputs a random point $\mathbf{x}^{\mathrm{alg}} \in \{-1, +1\}^n$ with law $\mu_{\mathbf{A}}^{\mathrm{alg}}$ such that, with probability $1 - o(1)$ over $\mathbf{A} \sim \mathrm{GOE}(n)$,

$$W_{2,n}(\mu_{\mathbf{A}}^{\mathrm{alg}}, \mu_{\mathbf{A}}) \le O(\varepsilon).$$

The parallel runtime of the algorithm is $\mathrm{poly}(\log(n/\varepsilon))$, and its total work is $\mathrm{poly}(n/\varepsilon)$.

▶ **Theorem 16** (p-spin Model Parallelization). *For any $\varepsilon \geq \varepsilon_n$ and temperature coefficients $\{\beta_p\}_{p\geq 2}$ satisfying*

$$\mathfrak{C}(\beta) \coloneqq \sum_{p=2}^{P} \beta_p \sqrt{p^3 \ln(p)} < \gamma_0 \qquad \mathfrak{D}(\beta) = \sum_{p=2}^{P} \beta_p \sqrt{2^p p^3 \ln(p)} < \infty$$

there exist parameters (Theorem 14)

$$\eta, K_{\mathrm{AMP}}, K_{\mathrm{NGD}}, K = \mathrm{poly}(1/\varepsilon), \qquad \delta = \mathrm{poly}(\varepsilon), \qquad t = K\delta = O(\log(1/\varepsilon)),$$

such that the following holds. Let $R = O_{\mathfrak{C}(\beta)}\big(\log^2(1/\varepsilon)\big)$ be the number of Picard iterations. Given query access to the p-spin Hamiltonian as defined in Equation (2), Algorithm 4, run with parameters $(\eta, K_{\mathrm{AMP}}, K_{\mathrm{NGD}}, K, \delta)$, outputs a random point $\mathbf{x}^{\mathrm{alg}} \in \{-1, +1\}^n$ with law $\mu_{H_n}^{\mathrm{alg}}$ such that, with probability $1 - o(1)$ over the Gaussian coefficients $\{g_J\}_{J\subseteq[n]}$,

$$W_{2,n}(\mu_{H_n}^{\mathrm{alg}}, \mu_{H_n}) \leq O(\varepsilon).$$

Here, ε is the error incurred by the sequential algorithm in Theorem 14, while $\mathrm{poly}(\varepsilon)$ is the additional error incurred by replacing the sequential updates with Picard iteration.

The parallel runtime of the algorithm is $\mathrm{poly}(\log(n/\varepsilon))$ and its total work is $\mathrm{poly}(n, 1/\varepsilon)$.

4.3 TAP-AMP in $\mathrm{polylog}(n)$ Parallel Time

The TAP-AMP mean-approximations can be computed in $\mathrm{polylog}(n/\varepsilon)$ parallel depth. The TAP-AMP algorithms Algorithm 7 and Algorithm 5 consist of AMP iterations followed by NGD iterations. Each individual AMP or NGD iteration is computable in parallel with depth $O(\log n)$; thus, it suffices to prove $\widetilde{O}(1)$ bounds on the numbers of AMP and NGD iterations. Since this proof is rather involved and does not affect the rest of the argument, we give only a brief sketch here; the complete details appear in the full version of this paper.

▶ **Proposition 17.** *The mean-approximation algorithms Algorithm 7 and Algorithm 5 are computable with parallel depth $\mathrm{polylog}(n/\varepsilon)$, given query access to the mixed p-spin Hamiltonian and the parameter choices specified in Theorem 14 for attaining normalized Wasserstein error ε.*

4.4 Lipschitz Property of the TAP-AMP Mean Approximation for the SK Model

In this section, we prove that, for the SK model at high temperature $\beta < \beta_0 \approx 0.3753$, the TAP-AMP algorithm (Algorithm 7) for mean approximation is $O\Big(\frac{1}{\beta_0 - \beta}\Big)$-Lipschitz with respect to the tilt parameter, or external field, with high probability. We provide an analogous result for the mixed p-spin model in the full version of this paper.

Algorithm 7 Mean of the Tilted Ising Measure [17].

