2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS) | 978-1-6654-5519-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/FOCS54457.2022.00038

Sampling from the Sherrington-Kirkpatrick Gibbs measure via algorithmic stochastic localization

Ahmed El Alaoui
Department of Statistics and Data Science
Cornell University
Ithaca, NY
elalaoui@cornell.edu

Andrea Montanari
Department of Statistics
Stanford University
Stanford, CA
montanar@stanford.edu

Mark Sellke
Department of Mathematics
Stanford University
Stanford, USA
msellke@stanford.edu

***Abstract*—We consider the Sherrington-Kirkpatrick model of spin glasses at high-temperature and no external field, and study the problem of sampling from the Gibbs distribution μ in polynomial time. We prove that, for any inverse temperature $\beta < 1/2$, there exists an algorithm with complexity $O(n^2)$ that samples from a distribution $\mu^{\rm alg}$ which is close in normalized Wasserstein distance to μ. Namely, there exists a coupling of μ and $\mu^{\rm alg}$ such that if $(\boldsymbol{x}, \boldsymbol{x}^{\rm alg}) \in \{-1,+1\}^n \times \{-1,+1\}^n$ is a pair drawn from this coupling, then $n^{-1}\,\mathbb{E}\{\|\boldsymbol{x} - \boldsymbol{x}^{\rm alg}\|_2^2\} = o_n(1)$. The best previous results, by Bauerschmidt and Bodineau [BB19] and by Eldan, Koehler, Zeitouni [EKZ21], implied efficient algorithms to approximately sample (under a stronger metric) for $\beta < 1/4$.**

We complement this result with a negative one, by introducing a suitable "stability" property for sampling algorithms, which is verified by many standard techniques. We prove that no stable algorithm can approximately sample for $\beta > 1$, even under the normalized Wasserstein metric. Our sampling method is based on an algorithmic implementation of stochastic localization, which progressively tilts the measure μ towards a single configuration, together with an approximate message passing algorithm that is used to approximate the mean of the tilted measure.

I. Introduction

The Sherrington-Kirkpatrick (SK) Gibbs measure is the probability distribution on $\Sigma_N = \{-1,+1\}^N$ given by

$$\mu_{\boldsymbol{A}}(\boldsymbol{x}) = \frac{1}{Z(\beta, \boldsymbol{A})}\, \exp\left\{\frac{\beta}{2}\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle\right\}, \qquad \text{(I.1)}$$

where $\beta \geq 0$ is an inverse temperature parameter and $\boldsymbol{A} \sim \mathsf{GOE}(N)$; i.e., $\boldsymbol{A}$ is symmetric. This means that $A_{ij} \sim \mathcal{N}(0, 1/N)$ are i.i.d. for $1 \leq i \leq j \leq N$, and the diagonal entries $A_{ii} \sim \mathcal{N}(0, 2/N)$ are i.i.d. for $1 \leq i \leq N$. The parameter β is fixed and often suppressed.

In this paper, we consider the problem of efficiently sampling from the measure (I.1). Namely, we seek a randomized algorithm that accepts as input $\boldsymbol{A}$ and generates $\boldsymbol{x}^{\rm alg} \sim \mu^{\rm alg}_{\boldsymbol{A}}$, such that: (i) the algorithm runs in polynomial time for any $\boldsymbol{A}$; (ii) the distribution $\mu^{\rm alg}_{\boldsymbol{A}}$ is close to $\mu_{\boldsymbol{A}}$ for typical realizations of $\boldsymbol{A}$. Given a bounded distance $\mathsf{dist}(\mu, \nu)$ between probability distributions μ, ν, this can be formalized by requiring $\mathbb{E}[\mathsf{dist}(\mu_{\boldsymbol{A}}, \mu^{\rm alg}_{\boldsymbol{A}})] = o_N(1)$.

Gibbs sampling (also known in this context as Glauber dynamics) provides an algorithm to approximately sample from $\mu_{\boldsymbol{A}}$. However, standard techniques to bound its mixing time (e.g., Dobrushin condition [AH87]) only imply polynomial mixing for a vanishing interval of temperatures $\beta = O(N^{-1/2})$. By contrast, physicists [SZ81], [MPV87] predict fast convergence to equilibrium (at least for certain observables) for all $\beta < 1$.

Significant progress on this question was achieved only recently. In [BB19], Bauerschmidt and Bodineau showed that, for $\beta < 1/4$, the measure $\mu_{\boldsymbol{A}}$ can be decomposed into a log-concave mixture of product measures. They use this decomposition to prove that $\mu_{\boldsymbol{A}}$ satisfies a log-Sobolev inequality, although not for the Dirichlet form of Glauber dynamics.[1] [EKZ21] prove

A.M. was supported by ONR grant N00014-18-1-2729. M.S. was supported by graduate research fellowships from the NSF and Stanford. A.M. and M.S. were supported by NSF award CCF-2006489. Part of this work was conducted while the authors were visiting the Simons Institute for the Theory of Computing.

[1]Their result immediately suggests a sampling algorithm: sample from the log-concave mixture using Langevin dynamics, and then from the corresponding component using the product formula.

978-1-6654-5519-0/22/$31.00 ©2022 IEEE
DOI 10.1109/FOCS54457.2022.00038

that, in the same region $\beta < 1/4$, $\mu_{\boldsymbol{A}}$ satisfies a Poincaré inequality for the Dirichlet form of Glauber dynamics. Hence Glauber dynamics mixes in $O(N^2)$ spin flips in total variation distance. This mixing time estimate was improved to $O(N \log N)$ by [AJK+21] using a modified log Sobolev inequality, see also [CE22, Corollary 51]. The aforementioned results apply deterministically to any matrix $\boldsymbol{A}$ satisfying $\beta(\lambda_{\max}(\boldsymbol{A}) - \lambda_{\min}(\boldsymbol{A})) \le 1 - \varepsilon$ (for constant $\varepsilon > 0$).

For *spherical* spin glasses, it is shown in [GJ19] that Langevin dynamics have a polynomial spectral gap at high temperature. Meanwhile [BAJ18] proves that at sufficiently low temperature and under an overlap gap condition, the mixing times of Glauber and Langevin dynamics are exponentially large in Ising and spherical spin glasses, respectively.

In this paper we develop a different approach which is not based on a Monte Carlo Markov Chain strategy. We build on the well known remark that approximate sampling can be reduced to approximate computation of expectations of the measure $\mu_{\boldsymbol{A}}$, and of a family of measures obtained from $\mu_{\boldsymbol{A}}$. One well known method to achieve this reduction is via sequential sampling [JVV86], [CDHL05], [BD11]. A sequential sampling approach to $\mu_{\boldsymbol{A}}$ would proceed as follows. Order the variables $x_1, \dots, x_N \in \{-1, +1\}$ arbitrarily. At step i compute the marginal distribution of x_i, conditional to $x_1, \dots, x_{i-1}$ taking the previously chosen values: $p_s^{(i)} := \mu_{\boldsymbol{A}}(x_i = s | x_1, \dots, x_{i-1})$, $s \in \{-1, +1\}$. Fix $x_i = +1$ with probability $p_{+1}^{(i)}$ and $x_i = -1$ with probability $p_{-1}^{(i)}$.

We follow a different route, which is similar in spirit, but that we find more convenient technically, and of potential practical interest. Our approach is motivated by the stochastic localization process [Eld20]. Given any probability measure μ on $\mathbb{R}^N$ with finite second moment, positive time $t > 0$, and vector $\boldsymbol{y} \in \mathbb{R}^N$, define the tilted measure

$$\mu_{\boldsymbol{y},t}(\mathrm{d}\boldsymbol{x}) := \frac{1}{Z(\boldsymbol{y})} e^{\langle \boldsymbol{y}, \boldsymbol{x} \rangle - \frac{t}{2}\|\boldsymbol{x}\|_2^2} \, \mu(\mathrm{d}\boldsymbol{x}) \,, \tag{I.2}$$

and let its mean vector be

$$\boldsymbol{m}(\boldsymbol{y}, t) := \int_{\mathbb{R}^N} \boldsymbol{x} \, \mu_{\boldsymbol{y},t}(\mathrm{d}\boldsymbol{x}) \,. \tag{I.3}$$

Consider the stochastic differential equation[2] (SDE)

$$\mathrm{d}\boldsymbol{y}(t) = \boldsymbol{m}(\boldsymbol{y}(t), t)\mathrm{d}t + \mathrm{d}\boldsymbol{B}(t), \qquad \boldsymbol{y}(0) = 0 \,, \tag{I.4}$$

[2]If μ is has finite variance, then $\boldsymbol{y} \to \boldsymbol{m}(\boldsymbol{y}, t)$ is Lipschitz and so this SDE is well posed with unique strong solution.

where $(\boldsymbol{B}(t))_{t \ge 0}$ is a standard Brownian motion in $\mathbb{R}^N$. Then, the measure-valued process $(\mu_{\boldsymbol{y}(t),t})_{t \ge 0}$ is a martingale and (almost surely) $\mu_{\boldsymbol{y}(t),t} \Rightarrow \delta_{\boldsymbol{x}^\star}$ as $t \to \infty$, for some random $\boldsymbol{x}^\star$ (i.e. the measure localizes). As a consequence of the martingale property, $\mathbb{E}[\int \varphi(\boldsymbol{x}) \mu_{\boldsymbol{y}(t),t}(\mathrm{d}\boldsymbol{x})]$ is a constant for any bounded continuous function φ, whence $\mathbb{E}[\varphi(\boldsymbol{x}^\star)] = \int \varphi(\boldsymbol{x}) \mu(\mathrm{d}\boldsymbol{x})$. In other words, $\boldsymbol{x}^\star$ is a sample from μ. For further information on this process, we refer to Section III.

In order to use this process as an algorithm to sample from the SK measure $\mu = \mu_{\boldsymbol{A}}$, we need to overcome two problems:

- *Discretization.* We need to discretize the SDE (I.4) in time, and still guarantee that the discretization closely tracks the original process. This is of course possible only if the map $\boldsymbol{y} \mapsto \boldsymbol{m}(\boldsymbol{y}, t)$ is sufficiently regular.
- *Mean computation.* We need to be able to compute the mean vector $\boldsymbol{m}(\boldsymbol{y}, t)$ efficiently. To this end, we use an approximate message passing (AMP) algorithm for which we can leverage earlier work [DAM17] to establish that $\|\boldsymbol{m}(\boldsymbol{y}) - \widehat{\boldsymbol{m}}_{\mathsf{AMP}}(\boldsymbol{y})\|_2^2/N = o_N(1)$ along the algorithm trajectory. (Note that the SK measure is supported on vectors with $\|\boldsymbol{x}\|_2^2 = N$, and hence the quadratic component of the tilt in Eq. (I.2) drops out. We will therefore write $\boldsymbol{m}(\boldsymbol{y})$ or $\boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y})$ instead $\boldsymbol{m}(\boldsymbol{y}, t)$ for the mean of the Gibbs measure.)

