

Minimal model of permutation symmetry in unsupervised learning

Tianqi Hou

Department of Physics, the Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

K. Y. Michael Wong

Department of Physics, the Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

Haiping Huang[‡]

PMI Lab, School of Physics, Sun Yat-sen University, Guangzhou 510275, People's Republic of China

E-mail: huanghp7@mail.sysu.edu.cn

Abstract. Permutation of any two hidden units yields invariant properties in typical deep generative neural networks. This permutation symmetry plays an important role in understanding the computation performance of a broad class of neural networks with two or more hidden units. However, a theoretical study of the permutation symmetry is still lacking. Here, we propose a minimal model with only two hidden units in a restricted Boltzmann machine, which aims to address how the permutation symmetry affects the critical learning data size at which the concept-formation (or spontaneous symmetry breaking in physics language) starts, and moreover semi-rigorously prove a conjecture that the critical data size is independent of the number of hidden units once this number is finite. Remarkably, we find that the embedded correlation between two receptive fields of hidden units reduces the critical data size. In particular, the weakly-correlated receptive fields have the benefit of significantly reducing the minimal data size that triggers the transition, given less noisy data. Inspired by the theory, we also propose an efficient fully-distributed algorithm to infer the receptive fields of hidden units. Furthermore, our minimal model reveals that the permutation symmetry can also be spontaneously broken following the spontaneous symmetry breaking. Overall, our results demonstrate that the unsupervised learning is a progressive combination of spontaneous symmetry breaking and permutation symmetry breaking which are both spontaneous processes driven by data streams (observations). All these effects can be analytically probed based on the minimal model, providing theoretical insights towards understanding unsupervised learning in a more general context.

[‡] correspondence author

1. Introduction

Unsupervised learning is defined as the process of searching for latent features in raw (unlabeled) data, and thus serves as a fundamental property of the cerebral cortex of the brain [1,2]. To understand unsupervised learning from a neural network perspective, restricted Boltzmann machine (RBM) is proposed. RBM is a two-layered neural network with one layer called the visible layer, and the other called the hidden layer. No lateral connections exist in each layer. The connections between visible and hidden layers are called synaptic weights, representing encoded latent features in the observed data. The process of learning the synaptic weights from the unlabeled data (also called training) mimics the unsupervised learning. RBM is thus receiving substantial research interests both from machine learning and statistical physics communities [3–13].

Training of RBM relies on the maximum likelihood principle via a gradient ascent procedure. The mean activity of each neuron or correlations between visible and hidden neurons can be estimated by either truncated Gibbs sampling [3,14] or advanced mean-field methods [6,7]. However, the gradient ascent method is difficult to analyze and thus not amenable for a theoretical model. Therefore, based on the probabilistic graphical model framework, one-bit RBM where only one hidden neuron is considered was proposed to address a fundamental issue of unsupervised learning [15], i.e., how many data samples are needed for a successful learning. This work revealed a continuous spontaneous symmetry breaking (SSB) transition separating a random-guess phase from a concept-formation phase at a critical value of the amount of provided samples (data size) [16], which is similar to the retarded learning phase transition observed in a generalized Hopfield model of pattern learning [17]. This conclusion is later generalized to RBM with generic priors [18,19], and synapses of ternary values [20]. However, it is still challenging to handle the case of multiple hidden neurons from the perspective of understanding the learning process as a phase transition. In the presence of multiple hidden neurons, permutation symmetry appears, i.e., the model of the observed data is invariant with respect to exchange of arbitrary two hidden neurons. In addition, the permutation symmetry is a common feature in many modern neural network architectures [21]. Therefore, understanding how the permutation symmetry affects the concept-formation process is important, which may provide us core mechanisms of unsupervised learning.

Here, we propose a minimal model of the permutation symmetry in unsupervised learning, based on mean-field approximations. We show that it is possible to theoretically understand the permutation symmetry using physics approximations. To be more precise, we consider a RBM with two hidden neurons, and embed a latent feature that generates a certain number of data samples through Gibbs samplings of the original model [15]. Then, the data samples are learned by a theory-inspired algorithm, and finally the learned synaptic weights (a latent feature vector) are compared with the embedded ones, to test whether SSB applies to the minimal model, and in addition investigate key factors affecting the critical data size for learning and moreover how

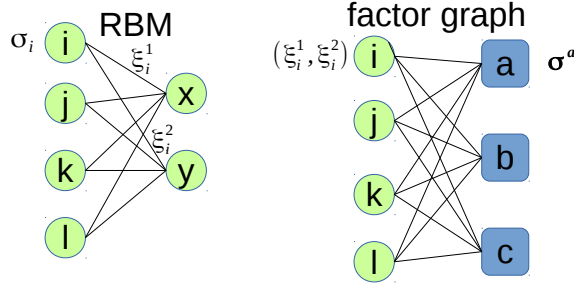


Figure 1. (Color online) A schematic illustration of the minimal model. $N = 4$ in this example (say, i, j, k and l). (Left panel) The original model with only two hidden neurons (say, x and y). (Right panel) The corresponding factor graph where the data node is indicated by a square, and the paired-synapses (feature vector) is indicated by a circle. In this example, $M = 3$ (say, a, b and c). The circle is an augmented version of single synapse considered in the one-bit RBM [15].

the permutation symmetry affects the learning process. We first apply the cavity approximation in statistical mechanics of disordered systems [22] to derive the learning algorithm from a Bayesian inference perspective, whose computation performances in single instances of the model are then predicted by a replica theory. This theory introduces many copies of the original model, and the interaction between any two copies is characterized by a set of self-consistent mean-field equations, from which the critical data size for learning is determined, and moreover whether the permutation symmetry can be spontaneously broken is clarified.

2. Model definition and mean-field methods

2.1. Minimal model of permutation symmetry

In this study, we use the RBM defined above with two hidden neurons (Fig. 1) to learn embedded features in input data samples, which are raw unlabeled data. Each data sample is specified by an Ising-like spin configuration $\boldsymbol{\sigma} = \{\sigma_i = \pm 1\}_{i=1}^N$ where N is the input dimensionality. A collection of M samples is denoted as $\{\boldsymbol{\sigma}^a\}_{a=1}^M$. Synaptic values connecting visible and hidden neurons are characterized by $\boldsymbol{\xi}$, where each component takes a binary value (± 1) as well. Because of two hidden neurons, $\boldsymbol{\xi} = (\boldsymbol{\xi}^1, \boldsymbol{\xi}^2)$ where the superscript indicates the hidden neuron's index. $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$ are also called receptive fields of the first and second hidden neurons, respectively. Statistical properties of this RBM are thus described by the Boltzmann distribution [6]

$$P(\boldsymbol{\sigma}) = \frac{1}{Z(\boldsymbol{\xi})} \cosh(\beta X) \cosh(\beta Y), \quad (1)$$

where $X = \frac{1}{\sqrt{N}} \boldsymbol{\xi}^1 \cdot \boldsymbol{\sigma}$, $Y = \frac{1}{\sqrt{N}} \boldsymbol{\xi}^2 \cdot \boldsymbol{\sigma}$, and $Z(\boldsymbol{\xi})$ is the partition function depending on the feature $\boldsymbol{\xi}$. Note that the two hidden neurons' activities (± 1) have been marginalized out. The scaling factor $\frac{1}{\sqrt{N}}$ ensures that the argument of the hyperbolic cosine function is of

the order of unity. σ can be arbitrary one of the M samples. When the embedded feature is randomly generated, the inverse-temperature β tunes the noise level of generated data samples from the feature. Clearly, the data distribution is invariant with respect to (w.r.t) the exchange of the hidden neurons, which is called the permutation symmetry in this paper. The required number of hidden neurons to yield this symmetry is at least two, therefore, this setup defines a minimal model to study the permutation symmetry in unsupervised learning.

In this model, the embedded feature follows the distribution $P(\xi) = P(\xi^1)P(\xi^2|\xi^1)$ in which $P(\xi^1) = \prod_{i=1}^N [\frac{1}{2}\delta(\xi_i^1 - 1) + \frac{1}{2}\delta(\xi_i^1 + 1)]$ together with $P(\xi^2|\xi^1) = \prod_{i=1}^N [p_d\delta(\xi_i^2 = -\xi_i^1) + (1 - p_d)\delta(\xi_i^2 = \xi_i^1)]$, where p_d controls the fraction of components taking different values in the two feature maps associated with the two hidden neurons.

Given the M data samples, one gets the posterior probability of the embedded feature according to the Bayes' rule:

$$\begin{aligned} P(\xi|\{\sigma^a\}_{a=1}^M) &= \frac{\prod_a P(\sigma^a|\xi)}{\sum_{\xi} \prod_a P(\sigma^a|\xi)} \\ &= \frac{1}{\Omega} \prod_a \frac{1}{Z(\xi^1, \xi^2)} \cosh\left(\frac{\beta}{\sqrt{N}}\xi^1 \cdot \sigma^a\right) \cosh\left(\frac{\beta}{\sqrt{N}}\xi^2 \cdot \sigma^a\right), \end{aligned} \quad (2)$$

where Ω is the partition function of the minimal model. For simplicity, a uniform prior for ξ is assumed, i.e., we have no prior knowledge about ξ , although there may exist correlations between two feature maps. In addition, we use the same temperature as that used to generate data. Because we do not use the true prior $\prod_i P_i(\xi_i^1, \xi_i^2|p_d)$, the current setting does not require the value of p_d , thereby is not the Bayes-optimal setting which corresponds to Nishimori condition in physics [16]. Therefore, using the uniform prior is more computationally challenging. We leave a detailed analysis of the Bayes-optimal setting in a future work.

One obstacle to compute the posterior probability is the nested partition function $Z(\xi^1, \xi^2)$. Fortunately, this partition function can be simplified in the large- N limit. More precisely,

$$\begin{aligned} Z(\xi^1, \xi^2) &= \sum_{\sigma} \cosh\left(\frac{\beta}{\sqrt{N}}\xi^1 \cdot \sigma\right) \cosh\left(\frac{\beta}{\sqrt{N}}\xi^2 \cdot \sigma\right) \\ &= \frac{1}{2} \sum_{\sigma} [\cosh(X + Y) + \cosh(X - Y)] \\ &= \frac{1}{2} \left[\prod_i 2 \cosh\left(\frac{\beta}{\sqrt{N}}(\xi_i^1 + \xi_i^2)\right) + \prod_i 2 \cosh\left(\frac{\beta}{\sqrt{N}}(\xi_i^1 - \xi_i^2)\right) \right] \\ &\simeq 2^N e^{\beta^2 Q} \cosh(\beta^2 Q), \end{aligned} \quad (3)$$

where we have used $\ln \cosh(x) \simeq \frac{x^2}{2}$ for small x to arrive at the final equality, and $Q \equiv \frac{1}{N} \sum_i \xi_i^1 \xi_i^2$, which is exactly the overlap between the two feature maps. To sum up, we move all the irrelevant constants into the partition function Ω , the posterior

probability can then be rewritten as

$$P(\xi|\{\sigma^a\}_{a=1}^M) = \frac{1}{\Omega} \prod_a \frac{1}{\cosh(\beta^2 Q)} \cosh\left(\frac{\beta}{\sqrt{N}} \xi^1 \cdot \sigma^a\right) \cosh\left(\frac{\beta}{\sqrt{N}} \xi^2 \cdot \sigma^a\right), \quad (4)$$

which forms the Boltzmann distribution of our minimal model. In this paper, we consider the case of $M = \alpha N$ where α specifies the data (constraint) density.

2.2. Cavity approximation to handle the posterior probability

In what follows, we compute the maximizer of the posterior marginals (MPM) estimator $(\hat{\xi}_i^1, \hat{\xi}_i^2) = \arg \max_{\xi_i^1, \xi_i^2} P_i(\xi_i^1, \xi_i^2)$ [16], where the feature map of each hidden neuron is combined and the prediction is thus the augmented version of the inferred feature vector in the one-bit RBM [15]. Hence, the task is to compute marginal probabilities, i.e., $P_i(\xi_i^1, \xi_i^2)$, which is intractable due to the interaction among data constraints (the product over a in Eq. (4)). However, by mapping the original model (Eq. (4)) onto a graphical model (Fig. 1), where data constraints and paired-synapses are treated respectively as factor (data) nodes and variable nodes, one can estimate the marginal probability by running a message passing iteration among factor and variable nodes, as we shall explain below. The key assumption is that the paired-synapses on the graphical model are weakly correlated, which is called the Bethe approximation [23] in physics.

We first define a cavity probability $P_{i \rightarrow a}(\xi_i^1, \xi_i^2)$ with the data node a removed. Under the weak correlation assumption, $P_{i \rightarrow a}(\xi_i^1, \xi_i^2)$ obeys a self-consistent equation:

$$P_{i \rightarrow a}(\xi_i^1, \xi_i^2) = \frac{1}{Z_{i \rightarrow a}} \prod_{b \in \partial i \setminus a} \mu_{b \rightarrow i}(\xi_i^1, \xi_i^2), \quad (5a)$$

$$\begin{aligned} \mu_{b \rightarrow i}(\xi_i^1, \xi_i^2) &= \sum_{\xi_i^1, \xi_i^2} \frac{1}{\cosh\left(\beta^2 Q_c + \frac{\beta^2}{N} \xi_i^1 \xi_i^2\right)} \cosh\left(\beta X_b + \frac{\beta}{\sqrt{N}} \xi_i^1 \sigma_i^b\right) \cosh\left(\beta Y_b + \frac{\beta}{\sqrt{N}} \xi_i^2 \sigma_i^b\right) \\ &\times \prod_{j \in \partial b \setminus i} P_{j \rightarrow b}(\xi_j^1, \xi_j^2), \end{aligned} \quad (5b)$$

where $Z_{i \rightarrow a}$ is a normalization constant, $\partial i \setminus a$ denotes neighbors of the feature node i except the data node a , $\partial b \setminus i$ denotes neighbors of the data node b except the feature node i , and the auxiliary quantity $\mu_{b \rightarrow i}(\xi_i^1, \xi_i^2)$ denotes the contribution from data node b to feature node i given the value of (ξ_i^1, ξ_i^2) [6, 15]. Products in Eq. (5) result from the weak correlation assumption. In addition, $X_b \equiv \frac{1}{\sqrt{N}} \sum_{j \neq i} \xi_j^1 \sigma_j^b$, $Y_b \equiv \frac{1}{\sqrt{N}} \sum_{j \neq i} \xi_j^2 \sigma_j^b$, and the cavity version of Q is defined as $Q_c \equiv \frac{1}{N} \sum_{j \neq i} \xi_j^1 \xi_j^2$, which can be further replaced by its typical value obtained by the average over the cavity probability (to be shown below). Although this is a crude approximation, it works quite well in practice.

Still, the above self-consistent equation is intractable due to the summation to estimate $\mu_{b \rightarrow i}$. Nevertheless, a careful inspection reveals that X_b and Y_b are approximately correlated Gaussian random variables due to the central limit theorem.

