Replica symmetry breaking and clustering phase transitions in undersampled restricted Boltzmann machines

Jorge Fernandez-De-Cossio-Diaz, Thomas Tulinski, Simona Cocco, Rémi Monasson

Replica symmetry breaking and clustering phase transitions in undersampled restricted Boltzmann machines

Jorge Fernandez-de-Cossio-Diaz, Thomas Tulinski, Simona Cocco, Rémi Monasson*
Laboratory of Physics of the Ecole Normale Supérieure, PSL & CNRS UMR8023, Sorbonne Université, Université Paris Cité, 24 rue Lhomond, Paris, France
(Dated: February 8, 2024)

Restricted Boltzmann machines (RBMs) are among the simplest unsupervised models implementing data/representation duality. The learning curves of RBMs trained on structured data are nevertheless difficult to characterize analytically, in part due to the presence of a partition function that depends on the trainable parameters. In this work, we present the exact solution of RBMs trained on structured data in the undersampled regime. The solution involves gradual symmetry breaking among the hidden units for decreasing regularization strength, as they specialize to finer-level details of the data. Hidden units form extensive blocks with identical weight parameters. Trained RBMs with different block sizes are separated by large barriers in the posterior distribution of the weights, which makes the optimal block size inaccessible during training with local gradient descent.

I. INTRODUCTION

Over the past decades, concepts and techniques from the statistical physics of disordered systems have proven extraordinarily successful in tackling complex problems outside the traditional realm of physics [1]. One striking illustration is provided by applications to machine learning. In supervised learning problems such as classification, the generalization properties of machine learning models could be precisely characterized depending on the network architecture, the optimization algorithm, and the data structure, see [2] for an early review. Statistical mechanics could also be applied to simple unsupervised learning setups, such as principal component analysis and its extensions [3].

Despite these successes, more sophisticated unsupervised architectures, aiming at modeling the distribution of data, have remained largely out of reach with statistical physics tools. Among them, probabilistic graphical models, which capture the pattern of dependencies between variables in high-dimensional data, are particularly hard to analyze. Informally speaking, the origin of the difficulties is the need to deal with the normalization of the model distribution; this partition function is computationally intractable and depends on a large number of model parameters, such as the interactions between variables to be optimized over learning.

In the present paper, we address this problem for a special case of graphical models, the so-called restricted Boltzmann machines (RBMs). RBMs are defined on a bipartite interaction graph, with one layer carrying data and the other extracting latent factors, and are one of the simplest architecture implementing the data/representation duality. Despite the simplicity of the architecture, RBM are competitive with deeper networks in many applications of interest, in particular in computational biology [4, 5].

* jorge.cossio@phys.ens.fr

We present an analytical derivation with the replica method of the learning curves of RBM in the case where the amount of data is much smaller than the size of the model. This undersampled regime is relevant in many contexts [6]. We explicit the meaning of the order parameters appearing in the replica analysis, and introduce a replica symmetry broken solution of the implicit equations they fulfill. The breaking of symmetry is then related to the nature of the representations of the data, in terms of multiple blocks of hidden units in the RBM. We show that, as the values of the RBM hyperparameters (aspect ratio, strength of regularization) are varied, a cascade of phase transitions arises, corresponding to different levels of coarse-graining of the data.

We now outline the content of the paper. Section II introduces RBMs in general. Section III gives the replica solution to the problem of an RBM with a visible layer of width N, a hidden layer of width M, learning from a dataset of size K, in the regime where $N, M \to \infty$ at fixed ratio α, with K arbitrary though finite. Section IV provides an interpretation of the replica solution through an independent albeit equivalent calculation. Section V then gives the analytical form of the weights learned by the RBM after training and describes the nature of the cascade of phase transitions as the regularization varies. Lastly, section VI defines a hierarchical structured dataset model, that we use as training data for the RBMs in numerical experiments.

Technical details are provided in the Appendices. Appendices A through E cover the replica calculation, together with the replica-symmetric ansatz and replica-symmetry-breaking ansatz for finite datasets, generic and hierarchically structured, alongside the derivation of the equivalence with the alternative calculation. Appendix H shows that these weights are actually a stable maximum of the posterior distribution of the weights, while Appendix I confirms that the global maxima weights are low-rank. Finally, Appendix J gives a linear programming formulation of the low-regularization phase used to bootstrap the numerical solutions of Eqs. (11)–(17).

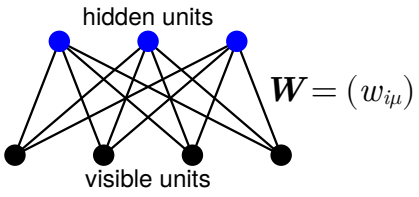

FIG. 1. Restricted Boltzmann machines (RBM). The RBM is defined by a set of N visible units (bottom, in black), and a set of M hidden units (top, in blue), connected by a set of weights $\mathbf{W} = (w_{i\mu})$, where $i \in \{1, \dots, N\}$ and $\mu \in \{1, \dots, M\}$. The units interact according to the energy function (1), which only allows interactions between pairs of visible and hidden units.

II. RESTRICTED BOLTZMANN MACHINES

A Restricted Boltzmann machines (RBM) is a simple two-layer network, composed of a layer of N visible units and a layer of M hidden units. See Figure 1. We will consider spin-valued units in both cases, and define the energy function as:

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i=1}^{N}\sum_{\mu=1}^{M} w_{i\mu} v_i h_\mu \tag{1}$$

where $\mathbf{v} = (v_1, \dots, v_N) \in \{\pm 1\}^N$ denotes a configuration of the visible layer, $\mathbf{h} = (h_1, \dots, h_M) \in \{\pm 1\}^M$ a configuration of the hidden layer, and $w_{i\mu}$ are the weights. The weights are the only parameters we consider in this work and must be trained on data. For simplicity, we will not consider biasing potentials on the units.

In turn, the RBM defines a probability distribution over configurations according to a Boltzmann law,

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

where

$$Z = \operatorname*{tr}_{\mathbf{v}\in\{\pm1\}^N} \operatorname*{tr}_{\mathbf{h}\in\{\pm1\}^M} \exp(-E(\mathbf{v}, \mathbf{h})) \tag{3}$$

is the partition function.

The RBM is trained by maximizing the likelihood of observed data. The data are usually modeled as particular configurations of the visible layer. The hidden units are treated as latent variables, and must be marginalized out to define the marginal distribution of the visible variables,

$$P_V(\mathbf{v}) = \operatorname*{tr}_{\mathbf{h}\in\{\pm1\}^M} P(\mathbf{v}, \mathbf{h}) \tag{4}$$

also called the likelihood.

A generic training dataset consists of K spin N-dimensional configurations,

$$\mathcal{D} = \{\boldsymbol{\xi}^1, \dots, \boldsymbol{\xi}^K\} \subset \{\pm 1\}^N \tag{5}$$

We then define a regularized log-likelihood, scaled by $1/K$ for convenience:

$$\mathcal{L}(\mathbf{W}) = \frac{1}{K}\sum_{k=1}^{K} \ln P_V(\boldsymbol{\xi}^k) - \frac{N\gamma}{2}\sum_{i\mu} w_{i\mu}^2 \tag{6}$$

that we view as a function of the RBM weight matrix, $\mathbf{W} = (w_{i\mu}) \in \mathbb{R}^{N\times M}$. A regularization term (with $\gamma \geq 0$) is usually introduced to avoid overfitting. In a Bayesian setting, this term can also be interpreted as arising from a prior distribution over the weights. Lastly, the trained weights of the RBM are defined by maximizing $\mathcal{L}(\mathbf{W})$,

$$\mathbf{W}^* = \operatorname*{argmax}_{\mathbf{W}} \mathcal{L}(\mathbf{W}) \tag{7}$$

Although finding the global maxima of $\mathcal{L}(\mathbf{W})$ might be desirable, we will be satisfied with local maxima. See Appendix I for some properties of global maxima. Although the value of $\mathcal{L}(\mathbf{W})$ is invariant under permutations of hidden units, the optimal $\mathbf{W}^*$ usually breaks this symmetry and a number of hidden units with specialized roles appear.

In what follows, we shall be concerned with an *undersampled regime*, where the width of the visible and hidden layers of the RBM, N, M grow without bound at a fixed aspect ratio $\alpha = M/N$, while the number of data points K remains finite.

III. REPLICA FORMALISM

Finding the *maximum a posteriori* (MAP) weights from (7) is a non-trivial task. In order to make progress, we will employ the replica trick from disordered systems [1].The first step consists of writing the Bayesian evidence of the data as a family of integrals parametrized by a non-negative parameter β,

$$\mathcal{Y}^{(\beta)}(\mathcal{D}) = \int [P_V(\mathcal{D}|\mathbf{W})]^\beta \, P_0(\mathbf{W}) d\mathbf{W}, \tag{8}$$

where $P_V(\mathcal{D}|\mathbf{W})$ is the likelihood (6) (note that we now make explicit the dependence on the weights $\mathbf{W}$), while $P_0(\mathbf{W})$ is a prior over the weights, that we take to be of the Gaussian form:

$$P_0(\mathbf{W}) = \prod_{i\mu} \sqrt{\frac{NK\beta\gamma}{2\pi}} \exp\left(-\frac{NK\beta\gamma}{2} w_{i\mu}^2\right) \tag{9}$$

Note that for $\beta = 1$, this choice is consistent with the L2 regularization introduced in Eq. (6). On the other hand, for large β, the integration in Eq. (8) concentrates around values of the weights $\mathbf{W}$ that maximize the log-posterior, *i.e.*, the solutions of (7). We can interpret β as an inverse temperature, controlling how much $\mathcal{Y}^{(\beta)}(\mathcal{D})$ is constrained to focus on the $\mathbf{W}^*$ point.

The main difficulty of the integration in (8) arises from the partition function, $Z(\mathbf{W})$, that appears as a denominator of the likelihood in the integrand. To handle this

difficult term, we can employ the replica trick [1]. If we consider n identical copies of the system, we can write:

$$\mathcal{Y}_n^{(\beta)}(\mathcal{D}) = \int d\mathbf{W}\, P_0(\mathbf{W}) \left[\operatorname*{tr}_{\{H_\mu^k\}} \exp\left(\sum_k \sum_{i\mu} w_{i\mu} \xi_i^k H_\mu^k \right) \right]^\beta [Z(\mathbf{W})]^{n\beta}, \quad (10)$$

The configuration spin variables H_μ^k appearing here, correspond to the hidden representations sampled conditioned on the datum $\boldsymbol{\xi}^k$. Then, as we take the limit $n \to -K$, we recover the original expression for the evidence (8). Following the replica formalism, we treat n, β as integers, and then only take their corresponding limits as the last step of the calculation.

To carry on with the replica calculation, it is necessary to introduce an *ansatz* for the form of the order parameters. In many systems, this has a replica symmetry breaking form, where the initially identical replicas introduced in (10) have non-equivalent overlaps with each other [7]. Since we expect that the trained RBM will develop energy minima tracking the data points, we propose a scheme where the $n\beta$ replicas split into a number Ω of subsets containing $n\beta/\Omega$ replicas each, such that replicas within the same subset are symmetric. After some calculations (see Appendix A for details), the following overlap order parameters are found:

$$Q_k^\omega = \frac{1}{N} \sum_i \xi_i^k \langle v_i \rangle^\omega \quad (11)$$

$$q_{\omega\omega'} = \frac{1}{N} \sum_i \langle v_i \rangle^\omega \langle v_i \rangle^{\omega'} \quad (12)$$

$$P_k^\omega = \frac{1}{M} \sum_\mu \langle h_\mu \rangle^\omega \langle H_\mu^k \rangle \quad (13)$$

$$p_{\omega\omega'} = \frac{1}{M} \sum_\mu \langle h_\mu \rangle^\omega \langle h_\mu \rangle^{\omega'} \quad (14)$$

Here, $\omega = 1, \dots, \Omega$ indexes the group of replicas, while $\langle v_i \rangle^\omega$ and $\langle h_\mu \rangle^\omega$ are the average spontaneous activities of the RBM units within a replica subset. The order parameters then have the following interpretations: Q_k^ω, P_k^ω measure the overlap between the RBM spontaneous activity and the data, while $q_{\omega\omega'}, p_{\omega\omega'}$ measure the overlap of the spontaneous activities of different replica subsets.

The overlaps (11)–(14) are found to satisfy the following set of self-consistent equations:

$$\langle v_i \rangle^\omega = \tanh\left(\frac{\alpha}{K\gamma} \sum_k P_k^\omega \xi_i^k - \frac{\alpha}{\Omega\gamma} \sum_{\omega'} p_{\omega\omega'} \langle v_i \rangle^{\omega'} \right) \quad (15)$$

$$\langle h_\mu \rangle^\omega = \tanh\left(\frac{1}{K\gamma} \sum_k \eta_k Q_k^\omega - \frac{1}{\Omega\gamma} \sum_{\omega'} q_{\omega\omega'} \langle h_\mu \rangle^{\omega'} \right) \quad (16)$$

$$\langle H^k \rangle_\mu = \tanh\left(\frac{1}{K\gamma} \sum_l x_{kl} \langle H_\mu^l \rangle - \frac{1}{\Omega\gamma} \sum_\omega Q_k^\omega \langle h_\mu \rangle^\omega \right) \quad (17)$$

where

$$x_{kl} = \frac{1}{N} \sum_i \xi_i^k \xi_i^l \quad (18)$$

is the overlap matrix of the data points.