To our knowledge, ours is the first algorithmic implementation of the stochastic localization process, although a recent paper by Nam, Sly and Zhang [NSZ22] uses this process (without naming it as such) to show that the Ising measure on the infinite regular tree is a factor of IID process up to a constant factor away from the Kesten–Stigum, or "reconstruction", threshold. Their construction can easily be transformed into a sampling algorithm.

In order to state our results, we define the normalized 2-Wasserstein distance between two probability measures μ, ν on $\mathbb{R}^N$ with finite second moments as

$$W_{2,N}(\mu, \nu)^2 = \inf_{\pi \in \mathcal{C}(\mu,\nu)} \frac{1}{N} \mathbb{E}_\pi \left[\left\| \boldsymbol{X} - \boldsymbol{Y} \right\|_2^2 \right], \tag{I.5}$$

where the infimum is over all couplings $(\boldsymbol{X}, \boldsymbol{Y}) \sim \pi$ with marginals $\boldsymbol{X} \sim \mu$ and $\boldsymbol{Y} \sim \nu$.

In this paper, we establish two main results.

Sampling algorithm for $\beta < 1/2$. We prove that the strategy outlined above yields an algorithm with

complexity $O(N^2)$, which samples from a distribution $\mu_{\boldsymbol{A}}^{\text{alg}}$ with $W_{2,N}(\mu_{\boldsymbol{A}}^{\text{alg}}, \mu_{\boldsymbol{A}}) = o_{N,\mathbb{P}}(1)$.

Hardness for stable algorithms, for $\beta > 1$. We prove that no algorithm satisfying a certain *stability* property can sample from the SK measure (under the same criterion $W_{2,N}(\mu_{\boldsymbol{A}}^{\text{alg}}, \mu_{\boldsymbol{A}}) = o_{N,\mathbb{P}}(1)$) for $\beta > 1$, i.e., when replica symmetry is broken. Roughly speaking, stability formalizes the notion that the algorithm output behaves continuously with respect to $\boldsymbol{A}$.

It is worth pointing out that we expect our algorithm to be successful (in the sense described above) for all $\beta < 1$ and that closing the gap between $\beta = 1/2$ and $\beta = 1$ should be within reach of existing techniques, at the price of a longer technical argument. We expound on this point in Remark II.1 further below, and in Section VI.

The hardness results for $\beta > 1$ are proven using the notion of disorder chaos, in a similar spirit to the use of the *overlap gap property* for random optimization, estimation, and constraint satisfaction problems [GS14], [RV17], [GS17], [CGPR19], [GJW20], [Wei22], [GK21], [BH21], [HS21]. While the overlap gap property has been used to rule out stable algorithms for this class of problems, and variants have been used to rule out efficient sampling by specific Markov chain algorithms, to the best of our knowledge we are the first to rule out stable sampling algorithms using these ideas. In sampling there is no hidden solution or set of solutions to be found, and therefore no notion of an overlap gap in the most natural sense. Instead, we argue directly that the distribution to be sampled from is unstable in a $W_{2,N}$ sense at low temperature, and hence cannot be approximated by any stable algorithm.

II. Main Results

A. Sampling algorithm for $\beta < 1/2$

In this section we describe the sampling algorithm, and formally state the result of our analysis. As pointed out in the introduction, a main component is the computation of the mean of the tilted SK measure:

$$\mu_{\boldsymbol{A},\boldsymbol{y}}(\boldsymbol{x}) = \frac{\exp\left(\frac{\beta}{2}\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle + \langle \boldsymbol{y}, \boldsymbol{x}\rangle\right)}{Z(\boldsymbol{A}, \boldsymbol{y})} \qquad \text{(II.1)}$$
$$\boldsymbol{x} \in \{-1, +1\}^N .$$

We describe the algorithm to approximate this mean in Section II-A1, the overall sampling procedure (which uses this estimator as a subroutine) in Section II-A2, and our Wasserstein-distance guarantee in Section II-A3.

Algorithm 1: Mean of the tilted Gibbs measure

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

$\widehat{\boldsymbol{m}}^{-1} = \boldsymbol{z}^0 = 0$,
for $k = 0, \cdots, K_{\mathsf{AMP}} - 1$ **do**
  $\widehat{\boldsymbol{m}}^k = \tanh(\boldsymbol{z}^k)$, $\quad \mathsf{b}_k = \frac{\beta^2}{N}\sum_{i=1}^N \left(1 - \tanh^2(z_i^k)\right)$,
  $\boldsymbol{z}^{k+1} = \beta \boldsymbol{A}\widehat{\boldsymbol{m}}^k + \boldsymbol{y} - \mathsf{b}_k \widehat{\boldsymbol{m}}^{k-1}$,
$\boldsymbol{u}^0 = \boldsymbol{z}^{K_{\mathsf{AMP}}}$,
for $k = 0, \cdots, K_{\mathsf{NGD}} - 1$ **do**
  $\boldsymbol{u}^{k+1} = \boldsymbol{u}^k - \eta \cdot \nabla \mathscr{F}_{\mathsf{TAP}}(\widehat{\boldsymbol{m}}^{+,k}; \boldsymbol{y}, q)$,
  $\widehat{\boldsymbol{m}}^{+,k+1} = \tanh(\boldsymbol{u}^{k+1})$,
return $\widehat{\boldsymbol{m}}^{+,K_{\mathsf{NGD}}}$

1) Approximating the mean of the Gibbs measure: We will denote our approximation of the mean of the Gibbs measure $\mu_{\boldsymbol{A},\boldsymbol{y}}$ by $\widehat{\boldsymbol{m}}(\boldsymbol{A}, \boldsymbol{y})$, while the actual mean will be $\boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y})$.

The algorithm to compute $\widehat{\boldsymbol{m}}(\boldsymbol{A}, \boldsymbol{y})$ is given in Algorithm 1, and is composed of two phases:

1) An Approximate Message Passing (AMP) algorithm is run for K_{AMP} iterations and constructs a first estimate of the mean. We denote by $\mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; k)$ the estimate produced after k AMP iterations
$$\mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; k) := \widehat{\boldsymbol{m}}^k . \qquad \text{(II.2)}$$
2) Natural gradient descent (NGD) is run for K_{NGD} iterations with initialization given by vector computed at the end of the first phase. This phase attempts to minimize the following version of the TAP free energy (for a specific value of q):
$$\mathscr{F}_{\mathsf{TAP}}(\boldsymbol{m}; \boldsymbol{y}, q) := -\frac{\beta}{2}\langle \boldsymbol{m}, \boldsymbol{A}\boldsymbol{m}\rangle - \langle \boldsymbol{y}, \boldsymbol{m}\rangle - \sum_{i=1}^N h(m_i) - \frac{N\beta^2(1-q)(1+q-2Q(\boldsymbol{m}))}{4},$$
$$Q(\boldsymbol{m}) = \frac{1}{N}\|\boldsymbol{m}\|^2,$$
$$h(m) = -\frac{1+m}{2}\log\left(\frac{1+m}{2}\right) - \frac{1-m}{2}\log\left(\frac{1-m}{2}\right).$$

The second stage is motivated by the TAP (Thouless-Anderson-Palmer) equations for the Gibbs mean of a

high-temperature spin glass [MPV87], [Tal10]. Essentially by construction, stationary points for the function $\mathscr{F}_{\text{TAP}}(\boldsymbol{m};\boldsymbol{y},q)$ satisfy the TAP equations, and we show that the first stage above constructs an approximate stationary point for $\mathscr{F}_{\text{TAP}}(\boldsymbol{m};\boldsymbol{y},q)$. The effect of the second stage is therefore numerically small, but it turns out to reduce the error incurred by discretizing time in line 6 of Algorithm 2.

Let us emphasize that this two-stage construction is considered for technical reasons. Indeed a simpler algorithm, that runs AMP for a larger number of iteration, and does not run NGD at all, is expected to work but our arguments do not go through. The hybrid algorithm above allows us to exploit known properties of AMP (precise analysis via state evolution) and $\mathscr{F}_{\text{TAP}}(\boldsymbol{m};\boldsymbol{y},q)$ (Lipschitz continuity of the minimizer in $\boldsymbol{y}$).

Algorithm 2: APPROXIMATE SAMPLING FROM THE SK GIBBS MEASURE

Input: Data $\boldsymbol{A}\in\mathbb{R}^{N\times N}$, parameters $(\beta,\eta,K_{\mathsf{AMP}},K_{\mathsf{NGD}},L,\delta)$

1 $\widehat{\boldsymbol{y}}_0=0$,
2 **for** $\ell=0,\cdots,L-1$ **do**
3 Draw $\boldsymbol{w}_{\ell+1}\sim\mathcal{N}(0,\boldsymbol{I}_N)$ independent of everything so far;
4 Set $q=q_*(\beta,t=\ell\delta)$;
5 Set $\widehat{\boldsymbol{m}}(\boldsymbol{A},\widehat{\boldsymbol{y}}_\ell)$ the output of Algorithm 1, with parameters $(\beta,\eta,q,K_{\mathsf{AMP}},K_{\mathsf{NGD}})$;
6 Update $\widehat{\boldsymbol{y}}_{\ell+1}=\widehat{\boldsymbol{y}}_\ell+\widehat{\boldsymbol{m}}(\boldsymbol{A},\widehat{\boldsymbol{y}}_\ell)\,\delta+\sqrt{\delta}\,\boldsymbol{w}_{\ell+1}$
7 Set $\widehat{\boldsymbol{m}}(\boldsymbol{A},\widehat{\boldsymbol{y}}_L)$ the output of Algorithm 1, with parameters $(\eta,q,K_{\mathsf{AMP}},K_{\mathsf{NGD}})$;
8 Draw $\{x_i^{\text{alg}}\}_{i\le N}$ conditionally independent with $\mathbb{E}[x_i^{\text{alg}}|\boldsymbol{y},\{\boldsymbol{w}_\ell\}]=\widehat{m}_i(\boldsymbol{A},\widehat{\boldsymbol{y}}_L)$
9 **return** $\boldsymbol{x}^{\text{alg}}$

2) Sampling via stochastic localization: Our sampling algorithm is presented as Algorithm 2. The algorithm makes uses of constants $q_k:=q_k(\beta,t)$. With $W\sim\mathcal{N}(0,1)$ a standard Gaussian, these constants are defined for $k,\beta,t\ge 0$ by the recursion

$$q_{k+1}=\mathbb{E}\Big[\tanh\big(\beta^2 q_k+t+\sqrt{\beta^2q_k+t}\,W\big)^2\Big],$$
$$q_0=0\,,\quad q_*=\lim_{k\to\infty}q_k\,.$$

This iteration can be implemented via a one-dimensional integral, and the limit q_* is approached exponentially fast in k (see Lemma V.3 below). The values $q_*(\beta,t=\ell\delta)$ for $\ell\in\{0,\dots,L\}$ can be precomputed and are independent of the input $\boldsymbol{A}$. For the sake of simplicity, we will neglect errors in this calculation.

