Barriers and Dynamical Paths in Alternating Gibbs Sampling of Restricted Boltzmann Machines

Clément Roussel, Simona Cocco, Rémi Monasson

.

Barriers and Dynamical Paths in Alternating Gibbs Sampling of Restricted Boltzmann Machines

Clément Roussel, Simona Cocco, Rémi Monasson
Laboratory of Physics of the Ecole Normale Supérieure, CNRS UMR 8023 & PSL Research, Sorbonne Université, 24 rue Lhomond, 75005 Paris, France

Restricted Boltzmann Machines (RBM) are bi-layer neural networks used for the unsupervised learning of model distributions from data. The bipartite architecture of RBM naturally defines an elegant sampling procedure, called Alternating Gibbs Sampling (AGS), where the configurations of the latent-variable layer are sampled conditional to the data-variable layer, and vice versa. We study here the performance of AGS on several analytically tractable models borrowed from statistical mechanics. We show that standard AGS is not more efficient than classical Metropolis-Hastings (MH) sampling of the effective energy landscape defined on the data layer. However, RBM can identify meaningful representations of training data in their latent space. Furthermore, using these representations and combining Gibbs sampling with the MH algorithm in the latent space can enhance the sampling performance of the RBM when the hidden units encode weakly dependent features of the data. We illustrate our findings on three datasets: Bars and Stripes and MNIST, well known in machine learning, and the so-called Lattice Proteins, introduced in theoretical biology to study the sequence-to-structure mapping in proteins.

I. INTRODUCTION

Studying large heterogeneous and strongly interacting systems is a challenge common to various scientific fields. For decades, various numerical methods have been developed to sample high-dimensional configurations of such systems. Among these Monte-Carlo (MC) methods are one of the most powerful and versatile procedures [1, 2]. Statistical averages over a target distribution are evaluated through an average over a set of stochastic configurations, generated according to a dynamical sampling process. Nevertheless, it is a well-known issue that these methods can suffer from poor mixing: sampled configurations can be trapped in one of the regions of high probability, i.e., of low free energy, while other favorable regions are not dynamically explored. Therefore, it is of most importance to design sampling procedures capable of efficient exploration, allowing for fast transitions from one minimum of the free energy to another. For ferromagnetic systems, cluster algorithms, which identify and flip large clusters of spins at once achieve this objective [3–6].

Recently, machine learning algorithms have been developed to detect relevant MC updates in condensed matter models [7–12]. Artificial neural networks are used to efficiently generate (with MC methods) low-energy configurations of approximate versions of target Hamiltonians. Hereafter we focus on one well-known machine learning architecture for unsupervised learning, called Restricted Boltzmann Machines (RBM) [13–15]. As illustrated in Fig. 1(a), RBM are undirected graphical models constituted by two sets of interconnected random variables: a visible layer $\mathbf{v}$ that represents the data and a hidden layer $\mathbf{h}$ able to extract and explain their statistical features. RBM learn a joint Boltzmann distribution $P(\mathbf{v}, \mathbf{h})$ by maximizing the log-likelihood of the data configurations:

$$P(\mathbf{v}) = \int d\mathbf{h}\, P(\mathbf{v}, \mathbf{h}) \ , \tag{1}$$

where the joint distribution of visible and hidden configurations reads

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} \exp\left(-E(\mathbf{v}, \mathbf{h})\right) \ , \tag{2}$$

and the energy $E(\mathbf{v}, \mathbf{h})$ includes couplings between, but not within the layers. RBM have been widely studied from a statistical mechanics point of view [15–20], see [21] for a recent review.

The bipartite architecture of RBM suggests a natural procedure for sampling the marginal distribution $P(\mathbf{v})$. The method, called Alternating Gibbs Sampling (AGS), uses the conditional distributions $P(\mathbf{h}|\mathbf{v})$ and $P(\mathbf{v}|\mathbf{h})$ to sequentially sample the hidden and the visible spaces (see Fig. 1(b)). As the interaction graph is bipartite, the two conditional distributions factorize over the units of the sampled layer, which allows for independent draws of unit values (within a layer).

Despite its elegance and the simplicity of implementation, it is unclear whether AGS thermalizes substantially better than standard MC procedures in the effective energy landscape over the visible configurations,

$$E^{\text{eff}}(\mathbf{v}) = -\log P(\mathbf{v}) \ , \tag{3}$$

see Fig. 1(c). On the one hand, the conditional sampling of visible configurations through $P(\mathbf{v}|\mathbf{h})$ seems to allow for global moves in the $\mathbf{v}$ space, as with cluster algorithms. On the other hand, the conditional sampling of latent variables through $P(\mathbf{h}|\mathbf{v})$ indicates that their values reflect global features of visible configurations and could remain frozen when the system is stuck in free

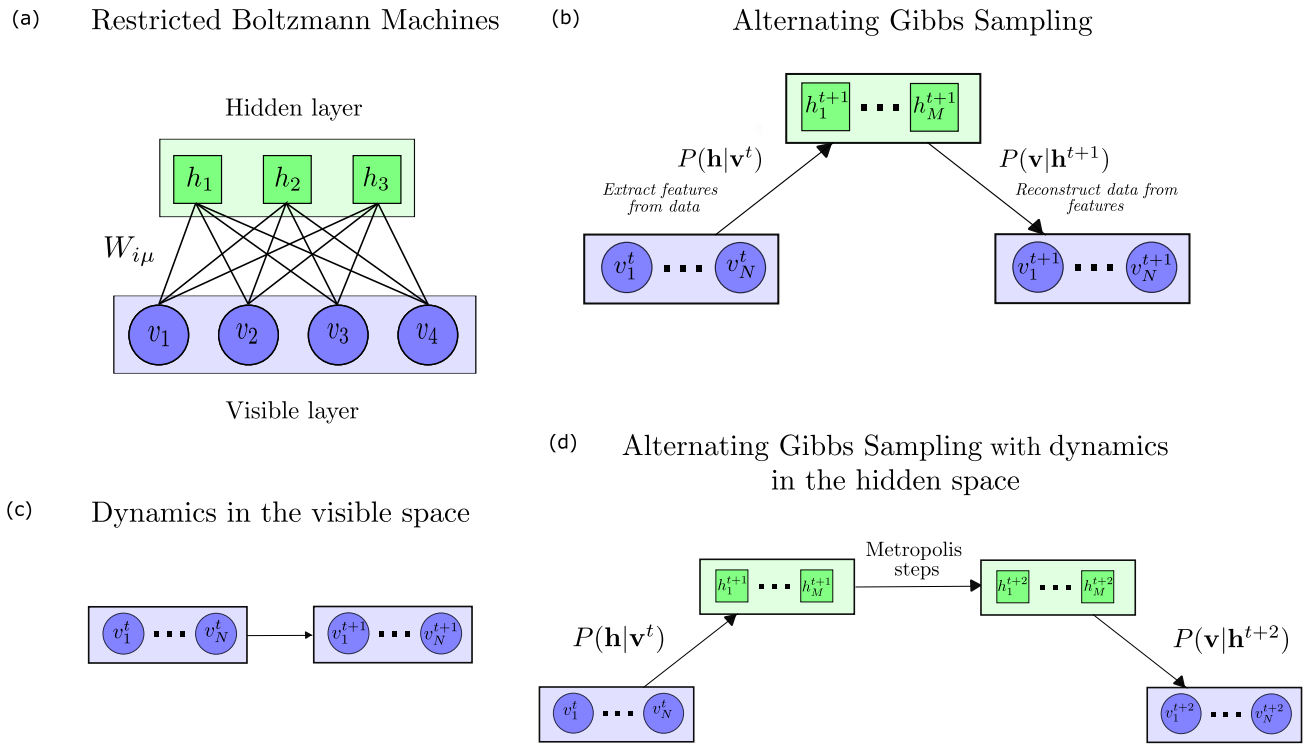

FIG. 1. Description of the Restricted Boltzmann Machines and of the different sampling algorithms. (a) Bipartite architecture of RBM, with the visible (blue) and hidden (green) layers. (b) Alternating Gibbs Sampling: hidden and visible configurations are conditionally sampled from one another. (c) Sampling dynamics in the landscape $E^{\text{eff}}(\mathbf{v})$. (d) Modified Alternating Gibbs Sampling with dynamics in the hidden configuration space.

energy minima. The purpose of the present work is to investigate this question on a few analytically tractable models. We show that canonical AGS is generally not more efficient than naive Metropolis-Hastings algorithm in the visible landscape $E^{\text{eff}}(\mathbf{v})$. However, the architecture of RBM offers two advantages with respect to the latter. First, the sampling paths joining one free energy minimum to another can be more easily interpreted in terms of trajectory in the hidden space than in the visible space. Secondly, we proposed an augmented version of AGS, in which intermediate moves in the hidden space are carried out (see Fig. 1(d)). We show that this new sampling procedure yields much reduced thermalization times if the statistical features attached to the hidden units are decorrelated enough.

Our paper is organized as follows. First, we define RBM, its sampling algorithm, the Alternating Gibbs Sampling between the visible and hidden layers, [22–24] and the different datasets we use for numerical experiments in Section II. Then, in Section III, we introduce the models under consideration and study how AGS samples them. In Section IV, we show how moving from one representation to another in the hidden space can help to sample. Finally, conclusions and perspectives are reported in Section V.

II. MODEL AND DATASETS

A. Restricted Boltzmann Machines

Restricted Boltzmann Machines are undirected probabilistic graphical models with two layers. A visible layer $\mathbf{v}$, which represents the data, is connected to a hidden layer $\mathbf{h}$ through a weight matrix W (see Fig. 1(a)). The visible layer includes N units v_i, and the hidden layer M units h_μ, which can take discrete or continuous values. The joint probability distribution of the visible configuration $\mathbf{v} = \{v_i\}_{i=1\dots N}$ and of the hidden configuration $\mathbf{h} = \{h_\mu\}_{\mu=1\dots M}$ is defined in Eq. 2. The energy $E(\mathbf{v}, \mathbf{h})$ is equal to

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{N}\sum_{\mu=1}^{M} W_{i\mu} v_i h_\mu + \sum_{\mu=1}^{M} \mathcal{U}_\mu(h_\mu) + \sum_{i=1}^{N} \mathcal{V}_i(v_i) \ . \tag{4}$$

In the formula above, $\mathcal{U}_\mu$ and $\mathcal{V}_i$ are potentials acting on, respectively, h_μ and v_i.

The effective energy over the visible configuration is obtained by marginalizing over the hidden units, see Eqs. (1) and (3), up to an additive constant, with the result

$$E^{\text{eff}}(\mathbf{v}) = \sum_{i=1}^{N} \mathcal{V}_i(v_i) - \sum_{\mu=1}^{M} \Gamma_\mu(I_\mu(\mathbf{v})) \ , \tag{5}$$

where $I_\mu(\mathbf{v}) = \sum_{i=1}^{N} W_{i\mu} v_i$ is the input received by hidden unit h_μ and $\Gamma_\mu(I) = \log\left(\int dh \exp\left(-\mathcal{U}_\mu(h) + hI\right)\right)$ is the cumulative generative function associated with the potential $\mathcal{U}_\mu$. Parameters $\Theta \equiv \{W_{i\mu}, \mathcal{U}_\mu, \mathcal{V}_i\}$ modulate the energy landscape $E^{\text{eff}}(\mathbf{v})$. RBM are known to be universal approximators (i.e., can approximate any distribution over the visible variables) when the number M of hidden units goes to infinity [25].

If the set of parameters Θ is known, the RBM model distribution is fully defined. However, expected values over the distribution are generally not tractable, and are estimated through MC methods. Different algorithms, based on Alternating Gibbs Sampling between the visible and hidden layers, are used to generate samples from $P(\mathbf{v})$, such as Contrastive Divergence [23], or Persistent Contrastive Divergence [24]. The pseudo-code of AGS is given in Algorithm 1 (see Fig. 1(c)). It is mainly composed of two steps:

- Starting from a visible configuration $\mathbf{v}^t$ at time t, a hidden configuration $\mathbf{h}^{t+1}$ is drawn from $P(\mathbf{h}|\mathbf{v}^t)$. This step can be seen as a stochastic feature extraction from the configuration $\mathbf{v}^t$.
- A new visible configuration $\mathbf{v}^{t+1}$ is drawn from $P\left(\mathbf{v}|\mathbf{h}^{t+1}\right)$. This step can be seen as a stochastic reconstruction of $\mathbf{v}$ from the latent configuration $\mathbf{h}^{t+1}$.

Note that AGS or its variations are also used during the learning phase. For a given training set of L samples, $\{\mathbf{v}^\ell\}_{\ell=1\dots L}$, the parameters Θ are found by maximizing the log-likelihood of the data, $\frac{1}{L}\sum_{\ell=1}^{L} \log P(\mathbf{v}^\ell) \equiv \langle \log P(\mathbf{v}) \rangle_{\text{data}}$. The maximization is done by gradient ascent. The general expression for the gradients is

$$\frac{\partial LL}{\partial \Theta} = -\left\langle \frac{\partial E^{\text{eff}}(\mathbf{v})}{\partial \Theta} \right\rangle_{\text{data}} + \left\langle \frac{\partial E^{\text{eff}}(\mathbf{v})}{\partial \Theta} \right\rangle_{\text{model}} , \quad (6)$$

where $\langle . \rangle_{\text{data}}$ denotes the expected value over the data and $\langle . \rangle_{\text{model}}$ over the model. We see that estimating the gradient requires computing averages over the RBM distribution at every step of the training process.

Algorithm 1: Alternating Gibbs Sampling

```
Pick v^0 in the training set;
for t ∈ [[0, T]] do
    h^{t+1} ~ P(h|v^t);
    v^{t+1} ~ P(v|h^{t+1});
end
```

B. Datasets

We use different datasets to illustrate our theoretical results. For all datasets, we train RBM using the learning algorithm of [15], available from https://github.com/jertubiana/PGM. We then study how AGS or other sampling algorithms sample the RBM distribution.

1. Bars and Stripes

Bars and Stripes (BAS) dataset [26] is made of $L \times L$ binary synthetic images which contain either exclusively bars or exclusively stripes. There are $2^{L+1} - 1$ possible configurations (see Fig. 2(a)).

2. MNIST

MNIST dataset [27] is a large dataset of 28×28 pixel images of handwritten digits. We limit ourselves to zeros and ones (Fig. 2(b)), two graphically far digits. We use the binarized version of MNIST: each pixel is white or black.