Input: Data $A \in \mathbb{R}^{n\times n}$, $y \in \mathbb{R}^n$, parameters $\beta, \eta > 0$, $q \in (0,1)$, iteration numbers $K_{\mathrm{AMP}}, K_{\mathrm{NGD}}$

$m_{-1} = z_0 = 0$

for $k = 0, \dots, K_{AMP} - 1$ **do**

  $m_k = \tanh(z_k), \quad b_k = \frac{\beta^2}{n}\sum_{i=1}^n (1 - \tanh^2((z_k)_i))$

  $z_{k+1} = \beta A m_k + y - b_k m_{k-1}$

$u^0 = z_{K_{\mathrm{AMP}}}$

for $k = 0, \dots, K_{NGD} - 1$ **do**

  $u_{k+1} = u_k - \eta \cdot \nabla \hat{\mathcal{F}}_{\mathrm{TAP}}(m_k^+; y, q)$

  $m_{k+1}^+ = \tanh(u_{k+1})$

return $m^+_{K_{\mathrm{NGD}}}$

Algorithm 7 consists of two phases: it first uses AMP iterations attempting to get close to the fixed point of the TAP iteration, and then it uses an NGD algorithm to close the gap to the fixed point as much as needed. Here, we will take a closer look at the two phases of Algorithm 7.

4.4.1 Phase 1: AMP phase

In the first phase, initialize $m_{-1} = 0$ and $z_0 = 0$. For $k = 0, 1, \dots, K_{\mathrm{AMP}} - 1$, set $m_k = \tanh(z_k)$ and $b_k = \frac{\beta^2}{n}\sum_{i\in[n]} \operatorname{sech}^2\big((z_k)_i\big)$, then update

$$z_{k+1} = \beta A m_k + y - b_k\, m_{k-1}. \tag{37}$$

4.4.2 Phase 2: NGD phase on TAP free energy

Initialize $u_0 := z_{K_{\mathrm{AMP}}}$. For $t = 0, 1, \dots, K_{\mathrm{NGD}} - 1$, set $m_t^+ = \tanh(u_t)$ and update

$$u_{t+1} = u_t - \eta\, \nabla_m \hat{\mathcal{F}}_{\mathrm{TAP}}(m_t^+; y, q). \tag{38}$$

The algorithm outputs $\hat{m}(A, y) := m^+_{K_{\mathrm{NGD}}} = \tanh(u_{K_{\mathrm{NGD}}})$.

▶ **Definition 18.** *For a given external field $y \in \mathbb{R}^n$ and interaction matrix $A \in \mathbb{R}^{n\times n}$, let $\hat{m}_{(K_{AMP}, K_{NGD})}(A, y)$ be the output of the algorithm after running the AMP iteration for K_{AMP} steps and then running NGD for K_{NGD} steps. We refer to $\hat{m}(y)$ as the tilted-mean computation for the external field y whenever A, K_{AMP}, and K_{NGD} are clear from the context.*

Our goal in this section is to find the temperature $\beta(A) = \beta(\|A\|_{\mathrm{op}})$ such that, for any two tilts y and $\widetilde{y}$, the TAP fixed point of the Ising model with interaction matrix A and temperature β satisfies

$$\|\hat{m}_{(K_{\mathrm{AMP}}, K_{\mathrm{NGD}})}(A, y) - \hat{m}_{(K_{\mathrm{AMP}}, K_{\mathrm{NGD}})}(A, \widetilde{y})\| \le \frac{\gamma(A)}{\beta(A) - \beta}\|y - \widetilde{y}\|.$$

▶ **Theorem 19** (Lipschitzness of AMP phase). *Assume $\gamma(A, \beta) := \beta\,\|A\|_{\mathrm{op}} + (1 + \frac{4}{3\sqrt{3}})\beta^2 < 1$. Then for all $k \ge 0$,*

$$\|\Delta z_k\| \le \frac{1}{1-\gamma}\,\|\Delta y\|, \qquad \|\Delta m_k\| \le \frac{1}{1-\gamma}\,\|\Delta y\|.$$

Lipschitzness of the full AMP+NGD algorithm for the SK model

We have proved so far that after an arbitrary number of iterations, $\|\Delta m_k\|$ is upper bounded by $\frac{1}{1-\gamma(A,\beta)}\|\Delta y\|$. Now, we prove that applying NGD after AMP preserves the same bound for the difference of the mean estimates. For convenience, and to avoid overloading the notation m_k from the AMP phase, we analyze the NGD step in the variables u_i defined by $m = \tanh(u)$ (equivalently $u = \operatorname{atanh}(m)$ component-wise).

▶ **Theorem 20** (TAP Lipschitz property)**.** *Assume* $\gamma(A,\beta) = \beta\|A\|_{\mathrm{op}} + (1+\frac{4}{3\sqrt{3}})\beta^2 < 1$ *and* $q \in (0,1)$*. Let* $0 < \eta \le 1$*. Then for every* $t \ge 0$*,*

$$\|\hat{m}(A,y) - \hat{m}(A,\widetilde{y})\| \le \frac{1}{1-\gamma(A,\beta)}\|y-\widetilde{y}\|.$$

▶ Remark 21. The key stability requirement for AMP and NGD iterations is

$$\gamma(A,\beta) = \beta\|A\|_{\mathrm{op}} + (1+\frac{4}{3\sqrt{3}})\beta^2 < 1.$$

In random matrix settings (e.g. $A \sim \mathrm{GOE}(n)$), one typically has $\|A\|_{\mathrm{op}} = 2 + o(1)$ with high probability, so any sufficiently small β ensuring $\gamma(2,\beta) < 1$ implies that the TAP-AMP mean approximation is $\frac{1}{1-\gamma(2,\beta)-c}$-Lipschitz with high probability for every sufficiently small constant c.