The core of the sampling procedure is step 6, which is a standard Euler discretization of the SDE (I.4), with step size δ, over the time interval $[0,T]$, $T=L\delta$. The mean of the Gibbs measure $\boldsymbol{m}(\boldsymbol{A},\boldsymbol{y})$ is replaced by the output of Algorithm 1 which we recall is denoted by $\widehat{\boldsymbol{m}}(\boldsymbol{A},\boldsymbol{y})$. We reproduce the Euler iteration here for future reference

$$\widehat{\boldsymbol{y}}_{\ell+1}=\widehat{\boldsymbol{y}}_\ell+\widehat{\boldsymbol{m}}(\boldsymbol{A},\widehat{\boldsymbol{y}}_\ell)\,\delta+\sqrt{\delta}\,\boldsymbol{w}_{\ell+1}\,. \tag{II.3}$$

The output of the iteration is $\widehat{\boldsymbol{m}}(\boldsymbol{A},\widehat{\boldsymbol{y}}_L)$, which should be thought of as an approximation of $\boldsymbol{m}(\boldsymbol{A},\boldsymbol{y}(T))$, $T=L\delta$, that is the mean of $\mu_{\boldsymbol{A},\boldsymbol{y}(T)}$. According to the discussion in the introduction, for large T, $\mu_{\boldsymbol{A},\boldsymbol{y}(T)}$ concentrates around $\boldsymbol{x}^\star\sim\mu_{\boldsymbol{A}}$. In other words, $\boldsymbol{m}(\boldsymbol{A},\boldsymbol{y}(T))$ is close to the corner $\boldsymbol{x}^\star$ of the hypercube. We round its coordinates independently to produce the output $\boldsymbol{x}^{\text{alg}}$.

3) Theoretical guarantee: Our main positive result is the following.

Theorem II.1. *For any $\varepsilon>0$ and $\beta_0<1/2$ there exist $\eta,K_{\mathsf{AMP}},K_{\mathsf{NGD}},L,\delta$ independent of N, so that the following holds for all $\beta\le\beta_0$. The sampling algorithm 2 takes as input $\boldsymbol{A}$ and parameters $(\eta,K_{\mathsf{AMP}},K_{\mathsf{NGD}},L,\delta)$ and outputs a random point $\boldsymbol{x}^{\text{alg}}\in\{-1,+1\}^N$ with law $\mu_{\boldsymbol{A}}^{\text{alg}}$ such that with probability $1-o_N(1)$ over $\boldsymbol{A}\sim\mathsf{GOE}(N)$,*

$$W_{2,N}(\mu_{\boldsymbol{A}}^{\text{alg}},\mu_{\boldsymbol{A}})\le\varepsilon\,. \tag{II.4}$$

The total complexity of this algorithm is $O(N^2)$.

Remark II.1. The condition $\beta<1/2$ arises because our proof requires the Hessian of the TAP free energy to be positive definite at its minimizer. A simple calculation yields

$$\nabla^2\mathscr{F}_{\text{TAP}}(\boldsymbol{m};\boldsymbol{y},q)=-\beta\boldsymbol{A}+\boldsymbol{D}(\boldsymbol{m})+\beta^2(1-q)\,\boldsymbol{I}_N\,,$$
$$\boldsymbol{D}(\boldsymbol{m}):=\operatorname{diag}\big(\{(1-m_i^2)^{-1}\}_{i\le N}\big).$$

A crude bound yields $\nabla^2\mathscr{F}_{\text{TAP}}(\boldsymbol{m};\boldsymbol{y},q)\succeq-\beta\boldsymbol{A}+\boldsymbol{I}_N\succeq(1-\beta\lambda_{\max}(\boldsymbol{A}))\boldsymbol{I}_N$. Since $\text{p-lim}_{N\to\infty}\lambda_{\max}(\boldsymbol{A})=2$ the desired condition holds trivially for $\beta<1/2$. However, we expect that a more careful treatment will reveal that the Hessian is locally positive in a neighborhood of the minimizer for all $\beta<1$.

After an initial version of this manuscript was made public, Celentano [Cel22] showed the above-mentioned local strong convexity of the TAP free energy for *all* $\beta<1$, thereby confirming that our sampling procedure succeeds up to the critical temperature.

B. Hardness for stable algorithms, for $\beta > 1$

The sampling algorithm 2 enjoys stability properties with respect to changes in the inverse temperature β and the matrix $\boldsymbol{A}$ which are shared by many natural efficient algorithms. We will use the fact that the actual Gibbs measure does not enjoy this stability property for $\beta > 1$ to conclude that sampling is hard for all stable algorithms.

Throughout this section, we denote the Gibbs and algorithmic output distributions by $\mu_{\boldsymbol{A},\beta}$ and $\mu^{\text{alg}}_{\boldsymbol{A},\beta}$ respectively to emphasize the dependence on β.

Definition II.1. *Let $\{\mathsf{ALG}_N\}_{N\geq 1}$ be a family of randomized sampling algorithms, i.e., measurable maps*

$$\mathsf{ALG}_N : (\boldsymbol{A}, \beta, \omega) \mapsto \mathsf{ALG}_N(\boldsymbol{A}, \beta, \omega) \in [-1, 1]^N ,$$

where ω is a random seed (a point in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$). Let $\boldsymbol{A}'$ and $\boldsymbol{A} \sim \mathsf{GOE}(N)$ be independent copies of the coupling matrix, and consider perturbations $\boldsymbol{A}_s = \sqrt{1-s^2}\boldsymbol{A} + s\boldsymbol{A}'$ for $s \in [0, 1]$. Finally, denote by $\mu^{\text{alg}}_{\boldsymbol{A}_s,\beta}$ the law of the algorithm output, i.e., the distribution of $\mathsf{ALG}_N(\boldsymbol{A}_s, \beta, \omega)$ when $\omega \sim \mathbb{P}$ independent of $\boldsymbol{A}_s, \beta$ which are fixed.

We say ALG_N is stable with respect to disorder*, at inverse temperature β, if*

$$\lim_{s\to 0} \operatorname*{p-lim}_{N\to\infty} W_{2,N}(\mu^{\text{alg}}_{\boldsymbol{A},\beta}, \mu^{\text{alg}}_{\boldsymbol{A}_s,\beta}) = 0 . \tag{II.5}$$

We say ALG_N is stable with respect to temperature *at inverse temperature β, if*

$$\lim_{\beta'\to\beta} \operatorname*{p-lim}_{N\to\infty} W_{2,N}(\mu^{\text{alg}}_{\boldsymbol{A},\beta}, \mu^{\text{alg}}_{\boldsymbol{A},\beta'}) = 0 . \tag{II.6}$$

As proved in the full version, Algorithm 2 is stable.

Theorem II.2. *2] For any $\beta \in (0, \infty)$ and fixed parameters (η, K_{AMP}, K_{NGD}, L, δ), Algorithm 2 is stable with respect to disorder and with respect to temperature.*

As a consequence, the Gibbs measures $\mu_{\boldsymbol{A},\beta}$ enjoy similar stability properties for $\beta < 1/2$, which amount (as discussed below) to the absence of chaos in both temperature and disorder:

Corollary II.2. *For any $\beta < 1/2$, the following properties hold for the Gibbs measure $\mu_{\boldsymbol{A},\beta}$ of the Sherrington-Kirkpatrick model, cf. Eq.* (I.1)*:*

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

2) $\lim_{\beta'\to\beta} \text{p-lim}_{N\to\infty} W_{2,N}(\mu_{\boldsymbol{A},\beta}, \mu_{\boldsymbol{A},\beta'}) = 0.$

Proof. Take $\varepsilon > 0$ arbitrarily small and choose parameters $(\eta, K_{\mathsf{AMP}}, K_{\mathsf{NGD}}, L, \delta)$ of Algorithm 2 with the desired tolerance ε so that Theorem II.1 holds. Combining with Theorem II.2 using the same parameters $(\eta, K_{\mathsf{AMP}}, K_{\mathsf{NGD}}, L, \delta)$ implies the result since ε is arbitrarily small. (Recall that $(\eta, K_{\mathsf{AMP}}, K_{\mathsf{NGD}}, L, \delta)$ can be chosen independent of β for $\beta \leq \beta_0 < 1/2$.) □

Remark II.2. We emphasize that Corollary II.2 makes no reference to an algorithm, and is instead a purely structural property of the Gibbs measure. The sampling algorithm, however, is the key tool of our proof.

Stability is related to chaos, which is a well studied and important property of spin glasses, see e.g. [Cha09], [Che13], [Cha14], [CHHS15], [CP18]. In particular, "disorder chaos" refers to the following phenomenon. Draw $\boldsymbol{x}^0 \sim \mu_{\boldsymbol{A},\beta}$ independently of $\boldsymbol{x}^s \sim \mu_{\boldsymbol{A}_s,\beta}$, and denote by $\mu^{(0,s)}_{\boldsymbol{A},\beta} := \mu_{\boldsymbol{A},\beta} \otimes \mu_{\boldsymbol{A}^s,\beta}$ their joint distribution. Disorder chaos holds at inverse temperature β if

$$\lim_{s\to 0} \lim_{N\to\infty} \mathbb{E}\, \mu^{(0,s)}_{\boldsymbol{A},\beta} \Big\{ \Big(\frac{1}{N} \langle \boldsymbol{x}^0, \boldsymbol{x}^s \rangle \Big)^2 \Big\} = 0 . \tag{II.7}$$

Note that disorder chaos is not necessarily a surprising property. For instance when $\beta = 0$, the distribution $\mu_{\boldsymbol{A}_s,\beta}$ is simply the uniform measure over the hypercube $\{-1, +1\}^N$ for all s, and this example exhibits disorder chaos in the sense of Eq. (II.7). In fact, the SK Gibbs measure exhibits disorder chaos at all $\beta \in [0, \infty)$ [Cha09]. However, for $\beta > 1$, Eq. (II.7) leads to a stronger conclusion.

Theorem II.3 (Disorder chaos in $W_{2,N}$ distance)**.** *For all $\beta > 1$,*

$$\inf_{s\in(0,1)} \liminf_{N\to\infty} \mathbb{E}\left[W_{2,N}(\mu_{\boldsymbol{A},\beta}, \mu_{\boldsymbol{A}_s,\beta}) \right] > 0 .$$

Finally, we obtain the desired hardness result by reversing the implication in Corollary II.2: no stable algorithm which can approximately sample from the measure $\mu_{\boldsymbol{A},\beta}$ in the $W_{2,N}$ sense for $\beta > 1$.

Theorem II.4. *Fix $\beta > 1$, and let $\{\mathsf{ALG}_N\}_{N\geq 1}$ be a family of randomized algorithms which is stable with respect to disorder as per Definition II.1 at inverse temperature β. Let $\mu^{\text{alg}}_{\boldsymbol{A},\beta}$ be the law of the output $\mathsf{ALG}_N(\boldsymbol{A}, \beta, \omega)$ conditional on $\boldsymbol{A}$. Then*

$$\liminf_{N\to\infty} \mathbb{E}\left[W_{2,N}(\mu^{\text{alg}}_{\boldsymbol{A},\beta},\ \mu_{\boldsymbol{A},\beta}) \right] > 0 .$$

We refer the reader to the full version for the proof. Let us remark that while our sampling algorithm allows us to conclude stability in both disorder and temperature,

only chaos in disorder is used to prove hardness. Indeed, temperature stability is known to hold below the replica-symmetry breaking temperature in pure spherical spin glasses [Sub17].