3. Lattice Protein

Lattice Proteins (LP) are artificial proteins used to investigate protein design [28, 29] and benchmarking inverse modeling procedures [30]. Proteins are sequences of amino acids, whose 3D structures encode their functionalities. In this model, a structure is defined as a self-avoiding path of 27 amino-acid-long chains ($\mathbf{v}$ represents a sequence) on the $3 \times 3 \times 3$ lattice cube. There are $\mathcal{N} = 103,406$ distinct structures (up to global symmetry). The probability that a protein sequence $\mathbf{v}$ folds in a given structure S is given by

$$P_{\text{nat}}(S|\mathbf{v}) = \frac{\exp\left(-E(\mathbf{v}, S)\right)}{\sum_{S'} \exp\left(-E(\mathbf{v}, S')\right)} , \quad (7)$$

where the energy of the sequence $\mathbf{v}$ in a structure S is defined through

$$E(\mathbf{v}, S) = \sum_{i<j} c^S_{i,j}\, E_{MJ}(v_i, v_j) . \quad (8)$$

In the previous formula, $c^S_{i,j} = 1$ if the sites i and j are in contact (neighbors on the cube) in structure S; there are 28 contacts between the amino acids for each structure[1]. Otherwise, $c^S_{i,j} = 0$. The pairwise energy $E_{MJ}(v_i, v_j)$ represents the physico-chemical interactions between the amino acids, given by the Miyazawa-Jernigan (MJ) potential [31]. Here, we focus on two structures, S_A and S_B, which define two protein families (Fig. 2(c)). For each

[1] Contacts along the chain are discarded, as their contribution to the energy is structure independent and, hence, does not affect the value of P_{nat}.

structure, we sample $\sim 10^4$ sequences that have a high probability to fold in this structure ($P_{\rm nat}(\mathbf{v}|S) > 0.99$) to build our datasets [30].

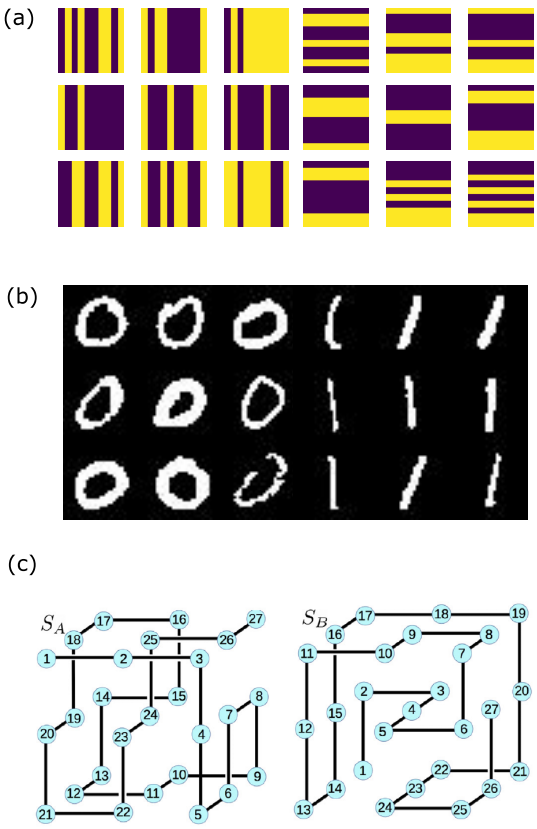

FIG. 2. (a) BAS: examples of bars (left) and stripes (right); here, $L = 10$. (b) MNIST: examples of hand-written 0 and 1 digits. (c) Lattice Proteins: two structures S_A and S_B defining two families of sequences having large P_{nat} with either fold, see Eq. (7). Structures from [30].

III. ALTERNATING GIBBS SAMPLING OF MULTI-MODAL DISTRIBUTIONS

This section examines how long it takes for AGS to sample complex energy landscapes with several states associated with multi-modal distributions. We consider first the Curie-Weiss model at low temperature, where two ferromagnetic states with opposite magnetizations coexist. We then turn to the case of the Hopfield model, in which different, uncorrelated states coexist. We finally study the general, more complex situation, in which multiple correlated states are present, and the optimal sampling paths follow a well-defined ordering of the states.

A. Case of bi-modal distribution

We consider the Curie-Weiss (CW) model over N spins, $v_i = \pm 1$. The energy function is defined as

$$E^{CW}(\mathbf{v}) = -\frac{w^2}{2N}\sum_{i,j=1}^{N} v_i v_j \ , \tag{9}$$

where w^2 plays the role of the inverse temperature. We start with the implementation of this mean-field model with RBM, before turning to a brief reminder of its properties and the study of the performance of AGS.

The CW model can be represented with a RBM with N visible units (with potentials $\mathcal{V}_i = 0$) and $M = 1$ hidden unit with a quadratic potential $\mathcal{U}(h) = \frac{h^2}{2}$. The weights $W_{i,\mu=1}$ are uniform and equal to $\frac{w}{\sqrt{N}}$ [2]. The energy of the RBM in Eq. (4) reads

$$E^{CW}(\mathbf{v},\mathbf{h}) = -\frac{w}{\sqrt{N}}\sum_{i=1}^{N} v_i\, h + \frac{h^2}{2} \ . \tag{10}$$

After integration over h, it is straightforward to check that the effective energy in Eq. (5) coincides with the CW energy in Eq. (9).

1. *Barriers and sampling time for MH procedures*

For $w^2 > 1$ and infinite-size limit $N \to \infty$, the average magnetization of the spins, $m = \frac{1}{N}\sum_{i=1}^{N} v_i$, spontaneously acquires a non zero value. The value of this order parameter is determined by minimizing the free energy (per spin), $f(m) = -\frac{w^2}{2}m^2 - \mathcal{S}(m)$, where

$$\mathcal{S}(m) = -\sum_{\sigma=\pm 1}\frac{1+\sigma m}{2}\log\left(\frac{1+\sigma m}{2}\right) \ , \tag{11}$$

is the entropy at fixed magnetization. The free energy $f(m)$ is an even function of m, with a double-well shape. The two opposite values of the spontaneous magnetization, roots of $f'(m^*) = 0$, define two collective states of the system. Notice that $m = 0$ is a local maximum of the free energy.

To go from one mode of the distribution to the other, a macroscopic number of spins has to be flipped. Local sampling processes, such as Metropolis-Hastings described in Algorithm 2[3] take exponential-in-N time to do so:

$$\tau \sim \exp\left(N\Delta f\right) \ , \quad \text{where} \quad \Delta f \equiv f(\pm m^*) - f(0) \ , \tag{12}$$

[2] We have checked that numerical experiments with RBM trained by gradient ascent on data sampled from the Curie-Weiss model converge to this solution.

[3] The specific choice of the Metropolis rule is irrelevant here; other choices, such as Glauber rule, [32], do not affect the leading behavior of τ.

is the free energy barrier between the minima $m = \pm m^*$ and the local maximum $m = 0$ of the free-energy landscape. Consequently, for large N, the system is stuck in one state/mode for long times, and thermalization is practically impossible.

Algorithm 2: Metropolis-Hastings algorithm

Pick $\mathbf{v}^0 \in \{-1,1\}^N$ at random ;
for $t \in [\![0, T]\!]$ **do**
  $\mathbf{v}^{'} = \mathbf{v}^t$;
  Choose $i \in [\![1, N]\!]$ uniformly at random;
  $v_i' = -v_i^t$;
  Generate a uniform random $u \in [0,1]$;
  if $u \leq \min\big(1, \exp\big[-(E^{\mathit{eff}}(\mathbf{v}') - E^{\mathit{eff}}(\mathbf{v}^t))\big]\big)$ **then**
    $\mathbf{v}^{t+1} = \mathbf{v}'$;
  else
    $\mathbf{v}^{t+1} = \mathbf{v}^t$;
  end
end

2. *Optimal sampling paths with AGS*

The AGS procedure can be entirely described in terms of the magnetizations m of the visible configurations and of the values h of the hidden unit. To get intensive quantities in the large N limit, we rescale $h \to h/\sqrt{N}$. The conditional configuration of the hidden unit h^{t+1} given a visible configuration with magnetization m^t then simply reads

$$P\big(h^{t+1}|m^t\big) = \frac{1}{\sqrt{2\pi/N}}\,\exp\left(-\frac{N}{2}\big(h^{t+1} - w\,m^t\big)^2\right)\,. \tag{13}$$

Some care must be taken to write the conditional distribution of the magnetization m^t given the hidden unit h^t. First, the conditional probability of $\mathbf{v}^t$ is

$$\begin{aligned} P\big(\mathbf{v}^t|h^t\big) &= \prod_{i=1}^{N} \frac{\exp\left(w\,h^t\,v_i^t\right)}{2\cosh(w\,h^t)} \\ &= \exp\big(N\big(w\,h^t\,m^t - \log 2\cosh(w\,h^t)\big)\big)\,, \end{aligned} \tag{14}$$

which depends on m^t as expected. Second, to turn the probability over visible configurations into a probability over magnetizations, we have to take into account the entropies of the latters. We end up with the normalized (to dominant order in N) conditional probability

$$\begin{aligned} P\big(m^t|h^t\big) &= \exp\big(N\big(w\,h^t\,m^t - \log 2\cosh(w\,h^t)\big)\big) \\ &\quad\times \exp\big(N\,\mathcal{S}(m^t)\big)\,. \end{aligned} \tag{15}$$

We may now express the probability to go from one minimum of the free energy landscape to the other in T steps of AGS. To do so, we compute the probability $P\left(m^T|m^0\right)$ that, given magnetization $m^0 = m^*$ at time $t = 0$, the dynamics associated with AGS reaches magnetization $m^T = -m^*$ at time $t = T$. This conditional probability may be computed by means of the saddle-point method in the thermodynamic limit $N \to \infty$ (for finite T):

$$P\left(m^T|m^0\right) = \int dh^1 \ldots dh^T \int dm^1 \ldots dm^{T-1} \prod_{t=0}^{T-1} P(m^{t+1}|h^{t+1})\,P(h^{t+1}|m^t) = \exp\left(-N \min_{\{m^t,h^t\}} \Phi\left(\{m^t,h^t\}\right)\right)\,,$$

where

$$\Phi\left(\{m^t,h^t\}\right) = \sum_{t=0}^{T-1} \delta\Phi\left(t \to t+1\right)\,, \tag{16}$$

and, according to Eqs. (13) and (15),

$$\begin{aligned} \delta\Phi\left(t \to t+1\right) &= \frac{1}{2}(h^{t+1} - w\,m^t)^2 \\ &\quad + \log\left(2\cosh\left(w\,h^{t+1}\right)\right) \\ &\quad - w\,m^{t+1}\,h^{t+1} - \mathcal{S}\big(m^{t+1}\big)\,. \end{aligned} \tag{17}$$

The set of magnetizations m^t and hidden-unit values h^t minimizing the action Φ in Eq. (16) define the most likely path, with AGS, capable of moving the system from one state to another in T alternating sampling steps. They are solutions of the following extremization equations for Φ, which must be fulfilled at all steps $1 \leq t \leq T-1$:

$$\begin{aligned} w(m^{t+1} + m^t) &= h^{t+1} + w\,\tanh(w\,h^{t+1})\,, \\ \operatorname{arctanh}\,(m^t) &= w(h^t + h^{t+1}) - w^2 m^t\,. \end{aligned} \tag{18}$$

An example of transition path obtained through brute force numerical minimization of $\Phi\left(\{m^t,h^t\}\right)$ is shown in Fig. 3(a). It is composed of two portions:

- an initial part of the trajectory ascending the free energy landscape from one stable state, say, $+m^*$ up to the free-energy local maximum, $m = 0$. This part is associated with an exponentially small probability, i.e., to a positive contribution to the action, $\delta\Phi > 0$ (see Fig. 3(b)).
- a final part of the trajectory descending the free energy landscape from the local maximum $m = 0$ down to the other stable state, say, $-m^*$. This stretch does not seem to contribute to the action, $\delta\Phi \simeq 0$ (see Fig. 3(b)).

As the number T of steps increases the total action decreases, as expected, and quickly converges towards a minimal value (see Fig. 3(c)). We show below that the

scenario above can be analytically understood when T is sent to infinity.

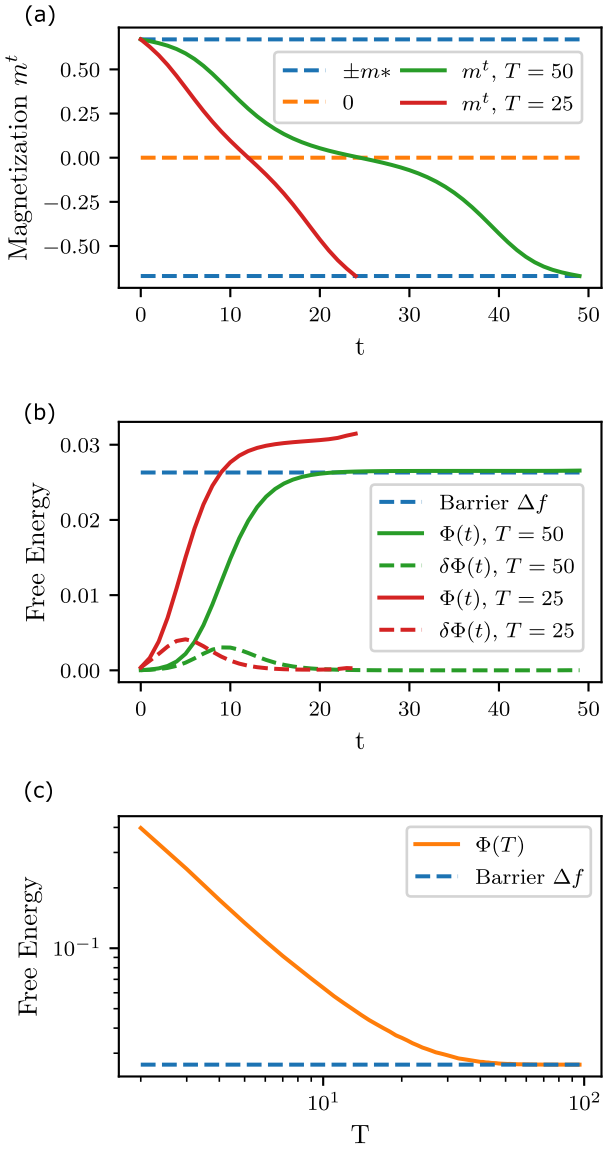

FIG. 3. Numerical minimization of $\Phi\left(\{m^t, h^t\}\right)$ for $w = 1.1$ with boundary conditions $m^0 = -m^T = m^*$. (a) Optimal time course of the magnetizations for $T = 25$ (red) and $T = 50$ (green) AGS steps. (b) Contributions $\delta\Phi(t)$ and full action $\Phi(t)$ as a function of the number of AGS steps for the optimal paths of duration $T = 25$ and $T = 50$. (c) Cost Φ of the optimal path as a function of T. For large T, Φ reaches from above a plateau equals to the free energy barrier Δf of the CW model, see Eq. (12). The convergence is exponentially fast, with decay time $T_{\rm decay} \sim 1/\log(w^2)$.