As a result, the intractable summation can be replaced by an integral which is easy to calculate in this model. We just need to compute the following mean, variance and covariance between these random variables.

$$G_{b \rightarrow i}^1 = \frac{1}{\sqrt{N}} \sum_{j \neq i} \sigma_j^b m_{j \rightarrow b}^1, \quad (6a)$$

$$G_{b \rightarrow i}^2 = \frac{1}{\sqrt{N}} \sum_{j \neq i} \sigma_j^b m_{j \rightarrow b}^2, \quad (6b)$$

$$\Gamma_{b \rightarrow i}^1 = \frac{1}{N} \sum_{j \neq i} (1 - (m_{j \rightarrow b}^1)^2), \quad (6c)$$

$$\Gamma_{b \rightarrow i}^2 = \frac{1}{N} \sum_{j \neq i} (1 - (m_{j \rightarrow b}^2)^2), \quad (6d)$$

$$\Xi_{b \rightarrow i} = \frac{1}{N} \sum_{j \neq i} (q_{j \rightarrow b} - m_{j \rightarrow b}^1 m_{j \rightarrow b}^2), \quad (6e)$$

where G and Γ denotes the mean and variance of the Gaussian random variable respectively, and the last quantity denotes the covariance between X_b and Y_b . The cavity magnetization is defined as $m_{j \rightarrow b}^{1,2} = \sum_{\xi_j^1, \xi_j^2} \xi_j^{1,2} P_{j \rightarrow b}(\xi_j^1, \xi_j^2)$, and the cavity correlation is defined as $q_{j \rightarrow b} = \sum_{\xi_j^1, \xi_j^2} \xi_j^1 \xi_j^2 P_{j \rightarrow b}(\xi_j^1, \xi_j^2)$. Finally, using the above parameters of the correlated Gaussian distribution, we rewrite $\mu_{b \rightarrow i}(\xi_i^1, \xi_i^2)$ as

$$\begin{aligned} \mu_{b \rightarrow i}(\xi_i^1, \xi_i^2) &= \frac{1}{\cosh\left(\beta^2 Q_{b \rightarrow i} + \frac{\beta^2}{N} \xi_i^1 \xi_i^2\right)} \iint Dx Dy \cosh\left(\beta \sqrt{\Gamma_{b \rightarrow i}^1} x + \beta G_{b \rightarrow i}^1 + \frac{\beta}{\sqrt{N}} \xi_i^1 \sigma_i^b\right) \\ &\times \cosh\left(\beta \sqrt{\Gamma_{b \rightarrow i}^2} (\psi x + \sqrt{1 - \psi^2} y) + \beta G_{b \rightarrow i}^2 + \frac{\beta}{\sqrt{N}} \xi_i^2 \sigma_i^b\right), \end{aligned} \quad (7)$$

where $Dx \equiv \frac{e^{-x^2/2} dx}{\sqrt{2\pi}}$, $\psi = \frac{\Xi_{b \rightarrow i}}{\sqrt{\Gamma_{b \rightarrow i}^1 \Gamma_{b \rightarrow i}^2}}$, and $Q_{b \rightarrow i} = \frac{1}{N} \sum_{j \neq i} q_{j \rightarrow b}$ stemming from Q_c in Eq. (5b) replaced by its cavity mean. The above integral representation of $\mu_{b \rightarrow i}(\xi_i^1, \xi_i^2)$ can be analytically estimated; for convenience, we define $u_{b \rightarrow i}(\xi_i^1, \xi_i^2) \equiv \ln \mu_{b \rightarrow i}(\xi_i^1, \xi_i^2)$. It is easy to show that

$$\begin{aligned} u_{b \rightarrow i}(\xi_i^1, \xi_i^2) &= \frac{\beta^2 \Gamma_{b \rightarrow i}^2 (1 - \psi^2)}{2} - \ln \left(2 \cosh \left(\beta^2 Q_{b \rightarrow i} + \frac{\beta^2 \xi_i^1 \xi_i^2}{N} \right) \right) + \frac{\beta^2}{2} \left(\sqrt{\Gamma_{b \rightarrow i}^1} + \sqrt{\Gamma_{b \rightarrow i}^2} \psi \right)^2 \\ &+ \ln \cosh \left(\beta G_{b \rightarrow i}^1 + \beta G_{b \rightarrow i}^2 + \frac{\beta}{\sqrt{N}} \sigma_i^b (\xi_i^1 + \xi_i^2) \right) \\ &+ \ln \left[1 + e^{-2\beta^2 \sqrt{\Gamma_{b \rightarrow i}^1 \Gamma_{b \rightarrow i}^2} \psi} \frac{\cosh \left(\beta G_{b \rightarrow i}^1 - \beta G_{b \rightarrow i}^2 + \frac{\beta}{\sqrt{N}} \sigma_i^b (\xi_i^1 - \xi_i^2) \right)}{\cosh \left(\beta G_{b \rightarrow i}^1 + \beta G_{b \rightarrow i}^2 + \frac{\beta}{\sqrt{N}} \sigma_i^b (\xi_i^1 + \xi_i^2) \right)} \right]. \end{aligned} \quad (8)$$

To close the iteration equation, we need to compute the cavity magnetization and

correlation as follows:

$$m_{i \rightarrow a}^1 = \frac{\sum_{\xi^1=\pm 1, \xi^2=\pm 1} \xi^1 e^{\sum_{b \in \partial i \setminus a} u_{b \rightarrow i}(\xi^1, \xi^2)}}{\sum_{\xi^1=\pm 1, \xi^2=\pm 1} e^{\sum_{b \in \partial i \setminus a} u_{b \rightarrow i}(\xi^1, \xi^2)}}, \quad (9a)$$

$$m_{i \rightarrow a}^2 = \frac{\sum_{\xi^1=\pm 1, \xi^2=\pm 1} \xi^2 e^{\sum_{b \in \partial i \setminus a} u_{b \rightarrow i}(\xi^1, \xi^2)}}{\sum_{\xi^1=\pm 1, \xi^2=\pm 1} e^{\sum_{b \in \partial i \setminus a} u_{b \rightarrow i}(\xi^1, \xi^2)}}, \quad (9b)$$

$$q_{i \rightarrow a} = \frac{\sum_{\xi^1=\pm 1, \xi^2=\pm 1} \xi^1 \xi^2 e^{\sum_{b \in \partial i \setminus a} u_{b \rightarrow i}(\xi^1, \xi^2)}}{\sum_{\xi^1=\pm 1, \xi^2=\pm 1} e^{\sum_{b \in \partial i \setminus a} u_{b \rightarrow i}(\xi^1, \xi^2)}}. \quad (9c)$$

$m_{i \rightarrow a}^{1,2}$ can be interpreted as the message passing from feature node i to data node a ($q_{i \rightarrow a}$ is also similarly interpreted), while $u_{b \rightarrow i}$ can be interpreted as the message passing from data node b to feature node i .

If the weak correlation assumption is self-consistent, starting from randomly initialized messages, the learning equations will converge to a fixed point corresponding to a thermodynamically dominant minimum of the Bethe free energy function [24], which is given by $-\beta f_{\text{Bethe}} = \frac{1}{N} \sum_i \Delta f_i - \frac{N-1}{N} \sum_a \Delta f_a$. The free energy contributions of variable node and data node are given respectively by:

$$\Delta f_i = \ln \sum_{\xi_i^1, \xi_i^2} \prod_{b \in \partial i} \mu_{b \rightarrow i}(\xi_i^1, \xi_i^2), \quad (10a)$$

$$\begin{aligned} \Delta f_a = & \frac{\beta^2 \Gamma_a^2 (1 - \tilde{\psi}^2)}{2} - \ln(2 \cosh(\beta^2 Q_a)) + \frac{\beta^2}{2} \left(\sqrt{\Gamma_a^1} + \sqrt{\Gamma_a^2} \tilde{\psi} \right)^2 \\ & + \ln \cosh(\beta G_a^1 + \beta G_a^2) + \ln \left[1 + e^{-2\beta^2 \Xi_a} \frac{\cosh(\beta G_a^1 - \beta G_a^2)}{\cosh(\beta G_a^1 + \beta G_a^2)} \right], \end{aligned} \quad (10b)$$

where $\tilde{\psi} = \frac{\Xi_a}{\sqrt{\Gamma_a^1 \Gamma_a^2}}$. The forms of $\Gamma_a^{1,2}$, $G_a^{1,2}$, Q_a and Ξ_a are similar to their cavity counterparts (e.g., in Eq. (8)), but with the only difference that the node i 's contribution is not excluded. Once the iteration converges, the MPM estimator predicts that $\hat{\xi}_i^1 = \text{sgn}(m_i^1)$ and $\hat{\xi}_i^2 = \text{sgn}(m_i^2)$, where the full (non-cavity) magnetization $m_i^{1,2}$ is computed taking into account all contributions of adjacent data nodes to the node i (see Eq. (9), and the symbol $\setminus a$ is thus removed).

2.3. Replica theory of the minimal model

To have an analytic argument about the critical threshold for spontaneous symmetry breaking, we calculate the free energy in the thermodynamic limit using the replica method. Instead of calculating a disorder average of $\ln \Omega$, the replica method computes the disorder average of an integer power of Ω , i.e., $\langle \Omega^n \rangle$. In physics, this corresponds to preparing n replicas of the original system; then the rescaled free energy density (multiplied by $-\beta$) can be obtained as [16]

$$-\beta f = \lim_{n \rightarrow 0, N \rightarrow \infty} \frac{\ln \langle \Omega^n \rangle}{nN}, \quad (11)$$

where the limits of $N \rightarrow \infty$ and $n \rightarrow 0$ have been exchanged, such that the thermodynamic limit can be taken first for applying the Laplace's method or saddle-point analysis [25], and the disorder average is taken over all possible samplings (data) and the random realizations of the true feature vector. The explicit form of $\langle \Omega^n \rangle$ reads

$$\begin{aligned} \langle \Omega^n \rangle = & \sum_{\{\boldsymbol{\sigma}^a, \boldsymbol{\xi}^{\text{true}}\}} \prod_i [P(\xi_i^{1,\text{true}}, \xi_i^{2,\text{true}})] \prod_a \frac{\cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{1,\text{true}} \cdot \boldsymbol{\sigma}^a\right) \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{2,\text{true}} \cdot \boldsymbol{\sigma}^a\right)}{2^N e^{\beta^2} \cosh(\beta^2 q)} \\ & \times \sum_{\{\boldsymbol{\xi}^{1,\gamma}, \boldsymbol{\xi}^{2,\gamma}\}} \prod_{a,\gamma} \frac{\cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{1,\gamma} \cdot \boldsymbol{\sigma}^a\right) \cosh\left(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{2,\gamma} \cdot \boldsymbol{\sigma}^a\right)}{\cosh(\beta^2 R^\gamma)}, \end{aligned} \quad (12)$$

where γ indicates the replica index, $\boldsymbol{\xi}^{\text{true}} \equiv (\boldsymbol{\xi}^{1,\text{true}}, \boldsymbol{\xi}^{2,\text{true}})$, $q \equiv \frac{1}{N} \boldsymbol{\xi}^{1,\text{true}} \cdot \boldsymbol{\xi}^{2,\text{true}}$, and $R^\gamma \equiv \frac{1}{N} \boldsymbol{\xi}^{1,\gamma} \cdot \boldsymbol{\xi}^{2,\gamma}$. Note that q is pre-determined and used to generate the random true feature maps, as also defined in section 2.1. We leave the technical details to Appendix A, and give the final result here. The free energy function reads,

$$\begin{aligned} -\beta f_{\text{RS}} = & -R\hat{R} - T_1\hat{T}_1 - \tau_1\hat{\tau}_1 - T_2\hat{T}_2 - \tau_2\hat{\tau}_2 + \frac{\hat{q}_1(q_1 - 1)}{2} + \frac{\hat{q}_2(q_2 - 1)}{2} + \frac{r\hat{r}}{2} \\ & + \alpha\beta^2 \left(1 - \frac{q_1 + q_2}{2}\right) - \alpha \ln\left(2 \cosh(\beta^2 R)\right) + \int D\mathbf{z} [\ln Z_{\text{eff}}]_{\boldsymbol{\xi}^{1,\text{true}}, \boldsymbol{\xi}^{2,\text{true}}} \quad (13) \\ & + \frac{\alpha e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh(\beta t_0) \cosh(\beta q t_0 + \beta \sqrt{1 - q^2} x_0) \ln Z_{\text{E}}, \end{aligned}$$

where $[\cdot]_{\boldsymbol{\xi}^{1,\text{true}}, \boldsymbol{\xi}^{2,\text{true}}}$ means an average w.r.t $P(\boldsymbol{\xi}^{1,\text{true}}, \boldsymbol{\xi}^{2,\text{true}})$, $D\mathbf{z} \equiv Dz_1 Dz_2 Dz_3$ (a standard Gaussian measure vector, as defined below Eq. (7)), and similarly $D\mathbf{t} \equiv Dt_0 Dx_0 Du Du'$. RS means the replica symmetry assumption we used to get the final result. This assumption implies that the order parameter (various kinds of overlaps, explicitly defined below) does not rely on its specific replica index. We assume that this assumption is able to describe the system as we shall show it leads to consistent predictions verified in algorithmic results of single instances. The auxiliary quantities Z_{eff} and Z_{E} are defined as follows,

$$Z_{\text{eff}} = \sum_{\xi^1, \xi^2} e^{b_1 \xi^1 + b_2 \xi^2 + b_3 \xi^1 \xi^2}, \quad (14a)$$

$$b_1 = \hat{T}_1 \xi^{1,\text{true}} + \hat{\tau}_2 \xi^{2,\text{true}} + \sqrt{\hat{q}_1 - \hat{r}/2} z_1 + \sqrt{\hat{r}/2} z_3, \quad (14b)$$

$$b_2 = \hat{T}_2 \xi^{2,\text{true}} + \hat{\tau}_1 \xi^{1,\text{true}} + \sqrt{\hat{q}_2 - \hat{r}/2} z_2 + \sqrt{\hat{r}/2} z_3, \quad (14c)$$

$$b_3 = \hat{R} - \hat{r}/2, \quad (14d)$$

$$Z_{\text{E}} = e^{\beta^2(R-r)} \cosh(\beta \Lambda_+) + e^{-\beta^2(R-r)} \cosh(\beta \Lambda_-), \quad (14e)$$

$$\Lambda_+ = (T_1 + \tau_1)t_0 + \frac{1}{\sqrt{1 - q^2}}(T_2 + \tau_2 - q(T_1 + \tau_1))x_0 + (B + \frac{r - A}{B})u + Ku', \quad (14f)$$

$$\Lambda_- = (T_1 - \tau_1)t_0 + \frac{1}{\sqrt{1 - q^2}}(\tau_2 - T_2 - q(T_1 - \tau_1))x_0 + (B - \frac{r - A}{B})u - Ku', \quad (14g)$$

where $A \equiv T_1 \tau_1 + \frac{(\tau_2 - T_1 q)(T_2 - \tau_1 q)}{1 - q^2}$, $B \equiv \sqrt{q_1 - T_1^2 - \frac{(\tau_2 - T_1 q)^2}{1 - q^2}}$, and $K \equiv \sqrt{q_2 - \tau_1^2 - \frac{(T_2 - \tau_1 q)^2}{1 - q^2} - \frac{(r - A)^2}{B^2}}$.

The associated (non-conjugated) saddle-point equations are expressed as

$$T_1 = [\xi^{1,\text{true}} \langle \xi^1 \rangle]_{\mathbf{z}, \xi^{1,\text{true}}, \xi^{2,\text{true}}} , \quad (15a)$$

$$T_2 = [\xi^{2,\text{true}} \langle \xi^2 \rangle]_{\mathbf{z}, \xi^{1,\text{true}}, \xi^{2,\text{true}}} , \quad (15b)$$

$$q_1 = [\langle \xi^1 \rangle^2]_{\mathbf{z}, \xi^{1,\text{true}}, \xi^{2,\text{true}}} , \quad (15c)$$

$$q_2 = [\langle \xi^2 \rangle^2]_{\mathbf{z}, \xi^{1,\text{true}}, \xi^{2,\text{true}}} , \quad (15d)$$

$$\tau_1 = [\xi^{1,\text{true}} \langle \xi^2 \rangle]_{\mathbf{z}, \xi^{1,\text{true}}, \xi^{2,\text{true}}} , \quad (15e)$$

$$\tau_2 = [\xi^{2,\text{true}} \langle \xi^1 \rangle]_{\mathbf{z}, \xi^{1,\text{true}}, \xi^{2,\text{true}}} , \quad (15f)$$

$$R = [\langle \xi^1 \xi^2 \rangle]_{\mathbf{z}, \xi^{1,\text{true}}, \xi^{2,\text{true}}} , \quad (15g)$$

$$r = [\langle \xi^1 \rangle \langle \xi^2 \rangle]_{\mathbf{z}, \xi^{1,\text{true}}, \xi^{2,\text{true}}} . \quad (15h)$$

Note that the average w.r.t the true features can be written explicitly by definition as $P(\xi^{1,\text{true}}, \xi^{2,\text{true}}) = \frac{p_d}{2}$ for both true components taking different values, and otherwise $P(\xi^{1,\text{true}}, \xi^{2,\text{true}}) = \frac{1 - p_d}{2}$. p_d is related to q by $p_d = \frac{1 - q}{2}$. The outer average also includes the disorder average over \mathbf{z} . The inner average $\langle \bullet \rangle$ indicates the thermal average under the partition function Z_{eff} (corresponding to a two-spin interaction Hamiltonian). This average is analytically tractable, e.g., $\langle \xi^1 \rangle = \frac{1}{Z_{\text{eff}}} \frac{\partial Z_{\text{eff}}}{\partial b_1} = \frac{\tanh b_1 + \tanh b_2 \tanh b_3}{1 + \tanh b_1 \tanh b_2 \tanh b_3}$. $\langle \xi^2 \rangle$ and $\langle \xi^1 \xi^2 \rangle$ can also be similarly computed.

We further comment that T_1 characterizes the typical overlap between inferred value of the first feature and its true counterpart, and likewise T_2 characterizes the typical overlap between the second feature and its ground truth; q_1 and q_2 characterize the sizes of the first and second feature spaces respectively; R characterizes the correlation of the two features within the same replica, while r is the correlation for different replicas; τ_1 characterizes the typical correlation between the first true feature and the inferred value of the second feature in an arbitrary replica, and likewise τ_2 characterizes the typical correlation between the second true feature and the inferred value of the first feature in an arbitrary replica. τ_1 and τ_2 are thus responsible for the permutation symmetry effect. Taken all together, $(T_1, T_2, q_1, q_2, R, r, \tau_1, \tau_2)$ forms the order parameter set of our model. Their exact mathematical definitions are given in the Appendix A.