C. Notations

We use $o_N(1)$ to indicate a quantity tending to 0 as $N \to \infty$. We use $o_{N,\mathbb{P}}(1)$ for a quantity tending to 0 in probability. If X is a random variable, then $\mathcal{L}(X)$ indicates its law. The quantity $C(\beta)$ refers to a constant depending on β. The uniform distribution on the interval $[a,b]$ is denoted by $\mathsf{Unif}([a,b])$.

D. Outline

The rest of the paper is organized as follows. Section IV introduces the planted model and its contiguity with the original model. We then analyze the AMP component of our algorithm in Section V, and the NGD component in Section VI. Finally, Section VII puts the various elements together and proves Theorem II.1.

III. Properties of Stochastic Localization

We collect in this section the main properties of the stochastic localization process needed for our analysis. We focus on the Gibbs measure (I.1), although most of the below generalizes to other probability measures in $\mathbb{R}^N$, under suitable tail conditions. Throughout this section, the matrix $\boldsymbol{A}$ is viewed as fixed.

Recalling the tilted measure $\mu_{\boldsymbol{A},\boldsymbol{y}}$ of Eq. (I.2), and the SDE of Eq. (I.4), we introduce the shorthand

$$\mu_t = \mu_{\boldsymbol{A},\boldsymbol{y}(t)}\,.$$

The following properties are well known. See for instance [ES22, Propositions 9, 10] or [Eld20].

Lemma III.1. *For all $t \geq 0$ and all $\boldsymbol{x} \in \{-1,+1\}^N$,*

$$\mathrm{d}\mu_t(\boldsymbol{x}) = \mu_t(\boldsymbol{x})\langle \boldsymbol{x} - \boldsymbol{m}_{\boldsymbol{A},\boldsymbol{y}(t)}, \mathrm{d}\boldsymbol{B}(t)\rangle\,. \tag{III.1}$$

As a consequence, for any function $\varphi : \mathbb{R}^N \to \mathbb{R}^m$, the process $\big(\mathbb{E}_{\boldsymbol{x}\sim\mu_t}\big[\varphi(\boldsymbol{x})\big]\big)_{t\geq 0}$ is a martingale.

Lemma III.2 ([Eld20])**.** *For all $t > 0$,*

$$\mathbb{E}\,\mathrm{cov}(\mu_t) \preceq \frac{1}{t}\boldsymbol{I}_N\,. \tag{III.2}$$

Lemma III.3. *For all $t > 0$,*

$$W_{2,N}\big(\mu_{\boldsymbol{A}}, \mathcal{L}(\boldsymbol{m}_{\boldsymbol{A},\boldsymbol{y}(t)})\big)^2 \leq \frac{1}{t}\,. \tag{III.3}$$

In particular, the mean vector $\boldsymbol{m}_{\boldsymbol{A},\boldsymbol{y}(t)}$ converges in distribution to a random vector $\boldsymbol{x}^\star \sim \mu_{\boldsymbol{A}}$ as $t \to \infty$.

IV. The planted model and contiguity

Let $\overline{\nu}$ be the uniform distribution over $\{-1,+1\}^N$ and consider the joint distribution of pairs $(\boldsymbol{x},\boldsymbol{A}) \in \{-1,+1\}^N \times \mathbb{R}^{N\times N}_{\text{sym}}$,

$$\mu_{\text{pl}}(\mathrm{d}\boldsymbol{x},\mathrm{d}\boldsymbol{A}) = \frac{1}{Z_{\text{pl}}}\,\exp\Big\{-\frac{N}{4}\Big\|\boldsymbol{A}-\frac{\beta\boldsymbol{x}\boldsymbol{x}^\top}{N}\Big\|_F^2\Big\}\,\overline{\nu}(\mathrm{d}\boldsymbol{x})\,\mathrm{d}\boldsymbol{A}\,, \tag{IV.1}$$

where $\mathrm{d}\boldsymbol{A}$ is the Lebesgue measure over the space of symmetric matrices $\mathbb{R}^{N\times N}_{\text{sym}}$, and the normalizing constant

$$Z_{\text{pl}} := \int \exp\Big\{-\frac{N}{4}\Big\|\boldsymbol{A}-\frac{\beta\boldsymbol{x}\boldsymbol{x}^\top}{N}\Big\|_F^2\Big\}\,\mathrm{d}\boldsymbol{A} \tag{IV.2}$$

is independent of $\boldsymbol{x} \in \{-1,+1\}^N$. It is easy to see by construction that the marginal distribution of $\boldsymbol{x}$ under μ_{pl} is $\overline{\nu}$, and the conditional law $\mu_{\text{pl}}(\,\cdot\,|\boldsymbol{x})$ is a rank-one spiked GOE model with spike $\beta\boldsymbol{x}\boldsymbol{x}^\top/N$. Namely, under $\mu_{\text{pl}}(\,\cdot\,|\boldsymbol{x})$, we have

$$\boldsymbol{A} = \frac{\beta}{N}\boldsymbol{x}\boldsymbol{x}^\top + \boldsymbol{W}\,,\quad \boldsymbol{W} \sim \mathsf{GOE}(N)\,. \tag{IV.3}$$

On the other hand, $\mu_{\text{pl}}(\,\cdot\,|\boldsymbol{A})$ is the SK measure $\mu_{\boldsymbol{A}}$.

The marginal of $\boldsymbol{A}$ under μ_{pl} is not the $\mathsf{GOE}(N)$ distribution μ_{GOE} but takes the form

$$\mu_{\text{pl}}(\mathrm{d}\boldsymbol{A}) = \frac{1}{Z_{\text{pl}}}\,e^{-\frac{N}{4}\|\boldsymbol{A}\|_F^2}\,Z_{\mathsf{SK}}(\boldsymbol{A})\,\mathrm{d}\boldsymbol{A} \tag{IV.4}$$

$$= \mu_{\mathsf{GOE}}(\mathrm{d}\boldsymbol{A})\,Z_{\mathsf{SK}}(\boldsymbol{A})\,, \tag{IV.5}$$

where $Z_{\mathsf{SK}}(\boldsymbol{A})$ is the (rescaled) partition function of the SK measure

$$Z_{\mathsf{SK}}(\boldsymbol{A}) = 2^{-n}\sum_{\boldsymbol{x}\in\{-1,+1\}^N}\exp\Big\{\frac{\beta}{2}\langle\boldsymbol{x},\boldsymbol{A}\boldsymbol{x}\rangle - \frac{\beta^2 N}{4}\Big\}\,. \tag{IV.6}$$

By a classical result [ALR87], $Z_{\mathsf{SK}}(\boldsymbol{A})$ has log-normal fluctuations for all $\beta < 1$:

Theorem IV.1 ([ALR87])**.** *Let $\beta < 1$, $\boldsymbol{A} \sim \mu_{\mathsf{GOE}}$ and $\sigma^2 = \frac{1}{4}(-\log(1-\beta^2) - \beta^2)$. Then*

$$Z_{\mathsf{SK}}(\boldsymbol{A}) \xrightarrow[n\to\infty]{\mathrm{d}} \exp(W)\,, \tag{IV.7}$$

where $W \sim \mathcal{N}\big(-\sigma^2, 2\sigma^2\big)$.

Therefore, by Le Cam's first lemma [VdV98, Lemma 6.4], $\mu_{\text{pl}}(\mathrm{d}\boldsymbol{A})$ and $\mu_{\mathsf{GOE}}(\mathrm{d}\boldsymbol{A})$ are mutually contiguous for all $\beta < 1$. For the purpose of our analysis we will need a stronger result about the joint distributions of $(\boldsymbol{A},\boldsymbol{y})$ under our "random" model and a planted model which we now introduce.

Recall that $\boldsymbol{m}(\boldsymbol{A},\boldsymbol{y})$ denotes the mean of the Gibbs measure $\mu_{\boldsymbol{A},\boldsymbol{y}}$ in Eq. (I.2). For a fixed $T \geq 0$, we define

two Borel distributions $\mathbb{P}$ (**planted**) and $\mathbb{Q}$ (**random**) on $(\boldsymbol{A}, \boldsymbol{y}) \in \mathbb{R}^{N\times N}_{\text{sym}} \times C([0,T], \mathbb{R}^N)$ as follows. For $t \in [0,T]$:

$$\mathbb{Q} \quad : \quad \begin{cases} \boldsymbol{A} & \sim & \mu_{\mathsf{GOE}}\,, \\ \boldsymbol{y}(t) & = & \displaystyle\int_0^t \boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(s))\,\mathrm{d}s + \boldsymbol{B}(t), \end{cases} \tag{IV.8}$$

$$\mathbb{P} \quad : \quad \begin{cases} \boldsymbol{x}_0 & \sim & \overline{\nu}\,, \\ \boldsymbol{A} & \sim & \mu_{\text{pl}}(\,\cdot\,|\,\boldsymbol{x}_0)\,, \\ \boldsymbol{y}(t) & = & t\boldsymbol{x}_0 + \boldsymbol{B}(t) \end{cases} \tag{IV.9}$$

where $(\boldsymbol{B}(t))_{t\ge 0}$ is a standard Brownian motion in $\mathbb{R}^N$ independent of everything else. Note the SDE defining the process $\boldsymbol{y} = (\boldsymbol{y}(t))_{t\in[0,T]}$ in Eq. (IV.8) is a restatement of the stochastic localization equation (I.4) applied to the SK measure $\mu_{\boldsymbol{A}}$. Using this it is not hard to verify the following.

Proposition IV.1. *For all $T \ge 0$ and $\beta \ge 0$, $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$ and for all $(\boldsymbol{A}, \boldsymbol{y}) \in \mathbb{R}^{N\times N}_{\text{sym}} \times C([0,T], \mathbb{R}^N)$,*

$$\frac{\mathrm{d}\,\mathbb{P}}{\mathrm{d}\,\mathbb{Q}}(\boldsymbol{A}, \boldsymbol{y}) = Z_{\mathsf{SK}}(\boldsymbol{A})\,.$$

Therefore, for all $\beta < 1$, $\mathbb{P}$ and $\mathbb{Q}$ are mutually contiguous. (I.e. for a sequence of events $\mathcal{E}_N$, $\lim_{N\to\infty}\mathbb{P}(\mathcal{E}_N) = 0$ if and only if $\lim_{N\to\infty}\mathbb{Q}(\mathcal{E}_N) = 0$.)

For the remainder of the proof of Theorem II.1, we work under the planted distribution $\mathbb{P}$. All results proven under $\mathbb{P}$ transfer to $\mathbb{Q}$ by contiguity.

