STOCHASTIC LOCALIZATION AND STABLE SAMPLING ALGORITHMS FOR THE SHERRINGTON-KIRKPATRICK GIBBS MEASURE

PETER LUO AND RUSHIL MALLARAPU

CONTENTS

1. INTRODUCTION

The Sherrington-Kirkpatrick model is a ubiquitous mathematical model of a spin glass, and is especially amenable to direct analysis. Unlike the spherical spin glasses studied in class, the SK model is described by a random distribution on the n-dimensional hypercube $\{-1,+1\}^n$.

Definition 1.1. The *Sherrington-Kirkpatrick Gibbs measure* on $\{-1,+1\}^n$ is

$$\mu_{\mathbf{A}}(\mathbf{x}) = \frac{1}{Z(\beta,\mathbf{A})}\exp\left(\frac{\beta}{2}\langle \mathbf{A}, \mathbf{x}^{\otimes 2}\rangle\right), \tag{1}$$

where $\mathbf{A} \sim \mathrm{GOE}(n)$ and $\beta \geq 0$ is inverse temperature.

Observe that this is effectively the Gibbs measure associated to a 2-spin Hamiltonian, where the GOE matrix absorbs the usual $n^{-(p-1)/2}$ scaling factor and the reference measure on $\{-1,+1\}^n$ is just the uniform one on vertices.

As might be expected, being able to sample from this random landscape is a question of foundational interest in statistical physics and high-dimensional optimization. Physicists expect fast sampling should only be possible in the "high-temperature" regime (when $\beta < 1$), whereas there should be a hardness threshold in the "low-temperature" regime ($\beta > 1$). This should be intuitive: in the more homogenized high-temperature setting, this says there should not be such severe bottlenecks in the energy landscape that algorithms like Gibbs sampling/Glauber dynamics get stuck. However, in the low-temperature setting (the limit $\beta \to \infty$ of which corresponds to optimization of the Hamiltonian), we are hopeless.

Much work has gone into studying or constructing fast sampling algorithms for Eq. (1). This mixing time is usually measured by algorithmic complexity needed to achieve a

Date: May 14, 2024.

certain "closeness" to stationarity. Here we will consider the normalized 2-Wasserstein distance:

$$W_{2,n}(\mu,\nu)^2 = \inf_{\pi\in\mathcal{C}(\mu,\nu)} \frac{1}{n}\mathbf{E}_\pi\|\mathbf{X}-\mathbf{Y}\|_2^2,$$

and ask that our sampling algorithm's output, with law $\mu_{\mathbf{A}}^{\mathrm{alg}}$ should satisfy $W_{2,n}(\mu_{\mathbf{A}}^{\mathrm{alg}},\mu_{\mathbf{A}}) = o_n(1)$ in probability. Other metrics, such as Total Variation (TV), are also used in the literature.

In addition, a good sampling algorithm should be stable, in the sense of its output varying "continuously" with its inputs. We can formalize this for the setting of our randomized SK sampling algorithms as follows: consider a family $\{\mathcal{A}_n\}$ of randomized sampling algorithms,

$$\mathcal{A}_n\colon(\mathbf{A},\beta,\omega)\mapsto\mathcal{A}_n(\mathbf{A},\beta,\omega);$$

Let $\mathbf{A},\mathbf{A}' \overset{\text{iid}}{\sim} \mathrm{GOE}(n)$ and for $s\in[0,1]$, define the perturbation $\mathbf{A}_s := \sqrt{1-s^2}\mathbf{A}+s\mathbf{A}'$ of A. Define, for fixed $(\mathbf{A}_s,\beta)$, the (random) law of the perturbed algorithm

$$\mu_{\mathbf{A}_s,\beta}^{\mathrm{alg}} := \mathcal{L}(\mathcal{A}_n(\mathbf{A},\beta,\omega)),\quad \omega \perp\!\!\!\perp (\mathbf{A}_s,\beta).$$

Definition 1.2. $\{\mathcal{A}_n\}$ is *disorder stable* at inverse temperature β if

$$\lim_{s\to 0}\operatorname*{p-lim}_{n\to\infty} W_{2,n}\big(\mu_{\mathbf{A},\beta}^{\mathrm{alg}},\mu_{\mathbf{A}_s,\beta}^{\mathrm{alg}}\big)=0.$$

It is *temperature stable* at inverse temperature β if

$$\lim_{\beta'\to\beta}\operatorname*{p-lim}_{n\to\infty} W_{2,n}\big(\mu_{\mathbf{A},\beta}^{\mathrm{alg}},\mu_{\mathbf{A},\beta'}^{\mathrm{alg}}\big)=0.$$

Stability has obvious practical benefits, but one additional boon is that the existence or non-existence of stable sampling algorithms in different temperature ranges lets us prove stability of the underlying SK measure; we will return to this point in Section 3.