Finally, the conjugated order parameters can also be derived from a saddle point

analysis of the free energy function, and they obey the following equations:

$$\hat{T}_1 = \alpha\beta^2 \langle \langle G_s^+ \rangle \rangle, \quad (16a)$$

$$\hat{T}_2 = \alpha\beta^2 \langle \langle \langle G_s^- \rangle \rangle \rangle, \quad (16b)$$

$$\hat{q}_1 = \alpha\beta^2 \langle (G_s^+)^2 \rangle, \quad (16c)$$

$$\hat{q}_2 = \alpha\beta^2 \langle (G_s^-)^2 \rangle, \quad (16d)$$

$$\hat{\tau}_1 = \alpha\beta^2 \langle \langle G_s^- \rangle \rangle, \quad (16e)$$

$$\hat{\tau}_2 = \alpha\beta^2 \langle \langle \langle G_s^+ \rangle \rangle \rangle, \quad (16f)$$

$$\hat{R} = \alpha\beta^2 \langle G_c^- \rangle - \alpha\beta^2 \tanh(\beta^2 R), \quad (16g)$$

$$\hat{r} = 2\alpha\beta^2 \langle G_s^+ G_s^- \rangle, \quad (16h)$$

where the average $\langle \bullet \rangle \equiv \frac{e^{-\beta^2}}{\cosh \beta^2 q} \int D\mathbf{t} \cosh(\beta t_0) \cosh(\beta q t_0 + \beta \sqrt{1-q^2} x_0) \bullet$, $\langle \langle \bullet \rangle \rangle \equiv \frac{e^{-\beta^2}}{\cosh \beta^2 q} \int D\mathbf{t} \sinh(\beta t_0) \cosh(\beta q t_0 + \beta \sqrt{1-q^2} x_0) \bullet$, and $\langle \langle \langle \bullet \rangle \rangle \rangle \equiv \frac{e^{-\beta^2}}{\cosh \beta^2 q} \int D\mathbf{t} \cosh(\beta t_0) \sinh(\beta q t_0 + \beta \sqrt{1-q^2} x_0) \bullet$. The auxiliary quantities are defined as follows,

$$G_c^- = \frac{e^{\beta^2(R-r)} \cosh \beta \Lambda_+ - e^{-\beta^2(R-r)} \cosh \beta \Lambda_-}{e^{\beta^2(R-r)} \cosh \beta \Lambda_+ + e^{-\beta^2(R-r)} \cosh \beta \Lambda_-}, \quad (17a)$$

$$G_s^+ = \frac{e^{\beta^2(R-r)} \sinh \beta \Lambda_+ + e^{-\beta^2(R-r)} \sinh \beta \Lambda_-}{e^{\beta^2(R-r)} \cosh \beta \Lambda_+ + e^{-\beta^2(R-r)} \cosh \beta \Lambda_-}, \quad (17b)$$

$$G_s^- = \frac{e^{\beta^2(R-r)} \sinh \beta \Lambda_+ - e^{-\beta^2(R-r)} \sinh \beta \Lambda_-}{e^{\beta^2(R-r)} \cosh \beta \Lambda_+ + e^{-\beta^2(R-r)} \cosh \beta \Lambda_-}. \quad (17c)$$

To sum up, Eqs. (15) and (16) construct a closed iterative equation (detailed derivations are given in the Appendix B), whose fixed point gives an approximate evaluation of the free energy. When all order parameters vanish (a trivial disordered state), the free energy has an analytic value expressed as $\alpha\beta^2 + \ln 4$ in agreement with $-\beta f_{\text{Bethe}}$ in the same trivial state. In addition, these saddle point equations in the case of $q = 0$ can be simplified to the result of Ref. [16] for unsupervised feature learning in a one-bit RBM (see the Appendix C). More precisely, when the true feature maps are orthogonal, we have $\Omega = \Omega_{\text{one-bit-RBM}}^2$, thus the free energy is two times as large as that of one-bit RBM.

Meanwhile, the converged order parameters from Eqs. (15) and (16), especially T_1 and T_2 can be compared with the algorithmic results, and can also be used to analytically derive the critical threshold α_c for unsupervised learning in this permutation-symmetry model. When the data size is not sufficient, we expect that the order parameters vanish, and in the small order-parameter limit, $T_1 \simeq \hat{T}_1 + q\hat{\tau}_2$, $\tau_2 \simeq \hat{\tau}_2 + \hat{T}_1 q$, $\hat{T}_1 \simeq \alpha\beta^4 [T_1 + \tau_2 \tanh(\beta^2 q)]$, and $\hat{\tau}_2 \simeq \alpha\beta^4 [T_1 \tanh(\beta^2 q) + \tau_2]$. Based on these four equations, it is easy to show that the critical learning threshold is given by

$$\alpha_c = \frac{\beta^{-4}}{1 + q \tanh(\beta^2 q) + |\tanh(\beta^2 q) + q|}. \quad (18)$$

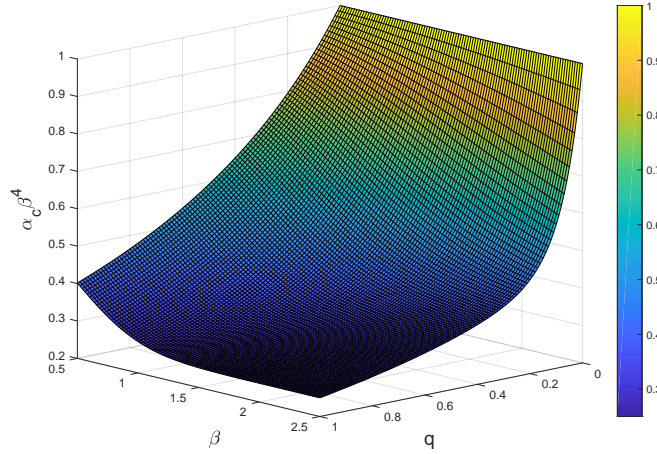


Figure 2. (Color online) The critical value of data size (Eq. (18)) for learning in our minimal model, as a function of the noise level β as well as the correlation level q .

As expected, the critical value for our current model does not depend on the sign of q . A detailed derivation of α_c is given in the Appendix D. In the correlation-free limit $q = 0$, the known result of the one-bit RBM, $\alpha_c = \beta^{-4}$ [15,16] is recovered. Thus, we first theoretically prove the conjecture made by empirical observations in Ref. [18] that once the number of hidden neurons is finite (not proportional to N), the critical learning threshold does not change with this finite number! However, Ref. [18] overlooks the effect of the potential correlation across true hidden features, which indeed affects the learning threshold. Remarkably, this effect, more natural than the ideal correlation-free case, is clearly captured by our theory (Eq. (18)). We show this effect in Fig. 2. We can see that for a fixed noise level, increasing q has the effect of decreasing the critical data size. That is, if the data set is created by strongly correlated receptive fields (feature maps), then the learning is relatively easy, or fewer samples are sufficient to trigger the phase transition of concept-formation. Conversely, if q is zero, the data is generated from independent feature maps, then a successful learning requires a much more larger dataset. We also observe that for a large β (less noisy data), a small correlation level q can already significantly reduce the minimal data size that triggers learning, as verified by the fact that around the small q region the larger the β is, the sharper the surface becomes.

Next, we analyze two interesting limits implied by the critical threshold equation (Eq. (18)). In the limit $|q| \rightarrow 1$, $\alpha_c \rightarrow \frac{1}{4}\beta^{-4}$ provided that β is relatively large such that $\tanh \beta^2 \simeq 1$. The second case is another limit $|q| \rightarrow 0$, i.e., q takes a small value but not zero, implying that a weak correlation among feature maps is maintained. Depending on the order of magnitude of q , we have the following result given a relatively large β :

$$\lim_{\beta \rightarrow \infty} \alpha_c \beta^4 = \begin{cases} 1 & \text{if } |q| \ll \beta^{-2}, \\ \frac{1}{1+|\tanh q_0|} & \text{if } q = q_0 \beta^{-2} \text{ or } |q| \sim \beta^{-2}, \\ \frac{1}{2(1+|q|)} & \text{if } |q| \gg \beta^{-2}. \end{cases} \quad (19)$$

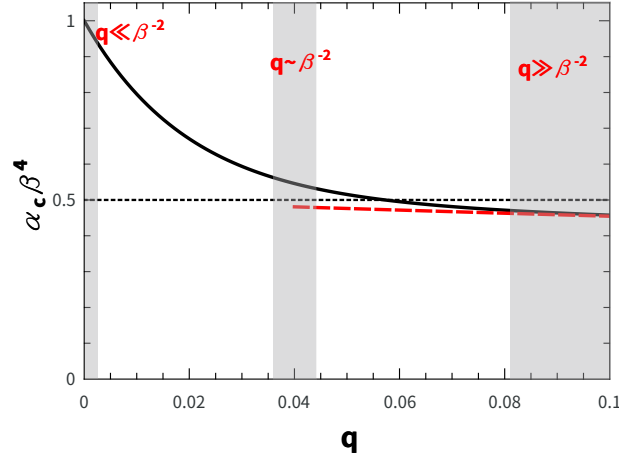


Figure 3. (Color online) The critical value of data size (Eq. (18)) as a function of the noise level q . We consider the weak-feature-correlation limit at different orders of magnitude compared with β^{-2} . We use the value of $\beta = 5$ for an example. The dashed line indicates the third case of Eq. (19).

Note that ∞ means any large value of β such that $\tanh \beta \simeq 1$ rather than a definite value of infinity. Eq. (19) reveals that once the two feature maps are weakly correlated, the minimal learning data size for a transition can be further (or even significantly) reduced compared to the correlation-free case, particularly in the case that q is not very small but still larger than the order of magnitude set by β^{-2} . We show this result in Fig. 3.

We thereby have a *significant hypothesis* for the triggering of concept-formation that a bit large (compared with β^{-2}) yet still small value of the correlation level is highly favored for unsupervised learning from a dataset of smaller size (compared with the correlation-free case). Regularization techniques such as locally enforcing feature orthogonality [26] and dropping some connections during training [27] have been introduced to deep learning. Weakly-correlated receptive fields are also favored from the perspective of neural computation, since the redundancy among synaptic weights is reduced and thus different feature detectors inside the network can encode efficiently stimuli features rather than capturing noise in the data. A similar decorrelation in hidden activities was recently theoretically analyzed in feedforward neural networks [28]. We hope our theoretical prediction can be verified in specific machine learning tasks, and even in neuroscience experiments where the relationship amongst the minimal data size for learning, the correlation level of synapses (or receptive fields) and the noise level in stimuli can be jointly established. Therefore, from the Bayesian learning perspective, the correlated-feature-map case yields a much lower threshold of phase transition towards the concept formation, in comparison with the correlation-free case [15, 16, 18].

Overall, the prediction quantitatively captures the learning behavior in both uncorrelated and correlated settings. In next section, we shall further verify this

conclusion by extensive numerical simulations on single instances of the minimal model.

3. Results and discussion

In this section, we study how the permutation symmetry between two hidden neurons affects the learning process, i.e., the spontaneous symmetry breaking transition of the concept-formation during unsupervised learning. We focus on whether the replica theory predicts the learning threshold related to the continuous transition. The learning threshold can be estimated from the message passing algorithmic results on single instances of the minimal model. We first randomly generate true feature maps with a pre-determined correlation level specified by q . The true feature maps are then used to generate M Monte-Carlo samples through a Gibbs sampling procedure [15] according to Eq. (1). Finally, these samples are used as a quenched disorder for the Bayesian inference of the true feature maps [16]. An overlap with the ground truth is computed and compared with the replica prediction. In fact, for comparison, we define the overlap [16], e.g., $T_1^{\text{MP}} = \frac{1}{N} \sum_i \xi_i^{\text{true}} m_i^1$ where MP means message passing, and m_i^1 takes into account the thermal average (uncertainty about the ground truth). Other overlaps can be similarly defined.

First, we compare the free energy function estimated under the Bethe approximation with that predicted by the replica theory (in the thermodynamic limit). As shown in Fig. 4 (a), the algorithmic results on finite-sized networks coincide very well with the theoretical predictions. This implies that, the approximation we used to derive the message passing equation for learning a RBM with two hidden units is reasonable, especially when the number of visible neurons is large.

Then, we study the evolution of the overlap with the ground truth as a function of the data density. As shown in Fig. 4 (b), we clearly see a continuous phase transition separating a disordered (symmetric) phase from an ordered (symmetry-broken) phase. This is consistent with previous works revealing the spontaneous symmetry breaking in unsupervised feature learning [15,16]. When the amount of provided data samples is not too large, the original model maintains its symmetry (i.e., equal probabilities for positive and negative assignments of synapses respectively). This is in an isotropic phase which does not capture any concept from the data samples. However, the increasing amount of data samples will break this symmetry through a continuous phase transition towards a non-trivial concept formation. During this process, in a practical message passing procedure, messages flowing in the factor graph are biased towards the true feature maps underlying the noisy data. Remarkably, the theory predicts the exact location of the phase transition point at which the messages running on a single instance start to polarize towards the true feature map. As predicted and observed, the learning threshold indeed decreases as the absolute value of q grows. For $|q| = 0.3$, $\alpha_c \simeq 0.596$; and for $|q| = 0.8$, α_c is significantly reduced to be about 0.334.

Due to the permutation symmetry, we find that after the SSB transition, there appear in sequence three non-trivial solutions to the saddle point equations (Eqs. (15)

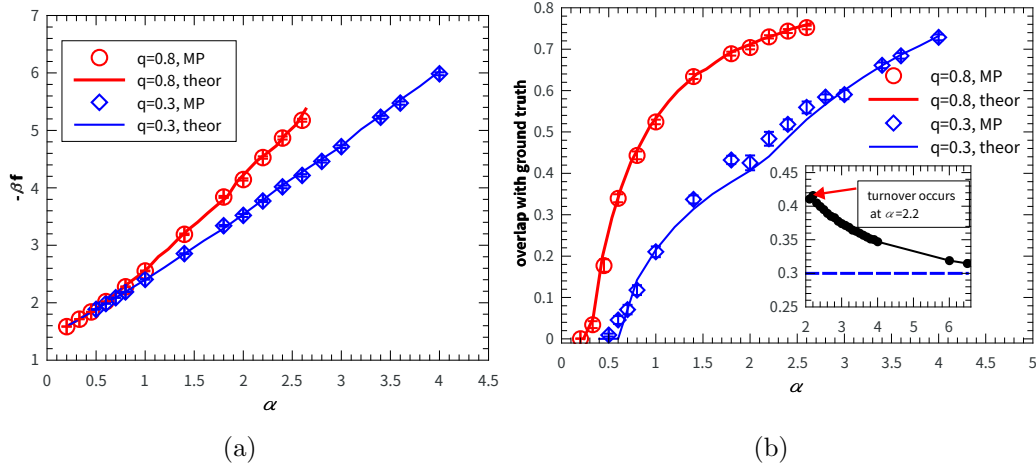


Figure 4. (Color online) Learning performances of the minimal model obtained from message passing (MP) algorithmic results in comparison with theory. We consider $\beta = 1$ with different values of q . MP is run on single instances of the minimal model with $N = 200$. The error bars characterize the standard deviation across different random realizations of the minimal model. (a) Rescaled free energy per neuron as a function of data density (data samples per neuron). (b) The overlap with the ground truth versus the data density. In the inset, we show the replica prediction of the permutation-type overlap (τ_1 or τ_2) obtained by an exchange of ξ^1 and ξ^2 in T_1 or T_2 for $q = 0.3$, provided that $(\xi^{1,\text{true}}, \xi^{2,\text{true}})$ is the planted feature. As expected, when the number of data samples is sufficiently large, this overlap tends to the embedded correlation q (indicated by the dashed line).

and (16)), as α increases for a given value of (β, q) . The first solution has the form of $q_1 = q_2 = r$ and $T_1 = T_2 = \tau_1 = \tau_2$. This solution is caused by the permutation symmetry and is dominant at the earlier stage of the post-SSB-transition. $q_1 = q_2 = r$ implies the permutation symmetry between inferred vectors of ξ^1 and ξ^2 ; and $T_1 = T_2 = \tau_1 = \tau_2$ implies the permutation symmetry between planted true feature vectors. In other words, after the SSB transition, the unsupervised learning increases the overlap with the ground truth through identifying the common components of the two true feature maps.