This set of equations admits several solutions, that correspond to multiple values of the weights that give local maxima of the posterior (7). We can recover the weights corresponding to a particular solution via (see Appendix A):

$$N\gamma w_{i\mu}^\star = \frac{1}{K} \sum_{k\mu} \xi_i^k \langle H^k \rangle_\mu - \frac{1}{\Omega} \sum_{\mu\omega} \langle v_i \rangle^\omega \langle h_\mu \rangle^\omega. \quad (19)$$

In the following sections we present an alternative approach that allows us to justify the replica-symmetric ansatz we have used to derive these equations, and the nature of the alternative solutions.

IV. INTERPRETATION OF THE REPLICA SOLUTION

To enligthen the results of the replica calculation of the previous section, we follow here an alternative mean-field approach. First we establish that the weights (7) must be of low-rank. In the undersampled regime, there typically are extensive numbers of sites $i \in \{1, \dots, N\}$ in the data that are indiscernible. The optimal weights (7) do not break the indiscernibility among visible sites. In other words, if $\xi_i^k = \xi_j^k$ for all k, then $w_{i\mu} = w_{j\mu}$ (for all μ). The weights attached to a given hidden unit can then be given as functions of the set of values that a spin takes across the data points, $w_{i\mu} = w_\mu(\xi_i^1, \dots, \xi_i^K)$. Moreover, $w_\mu(\boldsymbol{\sigma}) = -w_\mu(-\boldsymbol{\sigma})$. Note that these properties hold even for finite N, M. They depend crucially on the convexity of the L2-regularization used in (6), and might not hold for other types of regularization. See Appendix I for a derivation. As a consequence, the trained weights of the RBM are of finite rank (at most 2^{K-1}).

In practice, the actual rank of $\mathbf{W}$ is typically $\leq K$. In fact, we show in Appendix H that the trained weights can be given as linear combinations of the data, and provide stable local maxima of the regularized likelihood.

Since the rank of the weight matrix remains finite as $N, M \to \infty$, the partition function (3) is dominated by a finite number of saddle-points, each of which is characterized by a set of average spontaneous activities of the visible and hidden units. These spontaneous activities can be found by solving the following self-consistent mean-field equations:

$$\begin{aligned} \langle v_i \rangle^\omega &= \tanh\left(\sum_\mu w_{i\mu} \langle h_\mu \rangle^\omega\right) \\ \langle h_\mu \rangle^\omega &= \tanh\left(\sum_i w_{i\mu} \langle v_i \rangle^\omega\right) \end{aligned} \tag{20}$$

Equations (20) can be solved by fixed-point iteration (usually it helps to include an inertial term). The index $\omega \in \{1, \dots, \Omega\}$ then traverses the different solutions found. Since the RBM is trained to fit the data, we can initialize $\langle v_i \rangle^\omega$ to one of the data points in turn. At most $\Omega \leq K$ distinct saddle-points are found in this manner.

Differentiaton of (6) then results in a set of stationarity conditions for the trained weights:

$$N\gamma w_{i\mu} = \frac{1}{K} \sum_k \xi_i^k \tanh\left(\sum_j w_{j\mu} \xi_j^k\right) - \frac{1}{\Omega} \sum_\omega \langle v_i \rangle^\omega \langle h_\mu \rangle^\omega \tag{21}$$

that must hold for all $i \in \{1, \dots, N\}$ and all $\mu \in \{1, \dots, M\}$. See Appendix F for detailed calculations. The trained weights are such that Eqs. (21) and (20) are both simultaneously satisfied. These equations are equivalent to Eqs. (19), (15)–(17) (see Appendix C).

This derivation confirms that the replica-symmetric subsets *ansatz* introduced earlier correspond to saddle-points in the free energy of the trained RBM.

V. PHASE DIAGRAM OF THE TRAINED RBM

We now study how the trained weights evolve as the regularization strength γ is varied. Several phase transitions are found to occur at certain critical values of γ, where $w_{i\mu}$ changes rank. These transitions are to be interpreted as clustering events, where similar data points become merged. The precise location and structure of the transitions depend on details of the data. In this section, some general remarks are stated, valid for any dataset. Then in Section VI we consider a specific data model in more detail.

Over-regularized regime

At sufficiently large values of γ, the trained weights vanish. This *over-regularized regime* occurs for $K\gamma \geq \lambda_{\max}$, where $\lambda_{\max}$ is the largest eigenvalue of the overlap matrix of the data, Eq. (18). This condition can be derived through a small weight expansion of $\mathcal{L}(\mathbf{W})$ from Eq. (6), and considering the stability of the zero weights to second-order.

Paramagnetic phase

As γ decreases below $\lambda_{\max}$, the weights become non-zero but initially remain small. The saddle-point equations (20) admit non-zero solutions only if the matrix

$$\sum_\mu w_{i\mu} w_{j\mu} \tag{22}$$

has at least one eigenvalue > 1. This condition can be derived by considering an expansion of Eqs. (20) in the small unit activities. Otherwise, the RBM is found in a paramagnetic phase ($\langle v_i \rangle = 0$, $\langle h_\mu \rangle = 0$), and Eq. (21) simplifies to:

$$N\gamma w_{i\mu} = \frac{1}{K} \sum_k \xi_i^k \tanh\left(\sum_i w_{j\mu} \xi_j^k\right) \tag{23}$$

Eq. (23) decouples for different μ, and therefore all hidden units adopt identical weights after training, $w_{i\mu} = w_i$. In other terms, the symmetry of hidden units remains unbroken in the paramagnetic phase [8]. When w_i is still near-zero, it will be proportional to the top eigenvector of the overlap matrix Eq. (18), but as it grows the non-linearity of Eq. (23) will usually cause w_i to deviate from this direction.

Eq. (23) can be solved for w_i by fixed-point iteration. We can then check *a posteriori* whether Eq. (22) develops an eigenvalue > 1, at which point the spontaneous activity of the RBM must be considered [9].

Weakly regularized regime

At the opposite extreme of sufficiently small γ, the left-hand side term in Eq. (21) becomes negligible. The saddle-points of Eqs. (20) tend to approximate the data points, $\langle v_i \rangle^k \approx \xi_i^k$ (with $\Omega = K$). As the weights become large, the tanh approaches a sign function, and we obtain the self-consistent equations:

$$\hat{\xi}_\mu^k = \operatorname{sign}\left(\sum_i w_{i\mu} \xi_i^k\right), \quad \xi_i^k = \operatorname{sign}\left(\sum_\mu w_{i\mu} \hat{\xi}_\mu^k\right) \tag{24}$$

where $\hat{\xi}_\mu^k = \operatorname{sign}(\langle h_\mu \rangle^k)$. Eqs. (24) resemble the Bidirectional Associative Memory (BAM) model, introduced by Kosko in 1988 [10, 11], and subsequently studied by other authors (also by statistical physics methods) [12–17]. The weights $w_{i\mu}$ can be interpreted as storing a mapping between pairs of visible and hidden patterns, $\boldsymbol{\xi}^k \leftrightarrow \hat{\boldsymbol{\xi}}^k$. Whereas in the original BAM [10, 11], the

patterns themselves are used to define the matrix $w_{i\mu}$ through a Hebbian-like rule, here $w_{i\mu}$ is trained by maximum likelihood according to Eq. (7).

Equations (24) can also be written as inequalities:

$$\sum_i w_{i\mu}\xi_i^k\hat{\xi}_\mu^k \geq 0, \quad \sum_\mu w_{i\mu}\xi_i^k\hat{\xi}_\mu^k \geq 0 \tag{25}$$

Given $\{\hat{\xi}_\mu^k\}$, the inequalities (25) define a convex region in weight space (the intersection of $(N+M)K$ half-spaces). Regions corresponding to different assignments of $\{\hat{\xi}_\mu^k\}$ have disjoint interiors (neighboring regions meet only in the hyperplane that separates them). Not all assignments of $\{\hat{\xi}_\mu^k\}$ are feasible, since the bidirectional linear separability conditions (25) might not be satisfiable.

As $\gamma \to 0$, there is one dominant stationary weight within a feasible region, given asymptotically by:

$$w_{i\mu} \propto \sum_k \xi_i^k\hat{\xi}_\mu^k \tag{26}$$

with a norm diverging like $1/\gamma$ as $\gamma \to 0$. Eq. (26) recovers Kosko's original Hebbian-like learning rule of the BAM. We also recover the alternative "weighted learning rules" of [18], a generalization of Kosko's original prescription [19], as subdominant modes that decay in comparison to Eq. (26) as $\gamma \to 0$. See Appendix G for more details.

In this regime the weights are of rank K. We can call this regime the BAM phase due to its similarities with Kosko's model [10, 11]. Note that solutions in this phase can be labeled by the feasible assignment of the signs $\{\hat{\xi}_\mu^k\}$.

The intermediate γ regime

An application of the implicit function theorem [20] to Eq. (7) shows that the weights are continuously differentiable functions of γ (except at the origin). See Appendix G. There are a number of different parametric curves $w_{i\mu}(\gamma)$, describing the different solutions of Eq. (7). These curves intersect only at the origin [21] (for $K\gamma \geq \lambda_{\max}$, as noted before), where they are all tangent to the top eigenvector of x_{kl} (as pointed out below Eq. (23)). As $\gamma \to 0$ the curves separate and diverge to infinity, and are eventually described by the BAM regime of the previous paragraphs. Each curve can be identified by the signs $\{\hat{\xi}_\mu^k\}$ in the feasible region (*c.f.* Eq. (25)) it eventually enters.

As γ decreases from $\lambda_{\max}/K$ to 0, the weights interpolate between the paramagnetic solution of Eq. (23), and eventually become parallel to one of the BAM directions Eqs. (26). A cascade of phase transitions occurs during this interpolation, since the weights must increase their rank from 1 in the paramagnetic phase to K when the asymptotic BAM phase is reached. These transitions are associated with spontaneous breaking of the permutation symmetry of hidden units and clusterings of the most similar data points.

In both the paramagnetic phase and the BAM phase, the trained weights are formed as linear combinations of the data points. This suggests that the entire solution curve (for all γ) lies in the span of the data. We confirm in Appendix H that this is the case.

Replica symmetry breaking

RSB, which is a breaking of the permutation symmetry of replica, turns out to be associated to a breaking of the permutation symmetry of the hidden units, a specialization phenomenon. Set of replica preserving such a permutation symmetry represent RBMs falling in the same free energy minimum, with different sets of replica falling in different free energy minima, all of which must be accounted for in the moment-matching condition (21). Only the paramagnetic phase and the rank-1 phase are replica symmetric. As the rank of the weight matrix increases, and the number Ω of saddle-points in (20) increases beyond 1, a Ω-RSB *ansatz* needs to be performed to obtain consistent equations. Details about this calculation are provided in Appendix A.

VI. HIERARCHICALLY STRUCTURED DATA

The actual solution of (7) is crucially dependent on the structure of the data.

As an example, we considered a hierarchical dataset, where the data points are arranged in nested clusters. See Figure 2. We start from an ancestral sequence $\in \{\pm 1\}^N$, which, for simplicity, can be taken as $(+1, \ldots, +1)$. We generate K_1 children sequences by flipping the sites of the ancestral sequence independently, with a probability p_1. In turn, the sites of each of these sequences are flipped independently with probability p_2, so that from each of the K_1 previous sequences, K_2 descendant sequences are generated. This process can be continued indefinitely, up to L hierarchical levels. The resulting tree (resembling a phylogenetic tree) has $K = K_1 \times \cdots \times K_L$ leaves, which constitute our training data (5). The entire procedure is fully specified by: i) the ancestral sequence that we take to be $(+1, \ldots, +1)$, ii) the numbers of descendants per level $K_1, \ldots, K_L$, and iii) the mutation probabilities $p_1, \ldots, p_L$.

If two data points k, l meet at level ℓ of the tree, their overlap is given by:

$$x_{kl} = \prod_{j=\ell+1}^{L} (1-2p_j)^2 = x_\ell \tag{27}$$

which depends only on ℓ. An example is shown in Figure 2 for two levels, with $K_1 = 2$, $K_2 = 3$, and $p_1 = 0.2$, $p_2 = 0.1$.