Much of the literature on fast SK sampling focuses on Markov chain-based methods, such as Gibbs sampling (known as Glauber dynamics in the physics literature). This algorithm computes the conditional distribution of the coordinates of a sample and iteratively updates them, and traditional techniques for controlling its mixing time (such as the Dobrushin condition) only work in the asymptotically vanishing interval of $\beta\le O(n^{-1/2})$ [AH87]. More recent work on logarithmic Sobolev inequalities for the SK Gibbs measure in the higher temperature regime of $\beta<1/4$ proves stronger dimensionless bounds for the spectral gap of Gibbs sampling [BB19; EKZ21]. This implies Gibbs sampling mixes in $O(n^2)$ flips in TV distance when $\beta<1/4$. In fact, it is possible to improve this to $O(n\log n)$ via some modifications to the argument [Ana+21].

While these Markov chain methods are both well-motivated and computationally tractable, other approaches are possible. In this report, we will describe a stable sampling algorithm for Eq. (1) constructed in [EMS24], which is based on a different sampling framework inspired by *stochastic localization* and is highly amenable to direct analysis. We will first discuss the idea of stochastic localization and how it gives a natural way of constructing sampling algorithms, and then reproduce their proof that this algorithm has fast mixing in 2-Wasserstein distance for all $\beta<1/2$ and is both disorder and temperature stable in this range. Later, this guarantee was improved to show the [EMS24] algorithm

has fast mixing for *all* $\beta < 1$ via careful—but direct—analysis on the local convexity of the TAP free energy [Cel22].

Remark 1.3. The other major result from this paper is that *no* stable algorithm for $\beta > 1$ can exist. While this fact is of immense theoretical importance, it follows from a direct analysis of the Parisi formula, and is conceptually separate from the stochastic localization theory and the latter's connection to sampling algorithms. Thus, we've chosen to focus on the algorithm guarantees in the high temperature regime in this report.

2. Overview of Stochastic Localization

Stochastic localization is a general sampling technique involving a sequence of *random measures* μ_t. Now, suppose that we want to sample from some distribution μ in $\mathbf{R}^n$.

Definition 2.1. Let $\mathbf{x}_* \sim \mu$. A *stochastic localization* $\{\mu_t\}_{t\geq 0}$ is a stochastic process valued in probability measures on $\mathbf{R}^n$ such that

(1) (Localization) As $t \to \infty$, $\mu_t \implies \delta_{\mathbf{x}_*}$. More precisely, for any bounded and continuous function f,
$$\int f(\mathbf{x})\mu_t(d\mathbf{x}) \to \int f(\mathbf{x})\mu(d\mathbf{x})$$
in distribution.
(2) (Martingale) μ_t is a continuous-time martingale.

To be more precise, we define the expectation of a random measure as follows:

Definition 2.2 (Intensity Measure)**.** If ξ is a random measure on $\mathbf{R}^n$, then there exists a measure $\mathbf{E}\,\xi$, known as the *intensity measure*, such that
$$\mathbf{E}\left[\int f(\mathbf{x})\xi(d\mathbf{x})\right] = \int f(\mathbf{x})\,\mathbf{E}\,\xi(d\mathbf{x})$$
for all measureable functions f.

This definition can be extended to conditional expectation in the usual way. So, the martingale condition translates to
$$\mathbf{E}(\mu_t|\{\mu_r\}_{0\leq r\leq s}) = \mu_s$$
for all $s \leq t$. For a more rigorous treatment of random measures, we refer the reader to [Kal17]. We now present the general notion of stochastic localization [Mon23], and then specify it to the problem at hand.

Definition 2.3. An *observation process* $(\mathbf{Y}_t)_{t\geq 0}$ with respect to a random variable $\mathbf{x}_* \sim \mu$ satisfies the property that for all $t_1 < t_2 < \cdots < t_k$, we have the following equality in distribution for $i = 2, 3, \ldots, k$:
$$\mathcal{L}(\mathbf{Y}_{t_{i-1}}|\mathbf{x}_*, \mathbf{Y}_{t_i}, \ldots, \mathbf{Y}_{t_k}) = \mathcal{L}(\mathbf{Y}_{t_{i-1}}|\mathbf{Y}_{t_i}).$$
Equivalently,
$$\mathbf{x}_*, \mathbf{Y}_{t_k}, \mathbf{Y}_{t_{k-1}}, \ldots, \mathbf{Y}_{t_1}$$
forms a Markov chain.

The Markov chain condition is one way to formalize the notion of the observations $\mathbf{Y}_t$ becoming more "informative" about x_* as t grows. In other words, each $\mathbf{Y}_t$ can be viewed as a noisy measurement of μ which gets less noisy as $t \to \infty$. To connect back to stochastic localization, we have the following claim:

Claim 2.4. *Let $\mathbf{Y}_t$ be an observation process with respect to $\mathbf{x}_*$. Then,*

$$\mu_t = \mathcal{L}(\mathbf{x}_*|\mathbf{Y}_t)$$

is a stochastic localization.