The SSB phase is stable until a point where the unsupervised learning starts to predict the different components (as shown in Fig. 5 (a)), thereby breaking the symmetry between the inferred values of ξ^1 and ξ^2 . This point is thus referred to as $\alpha_c^{\text{PSB}_s}$, namely, the permutation symmetry breaking (PSB) for student/inferred features. However, the inferred features are equally likely to be either $(\xi^{1,\text{true}}, \xi^{2,\text{true}})$ or its permutation, thus the permutation symmetry between planted/teacher features (i.e., $\xi^{1,\text{true}}$ and $\xi^{2,\text{true}}$) is still preserved, therefore, after the PSB_s transition, we have the second solution: $q_1 = q_2 \neq r$, and $T_1 = T_2 = \tau_1 = \tau_2$. This second solution is stable until there is a turnover of the trend of the permutation-type overlap or the order parameter r (as shown in the insets of Fig. 4 (b) and Fig. 5 (a) respectively). At this turnover, r starts to decrease, thereby reducing the permutation-type overlap towards the true

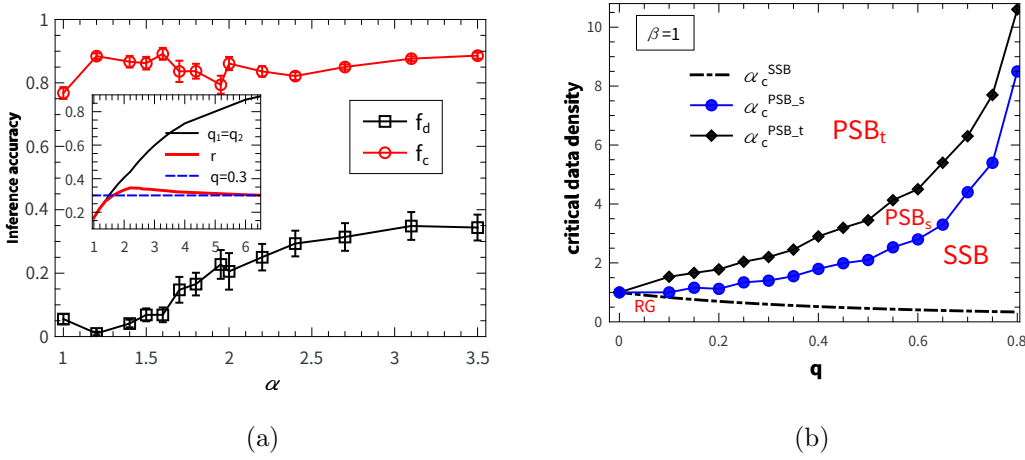


Figure 5. (Color online) Phase diagram of unsupervised learning. (a) Learning success probabilities on common feature components (f_c) and distinct components (f_d) show that the permutation symmetry can be spontaneously broken as predicted by the replica theory (the inset of Fig. 4 (b)). Learning success of one component implies that the inferred value of that component matches with the true one. The inset shows the replica result of q_1 (or q_2) and r versus the data density. MP is run on single instances of the minimal model with $N = 200$, $\beta = 1$ and $q = 0.3$. The error bars characterize the standard deviation across different random realizations of the minimal model. (b) Three phases (random-guess (RG), spontaneous symmetry breaking (SSB) and permutation symmetry breaking (PSB including PSB_s for inferred/student features and PSB_t for planted/teacher features)) are separated by three curves— $\alpha_c^{SSB}(q)$, $\alpha_c^{PSB_s}(q)$ and $\alpha_c^{PSB_t}(q)$. $\alpha_c^{PSB_s}$ can only be determined by numerically computing the deviation between q_1 (or q_2) and r , and $\alpha_c^{PSB_t}$ is indicated by the point where r starts to decrease (or the difference between T_1 (or T_2) and τ_1 (or τ_2) starts to appear). We consider $\beta = 1$ with different values of q .

correlation level of the two feature maps (Fig. 4 (b)). After this transition indicated by the turnover, even the permutation symmetry between the teacher’s features is broken, as the unsupervised learning is able to distinguish the two feature maps underlying the raw data samples. Therefore, we observe the final solution that has the form of either $q_1 = q_2 = T_1 = T_2$ or $q_1 = q_2 = \tau_1 = \tau_2$ (two sub-forms), which is dominant at the later stage of the unsupervised learning. Accordingly, this second PSB transition is called PSB_t transition with broken symmetry for teacher’s features.

We remark that this third type non-trivial solution can be deduced from the fact that as α grows, the inferred feature map gets close to the true feature map and thus both the inferred and true feature maps follow the same posterior probability of the learning process. Note that for this solution, these two sub-forms share the same free energy. One inferred feature map has the freedom of choosing to match the first or second true feature map. This choice does not change the overall free energy. Therefore, at the later stage, one solution of a larger overlap corresponds to the case that the inferred feature maps are matched with their true counterparts; the other smaller overlap corresponds to the case that the inferred feature maps are matched with their interchanged (permuted)

counterparts. In particular, the latter one should converge to the embedded correlation q as the learning converges to the true feature maps. We show one example in the case of $q = 0.3$ (see the inset of Fig. 4 (b)), which is consistent with the replica prediction (the inset of Fig. 5 (a)). Note that these two kinds of overlaps are very hard to distinguish at the earlier stage, since the inferred feature maps are only partially correct compared with their true counterparts. Indeed, in the algorithmic simulations, we observe two solutions of different overlaps (one is larger than the other), which are very easy to distinguish at the later stage of the learning. In addition, considering the sign of each inferred (or true) feature vector, we have other multiple types of solutions, since a sign-reversed assignment of the feature map does not change the posterior probability (Eq. (4)). For simplicity, we do not consider the sign-reversed case in the analysis.

Permutation symmetry of different types can be spontaneously broken. Both transitions are continuous. We summarize these qualitatively different transitions in Fig. 5 (b), showing that as q grows, the gap between α_c^{SSB} and α_c^{PSB} increases, thereby demonstrating that the small- q region is beneficial not only in the sense of a significant decrease of data size triggering the concept-formation but also in the sense of a small data size triggering the permutation symmetry breaking. Therefore, not only can increasing the number of data samples (observations) drive a spontaneous symmetry breaking, but increasing the observations also drive a permutation symmetry breaking which finally leads to a perfect reconstruction of the embedded feature maps.

In addition, when one tries to run the message passing from a perturbed true feature map, it is observed that its basin is quite large, i.e., it is stable in the presence of a sufficiently large dataset. But if the perturbation is large enough (e.g., crossing the sign boundary), the passing message is able to converge to other types of fixed points (e.g., the sign-reversed or permutation-symmetric ones).

4. Summary

In conclusion, we propose a minimal model of the permutation symmetry in a simple unsupervised learning system— a simple RBM with only two hidden units. Using statistical mechanics tools developed in theory of disordered systems, we reveal very rich properties of this model. Effects of permutation symmetry among hidden units have been studied in supervised learning in multilayer neural networks [29,30]. Here we are the first to consider its effects on unsupervised learning, focusing on spontaneous symmetry breaking and the critical data size that triggers the transition.

First, we derive an efficient algorithm to infer the embedded feature vectors from a given dataset (Monte-Carlo samples in this paper), according to the Bayesian principle. This algorithm is fully distributed, in that the computation is implemented in terms of local passing messages among feature nodes and data (constraint) nodes. The algorithmic results can be used as a test of the replica theory in the thermodynamic limit.

Second, the behavior of the algorithm, especially the critical data size at which

a continuous phase transition from a disordered (symmetric) phase to an ordered (symmetry-broken) phase occurs, can be predicted by a replica theory of the minimal model. Due to the existence of two hidden units, we have to manipulate eight order parameters that characterize all possible typical correlations in the replica space itself as well as between replica space and true-feature space. In addition to the spontaneous symmetry breaking, our theory predicts another later transition where the permutation symmetry is spontaneously broken and thus the identity of the feature map can be captured during unsupervised learning. *By studying the minimal model, we are able to interpret the unsupervised learning, a fundamental process governing artificial and biological intelligence, as a progressive combination of SSB and PSB (including two types—PSB_s and PSB_t), both of which are driven by increasing the data size (or observations).*

Based on the replica computation, we have three contributions. (i) We analytically prove that the critical data size for the phase transition indeed does not depend on the number of hidden units (a finite number case), and this conclusion was empirically observed in a previous work [18]. The critical value $\alpha_c = \beta^{-4}$ for learning in a one-bit RBM [15, 16] can directly apply to the RBM with two hidden units, once no correlations are embedded into the two feature vectors of hidden units. A detailed proof is also given in the Appendix C. (ii) In addition, we reveal the correlation level of embedded true feature vectors reduces the critical data size, characterized by a simple formula (see Eq. (18)). As an example, in the very large β limit (small variability in the data space), the critical data size for $q = -1$ is reduced to only one-fourth of that in the correlation-free case. In another limit $q \rightarrow 0$, depending on the order of magnitude of q compared with β^{-2} , the necessary data size that triggers the transition can be reduced to one half of β^{-4} or even less, although the receptive fields are weakly-correlated. This prediction qualitatively coincides with the observation that humans or non-human animals do not need a large amount of data samples to learn a concept from structured examples (e.g. natural images are highly structured as various levels of correlations are embedded) [31]. We expect that this quantitative prediction of the critical learning data threshold can be shown to hold in a generic case with arbitrary levels of receptive-field-correlation, and with an arbitrary finite number of hidden units, in particular for generative models of neural networks. (iii) Our theory predicts that an additional spontaneous permutation symmetry breaking follows the spontaneous symmetry breaking, leading to another benefit of the small- q regime where the data size triggering the permutation symmetry breaking is smaller compared with the large- q regime (Fig. 5).

Our study also encourages several interesting future directions. First, by analogy with a Bayesian iteration derived in our previous work [16] to predict the true noise level (the hyper-parameter β) in a dataset, one can also derive the iteration equation for predicting both correlation level q and noise level β for an arbitrary dataset. Second, one can also verify the predicted value of α_c in a practical neural network architecture with inferred q and β . This can be carried out in a more complex RBM, by comparing the predicted value of the critical data size with the observed one for a successful

unsupervised learning task. Furthermore, the prediction of our theory about the benefit of weakly-correlated feature maps compared with the correlation-free situation can also be tested in practical artificial neural networks and even in biological neural networks. Finally, it would be very interesting to verify in a more general setting whether the progressive combination of SSB and PSB is the underlying mechanism of unsupervised learning. In this sense, our minimal model paves the way towards understanding the fundamentally important unsupervised learning process, by addressing the role of the permutation symmetry and moreover the critical learning data sizes (related to both SSB and PSB) in a simple unsupervised learning system.

Acknowledgments

We would like to thank referees for their inspiring comments. This research was supported by the start-up budget 74130-18831109 of the 100-talent- program of Sun Yat-sen University (H.H.), the NSFC (Grant No. 11805284) (H.H.), and grants from the Research Grants Council of Hong Kong (16322616 and 16396817) (T.H. and M.W.)

Appendix A. Computation of $\langle \Omega^n \rangle$

Here, we show details to compute $\langle \Omega^n \rangle$, which is defined as:

$$\begin{aligned} \langle \Omega^n \rangle = & \sum_{\{\boldsymbol{\xi}^{true}, \boldsymbol{\sigma}^a\}} \prod_{i=1}^N [P(\xi_i^{1,true}, \xi_i^{2,true})] \prod_{a=1}^M \frac{\cosh(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{1,true} \boldsymbol{\sigma}^a) \cosh(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{2,true} \boldsymbol{\sigma}^a)}{2^N e^{\beta^2} \cosh(\beta^2 q)} \\ & \times \sum_{\{\boldsymbol{\xi}^{1,\gamma}, \boldsymbol{\xi}^{2,\gamma}\}} \prod_{a,\gamma} \frac{\cosh(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{1,\gamma} \boldsymbol{\sigma}^a) \cosh(\frac{\beta}{\sqrt{N}} \boldsymbol{\xi}^{2,\gamma} \boldsymbol{\sigma}^a)}{\cosh(\beta^2 R^\gamma)}, \end{aligned} \quad (\text{A.1})$$

where γ indicates the replica index, $\boldsymbol{\xi}^{true} = \{\boldsymbol{\xi}^{1,true}, \boldsymbol{\xi}^{2,true}\}$, the overlap $q = \frac{1}{N} \boldsymbol{\xi}^{1,true} \boldsymbol{\xi}^{2,true}$, and $R^\gamma = \frac{1}{N} \boldsymbol{\xi}^{1,\gamma} \boldsymbol{\xi}^{2,\gamma}$.

To further calculate $\langle \Omega^n \rangle$, we need to define the order parameters as follows:

$$T_1^\gamma = \frac{1}{N} \boldsymbol{\xi}^{1,true} \boldsymbol{\xi}^{1,\gamma}, \quad T_2^\gamma = \frac{1}{N} \boldsymbol{\xi}^{2,true} \boldsymbol{\xi}^{2,\gamma}, \quad (\text{A.2a})$$

$$\tau_1^\gamma = \frac{1}{N} \boldsymbol{\xi}^{1,true} \boldsymbol{\xi}^{2,\gamma}, \quad \tau_2^\gamma = \frac{1}{N} \boldsymbol{\xi}^{2,true} \boldsymbol{\xi}^{1,\gamma}, \quad (\text{A.2b})$$

$$q_1^{\gamma,\gamma'} = \frac{1}{N} \boldsymbol{\xi}^{1,\gamma} \boldsymbol{\xi}^{1,\gamma'}, \quad q_2^{\gamma,\gamma'} = \frac{1}{N} \boldsymbol{\xi}^{2,\gamma} \boldsymbol{\xi}^{2,\gamma'}, \quad (\text{A.2c})$$

$$R^\gamma = \frac{1}{N} \boldsymbol{\xi}^{1,\gamma} \boldsymbol{\xi}^{2,\gamma}, \quad r^{\gamma,\gamma'} = \frac{1}{N} \boldsymbol{\xi}^{1,\gamma} \boldsymbol{\xi}^{2,\gamma'}. \quad (\text{A.2d})$$

Note that these order parameters can be used to evaluate the disorder average in

Eq. (A.1). We then proceed as follows,

$$\begin{aligned}
\langle \Omega^n \rangle &= \sum_{\{\sigma^a, \xi^{true}\}} \prod_{i=1}^N P(\xi_i^{1,true}, \xi_i^{2,true}) \sum_{\{\xi^{1,\gamma}, \xi^{2,\gamma}\}} \int \prod_{\gamma=1}^n dR^\gamma \delta(\xi^{1,\gamma} \xi^{2,\gamma} - NR^\gamma) \\
&\times \int \prod_{\gamma=1}^n dT_1^\gamma \delta(\xi^{1,true} \xi^{1,\gamma} - NT_1^\gamma) \int \prod_{\gamma=1}^n dT_2^\gamma \delta(\xi^{2,true} \xi^{2,\gamma} - NT_2^\gamma) \int \prod_{\gamma=1}^n d\tau_1^\gamma \delta(\xi^{1,true} \xi^{2,\gamma} - N\tau_1^\gamma) \\
&\times \int \prod_{\gamma=1}^n d\tau_2^\gamma \delta(\xi^{2,true} \xi^{1,\gamma} - N\tau_2^\gamma) \int \prod_{\gamma < \gamma'} d q_1^{\gamma, \gamma'} \delta(\xi^{1,\gamma} \xi^{1, \gamma'} - N q_1^{\gamma, \gamma'}) \int \prod_{\gamma < \gamma'} d q_2^{\gamma, \gamma'} \delta(\xi^{2,\gamma} \xi^{2, \gamma'} - N q_2^{\gamma, \gamma'}) \\
&\times \int \prod_{\gamma < \gamma'} d r^{\gamma, \gamma'} \delta(\xi^{1,\gamma} \xi^{2, \gamma'} - N r^{\gamma, \gamma'}) \prod_{a=1}^M \left\{ \frac{\cosh(\beta X_a^0) \cosh(\beta Y_a^0)}{2^N e^{\beta^2} \cosh(\beta^2 q)} \prod_{\gamma=1}^n \frac{\cosh(\beta X_a^\gamma) \cosh(\beta Y_a^\gamma)}{\cosh(\beta^2 R^\gamma)} \right\} \\
&= \sum_{\{\sigma^a, \xi^{true}\}} \prod_{i=1}^N P(\xi_i^{1,true}, \xi_i^{2,true}) \sum_{\{\xi^{1,\gamma}, \xi^{2,\gamma}\}} \int \prod_{\gamma=1}^n \left(\frac{dR^\gamma d\hat{R}^\gamma}{2\pi} \right) \int \prod_{\gamma=1}^n \left(\frac{dT_1^\gamma d\hat{T}_1^\gamma}{2\pi} \right) \int \prod_{\gamma=1}^n \left(\frac{dT_2^\gamma d\hat{T}_2^\gamma}{2\pi} \right) \\
&\times \int \prod_{\gamma=1}^n \left(\frac{d\tau_1^\gamma d\hat{\tau}_1^\gamma}{2\pi} \right) \int \prod_{\gamma=1}^n \left(\frac{d\tau_2^\gamma d\hat{\tau}_2^\gamma}{2\pi} \right) \int \prod_{\gamma < \gamma'} \left(\frac{d q_1^{\gamma, \gamma'} d\hat{q}_1^{\gamma, \gamma'}}{2\pi} \right) \int \prod_{\gamma < \gamma'} \left(\frac{d q_2^{\gamma, \gamma'} d\hat{q}_2^{\gamma, \gamma'} d r^{\gamma, \gamma'} d\hat{r}^{\gamma, \gamma'}}{4\pi^2} \right) \\
&\times \exp \left(\sum_{\gamma=1}^n i \hat{R}^\gamma (\xi^{1,\gamma} \xi^{2,\gamma} - NR^\gamma) + \sum_{\gamma=1}^n i \hat{T}_1^\gamma (\xi^{1,\gamma} \xi^{1,true} - NT_1^\gamma) + \sum_{\gamma=1}^n i \hat{T}_2^\gamma (\xi^{2,\gamma} \xi^{2,true} - NT_2^\gamma) \right) \\
&\times \exp \left(\sum_{\gamma=1}^n i \hat{\tau}_1^\gamma (\xi^{1,true} \xi^{2,\gamma} - N\tau_1^\gamma) + \sum_{\gamma=1}^n i \hat{\tau}_2^\gamma (\xi^{2,true} \xi^{1,\gamma} - N\tau_2^\gamma) + \sum_{\gamma < \gamma'} \hat{q}_1^{\gamma, \gamma'} (\xi^{1,\gamma} \xi^{1, \gamma'} - N q_1^{\gamma, \gamma'}) \right) \\
&\times \exp \left(\sum_{\gamma < \gamma'} \hat{q}_2^{\gamma, \gamma'} (\xi^{2,\gamma} \xi^{2, \gamma'} - N q_2^{\gamma, \gamma'}) + \sum_{\gamma < \gamma'} \hat{r}^{\gamma, \gamma'} (\xi^{1,\gamma} \xi^{2, \gamma'} - N r^{\gamma, \gamma'}) \right) \\
&\times \prod_{a=1}^M \left\{ \frac{\cosh(\beta X_a^0) \cosh(\beta Y_a^0)}{2^N e^{\beta^2} \cosh(\beta^2 q)} \prod_{\gamma=1}^n \frac{\cosh(\beta X_a^\gamma) \cosh(\beta Y_a^\gamma)}{\cosh(\beta^2 R^\gamma)} \right\}, \tag{A.3}
\end{aligned}$$

where we have defined $X_a^0 = \frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i^{1,true} \sigma_i^a$, $Y_a^0 = \frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i^{2,true} \sigma_i^a$, and $X_a^\gamma = \frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i^{1,\gamma} \sigma_i^a$, $Y_a^\gamma = \frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i^{2,\gamma} \sigma_i^a$. To get the second equality, we have used the integral representation of the delta function $\delta(x) = \int \frac{d\hat{x}}{2\pi} e^{i\hat{x}x}$.