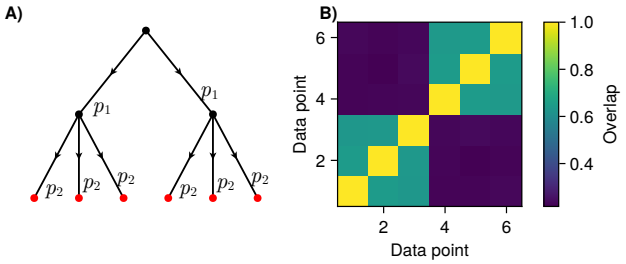

FIG. 2. Hierarchical data model. **A)** Data generation process across a phylogenetic tree. Root node corresponds to an ancestral spin sequence. Descendant nodes have a probability p_1 of introducing a "mutation" (*i.e.*, a spin flip) per site. In turn, descendants of these nodes introduce mutations with another probability p_2. The nodes at the lowest level (the leafs, shown in red) are the final data points. **B)** Heatmap of the resulting overlap matrix, as defined in (27), for $K_1 = 2$, $K_2 = 3$, $p_1 = 0.2$, $p_2 = 0.1$.

Numerical solution of the equations

To solve the equations numerically, we start from the BAM phase. We first find a feasible assignment of the signs $\{\hat{\xi}_\mu^k\}$, by a combination of exhaustive search and linear programming. See Appendix J for details. Then γ is increased slowly, and equations (20) and (21) are solved by iteration (with an inertial term). For each new value of γ, the previously found solution is used to hotstart the new fixed-point iteration.

Simulation results

We first checked that the steps described above to solve equations (7) yields results that are consistent with standard methods of training RBMs. To do this, we generated data according to the hierarchical model described in Section VI, considering an example tree with two levels, $K_1 = 2$, $K_2 = 3$, and with $p_1 = 0.2$, $p_2 = 0.1$. We then trained an RBM using Persistent Contrastive divergence [23]; see [24, 25] for implementation details. Here $\alpha = 0.9$ ($N = 1000$, $M = 9000$), and $\gamma = 0.01$. The resulting weight matrix from PCD has 5 non-zero singular values. Note that although for the hierarchical data we would expect degenerate eigenvalues for large N, at finite N finite sampling effects result in level splitting.

After training the RBM, we clustered the hidden unit weights using K-means [22], with different number of clusters. Panel A of Figure 3 shows the variance explained as a function of the number of clusters, and demonstrates that in this five hidden unit cluster are sufficient to fit the RBM weights, with all units within a block having approximately identical weights. We then solved equations (20) and (21) under the *ansatz* there are 5 kinds of hidden units, and initializing close to the k-means centers of the trained RBM weights. We then compare in Panel B of Figure 3 the theoretical prediction obtained, with the actual RBM weights from PCD. The agreement is excellent, which confirms the theoretical analysis of the previous section. Finally, Panel C shows a heatmap of the weights of the hidden units projected onto the data space, after suitable reordering of the hidden unit indices (using the K-mean clustering assignments). The panel clearly demonstrates the block structure of the trained weights.

We then inspected the behavior of the solution as γ changes. Figure 4A plots the overlaps between the saddle-points (20) of spontanous activity of the RBM and the data points, for a specific realization of the hierarchical data. Because of the hierarchical symmetries, the overlap matrix is defined by three free parameters, m_0, m_1, m_2, corresponding to the possible distances along the tree. In this example, two phase transitions occur, at the values of γ where these three parameters split. As $\gamma \to 0$, the weight matrix of the trained RBM approaches the BAM regime. We compare in Panels 4B-D the theoretical weights (26) to the actual weights observed after training the RBM for different values of γ. It is seen that as γ decreases, the agreement improves, suggesting that the solution (26) is approached eventually.

VII. DISCUSSION

In this work, we have studied the exact solution of the weights of a regularized RBM, trained by maximum likelihood on a set consisting of a finite number of data points. We first solved the problem by the replica method, introducing a particular replica-symmetry broken *ansatz*, where replicas are grouped into subsets that track different data points. Next, we presented an alternative solution, where we track how the weights interpolate between a paramagnetic and a weakly regularized regime as the regularization varies. In fact, both approaches lead to equivalent solutions, confirming our replica *ansatz*. The symmetric replica subsets are found to arise from the free-energy wells carved by the data points.

Although we have not been able to prove rigorously that the solutions studied here are the only possible solutions, they appear to be the most commonly found when numerically training the RBM by standard methods (CD, PCD). One explanation is that training of the RBM usually begins with small weights, where the weights are initially attracted to the top eigenvector of the data (like in (22)), and after that are likely to follow a path toward some of the solutions studied here. In addition, failing to produce saddle-points near the data (which is the main assumption we have made here) would necessarily lead to a lower likelihood.

Our calculation can be extended in several directions. It will be interesting to consider the consequences of biases in the units, or more generic potentials (e.g., ReLU [26], dReLU [4]). Importantly, the number of data K has

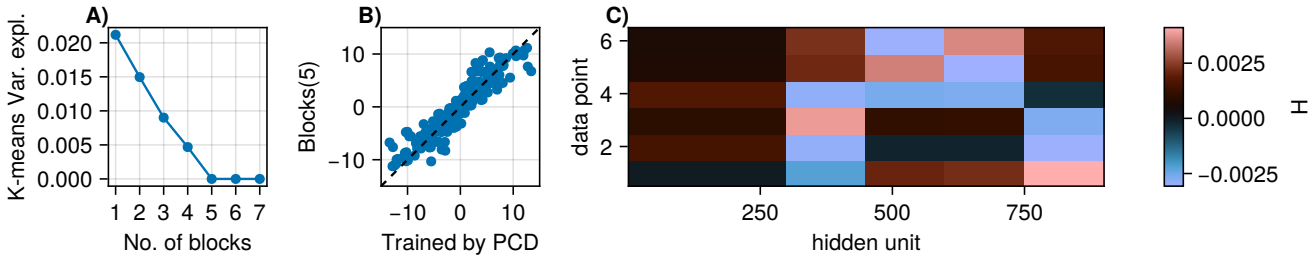

FIG. 3. Comparison of theory vs. PCD training of RBMs. **A)** We clusterize the weights attached to different hidden units using K-means [22], for different number of clusters, and compute the variance explained by the clustering. In this example, there are five blocks of identical units. **B)** Solving the mean-field equations in the manner described in the text, yields predictions consistent with the weights found by training the RBM with standard methods (Persistent Contrastive Divergence [23]). **C)** Projection of the trained RBM weights onto the vector space of the data points. Rows correspond to the data points, and columns to the hidden units (after reordering).

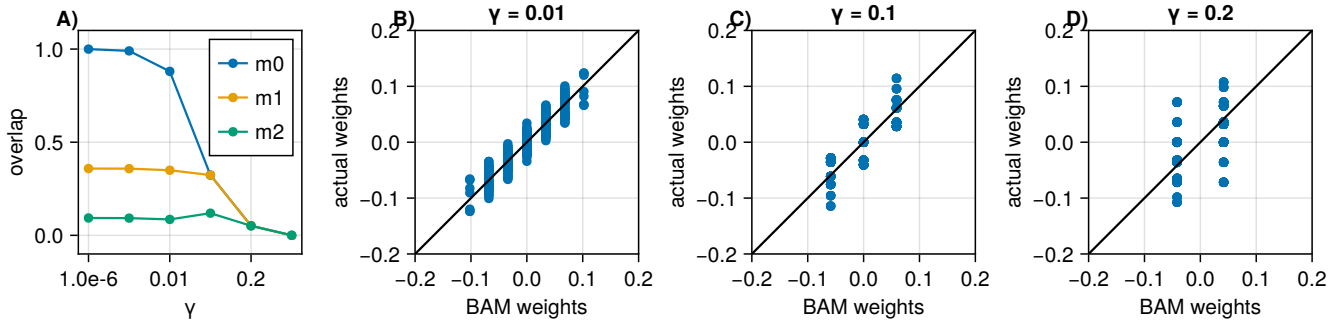

FIG. 4. Clustering phase transitions and asymptotic BAM regime. **A)** Overlaps between the spontaneous RBM activity saddle-points (solutions of (20)) and the data points, as a function of γ for the Hierarchical data model. Parameters of the simulations: $K_1 = 2, K_2 = 3$, $p_1 = 0.25, p_2 = 0.2$, $N = 100000$, $\alpha = 0.8$. For hierarchical data the matrix of overlaps also has a hierarchical structure, with only three free parameters m_0 (overlap of a saddle-point and its closest data point), and m_1, m_2 (overlaps of a saddle-point and second and third nearest data points). There are two phase transitions indicated by the values of γ where the three overlaps split. **B-D)** As $\gamma \to 0$, the weights approach the asymptotic form (26).

been kept finite here, while moving to the regime of K comparable to N, M, seems to pose more fundamental difficulties. The replica method is likely to help, but as we have shown here even in the simple finite K case, replica symmetry breaking occurs at multiple levels, suggesting a full-RSB phase as $K \to \infty$. We leave this question to future work.

ACKNOWLEDGMENTS

J.FdCD. acknowledges support from Université PSL, through the AI Junior Fellow program. T.T. acknowledges funding from Memnet ANR-22-CE16-0005.

[1] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, Vol. 9 (World Scientific Publishing Company, 1987).
[2] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, 2001).
[3] P. Reimann, C. V. den Broeck, and G. J. Bex, A gaussian scenario for unsupervised learning, Journal of Physics A: Mathematical and General **29**, 3521 (1996).
[4] J. Tubiana, S. Cocco, and R. Monasson, Learning protein constitutive motifs from sequence data, Elife **8**, e39397 (2019).
[5] B. Bravi, A. Di Gioacchino, J. Fernandez-de Cossio-Diaz, A. M. Walczak, T. Mora, S. Cocco, and R. Monasson, A transfer-learning approach to predict antigen immunogenicity and t-cell receptor specificity, ELife **12**, e85126 (2023).
[6] C. W. Lynn, Q. Yu, R. Pang, S. E. Palmer, and

W. Bialek, Exact minimax entropy models of large-scale neuronal activity, arXiv preprint arXiv:2402.00007 (2023).
[7] G. Parisi, Infinite number of order parameters for spin-glasses, Physical Review Letters **43**, 1754 (1979).
[8] If Eqs. (23) admit several solutions, the one maximizing $\mathcal{L}(\mathbf{W})$ is preferred. In degenerate situations, multiple solutions might have the same value of $\mathcal{L}(\mathbf{W})$, but we do not consider this scenario here.
[9] Note that Eq. (22) simplifies to $M\sum_i w_i w_j$, which has the eigenvalue $M\sum_i w_i^2$.
[10] B. Kosko, Bidirectional associative memories, IEEE Transactions on Systems, man, and Cybernetics **18**, 49 (1988).
[11] B. Kosko, Adaptive bidirectional associative memories, Appl. Opt. **26**, 4947 (1987).
[12] J. Kurchan, L. Peliti, and M. Saber, A statistical investigation of bidirectional associative memories (BAM), Journal de Physique I **4**, 1627 (1994).
[13] A. Barra, G. Catania, A. Decelle, and B. Seoane, Thermodynamics of bidirectional associative memories, Journal of Physics A: Mathematical and Theoretical **56**, 205005 (2023).
[14] M. S. Centonze, I. Kanter, and A. Barra, Statistical mechanics of learning via reverberation in bidirectional associative memories, Physica A: Statistical Mechanics and its Applications , 129512 (2024).
[15] B. Lenze, Improving Leung's bidirectional learning rule for associative memories, IEEE Transactions on Neural Networks **12**, 1222 (2001).
[16] C. Leung and K. Cheung, Householder encoding for discrete bidirectional associative memory, in *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks* (IEEE, 1991) pp. 237–241.
[17] Z.-O. Wang, A bidirectional associative memory based on optimal linear associative memory, IEEE transactions on computers **45**, 1171 (1996).
[18] T. Wang, X. Zhuang, and X. Xing, Weighted learning of bidirectional associative memories by global minimization, IEEE transactions on neural networks **3**, 1010 (1992).
[19] Y.-F. Wang, J. B. Cruz, and J. Mulligan, Guaranteed recall of all training pairs for bidirectional associative memory, IEEE transactions on Neural Networks **2**, 559 (1991).
[20] R. Courant, F. John, A. A. Blank, and A. Solomon, *Introduction to calculus and analysis*, Vol. 1 (Springer, 1965).
[21] V. I. Arnold, *Ordinary differential equations* (Springer Science & Business Media, 1992).
[22] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer-Verlag, Berlin, Heidelberg, 2006).
[23] T. Tieleman, Training restricted boltzmann machines using approximations to the likelihood gradient, in *Proceedings of the 25th international conference on Machine learning* (2008) pp. 1064–1071.
[24] J. Fernandez-de Cossio-Diaz, S. Cocco, and R. Monasson, Disentangling representations in restricted boltzmann machines without adversaries, Phys. Rev. X **13**, 021003 (2023).
[25] C. Roussel, J. Fernandez-de-Cossio-Diaz, S. Cocco, and R. Monasson, Accelerated sampling with stacked restricted boltzmann machines, in *International Conference on Learning Representations* (2024).
[26] V. Nair and G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in *Proceedings of the 27th international conference on machine learning (ICML-10)* (2010) pp. 807–814.
[27] G. S. Hartnett, E. Parker, and E. Geist, Replica symmetry breaking in bipartite spin glasses and neural networks, Phys. Rev. E **98**, 022116 (2018).
[28] S. P. Boyd and L. Vandenberghe, *Convex optimization* (Cambridge university press, 2004).
[29] F. Harary, The automorphism group of a hypercube., J. Univers. Comput. Sci. **6**, 136 (2000).
[30] R. J. Vanderbei *et al.*, *Linear programming* (Springer, 2020).