For the problem of Gibbs sampling, define the observation process

$$\mathbf{Y}_t = t\mathbf{x}_* + W_t,$$

where W_t is a standard Brownian motion independent of x_*, and define μ_t as in Claim 2.4. We can also express the observation process as

$$\mathbf{Y}_t \overset{d}{=} t\mathbf{x}_* + \sqrt{t}\mathbf{Z}$$

where $\mathbf{Z}$ is a Normal random variable in $\mathbf{R}^n$ with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_n$, independent of $\mathbf{x}_*$. By Bayes' rule, we can directly calculate the density μ_t, as follows:

$$\begin{aligned}
\mu_t(d\mathbf{x}) &\propto \varphi\left(\frac{\mathbf{Y}_t - t\mathbf{x}}{\sqrt{t}}\right)\mu(d\mathbf{x}) \\
&\propto \exp\left(-\frac{1}{2}\left\|\frac{\mathbf{Y}_t - t\mathbf{x}}{\sqrt{t}}\right\|_2^2\right)\mu(d\mathbf{x}) \\
&= \exp\left(-\frac{1}{2t}\|\mathbf{Y}_t - t\mathbf{x}\|_2^2\right)\mu(d\mathbf{x}) \\
&\propto \exp\left(\langle \mathbf{Y}_t, \mathbf{x}\rangle - \frac{t}{2}\|\mathbf{x}\|_2^2\right)\mu(d\mathbf{x}).
\end{aligned}$$

So, for this particular observation process, μ_t is a random tilt of μ. We see that as t tends to infinity, $\mu_t \implies \delta_{\mathbf{x}_*}$. Lastly, the measure we want to sample from is the SK Gibbs measure, so we set $\mu = \mu_{\mathbf{A}}$ and obtain

$$\mu_t(d\mathbf{x}) \propto \exp\left(\langle \mathbf{Y}_t, \mathbf{x}\rangle - \frac{t}{2}\|\mathbf{x}\|_2^2\right)\exp\left(\frac{\beta}{2}\langle \mathbf{A}, \mathbf{x}^{\otimes 2}\rangle\right) \propto \exp\left(\langle \mathbf{Y}_t, \mathbf{x}\rangle + \frac{\beta}{2}\langle \mathbf{A}, \mathbf{x}^{\otimes 2}\rangle\right).$$

The $-\frac{t}{2}\|\mathbf{x}\|_2^2$ term is absorbed into the normalization because $\mu_{\mathbf{A}}$ is supported on the set of vertices of the hypercube, which has constant norm.

From these definitions alone, it may not be clear why stochastic localization is useful, especially in the context of Gibbs sampling. It seems like we've turned the problem of sampling from $\mu_{\mathbf{A}}$ into a problem of sampling $\mathbf{Y}_t$, which depends on $\mu_{\mathbf{A}}$. In addition, $(\mu_t)_{t\geq 0}$ is a sequence of random measures, which adds another layer of complexity. However, the following connection to the theory of stochastic differential equations makes it clear why this idea is useful:

Proposition 2.5. *[LS77] Suppose $\mu = \mu_{\mathbf{A}}$ has finite second moment and let $\mathbf{x}_* \sim \mu$. Then $(\mathbf{Y}_t)_{t\geq 0}$ with initial condition $Y_0 = \mathbf{0}$ is the unique solution of the SDE*

$$\mathrm{d}\mathbf{Y}_t = \mathbf{m}(\mathbf{Y}_t; t)\,\mathrm{d}t + \mathrm{d}\mathbf{B}_t$$

where

$$\mathbf{m}(\mathbf{Y}_t; t) = \mathbf{E}\,\mu_t = \mathbf{E}\Big[\mathbf{x}_* \big| t\mathbf{x}_* + \sqrt{t}\mathbf{Z} = \mathbf{Y}_t\Big]$$

is the posterior mean of μ given the observation $\mathbf{Y}_t$.

Thus, to obtain samples for μ, one could discretize the SDE and output $\mathbf{Y}_T/T$ or $\mathbf{m}(\mathbf{Y}_T; T)$ for some large T. We focus on the latter option, since it is used in Algorithms 1 and 2 from [EMS24]. For consistency with the notation of the two algorithms in the next section, we change our notation slightly, emphasizing the randomness from $\mathbf{A}$, the GOE matrix that parametrizes the SK Gibbs measure. From now on, we will write $\mathbf{m}(\mathbf{A}, \mathbf{Y}_t)$ instead of $\mathbf{m}(\mathbf{Y}_t; t)$ to denote the mean of μ_t. Now, we provide a bound on the normalized Wasserstein 2-distance between $\mathbf{m}(\mathbf{A}, \mathbf{Y}_t)$ and $\mu_{\mathbf{A}}$, by way of two lemmas.

Lemma 2.6. *[Eld19] For all $t > 0$,*

$$\mathbf{E}\,\mathrm{Cov}(\mu_t) \preceq \frac{1}{t}\mathbf{I}_n,$$

where $\preceq$ is the Loewner ordering, defined by $A \preceq B \iff B - A$ is positive semi-definite, and the covariance matrix of μ_t is defined as

$$\mathrm{Cov}(\mu_t) = \int \mathbf{x}^{\otimes 2} d\mu_t(\mathbf{x}) - \left(\int \mathbf{x} d\mu_t(\mathbf{x})\right)^{\otimes 2}$$