To compute the free energy value, we assume a simple ansatz, i.e., all order parameters do not depend on their specific replica indexes, which is called the replica-symmetry assumption. To be more precise, we assume

$$R^\gamma = R, \quad i \hat{R}^\gamma = \hat{R}, \tag{A.4a}$$

$$T_1^\gamma = T_1, \quad i \hat{T}_1^\gamma = \hat{T}_1, \tag{A.4b}$$

$$T_2^\gamma = T_2, \quad i \hat{T}_2^\gamma = \hat{T}_2, \tag{A.4c}$$

$$\tau_1^\gamma = \tau_1, \quad i \hat{\tau}_1^\gamma = \hat{\tau}_1, \tag{A.4d}$$

$$\tau_2^\gamma = \tau_2, \quad i \hat{\tau}_2^\gamma = \hat{\tau}_2, \tag{A.4e}$$

for any γ . We also assume that

$$q_1^{\gamma,\gamma'} = q_1, \quad i\hat{q}_1^{\gamma,\gamma'} = \hat{q}_1, \quad (\text{A.5a})$$

$$q_2^{\gamma,\gamma'} = q_2, \quad i\hat{q}_2^{\gamma,\gamma'} = \hat{q}_2, \quad (\text{A.5b})$$

$$r^{\gamma,\gamma'} = r, \quad i\hat{r}^{\gamma,\gamma'} = \hat{r}, \quad (\text{A.5c})$$

for any γ and γ' . Then we can express $\langle \Omega^n \rangle$ as :

$$\langle \Omega^n \rangle = \int d\mathcal{O} d\hat{\mathcal{O}} e^{N\mathcal{A}(\mathcal{O}, \hat{\mathcal{O}}, \alpha, \beta, n)}. \quad (\text{A.6})$$

In the thermodynamics limit, $\langle \Omega^n \rangle$ can be approximated as $e^{N\mathcal{A}(\mathcal{O}_*, \hat{\mathcal{O}}_*, \alpha, \beta, n)}$ (namely, the saddle-point method), where \mathcal{O}_* and $\hat{\mathcal{O}}_*$ represent all non-conjugated order parameters and conjugated order parameters evaluated at the maximal value of the action, respectively. The expression for the action $\mathcal{A}(\mathcal{O}, \hat{\mathcal{O}}, \alpha, \beta, n)$ can be written by

$$\begin{aligned} \mathcal{A} = & -nR\hat{R} - nT_1\hat{T}_1 - nT_2\hat{T}_2 - n\tau_1\hat{\tau}_1 - n\tau_2\hat{\tau}_2 - \frac{n(n-1)}{2}q_1\hat{q}_1 \\ & - \frac{n(n-1)}{2}q_2\hat{q}_2 - \frac{n(n-1)}{2}r\hat{r} + G_S + \alpha G_E, \end{aligned} \quad (\text{A.7})$$

where G_S is the entropy term, and G_E is the energy term.

To derive the entropy term G_S , we use the following identities [32]:

$$\sum_{\gamma < \gamma'} \xi^{1,\gamma} \xi^{1,\gamma'} = \frac{1}{2} \left(\sum_{\gamma} \xi^{1,\gamma} \right)^2 - \frac{1}{2} \sum_{\gamma} (\xi^{1,\gamma})^2, \quad (\text{A.8a})$$

$$\sum_{\gamma < \gamma'} \xi^{2,\gamma} \xi^{2,\gamma'} = \frac{1}{2} \left(\sum_{\gamma} \xi^{2,\gamma} \right)^2 - \frac{1}{2} \sum_{\gamma} (\xi^{2,\gamma})^2, \quad (\text{A.8b})$$

$$\begin{aligned} \sum_{\gamma < \gamma'} \xi^{1,\gamma} \xi^{2,\gamma'} &= \frac{1}{2} \sum_{\gamma, \gamma'} \xi^{1,\gamma} \xi^{2,\gamma'} - \frac{1}{2} \sum_{\gamma} \xi^{1,\gamma} \xi^{2,\gamma} \\ &= \frac{1}{4} \left(\sum_{\gamma} \xi^{1,\gamma} + \sum_{\gamma'} \xi^{2,\gamma'} \right)^2 - \frac{1}{4} \left(\sum_{\gamma} \xi^{1,\gamma} \right)^2 - \frac{1}{4} \left(\sum_{\gamma'} \xi^{2,\gamma'} \right)^2 - \frac{1}{2} \sum_{\gamma} \xi^{1,\gamma} \xi^{2,\gamma}. \end{aligned} \quad (\text{A.8c})$$

The above non-linear terms can be reduced to linear terms in the exponential functions of Eq. (A.6) by the Hubbard-Stratonovich transformation $\int Dte^{bt} = e^{\frac{1}{2}b^2}$. Then, we

obtain G_S as :

$$\begin{aligned}
G_S &= \ln \left[\sum_{\{\xi^1, \gamma, \xi^2, \gamma\}} \exp \left(\hat{R} \sum_{\gamma=1}^n \xi^1, \gamma \xi^2, \gamma + \hat{T}_1 \sum_{\gamma=1}^n \xi^1, \gamma \xi^{1, \text{true}} + \hat{T}_2 \sum_{\gamma=1}^n \xi^2, \gamma \xi^{2, \text{true}} + \hat{\tau}_1 \sum_{\gamma=1}^n \xi^{1, \text{true}} \xi^{2, \gamma} \right) \right. \\
&\quad \times \exp \left(\hat{\tau}_2 \sum_{\gamma=1}^n \xi^1, \gamma \xi^{2, \text{true}} + \hat{q}_1 \sum_{\gamma < \gamma'} \xi^1, \gamma \xi^{1, \gamma'} + \hat{q}_2 \sum_{\gamma < \gamma'} \xi^2, \gamma \xi^{2, \gamma'} + \hat{r} \sum_{\gamma < \gamma'} \xi^1, \gamma \xi^{2, \gamma'} \right) \left. \right]_{\xi^{1, \text{true}}, \xi^{2, \text{true}}} \\
&= \ln \left[\sum_{\{\xi^1, \gamma, \xi^2, \gamma\}} \exp \left(\frac{(\hat{q}_1 - \frac{\hat{r}}{2})}{2} \left(\sum_{\gamma} \xi^1, \gamma \right)^2 + \frac{(\hat{q}_2 - \frac{\hat{r}}{2})}{2} \left(\sum_{\gamma} \xi^2, \gamma \right)^2 + \hat{T}_1 \sum_{\gamma} \xi^1, \gamma \xi^{1, \text{true}} \right) \right. \\
&\quad \times \exp \left(\frac{\hat{r}}{4} \left(\sum_{\gamma} \xi^1, \gamma + \sum_{\gamma'} \xi^{2, \gamma'} \right)^2 + \hat{T}_2 \sum_{\gamma} \xi^2, \gamma \xi^{2, \text{true}} + \left(\hat{R} - \frac{\hat{r}}{2} \right) \sum_{\gamma} \xi^1, \gamma \xi^{2, \gamma} \right) \\
&\quad \times \exp \left(\hat{\tau}_1 \sum_{\gamma} \xi^{1, \text{true}} \xi^{2, \gamma} + \hat{\tau}_2 \sum_{\gamma} \xi^{2, \text{true}} \xi^{1, \gamma} - \frac{n}{2} \hat{q}_1 - \frac{n}{2} \hat{q}_2 \right) \left. \right]_{\xi^{1, \text{true}}, \xi^{2, \text{true}}} \\
&= \ln \left[\sum_{\{\xi^1, \gamma, \xi^2, \gamma\}} \int D\mathbf{z} \exp \left(\sum_{\gamma} \sqrt{\hat{q}_1 - \frac{\hat{r}}{2}} \xi^1, \gamma z_1 + \sum_{\gamma} \sqrt{\hat{q}_2 - \frac{\hat{r}}{2}} \xi^2, \gamma z_2 + \sqrt{\frac{\hat{r}}{2}} z_3 \left(\sum_{\gamma} \xi^1, \gamma + \sum_{\gamma'} \xi^{2, \gamma'} \right) \right) \right. \\
&\quad \times \exp \left(\hat{T}_1 \sum_{\gamma} \xi^{1, \text{true}} \xi^{1, \gamma} + \hat{T}_2 \sum_{\gamma} \xi^{2, \gamma} \xi^{2, \text{true}} + \hat{\tau}_1 \sum_{\gamma} \xi^{1, \text{true}} \xi^{2, \gamma} \right) \\
&\quad \times \exp \left(\hat{\tau}_2 \sum_{\gamma} \xi^{2, \text{true}} \xi^{1, \gamma} + \left(\hat{R} - \frac{\hat{r}}{2} \right) \sum_{\gamma} \xi^1, \gamma \xi^{2, \gamma} - \frac{n}{2} \hat{q}_1 - \frac{n}{2} \hat{q}_2 \right) \left. \right]_{\xi^{1, \text{true}}, \xi^{2, \text{true}}} .
\end{aligned} \tag{A.9}$$

Finally, we can express the entropy term G_S in a compact form as

$$G_S = \ln \left[\int D\mathbf{z} \left(\sum_{\xi^1, \xi^2} e^{b_1 \xi^1 + b_2 \xi^2 + b_3 \xi^1 \xi^2} \right)^n \right]_{\xi^{1, \text{true}}, \xi^{2, \text{true}}} - \frac{n}{2} \hat{q}_1 - \frac{n}{2} \hat{q}_2, \tag{A.10}$$

where we have defined $D\mathbf{z} = Dz_1 Dz_2 Dz_3$, and the auxiliary variables b_1, b_2 , and b_3 as

$$b_1 = \sqrt{\hat{q}_1 - \frac{\hat{r}}{2}} z_1 + \sqrt{\frac{\hat{r}}{2}} z_3 + \hat{T}_1 \xi^{1, \text{true}} + \hat{\tau}_2 \xi^{2, \text{true}}, \tag{A.11a}$$

$$b_2 = \sqrt{\hat{q}_2 - \frac{\hat{r}}{2}} z_2 + \sqrt{\frac{\hat{r}}{2}} z_3 + \hat{T}_2 \xi^{2, \text{true}} + \hat{\tau}_1 \xi^{1, \text{true}}, \tag{A.11b}$$

$$b_3 = \hat{R} - \frac{\hat{r}}{2}. \tag{A.11c}$$

We remark that in the expression of G_S , the inner summation over ξ^1, ξ^2 can be thought as a two-spin interaction partition function, which is defined as Z_{eff} in the main text.

$[\bullet]_{\xi^{1,true}, \xi^{2,true}}$ means an average w.r.t $P(\xi^{1,true}, \xi^{2,true})$ which is also defined in the main text.

Next, we turn to compute the energy term G_E . The expression of G_E is given by

$$G_E = \ln \left\langle \frac{\cosh(\beta X^0) \cosh(\beta Y^0)}{\cosh(\beta^2 q)} \prod_{\gamma=1}^n \frac{\cosh(\beta X^\gamma) \cosh(\beta Y^\gamma)}{\cosh(\beta^2 R^\gamma)} \right\rangle, \quad (\text{A.12})$$

where $\langle \bullet \rangle$ defines the disorder average. $X^0, Y^0, X^\gamma, Y^\gamma$ are correlated Gaussian random variables, which are the same as before but data index a has been dropped off. They have zero mean and unit variance. Their covariances are determined by the aforementioned order parameters as follows:

$$\langle X^0 Y^0 \rangle = q, \quad \langle X^0 X^\gamma \rangle = T_1, \quad \langle X^0 Y^\gamma \rangle = \tau_1, \quad (\text{A.13a})$$

$$\langle X^\gamma X^{\gamma'} \rangle = q_1, \quad \langle Y^\gamma Y^{\gamma'} \rangle = q_2, \quad \langle X^\gamma Y^\gamma \rangle = R, \quad (\text{A.13b})$$

$$\langle Y^0 Y^\gamma \rangle = T_2, \quad \langle Y^0 X^\gamma \rangle = \tau_2, \quad \langle X^\gamma Y^{\gamma'} \rangle = r. \quad (\text{A.13c})$$

The random variables $X^0, Y^0, X^\gamma, Y^\gamma$ can thus be parameterized by six standard Gaussian variables of zero mean and unit variance $(t_0, x_0, u, u', y_\gamma, \omega_\gamma)$ as follows,

$$X^0 = t_0, \quad (\text{A.14a})$$

$$Y^0 = q t_0 + \sqrt{1 - q^2} x_0, \quad (\text{A.14b})$$

$$X^\gamma = T_1 t_0 + \frac{\tau_2 - T_1 q}{\sqrt{1 - q^2}} x_0 + B u + \sqrt{1 - q_1} \omega_\gamma, \quad (\text{A.14c})$$

$$Y^\gamma = \tau_1 t_0 + \frac{T_2 - \tau_1 q}{\sqrt{1 - q^2}} x_0 + \frac{r - A}{B} u + \frac{R - r}{\sqrt{1 - q_1}} \omega_\gamma + K u' + \sqrt{1 - q_2 - \frac{(R - r)^2}{1 - q_1}} y_\gamma, \quad (\text{A.14d})$$

where $A = T_1 \tau_1 + \frac{(\tau_2 - T_1 q)(T_2 - \tau_1 q)}{1 - q^2}$, $B = \sqrt{q_1 - (T_1)^2 - \frac{(\tau_2 - T_1 q)^2}{1 - q^2}}$, and $K = \sqrt{q_2 - (\tau_1)^2 - \frac{(T_2 - \tau_1 q)^2}{1 - q^2} - \left(\frac{r - A}{B}\right)^2}$. One can easily verify that the above parameterization satisfies their covariance structures. Therefore, the G_E term can be calculated by a standard Gaussian integration given by

$$G_E = \ln \left[\int D t_0 D x_0 D u D u' \frac{\cosh(\beta t_0) \cosh \beta(q t_0 + \sqrt{1 - q^2} x_0)}{\cosh(\beta^2 q)} \right. \\ \times \left(\int D \omega D y \frac{1}{\cosh(\beta^2 R)} \cosh \beta \left(T_1 t_0 + \frac{\tau_2 - T_1 q}{\sqrt{1 - q^2}} x_0 + B u + \sqrt{1 - q_1} \omega \right) \right. \\ \left. \times \cosh \beta \left(\tau_1 t_0 + \frac{T_2 - \tau_1 q}{\sqrt{1 - q^2}} x_0 + \frac{r - A}{B} u + \frac{R - r}{\sqrt{1 - q_1}} \omega + K u' + C y \right) \right)^n \left. \right], \quad (\text{A.15})$$

where $C \equiv \sqrt{1 - q_2 - \frac{(R-r)^2}{1-q_1}}$. To proceed, we first define the auxiliary quantities as :

$$\Lambda_+ = (T_1 + \tau_1)t_0 + \frac{(T_2 + \tau_2) - q(T_1 + \tau_1)}{\sqrt{1 - q^2}}x_0 + (B + \frac{r - A}{B})u + Ku', \quad (\text{A.16a})$$

$$\Lambda_- = (T_1 - \tau_1)t_0 + \frac{(\tau_2 - T_2) - q(T_1 - \tau_1)}{\sqrt{1 - q^2}}x_0 + (B - \frac{r - A}{B})u - Ku'. \quad (\text{A.16b})$$

Then we compute the integral inside the power n defined by I whose result is given by

$$\begin{aligned} I &\equiv \int D\omega Dy \left[\cosh \beta(\tau_1 t_0 + \frac{T_2 - \tau_1 q}{\sqrt{1 - q^2}}x_0 + \frac{r - A}{B}u + \frac{R - r}{\sqrt{1 - q_1}}\omega + Ku' + \sqrt{1 - q_2 - \frac{(R - r)^2}{1 - q_1}}y) \right. \\ &\quad \times \cosh \beta(T_1 t_0 + \frac{\tau_2 - T_1 q}{\sqrt{1 - q^2}}x_0 + Bu + \sqrt{1 - q_1}\omega) \left. \right] \\ &= \frac{1}{4} \int D\omega Dy \left[e^{\beta\{\Lambda_+ + (\sqrt{1 - q_1} + \frac{R - r}{\sqrt{1 - q_1}})\omega + \sqrt{1 - q_2 - \frac{(R - r)^2}{1 - q_1}}y\}} + e^{-\beta\{\Lambda_+ + (\sqrt{1 - q_1} + \frac{R - r}{\sqrt{1 - q_1}})\omega + \sqrt{1 - q_2 - \frac{(R - r)^2}{1 - q_1}}y\}} \right. \\ &\quad \left. + e^{\beta\{\Lambda_- + (\sqrt{1 - q_1} - \frac{R - r}{\sqrt{1 - q_1}})\omega - \sqrt{1 - q_2 - \frac{(R - r)^2}{1 - q_1}}y\}} + e^{-\beta\{\Lambda_- + (\sqrt{1 - q_1} - \frac{R - r}{\sqrt{1 - q_1}})\omega - \sqrt{1 - q_2 - \frac{(R - r)^2}{1 - q_1}}y\}} \right] \\ &= \frac{1}{2} e^{\beta^2(1 - \frac{q_1 + q_2}{2})} \left[e^{\beta^2(R - r)} \cosh(\beta\Lambda_+) + e^{-\beta^2(R - r)} \cosh(\beta\Lambda_-) \right]. \end{aligned} \quad (\text{A.17})$$