3. *Analytical expressions of the optimal trajectories in the $T \to \infty$ limit*

In the infinite T limit, the equations of motion (18) admit two distinct solutions that correspond to the two-fold behavior empirically observed for finite T.

a. Instanton-like trajectories. The ascending trajectories correspond to instantons, connecting a local minimum of the free energy to the local maximum, and are described by

$$\begin{aligned} m^{t+1} &= \frac{1}{w^2} \operatorname{arctanh}\left(m^t\right) , \\ h^{t+1} &= w\, m^{t+1} . \end{aligned} \tag{19}$$

Inserting these equations into Eq. (17), the contribution to the action associated to one AGS step reads, after some algebra,

$$\delta\Phi = f\left(m^{t+1}\right) - f\left(m^t\right) , \tag{20}$$

where $f(m)$ is the free energy of the CW model for magnetization m. The only stable fixed point of this dynamics is the local maximum of $f(m)$ in $m = 0$. Starting from $m^0 = m^*$, the dynamics converges to $m = 0$ for $T \to \infty$. Along this path, $\Phi\left(\{m^t, h^t\}\right) \underset{T\to\infty}{\longrightarrow} f(0) - f(m^*) = \Delta f$ (see Fig. 3(c)). Hence this path has a log-probability (per variable) equal to minus the free energy barrier separating the minima of the landscape.

b. Thermalization-like trajectories. The descending portion of the trajectory corresponds to relaxation towards the other minimum of the free energy and is described by the following solution of the extremization equations:

$$\begin{aligned} m^{t+1} &= \tanh\left(w^2\, m^t\right) , \\ h^{t+1} &= w\, m^t . \end{aligned} \tag{21}$$

We find that the contribution of an alternating step of AGS to the action vanishes:

$$\delta\Phi = 0 . \tag{22}$$

The stable fixed points of the dynamics are the two minima of $f(m)$. Starting from $m^0 = 0$ at time $t = 0$, the dynamics converges, when $T \to \infty$, to the spontaneous magnetization $\pm m^*$ associated to the minima of $f(m)$. Along this relaxation part of the trajectory, $\Phi\left(\{m^t, h^t\}\right) = 0$.

As a summary, the probability that a sequence of T steps of Alternating Gibbs Sampling brings the system from one minimum of the free energy to the other is given, to the dominant order in N, by $\exp\left(-N\Delta f\right)$. This result holds when N and T are very large (but with $T \ll N$). We conclude that it will take the same time τ as with the MH procedure, see Eq. (12), for the system to switch state. In other words, AGS is as inefficient as MH for sampling the bi-modal distribution associated with the CW model.

B. Case of unstructured multi-modal distribution

We now consider the case of a multi-modal distribution, where more than two states have high probabilities.

1. Hopfield model

Let us call $\boldsymbol{\xi}^\mu$ ($\mu = 1...M$) the centers of the states, which we suppose to be orthogonal in the infinite N limit. We assume that $\xi_i^\mu = \pm 1$. The order parameter is the M-dimensional vector of magnetizations along the centers, called patterns,

$$m^\mu = \frac{1}{N}\sum_{i=1}^{N} \langle v_i \rangle \xi_i^\mu \ . \tag{23}$$

We will hereafter consider the limit $\frac{M}{N} \to 0$. To be more precise, the energy over the visible configurations corresponds to the Hopfield model [33], and is defined through

$$E^{\rm Hop}(\mathbf{v}) = -\frac{w^2}{2N}\sum_{i,j=1}^{N}\left(\sum_{\mu=1}^{M}\xi_i^\mu\xi_j^\mu\right) v_i v_j \ , \tag{24}$$

where w^2 is the inverse temperature of the model. The free energy (per site) is given by

$$f(\mathbf{m}) = -\frac{w^2}{2}\sum_{\mu=1}^{M} m_\mu^2 - \mathcal{S}^{\rm Hop}(\mathbf{m}) \ , \tag{25}$$

where $\mathcal{S}^{\rm Hop}(\mathbf{m})$ denotes the entropy of the visible configurations at fixed magnetizations. It can be computed from the following Legendre formula

$$\mathcal{S}^{\rm Hop}(\mathbf{m}) = \min_{\boldsymbol{\lambda}}\left(\frac{1}{N}\sum_{i=1}^{N}\log 2\cosh\left(\sum_{\mu=1}^{M}\xi_i^\mu\lambda_\mu\right) - \sum_{\mu=1}^{M}\lambda_\mu m_\mu\right) \ . \tag{26}$$

The minimum is reached in the unique $\boldsymbol{\lambda}^*$ such that

$$m_\mu = \frac{1}{N}\sum_i \xi_i^\mu \tanh\left(\sum_\nu \xi_i^\nu \lambda_\nu^*\right) \ , \tag{27}$$

for all μ's. $\mathcal{S}^{\rm Hop}(\mathbf{m})$ can be expressed as a function of $\boldsymbol{\lambda}^*$ and the binary entropy $\mathcal{S}(m)$ in Eq. (11):

$$\mathcal{S}^{\rm Hop}(\mathbf{m}) = \frac{1}{N}\sum_i \mathcal{S}\left(\tanh\left(\sum_\mu \xi_i^\mu \lambda_\mu^*\right)\right) \ . \tag{28}$$

The Hopfield model can be represented with a RBM with N visible units (with potentials $\mathcal{V}_i = 0$) and M hidden units subject to the quadratic potential $\mathcal{U}(h) = \frac{h^2}{2}$ [16, 34, 35]. The energy of the RBM in Eq. (4) reads

$$E^{Hop}(\mathbf{v},\mathbf{h}) = -\sum_{i,\mu} W_{i\mu} v_i \, h_\mu + \sum_\mu \frac{h_\mu^2}{2} \ . \tag{29}$$

It is straightforward to check, after integration over the M hidden units, that the effective energy in Eq. (5) coincides with the Hopfield energy in Eq. (24) provided the weights fulfill the constraints

$$\sum_\mu W_{i\mu} W_{j\mu} = \frac{w^2}{N}\sum_\mu \xi_i^\mu \xi_j^\mu \ . \tag{30}$$

These conditions do not uniquely define the weight matrix $\mathbf{W}$. The energy is invariant under any transformation $\mathbf{W} \to \mathbf{W} \times \mathbf{O}$, where $\mathbf{O}$ is an orthogonal matrix. We choose for now the following parametrization for the weight matrix $\mathbf{W}$:

$$W_{i\mu} = \frac{w}{\sqrt{N}}\,\xi_i^\mu \ . \tag{31}$$

Alternative choices will be discussed later.

2. Optimal sampling with AGS

The AGS procedure can be entirely described in terms of M magnetizations $\mathbf{m}$ of the visible configurations and of the values $\mathbf{h}$ of the M hidden units. As in the case of the CW model, to get intensive quantities in the large N limit, we rescale $\mathbf{h} \to \mathbf{h}/\sqrt{N}$. The conditional configuration of the hidden unit $\mathbf{h}^{t+1}$ given a visible configuration with magnetization $\mathbf{m}^t$ is factorized, and reads:

$$P\big(h_\mu^{t+1}|\mathbf{m}^t\big) = \frac{1}{\sqrt{2\pi/N}}\,\exp\left(-\frac{N}{2}\Big(h_\mu^{t+1} - w\,m_\mu^t\Big)^2\right) \ . \tag{32}$$

The conditional probability of $\mathbf{m}^t$ given the hidden unit $\mathbf{h}^t$ can be easily written to the leading order in N, with the result

$$P\big(m_\mu^t|h_\mu^t\big) = \exp\left(-\sum_{i=1}^{N}\log 2\cosh\left(w\sum_{\mu=1}^{M}\xi_i^\mu h_\mu^t\right)\right) \times \exp\left(N\left(w\sum_{\mu=1}^{M} h_\mu^t\, m_\mu^t + \mathcal{S}^{\rm Hop}(\mathbf{m}^t)\right)\right) \ . \tag{33}$$

Similarly to the CW case, the probability of going from one minimum of the free energy landscape to another in T steps of AGS can be expressed as

$$P\left(\mathbf{m}^T|\mathbf{m}^0\right) = \exp\left(-N\min_{\{m^t,h^t\}}\Phi\left(\{\mathbf{m}^t,\mathbf{h}^t\}\right)\right) \ , \tag{34}$$

where the action $\Phi\left(\{\mathbf{m}^t,\mathbf{h}^t\}\right)$ is the sum of

$$\delta\Phi\left(t\to t+1\right) = \frac{1}{2}\sum_\mu (h_\mu^{t+1} - w\,m_\mu^t)^2 + \frac{1}{N}\sum_i \log 2\cosh\left(w\sum_\mu \xi_i^\mu h_\mu^{t+1}\right) - w\sum_\mu m_\mu^{t+1}\,h_\mu^{t+1} - \mathcal{S}^{\rm Hop}(\mathbf{m}^{t+1}) \ . \tag{35}$$

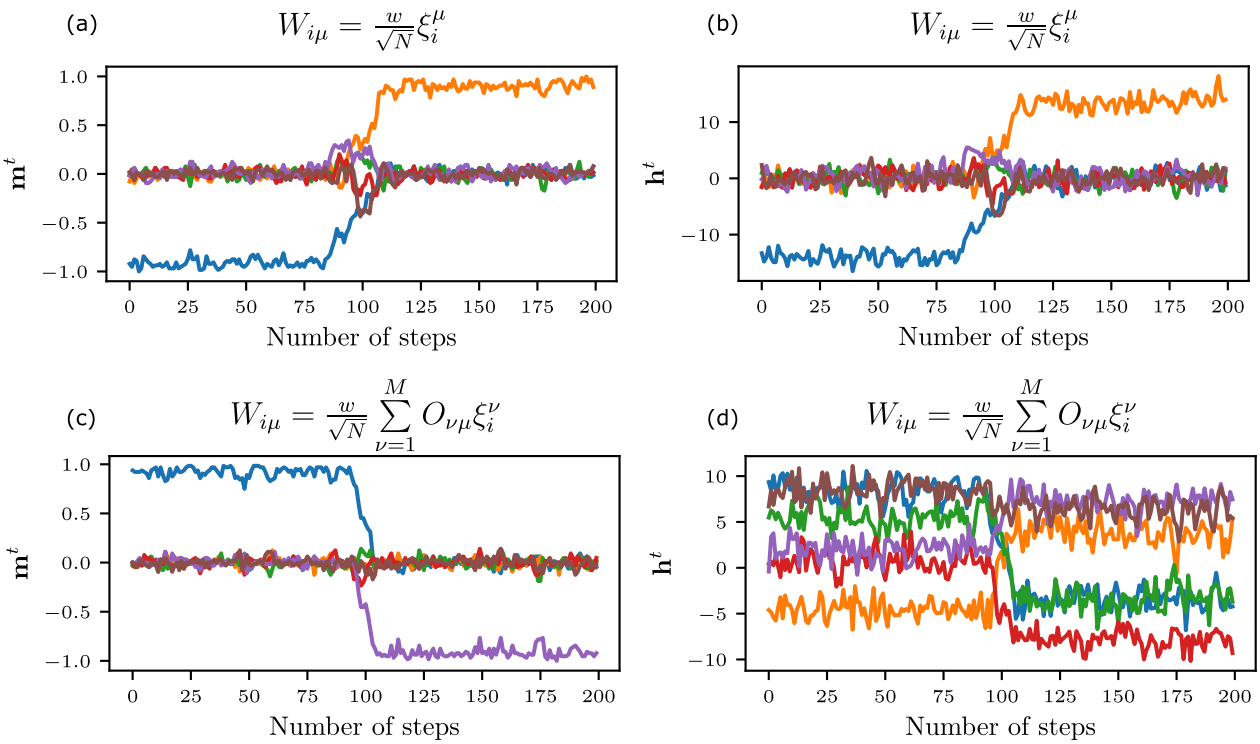

FIG. 4. Hopfield model over $N = 128$ sites, with $M = 6$ patterns and $w = 1.35$; each color refers to one index μ. Examples of transition between two states for $W_{i\mu} = \frac{w}{\sqrt{N}}\xi_i^\mu$ (panels a-b) and for $W_{i\mu} = \frac{w}{\sqrt{N}}\sum_\nu O_{\nu\mu}\xi_i^\nu$ (panels c-d). (a-c) Magnetizations m_μ along the patterns as functions of the number of AGS steps. (b-d) Hidden unit values h_μ as functions of the number of AGS steps for the same transitions as in panels (a-c).

The set of magnetizations $\mathbf{m}^t$ and hidden-unit values $\mathbf{h}^t$ minimizing the action Φ define the most likely path interpolating between two states in T AGS steps. They are solutions of the following extremization equations for Φ, which must be fulfilled at all steps $1 \leq t \leq T-1$:

$$(\lambda^*)_\mu^t = w(h_\mu^t + h_\mu^{t+1}) - w^2 m_\mu^t \ , \tag{36}$$

$$w(m_\mu^{t+1} + m_\mu^t) \ = h_\mu^{t+1} + \frac{w}{N}\sum_i \xi_i^\mu \tanh\left(w\sum_\nu \xi_i^\nu h_\nu^{t+1}\right) \ .$$

3. *Analytical expressions of the optimal trajectories in the $T \to \infty$ limit*

As for the CW model, we find

a. Instanton-like trajectories , defined by

$$h_\mu^{t+1} = w\, m_\mu^{t+1} = \frac{1}{w}(\lambda^*)_\mu^t \ , \tag{37}$$

$$m_\mu^t = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \tanh\left(w^2 \sum_{\mu=1}^{M} \xi_i^\mu m_\mu^{t+1}\right) \ .$$

The contribution to the action associated with this AGS step reads

$$\delta\Phi = f\left(\mathbf{m}^{t+1}\right) - f\left(\mathbf{m}^t\right) \ . \tag{38}$$

b. Thermalization-like trajectories , corresponding to

$$h_\mu^{t+1} = w\, m_\mu^t = \frac{1}{w}(\lambda^*)_\mu^{t+1} \ , \tag{39}$$

$$m_\mu^{t+1} = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \tanh\left(w^2 \sum_{\mu=1}^{M} \xi_i^\mu m_\mu^t\right) \ .$$

The contribution to the action associated with such an AGS step vanishes:

$$\delta\Phi = 0 \ . \tag{40}$$

c. Orthogonal transformation of the weight matrix. The computation can be repeated for a weight matrix $\tilde{\mathbf{W}} = \mathbf{W} \times \mathbf{O}$ where $\mathbf{O}$ is an orthogonal matrix. In the limit $T \to \infty$, instanton-like and themalization-like trajectories are found, and contributions to the action for both trajectories are the same as for $\mathbf{W}$. Therefore, the barriers are identical for all rotations $\mathbf{O}$. However, contrary to the previous case where the hidden unit h_μ codes for the magnetization m_μ only ($h_\mu = w\, m_\mu$), under an orthogonal transformation of the weight matrix, the hidden unit h_μ represents a superposition: $h_\mu = w\sum_{\nu=1}^{M} O_{\nu\mu} m_\nu$.