Proof. Using the representation $\mathbf{Y}_t = t\mathbf{x}_* + \sqrt{t}\mathbf{Z}$, we have

$$\mathbf{E}\,\mathrm{Cov}(\mu_t) \preceq \mathbf{E}\left[\left(\mathbf{x}_* - \frac{\mathbf{Y}_t}{t}\right)^{\otimes 2}\right] = \mathbf{E}\left[\left(\frac{\mathbf{Z}}{\sqrt{t}}\right)^{\otimes 2}\right] = \frac{1}{t}\,\mathbf{E}[\mathbf{Z}^{\otimes 2}] = \frac{1}{t}\mathbf{I}_n.$$

//

Lemma 2.7. *For all $t > 0$,*

$$W_{2,n}(\mu_{\mathbf{A}}, \mathcal{L}(\mathbf{m}(\mathbf{A}, \mathbf{Y}_t)))^2 \leq \frac{1}{t}.$$

Proof. Recall that a positive semi-definite (PSD) matrix must have all nonnegative eigenvalues. The trace, being the sum of the eigenvalues, must also be nonnegative. By Lemma 2.6, $\frac{1}{t}\mathbf{I}_n - \mathbf{E}\,\mathrm{Cov}(\mu_t)$ is PSD and thus

$$\mathrm{tr}(\mathbf{E}\,\mathrm{Cov}(\mu_t)) \leq \mathrm{tr}\left(\frac{1}{t}\mathbf{I}_n\right) = \frac{n}{t},$$

due to the trace being additive. We further develop the LHS. By linearity, we can switch the trace and the expectation. Then, we can write the trace of the covariance matrix as an expected norm, as follows:

$$\mathrm{tr}(\mathbf{E}\,\mathrm{Cov}(\mu_t)) = \mathbf{E}[\mathrm{tr}\,\mathrm{Cov}(\mu_t)] = \mathbf{E}\Big[\mathbf{E}_{\mathbf{x}\sim\mu_t}\big[\|\mathbf{x} - \mathbf{m}(\mathbf{A}, \mathbf{Y}_t)\|_2^2\big]\Big].$$

Pattern-matching to the Wasserstein distance, we see that

$$\begin{aligned}\mathbf{E}\Big[W_{2,n}\big(\mu_t, \delta_{\mathbf{m}(\mathbf{A},\mathbf{Y}_t)}\big)^2\Big] &= n\,\mathbf{E}\Big[\inf_{\pi\in\mathcal{C}(\mu_t,\delta_{\mathbf{m}(\mathbf{A},\mathbf{Y}_t)})}\mathbf{E}_\pi\|\mathbf{X}-\mathbf{Y}\|_2^2\Big]\\ &\le n\,\mathbf{E}\Big[\mathbf{E}_{\mathbf{x}\sim\mu_t}\Big[\|\mathbf{x}-\mathbf{m}(\mathbf{A},\mathbf{Y}_t)\|_2^2\Big]\Big].\end{aligned}$$

Lastly, recall the density

$$\mu_t(d\mathbf{x}) \propto \exp\Big(\langle \mathbf{Y}_t, \mathbf{x}\rangle + \frac{\beta}{2}\langle \mathbf{A}, \mathbf{x}^{\otimes 2}\rangle\Big) \implies \mu_0(d\mathbf{x}) \propto \exp\Big(\frac{\beta}{2}\langle \mathbf{A}, \mathbf{x}^{\otimes 2}\rangle\Big),$$

so $\mu_0 = \mu_A$. Since μ_t is a martingale, then we have $\mathbf{E}\,\mu_t = \mu_0 = \mu_{\mathbf{A}}$. Jensen on the convex function of measures $(p,q) \mapsto W_{2,n}(p,q)^2$ gives the following upper-bound on the quantity of interest:

$$\begin{aligned}W_{2,n}(\mu_{\mathbf{A}}, \mathcal{L}(\mathbf{m}(\mathbf{A},\mathbf{Y}_t)))^2 &= W_{2,n}\big(\mathbf{E}\,\mu_t, \mathbf{E}\,\delta_{\mathbf{m}(\mathbf{A},\mathbf{y}_t)}\big)^2\\ &\le \mathbf{E}\Big[W_{2,n}\big(\mu_t, \delta_{\mathbf{m}(\mathbf{A},\mathbf{y}_t)}\big)^2\Big].\end{aligned}$$

Combining this inequality with the bounds we developed earlier yields the desired result. //

This lemma implies that $\mathbf{m}(\mathbf{A},\mathbf{Y}_t)$ converges in distribution to $\mu_{\mathbf{A}}$, which is an important theoretical guarantee.

Lastly, we address two practicality concerns regarding the computation of $\mathbf{m}(\mathbf{A},\mathbf{Y}_t)$.

(1) (Discretization) The first step in obtaining our samples was to discretize the SDE from 2.5. Which numerical method is best-suited for this discretization? While there are many methods available, the basic Euler discretization is used in Algorithm 2.
(2) (Estimating $\mathbf{m}(\mathbf{A},\mathbf{Y}_t)$) It is computationally intractable to sample from $\mathbf{m}(\mathbf{A},\mathbf{Y}_t) = \mathbf{E}\,\mu_t$. However, there is a known algorithm to estimate this mean accurately, called Approximate Message Passing (AMP).