For simplicity, we also define the following auxiliary quantities Z_E, G_c^-, G_s^+, G_s^- :

$$\begin{aligned} Z_E &= e^{\beta^2(R - r)} \cosh(\beta\Lambda_+) + e^{-\beta^2(R - r)} \cosh(\beta\Lambda_-), \\ G_c^- &= \frac{e^{\beta^2(R - r)} \cosh(\beta\Lambda_+) - e^{-\beta^2(R - r)} \cosh(\beta\Lambda_-)}{e^{\beta^2(R - r)} \cosh(\beta\Lambda_+) + e^{-\beta^2(R - r)} \cosh(\beta\Lambda_-)}, \\ G_s^+ &= \frac{e^{\beta^2(R - r)} \sinh(\beta\Lambda_+) + e^{-\beta^2(R - r)} \sinh(\beta\Lambda_-)}{e^{\beta^2(R - r)} \cosh(\beta\Lambda_+) + e^{-\beta^2(R - r)} \cosh(\beta\Lambda_-)}, \\ G_s^- &= \frac{e^{\beta^2(R - r)} \sinh(\beta\Lambda_+) - e^{-\beta^2(R - r)} \sinh(\beta\Lambda_-)}{e^{\beta^2(R - r)} \cosh(\beta\Lambda_+) + e^{-\beta^2(R - r)} \cosh(\beta\Lambda_-)}. \end{aligned} \quad (\text{A.18})$$

Following the replica trick, we can get:

$$\lim_{n \rightarrow 0} \frac{G_E}{n} = \frac{\int Dt_0 Dx_0 Du Du' \frac{\cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1 - q^2}x_0)}{\cosh \beta^2 q} \ln \left[\frac{I}{\cosh \beta^2 R} \right]}{\int Dt_0 Dx_0 Du Du' \frac{\cosh(\beta t_0) \cosh \beta(qt_0 + \sqrt{1 - q^2}x_0)}{\cosh(\beta^2 q)}}, \quad (\text{A.19})$$

where the integral in the denominator can be exactly computed with the result given by

$$\begin{aligned} &\int Dt_0 Dx_0 Du Du' \cosh(\beta t_0) \cosh \beta(qt_0 + \sqrt{1 - q^2}x_0) \\ &= \frac{1}{2} \left(e^{\frac{\beta^2}{2}(1 - q)^2 + \frac{\beta^2}{2}(1 - q^2)} + e^{\frac{\beta^2}{2}(1 + q)^2 + \frac{\beta^2}{2}(1 - q^2)} \right) = e^{\beta^2} \cosh \beta^2 q. \end{aligned} \quad (\text{A.20})$$

Finally, by collecting all the above relevant terms, we have the following estimation of $\langle \Omega^n \rangle$ given by

$$\begin{aligned} \langle \Omega^n \rangle = & \int d\mathcal{O} d\hat{\mathcal{O}} \exp \left(-NnR\hat{R} - NnT_1\hat{T}_1 - NnT_2\hat{T}_2 - Nn\tau_2\hat{\tau}_2 - \frac{N}{2}n(n-1)q_1\hat{q}_1 \right) \\ & \times \exp \left(-\frac{N}{2}n(n-1)q_2\hat{q}_2 - \frac{N}{2}n(n-1)r\hat{r} - \frac{nN}{2}\hat{q}_1 - \frac{nN}{2}\hat{q}_2 + N \ln \left[\int D\mathbf{z} Z_{\text{eff}}^n \right]_{\xi^{1,true}, \xi^{2,true}} \right. \\ & \left. + \alpha N \ln \left\{ \int D\mathbf{t} \frac{\cosh(\beta t_0) \cosh \beta(qt_0 + \sqrt{1-q^2}x_0)}{\cosh(\beta^2 q)} \left[\frac{I}{\cosh(\beta^2 R)} \right]^n \right\} \right), \end{aligned} \quad (\text{A.21})$$

where in shorthand $D\mathbf{t} = Dt_0 Dx_0 Du Du'$. By computing $\lim_{n \rightarrow 0} \frac{\ln \langle \Omega^n \rangle}{n}$ and using Eq. (A.19), we get the expression $F_\beta = -\beta f_{RS}$ as

$$\begin{aligned} F_\beta = & -R\hat{R} - T_1\hat{T}_1 - T_2\hat{T}_2 - \tau_1\hat{\tau}_1 - \tau_2\hat{\tau}_2 + \frac{\hat{q}_1}{2}(q_1 - 1) + \frac{\hat{q}_2}{2}(q_2 - 1) \\ & + \frac{r\hat{r}}{2} + \int D\mathbf{z} [\ln Z_{\text{eff}}]_{\xi^{1,true}, \xi^{2,true}} - \alpha \ln(2 \cosh(\beta^2 R)) + \alpha \beta^2 \left(1 - \frac{q_1 + q_2}{2} \right) \\ & + \frac{\alpha e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) \ln Z_E. \end{aligned} \quad (\text{A.22})$$

Note that we have used $\lim_{n \rightarrow 0} \frac{\ln[\int D\mathbf{z} Z_{\text{eff}}^n]_{\xi^{1,true}, \xi^{2,true}}}{n} = \int D\mathbf{z} [\ln Z_{\text{eff}}]_{\xi^{1,true}, \xi^{2,true}}$ to arrive at the final expression.

Appendix B. Derivation of saddle-point equations

By the saddle-point analysis, these non-conjugated order parameters \mathcal{O} should obey the following stationary conditions:

$$\frac{\partial F_\beta}{\partial R} = 0, \quad \frac{\partial F_\beta}{\partial r} = 0, \quad \frac{\partial F_\beta}{\partial q_1} = 0, \quad \frac{\partial F_\beta}{\partial q_2} = 0, \quad (\text{B.1a})$$

$$\frac{\partial F_\beta}{\partial T_1} = 0, \quad \frac{\partial F_\beta}{\partial T_2} = 0, \quad \frac{\partial F_\beta}{\partial \tau_1} = 0, \quad \frac{\partial F_\beta}{\partial \tau_2} = 0. \quad (\text{B.1b})$$

Similarly, for conjugated order parameters $\hat{\mathcal{O}}$, the following stationary conditions should be satisfied as:

$$\frac{\partial F_\beta}{\partial \hat{R}} = 0, \quad \frac{\partial F_\beta}{\partial \hat{r}} = 0, \quad \frac{\partial F_\beta}{\partial \hat{T}_1} = 0, \quad \frac{\partial F_\beta}{\partial \hat{T}_2} = 0, \quad (\text{B.2a})$$

$$\frac{\partial F_\beta}{\partial \hat{q}_1} = 0, \quad \frac{\partial F_\beta}{\partial \hat{q}_2} = 0, \quad \frac{\partial F_\beta}{\partial \hat{\tau}_1} = 0, \quad \frac{\partial F_\beta}{\partial \hat{\tau}_2} = 0. \quad (\text{B.2b})$$

We first evaluate the self-consistent equations those non-conjugated order parameters obey. For R , we have the following equation as

$$\frac{\partial F_\beta}{\partial R} = -R + \left[\int D\mathbf{z} \frac{\partial \ln Z_{\text{eff}}}{\partial R} \right]_{\xi^{1,true}, \xi^{2,true}} = 0. \quad (\text{B.3})$$

Thus the saddle-point equation of R is given by

$$R = [\langle \xi^1 \xi^2 \rangle]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}, \quad (\text{B.4})$$

where the thermal average $\langle \bullet \rangle$ is computed under the partition function Z_{eff} (a two-spin interaction partition function), and the outer average indicates the disorder average over Gaussian random variables \mathbf{z} and the distribution $P(\xi^{1,true}, \xi^{2,true})$.

Similarly, for the order parameter T_1 , we have the following equation as

$$\frac{\partial F_\beta}{\partial \hat{T}_1} = -T_1 + \int D\mathbf{z} \left[\frac{1}{Z_{\text{eff}}} \frac{\partial Z_{\text{eff}}}{\partial \hat{T}_1} \right]_{\xi^{1,true}, \xi^{2,true}} = 0. \quad (\text{B.5})$$

Noting that $\frac{\partial Z_{\text{eff}}}{\partial \hat{T}_1} = \sum_{\xi^1, \xi^2} \xi^{1,true} \xi^1 e^{b_1 \xi^1 + b_2 \xi^2 + b_3 \xi^1 \xi^2}$, we get the the final expression of T_1 as

$$T_1 = [\langle \xi^1 \rangle \xi^{1,true}]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}. \quad (\text{B.6})$$

The expressions of T_2, τ_1 and τ_2 can be derived in the same way as follows:

$$T_2 = [\langle \xi^2 \rangle \xi^{2,true}]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}, \quad (\text{B.7})$$

$$\tau_1 = [\langle \xi^2 \rangle \xi^{1,true}]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}, \quad (\text{B.8})$$

$$\tau_2 = [\langle \xi^1 \rangle \xi^{2,true}]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}. \quad (\text{B.9})$$

Next, we turn to the saddle-point equation of q_1 , i.e.,

$$\frac{\partial F_\beta}{\partial \hat{q}_1} = \frac{1}{2}(q_1 - 1) + \int D\mathbf{z} \left[\frac{1}{Z_{\text{eff}}} \frac{\partial Z_{\text{eff}}}{\partial \hat{q}_1} \right]_{\xi^{1,true}, \xi^{2,true}} = 0. \quad (\text{B.10})$$

Noticing that $\frac{\partial Z_{\text{eff}}}{\partial \hat{q}_1} = \frac{1}{2}(\hat{q}_1 - \frac{\hat{r}}{2})^{-\frac{1}{2}} \sum_{\xi^1, \xi^2} \xi^1 z_1 e^{b_1 \xi^1 + b_2 \xi^2 + b_3 \xi^1 \xi^2}$, we get the expression of q_1 as

$$q_1 - 1 + (\hat{q}_1 - \frac{\hat{r}}{2})^{-\frac{1}{2}} [\langle \xi^1 \rangle z_1]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}} = 0. \quad (\text{B.11})$$

To proceed, we use the following identity

$$\int Dz f(z) z = \int Dz f'(z), \quad (\text{B.12})$$

where $f(z)$ is any differentiable function of z . Thus we have the following equality as

$$[\langle \xi^1 \rangle z_1]_{\mathbf{z}} = \left[\frac{\partial}{\partial z_1} \left(\frac{\sum_{\xi^1, \xi^2} \xi^1 e^{b_1 \xi^1 + b_2 \xi^2 + b_3 \xi^1 \xi^2}}{Z_{\text{eff}}} \right) \right]_{\mathbf{z}} = \sqrt{\hat{q}_1 - \frac{\hat{r}}{2}} [1 - \langle \xi^1 \rangle^2]_{\mathbf{z}}. \quad (\text{B.13})$$

Finally, the expression of q_1 is given by

$$q_1 = [\langle \xi^1 \rangle^2]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}. \quad (\text{B.14})$$

Similarly, q_2 should obey the following equation given by

$$q_2 = [\langle \xi^2 \rangle^2]_{\mathbf{z}, \xi^1, true, \xi^2, true}. \quad (\text{B.15})$$

Following the same spirit, we get the following stationary condition for r as

$$\frac{r}{2} + \int D\mathbf{z} \left[\frac{\partial}{\partial \hat{r}} \ln Z_{eff} \right]_{\xi^1, true, \xi^2, true} = 0. \quad (\text{B.16})$$

Note that

$$\begin{aligned} \frac{\partial}{\partial \hat{r}} \ln Z_{eff} = & -\frac{1}{4}(\hat{q}_1 - \frac{\hat{r}}{2})^{-\frac{1}{2}} \langle \xi^1 \rangle z_1 + \frac{1}{4} \left(\frac{\hat{r}}{2} \right)^{-\frac{1}{2}} \langle \xi^1 \rangle z_3 \\ & - \frac{1}{4}(\hat{q}_2 - \frac{\hat{r}}{2})^{-\frac{1}{2}} \langle \xi^2 \rangle z_2 + \frac{1}{4} \left(\frac{\hat{r}}{2} \right)^{-\frac{1}{2}} \langle \xi^2 \rangle z_3 - \frac{1}{2} \langle \xi^1 \xi^2 \rangle. \end{aligned} \quad (\text{B.17})$$

By applying Eq. (B.12), we can obtain the following three identities as

$$\begin{aligned} [\langle \xi^2 \rangle z_2]_{\mathbf{z}} &= \sqrt{\hat{q}_2 - \frac{\hat{r}}{2}} (1 - [\langle \xi^2 \rangle^2]_{\mathbf{z}}), \\ [\langle \xi^1 \rangle z_3]_{\mathbf{z}} &= \sqrt{\frac{\hat{r}}{2}} (1 - [\langle \xi^1 \rangle^2]_{\mathbf{z}} + [\langle \xi^1 \xi^2 \rangle]_{\mathbf{z}} - [\langle \xi^1 \rangle \langle \xi^2 \rangle]_{\mathbf{z}}), \\ [\langle \xi^2 \rangle z_3]_{\mathbf{z}} &= \sqrt{\frac{\hat{r}}{2}} (1 - [\langle \xi^2 \rangle^2]_{\mathbf{z}} + [\langle \xi^1 \xi^2 \rangle]_{\mathbf{z}} - [\langle \xi^1 \rangle \langle \xi^2 \rangle]_{\mathbf{z}}). \end{aligned} \quad (\text{B.18})$$

Using the above three identities together with Eq. (B.13), we get the expression of the saddle-point equation for r as follows

$$r = [\langle \xi^1 \rangle \langle \xi^2 \rangle]_{\mathbf{z}, \xi^1, true, \xi^2, true}. \quad (\text{B.19})$$

Given the result that $Z_{eff} = 2e^{b_3} \cosh(b_1 + b_2) + 2e^{-b_3} \cosh(b_1 - b_2)$, the thermal average like $\langle \xi^1 \rangle$, $\langle \xi^2 \rangle$, and $\langle \xi^1 \xi^2 \rangle$ can be easily calculated as follows:

$$\begin{aligned} \langle \xi^1 \xi^2 \rangle_{Z_{eff}} &= \frac{\partial}{\partial b_3} \ln Z_{eff} \\ &= \frac{e^{b_3} \cosh(b_1 + b_2) - e^{-b_3} \cosh(b_1 - b_2)}{e^{b_3} \cosh(b_1 + b_2) + e^{-b_3} \cosh(b_1 - b_2)} \\ &= \frac{e^{b_3} (\cosh b_1 \cosh b_2 + \sinh b_1 \sinh b_2) - e^{-b_3} (\cosh b_1 \cosh b_2 - \sinh b_1 \sinh b_2)}{e^{b_3} (\cosh b_1 \cosh b_2 + \sinh b_1 \sinh b_2) + e^{-b_3} (\cosh b_1 \cosh b_2 - \sinh b_1 \sinh b_2)}, \\ &= \frac{\sinh b_3 \cosh b_1 \cosh b_2 + \cosh b_3 \sinh b_1 \sinh b_2}{\cosh b_3 \cosh b_1 \cosh b_2 + \sinh b_3 \sinh b_1 \sinh b_2} \\ &= \frac{\tanh b_3 + \tanh b_1 \tanh b_2}{1 + \tanh b_1 \tanh b_2 \tanh b_3}, \end{aligned} \quad (\text{B.20})$$

and

$$\begin{aligned}
\langle \xi^1 \rangle_{Z_{eff}} &= \frac{\partial}{\partial b_1} \ln Z_{eff} \\
&= \frac{e^{b_3} \sinh(b_1 + b_2) + e^{-b_3} \sinh(b_1 - b_2)}{e^{b_3} \cosh(b_1 + b_2) + e^{-b_3} \cosh(b_1 - b_2)} \\
&= \frac{e^{b_3} (\sinh b_1 \cosh b_2 + \cosh b_1 \sinh b_2) + e^{-b_3} (\sinh b_1 \cosh b_2 - \cosh b_1 \sinh b_2)}{e^{b_3} (\cosh b_1 \cosh b_2 + \sinh b_1 \sinh b_2) + e^{-b_3} (\cosh b_1 \cosh b_2 - \sinh b_1 \sinh b_2)} \\
&= \frac{\cosh b_3 \sinh b_1 \cosh b_2 + \sinh b_3 \cosh b_1 \sinh b_2}{\cosh b_3 \cosh b_1 \cosh b_2 + \sinh b_3 \sinh b_1 \sinh b_2} \\
&= \frac{\tanh b_1 + \tanh b_2 \tanh b_3}{1 + \tanh b_1 \tanh b_2 \tanh b_3},
\end{aligned} \tag{B.21}$$

and finally

$$\begin{aligned}
\langle \xi^2 \rangle_{Z_{eff}} &= \frac{\partial}{\partial b_2} \ln Z_{eff} \\
&= \frac{e^{b_3} \sinh(b_1 + b_2) - e^{-b_3} \sinh(b_1 - b_2)}{e^{b_3} \cosh(b_1 + b_2) + e^{-b_3} \cosh(b_1 - b_2)} \\
&= \frac{e^{b_3} (\sinh b_1 \cosh b_2 + \cosh b_1 \sinh b_2) - e^{-b_3} (\sinh b_1 \cosh b_2 - \cosh b_1 \sinh b_2)}{e^{b_3} (\cosh b_1 \cosh b_2 + \sinh b_1 \sinh b_2) + e^{-b_3} (\cosh b_1 \cosh b_2 - \sinh b_1 \sinh b_2)} \\
&= \frac{\cosh b_2 \sinh b_1 \sinh b_3 + \sinh b_2 \cosh b_1 \cosh b_3}{\cosh b_3 \cosh b_1 \cosh b_2 + \sinh b_3 \sinh b_1 \sinh b_2} \\
&= \frac{\tanh b_2 + \tanh b_1 \tanh b_3}{1 + \tanh b_1 \tanh b_2 \tanh b_3}.
\end{aligned} \tag{B.22}$$

In case of $\hat{r} < 0$, we can re-parameterize b_1 and b_2 as

$$b_1 = \sqrt{\hat{q}_1} z_1 + \hat{T}_1 \xi^{1,true} + \hat{\tau}_2 \xi^{2,true}, \tag{B.23a}$$

$$b_2 = \sqrt{\hat{q}_2} (\psi z_1 + \sqrt{1 - \psi^2} z_2) + \hat{T}_2 \xi^{2,true} + \hat{\tau}_1 \xi^{1,true}, \tag{B.23b}$$

$$\psi = \frac{\hat{r}}{2\sqrt{\hat{q}_1 \hat{q}_2}}. \tag{B.23c}$$

We remark that this re-parameterization does not change the final results of multidimensional Gaussian integrations in the saddle-point equation.