Appendix A: Replica calculation

We are interested in the *maximum a posteriori* (MAP) weights, i.e. the weights $\mathbf{W}^\star$ maximizing the posterior

$$P(\mathbf{W}|\mathcal{D}) = \frac{1}{\mathcal{Y}(\mathcal{D})} P(\mathcal{D}|\mathbf{W}) P_0(\mathbf{W}). \tag{A1}$$

In the above,

$$P_V(\mathcal{D}|\mathbf{W}) = \prod_{k=1}^{K} \frac{1}{Z(\mathbf{W})} \operatorname*{tr}_{\{H_\mu^k\}} \exp\left(\sum_{i=1}^{N} \sum_{\mu=1}^{M} w_{i\mu} \xi_i^k H_\mu^k \right) \tag{A2}$$

is the likelihood. The prior

$$P_0(\mathbf{W}) = \prod_{i\mu} \sqrt{\frac{NK\beta\gamma}{2\pi}} \exp\left(-\frac{NK\beta\gamma}{2} w_{i\mu}^2 \right) \tag{A3}$$

assumes that the weights are i.i.d. Gaussian variables of mean 0 and variance $1/NK\beta\gamma$, with β a non-negative parameter. The normalization

$$\mathcal{Y}(\mathcal{D}) = \int P_V(\mathcal{D}|\mathbf{W}) P_0(\mathbf{W}) d\mathbf{W} \tag{A4}$$

is the evidence. We consider the integral

$$\mathcal{Y}^{(\beta)}(\mathcal{D}) = \int \left[P_V(\mathcal{D}|\mathbf{W}) \right]^\beta P_0(\mathbf{W}) d\mathbf{W}, \tag{A5}$$

since for $\beta \to \infty$ it concentrates around the MAP weights,

$$\frac{1}{\beta} \ln \mathcal{Y}^{(\beta)}(\mathcal{D}) \sim \ln P_V(\mathcal{D}|\mathbf{W}^\star) - \frac{NK\gamma}{2} \sum_{i\mu} (w_{i\mu}^\star)^2, \tag{A6}$$

where we neglected irrelevant additive constant terms. This is true when the $\beta \to \infty$ limit is taken at N finite. In the replica calculation below we take $N \to \infty$ first, and thus assume that the two limits commute. We can write

$$\mathcal{Y}^{(\beta)}(\mathcal{D}) = \lim_{n \to -K} \mathcal{Y}_n^{(\beta)}(\mathcal{D}) \tag{A7}$$

where

$$\mathcal{Y}_n^{(\beta)}(\mathcal{D}) = \int \left[\operatorname*{tr}_{\{H_\mu^k\}} \exp\left(\sum_k \sum_{i\mu} w_{i\mu} \xi_i^k H_\mu^k \right) \right]^\beta [Z(\mathbf{W})]^{n\beta} P_0(\mathbf{W}) d\mathbf{W}, \tag{A8}$$

and assume that β, n are positive integers. This is the replica trick, leading to the replicated evidence,

$$\mathcal{Y}_n^{(\beta)}(\mathcal{D}) = \int \operatorname*{tr}_{\{H_\mu^{k\kappa}\}} \exp\left(\sum_{k=1}^{K} \sum_{\kappa=1}^{\beta} \sum_{i\mu} w_{i\mu} \xi_i^k H_\mu^{k\kappa} \right) \operatorname*{tr}_{\{v_i^{a\kappa}\},\{h_\mu^{a\kappa}\}} \exp\left(\sum_{a=1}^{n} \sum_{\kappa=1}^{\beta} \sum_{i\mu} w_{i\mu} v_i^{a\kappa} h_\mu^{a\kappa} \right) \mathrm{d}\mathbf{W}, \tag{A9}$$

where we have introduced $n\beta$ equilibrium configurations of the RBM spontaneous visible and hidden activities, $v_i^{a\kappa}$ and $h_\mu^{a\kappa}$, $a = 1, \ldots, n$, $\kappa = 1, \ldots, \beta$, along with $K\beta$ equilibrium configurations of the RBM hidden representations of the data, $H^{k\kappa}$, $k = 1, \ldots, K$, $\kappa = 1, \ldots, \beta$. In both cases these configurations, so-called *replica*, are assumed to be sampled independently from the model at fixed weights $\mathbf{W}$. Performing the Gaussian integrals introduces two-body couplings between replica,

$$\mathcal{Y}_n^{(\beta)}(\mathcal{D}) = \operatorname*{tr}_{\{H_\mu^{k\tau}\},\{v_i^{a\tau}\},\{h_\mu^{a\tau}\}} \exp\left[\frac{1}{NK\beta\gamma} \sum_{i\mu} \left(\sum_{(k\kappa)<(l\lambda)} \xi_i^k \xi_i^l H_\mu^{k\kappa} H_\mu^{l\lambda} + \sum_{a\kappa\lambda,k} v_i^{a\kappa} \xi_i^k h_\mu^{a\kappa} H_\mu^{k\lambda} + \sum_{(a\kappa)<(b\lambda)} v_i^{a\kappa} v_i^{b\lambda} h_\mu^{a\kappa} h_\mu^{b\lambda} + \frac{\beta}{2}(K+n) \right) \right. \tag{A10}$$

where $(k\kappa) < (l\lambda)$ and $(a\kappa) < (b\lambda)$ are lexicographic orders,

$$\sum_{(k\kappa)<(l\lambda)} V_i^{k\kappa} V_i^{l\lambda} H_\mu^{k\kappa} H_\mu^{l\lambda} = \frac{1}{2}\left(\sum_{k\kappa l\lambda} V_i^{k\kappa} V_i^{l\lambda} H_\mu^{k\kappa} H_\mu^{l\lambda} - K\beta\right),$$

$$\sum_{(a\kappa)<(b\lambda)} v_i^{a\kappa} v_i^{b\lambda} h_\mu^{a\kappa} h_\mu^{b\lambda} = \frac{1}{2}\left(\sum_{a\kappa b\lambda} v_i^{a\kappa} v_i^{b\lambda} h_\mu^{a\kappa} h_\mu^{b\lambda} - n\beta\right).$$

We identify the order parameters

$$Q_k^{a\kappa} = \frac{1}{N}\sum_i v_i^{a\kappa}\xi_i^k, \qquad q_{b\lambda}^{a\kappa} = \frac{1}{N}\sum_i v_i^{a\kappa} v_i^{b\lambda} \tag{A11}$$

$$P_{k\lambda}^{a\kappa} = \frac{1}{M}\sum_\mu h_\mu^{a\kappa} H_\mu^{k\lambda}, \qquad p_{b\lambda}^{a\kappa} = \frac{1}{M}\sum_\mu h_\mu^{a\kappa} h_\mu^{b\lambda}, \tag{A12}$$

$$y_{l\lambda}^{k\kappa} = \frac{1}{M}\sum_\mu H_\mu^{k\kappa} H_\mu^{l\lambda}. \tag{A13}$$

which are overlap tensors. Denoting $x_{kl} = \frac{1}{N}\sum_i \xi_i^k \xi_i^l$ the overlap matrix of the data we have, up to irrelevant constant factors,

$$\mathcal{Y}_n^{(\beta)}(\mathcal{D}) = \int \mathcal{N}_v(\mathcal{Q}_v)\mathcal{N}_h(\mathcal{Q}_h) d\mathcal{Q}_v d\mathcal{Q}_h \exp\left[\frac{M}{K\beta\gamma}\left(\sum_{(k\kappa)<(l\lambda)} x_{kl} y_{l\lambda}^{k\kappa} + \sum_{a\kappa\lambda,k} Q_k^{a\kappa} P_{k\lambda}^{a\kappa} + \sum_{(a\kappa)<(b\lambda)} q_{b\lambda}^{a\kappa} p_{b\lambda}^{a\kappa}\right)\right], \tag{A14}$$

where $\mathcal{Q}_v = \{Q_k^{a\kappa}, q_{b\lambda}^{a\kappa}\}$ and $\mathcal{Q}_h = \{y_{l\lambda}^{k\kappa}, P_{k\lambda}^{a\kappa}, p_{b\lambda}^{a\kappa}\}$ including off-diagonal entries only once, and

$$\mathcal{N}_v(\mathcal{Q}_v) = \operatorname*{tr}_{\{v_i^{a\kappa}\}}\left[\prod_{a\kappa,k}\delta\left(Q_k^{a\kappa} - \frac{1}{N}\sum_i v_i^{a\kappa}\xi_i^k\right)\right]\left[\prod_{(a\kappa)<(b\lambda)}\delta\left(q_{b\lambda}^{a\kappa} - \frac{1}{N}\sum_i v_i^{a\kappa} v_i^{b\lambda}\right)\right], \tag{A15}$$

$$\begin{aligned}\mathcal{N}_h(\mathcal{Q}_h) = \operatorname*{tr}_{\{H_\mu^k\},\{h_\mu^{a\kappa}\}}&\left[\prod_{(k\kappa)<(l\lambda)}\delta\left(y_{l\lambda}^{k\kappa} - \frac{1}{M}\sum_\mu H_\mu^{k\kappa} H_\mu^{l\lambda}\right)\right]\left[\prod_{a\kappa\lambda,k}\delta\left(P_{k\lambda}^{a\kappa} - \frac{1}{M}\sum_\mu h_\mu^{a\kappa} H_\mu^{k\lambda}\right)\right] \\ &\times\left[\prod_{(a\kappa)<(b\lambda)}\delta\left(p_{b\lambda}^{a\kappa} - \frac{1}{M}\sum_\mu h_\mu^{a\kappa} h_\mu^{b\lambda}\right)\right],\end{aligned} \tag{A16}$$

are entropic factors. Following standard field-theoretic manipulations,

$$\mathcal{N}_v(\mathcal{Q}_v) = \int e^{N\mathcal{S}_v(\mathcal{Q}_v,\hat{\mathcal{Q}}_v)} d\hat{\mathcal{Q}}, \tag{A17}$$

$$\mathcal{N}_h(\mathcal{Q}_h) = \int e^{M\mathcal{S}_h(\mathcal{Q}_h,\hat{\mathcal{Q}}_h)} d\hat{\mathcal{Q}}_h, \tag{A18}$$

where we introduced the entropies

$$\mathcal{S}_v(\mathcal{Q}_v,\hat{\mathcal{Q}}_v) = -\sum_{a\kappa,k} Q_k^{a\kappa}\hat{Q}_k^{a\kappa} - \sum_{(a\kappa)<(b\lambda)} q_{b\lambda}^{a\kappa}\hat{q}_{b\lambda}^{a\kappa} + \frac{1}{N}\sum_i \ln z_i^v, \tag{A19}$$

$$\mathcal{S}_h(\mathcal{Q}_h,\hat{\mathcal{Q}}_h) = -\sum_{(k\kappa)<(l\lambda)} y_{l\lambda}^{k\kappa}\hat{y}_{l\lambda}^{k\kappa} - \sum_{a\kappa\lambda,k} P_{k\lambda}^{a\kappa}\hat{P}_{k\lambda}^{a\kappa} - \sum_{(a\kappa)<(b\lambda)} p_{b\lambda}^{a\kappa}\hat{p}_{b\lambda}^{a\kappa} + \ln z^h, \tag{A20}$$

with

$$z_i^v = \operatorname*{tr}_{\{v^{a\kappa}\}}\exp\left(\sum_{a\kappa,k}\hat{Q}_k^{a\kappa}\xi_i^k v^{a\kappa} + \sum_{(a\kappa)<(b\lambda)}\hat{q}_{b\lambda}^{a\kappa} v^{a\kappa} v^{b\lambda}\right), \tag{A21}$$

$$z^h = \operatorname*{tr}_{\{H^{k\kappa}\},\{h^{a\kappa}\}} \exp\left(\sum_{(k\kappa)<(l\lambda)} \hat{y}_{l\lambda}^{k\kappa} H^{k\kappa} H^{l\lambda} + \sum_{a\kappa\lambda,k} \hat{P}_{k\lambda}^{a\kappa} h^{a\kappa} H^{k\lambda} + \sum_{(a\kappa)<(b\lambda)} \hat{p}_{b\lambda}^{a\kappa} h^{a\kappa} h^{b\lambda} \right). \tag{A22}$$