Theorem 2.8. *[DAM16] The AMP algorithm gives an estimate $\widehat{\mathbf{m}}_{AMP}(\mathbf{y})$ such that*

$$\|\mathbf{m}(\mathbf{A},\mathbf{y}) - \widehat{\mathbf{m}}_{AMP}(\mathbf{y})\|_2^2/n = o_n(1)$$

These two techniques are key to the algorithms that follow.

3. The Sampling Algorithm

With this theoretical background in hand, we can describe the desired sampling algorithm. The first, an auxiliary algorithm, implements an efficient sampler of the conditional mean of the linear-tilt localization from the previous section.

The first step runs an approximate message passing (AMP) algorithm to produce a rough estimate of the mean. The second step performs natural gradient descent (NGD) to minimize the Thouless-Anderson-Palmer free energy—the $\mathcal{F}_{\text{TAP}}$ term—minimizers of which correspond to means of the Gibbs free energy. As we will see in the proof, this two-step construction is primarily considered for technical reasons. State analysis (for the AMP loop) and convergence estimates (for the NGD loop) interact particularly well,

Algorithm 1: Mean of the tilted Gibbs measure [EMS24, Alg. 1]

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

$\widehat{\mathbf{m}}^{-1} = \mathbf{z}^0 = 0$;
for $k = 0, \dots, K_{\mathrm{AMP}} - 1$ **do**
  $\widehat{\mathbf{m}}^k = \tanh(\mathbf{z}^k)$;
  $b_k = \frac{\beta^2}{n} \sum_{i=1}^n (1 - \tanh^2(z_i^k))$;
  $\mathbf{z}^{k+1} = \beta \mathbf{A} \widehat{\mathbf{m}}^k + \mathbf{y} - \mathbf{b}_k \widehat{m}^{k-1}$;
end
$\mathbf{u}^0 = \mathbf{z}^{K_{\mathrm{AMP}}}$;
for $k = 0, \dots, K_{\mathrm{NGD}} - 1$ **do**
  $\mathbf{u}^{k+1} = \mathbf{u}^k - \eta \cdot \nabla \mathcal{F}_{\mathrm{TAP}}(\widehat{\mathbf{m}}^{+,k}; \mathbf{y}, q)$;
  $\widehat{\mathbf{m}}^{+,k+1} = \tanh(\mathbf{u}^{k+1})$;
end
return $\widehat{m}^{+,K_{\mathrm{NGD}}}$

but the authors of [EMS24] expect similar performance would be achieved by simply running more AMP iterations.

The main loop, Algorithm 2, implements an Euler discretization of the stochastic localization SDE, using Algorithm 1 to update the drift coefficients while adding in Gaussian noise. The parameters q_* referenced are constants which can be precomputed from knowledge of β and the timestep δ, and are only used for technical reasons.

Algorithm 2: Sampling from Gibbs measure [EMS24, Alg. 2]

Input: Data $\mathbf{A} \in \mathbf{R}^{n\times n}$, parameters $\beta, \eta, K_{\mathrm{AMP}}, K_{\mathrm{NGD}}, L, \delta$

$\hat{\mathbf{y}}_0 = 0$;
for $\ell = 0, \dots, L - 1$ **do**
  $\mathbf{w}_{\ell+1} \sim \mathcal{N}(0, \mathbf{I}_n)$;
  $\widehat{\mathbf{m}}(\mathbf{A}, \hat{\mathbf{y}}_\ell) = \text{Alg. 1}(\beta, \eta, q_*(\beta, \ell\delta))$;
  $\hat{\mathbf{y}}_{\ell+1} = \hat{\mathbf{y}}_\ell + \widehat{\mathbf{m}}(\mathbf{A}, \hat{\mathbf{y}}_\ell)\delta + \sqrt{\delta}\mathbf{w}_{\ell+1}$;
end
$\widehat{\mathbf{m}}(\mathbf{A}, \hat{\mathbf{y}}_L) = \text{Alg. 1}(\beta, \eta, q_*(\beta, (L-1)\delta))$;
$\{x_i^{\mathrm{alg}}\}_{i\le n}$ drawn conditionally independent with $\mathbf{E}[x_i^{\mathrm{alg}} \mid \mathbf{y}, \{\mathbf{w}_\ell\}] = \widehat{m}_i(\mathbf{A}, \hat{\mathbf{y}}_\ell)$;
return $\mathbf{x}^{\mathrm{alg}}$

The positive result of [EMS24] is that for $\beta < 1/2$, Algorithm 2 converges and is disorder and temperature stable. Namely, we have

Theorem 3.1. *For any $\epsilon > 0, \beta_0 < 1/2$, there exist parameters independent of n such that for $\beta \le \beta_0$, the law $\mu_{\mathbf{A}}^{\mathrm{alg}}$ of the output of Algorithm 2 has*

$$W_{2,n}(\mu_{\mathbf{A}}^{\mathrm{alg}}, \mu_{\mathbf{A}}) \le \epsilon \text{ w.p. } 1 - o_n(1) \text{ over } \mathbf{A} \sim \mathrm{GOE}(n),$$

with total complexity $O(n^2)$ [EMS24, Thm. 2.1].