To sum up, the saddle-point equations for non-conjugated order parameters are

given by

$$T_1 = [\xi^{1,true} \langle \xi^1 \rangle]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}, \quad (\text{B.24a})$$

$$T_2 = [\xi^{2,true} \langle \xi^2 \rangle]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}, \quad (\text{B.24b})$$

$$q_1 = [\langle \xi^1 \rangle^2]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}, \quad (\text{B.24c})$$

$$q_2 = [\langle \xi^2 \rangle^2]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}, \quad (\text{B.24d})$$

$$\tau_1 = [\xi^{1,true} \langle \xi^2 \rangle]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}, \quad (\text{B.24e})$$

$$\tau_2 = [\xi^{2,true} \langle \xi^1 \rangle]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}, \quad (\text{B.24f})$$

$$R = [\langle \xi^1 \xi^2 \rangle]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}, \quad (\text{B.24g})$$

$$r = [\langle \xi^1 \rangle \langle \xi^2 \rangle]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}}. \quad (\text{B.24h})$$

Next, we derive the saddle-point equations for those conjugated order parameters. For \hat{R} , we obtain the saddle point equation as

$$\frac{\partial F_\beta}{\partial R} = -\hat{R} - \alpha \beta^2 \tanh(\beta^2 R) + \frac{\alpha e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh \beta t_0 \cosh \beta (qt_0 + \sqrt{1-q^2} x_0) \frac{\partial}{\partial R} \ln Z_E = 0, \quad (\text{B.25})$$

where $\frac{\partial}{\partial R} \ln Z_E = \beta^2 \frac{e^{\beta^2(R-r)} \cosh(\beta \Lambda_+) - e^{-\beta^2(R-r)} \cosh(\beta \Lambda_-)}{e^{\beta^2(R-r)} \cosh(\beta \Lambda_+) + e^{-\beta^2(R-r)} \cosh(\beta \Lambda_-)} = \beta^2 G_c^-$. Therefore, the saddle-point equation of \hat{R} is given by

$$\hat{R} = \frac{\alpha \beta^2 e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} [\cosh \beta t_0 \cosh \beta (qt_0 + \sqrt{1-q^2} x_0) G_c^- - \alpha \beta^2 \tanh(\beta^2 R)]. \quad (\text{B.26})$$

For convenience, we define the measure $\langle \bullet \rangle$ as $\frac{e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh \beta t_0 \cosh \beta (qt_0 + \sqrt{1-q^2} x_0) \bullet$. As a result,

$$\hat{R} = \alpha \beta^2 \langle G_c^- \rangle - \alpha \beta^2 \tanh(\beta^2 R). \quad (\text{B.27})$$

For \hat{T}_1 , we have the following condition

$$\frac{\partial F_\beta}{\partial T_1} = -\hat{T}_1 + \frac{\alpha e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh \beta t_0 \cosh \beta (qt_0 + \sqrt{1-q^2} x_0) \frac{\partial}{\partial T_1} \ln Z_E = 0. \quad (\text{B.28})$$

To proceed, we first get the derivation of Λ_+ and Λ_- w.r.t T_1 as follows

$$\frac{\partial \Lambda_+}{\partial T_1} = t_0 - \frac{q}{\sqrt{1-q^2}} x_0 + \frac{\partial}{\partial T_1} \left(B + \frac{r-A}{B} \right) u + \frac{\partial K}{\partial T_1} u', \quad (\text{B.29a})$$

$$\frac{\partial \Lambda_-}{\partial T_1} = t_0 - \frac{q}{\sqrt{1-q^2}} x_0 + \frac{\partial}{\partial T_1} \left(B - \frac{r-A}{B} \right) u - \frac{\partial K}{\partial T_1} u'. \quad (\text{B.29b})$$

Then, the derivation of $\ln Z_E$ w.r.t T_1 can be simplified into the form as

$$\frac{\partial \ln Z_E}{\partial T_1} = \beta \left[G_s^+ t_0 - \frac{q}{\sqrt{1-q^2}} G_s^+ x_0 + \frac{\partial B}{\partial T_1} G_s^+ u + \frac{\partial}{\partial T_1} \left(\frac{r-A}{B} \right) G_s^- u + \frac{\partial K}{\partial T_1} G_s^- u' \right]. \quad (\text{B.30})$$

To further simplify the result, we need to evaluate the following equations. The first one is derived by applying Eq. (B.12) as

$$\begin{aligned}
& \int D\mathbf{t} \cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) G_s^+ t_0 \\
&= \int D\mathbf{t} \frac{\partial}{\partial t_0} \left(\cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) G_s^+ \right) \\
&= \beta \int D\mathbf{t} \left[\sinh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) + q \cosh \beta t_0 \sinh \beta(qt_0 + \sqrt{1-q^2}x_0) \right] G_s^+ \\
&+ \beta \int D\mathbf{t} \cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) \left[T_1 + \tau_1 G_c^- - T_1 (G_s^+)^2 - \tau_1 G_s^+ G_s^- \right].
\end{aligned} \tag{B.31}$$

The second one is derived as

$$\begin{aligned}
& \int D\mathbf{t} \cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) G_s^+ x_0 \\
&= \int D\mathbf{t} \frac{\partial}{\partial x_0} \left(\cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) G_s^+ \right) \\
&= \beta \sqrt{1-q^2} \int D\mathbf{t} \cosh \beta t_0 \sinh \beta(qt_0 + \sqrt{1-q^2}x_0) G_s^+ \\
&+ \frac{\beta}{\sqrt{1-q^2}} \int D\mathbf{t} \cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) \\
&\times \left[(\tau_2 - qT_1) + (T_2 - q\tau_1) G_c^- - (\tau_2 - qT_1) (G_s^+)^2 - (T_2 - q\tau_1) G_s^+ G_s^- \right].
\end{aligned} \tag{B.32}$$

The third one is derived as

$$\begin{aligned}
& \int D\mathbf{t} \cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) G_s^+ u \\
&= \int D\mathbf{t} \frac{\partial}{\partial u} \left(\cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) G_s^+ \right) \\
&= \beta \int D\mathbf{t} \cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) \left[B + \frac{r-A}{B} G_c^- - B (G_s^+)^2 - \frac{r-A}{B} G_s^+ G_s^- \right].
\end{aligned} \tag{B.33}$$

The fourth one is derived as

$$\begin{aligned}
& \int D\mathbf{t} \cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) G_s^- u \\
&= \int D\mathbf{t} \frac{\partial}{\partial u} \left(\cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) G_s^- \right) \\
&= \beta \int D\mathbf{t} \cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) \left[B G_c^- + \frac{r-A}{B} - \frac{r-A}{B} (G_s^-)^2 - B G_s^+ G_s^- \right].
\end{aligned} \tag{B.34}$$

The last one is given by

$$\begin{aligned}
& \int D\mathbf{t} \cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) G_s^- u' \\
&= \int D\mathbf{t} \frac{\partial}{\partial u'} \left(\cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) G_s^- \right) \\
&= \beta K \int D\mathbf{t} \cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) \left[1 - (G_s^-)^2 \right].
\end{aligned} \tag{B.35}$$

Through a bit lengthy algebraic manipulations, we get from Eq. (B.28)

$$\hat{T}_1 = \frac{\alpha \beta^2 e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \sinh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) G_s^+. \tag{B.36}$$

We thus define another measure $\langle\langle \bullet \rangle\rangle = \frac{e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \sinh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) \bullet$, and it then follows that

$$\hat{T}_1 = \alpha \beta^2 \langle\langle G_s^+ \rangle\rangle. \tag{B.37}$$

Similarly, we can obtain the saddle-point equation of $\hat{\tau}_1$ as

$$\hat{\tau}_1 = \alpha \beta^2 \langle\langle G_s^- \rangle\rangle. \tag{B.38}$$

Next we turn to the saddle-point equations for \hat{T}_2 and $\hat{\tau}_2$. We first get the derivation of Λ_+ and Λ_- w.r.t T_2 as

$$\frac{\partial \Lambda_+}{\partial T_2} = \frac{x_0}{\sqrt{1-q^2}} - \frac{1}{B} \frac{\partial A}{\partial T_2} u + \frac{\partial K}{\partial T_2} u', \tag{B.39a}$$

$$\frac{\partial \Lambda_-}{\partial T_2} = -\frac{x_0}{\sqrt{1-q^2}} + \frac{1}{B} \frac{\partial A}{\partial T_2} u - \frac{\partial K}{\partial T_2} u'. \tag{B.39b}$$

Based on the above equations, we get the derivation of $\ln Z_E$ w.r.t T_2 given by

$$\frac{\partial \ln Z_E}{\partial T_2} = \beta \left[\frac{x_0}{\sqrt{1-q^2}} G_s^- - \frac{1}{B} \frac{\partial A}{\partial T_2} G_s^- u + \frac{\partial K}{\partial T_2} G_s^- u' \right]. \tag{B.40}$$

Then we have

$$\hat{T}_2 = \frac{\alpha \beta e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) \left[\frac{x_0}{\sqrt{1-q^2}} G_s^- - \frac{1}{B} \frac{\partial A}{\partial T_2} G_s^- u + \frac{\partial K}{\partial T_2} G_s^- u' \right]. \tag{B.41}$$

For a further simplification, we need to derive the following identity as

$$\begin{aligned}
& \int D\mathbf{t} \cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) G_s^- x_0 \\
&= \int D\mathbf{t} \frac{\partial}{\partial x_0} \left(\cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) G_s^- \right) \\
&= \beta \sqrt{1-q^2} \int D\mathbf{t} \cosh \beta t_0 \sinh \beta(qt_0 + \sqrt{1-q^2}x_0) G_s^- \\
&+ \frac{\beta}{\sqrt{1-q^2}} \int D\mathbf{t} \cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) \\
&\times \left[(\tau_2 - qT_1) G_s^- - (\tau_2 - qT_1) G_s^+ G_s^- - (T_2 - q\tau_1) (G_s^-)^2 + (T_2 - q\tau_1) \right].
\end{aligned} \tag{B.42}$$

Using Eq. (B.42) together with Eq. (B.34) and Eq. (B.35), we finally arrive at the saddle-point equation of \hat{T}_2 :

$$\hat{T}_2 = \frac{\alpha\beta^2 e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh \beta t_0 \sinh \beta(qt_0 + \sqrt{1-q^2}x_0) G_s^-. \quad (\text{B.43})$$

We thus define the third measure $\langle\langle\bullet\rangle\rangle = \frac{e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh \beta t_0 \sinh \beta(qt_0 + \sqrt{1-q^2}x_0) \bullet$. We then write the saddle-point equation in a compact form as

$$\hat{T}_2 = \alpha\beta^2 \langle\langle G_s^- \rangle\rangle. \quad (\text{B.44})$$

Similarly, we obtain the saddle-point equation for $\hat{\tau}_2$ as

$$\hat{\tau}_2 = \alpha\beta^2 \langle\langle G_s^+ \rangle\rangle. \quad (\text{B.45})$$

Then we turn to the saddle-point equations of \hat{q}_1 and \hat{q}_2 . From $\frac{\partial F_\beta}{\partial q_1} = 0$, we get

$$\frac{1}{2}\hat{q}_1 - \frac{\alpha\beta^2}{2} + \frac{\alpha\beta e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) \frac{\partial \ln Z_E}{\partial q_1} = 0. \quad (\text{B.46})$$

The derivation of $\ln Z_E$ w.r.t q_1 is given by

$$\frac{\partial \ln Z_E}{\partial q_1} = \frac{\partial B}{\partial q_1} G_s^+ u + \frac{\partial}{\partial q_1} \left(\frac{r-A}{B} \right) G_s^- u + \frac{\partial K}{\partial q_1} G_s^- u'. \quad (\text{B.47})$$

Using Eq. (B.33), Eq. (B.34) and Eq. (B.35), we get the saddle-point equation of \hat{q}_1 as

$$\begin{aligned} \hat{q}_1 &= \frac{\alpha\beta^2 e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh(\beta t_0) \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) (G_s^+)^2 \\ &= \alpha\beta^2 \langle (G_s^+)^2 \rangle. \end{aligned} \quad (\text{B.48})$$

Similarly, we can derive the saddle-point equation for \hat{q}_2 as

$$\begin{aligned} \hat{q}_2 &= \frac{\alpha\beta^2 e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh(\beta t_0) \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) (G_s^-)^2 \\ &= \alpha\beta^2 \langle (G_s^-)^2 \rangle. \end{aligned} \quad (\text{B.49})$$

Lastly, we derive the saddle-point equation for \hat{r} as

$$\frac{\hat{r}}{2} + \frac{\alpha e^{-\beta^2}}{\cosh \beta^2 q} \int D\mathbf{t} \cosh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) \frac{\partial \ln Z_E}{\partial r} = 0. \quad (\text{B.50})$$

Noting that $\frac{\partial \ln Z_E}{\partial r} = -\beta^2 G_c^- + \beta \left(\frac{1}{B} G_s^- u + \frac{\partial K}{\partial r} G_s^- u' \right)$, we get the saddle-point equation of \hat{r} as

$$\hat{r} = 2\alpha\beta^2 \langle G_s^+ G_s^- \rangle. \quad (\text{B.51})$$

Note that Eq. (B.34) and Eq. (B.35) are used to derive the final result.