4. Transition paths between Mattis states

In the thermodynamic limit, the $\boldsymbol{\xi}^\mu$ are orthogonal. The free energy landscape $f(\mathbf{m})$ (Eq. (25)) exhibit a large variety of critical points when $w^2 > 1$ [36, 37], defined through Eq. (23), with

$$\langle v_i \rangle = \tanh\left(w^2 \sum_{\mu=1}^{M} \xi_i^\mu m_\mu \right) . \tag{41}$$

Global minima of Eq. (25) are reached for magnetization with only one non zero component, called Mattis states [38]. Numerical experiments for finite N exhibit transitions between the Mattis states, for all orthogonal transformation $\tilde{\mathbf{W}} = \mathbf{W} \times \mathbf{O}$ (see Figs. 4(a) & (c)). However, the hidden representations of the path between Mattis states may be easy or difficult to interpret depending on the orthogonal transformation (see Figs. 4(b) & (d)).

Furthermore, as for CW, for large T and N (with $T \ll N$), the probability to go from one Mattis state to another scale as $\exp(-N\Delta f)$. The barrier Δf depends on w and is always positive for $w^2 > 1$ [36]. Therefore AGS is as inefficient as MH for sampling the Hopfield model.

C. Case of structured multi-modal distributions

We now turn to a more complex case of multi-modal distributions, in which the states do not occupy uncorrelated pockets of configurations in the visible space but are structured. In addition, contrary to the previous models, the hidden units h_μ, which can be discrete or continuous, are now subject to an arbitrary, not necessarily quadratic potential $\mathcal{U}_\mu(h_\mu)$. Common potentials in the machine learning community are Bernoulli or ReLU potentials [15, 39], see Appendix A.

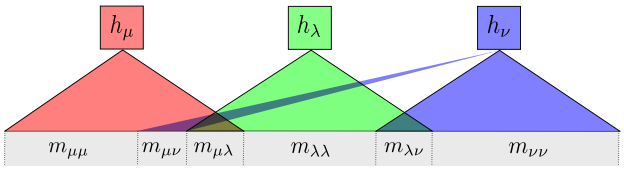

FIG. 5. Illustration of the structured model for $M = 3$ hidden units. The structural overlap matrix $\boldsymbol{\alpha}$ divides the visible layer into six different areas labeled by μ, ν, with $1 \leq \mu \leq \nu \leq M$. For each area, we define the corresponding normalized magnetization $m_{\mu\nu}$.

The $N \to \infty$ visible units v_i are ± 1 variables, and no potential acts on them ($\mathcal{V}_i = 0$). A visible unit v_i is connected to one or two hidden units with equal weights $\frac{w}{\sqrt{N}}$, following a pattern of connections shown in Fig. 5. We define the adjacency matrix $\boldsymbol{a}$ of our model as:

$$a_{i\mu} = \begin{cases} 1 & \text{if } W_{i\mu} = \frac{w}{\sqrt{N}} \\ 0 & \text{otherwise.} \end{cases} \tag{42}$$

From the adjacency matrix $\boldsymbol{a}$, we define the overlap matrix $\boldsymbol{\alpha}$ and the magnetization matrix $\boldsymbol{m}$:

$$\alpha_{\mu\mu} = \frac{1}{N} \sum_{i=1}^{N} a_{i\mu} \prod_{\nu \neq \mu} (1 - a_{i\nu}) , \tag{43}$$

$$\alpha_{\mu\nu} = \frac{1}{N} \sum_{i=1}^{N} a_{i\mu} a_{i\nu} , \tag{44}$$

$$m_{\mu\mu} = \frac{1}{\alpha_{\mu\mu} N} \sum_{i=1}^{N} \langle v_i \rangle \, a_{i\mu} \prod_{\nu \neq \mu} (1 - a_{i\nu}) , \tag{45}$$

$$m_{\mu\nu} = \frac{1}{\alpha_{\mu\nu} N} \sum_{i=1}^{N} \langle v_i \rangle \, a_{i\mu} a_{i\nu} . \tag{46}$$

In other words, there are $\alpha_{\mu\mu} N$ visible units connected only to h_μ, and $\alpha_{\mu\nu} N$ visible units connected to both h_μ and h_ν. The overlap matrix $\boldsymbol{\alpha}$ partitions the visible layer into $\frac{M(M+1)}{2}$ subsets with associated magnetizations $\mathbf{m}$ (Fig. 5).

It is straightforward to write down the free energy per variable $f(\mathbf{m})$ as a function of the $\frac{M(M+1)}{2}$ magnetizations, with the result

$$f(\mathbf{m}) = -\sum_{\mu=1}^{M} \hat{\Gamma}_\mu \left(w \sum_{\nu=1}^{M} \alpha_{\mu\nu} \, m_{\mu\nu} \right) - \sum_{\nu \leq \mu} \alpha_{\mu\nu} \, \mathcal{S}(m_{\mu\nu}) , \tag{47}$$

where $\hat{\Gamma}_\mu$ is the rescaled cumulative generative function associated with the hidden potential $\mathcal{U}_\mu$, see Eq. (5) and Appendix A, and $\mathcal{S}(m)$ is the entropy associated to a single ± 1 variable with magnetization m. The minima of $f(\mathbf{m})$ obey the following self-consistent equations,

$$\begin{aligned} m^*_{\mu\mu} &= \tanh\left(w \, f_\mu \left(I^*_\mu \right) \right) , \\ m^*_{\mu\nu} &= \frac{m^*_{\mu\mu} + m^*_{\nu\nu}}{1 + m^*_{\mu\mu} m^*_{\nu\nu}} , \end{aligned} \tag{48}$$

where $I^*_\mu = w \sum_{\nu=1}^{M} \alpha_{\mu\nu} m^*_{\mu\nu}$ is the input received by the hidden unit h_μ and $f_\mu = \hat{\Gamma}'_\mu$ is the transfer function associated with hidden unit h_μ.

1. Optimal sampling paths with AGS

We may now express the conditional probabilities of the magnetization matrix $\mathbf{m}$ (of dimension $M \times M$) and of the hidden-unit value vector $\mathbf{h}$ (of dimension M) following what was done for the simpler models in the previous sections. We first write the conditional probability of the hidden configuration given a set of visible activities,

$$\begin{aligned} P\left(h^{t+1}_\mu | \mathbf{m}^t\right) &= \frac{\exp\left(-N\left(\mathcal{U}_\mu(h^{t+1}_\mu) - h^{t+1}_\mu I^t_\mu\right)\right)}{\int dh \exp\left(-N\left(\mathcal{U}_\mu(h) - h I^t_\mu\right)\right)} \\ &\simeq \exp\left(-N\left(\mathcal{U}_\mu(h^{t+1}_\mu) - h^{t+1}_\mu I^t_\mu\right)\right) \\ &\times \exp\left(-N \hat{\Gamma}_\mu \left(I^t_\mu\right)\right) , \end{aligned} \tag{49}$$

where we have defined the input $I_\mu^t = w \sum_{\nu=1}^{M} \alpha_{\mu\nu} m_{\mu\nu}^t$ received by the hidden unit h_μ given the magnetization matrix $\mathbf{m}^t$.

In turn, we write the conditional probability over magnetizations given the set of hidden-unit values (to dominant order in N),

$$P(\mathbf{m}^t|\mathbf{h}^t) \simeq \exp\left(N\left(\sum_{\mu=1}^{M} I_\mu^t h_\mu^t - \alpha_{\mu\mu} \log 2\cosh\left(w\, h_\mu^t\right) - \sum_{\mu\leq\nu} \alpha_{\mu\nu} \log 2\cosh\left(w(h_\mu^t + h_\nu^t)\right) + \alpha_{\mu\nu}\, \mathcal{S}(m_{\mu\nu}^t) \right)\right) . \quad (50)$$

The probability to go from one minimum of the free energy landscape to another in T steps of AGS, $P\left(\mathbf{m}^T|\mathbf{m}^0\right)$, takes the same form as Eq. (16), where the action $\Phi\left(\{\mathbf{m}^t, \mathbf{h}^t\}\right)$ is the sum of

$$\begin{aligned}
\delta\Phi\left(t \to t+1\right) = &\sum_{\mu=1}^{M} \mathcal{U}_\mu(h_\mu^{t+1}) + \sum_{\mu=1}^{M} \hat{\Gamma}_\mu\left(I_\mu^t\right) \\
&+ \sum_{\mu=1}^{M} \alpha_{\mu\mu}\, \log 2\cosh\left(w\, h_\mu^{t+1}\right) \\
&+ \sum_{\mu\leq\nu} \alpha_{\mu\nu} \log 2\cosh\left(w(h_\mu^{t+1} + h_\nu^{t+1})\right) \\
&- \sum_{\mu=1}^{M} (I_\mu^{t+1} + I_\mu^t) h_\mu^{t+1} \\
&- \sum_{\mu\leq\nu} \alpha_{\mu\nu}\, \mathcal{S}(m_{\mu\nu}^{t+1}) \ .
\end{aligned} \quad (51)$$

Notice that the previous expression extends the model studied in Section III A, which can be recovered for $M = 1$, $\alpha_{11} = 1$ with a quadratic potential $\mathcal{U}(h) = \frac{h^2}{2}$.

We show the best path found through minimization of Φ in the case of $M = 2$ hidden units, quadratic $\mathcal{U}(h)$, $w > 1$, and small positive overlap α_{12}. The free energy landscape $f(\mathbf{m})$ represents two coupled Curie-Weiss models (see Fig. 6(a)), and displays two global minima and two local minima. The green trajectory shows the most likely path connecting the two global minima in $T = 100$ steps. Along this path, m_{11}^t and m_{22}^t, and therefore h_1^t and h_2^t, have asymmetric behaviors. In contradistinction, trajectories along which m_{11}^t and m_{22}^t are equal have exponentially smaller probabilities, see the red path. We elucidate this behavior below.

2. *Optimal trajectories in the $T \to \infty$ limit*

The set of magnetizations $\mathbf{m}^t$ and hidden-unit values $\mathbf{h}^t$ minimizing the action Φ define the most likely path, with AGS, capable of moving the system from one state to another in T alternating sampling steps. They are solutions of the following extremization equations for Φ, which must be fulfilled at all steps $1 \leq t \leq T-1$:

$$\begin{aligned}
I_\mu^t + I_\mu^{t+1} = &\mathcal{U}_\mu{}'(h_\mu^{t+1}) + w\, \alpha_{\mu\mu}\, \tanh(w h_\mu^{t+1}) \\
&+ w \sum_{\nu\neq\mu} \alpha_{\mu\nu} \tanh\left(w\left(h_\mu^{t+1} + h_\nu^{t+1}\right)\right) \ , \quad (52)
\end{aligned}$$

$$m_{\mu\mu}^t = \tanh\left(w\left(h_\mu^{t+1} + h_\mu^t\right) - w\, \hat{\Gamma}_\mu'\left(I_\mu^t\right)\right) \ , \quad (53)$$

$$m_{\mu\nu}^t = \frac{m_{\mu\mu}^t + m_{\nu\nu}^t}{1 + m_{\mu\mu}^t m_{\nu\nu}^t} \ . \quad (54)$$

In the infinite T limit, these equations of motion admit two distinct solutions.

a. Instanton-like solutions correspond to an increase of free energy from a local minimum, to a saddle-point of $f(\mathbf{m})$. These solutions can be written as:

$$\begin{aligned}
h_\mu^{t+1} &= f_\mu(I_\mu^{t+1}) \ , \quad (55) \\
m_{\mu\mu}^t &= \tanh(w f_\mu(I_\mu^{t+1})) \ .
\end{aligned}$$

Inserting these equations into Eq. (51):

$$\delta\Phi = f(\mathbf{m}^{t+1}) - f(\mathbf{m}^t) \ . \quad (56)$$

b. Thermalization-like solution makes the free energy decrease until a local minimum is reached. The relaxation solution can be written as:

$$\begin{aligned}
h_\mu^{t+1} &= f_\mu(I_\mu^t) \ , \quad (57) \\
m_{\mu\mu}^{t+1} &= \tanh(w\, f_\mu(I_\mu^t)) \ .
\end{aligned}$$

Inserting these equations into Eq. (51):

$$\delta\Phi = 0 \ . \quad (58)$$

While instantonic and thermalization trajectories are, strictly speaking, defined for $T \to \infty$ qualitatively analogous bouts of trajectories are observed for finite T, see Fig. 6(b) & (c) for the $M = 2$ example above. The green and the red paths are each composed of a sequence of instantonic and thermalization stretches. In the case of the red path, starting from a global minimum, the instantonic dynamics leads to the global maximum of $f(\mathbf{m})$. The relaxation dynamics then brings the system down to the other global minimum. In the reencase of the green path, starting from a global minimum, the instantonic solution leads to a saddle point of $f(\mathbf{m})$, which is

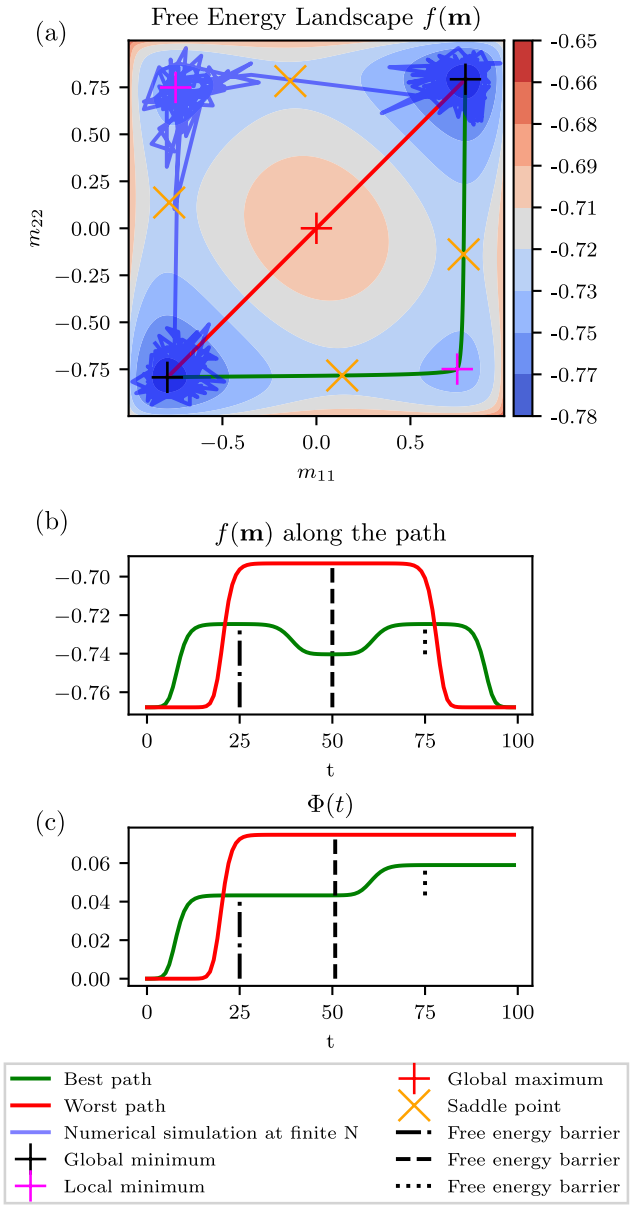

FIG. 6. (a) Free energy landscape for a coupled Curie-Weiss model with two global minima and two local minima. $M = 2$, $\mathcal{U}(h) = \frac{h^2}{2}$, $w = 1.15\sqrt{2}$ and $\alpha_{12} = 0.02$. Among the many paths connecting the two global minima in $T = 100$ steps, the green path is the optimal one. The red path is another path, along which both magnetizations m_{11} and m_{22} are equal at all times. The blue path is a representative trajectory found by simulating AGS for $N = 400$ and 10^5 steps. (b) Free energy $f(\mathbf{m}^t)$ along the different paths. (c) Cost $\Phi\left(\{\mathbf{m}^t, \mathbf{h}^t\}\right)$ for the different paths.

unstable for the instantonic and the thermalization dynamics. Then, the relaxation dynamics leads to a local minimum of $f(\mathbf{m})$. Through another pair of instantonic/relaxation dynamics, the second global minimum is finally reached. Thus, for the green and the red paths, the action $\Phi\left(\{\mathbf{m}^t, \mathbf{h}^t\}\right)$ corresponds to the sum of the free energy barriers along the paths (Figs. 6(b) & (c)). These theoretical findings are corroborated by running AGS on a RBM with $N = 400$ spins, with the same overlap matrix $\boldsymbol{\alpha}$. Along the transition path allowing the RBM to interpolate from one global state to the other, hidden units are preferentially flipped one by one, see the blue path in Fig. 6(a).