Note that z_i^v depends on i through the data points. For large N, we get the saddle-point equations extremizing $\mathcal{S}_v$, $\mathcal{S}_h$ in $\hat{\mathcal{Q}}_v$, $\hat{\mathcal{Q}}_h$,

$$Q_k^{a\kappa} = \frac{1}{N} \sum_i \xi_i^k \langle v^{a\kappa} \rangle_i, \qquad q_{b\lambda}^{a\kappa} = \frac{1}{N} \sum_i \langle v^{a\kappa} v^{b\lambda} \rangle_i, \tag{A23}$$

$$P_{k\lambda}^{a\kappa} = \langle h^{a\kappa} H^{k\lambda} \rangle, \qquad p_{b\lambda}^{a\kappa} = \langle h^{a\kappa} h^{b\lambda} \rangle, \tag{A24}$$

$$\tag{A25}$$

$$y_{l\lambda}^{k\kappa} = \langle H^{k\kappa} H^{l\lambda} \rangle. \tag{A26}$$

where $\langle \cdot \rangle_i$ and $\langle \cdot \rangle$ denote an average under the density defined by the partition function z_i^v and z^h respectively. Ignoring irrelevant constants, we can write

$$\mathcal{Y}_n^{(\beta)}(\mathcal{D}) = \int \exp(N\Phi(\mathcal{Q}_v, \mathcal{Q}_h, \hat{\mathcal{Q}}_v, \hat{\mathcal{Q}}_h)) d\mathcal{Q}_v d\mathcal{Q}_h d\hat{\mathcal{Q}}_v d\hat{\mathcal{Q}}_h \tag{A27}$$

where

$$\Phi = \frac{\alpha}{K\beta\gamma} \left(\sum_{(k\kappa)<(l\lambda)} x_{kl} y_{l\lambda}^{k\kappa} + \sum_{a\kappa\lambda,k} Q_k^{a\kappa} P_{k\lambda}^{a\kappa} + \sum_{(a\kappa)<(b\lambda)} q_{b\lambda}^{a\kappa} p_{b\lambda}^{a\kappa} \right) + \mathcal{S}_v(\mathcal{Q}_v, \hat{\mathcal{Q}}_v) + \alpha \mathcal{S}_h(\mathcal{Q}_h, \hat{\mathcal{Q}}_h). \tag{A28}$$

For large N, we get saddle-point equations extremizing Φ in $\mathcal{Q}_v, \mathcal{Q}_h$,

$$\hat{Q}_k^{a\kappa} = \frac{\alpha}{K\beta\gamma} \sum_\lambda P_{k\lambda}^{a\kappa}, \qquad \hat{q}_{b\lambda}^{a\kappa} = \frac{\alpha}{K\beta\gamma} p_{b\lambda}^{a\kappa}, \tag{A29}$$

$$\hat{P}_{k\lambda}^{a\kappa} = \frac{1}{K\beta\gamma} Q_k^{a\kappa}, \qquad \hat{p}_{b\lambda}^{a\kappa} = \frac{1}{K\beta\gamma} q_{b\lambda}^{a\kappa}, \tag{A30}$$

$$\hat{y}_{l\lambda}^{k\kappa} = \frac{1}{K\beta\gamma} x_{kl}. \tag{A31}$$

Looking at z^h and the saddle-point equations for $\hat{y}_{l\lambda}^{k\kappa}$, we see that, by symmetry, the moments $\langle H^{k\kappa} H^{l\lambda} \rangle$ for $(k\kappa) < (l\lambda)$ cannot depend on κ, λ, hence $y_{l\lambda}^{k\kappa} = y^{kl}$. Similarly $\langle h^{a\kappa} H^{k\lambda} \rangle$ for all $a\kappa\lambda$, k, must be independent of λ, hence $P_{k\lambda}^{a\kappa} = P_k^{a\kappa}$. Eliminating $\hat{\mathcal{Q}}_v$, $\hat{\mathcal{Q}}_h$, in favor of $\mathcal{Q}_v$, $\mathcal{Q}_h$, we find that, at a saddle-point,

$$\Phi = -\frac{\alpha}{K\gamma} \left(\sum_{a\kappa,k} Q_k^{a\kappa} P_k^{a\kappa} + \frac{1}{\beta} \sum_{(a\kappa)<(b\lambda)} q_{b\lambda}^{a\kappa} p_{b\lambda}^{a\kappa} \right) + \ln z_i^v + \alpha \ln z^h, \tag{A32}$$

with

$$z_i^v = \operatorname*{tr}_{\{v^{a\kappa}\}} \exp\left[\frac{\alpha}{K\gamma} \left(\sum_{a\kappa,k} P_k^{a\kappa} \xi_i^k v^{a\kappa} + \frac{1}{\beta} \sum_{(a\kappa)<(b\lambda)} p_{b\lambda}^{a\kappa} v^{a\kappa} v^{b\lambda} \right) \right], \tag{A33}$$

$$z^h = \operatorname*{tr}_{\{H^{k\kappa}\},\{h^{a\kappa}\}} \exp\left[\frac{1}{K\beta\gamma} \left(\sum_{(k\kappa)<(l\lambda)} x_{kl} H^{k\kappa} H^{l\lambda} + \sum_{a\kappa\lambda,k} Q_k^{a\kappa} h^{a\kappa} H^{k\lambda} + \sum_{(a\kappa)<(b\lambda)} q_{b\lambda}^{a\kappa} h^{a\kappa} h^{b\lambda} \right) \right]. \tag{A34}$$

Appendix B: Replica-symmetry-breaking ansatz

We now make an RSB ansatz, assuming replica symmetry only within a finite number Ω of sets of $n\beta/\Omega$ replica each, with broken replica symmetry across these sets. For $(a\kappa) = 1, \ldots, n\beta$, and $(a\kappa) < (b\lambda)$,

$$Q_k^{a\kappa} = Q_k^\omega \qquad P_k^{a\kappa} = P_k^\omega, \quad (a\kappa) \in \omega, \tag{B1}$$

$$q_{b\lambda}^{a\kappa} = q_{\omega\omega'} \qquad p_{b\lambda}^{a\kappa} = p_{\omega\omega'} \quad (a\kappa) \in \omega, (b\lambda) \in \omega', \tag{B2}$$

where $\omega = 1, \ldots, \Omega$ and $\omega < \omega'$. We get

$$z_i^v = \underset{\{v^{a\kappa}\}}{\mathrm{tr}} \exp\left[\frac{\alpha}{K\gamma}\left(\sum_{\omega,k} P_k^\omega \xi_i^k \sum_{(a\kappa)\in\omega} v^{a\kappa} + \frac{1}{2\beta}\sum_{\omega<\omega'} p_{\omega\omega'} \sum_{(a\kappa)\in\omega} v^{a\kappa} \sum_{(b\lambda)\in\omega'} v^{b\lambda} - \frac{n}{2}\right)\right] \tag{B3}$$

$$z^h = \underset{\{H^{k\kappa}\},\{h^{a\kappa}\}}{\mathrm{tr}} \exp\left[\frac{1}{K\beta\gamma}\left(\frac{1}{2}\sum_{k\kappa l\lambda} x_{kl} H^{k\kappa} H^{l\lambda} + \sum_\omega \sum_{k\lambda} Q_k^\omega H^{k\lambda} \sum_{(a\kappa)\in\omega} h^{a\kappa} + \frac{1}{2}\sum_{\omega<\omega'} q_{\omega\omega'} \sum_{(a\kappa)\in\omega} h^{a\kappa} \sum_{(b\lambda)\in\omega'} h^{b\lambda} - \frac{\beta}{2}(K+n)\right)\right] \tag{B4}$$

Making a few Hubbard-Stratonovich transforms to decouple the unary (multivariate Gaussian identity) and binary quadratic forms ([12, 13, 27]),

$$z_i^v = \underset{\{v^{a\kappa}\}}{\mathrm{tr}} \int \exp\left\{\frac{\alpha}{\gamma}\left[\sum_\omega \left(\frac{1}{K}\sum_k P_k^\omega \xi_i^k - \zeta_\omega\right) \sum_{(a\kappa)\in\omega} v^{a\kappa} - \frac{K\beta}{2}\sum_{\omega<\omega'} (p^{-1})_{\omega\chi} \zeta_\omega^v \zeta_\chi^v\right]\right\} \mathrm{d}\boldsymbol{\zeta} \tag{B5}$$

$$\begin{aligned} z^h = \underset{\{H^{k\tau}\},\{h^{a\tau}\}}{\mathrm{tr}} \int \exp\Bigg\{\frac{1}{\gamma}\Bigg[\beta\sum_k \chi_k\eta_k - \frac{K\beta}{2}\sum_{kl}(x^{-1})_{kl}\varsigma_k\varsigma_l - \frac{K\beta}{2\gamma}\sum_{\omega<\omega'}(q^{-1})_{\omega\omega'}\vartheta_\omega\vartheta_{\omega'} + \sum_{k\tau}(\varsigma_k - \chi_k)H^{k\kappa} \\ + \left(\frac{1}{K}\sum_{k\omega}\eta_k Q_k^\omega - \sum_\omega \vartheta_\omega\right)\sum_{(a\kappa)\in\omega} h^{a\kappa}\Bigg]\Bigg\} \mathrm{d}\boldsymbol{\varsigma}\mathrm{d}\boldsymbol{\chi}\mathrm{d}\boldsymbol{\eta}\mathrm{d}\boldsymbol{\vartheta} \end{aligned} \tag{B6}$$

Computing the traces and taking the $n \to -K$ limit,

$$z_i^v = \int \exp\left\{\beta\left[-\frac{K}{\Omega}\sum_\omega \ln\cosh\left(\frac{\alpha}{\gamma}\left(\frac{1}{K}\sum_k P_k^\omega \xi_i^k - \zeta_\omega\right)\right) - \frac{K\alpha}{2\gamma}\sum_{\omega<\omega'}(p^{-1})_{\omega\omega'}\zeta_\omega\zeta_{\omega'}\right]\right\} d\boldsymbol{\zeta} \tag{B7}$$

$$\begin{aligned} z^h = \int \exp\Bigg\{\beta\Bigg[\frac{1}{\gamma}\sum_k \chi_k\eta_k + \sum_k \ln\cosh\left(\frac{1}{\gamma}\left(\varsigma_k - \chi_k\right)\right) - \frac{K}{2\gamma}\left(\sum_{kl}(x^{-1})_{kl}\varsigma_k\varsigma_l + \sum_{\omega<\omega'}(q^{-1})_{\omega\omega'}\vartheta_\omega\vartheta_\chi\right) \\ - \frac{K}{\Omega}\sum_\omega \ln\cosh\left(\frac{1}{\gamma}\left(\frac{1}{K}\sum_k \eta_k Q_k^\omega - \vartheta_\omega\right)\right)\Bigg]\Bigg\} d\boldsymbol{\vartheta} d\boldsymbol{\varsigma} d\boldsymbol{\chi} d\boldsymbol{\eta} \end{aligned} \tag{B8}$$

For large β, we get the following saddle-point equations extremizing z_i^v and z^h in $\boldsymbol{\zeta}$ and $\boldsymbol{\vartheta}$, $\boldsymbol{\varsigma}$, $\boldsymbol{\chi}$, $\boldsymbol{\eta}$ respectively,

$$\zeta_{\omega'} = \frac{1}{\Omega}\sum_\omega p_{\omega\omega'} \tanh\left(\frac{\alpha}{\gamma}\left(\frac{1}{K}\sum_k P_k^\omega \xi_i^k - \zeta_\omega\right)\right) \tag{B9}$$

$$\vartheta_{\omega'} = \frac{1}{\Omega}\sum_\omega q_{\omega\omega'} \tanh\left(\frac{1}{\gamma}\left(\frac{1}{K}\sum_k \eta_k Q_k^\omega - \vartheta_\omega\right)\right) \tag{B10}$$

$$\eta_k = \tanh\left(\frac{1}{\gamma}\left(\varsigma_k - \chi_k\right)\right) \tag{B11}$$

$$\chi_k = \frac{1}{\Omega}\sum_\omega Q_k^\omega \tanh\left(\frac{1}{\gamma}\left(\frac{1}{K}\sum_l \eta_l Q_l^\omega - \vartheta_\omega\right)\right) \tag{B12}$$

$$\varsigma_l = \frac{1}{K}\sum_k x_{kl} \tanh\left(\frac{1}{\gamma}\left(\varsigma_k - \chi_k\right)\right) \tag{B13}$$