Theorem 3.2. *For any β and fixed parameters, Algorithm 2 is stable with respect to disorder and temperature [EMS24, Thm. 2.3].*

As mentioned above, one benefit of Theorems 3.1 and 3.2 is that we can transfer stability of the sampling algorithm to stability of the Gibbs measure itself:

Corollary 3.3. *For any $\beta < 1/2$, we have*

1. $\lim_{s\to 0} \text{p-lim}_{n\to\infty} W_{2,n}(\mu_{\mathbf{A},\beta}, \mu_{\mathbf{A}_s,\beta}) = 0.$
2. $\lim_{\beta'\to\beta} \text{p-lim}_{n\to\infty} W_{2,n}(\mu_{\mathbf{A},\beta}, \mu_{\mathbf{A},\beta'}) = 0.$

Proof. By the triangle inequality, we can write

$$W_{2,n}(\mu_{\mathbf{A},\beta}, \mu_{\mathbf{A}_s,\beta}) \le W_{2,n}\big(\mu_{\mathbf{A},\beta}, \mu^{\text{alg}}_{\mathbf{A},\beta}\big) + W_{2,n}\big(\mu^{\text{alg}}_{\mathbf{A},\beta}, \mu^{\text{alg}}_{\mathbf{A}_s,\beta}\big) + W_{2,n}\big(\mu^{\text{alg}}_{\mathbf{A}_s,\beta}, \mu_{\mathbf{A}_s,\beta}\big),$$

and similarly for $W_{2,n}\big(\mu_{\mathbf{A},\beta}, \mu_{\mathbf{A},\beta'}\big)$. By Theorem 3.1, the first and third terms go to 0 in the $n \to \infty$, and by Theorem 3.2, the second term vanishes in the disorder/temperature limit. //

Note that this stability result makes no reference to the algorithm, but the proof uses this algorithm's existence and properties in a fundamental way, similar to what we've seen in class for Langevin dynamics.

3.1. **Proof of Theorem 3.1.** The key idea behind showing convergence of Algorithm 2 is that the AMP and NGD iterations of Algorithm 1 is a good enough approximation of $\mathbf{m}(\mathbf{A}, \mathbf{y})$. In particular, we have the following proposition:

Proposition 3.4. *For $\beta < 1/2$, $T > 0$, there exists a constant $C < \infty$ such that*

$$\frac{1}{\sqrt{n}} \|\widehat{\mathbf{m}}(\mathbf{A}, \hat{\mathbf{y}}_L) - \mathbf{m}(\mathbf{A}, \mathbf{y}_T)\| \le C\sqrt{\delta} + o_n(1)$$

with probability $1 - o_n(1)$.

Proof. See [EMS24, p. 4.14]. The key inputs are Lipschitz continuity of the output of the AMP algorithm and gradient bounds on the TAP Hessian; the latter is where the $\beta < 1/2$ assumption is used. The rest of the proof is heavily computational and relies on analyzing the convergence rate of the NGD loop to show that after sufficiently many iterations, we can guarantee closeness of the true conditional mean $\mathbf{m}$ and the computed mean $\widehat{\mathbf{m}}$ with high probability. //

With this, the proof is relatively simple. We will need the following lemma to control the rounding step in line 8 of Algorithm 2.

Proposition 3.5. *Let μ_1, μ_2 be distributions on $[-1, 1]^n$, $\mathbf{m}_i \sim \mu_i$, and $\mathbf{x}_i \in \{\pm 1\}^n$ randomized roundings of $\mathbf{m}_i$, for $i = 1, 2$. Then*

$$W_{2,n}(\mathcal{L}(\mathbf{x}_1), \mathcal{L}(\mathbf{x}_2)) \le 2\sqrt{W_{2,n}(\mu_1, \mu_2)}.$$

Proof. Let $(\mathbf{m}_1, \mathbf{m}_2)$ be drawn according to a $W_{2,n}$-optimal coupling of (μ_1, μ_2). Denote by $\mathbf{m}_i^j$ the coordinates of this vector. Now, we can explicitly couple the roundings $\mathbf{x}_i^j$ of $\mathbf{m}_i^j$ by picking $U_1, \dots, U_n \overset{\text{iid}}{\sim} \text{Unif}([0,1])$ and setting

$$\mathbf{x}_i^j = \begin{cases} +1 & U_j \le \frac{1+\mathbf{m}_i^j}{2}, \\ -1 & \text{else}, \end{cases}$$

for $i = 1, 2$ and $1 \le j \le n$. Then,

$$\frac{1}{n}\mathbf{E}\big[\|\mathbf{x}_1 - \mathbf{x}_2\|^2 \mid (\mathbf{m}_1, \mathbf{m}_2)\big] = \frac{2}{n}\sum_{i=1}^n |\mathbf{m}_1^j - \mathbf{m}_2^j| \le 2\sqrt{\frac{1}{n}\|\mathbf{m}_1 - \mathbf{m}_2\|^2}$$

the first line by choosing a Uniform coupling and the second by Cauchy-Schwarz. Taking expectations of both sides (and using Jensen's inequality), we have upper bounded $W_{2,n}(\mathcal{L}(\mathbf{x}_1), \mathcal{L}(\mathbf{x}_2))$ by $2\sqrt{W_{2,n}(\mu_1, \mu_2)}$, thus completing the proof. //