3. *Dependence of barrier upon structural overlap α*

This section examines the influence of the structural overlap on the free-energy barrier (and on the transition time) separating states. For the sake of simplicity, we focus on the case of $M = 2$ hidden units subject to quadratic potentials and restrict ourselves to small overlap values, $\alpha = \alpha_{12} \ll 1$. For $\alpha = 0$ the two global minima of $f(\mathbf{m})$ are $\mathbf{m}^*$ and $-\mathbf{m}^*$, where

$$\mathbf{m}^* = \begin{bmatrix} m_{11} = m^* \\ m_{22} = m^* \end{bmatrix}. \tag{59}$$

An optimal path between these two global minima follows the sequence of critical points:

$$\begin{bmatrix} m^* \\ m^* \end{bmatrix} \to \begin{bmatrix} 0 \\ m^* \end{bmatrix} \to \begin{bmatrix} -m^* \\ m^* \end{bmatrix} \to \begin{bmatrix} -m^* \\ 0 \end{bmatrix} \to \begin{bmatrix} -m^* \\ -m^* \end{bmatrix}, \tag{60}$$

and, for large T, $\Phi(T)$ equals the sum of the free-energy barriers along the path

$$\begin{aligned} \Phi &= -f\left(\begin{bmatrix} m^* \\ m^* \end{bmatrix}\right) + 2f\left(\begin{bmatrix} 0 \\ m^* \end{bmatrix}\right) - f\left(\begin{bmatrix} -m^* \\ m^* \end{bmatrix}\right) \\ &= -\log 2 + \frac{w^2}{2}(m^*)^2 + \mathcal{S}(m^*) \,. \end{aligned} \tag{61}$$

Assume now we make small changes to the weight and overlap values, *i.e.* $w \to w + dw, \alpha \to d\alpha$. We denote the displacement of the critical points of $f(\mathbf{m})$ by $d\mathbf{m}$, and the variations of the free energy by $df(\mathbf{m})$. We will consider only contributions to the first order in $d\alpha$ and dw,

$$d\mathbf{m} = \mathbf{m}^w dw + \mathbf{m}^\alpha d\alpha \,, \tag{62}$$

$$df(\mathbf{m}) = f^w(\mathbf{m}) dw + f^\alpha(\mathbf{m}) d\alpha \,. \tag{63}$$

Expressions for $\mathbf{m}^w$, $\mathbf{m}^\alpha$, $f^w(\mathbf{m})$ and $f^\alpha(\mathbf{m})$ are given in Appendix B.

As the variation of α changes the critical points of $f(\mathbf{m})$, we have to change w in order to keep fixed the two global minima $\pm\mathbf{m}^*$ of $f(\mathbf{m})$. Therefore, the variation of the cost Φ between an optimal path for $\alpha = d\alpha$ and one for $\alpha = 0$ defined in Eq. (60) reads

$$d\Phi = -df\left(\begin{bmatrix} m^* \\ m^* \end{bmatrix}\right) + 2\,df\left(\begin{bmatrix} 0 \\ m^* \end{bmatrix}\right) - df\left(\begin{bmatrix} -m^* \\ m^* \end{bmatrix}\right) . \tag{64}$$

As we observe in Fig. 7, a small overlap α reduces the cost for a wide range of w and therefore helps reduce the transition time between the global minima of f.

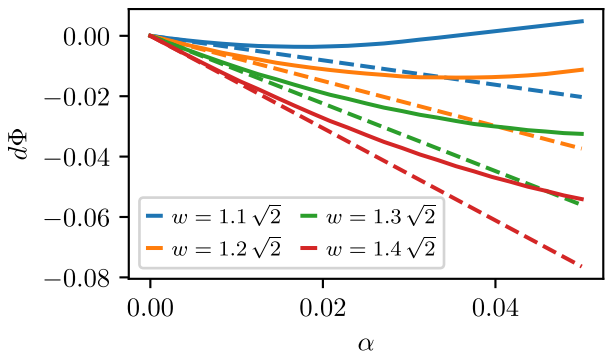

FIG. 7. Solid lines: numerical evaluation of $d\Phi$. Dashed lines: first order perturbation theory evaluated with Eq. (64).

4. *Time ordering of hidden-unit changes on sampling path*

As the optimal paths for the Alternating Gibbs Sampling are the ones that minimize the sum of the free energy barriers along the paths, the optimal paths depend strongly on the overlap matrix between the hidden units. If we impose a 1d structure with periodic boundary conditions for the overlap matrix, i.e $\alpha_{\mu\nu} = \alpha$ for $\nu = \mu - 1$ and $\nu = \mu + 1, \alpha_{\mu\mu} = \frac{1}{M} - \frac{M-1}{2}\alpha > 0$ (the hidden units are on a circle and have an overlap only with their two neighbors), the optimal path corresponds to an asymmetric behavior of the hidden units: they evolve one by one, according to their orders on the circle (h_μ evolves then $h_{\mu+1}$ then $h_{\mu+2}$ ...), see Fig. 8.

D. Numerical experiments

We train RBM with the datasets defined in Section II B, then test the performances of Alternating Gibbs Sampling. The different RBM can generate high-quality configurations, but the dynamics associated with Gibbs sampling struggles to mix efficiently between the data modes.

1. *BAS*

We train RBM with $2L$ real hidden units subject to quadratic potentials and ± 1 visible units. A L_1 regularization is added to the log-likelihood to enforce the sparsity of the weights. With this regularization, each hidden unit focuses on a given bar or a given stripe, see Section IV B for further details. Hidden units identify the relevant degrees of freedom of the visible units. For an image of bars, hidden units encoding the bars are strongly magnetized, and the hidden units encoding the stripes are weakly magnetized (they are silent). It is essential to use real hidden units because each hidden unit must have more than two equilibrium positions (strongly magnetized with positive or negative value, and weakly magnetized with positive or negative value). This behavior is not possible with discrete units like Bernoulli or Spin. AGS is inefficient for large L and long training, and the dynamics gets stuck in a bar or stripe configuration (Fig. 9). For short training, dynamics can escape from a given configuration but sampled configurations are noisy.

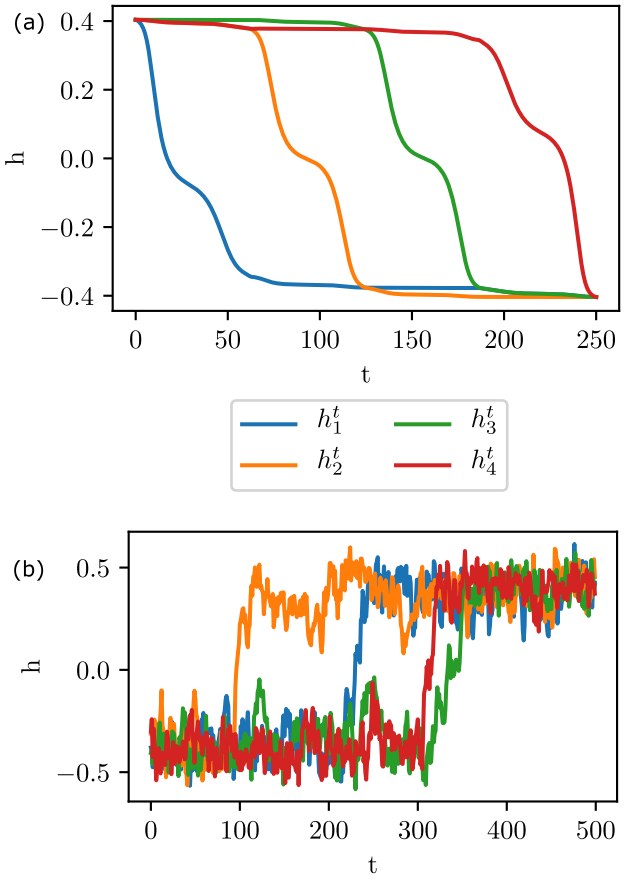

FIG. 8. Sampling paths for structured states. $M = 4$ hidden units are arranged on a ring, with $w = 2.2$ and $\alpha = 0.02$. (a) Numerical minimization of $\Phi\left(\{\mathbf{m}^t, \mathbf{h}^t\}\right)$ for $T = 250$. Hidden units are flipped according to their ordering on the ring ($h_1 \to h_2 \to h_3 \to h_4$). There are $2M$ equivalent optimal paths. (b) Numerical experiment on a RBM with $N = 400$ visible units. Hidden units are flipped according to their ordering on the ring ($h_2 \to h_1 \to h_4 \to h_3$).

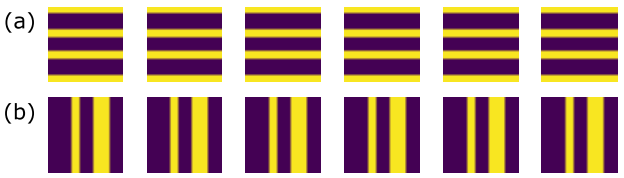

FIG. 9. Example of configurations obtain with AGS starting from a stripe (a) or a bar (b). 1000 steps between each frame.

2. MNIST 0/1

We train Spin-Spin RBM (hidden and visible units are ± 1 spins). The weights of the RBM encode the digits strokes. Zeros have many strokes in common, and so have ones. Therefore, the hidden representations of each digit are close to each other (in terms of Hamming distance). AGS is efficient to sample within a digit class and generate high-quality data, see Figs. 10(a) & (b). However, hidden representations of the zeros and the ones are far away from each other. Therefore, many hidden units should be simultaneously flipped to go from one class to another, which is very unlikely with AGS: the dynamics remains confined to one digit class, see Fig. 10(c). Notice that this observation crucially depends on the restriction of MNIST to 0-1 digits done here. RBM trained on all ten digits sample much more efficiently all classes and can reach 1 from 0 or vice versa [15, 40], as other digits carve interpolating paths in the energy landscape.

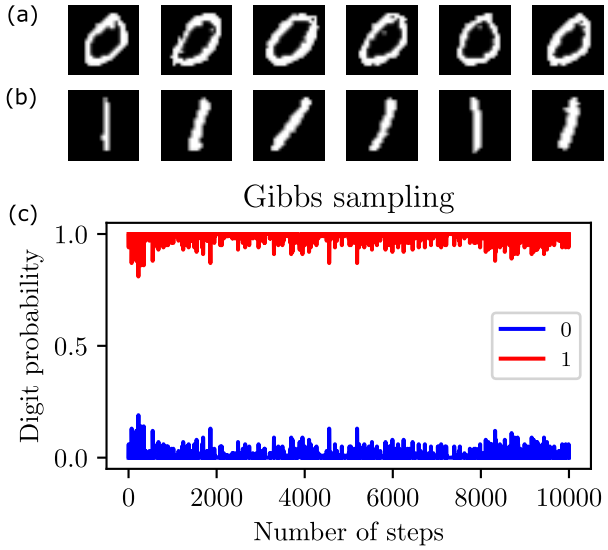

FIG. 10. Examples of digits obtained with AGS starting from a 0 (a) and from a 1 (b); 1000 steps between each frame. (c) Probabilities that the visible unit configurations sampled by the RBM at different times are 0 (blue) or 1 (red), estimated by a random forest classifier trained on 0-1 data [41, 42]. The dynamics is stuck in a given mode.

3. Lattice Proteins

To encode amino acids (which may take 20 values), we introduce RBM with categorical (Potts) visible units. Couplings between the hidden layer and the visible layers are represented by a $M \times N \times 20$ tensor. Thus, the energy of the RBM can be written as:

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{N}\sum_{\mu=1}^{M} W_{i\mu}(v_i)h_\mu + \sum_{\mu=1}^{M}\mathcal{U}_\mu(h_\mu) + \sum_{i=1}^{N}\mathcal{V}_i(v_i)\ .$$

The weights of the RBM encode the constraints, such as contacts between different amino acids defined by the structure. Contrary to the two previous examples, to generate high-quality proteins with the RBM, i.e., proteins with a high probability to fold in a given structure, the landscape has to be sampled at low temperatures. Using the trick introduced in [43], we copy each hidden unit $\beta \in \mathbb{N}$ times and multiply the visible fields by the same factor β:

$$P_\beta(\mathbf{v}) \propto \int \prod_{\mu=1}^{M}\prod_{c=1}^{\beta} P(\mathbf{v}|h_\mu^c) = P(\mathbf{v})^\beta\ . \qquad (65)$$

With this modification, it is possible to sample the landscape $P(\mathbf{v})$ at inverse temperature β. RBM generate high-quality proteins but struggles to mix between two families with essentially dissimilar contact maps, such as structures S_A and S_B defined in Fig. 2, see Fig. 11. Many hidden units would have to change at once, a very unlikely update with AGS to go from one family to another.