Substituting the averages of v_i^ω, h^ω, and H_k^ω over the density defined by z_i^v and z^h respectively,

$$\langle v^\omega \rangle_i = \tanh\left(\frac{\alpha}{\gamma}\left(\frac{1}{K}\sum_k P_k^\omega \xi_i^k - \zeta_\omega\right)\right), \tag{B14}$$

$$\langle h^\omega \rangle = \tanh\left(\frac{1}{\gamma}\left(\frac{1}{K}\sum_k \eta_k Q_k^\omega - \vartheta_\omega\right)\right), \tag{B15}$$

$$\langle H^k \rangle = \tanh\left(\frac{1}{\gamma}\left(\varsigma_k - \chi_k\right)\right), \tag{B16}$$

gives

$$\langle v^\omega \rangle_i = \tanh\left(\frac{\alpha}{\gamma}\left(\frac{1}{K}\sum_k P_k^\omega \xi_i^k - \frac{1}{\Omega}\sum_{\omega'} p_{\omega\omega'}\langle v^{\omega'}\rangle_i\right)\right) \tag{B17}$$

$$\langle h^\omega \rangle_\mu = \tanh\left(\frac{1}{\gamma}\left(\frac{1}{K}\sum_k \eta_k Q_k^\omega - \frac{1}{\Omega}\sum_{\omega'} q_{\omega\omega'}\langle h^{\omega'}\rangle_\mu\right)\right) \tag{B18}$$

$$\langle H^k \rangle_\mu = \tanh\left(\frac{1}{\gamma}\left(\frac{1}{K}\sum_l x_{kl}\langle H_l\rangle_\mu - \frac{1}{\Omega}\sum_\omega Q_k^\omega \langle h^\omega\rangle_\mu\right)\right) \tag{B19}$$

where to account for possibly multiple solutions to the self-consistent equations of the hidden activities, we added a mute index μ to the law defined by z^h. This is an abbreviation for a similar calculation, where we add to the prior a linear source field, $\beta\varepsilon\sum_{i\mu}\sigma_i\tau_\mu w_{i\mu}$, for $\boldsymbol{\sigma}$, $\boldsymbol{\tau}$, living on the hypercube in dimensions N, M, respectively, and ε an $O(1)$ non-negative real number that we send to zero at the end. The system (15), (16), (17) closes onto the order parameters via

$$Q_k^\omega = \frac{1}{N}\sum_i \xi_i^k \langle v^\omega\rangle_i, \qquad q_{\omega\omega'} = \frac{1}{N}\sum_i \langle v^\omega\rangle_i\langle v^{\omega'}\rangle_i, \tag{B20}$$

$$P_k^\omega = \frac{1}{M}\sum_\mu \langle h^\omega\rangle_\mu\langle H_k\rangle_\mu, \qquad p_{\omega\omega'} = \frac{1}{M}\sum_\mu \langle h^\omega\rangle_\mu\langle h^{\omega'}\rangle_\mu. \tag{B21}$$

Appendix C: Consistency of the RSB equations

Eqs. (15), (16), (17) were derived using the powerful albeit non-rigorous replica method. They also turn out to be equivalent to Eqs. (20), (21), from the main text (F). Substituting Eqs. (11)–(14) into (15) and (16),

$$\langle v^\omega \rangle_i = \tanh\left(\frac{1}{N\gamma}\sum_\mu\left(\frac{1}{K}\sum_k \xi_i^k\langle H^k\rangle_\mu - \frac{1}{\Omega}\sum_{\omega'}\langle v^{\omega'}\rangle_i\langle h^{\omega'}\rangle_\mu\right)\langle h^\omega\rangle_\mu\right), \tag{C1}$$

$$\langle h^\omega \rangle_\mu = \tanh\left(\frac{1}{N\gamma}\sum_i\left(\frac{1}{K}\sum_k \xi_i^k\langle H^k\rangle_\mu - \frac{1}{\Omega}\sum_{\omega'}\langle v^{\omega'}\rangle_i\langle h^{\omega'}\rangle_\mu\right)\langle v^\omega\rangle_i\right), \tag{C2}$$

as well as the definition of x_{kl} and Eqs. (11)–(14) for Q_k^ω into eq. (17),

$$\langle H^k \rangle_\mu = \tanh\left(\frac{1}{N\gamma}\sum_i \xi_i^k\left(\frac{1}{K}\sum_l \xi_i^l\langle H_\mu^l\rangle - \frac{1}{\Omega}\sum_\omega\langle v^\omega\rangle_i\langle h^\omega\rangle_\mu\right)\right) \tag{C3}$$

we identify the RHS of eq. (21), remembering that

$$\langle H^k \rangle_\mu = \tanh\left(\sum_i w_{i\mu}\xi_i^k\right), \tag{C4}$$

thus giving eqs. (20), (21).

Appendix D: Replica-symmetric ansatz

Under the replica-symmetric (RS) ansatz

$$Q_k^{a\kappa} = Q_k, \quad P_k^{a\kappa} = P_k, \quad q_{b\lambda}^{a\kappa} = q, \quad p_{b\lambda}^{a\kappa} = p, \tag{D1}$$

for all a, κ, k and $(a\kappa) < (b\lambda)$ we find, following similar steps as in Appendix B,

$$\langle v\rangle_i = \tanh\left(\frac{\alpha}{\gamma}\left(\frac{1}{K}\sum_k P_k V_i^k - p\langle v\rangle_i\right)\right), \tag{D2}$$

$$\langle h\rangle_\mu = \tanh\left(\frac{1}{\gamma}\left(\frac{1}{K}\sum_k Q_k\langle H^k\rangle_\mu - q\langle h\rangle_\mu\right)\right), \tag{D3}$$

$$\langle H^k\rangle_\mu = \tanh\left(\frac{1}{\gamma}\left(\frac{1}{K}\sum_l x_{kl}\langle H_\mu^l\rangle - Q_k\langle h\rangle_\mu\right)\right), \tag{D4}$$

$$Q_k = \frac{1}{N}\sum_i V_i^k\langle v\rangle_i, \qquad q = \frac{1}{N}\sum_i \langle v\rangle_i^2, \tag{D5}$$

$$P_k = \frac{1}{M}\sum_\mu \langle h\rangle_\mu\langle H^k\rangle_\mu, \qquad p = \frac{1}{M}\sum_\mu \langle h\rangle_\mu^2. \tag{D6}$$

We see that the RS ansatz can only hope at best to describe the paramagnetic and the rank-1 phases.

Appendix E: RSB-ℓ ansatz for hierarchically structured data

We now propose an RSB-ℓ ansatz, $\ell = 1, \dots, L$, for hierarchically structured data (VI) corresponding to a tree of depth L. We parametrize the data index k in terms of the data tree coordinates $k = (k_0, k_1, \dots, k_L)$, $k_\ell = 1, \dots, C_\ell$, where k_L is the leaf index, $C_L = K$, and in what follows we drop the trivial root index $k_0 = C_0 = 1$. Likewise, we parametrize the RSB index ω in terms of a replica tree coordinate $\omega = (\omega_1, \dots, \omega_L)$. With these notations, we assume

$$Q_{k_1,\dots,k_L}^{\omega_1,\dots\omega_L} = \begin{cases} Q_0 & \text{if} \quad k_1 = \omega_1, \dots, k_L = \omega_L, \\ Q_1 & \text{if} \quad k_1 = \omega_1, \dots, k_L \neq \omega_L, \\ \dots & \\ Q_L & \text{if} \quad k_1 \neq \omega_1, \dots, k_L \neq \omega_L. \end{cases} \qquad P_{k_1,\dots,k_L}^{\omega_1,\dots\omega_L} = \begin{cases} P_0 & \text{if} \quad k_1 = \omega_1, \dots, k_L = \omega_L, \\ P_1 & \text{if} \quad k_1 = \omega_1, \dots, k_L \neq \omega_L, \\ \dots & \\ P_L & \text{if} \quad k_1 \neq \omega_1, \dots, k_L \neq \omega_L. \end{cases} \tag{E1}$$

$$q_{\omega_1',\dots\omega_L'}^{\omega_1,\dots\omega_L} = \begin{cases} q_0 & \text{if} \quad \omega_1 = \omega_1', \dots, \omega_L = \omega_L', \\ q_1 & \text{if} \quad \omega_1 = \omega_1, \dots, \omega_L \neq \omega_L', \\ \dots & \\ q_L & \text{if} \quad \omega_1 \neq \omega_1', \dots, \omega_L \neq \omega_L'. \end{cases} \qquad p_{\omega_1',\dots\omega_L'}^{\omega_1,\dots\omega_L} = \begin{cases} p_0 & \text{if} \quad \omega_1 = \omega_1', \dots, \omega_L = \omega_L', \\ p_1 & \text{if} \quad \omega_1 = \omega_1, \dots, \omega_L \neq \omega_L', \\ \dots & \\ p_L & \text{if} \quad \omega_1 \neq \omega_1', \dots, \omega_L \neq \omega_L'. \end{cases} \tag{E2}$$

In terms of the parameters Q_ℓ, q_ℓ, P_ℓ, p_ℓ, the self-consistent equations for the local magnetizations are

$$\begin{aligned} \langle v^{\omega_1\dots\omega_L}\rangle_i = \tanh\Bigg[\frac{\alpha}{\gamma}\Bigg(&\frac{P_0}{K}\xi_i^{\omega_1\dots\omega_L} - \frac{p_0}{\Omega}\langle v^{\omega_1\dots\omega_L}\rangle_i + \frac{P_1}{K}\sum_{k_L(\neq\omega_L)}\xi_i^{\omega_1\dots k_L} - \frac{p_1}{\Omega}\sum_{\omega_L'(\neq\omega_L)}\langle v^{\omega_1\dots\omega_L'}\rangle_i \\ &+\frac{P^{(2)}}{K}\sum_{k_L(\neq\omega_L)k_{L-1}(\neq\omega_{L-1})}\xi_i^{\omega_1\dots k_{L-1}k_L} - \frac{p_2}{\Omega}\sum_{\omega_L'(\neq\omega_L)\omega_{L-1}'(\neq\omega_{L-1})}\langle v^{\omega_1\dots\omega_{L-1}'\omega_L'}\rangle_i + \dots \\ &+\frac{P_L}{K}\sum_{k_L(\neq\omega_L)k_{L-1}(\neq\omega_{L-1})\dots k_1(\neq\omega_1)}\xi_i^{k_1\dots k_{L-1}k_L} - \frac{p_L}{\Omega}\sum_{\omega_L'(\neq\omega_L)\omega_{L-1}'(\neq\omega_{L-1})\dots\omega_1'(\neq\omega_1)}\langle v^{\omega_1'\dots\omega_{L-1}'\omega_L'}\rangle_i\Bigg)\Bigg], \end{aligned} \tag{E3}$$

$$\langle h^{\omega_1\dots\omega_L}\rangle_\mu = \tanh\left[\frac{\alpha}{\gamma}\left(\frac{Q_0}{K}\langle H^{\omega_1\dots\omega_L}\rangle_\mu - \frac{q_0}{\Omega}\langle h^{\omega_1\dots\omega_L}\rangle_\mu + \frac{Q_1}{K}\sum_{k_L(\neq\omega_L)}\langle H^{\omega_1\dots k_L}\rangle_\mu - \frac{q_1}{\Omega}\sum_{\omega'_L(\neq\omega_L)}\langle h^{\omega_1\dots\omega'_L}\rangle_\mu \right.\right.$$
$$+\frac{Q_2}{K}\sum_{k_L(\neq\omega_L)k_{L-1}(\neq\omega_{L-1})}\langle H^{\omega_1\dots k_{L-1}k_L}\rangle_\mu - \frac{q_2}{\Omega}\sum_{\omega'_L(\neq\omega_L)\omega'_{L-1}(\neq\omega_{L-1})}\langle h^{\omega_1\dots\omega'_{L-1}\omega'_L}\rangle_\mu + \dots \quad \text{(E4)}$$
$$\left.\left.+\frac{Q_L}{K}\sum_{k_L(\neq\omega_L)k_{L-1}(\neq\omega_{L-1})\dots k_1(\neq\omega_1)}\langle H^{k_1\dots k_{L-1}k_L}\rangle_\mu - \frac{q_L}{\Omega}\sum_{\omega'_L(\neq\omega_L)\omega'_{L-1}(\neq\omega_{L-1})\dots\omega'_1(\neq\omega_1)}\langle h^{\omega'_1\dots\omega'_{L-1}\omega'_L}\rangle_\mu\right)\right],$$

$$\langle H^{k_1\dots k_L}\rangle_\mu = \tanh\left[\frac{1}{\gamma}\left(\frac{1}{K}\langle H^{k_1\dots k_L}\rangle_\mu - \frac{Q_0}{\Omega}\langle h^{k_1\dots k_L}\rangle_\mu + \frac{x_1}{K}\sum_{l_L(\neq k_L)}\langle H^{k_1\dots l_L}\rangle_\mu - \frac{Q_1}{\Omega}\sum_{\omega_L(\neq k_L)}\langle h^{k_1\dots\omega_L}\rangle_\mu \right.\right.$$
$$+\frac{x_2}{K}\sum_{l_L(\neq k_L)l_{L-1}(\neq k_{L-1})}\langle H^{k_1\dots l_{L-1}l_L}\rangle_\mu - \frac{Q_2}{\Omega}\sum_{\omega_L(\neq k_L)\omega_{L-1}(\neq k_{L-1})}\langle h^{k_1\dots\omega_{L-1}\omega_L}\rangle_\mu\Big)\Big] + \dots \quad \text{(E5)}$$
$$\left.\left.+\frac{x_L}{K}\sum_{l_L(\neq k_L)l_{L-1}(\neq k_{L-1})\dots l_1(\neq k_1)}\langle H^{l_1\dots l_{L-1}l_L}\rangle_\mu - \frac{Q_L}{\Omega}\sum_{\omega_L(\neq k_L)\omega_{L-1}(\neq k_{L-1})\dots\omega_1(\neq k_1)}\langle h^{\omega_1\dots\omega_{L-1}\omega_L}\rangle_\mu\right)\right].$$