▶ **Corollary 22.** *Fix* $\varepsilon > 0$*. If the SK inverse temperature* β *is at most* $\beta_0 - \varepsilon$*, where* $\beta_0 \approx 0.3753$ *is the positive root of the equation*

$$\gamma(2,\beta) = 2\beta + (1+\frac{4}{3\sqrt{3}})\beta^2 = 1,$$

then the tilted-mean computation in Algorithm 7 is $O(\frac{1}{\varepsilon})$*-Lipschitz for the SK model with high probability.*

Proof. We know that $\|A\|_{\mathrm{op}}$ is concentrated around 2, and the probability that it exceeds $2+\varepsilon/\beta_0$ is $e^{-\Omega_\varepsilon(n)}$ [9]. Recall that β_0 is the positive solution of the equation $2x+(1+\frac{4}{3\sqrt{3}})x^2 = 1$. Thus, we only need to show that the parameter $\gamma = \beta\|A\|_{\mathrm{op}} + (1+\frac{4}{3\sqrt{3}})\beta^2 \le 1 - \Omega(\varepsilon)$ with high probability.

$$\begin{aligned}\beta\|A\|_{\mathrm{op}} + (1+\frac{4}{3\sqrt{3}})\beta^2 &\le \beta(2+\varepsilon/\beta_0) + (1+\frac{4}{3\sqrt{3}})\beta^2\\ &\le \varepsilon + 2\beta + (1+\frac{4}{3\sqrt{3}})\beta^2\\ &\le -\varepsilon + 2\beta_0 + (1+\frac{4}{3\sqrt{3}})\beta_0^2 = 1-\varepsilon\end{aligned}$$

Since $\mathrm{P}\{\|A\|_{\mathrm{op}} > 2+\varepsilon/\beta_0\} \le e^{-\Omega_\varepsilon(n)}$, the tilted-mean computation is $O(\frac{1}{\varepsilon})$-Lipschitz with high probability if $\beta \le \beta_0 - \varepsilon$. ◀

References

1 Arka Adhikari, Christian Brennecke, Changji Xu, and Horng-Tzer Yau. Spectral gap estimates for mixed p-spin models at high temperature. *arXiv preprint*, 2022. arXiv:2208.07844.

2 Nima Anari, Carlo Baronio, CJ Chen, Alireza Haqi, Frederic Koehler, Anqi Li, and Thuy-Duong Vuong. Parallel sampling via autospeculation. *arXiv preprint*, 2025. doi:10.48550/arXiv.2511.07869.

3 Nima Anari, Sinho Chewi, and Thuy-Duong Vuong. Fast parallel sampling under isoperimetry. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 161–185. PMLR, 2024. URL: https://proceedings.mlr.press/v247/anari24a.html.

4 Nima Anari, Ruiquan Gao, and Aviad Rubinstein. Parallel sampling via counting. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 537–548, 2024. doi:10.1145/3618260.3649744.

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

6 Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Entropic independence: optimal mixing of down-up random walks. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1418–1430, 2022. doi:10.1145/3519935.3520048.

7 Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Universality of spectral independence with applications to fast mixing in spin glasses. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 5029–5056. SIAM, 2024. doi:10.1137/1.9781611977912.181.

8 Nima Anari, Frederic Koehler, and Thuy-Duong Vuong. Trickle-down in localization schemes and applications. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1094–1105, 2024. doi:10.1145/3618260.3649622.

9 Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2010.

10 Michael Celentano. Sudakov–fernique post-amp, and a new proof of the local convexity of the tap free energy. *The Annals of Probability*, 52(3):923–954, 2024.

11 Xiaoyu Chen, Hongyang Liu, Yitong Yin, and Xinyuan Zhang. Efficient parallel ising samplers via localization schemes. *arXiv preprint*, 2025. doi:10.48550/arXiv.2505.05185.

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

13 Laszlo Csanky. Fast parallel matrix inversion algorithms. In *16th Annual Symposium on Foundations of Computer Science (sfcs 1975)*, pages 11–12. IEEE, 1975.

14 PL Dobruschin. The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory of Probability & Its Applications*, 13(2):197–224, 1968.