IV. ALTERNATING GIBBS SAMPLING AND DYNAMICS IN THE LATENT SPACE

A. Principle of the algorithm

We have shown in the previous Section III that AGS was as efficient as the local MH procedure to sample the landscape over the visible configurations, defined by the effective energy $E^{\text{eff}}(\mathbf{v})$. However, RBM offer more than this landscape, and it is natural to wonder if the representations of data could be exploited to enhance sampling performance. To do so, we propose a sampling algorithm combining AGS and moves in the *hidden unit* space, see Fig. 1(d) and Algorithm 3. The main idea is to exploit the fact that hidden units can encode specific features of the data. By doing Metropolis steps in the hidden space, we try to flip the hidden units one by one, or by blocks, for switching on/off the features they encode. This flipping procedure must obviously preserve detailed balance. We therefore need to know the effective energy over hidden configurations, $E^{\text{eff}}(\mathbf{h})$, defined by marginalizing the joint distribution $P(\mathbf{v},\mathbf{h})$ over the visible variables:

$$E^{\text{eff}}(\mathbf{h}) = -\log\left(\int d\mathbf{v}\, P(\mathbf{v},\mathbf{h})\right)\ . \qquad (66)$$

We can gain intuition about the exponential speed up offered by the algorithm in the latent space by considering first the CW model. In the absence of any bias (external field) between the + and − states of the visible variables, the effective energy $E^{\text{eff}}(\mathbf{h})$ is an even function of the hidden unit value h. A step of the sampling algorithm in the hidden space, see Algorithm 3, has thus probability $\frac{1}{2}$ to flip the hidden unit. Sampling back the visible layer will change the state of a macroscopic number of visible variables. Using MH algorithm in the hidden space is similar to using cluster algorithms for the

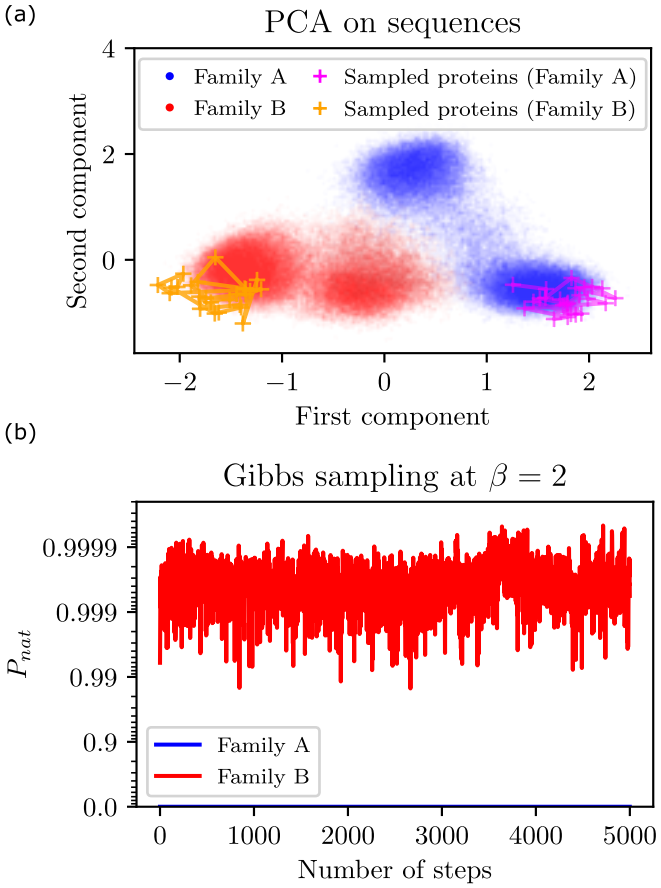

FIG. 11. (a) Principal Component Analysis in the sequence spaces, showing the cluster structure of each family (blue and red colors). Fuchsia and orange paths are the projection of sampled proteins with AGS, starting respectively from a protein in family A and B. Sampled proteins are stuck in a given family; 250 Gibbs steps between each cross. This number of steps is larger than the decorrelation time estimated from the Hamming distance between sequences $\mathbf{v}^t$. (b) $P_{\rm nat}(\mathbf{v}|S)$ of sampled proteins with AGS, for S_A and S_B, for an initial protein in the S_B family (orange path in panel a). RBM generates high-quality and diverse proteins, which are different from the training data.

visible spins [3, 4]. For ferromagnetic models, these algorithms are known to be much more efficient than local MH over spins [44–46]. The latent variable is here attached to the relevant collective mode (global reversal) of the spin variables.

For the mean-field structured models defined in Section III C, as long as the overlap between the hidden units is weak, the hidden units could be flipped one by one for moderate system size N. We define the potential acting on one hidden unit, say h_μ, conditional to the other units $\mathbf{h}_{\neg\mu}$ through

$$e_\mu(h_\mu|\mathbf{h}_{\neg\mu}) = \frac{1}{N} E^{\rm eff}\big(\mathbf{h} = (h_\mu, \mathbf{h}_{\neg\mu})\big) \ . \qquad (67)$$

Each flip of a hidden unit corresponds to a move from one local minimum to another in the landscape $e_\mu(h_\mu|\mathbf{h}_{\neg\mu})$,

Algorithm 3: Alternating Gibbs Sampling with Metropolis-Hastings steps in latent space

Pick $\mathbf{v}^0$ in the training set;
for $t \in [\![0, T]\!]$ **do**
  $\mathbf{h}^{t+1} \sim P(\mathbf{h}|\mathbf{v}^t)$;
  π = random permutation of $[\![1, M]\!]$;
  for $i = 1...M$ **do**
    $\mu = \pi(i)$;
    $h_\mu^{t+1} \sim P(h_\mu|\mathbf{h}_{\neg\mu}^{t+1})$;
  end
  $\mathbf{v}^{t+1} \sim P\left(\mathbf{v}|\mathbf{h}^{t+1}\right)$;
end

see Fig. 12. Metropolis steps in the hidden space can speed up the dynamics: the free energy barrier for Metropolis-Hastings in the hidden space, $N\Delta e_{\rm MH}$, where

$$\Delta e_{\rm MH} = -\frac{1}{N} \log \left[\frac{\int_0^\infty dh\, e^{-N e_\mu(h|\mathbf{h}_{\neg\mu})}}{\int_{-\infty}^0 dh\, e^{-N e_\mu(h|\mathbf{h}_{\neg\mu})}} \right] , \qquad (68)$$

is smaller than the free energy barrier $N\Delta f$ 'seen' by Alternating Gibbs Sampling.

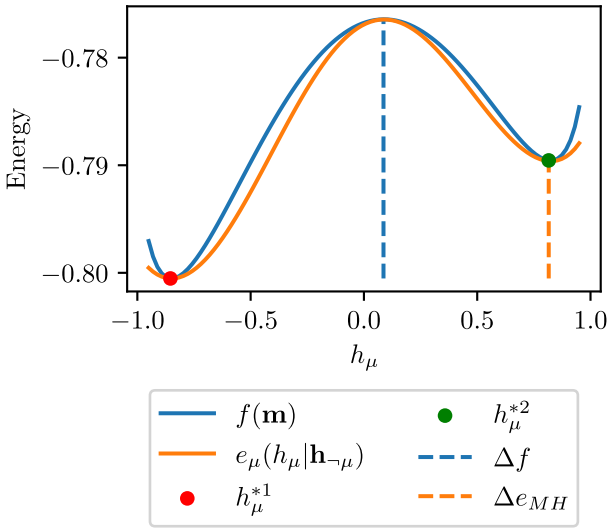

FIG. 12. Barriers in a structured model with $M = 5$ hidden units, with $w = 1.2\sqrt{5}$, $\alpha_{\mu\nu} = 0.03$ for all pairs $\mu \neq \nu$. All hidden units are frozen except h_μ. For small overlap between the hidden units, the potential $e_\mu(h_\mu|\mathbf{h}_{\neg\mu})$ has two local minima for two different values of h_μ, h_μ^{1*} and h_μ^{2*}. By sampling back the visible layer $P(\mathbf{m}|\mathbf{h})$, we see that there are two local minima for $f(\mathbf{m})$. Flipping the hidden unit h_μ allows one to go from one local minimum to another. The free energy barrier in the hidden space with Metropolis-Hastings algorithm $\Delta e_{\rm MH}$ is smaller than the free energy barrier of the Alternating Gibbs Sampling Δf.

B. Application to BAS

We train RBM trained on BAS with a L_1 regularization to enforce the sparsity of the weights. Thanks to the regularization, each hidden unit focuses on a given bar or a given stripe, see Fig. 13(a). The change $h_\mu \leftarrow -h_\mu$ leaves the energy $E^{\rm eff}(\mathbf{h})$ unchanged: a bar or a stripe can be present or not, see Fig. 13(c). We use a Gibbs sampling in the hidden space where one hidden unit is updated according to Algorithm 3. Our algorithm efficiently switches on/off these hidden units, see Fig. 14(a).

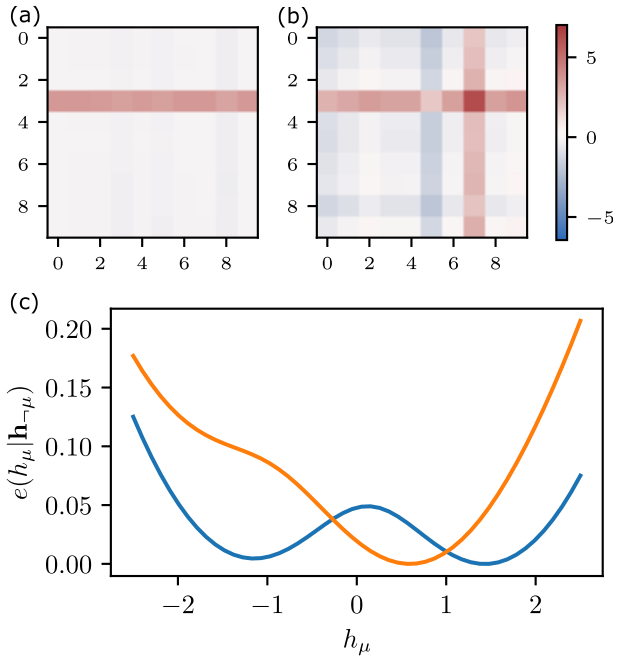

FIG. 13. Example of weights learned by RBM on BAS, $L = 10$. (a) With L_1 regularization. Each hidden unit focuses on a bar or stripe. (b) Without L_1 regularization. Each hidden unit focuses on several bars and stripes. (c) Potential $e_\mu(h_\mu|\mathbf{h}_{\neg\mu})$ for $\mathbf{h}$ associated with a stripe image; the minimum of the energy is set to zero. Solid blue line: hidden unit h_μ encoding a stripe; the two minima coding from the on/off stripe have roughly the same energy. Solid orange line: hidden unit h_μ encoding a bar, the minimum encoding the on bar has an energy much higher than the one corresponding to the off bar.

Notice that, without regularization, each hidden unit would focus on several bars and stripes (Fig. 13(b)). In that case, allowing for steps in the hidden-unit space does not help, and our algorithm is inefficient (Fig. 14(b)).

C. Application to the Hopfield model

We have seen in Section III B that, for large enough weight amplitude w, the AGS dynamics is stuck in one Mattis state of the Hopfield model, i.e., the magnetization $\mathbf{m}$ has only one component different from zero in the infinite size limit. The behavior of the hidden-unit configurations depends on the prescription of the weights, which may or may not be aligned with the states $\boldsymbol{\xi}^\mu$ (Eq. 30).

FIG. 14. Visible configurations obtained with Alternating Gibbs Sampling and Metropolis-Hastings algorithm in the hidden space, $L = 10$. 25 Gibbs steps between each frame. (a) With L_1 regularization. (b) Without L_1 regularization.

1. *Aligned weights*

Let us first assume that the weights are aligned with the states, , *i.e.* that Eq. (31) holds. The effective energy over the hidden configurations reads

$$E^{\rm eff}(\mathbf{h}) = \sum_\mu \frac{h_\mu^2}{2} - \sum_i \log 2\cosh\left(\frac{w}{\sqrt{N}}\sum_\mu \xi_i^\mu h_\mu\right) . \tag{69}$$

Identifying $\frac{h_\mu}{w\sqrt{N}} = m_\mu$, the effective energy is equal to the free energy of the Hopfield model derived in [36] at inverse temperature w^2. The representations of the Mattis states are very simple in the hidden space of the RBM. In the presence of $\boldsymbol{\xi}^\mu$ on the visible layer, one hidden unit, say, $\mu = 1$, is strongly magnetized: $h_1 = \mathcal{O}(\sqrt{N})$. The $M-1$ other hidden units are weakly activated: $h_\nu = \mathcal{O}(1)$ for $\nu \geq 2$. $E^{\rm eff}(\mathbf{h})$ has $2M$ global minima corresponding to the $2M$ Mattis states.

a. Single unit potential. According to Eq. (69) the potential over the strongly magnetized hidden unit $\mu = 1$ reads, after rescaling $h_1 \to h_1/\sqrt{N}$,

$$e_1(h_1|\mathbf{h}_{\neg 1}) = \frac{h_1^2}{2} - \log 2\cosh(wh_1) , \tag{70}$$

up to an additive constant. This potential has two global, opposed minima for $w^2 > 1$. The situation is similar to the CW model studied above: MH steps in the hidden-unit space allow for efficient sampling on the states $\boldsymbol{\xi}^1$ and $-\boldsymbol{\xi}^1$.

The potential on the other hidden units $\nu \neq 1$ is given by, up to an irrelevant additive constant and in the large-N limit, after rescaling $h_\nu \to h_\nu/\sqrt{N}$,

$$e_\nu(h_\nu|\mathbf{h}_{\neg\nu}) = \frac{h_\nu^2}{2} - (1-m_1^2)\left(\frac{1}{N}\sum_i \xi_i^1\xi_i^\nu\right)h_\nu . \tag{71}$$

Sampling this quadratic potential allows to better explore the Mattis state around $\boldsymbol{\xi}^1$, but it does not help changing state.