V. Approximate Message Passing

In this section we analyze the AMP iteration of Algorithm 1, which we copy here for the reader's convenience

$$\widehat{\boldsymbol{m}}^{-1} = \boldsymbol{z}^0 = 0, \quad \widehat{\boldsymbol{m}}^k = \tanh(\boldsymbol{z}^k) \tag{V.1}$$
$$\mathsf{b}_k = \frac{\beta^2}{N}\sum_{i=1}^N \big(1 - \tanh^2(z_i^k)\big) \quad \forall k \ge 0\,,$$
$$\boldsymbol{z}^{k+1} = \beta \boldsymbol{A}\widehat{\boldsymbol{m}}^k + \boldsymbol{y} - \mathsf{b}_k\widehat{\boldsymbol{m}}^{k-1}\,.$$

When needed, we will specify the dependence on $\boldsymbol{A}, \boldsymbol{y}$ by writing $\widehat{\boldsymbol{m}}^k = \widehat{\boldsymbol{m}}^k(\boldsymbol{A}, \boldsymbol{y}) = \mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; k)$ and $\boldsymbol{z}^k = \boldsymbol{z}^k(\boldsymbol{A}, \boldsymbol{y})$. Throughout this section $(\boldsymbol{A}, \boldsymbol{y}) \sim \mathbb{P}$ will be distributed according to the planted model introduced above.

Our analysis will be based on the general state evolution result of [BM11], [JM13], which implies the following asymptotic characterization for the iterates. Set $\gamma_0(\beta, t) = 0, \Sigma_{0,i}(\beta, t) = 0$ and recursively define

$$\gamma_{k+1}(\beta, t) = \beta^2 \cdot \mathbb{E}\left[\tanh\left(\gamma_k(\beta, t) + t + G_k\right)\right], \tag{V.2}$$
$$\Sigma_{k+1,j+1}(\beta, t) = \beta^2 \cdot \mathbb{E}\left[\tanh\left(\gamma_k(\beta, t) + t + G_k\right) \cdot \tanh\left(\gamma_j(\beta, t) + t + G_j\right)\right], \tag{V.3}$$

where $(\boldsymbol{G}_j)_{j\le k}$ are jointly Gaussian, with zero mean and covariance $\boldsymbol{\Sigma}_{\le k} + t\mathbf{1}\mathbf{1}^\top$, $\boldsymbol{\Sigma}_{\le k} := (\Sigma_{ij})_{i,j\le k}$.

Proposition V.1 (Theorem 1 of [BM11])**.** *For $(\boldsymbol{A}, \boldsymbol{y}) \sim \mathbb{P}$ and any $k \in \mathbb{Z}_{\ge 0}$, the empirical distribution of the coordinate of the AMP iterates converges almost surely in $W_2(\mathbb{R}^{k+2})$ as follows:*

$$\frac{1}{N}\sum_{i=1}^N \delta_{(z_i^1, \cdots, z_i^k, x_i, y_i)} \xrightarrow[n\to\infty]{W_2}$$
$$\mathcal{L}\left(\boldsymbol{\gamma}_{\le k}(\beta, t)X + \boldsymbol{G} + Y\mathbf{1}, X, Y\right);$$
$$\boldsymbol{\gamma}_{\le k}(\beta, t) = \big(\gamma_1(\beta, t), \ldots, \gamma_k(\beta, t)\big), \quad \boldsymbol{G} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\le k})\,.$$

On the right-hand side, X is uniformly random in $\{-1, +1\}$, $Y = tX + \sqrt{t}W$ where $W \sim \mathcal{N}(0,1)$ and $X, \boldsymbol{G}, W$ are mutually independent.

As in [DAM17, Eqs. (69, 70)] we argue that the state evolution equations (V.2), (V.3) take a simple form thanks to our specific choice of AMP non-linearity $\tanh(\cdot)$. It will be convenient to use the notations

$$\widetilde{\gamma}_k(\beta, t) = \gamma_k(\beta, t) + t\,,$$
$$\widetilde{\Sigma}_{k,j}(\beta, t) = \Sigma_{k,j}(\beta, t) + t\,.$$

Proposition V.2. *For any $t \in \mathbb{R}_{\ge 0}$ and $k, j \in \mathbb{Z}_{\ge 0}$,*

$$\Sigma_{k,j}(\beta, t) = \gamma_{k\wedge j}(\beta, t)\,, \quad \text{and} \quad \widetilde{\Sigma}_{k,j}(\beta, t) = \widetilde{\gamma}_{k\wedge j}(\beta, t)\,.$$

Proof. The two claims are equivalent and we proceed by induction. The base case $k = 0$ holds by definition, so we may assume $\Sigma_{i,j}(\beta, t) = \gamma_{i\wedge j}(\beta, t)$ for $i, j \le k-1$. Set $Z_j = \gamma_j X + G_j$ where $\widetilde{\boldsymbol{G}} \sim \mathcal{N}(0, \widetilde{\boldsymbol{\Sigma}}_{\le k-1})$. Note that, by the induction hypothesis, Z_{k-1} is a sufficient statistic for X given $(Z_j)_{j\le k-1}$. Using Bayes' rule, and writing $\widetilde{\sigma}_{k-1}^2 := \widetilde{\Sigma}_{k-1,k-1}$, one easily computes that $\mathbb{E}[X|Z_{k-1}]$ equals

$$\frac{e^{\widetilde{\gamma}_{k-1}Z_{k-1}/\widetilde{\sigma}_{k-1}^2} - e^{-\widetilde{\gamma}_{k-1}Z_{k-1}/\widetilde{\sigma}_{k-1}^2}}{e^{\widetilde{\gamma}_{k-1}Z/\widetilde{\sigma}_{k-1}^2} + e^{-\widetilde{\gamma}_{k-1}Z/\widetilde{\sigma}_{k-1}^2}} = \tanh(Z_{k-1})\,.$$

Therefore using Eq. (V.2), the fact that $\tanh$ is an odd function and $WX \stackrel{\mathrm{d}}{=} W$,

$$\begin{aligned}\widetilde{\Sigma}_{k,j} &= \mathbb{E}\left[\mathbb{E}[X|Z_{k-1}]\,\mathbb{E}[X|Z_{j-1}]\right] \\ &\stackrel{(a)}{=} \mathbb{E}\left[X\,\mathbb{E}[X|Z_{j-1}]\right] \\ &= \mathbb{E}\left[X\tanh(\widetilde{\gamma}_{j-1}X + \widetilde{\sigma}^2_{j-1}W)\right] \\ &= \mathbb{E}\left[\tanh(\widetilde{\gamma}_{j-1} + \widetilde{\sigma}^2_{j-1}W)\right] = \gamma_j\,,\end{aligned}$$

where in step (a) we used the sufficient statistic property. This completes the inductive step and proof. □

Define the function $\mathsf{mmse} : \mathbb{R} \to \mathbb{R}$ given by

$$\begin{aligned}\mathsf{mmse}(\gamma) &\equiv 1 - \mathbb{E}\left[\tanh(\gamma + \sqrt{\gamma}W)^2\right] \\ &= 1 - \mathbb{E}\left[\mathbb{E}[X|\gamma X + \sqrt{\gamma}W]^2\right].\end{aligned}$$

It follows from Proposition V.2 that (V.2) and (V.3) can be expressed just in terms of the sequence $\gamma_k(\beta, t)$ defined by $\gamma_0(t) = 0$ and the recursion

$$\gamma_{k+1}(\beta, t) = \beta^2\big(1 - \mathsf{mmse}(\gamma_k(\beta, t) + t)\big). \tag{V.4}$$

Note that $\gamma_k(\beta, t)$ depends also on β, which is usually treated as constant. The following result (proved in the full version) details some useful properties of mmse.

Lemma V.3 ([DAM17, Lemma 6.1])**.** *The following properties hold, where $\{\gamma_k(\beta, t)\}_{k\geq 1}$ is as in* (V.4)*.*

(a) mmse *is differentiable, strictly decreasing, and convex in $\gamma \in \mathbb{R}_{\geq 0}$.*

(b) $\mathsf{mmse}(0) = 1$, $\mathsf{mmse}'(0) = -1$ *and* $\lim_{\gamma\to\infty} \mathsf{mmse}(\gamma) = 0$*.*

(c) For $t \geq 0$ there exists a non-negative solution $\gamma_* = \gamma_*(\beta, t)$ to the fixed point equation

$$\gamma_* = \beta^2(1 - \mathsf{mmse}(\gamma_* + t))\,. \tag{V.5}$$

The solution to this equation is unique for all $t > 0$.

(d) The function $(\beta, t) \mapsto \gamma_*(\beta, t)$ is differentiable for $t > 0$.

(e) For all $\beta < 1$ and $t > 0$,

$$1 - \beta^{2k} \leq \frac{\gamma_k(\beta, t)}{\gamma_*(\beta, t)} \leq 1\,. \tag{V.6}$$

(f) For $\beta < 1$ and $T > 0$, there exist constants $c(\beta, T), C(\beta, T) \in (0, \infty)$ such that, for all $t \in (0, T]$,

$$c(\beta, T) \leq \frac{\gamma_*(\beta, t)}{t} \leq C(\beta, T)\,. \tag{V.7}$$

(g) For $\beta < 1$ and any $t_1, t_2 \in (0, \infty)$,

$$\gamma_*(\beta, t_1) - \gamma_*(\beta, t_2) \leq \frac{\beta^2}{1 - \beta^2}|t_1 - t_2|. \tag{V.8}$$

Next for $(\boldsymbol{A}, \boldsymbol{y}) \sim \mathbb{P}$ and $\boldsymbol{x} \sim \mu_{\boldsymbol{A},\boldsymbol{y}(t)}$, define

$$\mathsf{MSE}_{\mathsf{AMP}}(k; \beta, t) = \operatorname*{p-lim}_{N\to\infty} \frac{1}{N}\,\mathbb{E}\left\|\boldsymbol{x} - \widehat{\boldsymbol{m}}^k(\boldsymbol{A}, \boldsymbol{y}(t))\right\|_2^2, \tag{V.9}$$

$$\widehat{\boldsymbol{m}}^k(\boldsymbol{A}, \boldsymbol{y}(t)) := \mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}(t); k)\,,$$

where the limit is guaranteed to exist by Proposition V.1.

Lemma V.4. *We have*

$$\mathsf{MSE}_{\mathsf{AMP}}(k; \beta, t) = 1 - \frac{\gamma_{k+1}(\beta, t)}{\beta^2}\,.$$

In particular,

$$\lim_{k\to\infty} \mathsf{MSE}_{\mathsf{AMP}}(k; \beta, t) = 1 - \frac{\gamma_*(\beta, t)}{\beta^2}\,.$$

Proof. By state evolution

$$\begin{aligned}\mathsf{MSE}_{\mathsf{AMP}}(k; \beta, t) &= \operatorname*{p-lim}_{N\to\infty} \frac{1}{N}\,\mathbb{E}\left\|\widehat{\boldsymbol{m}}^k(\boldsymbol{A}, \boldsymbol{y}(t)) - \boldsymbol{x}\right\|_2^2 \\ &= \mathbb{E}\left[\big(\tanh(\gamma_k X + \sigma_k W + Y) - X\big)^2\right] \\ &= \mathbb{E}\left[\big(\tanh(\widetilde{\gamma}_k X + \widetilde{\sigma}_k W) - X\big)^2\right] \\ &= 1 - 2\,\mathbb{E}[\tanh(\widetilde{\gamma}_k X + \widetilde{\sigma}_k W)X] \\ &\quad + \mathbb{E}[\tanh(\widetilde{\gamma}_k X + \widetilde{\sigma}_k W)^2] \\ &= 1 - 2\gamma_{k+1}/\beta^2 + \sigma^2_{k+1}/\beta^2 \\ &= 1 - \gamma_{k+1}/\beta^2,\end{aligned}$$

where the last line follows from Proposition V.2. □

The next Proposition shows that for any $t > 0$, the mean square error achieved by AMP is the same as the Bayes optimal error, i.e., the mean squared error achieved by the posterior expectation $\boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(t))$. The proof is based on an area law argument similar to [DAM17] and is omitted.