Finally, we can present the proof of algorithmic convergence

Proof of Theorem 3.1. Writing T for the total time to run the stochastic localization process and δ for the discretized timestep, to be determined, let $L = T/\delta$, $\mathbf{m} = \mathbf{m}(\mathbf{A}, \mathbf{y}_T)$ and $\widehat{\mathbf{m}} = \widehat{\mathbf{m}}(\mathbf{A}, \hat{\mathbf{y}}_L)$. Taking laws conditional on $\mathbf{A}$, we have

$$\begin{aligned} \mathbf{E}\, W_{2,n}(\mu_{\mathbf{A}}, \mathcal{L}(\widehat{\mathbf{m}})) &\le \mathbf{E}\, W_{2,n}(\mu_{\mathbf{A}}, \mathcal{L}(\mathbf{m})) + \mathbf{E}\, W_{2,n}(\mathcal{L}(\mathbf{m}), \mathcal{L}(\widehat{\mathbf{m}})) \\ &\le T^{-1/2} + C\sqrt{\delta} + o_n(1), \end{aligned}$$

the first term by Lemma 2.7, and the second by Proposition 3.4. For any $\epsilon > 0$, we can take T, n sufficiently large and δ sufficiently small to get

$$\mathbf{E}\, W_{2,n}(\mu_{\mathbf{A}}, \mathcal{L}(\widehat{\mathbf{m}})) \le \epsilon^2/4,$$

whence by Proposition 3.5,

$$\mathbf{E}\, W_{2,n}(\mu_{\mathbf{A}}, \mu_{\mathbf{A}}^{\text{alg}}) = \mathbf{E}\, W_{2,n}(\mu_{\mathbf{A}}, \mathcal{L}(\mathbf{x}^{\text{alg}})) \le \epsilon.$$

Ergo, by Markov's inequality, we can pick the algorithm parameters so $W_{2,n}(\mu_{\mathbf{A}}, \mu_{\mathbf{A}}^{\text{alg}}) \le \epsilon$ holds whp, thus completing the proof. //

3.2. **Proof of Theorem 3.2.** While the convergence of Algorithm 2 relied crucially on properties of the stochastic localization process, the stability is reliant on a more direct observation: Algorithm 2 is given by iterative updates which are "sufficiently Lipschitz." In particular, we have the following general proposition:

Proposition 3.6. *Suppose an algorithm $\mathcal{A}$ is given iteratively: for $0 \le k \le K-1$*

$$\begin{aligned} \mathbf{z}^{k+1} &= G_k((\mathbf{z}^j, \beta\mathbf{A}\mathbf{m}^j, \mathbf{A}\mathbf{m}^j, \beta^2\mathbf{m}^j, \mathbf{w}^j)_{0\le j\le k}), \\ \mathbf{m}^k &= \rho_k(\mathbf{z}^k), \\ \mathcal{A}(\mathbf{A}, \beta, \omega) &\coloneqq \mathbf{m}^K, \end{aligned}$$

where $\omega = (\mathbf{w}^0, \dots, \mathbf{w}^{K-1})$, $\mathbf{z}^0$ *and* $\mathbf{A}$ *are independent, and the functions* $G_k \colon (\mathbf{R}^n)^{5k+5} \to \mathbf{R}^n$, $\rho_k \colon \mathbf{R}^n \to [-1,1]^n$ *are* L_0*-Lipschitz for an* n*-independent constant. Then* $\mathcal{A}$ *is disorder and temperature stable.*

Proof. See [EMS24, p. 5.1]. The idea is to first use standard matrix multiplication estimates to establish that the inputs of G_k and ρ_k are Lipschitz (plus dimensional terms) with respect to $\mathbf{m}$ and β, with high probability. This lets us analyze the error sequence

$$A_k = \frac{1}{\sqrt{n}} \max_{j \le k} \left\| \mathbf{z}^{j+1}(\mathbf{A}_0, \beta) - \mathbf{z}^{j+1}(\mathbf{A}_s, \tilde{\beta}) \right\|,$$

which with high probability satisfies $A_0 = 0$ and

$$A_{k+1} \le L_0 k^{1/2} C(A_k + s + |\beta - \tilde{\beta}|),$$

for some constant C depending on β. Thus, taking expectations and letting n get large enough, we can conclude that the temperature and disorder limits vanish, as claimed. //

As expected, the proof of stability amounts to showing Algorithm 2 follows the format of Proposition 3.6.