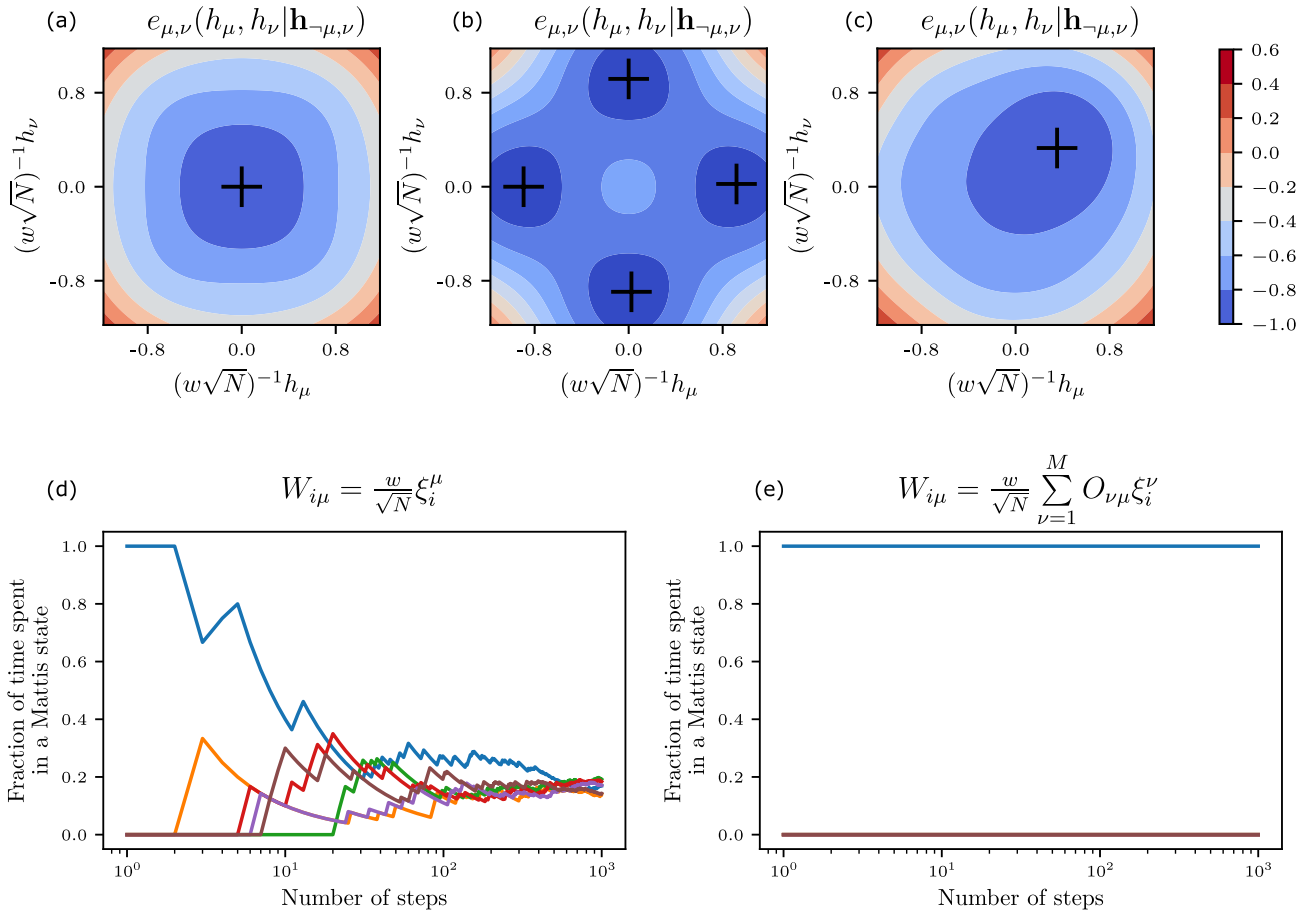

FIG. 15. Hopfield model encoded by a RBM with $N = 128$, $M = 6$ and $w = 1.5$ and orthogonal $\boldsymbol{\xi}^\mu$. (a), (b) and (c) represent the landscape $e_{\mu,\nu}(h_\mu, h_\nu|\mathbf{h}_{\neg\mu,\nu})$, where the $M-2$ other components of $\mathbf{h}$ are fixed. Black dots represent minima of the landscape. (a) $W_{i\mu} = \frac{w}{\sqrt{N}}\xi_i^\mu$. Initial configuration is h_λ strongly magnetized and $h_\mu \sim h_\nu = \mathcal{O}(1)$. Minimum is reached for $h_\mu \sim h_\nu = \mathcal{O}(1)$. (b) $W_{i\mu} = \frac{w}{\sqrt{N}}\xi_i^\mu$. Initial configuration is h_μ strongly magnetized and $h_\nu = \mathcal{O}(1)$. Four minima exist corresponding to the four possible Mattis states. (c) Case $W_{i\mu} = \frac{w}{\sqrt{N}}\sum_{\nu=1}^{N} O_{\mu\nu}\xi_i^\nu$. There exist only one minimum. (d) and (e) $\mathbf{v}^t$ are generated with AGS with MH steps in the hidden space. The fraction of time spent in a Mattis state is measured through time. (d) $W_{i\mu} = \frac{w}{\sqrt{N}}\xi_i^\mu$: the visible configuration $\mathbf{v}^t$ eventually visits all Mattis states with equal probabilities. (e) $W_{i\mu} = \frac{w}{\sqrt{N}}\sum_{\nu=1}^{N} O_{\mu\nu}\xi_i^\nu$: the dynamics gets stuck in a given Mattis state.

b. Two-unit potential. To speed up exploration of different states, we introduce the two-unit potentials

$$e_{\mu,\nu}(h_\mu, h_\nu|\mathbf{h}_{\neg\mu,\nu}) = \frac{1}{N}E^{\text{eff}}\big(\mathbf{h} = (h_\mu, h_\nu, \mathbf{h}_{\neg\mu,\nu})\big) \ , \quad (72)$$

where all but two hidden units are kept fixed. These potentials are plotted in Fig. 15. Two typical behaviors are encountered:

- μ, ν are both different from 1. The two-unit potential $e_{\mu,\nu}$ is simply the sum of the single-unit potentials e_μ and e_ν, see Eq. (71). Therefore $e_{\mu,\nu}$ has only one global minimum (Fig. 15(a)). Changing h_μ or h_ν does not allow for moving outside the state condensed $\boldsymbol{\xi}^1$.
- $\mu = 1$ and $\nu \neq 1$. Contrary to the previous case, h_1 is now a free parameter. Therefore, by tuning h_1 and h_ν, four global minima of $e_{1,\nu}$ can be reached, corresponding to the cases where h_1 or h_ν are strongly magnetized (with positive or negative values), see Fig. 15(b). We can exploit this structure by introducing a block Gibbs sampling in the hidden space, where two hidden units are updated simultaneously, see Algorithm 4. The dynamics can now explore all the Mattis states very efficiently, see Fig. 15(d).

Algorithm 4: Alternating Gibbs Sampling with Metropolis-Hastings updates of two hidden units

Pick $\mathbf{v}^0$ in the training set;
for $t \in [\![0, T]\!]$ **do**
  $\mathbf{h}^{t+1} \sim P(\mathbf{h}|\mathbf{v}^t)$;
  π = random pairing of $[\![1, M]\!]$, defining $M/2$ pairs of elements ;
  for $i \in [\![1, M/2]\!]$) **do**
    $\mu, \nu = \pi(i)$;
    $h_\mu^{t+1}, h_\nu^{t+1} \sim P(h_\mu, h_\nu|\mathbf{h}^{t+1}_{\neg\mu,\nu})$;
  end
  $\mathbf{v}^{t+1} \sim P\left(\mathbf{v}|\mathbf{h}^{t+1}\right)$;
end

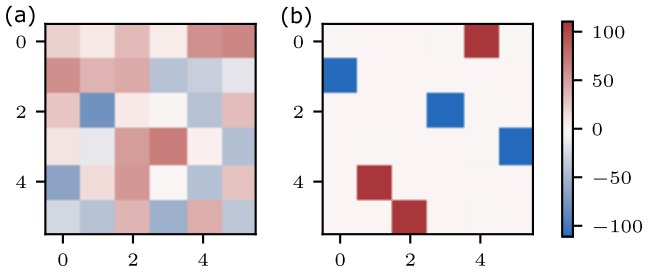

FIG. 16. Matrix product between the weight matrix $\mathbf{W}^T$ (size $M \times N$) and the matrix of patterns $\boldsymbol{\xi}$ (size $N \times M$). $N = 128$ and $M = 6$. (a) Without regularization, $\lambda_{\text{pen}} = 0$. Each pattern $\boldsymbol{\xi}^\mu$ has a delocalized representation in the hidden space. (b) With regularization, $\lambda_{\text{pen}} = 0.001$. Each pattern $\boldsymbol{\xi}^\mu$ strongly magnetized only one hidden unit.

2. *Rotated weights*

As already mentioned in Section III B, the conditions in Eq. 30 do not uniquely define the weight matrix $\mathbf{W}$. The Hopfield model energy is invariant under any transformation $\mathbf{W} \to \mathbf{W} \times \mathbf{O}$, where $\mathbf{O}$ is an orthogonal matrix. After this orthogonal transformation, the hidden representation of a Mattis state is delocalized: each component of $\mathbf{h}$ is strongly magnetized (of the order of $\sqrt{N}$). Single or two-unit potentials have one global minimum (Fig. 15(c)). Therefore, Metropolis-steps in the hidden space do not speed up sampling (Fig. 15(e)) unless all M hidden units are simultaneously updated.

Numerical experiments with RBM trained by gradient ascent on data sampled from the Hopfield model generally converge to a solution, where the hidden representation of a Mattis state is delocalized (Fig. 16(a)) [47]. By adding the following penalty term in the log-likelihood, it is possible to ensure that only one hidden unit is strongly magnetized and encodes for a specific pattern $\boldsymbol{\xi}^\mu$, see Fig. 16(b):

$$LL^{\text{pen}} = -\frac{\lambda_{\text{pen}}}{L} \sum_{\ell=1}^{L} \sum_{\mu \neq \nu} |f_\mu(\mathbf{v}^\ell) f_\nu(\mathbf{v}^\ell)| \ , \qquad (73)$$

where $\{\mathbf{v}^\ell\}_{\ell=1\ldots L}$ are the L samples in the training set. This penalty favors solutions where only one hidden unit is strongly magnetized. Its intensity is set by the parameter λ_{pen}.

V. CONCLUSION

This work presents a combination of analytical and numerical results on the dynamics defined by Alternating Gibbs Sampling of Restricted Boltzmann Machines and applied to several mean-field models. We have shown how this sampling procedure can find optimal transition paths between the local minima of the free-energy landscape over the visible configurations. However, large free-energy barriers, extensive in the system size, have to be crossed to go from one state to another. As a result, AGS is not more efficient than standard local Metropolis sampling of the effective energy of the visible configurations. Notice that our analytical results were derived in a double large-size setting, where the asymptotics on the size N of the system was considered first, and the time T of transition paths was made large afterward. In practice, the probabilities that these transitions paths successfully interpolate between states are exponentially small in N, which implies, in turn, that transitions almost surely happen on times scales growing exponentially in N (and equal to the inverse probabilities). As shown in Fig. 6(a), the system spends most of this exponential time attempting to escape local minima of the free energy landscapes, while transitions between the minima are actually fast (but rare).

The inability of AGS to outperform local sampling procedures in mixing between states calls for some comments. First, AGS, with Contrastive Divergence or Persistent Contrastive Divergence, remains an efficient training algorithm for RBM. These two procedures authorize initializations of the dynamics in different local minima close to the training data. Thus, even if AGS suffers from poor mixing between far away minima, the different minima close to the data may be well sampled. Second, AGS can be efficient when the different modes of data are connected through energy valleys. For example, AGS of RBM trained on all digits of MNIST can generate transition between 0 and 1. However, these transitions go through different intermediate states, which are other digits. When training RBM on zeros and ones only, as done in this paper, intermediate states do not exist: the two modes are not connected by low energy funnels, and transitions are unlikely to occur. Third, RBM are supposed to encode meaningful (hidden) representations, coding for collective features in the data. It is tempting to see these features as modes of excitation that could be flipped at once, similarly to what cluster algorithms achieve for ferromagnetic models.

In this context, we have shown how Metropolis-Hastings steps in the hidden space (in between the forward and backward passes of AGS) can enhance sampling

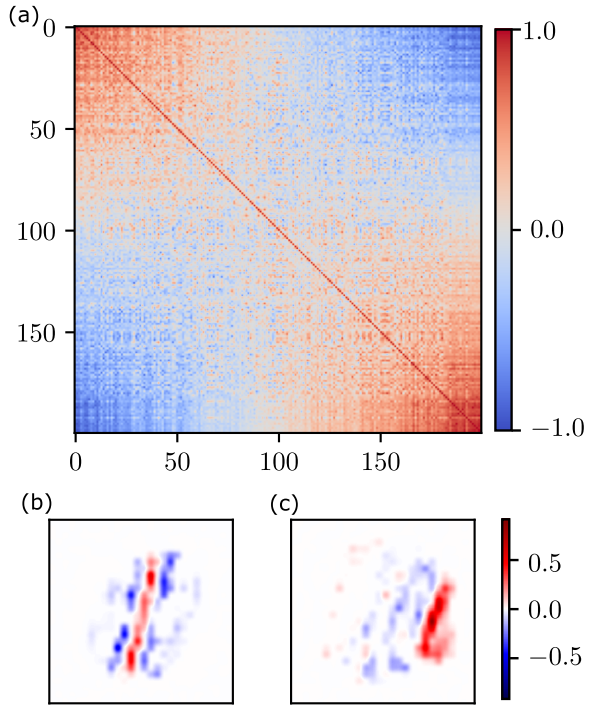

FIG. 17. RBM trained on MNIST 0/1, with $M = 200$ hidden units. (a) Correlation matrix of the inputs received by the hidden units on the training data. Hidden units are sorted according to the components of the top eigenvector on this matrix. Two clusters emerge, corresponding to 0's and 1's: each digit is attached to roughly half the hidden units. (b) Example of weights $\mathbf{W}_{i\mu}$ for a hidden unit μ associated to 1, corresponding to a stroke specific to 1. (c) Example of weights $\mathbf{W}_{i\mu}$ for a hidden unit μ associated to 0, corresponding to a stroke specific to 0.

performance when hidden units encode essentially independent features of the data. Updating of one or two (or a small number of) hidden units then allows for a macroscopic change of visible units and offers rapid mixing between states. We have illustrated this mechanism on the Bars and Stripes dataset and on the Hopfield model. In the latter case, the success of this procedure crucially depends on the specific set of weights output by the learning procedure, or, equivalently, on the nature of representations. MH updates in the hidden space are effective in the prototype-like regime, in which one or few strongly active hidden units rigidly determine the visible configurations [48]. This statement is expected to hold also in the so-called compositional regime, in which hidden-unit activity configurations are sparse, but the combinations of strongly activated latent variables are highly flexible and allow for a combinatorial number of visible states [15].

In the case of entangled representations, in which all (or a large number of) the hidden units are strongly magnetized (with different degrees of activation from one state to another), our combined AGS-MH procedure is inefficient, as flipping a small number of hidden units is unable to change the identity of the state. This phenomenon was illustrated on the Hopfield model in the case of 'rotated' weights, compare Figs. 15(d) & (e). In much the same way, MH updates of a small subset of the hidden units of RBM trained on MNIST 0/1 or Lattice Proteins do not significantly enhance mixing performances. Hidden units capture features of the data, such as digit strokes for MNIST, which are correlated. Changing state demands to tune a large number of hidden units, see Fig. 17. In other words, the very existence of collective modes of hidden units prevents the success of our AGS-MH procedure, which is local in the hidden space. Another illustration of these collective modes in the hidden space is provided by RBM trained on BAS. Even if our algorithm is efficient to sample within a given class (bars or stripes), it cannot go from one class to another. To go from an image of bars to an image of stripes, the hidden units encoding the bars have to be silent, and the hidden units encoding the stripes have to be strongly magnetized. These define two collective modes of the hidden units, which AGS-MH cannot change. We stress that the unability of AGS to achieve rapid mixing is not limited to mean-field-like models. Even in the case of RBM tailored to encode finite-dimensional models with high-order ferromagnetic interactions, AGS suffers from poor mixing, and efficient sampling could only be obtained by combining with cluster algorithms such as the Swendsen-Wang procedure [49]. In a forthcoming publication, we show how stack of RBM, with ideas proposed in [50, 51], can detect collective modes of hidden units and thus improve the sampling of the energy landscape.