Proposition V.5. *Fix $\beta < 1$ and $t \geq 0$. We have*

$$\lim_{N\to\infty} \frac{1}{N}\,\mathbb{E}\left[\left\|\boldsymbol{x} - \boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(t))\right\|_2^2\right] = \frac{\gamma_*(\beta, t)}{\beta^2}\,. \tag{V.10}$$

It follows that AMP approximately computes the posterior mean $\boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(t))$ in the following sense.

Proposition V.6. *Fix $\beta < 1$, $T > 0$ and let $t \in (0, T]$. Recalling that $\widehat{\boldsymbol{m}}^k(\boldsymbol{A}, \boldsymbol{y}(t)) := \mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}(t); k)$ denotes the AMP estimate after k iterations, and that $\boldsymbol{z}^k$ is defined by Eq.* (V.1)*, we have*

$$\lim_{k\to\infty} \sup_{t\in(0,T)} \operatorname*{p-lim}_{N\to\infty} \frac{\|\boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(t)) - \widehat{\boldsymbol{m}}^k(\boldsymbol{A}, \boldsymbol{y}(t))\|_2}{\|\boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(t))\|_2} = 0\,. \tag{V.11}$$

Moreover

$$\lim_{k\to\infty} \sup_{t\in(0,T)} \operatorname*{p-lim}_{N\to\infty} \frac{\|\boldsymbol{z}^{k+1} - \boldsymbol{z}^k\|}{\|\boldsymbol{z}^k\|} = 0\,. \tag{V.12}$$

Remark V.1. [CT21] shows a related result (under different conditions) where the external field vector $\boldsymbol{y}(t)$ is replaced by a multiple of the all-ones vector $h\mathbf{1}$.

We conclude this subsection with a lemma controlling the regularity of the posterior path $t \mapsto \boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(t))$, which will be useful later. We omit the proof, which is based on Doob's maximal inequality.

Lemma V.7. *Fix $\beta < 1$ and $0 \le t_1 < t_2 \le T$. Then*

$$\begin{aligned}&\operatorname*{p-lim}_{N\to\infty} \sup_{t\in[t_1,t_2]} \frac{1}{N}\big\|\boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(t)) - \boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(t_1))\big\|_2^2 \\ &= \operatorname*{p-lim}_{N\to\infty} \frac{1}{N}\big\|\boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(t_2)) - \boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(t_1))\big\|_2^2 \\ &= \big(\gamma_*(\beta, t_2) - \gamma_*(\beta, t_1)\big)/\beta^2.\end{aligned} \tag{V.13}$$

VI. Natural Gradient Descent

Algorithm 3: Natural Gradient Descent on $\mathscr{F}_{\text{TAP}}(\,\cdot\,; \boldsymbol{y}, q)$

Input: Initialization $\boldsymbol{u}^0 \in \mathbb{R}^N$, data $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, $\widehat{\boldsymbol{y}} \in \mathbb{R}^N$, step size $\eta > 0$, $q \in (0, 1)$, integer $K > 0$.

$\widehat{\boldsymbol{m}}^{+,0} = \tanh(\boldsymbol{u}^0)$.

for $k = 0, \cdots, K-1$ **do**

  $\boldsymbol{u}^{k+1} \leftarrow \boldsymbol{u}^k - \eta \cdot \nabla \mathscr{F}_{\text{TAP}}(\widehat{\boldsymbol{m}}^{+,k}; \boldsymbol{y}, q)$,

  $\widehat{\boldsymbol{m}}^{+,k+1} = \tanh(\boldsymbol{u}^{+,k+1})$,

return $\widehat{\boldsymbol{m}}^{+,K}$

Here we state Lemma VI.1, which shows that $\mathscr{F}_{\text{TAP}}(\boldsymbol{m}; \boldsymbol{y}, q)$ behaves well for $q = q_*(\beta, t)$ and for $\boldsymbol{m}$ in a neighborhood of $\widehat{\boldsymbol{m}}^{K_{\mathsf{AMP}}}$. Namely it has a unique local minimum $\boldsymbol{m}_* = \boldsymbol{m}_*(\boldsymbol{A}, \boldsymbol{y})$ in such a neighborhood, and NGD approximates $\boldsymbol{m}_*$ well for large number of iterations K. Crucially, the map $\boldsymbol{y} \mapsto \boldsymbol{m}_*$ will be Lipschitz. For reference, we reproduce Algorithm 3 (NGD), corresponding to lines 5-9 of Algorithm 1.

Lemma VI.1. *Let $\beta < \frac{1}{2}$, $c \in (0, 1-2\beta)$, and $T > 0$ be fixed. Then there exists $\varepsilon_0 = \varepsilon_0(\beta, T)$ such that, for all $\varepsilon \in (0, \varepsilon_0)$ there exists $K_{\mathsf{AMP}} = K_{\mathsf{AMP}}(\beta, T, \varepsilon)$ and $\rho_0 = \rho_0(\beta, T, \varepsilon)$ such that for all $\rho \in (0, \rho_0)$ there exists $K_{\mathsf{NGD}} = K_{\mathsf{NGD}}(\beta, T, \varepsilon, \rho)$, such that the following holds.*

Let $\widehat{\boldsymbol{m}}^{\mathsf{AMP}} = \mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}(t); K_{\mathsf{AMP}})$ be the output of the AMP after K_{AMP} iterations, when applied to $\boldsymbol{y}(t)$. Fix $K \ge K_{\mathsf{AMP}}$. With probability $1 - o_N(1)$ over $(\boldsymbol{A}, \boldsymbol{y}) \sim \mathbb{P}$, for all $t \in (0, T]$ and all $\widehat{\boldsymbol{y}} \in B\Big(\boldsymbol{y}(t), c\sqrt{\varepsilon t N}/4\Big)$, setting $q_* := q_*(\beta, t)$:

1) *The function*

$$\boldsymbol{m} \mapsto \mathscr{F}_{\text{TAP}}(\boldsymbol{m}; \widehat{\boldsymbol{y}}, q_*)$$

restricted to $B\Big(\widehat{\boldsymbol{m}}^{\mathsf{AMP}}, \sqrt{\varepsilon t N}\Big) \cap (-1, 1)^N$ has a unique stationary point

$$\boldsymbol{m}_*(\boldsymbol{A}, \widehat{\boldsymbol{y}}) \in B\Big(\widehat{\boldsymbol{m}}^{\mathsf{AMP}}, \sqrt{\varepsilon t N}/2\Big) \cap (-1, 1)^N$$

which is also a local minimum. In the case $\widehat{\boldsymbol{y}} = \boldsymbol{y}(t)$, $\boldsymbol{m}_*(\boldsymbol{A}, \boldsymbol{y}(t))$ also satisfies

$$\boldsymbol{m}_*(\boldsymbol{A}, \boldsymbol{y}) \in B\Big(\widehat{\boldsymbol{m}}^{k'}, \sqrt{\varepsilon t N}/2\Big) \cap (-1, 1)^N$$

for all $k' \in [K_{\mathsf{AMP}}, K]$, where $\widehat{\boldsymbol{m}}^{k'} = \mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}(t); k')$.

2) *The stationary point $\boldsymbol{m}_*(\boldsymbol{A}, \widehat{\boldsymbol{y}})$ satisfies (recall that $\boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y})$ denotes the mean of the Gibbs measure)*

$$\big\|\boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}) - \boldsymbol{m}_*(\boldsymbol{A}, \boldsymbol{y})\big\|_2 \le \rho\sqrt{tN}\,.$$

3) *The stationary point $\boldsymbol{m}_*$ obeys the following Lipschitz property for all $\widehat{\boldsymbol{y}}, \widehat{\boldsymbol{y}}' \in B\Big(\boldsymbol{y}(t), c\sqrt{\varepsilon t N}/4\Big)$:*

$$\big\|\boldsymbol{m}_*(\boldsymbol{A}, \widehat{\boldsymbol{y}}) - \boldsymbol{m}_*(\boldsymbol{A}, \widehat{\boldsymbol{y}}')\big\| \le c^{-1}\|\widehat{\boldsymbol{y}} - \widehat{\boldsymbol{y}}'\|\,. \tag{VI.1}$$

4) *There exists a learning rate $\eta = \eta(\beta, T, \varepsilon)$ such that the following holds. Let $\widehat{\boldsymbol{m}}^{\mathsf{NGD}}(\boldsymbol{A}, \widehat{\boldsymbol{y}})$ be the output of NGD (Algorithm 3), when run for K_{NGD} iterations with parameter q_*, $\widehat{\boldsymbol{y}}$, η. Assume that the initialization $\boldsymbol{u}^0$ satisfies*

$$\big\|\boldsymbol{u}^0 - \operatorname{arctanh}(\widehat{\boldsymbol{m}}^{\mathsf{AMP}})\big\| \le \frac{c\sqrt{\varepsilon t N}}{200}\,. \tag{VI.2}$$

Then the algorithm output satisfies

$$\big\|\widehat{\boldsymbol{m}}^{\mathsf{NGD}}(\boldsymbol{A}, \widehat{\boldsymbol{y}}) - \boldsymbol{m}_*(\boldsymbol{A}, \widehat{\boldsymbol{y}})\big\| \le \rho\sqrt{tN}\,. \tag{VI.3}$$

Lemma VI.1 is proved in the full version. We note that using a radius $\Theta(\sqrt{tN})$ neighborhood is crucial.

VII. Continuous limit and proof of Theorem II.1

We fix (β, T) and choose constants $K_{\mathsf{AMP}} = K_{\mathsf{AMP}}(\beta, T, \varepsilon)$, $\rho_0 = \rho_0(\beta, T, \varepsilon, K_{\mathsf{AMP}})$, $\rho \in (0, \rho_0)$ and $K_{\mathsf{NGD}} = K_{\mathsf{NGD}}(\beta, T, \varepsilon, \rho)$ so that Lemma VI.1 holds.

We couple the discretized process $(\widehat{\boldsymbol{y}}_\ell)_{\ell \geq 0}$ defined in Eq. (II.3) (line 6 of Algorithm 2) to the continuous time process $(\boldsymbol{y}(t))_{t \in \mathbb{R}_{\geq 0}}$ (cf. Eq. (IV.8)) via the driving noise, as follows:

$$\boldsymbol{w}_{\ell+1} = \frac{1}{\sqrt{\delta}} \int_{\ell\delta}^{(\ell+1)\delta} \mathrm{d}\boldsymbol{B}(t)\,. \tag{VII.1}$$

We denote by $\widehat{\boldsymbol{m}}(\boldsymbol{A}, \boldsymbol{y})$ the output of the mean estimation algorithm 1 on input $\boldsymbol{A}, \boldsymbol{y}$. Lemma VI.1 ensures that for any $t \in (0, T]$, with probability $1 - o_N(1)$,

$$\big\|\widehat{\boldsymbol{m}}(\boldsymbol{A}, \boldsymbol{y}(t)) - \boldsymbol{m}_*(\boldsymbol{A}, \boldsymbol{y}(t); q_*(\beta, t))\big\| \leq \rho\sqrt{tN}\,. \tag{VII.2}$$

Here and below we note explicitly the dependence of $\boldsymbol{m}_*$ on t via q_*. The next lemma provides a crude estimate on the Lipschitz continuity of AMP in its input.