Appendix F: Derivation of the stationarity conditions in the undersampled regime

As shown in Appendix I, the trained weights in the undersampled regime, will be at most of rank $\leq 2^K$ with K finite, but in practice their rank is always found to be $\leq K$. Let us then go ahead and assume that the RBM weights are rank K, given as

$$w_{i\mu} = \frac{1}{N}\sum_k \xi_i^k c_\mu^k \quad \text{(F1)}$$

for some coefficients c_μ^k. The input on hidden unit μ from a data point $\boldsymbol{\xi}^k$ equals $I_\mu^k = \sum_i w_{i\mu}\xi_i^k = \sum_l c_\mu^l x^{lk}$. In particular $\langle h_\mu|\boldsymbol{\xi}^k\rangle = \tanh(I_\mu^k) = \tanh\left(\sum_l c_\mu^l x^{lk}\right)$. The RBM energy writes:

$$-E(\mathbf{v},\mathbf{h}) = M\sum_k m_k n_k \quad \text{(F2)}$$

where

$$m_k(\mathbf{v}) = \frac{1}{N}\sum_i \xi_i^k v_i, \qquad n_k(\mathbf{h}) = \frac{1}{M}\sum_\mu c_\mu^k h_\mu, \quad \text{(F3)}$$

while the effective energy is, up to irrelevant factors,

$$-E_{\text{eff}}(\mathbf{v}) = \sum_\mu \ln\cosh\left(\sum_k m_k c_\mu^k\right). \quad \text{(F4)}$$

Using the Hubbard-Stratonovich transform,

$$P(\mathbf{v},\mathbf{h}) \propto \int \exp\left(-M\sum_{kl} m_k n_l + \alpha\sum_{ki} n_k\xi_i^k v_i + \sum_{k\mu} m_k c_\mu^k h_\mu\right) d\mathbf{m}d\mathbf{n}. \quad \text{(F5)}$$

For large N one has $Z \sim e^{-NF(\mathbf{m},\mathbf{n})}$ with the free energy

$$F(\mathbf{m},\mathbf{n}) = \alpha\sum_k m_k n_k - \frac{1}{N}\sum_i \ln\cosh\left(\alpha\sum_k n_k v_i^k\right) - \frac{1}{N}\sum_\mu \ln\cosh\left(\sum_k m_k h_\mu^k\right) \quad \text{(F6)}$$

Here $\mathbf{m} = (m_1, \ldots, m_K)$, $\mathbf{n} = (n_1, \ldots, n_K)$ are minima of $F(\mathbf{m}, \mathbf{n})$, solving the saddle-point equations

$$m_k = \frac{1}{N} \sum_i \xi_i^k \langle v_i \rangle, \qquad n_k = \frac{1}{M} \sum_\mu c_\mu^k \langle h_\mu \rangle, \tag{F7}$$

with

$$\langle v_i \rangle = \tanh\left(\alpha \sum_k n_k \xi_i^k \right), \quad \langle h_\mu \rangle = \tanh\left(\sum_k m_k c_\mu^k \right) \tag{F8}$$

The minima of F can be found by initializing at a Mattis state, $m_k = x_{kl}$ for some l, and iterating the above equations until convergence. The resulting saddle-point is the local minima where the RBM average activities $\langle v_i \rangle$ are closest to the data point ξ_i^k. Initializing at different rows of x_{kl} may sometimes lead to the same saddle-point, due to merging. Thus at most K saddle-points are found in this way, that we denote by m_k^ω, n_k^ω, with $\omega = 1, \ldots, \Omega \leq K$. Each saddle-point defines a thermodynamic state, with averages denoted $\langle v_i \rangle^\omega, \ldots$ and so on. Substituting,

$$\langle v_i \rangle = \tanh\left(\frac{\alpha}{M} \sum_{k\mu} \xi_i^k c_\mu^k \langle h_\mu \rangle \right), \tag{F9}$$

$$\langle h_\mu \rangle = \tanh\left(\frac{1}{N} \sum_{ki} \xi_i^k c_\mu^k \langle v_i \rangle \right). \tag{F10}$$

The RBM trained by maximum likelihood, satisfies the moment-matching conditions:

$$\langle v_i h_\mu \rangle_{\mathcal{D}} - \langle v_i h_\mu \rangle - N\gamma w_{i\mu} = 0. \tag{F11}$$

Substituting the ansatz,

$$\frac{1}{K} \sum_k \xi_i^k \tanh\left(\sum_l x_{kl} c_\mu^l \right) - \frac{1}{\Omega} \sum_\omega \langle v_i \rangle^\omega \langle h_\mu \rangle^\omega - \gamma \sum_k \xi_i^k c_\mu^k = 0. \tag{F12}$$

Appendix G: The weights as implicit functions of γ

We invoke the implicit function theorem [20], to the weights in Equation (7) with respect to γ, obtaining:

$$\sum_{j\nu} \frac{\partial^2 \mathcal{L}}{\partial w_{i\mu} \partial w_{j\nu}} \frac{dw_{j\nu}}{d\gamma} = N w_{i\mu} \tag{G1}$$

Since the weights are a local maximum of the likelihood, the Hessian $\partial^2 \mathcal{L} / (\partial w_{i\mu} \partial w_{j\nu})$ is negative definite. Thus $dw_{j\nu}/d\gamma$ can be found by inversion of (G1). Note that it vanishes only at the origin.

We can then integrate (G1) to obtain parametric curves $w_{i\mu}(\gamma)$ [21] that trace the different solutions of (7). For $K\gamma \geq \lambda_{\max}$, these curves intersect at the origin, which is the only fixed point of the flow (G1).

As γ approaches zero, the curves diverge to infinity, in a manner described by the BAM phase discussed in the text. In this regime, the Hessian simplifies to

$$\frac{\partial^2 \mathcal{L}}{\partial w_{i\mu} \partial w_{j\nu}} = -\frac{1}{K} \sum_k A_{i\mu}^k A_{j\nu}^k - N\gamma \delta_{ij} \delta_{\mu\nu} \tag{G2}$$

where $A_{i\mu}^k = \xi_i^k \hat{\xi}_\mu^k - \frac{1}{K} \sum_l \xi_i^l \hat{\xi}_\mu^l$, in the same notation of Section V in the main text. The asymptotic behavior of $w_{i\mu}(\gamma)$ for $\gamma \to 0$ is then ruled by the eigen-modes of (G2).

The term $\sum_k A_{i\mu}^k A_{j\nu}^k$ is a rank K positive semi-definite matrix in weight space. It has the same eigenvalues as

$$\sum_{i\mu} A_{i\mu}^k A_{i\mu}^l = \sum_{i\mu} \left(\xi_i^k \hat{\xi}_\mu^k - \frac{1}{K} \sum_t \xi_i^t \hat{\xi}_\mu^t \right) \left(\xi_i^l \hat{\xi}_\mu^l - \frac{1}{K} \sum_t \xi_i^t \hat{\xi}_\mu^t \right) \tag{G3}$$

$$= MN \left(x_{kl} y_{kl} - \frac{1}{K} \sum_t (x_{kt} y_{kt} + x_{lt} y_{lt}) + \frac{1}{K^2} \sum_{st} x_{st} y_{st} \right) \tag{G4}$$

where $x_{kl} = \frac{1}{N}\sum_i \xi_i^k \xi_i^l$ and $y_{kl} = \frac{1}{M}\sum_\mu \hat{\xi}_\mu^k \hat{\xi}_\mu^l$. If $\hat{r}_k$ is an eigenvector of $\sum_{i\mu} A_{i\mu}^k A_{i\mu}^l$, then $\hat{w}_{i\mu} = \sum_k A_{i\mu}^k \hat{r}_k$ is an eigenvector of $\sum_k A_{i\mu}^k A_{j\nu}^k$, with the same eigenvalue. This shows that the asymptotic weights, are of the form:

$$w_{i\mu} \propto \sum_k \hat{r}_k \left(\xi_i^k \hat{\xi}_\mu^k - \frac{1}{K} \sum_l \xi_i^l \hat{\xi}_\mu^l \right) = \sum_k w_k \xi_i^k \hat{\xi}_\mu^k \tag{G5}$$

where $w_k = \hat{r}_k - \frac{1}{K}\sum_l \hat{r}_l$. As mentioned in the main-text, this form corresponds to the "weighted learning rule" of BAM [18].

Now, the matrix $\sum_{i\mu} A_{i\mu}^k A_{i\mu}^l$ is a projection of $x_{kl} y_{kl}$ to the hyperplane orthogonal to the all ones vector, $\mathbf{1} = (1, \ldots, 1)$, which implies that it develops a zero eigenvalue, associated to the eigenvector $\mathbf{1}$. This happens to be its smallest eigenvalue, and will therefore eventually dominate the dynamics of (G1). Indeed, other modes would converge towards a finite value for the weights, whereas this mode is the one that leads to weights that diverge like $1/\gamma$ as $\gamma \to 0$. Therefore, (G5) simplifies to:

$$w_{i\mu} \propto \sum_k \xi_i^k \hat{\xi}_\mu^k \tag{G6}$$

the original learning rule proposed by Kosko [10].

Appendix H: Stability analysis

The Hessian of the log-likelihood writes:

$$\frac{\partial^2 \mathcal{L}}{\partial w_{i\mu} \partial w_{j\nu}} = \frac{1}{K} \sum_k \xi_i^k \xi_j^k (1 - \langle h_\mu | \mathbf{v}^k \rangle^2) \delta_{\mu\nu} - \langle v_i v_j h_\mu h_\nu \rangle + \langle v_i h_\mu \rangle \langle v_j h_\nu \rangle - N\gamma \delta_{ij} \delta_{\mu\nu} \tag{H1}$$

where $\langle h_\mu | \mathbf{v} \rangle = \tanh\left(\sum_i w_{i\mu} v_i\right)$.

We can consider a general perturbation $\tilde{w}_{i\mu} = w_{i\mu} + \epsilon_{i\mu}$ to the weights, where $\epsilon_{i\mu} = \epsilon_{i\mu}^{\parallel} + \epsilon_{i\mu}^{\perp}$ may have both parallel $\epsilon_{i\mu}^{\parallel}$ and perpendicular $\epsilon_{i\mu}^{\perp}$ components to the data. We will assume that the unperturbed weights $w_{i\mu}$ are fully contained in the span of the data, and that they are stationary,

$$\langle v_i h_\mu \rangle_{\mathrm{d}} - \langle v_i h_\mu \rangle - N\gamma w_{i\mu} = 0 \tag{H2}$$

We will also assume that $w_{i\mu}$ are locally stable within the space of the data. That is, $\mathcal{L}(\mathbf{W} + \epsilon^{\parallel}) \leq \mathcal{L}(\mathbf{W})$ to second-order, for any small perturbation $\epsilon^{\parallel}$ contained in the data span. After some algebra,

$$\mathcal{L}(\mathbf{W}) \approx \mathcal{L}(\mathbf{W} + \epsilon^{\parallel}) - \frac{1}{2} \sum_{ij} \sum_{\mu\nu} (\epsilon_{i\mu}^{\parallel} \epsilon_{j\nu}^{\perp} + \epsilon_{i\mu}^{\perp} \epsilon_{j\nu}^{\parallel} + \epsilon_{i\mu}^{\perp} \epsilon_{j\nu}^{\perp}) \langle v_i v_j h_\mu h_\nu \rangle - \frac{N\gamma}{2} \sum_{i\mu} (\epsilon_{i\mu}^{\perp})^2 \tag{H3}$$

In the thermodynamic limit of the undersampled regime we have considered in this paper, we have $\langle v_i h_\mu \rangle = \frac{1}{\Omega} \sum_\omega \langle v_i \rangle^\omega \langle h_\mu \rangle^\omega$, where $\omega \in \{1, \ldots, \Omega\}$ are the distinct saddle-points of the free energy (or states). If we assume that the sets of vectors $\{\langle v_i \rangle^\omega\}$ and $\{\langle h_\mu \rangle^\omega\}$ are both linearly independent, then each $\langle \mathbf{v} \rangle^\omega$ must be in the span of the data (because the colspan of $\langle v_i h_\mu \rangle$ is in the span of the data). But in the thermodynamic limit, we also have:

$$\langle v_i v_j h_\mu h_\nu \rangle = \frac{1}{\Omega} \sum_\omega \langle v_i \rangle^\omega \langle v_j \rangle^\omega \langle h_\mu \rangle^\omega \langle h_\nu \rangle^\omega \tag{H4}$$

If each of the $\langle v_i \rangle^\omega$ is also in the span of the data, the above simplifies to:

$$\mathcal{L} \approx \mathcal{L}(\mathbf{W} + \epsilon^{\parallel}) - \frac{N\gamma}{2} \sum_{i\mu} (\epsilon_{i\mu}^{\perp})^2 \leq \mathcal{L}(\mathbf{W}) \tag{H5}$$

which confirms the stability of the data parallel solution.