ACKNOWLEDGMENTS

We are grateful to J. Tubiana for interesting discussions. Furthermore, we acknowledge fundings from Direction générale de l'armement (C. Roussel's PhD grant) and from the Agence Nationale de la Recherche, RBM-Pro Project 17-CE30-0021.

Appendix A: General hidden-unit potentials

We consider below three different potentials acting on hidden units, and how they should scale when $N \to \infty$.

a. Quadratic potential

The quadratic potential is defined as $\mathcal{U}_\mu(h_\mu) = \frac{h_\mu^2}{2}$. In that case, we should rescale $h_\mu \to h_\mu/\sqrt{N}$. We get:

$$P\big(h_\mu|\mathbf{m}\big) = \frac{1}{\sqrt{2\pi/N}} \exp\left(-\frac{N}{2}\big(h_\mu - I_\mu\big)^2\right) \ , \tag{A1}$$

$$\hat{\Gamma}_\mu(I) = \frac{I^2}{2} \ , \quad f_\mu(I) = I \ . \tag{A2}$$

b. ReLU potential

We can use the so-called ReLU (Rectified Linear Unit) potential $\mathcal{U}_\mu(h_\mu) = \frac{1}{2}\gamma^+ h_\mu^{+2} + \theta^+ h_\mu^+$ where $h_\mu^+ = \max(h_\mu, 0)$, see for instance [43]. We should rescale $h_\mu \to h_\mu/\sqrt{N}$ and $\theta_\mu^+ \to \theta_\mu^+/\sqrt{N}$. We get:

$$P\big(h_\mu|\mathbf{m}\big) = \mathcal{TN}\left(N\frac{I_\mu - \theta_\mu^+}{\gamma^+}, \frac{1}{\gamma^+}, \mathcal{R}^+\right) \ , \tag{A3}$$

$$\hat{\Gamma}_\mu(I) = \max\left(0, \frac{1}{2}\left(\frac{I-\theta_\mu^+}{\gamma_\mu^+}\right)^2\right) \ , \quad f_\mu(I) = \max\left(0, \frac{I-\theta_\mu^+}{\gamma_\mu^+}\right) \ . \tag{A4}$$

$\mathcal{TN}(\mu, \sigma^2, \mathcal{R}^+)$ denotes the truncated Gaussian distribution of mode μ, width σ and support $\mathcal{R}^+$. This potential is called ReLU because its transfer function is a ReLU function.

c. Binary hidden units

If the hidden units are spinlike variables, i.e. $h_\mu \in \{-1, 1\}$, the potential can be written as a field $\mathcal{U}_\mu(h_\mu) = -c_\mu\, h_\mu$. In that case, we should rescale $c_\mu \to c_\mu/N$, $w \to w\sqrt{N}$. We get

$$P\big(h_\mu|\mathbf{m}\big) = \frac{1}{2}\big(1 + h_\mu \tanh\left(N\left(I_\mu + c_\mu\right)\right)\big) \ , \tag{A5}$$

$$\hat{\Gamma}_\mu(I) = |I + c_\mu| \ , \quad f_\mu(I) = \mathrm{sign}\left(I + c_\mu\right) \ . \tag{A6}$$

If the hidden units are Bernoulli units, i.e., $h_\mu \in \{0, 1\}$, the potential acting on the hidden units is the same as for spins variables, and we get:

$$P\big(h_\mu|\mathbf{m}\big) = \frac{\exp\big(N h_\mu\big(I_\mu + c_\mu\big)\big)}{1 + \exp\big(N\big(I_\mu + c_\mu\big)\big)} \ , \tag{A7}$$

$$\hat{\Gamma}_\mu(I) = \max(0, I + c_\mu) \ , \quad f_\mu(I) = H\left(I + c_\mu\right) \ . \tag{A8}$$

$H(x)$ is the Heaviside step function.

Appendix B: Expansion of barrier height to first order in parameter changes

By using first order perturbation theory with the self-consistent equation defined in Eq. (48), we end up with:

$$\mathbf{m}^\alpha = \begin{bmatrix} g_\alpha(m_{11}, m_{22}) \\ g_\alpha(m_{22}, m_{11}) \end{bmatrix} \quad , \qquad \mathbf{m}^w = \begin{bmatrix} g_w(m_{11}) \\ g_w(m_{22}) \end{bmatrix} \quad , \tag{B1}$$

with

$$g_\alpha(x, y) = \left(-\frac{x}{2} + \frac{x+y}{1+xy}\right)\left(\frac{2w^2(1-x^2)}{2 - w^2(1-x^2)}\right) \ , \tag{B2}$$

$$g_w(x) = wx\left(\frac{2(1-x^2)}{2 - w^2(1-x^2)}\right) \ . \tag{B3}$$

Inserting these results in the expression of $f(\mathbf{m})$ (Eq. 47) leads to:

$$f^{\alpha}(\mathbf{m}) = -\frac{w^2}{2} m_{11} \left(\frac{m_{11}+m_{22}}{1+m_{11}m_{22}} + \frac{g_\alpha(m_{11},m_{22}) - m_{11}}{2} \right) - \frac{w^2}{2} m_{22} \left(\frac{m_{11}+m_{22}}{1+m_{11}m_{22}} + \frac{g_\alpha(m_{22},m_{11}) - m_{22}}{2} \right) + \frac{\mathcal{S}(m_{11})+\mathcal{S}(m_{22})}{2} - \mathcal{S}(m_{12}) + \frac{g_\alpha(m_{11},m_{22})}{2} \text{arctanh}(m_{11}) + \frac{g_\alpha(m_{22},m_{11})}{2} \text{arctanh}(m_{22}) \quad , \tag{B4}$$

$$f^{w}(\mathbf{m}) = -\frac{w^2}{2} \left(m_{11} \frac{g_w(m_{11})}{2} + m_{22} \frac{g_w(m_{22})}{2} \right) - \frac{w}{4} \left(m_{11}^2 + m_{22}^2 \right) + \frac{g_w(m_{11})}{2} \text{arctanh}(m_{11}) + \frac{g_w(m_{22})}{2} \text{arctanh}(m_{22}) \quad . \tag{B5}$$

Appendix C: Sampling in the hidden space

Numerically, $P(h_\mu|\mathbf{h}_{\neg\mu})$ (Algorithm 3) and $P(h_\mu, h_\nu|\mathbf{h}_{\neg\mu,\nu})$ (Algorithm 4) are discretized and the new candidate is drawn from the discretized distribution with the tower sampling algorithm [52]. Let us denote the acceptance probability from a configuration $\mathbf{h}$ to a configuration $\mathbf{h}'$ by $A_h(\mathbf{h} \to \mathbf{h}')$. The Metropolis-Hastings algorithm and Gibbs sampling satisfy detailed balance in $E^{\text{eff}}(\mathbf{h})$, hence

$$P(\mathbf{h}) A_h(\mathbf{h} \to \mathbf{h}') = P(\mathbf{h}') A_h(\mathbf{h}' \to \mathbf{h}) \ . \tag{C1}$$

For the dynamics defined in Fig. 1(d), we have the following acceptance probability from a configuration $\mathbf{v}$ to a configuration $\mathbf{v}'$

$$A_v(\mathbf{v} \to \mathbf{v}') = \int d\mathbf{h} d\mathbf{h}' P(\mathbf{h}|\mathbf{v}) A_h(\mathbf{h} \to \mathbf{h}') P(\mathbf{v}'|\mathbf{h}') \ . \tag{C2}$$

Therefore,

$$\begin{aligned} P(\mathbf{v}) A_v(\mathbf{v} \to \mathbf{v}') &= \int d\mathbf{h} d\mathbf{h}' P(\mathbf{v}) P(\mathbf{h}|\mathbf{v}) A_h(\mathbf{h} \to \mathbf{h}') P(\mathbf{v}'|\mathbf{h}') \\ &= \int d\mathbf{h} d\mathbf{h}' P(\mathbf{v}) \frac{P(\mathbf{v},\mathbf{h})}{P(\mathbf{v})} \frac{P(\mathbf{h}') A_h(\mathbf{h}' \to \mathbf{h})}{P(\mathbf{h})} \frac{P(\mathbf{v}',\mathbf{h}')}{P(\mathbf{h}')} \\ &= \int d\mathbf{h} d\mathbf{h}' P(\mathbf{v}|\mathbf{h}) A_h(\mathbf{h}' \to \mathbf{h}) P(\mathbf{v}',\mathbf{h}') \\ &= P(\mathbf{v}') A_v(\mathbf{v}' \to \mathbf{v}) \ . \end{aligned} \tag{C3}$$

As a consequence, our algorithm satisfies the detailed balance condition.

[1] N. Metropolis and S. Ulam, Journal of the American Statistical Association **44**, 335 (1949).
[2] W. K. Hastings, Biometrika **57**, 97 (1970).
[3] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **58**, 86 (1987).
[4] U. Wolff, Phys. Rev. Lett. **62**, 361 (1989).
[5] J.-S. Wang and R. H. Swendsen, Physica A: Statistical Mechanics and its Applications **167**, 565 (1990).
[6] A. Barbu and S.-C. Zhu, IEEE Trans Pattern Anal Mach Intell **27**, 1239 (2005).
[7] J. Liu, Y. Qi, Z. Y. Meng, and L. Fu, Phys. Rev. B **95**, 041101 (2017).
[8] X. Y. Xu, Y. Qi, J. Liu, L. Fu, and Z. Y. Meng, Phys. Rev. B **96**, 041119 (2017).
[9] L. Huang and L. Wang, Phys. Rev. B **95**, 035105 (2017).
[10] Y. Nagai, H. Shen, Y. Qi, J. Liu, and L. Fu, Phys. Rev. B **96**, 161102 (2017).
[11] H. Shen, J. Liu, and L. Fu, Phys. Rev. B **97**, 205140 (2018).
[12] Y. Nagai, M. Okumura, and A. Tanaka, Phys. Rev. B **101**, 115111 (2020).
[13] P. Smolensky, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations.* (MIT Press Cambridge, MA, 1986) pp. 194–281.
[14] G. E. Hinton, Neural Comput **14**, 1771 (2002).
[15] J. Tubiana and R. Monasson, Phys. Rev. Lett. **118**, 138301 (2017).
[16] E. Agliari, A. Barra, A. Galluzzi, F. Guerra, and F. Moauro, Phys. Rev. Lett. **109**, 268101 (2012).

[17] A. Decelle, G. Fissore, and C. Furtlehner, EPL **119**, 60001 (2017).

[18] A. Decelle, G. Fissore, and C. Furtlehner, J Stat Phys **172**, 1576 (2018).

[19] A. Barra, G. Genovese, P. Sollich, and D. Tantari, Phys. Rev. E **97**, 022310 (2018).

[20] G. S. Hartnett, E. Parker, and E. Geist, Phys. Rev. E **98**, 022116 (2018).

[21] A. Decelle and C. Furtlehner, Chinese Phys. B 10.1088/1674-1056/abd160 (2020).

[22] S. Geman and D. Geman, IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-6**, 721 (1984).

[23] G. E. Hinton, S. Osindero, and Y.-W. Teh, Neural Computation **18**, 1527 (2006).

[24] T. Tieleman, in *Proceedings of the 25th international conference on Machine learning - ICML '08* (ACM Press, Helsinki, Finland, 2008) pp. 1064–1071.

[25] N. Le Roux and Y. Bengio, Neural Computation **20**, 1631 (2008).

[26] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, 2003).

[27] Y. LeCun, http://yann. lecun. com/exdb/mnist/ (1998).

[28] E. Shakhnovich and A. Gutin, J. Chem. Phys. **93**, 5967 (1990).

[29] L. Mirny and E. Shakhnovich, Annu. Rev. Biophys. Biomol. Struct. **30**, 361 (2001).

[30] H. Jacquin, A. Gilson, E. Shakhnovich, S. Cocco, and R. Monasson, PLOS Computational Biology **12**, e1004889 (2016).

[31] S. Miyazawa and R. L. Jernigan, J Mol Biol **256**, 623 (1996).

[32] R. J. Glauber, Journal of Mathematical Physics **4**, 294 (1963).

[33] J. J. Hopfield, PNAS **79**, 2554 (1982).

[34] A. Barra, A. Bernacchia, E. Santucci, and P. Contucci, Neural Networks **34**, 1 (2012).

[35] F. E. Leonelli, E. Agliari, L. Albanese, and A. Barra, Neural Networks **143**, 314 (2021).

[36] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985).

[37] D. J. Amit, *Modeling Brain Function: The World of Attractor Neural Networks* (Cambridge University Press, Cambridge, 1989).

[38] C. Procesi and B. Tirozzi, Int. J. Mod. Phys. B **04**, 143 (1990).

[39] V. Nair and G. E. Hinton, in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10 (Omnipress, Madison, WI, USA, 2010) pp. 807–814.

[40] G. Desjardins, A. Courville, Y. Bengio, P. Vincent, and O. Delalleau, in *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (MIT Press Cambridge, MA, 2010) pp. 145–152.

[41] T. K. Ho, in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Vol. 1 (1995) pp. 278–282 vol.1.

[42] L. Breiman, Machine Learning **45**, 5 (2001).

[43] J. Tubiana, S. Cocco, and R. Monasson, eLife **8**, e39397 (2019).

[44] T. S. Ray, P. Tamayo, and W. Klein, Phys. Rev. A **39**, 5949 (1989).

[45] N. Persky, R. Ben-Av, I. Kanter, and E. Domany, Phys. Rev. E **54**, 2351 (1996).

[46] Y. Long, A. Nachmias, W. Ning, and Y. Peres, *A power law of order 1/4 for critical mean field Swendsen-Wang dynamics*, Memoirs of the American Mathematical Society, Vol. 232 (American Mathematical Society, 2014).

[47] A. Decelle, S. Hwang, J. Rocchi, and D. Tantari, arXiv:1906.11988 [cond-mat] (2019), arXiv: 1906.11988.

[48] S. Cocco, R. Monasson, L. Posani, S. Rosay, and J. Tubiana, Physica A: Statistical Mechanics and its Applications Lecture Notes of the 14th International Summer School on Fundamental Problems in Statistical Physics, **504**, 45 (2018).

[49] N. Yoshioka, Y. Akagi, and H. Katsura, Phys. Rev. E **99**, 032113 (2019).

[50] Y. Bengio, G. Mesnil, Y. Dauphin, and S. Rifai, in *International Conference on Machine Learning* (PMLR, 2013) pp. 552–560.

[51] G. Desjardins, H. Luo, A. Courville, and Y. Bengio, arXiv:1410.0123 [cs, stat] (2014).

[52] W. Krauth, *Statistical Mechanics: Algorithms and Computations* (Oxford Master Series in Physics, 2006).