15 Martin Dyer, Alan Frieze, and Ravi Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM (JACM)*, 38(1):1–17, 1991. doi:10.1145/102782.102783.

16 Ahmed El Alaoui and Andrea Montanari. An information-theoretic view of stochastic localization. *IEEE Transactions on Information Theory*, 68(11):7423–7426, 2022. doi:10.1109/TIT.2022.3180298.

17 Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from the Sherrington-Kirkpatrick Gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 323–334. IEEE, 2022. doi:10.1109/FOCS54457.2022.00038.

18 Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Algorithmic stochastic localization for sampling from the p-spin model. *arXiv preprint*, 2023. arXiv:2307.07130.

19 Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from mean-field gibbs measures via diffusion processes. *Probability and Mathematical Physics*, 6(3):961–1022, 2025.

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

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

22 Jiaojiao Fan, Bo Yuan, and Yongxin Chen. Improved dimension dependence of a proximal algorithm for sampling. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1473–1521. PMLR, 2023. URL: https://proceedings.mlr.press/v195/fan23a.html.

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

24 Andreas Galanis, Alkis Kalavasis, and Anthimos Vardis Kandiros. On sampling from ising models with spectral constraints. *arXiv preprint*, 2024. arXiv:2407.07645, doi:10.48550/arXiv.2407.07645.

25 Andreas Galanis, Daniel Štefankovič, and Eric Vigoda. The complexity of approximating averages on bounded-degree graphs. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1345–1355. IEEE, 2020.

26 Reza Gheissari and Aukosh Jagannath. On the spectral gap of spherical spin glass dynamics. *arXiv preprint*, 2016. arXiv:1608.06609.

27 Roy J Glauber. Time-dependent statistics of the ising model. *Journal of mathematical physics*, 4(2):294–307, 1963.

28 Hengyuan Hu, Aniket Das, Dorsa Sadigh, and Nima Anari. Diffusion models are secretly exchangeable: Parallelizing ddpms via autospeculation. *arXiv preprint*, 2025. doi:10.48550/arXiv.2505.03983.

29 Brice Huang, Sidhanth Mohanty, Amit Rajaraman, and David X Wu. Weak poincaré inequalities, simulated annealing, and sampling from spherical spin glasses. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 915–923, 2025. doi:10.1145/3717823.3718236.

30 Brice Huang, Andrea Montanari, and Huy Tuan Pham. Sampling from spherical spin glasses in total variation via algorithmic stochastic localization. *arXiv preprint*, 2024. arXiv:2404.15651.

31 Mark Jerrum, Alistair Sinclair, and Eric Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *Journal of the ACM (JACM)*, 51(4):671–697, 2004. doi:10.1145/1008731.1008738.

32 Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical computer science*, 43:169–188, 1986. doi:10.1016/0304-3975(86)90174-X.

33 Chieh-Hsin Lai, Yang Song, Dongjun Kim, Yuki Mitsufuji, and Stefano Ermon. The principles of diffusion models. *arXiv preprint*, 2025. doi:10.48550/arXiv.2510.21890.

34 Holden Lee. Parallelising glauber dynamics. *arXiv preprint*, 2023. doi:10.48550/arXiv.2307.07131.

35 Yin Tat Lee and Santosh S Vempala. Eldan's stochastic localization and the kls conjecture: Isoperimetry, concentration and mixing. *Annals of Mathematics*, 199(3):1043–1092, 2024.

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

37 Marc Mézard, Giorgio Parisi, Miguel Angel Virasoro, and David J Thouless. Spin glass theory and beyond, 1988.

38 Andrea Montanari and Subhabrata Sen. A friendly tutorial on mean-field spin glass techniques for non-physicists. *Foundations and Trends in Machine Learning*, 17(1):1–173, 2024. doi:10.1561/2200000105.

39 Ketan Mulmuley, Umesh V Vazirani, and Vijay V Vazirani. Matching is as easy as matrix inversion. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, pages 345–354, 1987. doi:10.1145/28395.383347.

40 Dmitry Panchenko. The Sherrington-Kirkpatrick model: An overview. *Journal of Statistical Physics*, 149(2):362–383, 2012.

41 Giorgio Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180(3):378–384, 1981.

42 Lawrence Perko. *Differential equations and dynamical systems*, volume 7. Springer Science & Business Media, 2013.

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

44 Andy Shih, Suneel Belkhale, Stefano Ermon, Dorsa Sadigh, and Nima Anari. Parallel sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36:4263–4276, 2023.

45 Allan Sly and Nike Sun. The computational hardness of counting in two-spin models on d-regular graphs. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 361–369. IEEE, 2012. `doi:10.1109/FOCS.2012.56`.

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

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