Lemma VII.1. *Recall that* $\mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; k) \in \mathbb{R}^N$ *denotes the output of the AMP algorithm on input* $(\boldsymbol{A}, \boldsymbol{y})$*, after* k *iterations, cf. Eq.* (II.2)*. If* $\|\boldsymbol{A}\|_{\mathrm{op}} \leq 3$*, then, for any* $\boldsymbol{y}, \widehat{\boldsymbol{y}} \in \mathbb{R}^N$*,*

$$\begin{aligned}\big\| \operatorname{arctanh}\big(\mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; k)\big) - \operatorname{arctanh}\big(\mathsf{AMP}(\boldsymbol{A}, \widehat{\boldsymbol{y}}; k)\big)\big\|_2 \\ \leq k6^k \|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|_2\,.\end{aligned} \tag{VII.3}$$

Proof. For $0 \leq j \leq k$, set:

$$\begin{aligned}\boldsymbol{m}^j &= \mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; j), \quad \boldsymbol{z}^j = \operatorname{arctanh}(\boldsymbol{m}^j),\\ \mathsf{b}_j &= \frac{\beta^2}{N}\sum_{i=1}^N \big(1 - \tanh^2(z_i^j)\big),\\ \widehat{\boldsymbol{m}}^j &= \mathsf{AMP}(\boldsymbol{A}, \widehat{\boldsymbol{y}}; j), \quad \widehat{\boldsymbol{z}}^j = \operatorname{arctanh}(\widehat{\boldsymbol{m}}^j),\\ \widehat{\mathsf{b}}_j &= \frac{\beta^2}{N}\sum_{i=1}^N \big(1 - \tanh^2(\widehat{z}_i^j)\big).\end{aligned}$$

Using the AMP update equation (line 4 of Algorithm 1) and the fact that $\tanh(\,\cdot\,)$ is 1-Lipschitz, we obtain

$$\begin{aligned}\|\boldsymbol{z}^{j+1} - \widehat{\boldsymbol{z}}^{j+1}\| &\leq \|\beta\boldsymbol{A}(\boldsymbol{m}^j - \widehat{\boldsymbol{m}}^j)\| + \|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|\\ &\quad + \|\mathsf{b}_j\boldsymbol{m}^{j-1} - \mathsf{b}_j\widehat{\boldsymbol{m}}^{j-1}\|\\ &\quad + \|\mathsf{b}_j\widehat{\boldsymbol{m}}^{j-1} - \widehat{\mathsf{b}}_j\widehat{\boldsymbol{m}}^{j-1}\|\\ &\leq 3\beta\|\boldsymbol{z}^j - \widehat{\boldsymbol{z}}^j\| + \|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|\\ &\quad + \mathsf{b}_j\|\boldsymbol{z}^{j-1} - \widehat{\boldsymbol{z}}^{j-1}\| + |\mathsf{b}_j - \widehat{\mathsf{b}}_j|\sqrt{N}\,.\end{aligned}$$

Note that $|1 - \tanh^2(x)| \leq 1$ for all $x \in \mathbb{R}$ and $|b_j| \leq \beta^2$. Setting $E_j = \max_{i \leq j} \|\boldsymbol{z}^{i+1} - \widehat{\boldsymbol{z}}^{i+1}\|$, we find

$$\begin{aligned}E_{j+1} &\leq (3\beta^2 + 3\beta)E_j + \|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|\\ &\leq 6E_j + \|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|\,.\end{aligned}$$

It follows by induction that

$$E_j \leq j6^j\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|\,.$$

Setting $j = k$ concludes the proof. □

Define the random approximation errors

$$A_\ell := \frac{1}{\sqrt{N}}\big\|\widehat{\boldsymbol{y}}_\ell - \boldsymbol{y}(\ell\delta)\big\|\,, \tag{VII.4}$$

$$B_\ell := \frac{1}{\sqrt{N}}\big\|\widehat{\boldsymbol{m}}(\boldsymbol{A}, \widehat{\boldsymbol{y}}_\ell) - \boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(\ell\delta))\big\|\,. \tag{VII.5}$$

Note that $A_0 = B_0 = 0$. In the next lemma we bound the above quantities:

Lemma VII.2. *For* $\beta < 1/2$ *and* $T > 0$*, there exists a constant* $C = C(\beta) < \infty$*, and a deterministic non-negative sequence* $\alpha(N)$ *with* $\lim_{N\to\infty}\alpha(N) = 0$ *such that the following holds with probability* $1 - o_N(1)$*. For every* $\ell \geq 0$*,* $\delta \in (0, 1)$ *such that* $\ell\delta \leq T$*,*

$$A_\ell \leq Ce^{C\ell\delta}\ell\delta\big(\rho\sqrt{\ell\delta} + \sqrt{\delta}\big) + \alpha(N)\,, \tag{VII.6}$$

$$B_\ell \leq Ce^{C\ell\delta}\ell\delta\big(\rho\sqrt{\ell\delta} + \sqrt{\delta}\big) + C\rho\sqrt{\ell\delta} + \alpha(N)\,. \tag{VII.7}$$

Proof. Throughout the proof, we denote by $\alpha(N)$ a deterministic non-negative sequence $\alpha(N)$ with $\lim_{N\to\infty}\alpha(N) = 0$, which can change from line to line. Also, C will denote a generic constant that may depend on $\beta, T, K_{\mathsf{AMP}}$.

We induct on ℓ. As the base case is trivial, we assume the result for all $j \leq \ell$ and prove it for $\ell + 1$. We first claim that with probability $1 - o_N(1)$,

$$A_{\ell+1} \leq A_\ell + \delta B_\ell + C\delta^{3/2}. \tag{VII.8}$$

Indeed, using (VII.1) we find for $I_\ell = [\ell\delta, (\ell+1)\delta]$:

$$\begin{aligned}A_{\ell+1} - A_\ell &\leq n^{-1/2}\int_{I_\ell}\big\|\widehat{\boldsymbol{m}}(\boldsymbol{A}, \widehat{\boldsymbol{y}}_\ell) - \boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(t))\big\|\,\mathrm{d}t\\ &\leq \delta n^{-1/2}\Big(\big\|\widehat{\boldsymbol{m}}(\boldsymbol{A}, \widehat{\boldsymbol{y}}_\ell) - \boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(\ell\delta))\big\|\\ &\quad + \sup_{t\in I_\ell}\big\|\boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(t)) - \boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(\ell\delta))\big\|\Big)\\ &\leq \delta B_\ell + \delta\cdot\frac{\sup_{t\in I_\ell}\big\|\boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(t)) - \boldsymbol{m}(\boldsymbol{A}, \boldsymbol{y}(\ell\delta))\big\|}{\sqrt{n}}\\ &\leq \delta B_\ell + C(\beta)\delta^{3/2} + \alpha(N)\,,\end{aligned}$$

where the last line holds with high probability by Lemma V.7 and Eq. (V.8) of Lemma V.3. Using this bound and the inductive hypothesis on A_ℓ, B_ℓ we obtain

$$\begin{aligned} A_{\ell+1} &\le Ce^{C(\ell+1)\delta}\ell\delta(\rho\sqrt{\ell\delta}+\sqrt{\delta}) + C\rho\delta\sqrt{\ell\delta} \\ &\quad + C\delta^{3/2} + \alpha(N) \\ &\le Ce^{C(\ell+1)\delta}(\ell+1)\delta(\rho+\sqrt{\delta}) + \alpha(N)\,. \end{aligned}$$

This implies Eq. (VII.6) for $\ell+1$.

We next show that Eq. (VII.7) holds with ℓ replaced by $\ell+1$. By the bound (VII.6) for $\ell+1$, taking $\delta \le \delta(\beta,\varepsilon,K_{\mathsf{AMP}},T)$ and $\rho \in (0,\rho_0)$ $\rho = \rho(\beta,\varepsilon,K_{\mathsf{AMP}},T)$ ensures that

$$A_{\ell+1} \le \frac{c\sqrt{\varepsilon\ell\delta}}{200K_{\mathsf{AMP}}6^{K_{\mathsf{AMP}}}}\,,$$

where ε can be chosen an arbitrarily small constant. So by Lemma VII.1, we have with probability $1-o_N(1)$,

$$\begin{aligned} &\big|\big|\operatorname{arctanh}(\mathsf{AMP}(\boldsymbol{A},\boldsymbol{y}((\ell+1)\delta);K_{\mathsf{AMP}})) \\ &-\operatorname{arctanh}(\mathsf{AMP}(\boldsymbol{A},\widehat{\boldsymbol{y}}_{\ell+1};K_{\mathsf{AMP}}))\big|\big|_2 \\ &\le K_{\mathsf{AMP}}6^{K_{\mathsf{AMP}}}A_{\ell+1}\sqrt{N} \\ &\le \frac{c\sqrt{\varepsilon\ell\delta N}}{200}\,. \end{aligned}$$

By choosing $\varepsilon \le \varepsilon_0(\beta,T)$, we obtain that Lemma VI.1, part 4 applies. We thus find

$$\|\widehat{\boldsymbol{m}}(\boldsymbol{A},\widehat{\boldsymbol{y}}_{\ell+1}) - \boldsymbol{m}_*(\boldsymbol{A},\widehat{\boldsymbol{y}}_{\ell+1})\| \le \rho\sqrt{\ell\delta N}\,.$$

Using parts 3 and 2 respectively of Lemma VI.1 on the other terms below, by triangle inequality we obtain (writing for simplicity $q_\ell := q_*(\beta,\ell\delta)$)

$$\begin{aligned} &\|\widehat{\boldsymbol{m}}(\boldsymbol{A},\widehat{\boldsymbol{y}}_{\ell+1}) - \boldsymbol{m}(\boldsymbol{A},\boldsymbol{y}((\ell+1)\delta))\| \\ &\le \|\widehat{\boldsymbol{m}}(\boldsymbol{A},\widehat{\boldsymbol{y}}_{\ell+1}) - \boldsymbol{m}_*(\boldsymbol{A},\widehat{\boldsymbol{y}}_{\ell+1};q_{\ell+1})\| \\ &\quad + \|\boldsymbol{m}_*(\boldsymbol{A},\widehat{\boldsymbol{y}}_{\ell+1};q_{\ell+1}) - \boldsymbol{m}_*(\boldsymbol{A},\boldsymbol{y}((\ell+1)\delta);q_{\ell+1})\| \\ &\quad + \|\boldsymbol{m}_*(\boldsymbol{A},\boldsymbol{y}((\ell+1)\delta);q_{\ell+1}) - \boldsymbol{m}(\boldsymbol{A},\boldsymbol{y}((\ell+1)\delta))\| \\ &\le \big(\rho\sqrt{\ell\delta} + c^{-1}A_{\ell+1} + \rho\sqrt{\ell\delta} + \alpha(N)\big)\sqrt{N}\,. \end{aligned} \tag{VII.9}$$

In other words with probability $1-o_N(1)$,

$$B_{\ell+1} \le c^{-1}A_{\ell+1} + 2\rho\sqrt{\ell\delta} + \alpha(N)\,.$$

Together with the bound (VII.6) for $\ell+1$, this verifies the inductive step for (VII.7), concluding the proof. □

The following lemma ensures the final randomized rounding step in our sampling algorithm is benign.