Appendix I: The global maxima weights are low-rank

The data defines a grouping of sites $i = 1, \ldots, N$, into classes $\mathcal{I}$, such that $\xi_i^k = \xi_{\mathcal{I}}^k$ for all k, has the same value for all $i \in \mathcal{I}$. There are 2^K possible values that the K-tuple $(\xi_{\mathcal{I}}^1, \ldots, \xi_{\mathcal{I}}^K)$ can take, and therefore there are 2^K possible groups $\mathcal{I}$. Let $-\mathcal{I}$ denote the *opposite* class of $\mathcal{I}$, that is $\xi_{-\mathcal{I}}^k = -\xi_{\mathcal{I}}^k$ takes the negative values of the sites in $\mathcal{I}$.

Let $\mathcal{V}$ denote the space of vectors $\mathbf{w} = (w_i)$ satisfying $w_i = w_{\mathcal{I}}$ for all $i \in \mathcal{I}$, and $w_i = -w_{\mathcal{I}}$ whenever $i \in -\mathcal{I}$. In other words, w_i respects the same hyperoctahedral symmetries regarding the sites i as the data. It is easy to see that $\mathcal{V}$ is a vector space and that it contains the data points, $\boldsymbol{\xi}^k \in \mathcal{V}$.

Decompose the weights as $w_{i\mu} = w_{i\mu}^{\parallel} + w_{i\mu}^{\perp}$, where $\mathbf{w}_\mu^{\parallel} \in \mathcal{V}$, while $\mathbf{w}_\mu^{\perp} \perp \mathcal{V}$ is orthogonal to all vectors in $\mathcal{V}$. Note that this decomposition is unique. We analyze how the three terms in the log-likelihood depend on the parallel and orthogonal components of the weights.

The first term of the log-likelihood depends only on the parallel component,

$$\frac{1}{K}\sum_k \ln \operatorname*{tr}_{\mathbf{h}} \exp\left(\sum_{i\mu} w_{i\mu}\xi_i^k h_\mu\right) = \frac{1}{K}\sum_k \ln \operatorname*{tr}_{\mathbf{h}} \exp\left(\sum_{i\mu} w_{i\mu}^{\parallel}\xi_i^k h_\mu\right) \tag{I1}$$

because the dot product with the data annihilates any transverse components, $\sum_i w_{i\mu}\xi_i^k = \sum_i w_{i\mu}^{\parallel}\xi_i^k$.

For the regularization we have that

$$\|\mathbf{W}\|^2 = \|\mathbf{W}^{\parallel}\|^2 + \|\mathbf{W}^{\perp}\|^2 \geq \|\mathbf{W}^{\parallel}\|^2 \tag{I2}$$

by properties of the L2-norm (Pythagoras theorem). The regularization pushes towards decreasing any transverse norm $\|\mathbf{W}^{\perp}\|^2$.

The log-partition function $\ln Z(\mathbf{W})$ is a convex function of $\mathbf{W}$. This is a consequence of the fact that $\ln Z(\mathbf{W})$ is of the Log-Sum-Exp class [28].

It is also invariant under the action of the hyperoctahedral group [29]. More explicitly, let g be a member of the hyperoctahedral group. This means that g is specified by some permutation $\pi_1, \ldots, \pi_N$ of of the coordinate axes $1, \ldots, N$, and by N signs which correspond to reflections, $\sigma_1, \ldots, \sigma_N = \pm 1$. Let $g(\mathbf{W})$ be the weight matrix after the action of g,

$$g(\mathbf{W})_{i\mu} = \sigma_{\pi_i} w_{\pi_i \mu}$$

Then:

$$Z(g(\mathbf{W})) = \operatorname*{tr}_{\mathbf{v},\mathbf{h}} \exp\left(\sum_\mu \mathbf{v}^\top g(\mathbf{w}_\mu) h_\mu\right) = \operatorname*{tr}_{\mathbf{v},\mathbf{h}} \exp\left(\sum_\mu \mathbf{w}_\mu^\top g^{-1}(\mathbf{v}) h_\mu\right) = Z(\mathbf{W}) \tag{I3}$$

where g^{-1} denotes the inverse of g in the hyperoctahedral group. Since summing over the $g^{-1}(\mathbf{v})$ amounts to a reordering of the vertices of the hypercube, the sum has the same value (a sum is invariant to the order of the terms added), and therefore $Z(\mathbf{W}_g) = Z(\mathbf{W})$.

Among all g in the hyperoctahedral group, there are some that leave the data points invariant, that is, $g(\mathbf{v}^k) = \mathbf{v}^k$ for all $k = 1, \ldots, K$. The set of such g forms a subgroup, that we denote by $\mathcal{G}_{\mathcal{D}}$. Its members, $g \in \mathcal{G}_{\mathcal{D}}$, satisfy:

$$g(\boldsymbol{\xi}^k) = (\sigma_{\pi_1}\xi_{\pi_1}^k, \ldots, \sigma_{\pi_N}\xi_{\pi_N}^k) = (\xi_1^k, \ldots, \xi_N^k)$$

That is, the data vectors $\boldsymbol{\xi}^k$ are eigenvectors of the matrix associated to the transformation g and with eigenvalue one. The following $N \times K$ equations must then hold,

$$\sigma_{\pi_i}\xi_{\pi_i}^k = \xi_i^k \tag{I4}$$

for all $i = 1, \ldots, N$ and all $k = 1, \ldots, K$.

If $\sigma_{\pi_i} = 1$, this means that $\xi_{\pi_i}^1 = \xi_i^1, \ldots, \xi_{\pi_i}^K = \xi_i^K$. Thus if $i \in \mathcal{I}$ then $\pi_i \in \mathcal{I}$ must be in the same class. If $\sigma_{\pi_i} = -1$, then $\xi_{\pi_i}^1 = -\xi_i^1, \ldots, \xi_{\pi_i}^K = -\xi_i^K$, which means that if $i \in \mathcal{I}$ then $\pi_i \in -\mathcal{I}$ must belong to the opposite class.

In any case, if $i \in \mathcal{I}$, then $\pi_i \in \mathcal{I}$ or $\pi_i \in -\mathcal{I}$. In the first case $\sigma_{\pi_i} = +1$ and in the second case $\sigma_{\pi_i} = -1$. If we consider the restriction of π to sites belonging to two opposite classes $\mathcal{I}$ or $-\mathcal{I}$, we have that π is allowed to permute these sites freely. That is, we can consider any permutation of the sites in $\mathcal{I} \cup -\mathcal{I}$.

We can also view this algebraically, multiplying the equation $\sigma_{\pi_i}\xi_{\pi_i}^k = \xi_i^k$ by $\xi_{\pi_i}^k$, gives

$$\sigma_{\pi_i} = \xi_i^k \xi_{\pi_i}^k$$

which says the same thing: $\sigma_{\pi_i} = 1$ or $\sigma_{\pi_i} = -1$ according to whether π_i sends i to $\mathcal{I}$ or $-\mathcal{I}$, where $i \in \mathcal{I}$. The permutation π is only constrained to satisfy that $\xi_i^k \xi_{\pi_i}^k$ has the same value for all k.

Now consider an arbitrary vector $\mathbf{w} = (w_1, \dots, w_N)$. We construct $\tilde{\mathbf{w}}$ by considering all possible actions of $\mathcal{G}_\mathcal{D}$ on $\mathbf{w}$, and averaging:

$$\tilde{\mathbf{w}} = \frac{1}{|\mathcal{G}_\mathcal{D}|} \sum_{g \in \mathcal{G}_\mathcal{D}} g(\mathbf{w})$$

This vector is invariant to the action of any $g \in \mathcal{G}_\mathcal{D}$:

$$g(\tilde{\mathbf{w}}) = \frac{1}{|\mathcal{G}_\mathcal{D}|} \sum_{g' \in \mathcal{G}_\mathcal{D}} (g \circ g')(\mathbf{w}) = \frac{1}{|\mathcal{G}_\mathcal{D}|} \sum_{g' \in \mathcal{G}_\mathcal{D}} g(\mathbf{w}) = \tilde{\mathbf{w}}$$

because summing over $g \circ g'$ instead of g' amounts to just a re-ordering of the terms in the sum. It follows that, for all $g \in \mathcal{G}_\mathcal{D}$, we must satisfy the equations

$$\sigma_{\pi_i} \tilde{w}_{\pi_i} = \tilde{w}_i$$

for all $i = 1, \dots, N$. Suppose $i \in \mathcal{I}$. Then $\pi_i \in \mathcal{I}$ or $\pi_i \in -\mathcal{I}$, according to whether $\sigma_{\pi_i} = 1$ or $\sigma_{\pi_i} = -1$, respectively. Therefore, we must have $\tilde{w}_j = \tilde{w}_i$ for all $j \in \mathcal{I}$, while $\tilde{w}_j = -\tilde{w}_i$ for all $j \in -\mathcal{I}$. In other words, $\tilde{\mathbf{w}} \in \mathcal{V}$.

Now we can come back to our decomposition of the weights, $w_{i\mu} = w_{i\mu}^\parallel + w_{i\mu}^\perp$, with $\mathbf{w}_\mu^\parallel \in \mathcal{V}$ and $\mathbf{w}_\mu^\perp \perp \mathcal{V}$. Clearly,

$$\frac{1}{|\mathcal{G}_\mathcal{D}|} \sum_{g \in \mathcal{G}_\mathcal{D}} g(\mathbf{W}) = \mathbf{W}^\parallel$$

Next, by the convexity of $\ln Z(\mathbf{W})$,

$$\ln Z(\mathbf{W}^\parallel) = \ln Z \left(\frac{1}{|\mathcal{G}_\mathcal{D}|} \sum_{g \in \mathcal{G}_\mathcal{D}} g(\mathbf{W}) \right) \leq \frac{1}{|\mathcal{G}_\mathcal{D}|} \sum_{g \in \mathcal{G}_\mathcal{D}} \ln Z(g(\mathbf{W})) \tag{I5}$$

But we have seen above that $Z(g(\mathbf{W})) = Z(\mathbf{W})$ is invariant, for any g in the hyperoctahedral group. Therefore, we have that:

$$\ln Z(\mathbf{W}^\parallel) \leq \ln Z(\mathbf{W})$$

Finally, combining all three terms in the log-likelihood, we have that:

$$\mathcal{L}(\mathbf{W}) = \mathcal{L}(\mathbf{W}^\parallel + \mathbf{W}^\perp) \leq \mathcal{L}(\mathbf{W}^\parallel) \tag{I6}$$

We conclude that the log-likelihood is maximized by weights that lie inside $\mathcal{V}$. Note that this derivation remains valid for other regularizations in addition to L2, as long as the regularization is convex and invariant under the action of the hyperoctahedral group.

Appendix J: Linear programming formulation of BAM phase for small K

With the weights given by Eq. (26), the inequalities (25) write:

$$\hat{\xi}_\mu^k \sum_l x_{kl} \hat{\xi}_\mu^l \geq 0, \quad \xi_i^k \sum_l y_{kl} \xi_i^l \geq 0 \tag{J1}$$

The first inequality involves single hidden units, and defines a set of admissible vectors $\boldsymbol{\xi}_\mu = (\hat{\xi}_\mu^1, \dots, \hat{\xi}_\mu^K) \in \{\pm 1\}^K$, that do not change orthant when matrix x is applied. The second inequality involves only y_{kl}, not individual units. Now if we consider only admissible vectors, we can count how many copies of each such vector are present, say $M\psi_\mu$, where $\psi_\mu \geq 0$ and $\sum_\mu \psi_\mu = 1$. Then we can write $y_{kl} = \sum_\mu \psi_\mu \hat{\xi}_\mu^k \hat{\xi}_\mu^l$. Crucially, the first inequality is already satisfied (independently of ψ_μ) because we are considering now only admissible vectors. The second inequality writes:

$$\sum_l y_{kl} \xi_i^k \xi_i^l = \sum_{l\mu} \psi_\mu \hat{\xi}_\mu^k \hat{\xi}_\mu^l \xi_i^k \xi_i^l \geq 0 \tag{J2}$$

and must be satisfied for all k, i. Together with $\psi_\mu \geq 0$, $\sum_\mu \psi_\mu = 1$, we find that ψ_μ is confined to a convex polytope.

Lastly, under Eq. (26), the log-likelihood simplifies to:

$$\mathcal{L}(\mathbf{W}) \sim \frac{WM}{K}\left(2 - \frac{KW\gamma}{2}\right)\sum_{kl} x_{kl} y_{kl} \tag{J3}$$

where W denotes the (large) norm of the weights (*i.e.*, the missing proportionality constant in Eq. (26)). Maximizing (J3) with respect to the ψ_μ, under constraints (J2), is a linear program (LP), and can be easily solved for small K [30].