Proof of Theorem 3.2. We can write Algorithm 2 in the following form. Let $\ell \in \{0, \dots, L-1\}$ index the outer iterations; for each ℓ, the Algorithm 1 subroutine works as follows:

1. for $k = 0, \dots, K_{\text{AMP}} - 1$, run the AMP loop:
$$\mathbf{z}^{\ell,k+1} = \beta \mathbf{A} \tanh(\mathbf{z}^{\ell,k}) + \hat{\mathbf{y}}_\ell - \tanh(\mathbf{z}^{\ell,k-1}) \frac{\beta^2}{n} \sum_{i=1}^n \tanh'(z_i^{\ell,k}).$$
2. For $k = K_{\text{AMP}}, \dots, K_{\text{AMP}} + K_{\text{NGD}} - 1$, run
$$\mathbf{z}^{\ell,k+1} = \mathbf{z}^{\ell,k} + \eta\big[\beta \mathbf{A} \tanh(\mathbf{z}^{\ell,k}) + \mathbf{y}_\ell - \mathbf{z}^{\ell,k} - \beta^2(1 - q_\ell) \tanh(\mathbf{z}^{\ell,k})\big].$$
3. Finally, update
$$\hat{\mathbf{y}}_{\ell+1} = \hat{\mathbf{y}}_\ell + \widehat{\mathbf{m}}^{\ell, K_{\text{AMP}} + K_{\text{NGD}}} \delta + \sqrt{\delta} \mathbf{w}_{\ell+1}.$$

Thus, these are a sequence of iterative updates indexed by (ℓ, k), with randomness $\omega = (\mathbf{w}_1, \dots, \mathbf{w}_L)$, and $\rho_{\ell,k}(\mathbf{z}) = \tanh(\mathbf{z})$. Observe at each step that the functions $G_{\ell,k}$ defined by these update steps are Lipschitz, so $\hat{y}_\ell$ is updated in a Lipschitz way on the previous iterates. In fact, the hard part of the argument involves showing the Euler discretization updates are Lipschitz, which by unwinding the recursion can be done controlling the Lipschitz modulus of functions related to the AMP and NGD updates. Overall, we see that Algorithm 2 satisfies the conditions of Proposition 3.6, whence it is disorder and temperature stable. //

References

[LS77] R. S. Liptser and A. N. Shiryayev. *Statistics of Random Processes I*. New York, NY: Springer, 1977. ISBN: 9781475716672. DOI: 10.1007/978-1-4757-1665-8. URL: http://link.springer.com/10.1007/978-1-4757-1665-8.

[AH87] M. Aizenman and R. Holley. "Rapid Convergence to Equilibrium of Stochastic Ising Models in the Dobrushin Shlosman Regime". In: *Percolation Theory and Ergodic Theory of Infinite Particle Systems*. Ed. by Harry Kesten. Vol. 8. New York, NY: Springer New York, 1987, pp. 1–11. ISBN: 9781461387367. DOI: 10.1007/978-1-4613-8734-3_1. URL: http://link.springer.com/10.1007/978-1-4613-8734-3_1.

[DAM16] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. "Asymptotic mutual information for the balanced binary stochastic block model". en. In: *Information and Inference* (Dec. 2016), iaw017. ISSN: 2049-8764, 2049-8772. DOI: 10.1093/imaiai/iaw017. URL: https://academic.oup.com/imaiai/article-lookup/doi/10.1093/imaiai/iaw017.

[Kal17] Olav Kallenberg. *Random Measures, Theory and Applications*. en. Vol. 77. Probability Theory and Stochastic Modelling. Cham: Springer International Publishing, 2017. ISBN: 9783319415963. DOI: 10.1007/978-3-319-41598-7. URL: http://link.springer.com/10.1007/978-3-319-41598-7.

[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 (Apr. 2019). arXiv:1712.03676 [math-ph], pp. 2582–2588. ISSN: 00221236. DOI: 10.1016/j.jfa.2019.01.007. URL: http://arxiv.org/abs/1712.03676.

[Eld19] Ronen Eldan. "Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation". In: arXiv:1811.11530 (May 2019). arXiv:1811.11530 [math-ph]. URL: http://arxiv.org/abs/1811.11530.

[Ana+21] Nima Anari et al. "Entropic Independence I: Modified Log-Sobolev Inequalities for Fractionally Log-Concave Distributions and High-Temperature Ising Models". In: arXiv:2106.04105 (Nov. 2021). arXiv:2106.04105 [math-ph]. URL: http://arxiv.org/abs/2106.04105.

[EKZ21] Ronen Eldan, Frederic Koehler, and Ofer Zeitouni. "A Spectral Condition for Spectral Gap: Fast Mixing in High-Temperature Ising Models". In: arXiv:2007.08200 (Aug. 2021). arXiv:2007.08200 [math-ph]. URL: http://arxiv.org/abs/2007.08200.

[Cel22] Michael Celentano. "Sudakov-Fernique post-AMP, and a new proof of the local convexity of the TAP free energy". In: arXiv:2208.09550 (Aug. 2022). arXiv:2208.09550 [cs, math, stat]. URL: http://arxiv.org/abs/2208.09550.

[Mon23] Andrea Montanari. "Sampling, Diffusions, and Stochastic Localization". In: arXiv:2305.10690 (May 2023). arXiv:2305.10690 [cs]. URL: http://arxiv.org/abs/2305.10690.

[EMS24] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. "Sampling from the Sherrington-Kirkpatrick Gibbs measure via algorithmic stochastic localization". In: arXiv:2203.05093 (Feb. 2024). arXiv:2203.05093 [cond-mat]. URL: http://arxiv.org/abs/2203.05093.