Lemma VII.3. *Suppose probability distributions μ_1,μ_2 on $[-1,1]^N$ are given. Sample $\boldsymbol{m}_1 \sim \mu_1$ and $\boldsymbol{m}_2 \sim \mu_2$ and let $\boldsymbol{x}_1,\boldsymbol{x}_2 \in \{-1,+1\}^N$ be standard randomized roundings, respectively of $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$. (Namely, the coordinates of $\boldsymbol{x}_i$ are conditionally independent given $\boldsymbol{m}_i$, with $\mathbb{E}[\boldsymbol{x}_i|\boldsymbol{m}_i] = \boldsymbol{m}_i$.) Then*

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

Proof. Let $(\boldsymbol{m}_1,\boldsymbol{m}_2)$ be a $W_{2,N}$-optimal coupling between μ_1,μ_2. Couple $\boldsymbol{x}_1,\boldsymbol{x}_2$ by choosing i.i.d. uniform random variables $u_i \sim \mathsf{Unif}([0,1])$ for $i \in [n]$, and for $(i,j) \in [n]\times\{1,2\}$ setting

$$(\boldsymbol{x}_j)_i = \begin{cases} +1, & \text{if } u \le \frac{1+(\boldsymbol{m}_j)_i}{2}\,, \\ -1, & \text{else.} \end{cases}$$

Then it is not difficult to see that

$$\begin{aligned} \frac{1}{N}\,\mathbb{E}\left[\|\boldsymbol{x}_1-\boldsymbol{x}_2\|^2 \;|(\boldsymbol{m}_1,\boldsymbol{m}_2)\right] &= \frac{2}{N}\sum_{i=1}^N |(\boldsymbol{m}_1)_i - (\boldsymbol{m}_2)_i| \\ &\le 2\sqrt{\frac{1}{N}\|\boldsymbol{m}_1-\boldsymbol{m}_2\|^2}. \end{aligned}$$

Averaging over $(\boldsymbol{m}_1,\boldsymbol{m}_2)$ implies the result. □

Proof of Theorem II.1. Set $\ell = L = T/\delta$ and $\rho = \sqrt{\delta}$ in Eq. (VII.7). With all laws $\mathcal{L}(\,\cdot\,)$ conditional on $\boldsymbol{A}$ below, we find

$$\begin{aligned} &\mathbb{E}W_{2,N}(\mu_{\boldsymbol{A}},\mathcal{L}(\widehat{\boldsymbol{m}}(\boldsymbol{A},\widehat{\boldsymbol{y}}_L))) \\ &\quad\le \mathbb{E}\,W_{2,N}(\mu_{\boldsymbol{A}},\mathcal{L}(\boldsymbol{m}(\boldsymbol{A},\boldsymbol{y}(T)))) \\ &\qquad + \mathbb{E}\,W_{2,N}(\mathcal{L}(\boldsymbol{m}(\boldsymbol{A},\boldsymbol{y}(T))),\mathcal{L}(\widehat{\boldsymbol{m}}(\boldsymbol{A},\widehat{\boldsymbol{y}}_L))) \\ &\quad\le T^{-1/2} + C(\beta,T)\sqrt{\delta} + o_N(1). \end{aligned}$$

Here the first term was bounded by Eq. (III.3) in Section III and the second by Eq. (VII.7). Taking T sufficiently large, δ sufficiently small, and N sufficiently large, we obtain

$$\mathbb{E}\,W_{2,N}\big(\mu_{\boldsymbol{A}},\mathcal{L}(\widehat{\boldsymbol{m}}_{\mathsf{NGD}}(\boldsymbol{A},\widehat{\boldsymbol{y}}_L))\big) \le \frac{\varepsilon^2}{4}$$

for any desired $\varepsilon > 0$. Applying Lemma VII.3 shows

$$\mathbb{E}\,W_{2,N}(\mu_{\boldsymbol{A}},\boldsymbol{x}^{\mathrm{alg}}) \le \varepsilon\,.$$

The Markov inequality now implies that (II.4) holds with probability $1-o_N(1)$ as desired. □

The full version of this paper is available as an online preprint at https://arxiv.org/abs/2203.05093 .

REFERENCES

[AH87] Michael Aizenman and Richard Holley, *Rapid convergence to equilibrium of stochastic Ising models in the Dobrushin Shlosman regime*, Percolation theory and ergodic theory of infinite particle systems, Springer, 1987, pp. 1–11.

[AJK+21] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong, *Entropic Independence I: Modified Log-Sobolev Inequalities for Fractionally Log-Concave Distributions and High-Temperature Ising Models*, arXiv preprint arXiv:2106.04105 (2021).

[ALR87] Michael Aizenman, Joel L Lebowitz, and David Ruelle, *Some rigorous results on the Sherrington–Kirkpatrick spin glass model*, Communications in Mathematical Physics **112** (1987), no. 1, 3–20.

[BAJ18] Gérard Ben Arous and Aukosh Jagannath, *Spectral gap estimates in mean field spin glasses*, Communications in Mathematical Physics **361** (2018), no. 1, 1–52.

[BB19] Roland Bauerschmidt and Thierry Bodineau, *A very simple proof of the LSI for high temperature spin systems*, Journal of Functional Analysis **276** (2019), no. 8, 2582–2588.

[BD11] Joseph Blitzstein and Persi Diaconis, *A sequential importance sampling algorithm for generating random graphs with prescribed degrees*, Internet Mathematics **6** (2011), no. 4, 489–522.

[BH21] Guy Bresler and Brice Huang, *The Algorithmic Phase Transition of Random k-SAT for Low Degree Polynomials*, 2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS), 2021, pp. 298–309.

[BM11] Mohsen Bayati and Andrea Montanari, *The dynamics of message passing on dense graphs, with applications to compressed sensing*, IEEE Transactions on Information Theory **57** (2011), no. 2, 764–785.

[CDHL05] Yuguo Chen, Persi Diaconis, Susan P Holmes, and Jun S Liu, *Sequential Monte Carlo methods for statistical analysis of tables*, Journal of the American Statistical Association **100** (2005), no. 469, 109–120.

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

[Cel22] Michael Celentano, *Sudakov-fernique post-amp, and a new proof of the local convexity of the tap free energy*, arXiv preprint arXiv:2208.09550 (2022).

[CGPR19] Wei-Kuo Chen, David Gamarnik, Dmitry Panchenko, and Mustazee Rahman, *Suboptimality of local algorithms for a class of max-cut problems*, The Annals of Probability **47** (2019), no. 3, 1587–1618.

[Cha09] Sourav Chatterjee, *Disorder chaos and multiple valleys in spin glasses*, arXiv preprint arXiv:0907.3381 (2009).

[Cha14] ———, *Superconcentration and related topics*, vol. 15, Springer, 2014.

[Che13] Wei-Kuo Chen, *Disorder chaos in the Sherrington–Kirkpatrick model with external field*, The Annals of Probability **41** (2013), no. 5, 3345–3391.

[CHHS15] Wei-Kuo Chen, Hsi-Wei Hsieh, Chii-Ruey Hwang, and Yuan-Chung Sheu, *Disorder chaos in the spherical mean-field model*, Journal of Statistical Physics **160** (2015), no. 2, 417–429.

[CP18] Wei-Kuo Chen and Dmitry Panchenko, *Disorder chaos in some diluted spin glass models*, The Annals of Applied Probability **28** (2018), no. 3, 1356–1378.

[CT21] Wei-Kuo Chen and Si Tang, *On Convergence of the Cavity and Bolthausen's TAP Iterations to the Local Magnetization*, Communications in Mathematical Physics **386** (2021), no. 2, 1209–1242.

[DAM17] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari, *Asymptotic mutual information for the balanced binary stochastic block model*, Information and Inference: A Journal of the IMA **6** (2017), no. 2, 125–170.

[EKZ21] 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 (2021), 1–17.

[Eld20] Ronen Eldan, *Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation*, Probability Theory and Related Fields **176** (2020), no. 3, 737–755.

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

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

[GJW20] David Gamarnik, Aukosh Jagannath, and Alexander S. Wein, *Low-degree hardness of random optimization problems*, Proceedings of 61st FOCS, IEEE, 2020, pp. 131–140.

[GK21] David Gamarnik and Eren C. Kızıldağ, *Algorithmic obstructions in the random number partitioning problem*, arXiv preprint arXiv:2103.01369 (2021).

[GS14] David Gamarnik and Madhu Sudan, *Limits of local algorithms over sparse random graphs*, Proceedings of the 5th conference on Innovations in theoretical computer science, ACM, 2014, pp. 369–376.

[GS17] ———, *Performance of sequential local algorithms for the random NAE-K-sat problem*, SIAM Journal on Computing **46** (2017), no. 2, 590–619.

[HS21] Brice Huang and Mark Sellke, *Tight Lipschitz Hardness for Optimizing Mean Field Spin Glasses*, arXiv preprint arXiv:2110.07847 (2021).

[JM13] Adel Javanmard and Andrea Montanari, *State evolution for general approximate message passing algorithms, with applications to spatial coupling*, Information and Inference: A Journal of the IMA **2** (2013), no. 2, 115–144.

[JVV86] Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani, *Random generation of combinatorial structures from a uniform distribution*, Theoretical computer science **43** (1986), 169–188.

[MPV87] Marc Mézard, Giorgio Parisi, and Miguel A. Virasoro, *Spin glass theory and beyond*, World Scientific, 1987.

[NSZ22] Danny Nam, Allan Sly, and Lingfu Zhang, *Ising model on trees and factors of IID*, Communications in Mathematical Physics (2022), 1–38.

[RV17] Mustazee Rahman and Bálint Virág, *Local algorithms for independent sets are half-optimal*, The Annals of Probability **45** (2017), no. 3, 1543–1577.

[Sub17] Eliran Subag, *The geometry of the gibbs measure of pure spherical spin glasses*, Inventiones mathematicae **210** (2017), no. 1, 135–209.

[SZ81] Haim Sompolinsky and Annette Zippelius, *Dynamic theory of the spin-glass phase*, Physical Review Letters **47** (1981), no. 5, 359.

[Tal10] Michel Talagrand, *Mean field models for spin glasses: Volume i*, Springer-Verlag, Berlin, 2010.

[VdV98] Aad W Van der Vaart, *Asymptotic statistics*, vol. 3, Cambridge university press, 1998.

[Wei22] Alexander S Wein, *Optimal low-degree hardness of maximum independent set*, Mathematical Statistics and Learning (2022).