To sum up, the saddle-point equations of our minimal model are listed as follows

$$\hat{T}_1 = \alpha\beta^2 \langle \langle G_s^+ \rangle \rangle, \quad (\text{B.52a})$$

$$\hat{T}_2 = \alpha\beta^2 \langle \langle \langle G_s^- \rangle \rangle \rangle, \quad (\text{B.52b})$$

$$\hat{\tau}_1 = \alpha\beta^2 \langle \langle G_s^- \rangle \rangle, \quad (\text{B.52c})$$

$$\hat{\tau}_2 = \alpha\beta^2 \langle \langle \langle G_s^+ \rangle \rangle \rangle, \quad (\text{B.52d})$$

$$\hat{q}_1 = \alpha\beta^2 \langle (G_s^+)^2 \rangle, \quad (\text{B.52e})$$

$$\hat{q}_2 = \alpha\beta^2 \langle (G_s^-)^2 \rangle, \quad (\text{B.52f})$$

$$\hat{r} = 2\alpha\beta^2 \langle G_s^+ G_s^- \rangle, \quad (\text{B.52g})$$

$$\hat{R} = \alpha\beta^2 \langle G_c^- \rangle - \alpha\beta^2 \tanh(\beta^2 R). \quad (\text{B.52h})$$

Appendix C. The free energy function in the limit of $q = 0$

In the case of $q = 0$, the saddle point equation of the minimal model has the solution: $q_1 = q_2 = T_1 = T_2$ and other order parameters vanish. Thus, we can simplify Λ_+ and Λ_- as follows,

$$\Lambda_+ = T_1 t_0 + T_2 x_0 + \sqrt{q_1 - (T_1)^2} u + \sqrt{q_2 - (T_2)^2} u', \quad (\text{C.1a})$$

$$\Lambda_- = T_1 t_0 - T_2 x_0 + \sqrt{q_1 - (T_1)^2} u - \sqrt{q_2 - (T_2)^2} u'. \quad (\text{C.1b})$$

We then define $\chi_1 = T_1 t_0 + \sqrt{q_1 - (T_1)^2} u$, and $\chi_2 = T_2 x_0 + \sqrt{q_2 - (T_2)^2} u'$, so the saddle point equation of \hat{T}_1 is given by

$$\begin{aligned} \hat{T}_1 &= \alpha\beta^2 e^{-\beta^2} \int D\mathbf{t} \sinh \beta t_0 \cosh \beta x_0 \left[\frac{\sinh \beta \Lambda_+ + \sinh \beta \Lambda_-}{\cosh \beta \Lambda_+ + \cosh \beta \Lambda_-} \right] \\ &= \alpha\beta^2 e^{-\beta^2} \int D\mathbf{t} \sinh \beta t_0 \cosh \beta x_0 \left[\frac{\sinh \beta \chi_1 \cosh \beta \chi_2}{\cosh \beta \chi_1 \cosh \beta \chi_2} \right] \\ &= \alpha\beta^2 e^{-\frac{\beta^2}{2}} \int Dt_0 Du \sinh \beta t_0 \tanh \beta (T_1 t_0 + \sqrt{q_1 - (T_1)^2} u), \end{aligned} \quad (\text{C.2})$$

where we used the identity $\int Dx_0 \cosh(\beta x_0) = e^{\beta^2/2}$. Similarly, one can prove that $\hat{T}_1 = \hat{T}_2$. As for \hat{q}_1 , we will have

$$\begin{aligned} \hat{q}_1 &= \alpha\beta^2 e^{-\beta^2} \int D\mathbf{t} \cosh \beta t_0 \cosh \beta x_0 \left[\frac{\sinh \beta \Lambda_+ + \sinh \beta \Lambda_-}{\cosh \beta \Lambda_+ + \cosh \beta \Lambda_-} \right]^2 \\ &= \alpha\beta^2 e^{-\beta^2} \int D\mathbf{t} \cosh \beta t_0 \cosh \beta x_0 \left[\frac{\sinh \beta \chi_1 \cosh \beta \chi_2}{\cosh \beta \chi_1 \cosh \beta \chi_2} \right]^2 \\ &= \alpha\beta^2 e^{-\frac{\beta^2}{2}} \int Dt_0 Du \cosh \beta t_0 \tanh^2 \beta (T_1 t_0 + \sqrt{q_1 - (T_1)^2} u). \end{aligned} \quad (\text{C.3})$$

Similarly, one can prove that $\hat{q}_1 = \hat{q}_2$.

It is easy to prove that $\hat{\tau}_1 = 0$, $\hat{\tau}_2 = 0$ and $\hat{R} = 0$, $\hat{r} = 0$, then we can express b_1, b_2 and b_3 as :

$$b_1 = \hat{T}_1 \xi^{1,true} + \sqrt{\hat{q}_1} z_1, \quad (\text{C.4a})$$

$$b_2 = \hat{T}_2 \xi^{2,true} + \sqrt{\hat{q}_2} z_2, \quad (\text{C.4b})$$

$$b_3 = 0. \quad (\text{C.4c})$$

Therefore, $\int D\mathbf{z} [\ln Z_{eff}]_{\xi^{1,true}, \xi^{2,true}}$ can be simplified as $2 \int Dz \ln 2 \cosh(\hat{T}_1 + \sqrt{\hat{q}_1} z)$. T_1 becomes

$$\begin{aligned} T_1 &= \left[\int Dz_1 Dz_2 Dz_3 \xi^{1,true} \tanh(\hat{T}_1 \xi^{1,true} + \sqrt{\hat{q}_1} z_1) \right]_{\xi^{1,true}, \xi^{2,true}} \\ &= \int Dz_1 \frac{1}{2} \left[\tanh(\hat{T}_1 + \sqrt{\hat{q}_1} z_1) - \tanh(-\hat{T}_1 + \sqrt{\hat{q}_1} z_1) \right] \\ &= \int Dz_1 \frac{1}{2} \left[\tanh(\hat{T}_1 + \sqrt{\hat{q}_1} z_1) - \tanh(-\hat{T}_1 - \sqrt{\hat{q}_1} z_1) \right] \\ &= \int Dz_1 \tanh(\hat{T}_1 + \sqrt{\hat{q}_1} z_1). \end{aligned} \quad (\text{C.5})$$

One can easily prove that $T_1 = T_2$. Similarly for the order parameter q_2 , we can also get:

$$\begin{aligned} q_2 &= \left[\int Dz_2 \tanh^2(\hat{T}_2 \xi^{1,true} + \sqrt{\hat{q}_2} z_1) \right]_{\xi^{1,true}, \xi^{2,true}} \\ &= \frac{1}{2} \int Dz_2 \left[\tanh^2(\hat{T}_2 + \sqrt{\hat{q}_2} z_2) + \tanh^2(-\hat{T}_2 + \sqrt{\hat{q}_2} z_2) \right] \\ &= \int Dz_2 \tanh^2(\hat{T}_2 + \sqrt{\hat{q}_2} z_2). \end{aligned} \quad (\text{C.6})$$

It is easy to show that $q_1 = q_2$, and moreover $R = r = \tau_1 = \tau_2 = 0$. To sum up, we recover the saddle point equations of one-bit RBM reported in Ref. [16].

Next, we show the $q = 0$ version of the free energy function. It is easy to show that $Z_E = \cosh \beta(\chi_1 + \chi_2) + \cosh \beta(\chi_1 - \chi_2) = 2 \cosh \beta \chi_1 \cosh \beta \chi_2$. Therefore, we have the following integral

$$\begin{aligned} \alpha e^{-\beta^2} \int D\mathbf{t} \cosh \beta t_0 \cosh \beta x_0 \ln Z_E &= \alpha e^{-\beta^2} \int D\mathbf{t} \cosh \beta t_0 \cosh \beta x_0 \ln(2 \cosh \beta \chi_1 \cosh \beta \chi_2) \\ &= \alpha \ln 2 + 2\alpha e^{-\frac{\beta^2}{2}} \int Du Dt_0 \cosh \beta t_0 \ln \cosh \beta(T_1 t_0 + \sqrt{q_1 - (T_1)^2} u). \end{aligned} \quad (\text{C.7})$$

Collecting all the relevant terms, we can show that the free energy of our minimal model with $q = 0$ is two times as large as that of one-bit RBM, which can also be intuitively understood by the argument that the partition function factorizes as $\Omega = \Omega_{\text{one-bit-RBM}}^2$. Therefore we can conclude that the critical data size for spontaneous symmetry breaking does not change even if an additional hidden node is added. This conclusion seems to carry over to the case of more hidden nodes following the principle of the partition function's factorization.

Appendix D. Derivation of the critical data size (Eq. (18) in the main text)

We assume that near to the transition point, all order parameters are very small such that we can expand them to leading order. According to Eq. (B.24), when the critical point is approached from below, $\langle \xi^1 \rangle \simeq \tanh b_1 \simeq b_1$. Analogously, $\langle \xi^2 \rangle \simeq b_2$, and $\langle \xi^1 \xi^2 \rangle \simeq b_3$. We thus have the following equalities in this limit:

$$T_1 = [\xi^{1,true} \langle \xi^1 \rangle]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}} = \hat{T}_1 + q\hat{\tau}_2, \quad (\text{D.1})$$

$$\tau_2 = [\xi^{2,true} \langle \xi^1 \rangle]_{\mathbf{z}, \xi^{1,true}, \xi^{2,true}} = \hat{\tau}_2 + q\hat{T}_1. \quad (\text{D.2})$$

Similarly, in the limit of vanishing order parameters, we have the following approximation

$$\begin{aligned} G_s^+ &= \frac{e^{\beta^2(R-r)} \sinh(\beta\Lambda_+) + e^{-\beta^2(R-r)} \sinh(\beta\Lambda_-)}{e^{\beta^2(R-r)} \cosh(\beta\Lambda_+) + e^{-\beta^2(R-r)} \cosh(\beta\Lambda_-)} \\ &= \frac{\beta}{2}(\Lambda_+ + \Lambda_-). \end{aligned} \quad (\text{D.3})$$

Inserting this approximation into the saddle-point equations of \hat{T}_1 and $\hat{\tau}_2$, we obtain the approximate results of \hat{T}_1 and $\hat{\tau}_2$ as

$$\begin{aligned} \hat{T}_1 &= \alpha\beta^2 \langle \langle G_s^+ \rangle \rangle \simeq \frac{\alpha\beta^2 e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \sinh \beta t_0 \cosh \beta(qt_0 + \sqrt{1-q^2}x_0) \frac{\beta}{2}[\Lambda_+ + \Lambda_-] \\ &= \alpha\beta^4 [T_1 + \tanh(\beta^2 q)\tau_2], \\ \hat{\tau}_2 &= \alpha\beta^2 \langle \langle \langle G_s^+ \rangle \rangle \rangle \simeq \frac{\alpha\beta^2 e^{-\beta^2}}{\cosh(\beta^2 q)} \int D\mathbf{t} \cosh \beta t_0 \sinh \beta(qt_0 + \sqrt{1-q^2}x_0) \frac{\beta}{2}[\Lambda_+ + \Lambda_-] \\ &= \alpha\beta^4 [\tau_2 + \tanh(\beta^2 q)T_1]. \end{aligned} \quad (\text{D.4})$$

We recast the equations for all these four order parameters in a matrix form as

$$\begin{pmatrix} T_1 \\ \tau_2 \end{pmatrix} = \begin{pmatrix} 1 & q \\ q & 1 \end{pmatrix} \begin{pmatrix} \hat{T}_1 \\ \hat{\tau}_2 \end{pmatrix}, \quad (\text{D.5})$$

$$\begin{pmatrix} \hat{T}_1 \\ \hat{\tau}_2 \end{pmatrix} = \alpha\beta^4 \begin{pmatrix} 1 & \tanh(\beta^2 q) \\ \tanh(\beta^2 q) & 1 \end{pmatrix} \begin{pmatrix} T_1 \\ \tau_2 \end{pmatrix}. \quad (\text{D.6})$$

From the Eq. (D.5) and Eq. (D.6), T_1 and τ_2 can be calculated out as

$$\begin{pmatrix} T_1 \\ \tau_2 \end{pmatrix} = \alpha\beta^4 \begin{pmatrix} 1 + q \tanh(\beta^2 q) & q + \tanh(\beta^2 q) \\ q + \tanh(\beta^2 q) & 1 + q \tanh(\beta^2 q) \end{pmatrix} \begin{pmatrix} T_1 \\ \tau_2 \end{pmatrix} = \mathcal{M} \begin{pmatrix} T_1 \\ \tau_2 \end{pmatrix}, \quad (\text{D.7})$$

where the matrix \mathcal{M} is named the stability matrix, whose largest eigenvalue determines the critical value of the learning data size α_c . In detail, the stability matrix has two eigenvalues:

$$\lambda_+ = \alpha\beta^4 (1 + q \tanh(\beta^2 q) + |q + \tanh(\beta^2 q)|), \quad (\text{D.8})$$

$$\lambda_- = \alpha\beta^4 (1 + q \tanh(\beta^2 q) - |q + \tanh(\beta^2 q)|). \quad (\text{D.9})$$

The α_c can be read off from $\lambda_+ = 1$, i.e.,

$$\alpha_c = \frac{\beta^{-4}}{1 + q \tanh(\beta^2 q) + |q + \tanh(\beta^2 q)|}. \quad (\text{D.10})$$

An alternative way to understand that the smaller eigenvalue could not be used to determine α_c , is that it leads to a non-physical solution $\alpha_c = \frac{\beta^{-4}}{1 + q \tanh \beta^2 q - |q + \tanh(\beta^2 q)|}$. Because in a special case of large β limit and positive q , $\tanh(\beta^2 q) \simeq 1 - 2e^{-2\beta^2 q}$, then we have $\alpha_c \simeq \frac{e^{2\beta^2 q}}{2(1-q)\beta^4}$, which implies that this value tends to ∞ which is in contradiction with the expectation that learning should be easier given noise-free data.

References

- [1] D. Marr. A Theory for Cerebral Neocortex. *Proceedings of the Royal Society of London B: Biological Sciences*, 176(1043):161–234, 1970.
- [2] H.B. Barlow. Unsupervised learning. *Neural Computation*, 1:295–311, 1989.
- [3] G Hinton, S Osindero, and Y Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.
- [4] Nicolas Le Roux and Yoshua Bengio. Representational power of restricted boltzmann machines and deep belief networks. *Neural Comput.*, 20(6):1631–1649, 2008.
- [5] Adriano Barra, Alberto Bernacchia, Enrica Santucci, and Pierluigi Contucci. On the equivalence of hopfield networks and boltzmann machines. *Neural Networks*, 34:1–9, 2012.
- [6] Haiping Huang and Taro Toyoizumi. Advanced mean-field theory of the restricted boltzmann machine. *Phys. Rev. E*, 91:050101, 2015.
- [7] Marylou Gabrie, Eric W Tramel, and Florent Krzakala. Training restricted boltzmann machine via the thouless-anderson-palmer free energy. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 640–648. Curran Associates, Inc., 2015.
- [8] Marc Mézard. Mean-field message-passing equations in the hopfield model and its generalizations. *Phys. Rev. E*, 95:022117, 2017.
- [9] J. Tübiana and R. Monasson. Emergence of compositional representations in restricted boltzmann machines. *Phys. Rev. Lett.*, 118:138301, 2017.
- [10] J. Song, M. Marsili, and J. Jo. Resolution and relevance trade-offs in deep learning. *J. Stat. Mech*, 2018:123406, 2018.
- [11] A. Decelle, G. Fissore, and C. Furtlehner. Spectral dynamics of learning in restricted Boltzmann machines. *EPL (Europhysics Letters)*, 119(6):60001, 2017.
- [12] Domingos S. P. Salazar. Nonequilibrium thermodynamics of restricted boltzmann machines. *Phys. Rev. E*, 96:022131, 2017.
- [13] Gavin S. Hartnett, Edward Parker, and Edward Geist. Replica symmetry breaking in bipartite spin glasses and neural networks. *Phys. Rev. E*, 98:022116, 2018.
- [14] Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In W.W. Cohen, A. McCallum, and S.T. Roweis, editors, *Proceedings of the 25th International Conference on Machine Learning*, pages 1064–1071. ACM, New York, NY, USA, 2008.
- [15] Haiping Huang and Taro Toyoizumi. Unsupervised feature learning from finite data by message passing: Discontinuous versus continuous phase transition. *Phys. Rev. E*, 94:062310, 2016.
- [16] Haiping Huang. Statistical mechanics of unsupervised feature learning in a restricted Boltzmann machine with binary synapses. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(5):053302, 2017.

- [17] S. Cocco, R. Monasson, and V. Sessak. High-dimensional inference with the generalized hopfield model: Principal component analysis and corrections. *Phys. Rev. E*, 83:051123, 2011.
- [18] Adriano Barra, Giuseppe Genovese, Peter Sollich, and Daniele Tantari. Phase transitions in restricted boltzmann machines with generic priors. *Phys. Rev. E*, 96:042156, 2017.
- [19] Adriano Barra, Giuseppe Genovese, Peter Sollich, and Daniele Tantari. Phase diagram of restricted boltzmann machines and generalized hopfield networks with arbitrary priors. *Phys. Rev. E*, 97:022310, 2018.
- [20] Haiping Huang. Role of zero synapses in unsupervised feature learning. *Journal of Physics A: Mathematical and Theoretical*, 51:08LT01, 2018.
- [21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016.
- [22] M. Mézard and G. Parisi. The bethe lattice spin glass revisited. *Eur. Phys. J. B*, 20:217, 2001.
- [23] M. Mézard and A. Montanari. *Information, Physics, and Computation*. Oxford University Press, Oxford, 2009.
- [24] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free energy approximations and generalized belief propagation algorithms. *IEEE Trans Inf Theory*, 51:2282–2312, 2005.
- [25] H. Nishimori. *Statistical Physics of Spin Glasses and Information Processing: An Introduction*. Oxford University Press, Oxford, 2001.
- [26] Pau Rodríguez, Jordi González, Guillem Cucurull, Josep M. Gonfaus, and Xavier Roca. Regularizing CNNs with locally constrained decorrelations. *arXiv:1611.01967*, 2016.
- [27] Li Wan, Matthew Zeiler, Sixin Zhang, Yann L. Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In Sanjoy Dasgupta and David Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 1058–1066. JMLR Workshop and Conference Proceedings, 2013.
- [28] Haiping Huang. Mechanisms of dimensionality reduction and decorrelation in deep neural networks. *Phys. Rev. E*, 98:062313, 2018.
- [29] A. Engel, H. M. Köhler, F. Tschepke, H. Vollmayr, and A. Zippelius. Storage capacity and learning algorithms for two-layer neural networks. *Phys. Rev. A*, 45:7590–7609, 1992.
- [30] E. Barkai, D. Hansel, and H. Sompolinsky. Broken symmetries in multilayered perceptrons. *Phys. Rev. A*, 45:4146–4161, 1992.
- [31] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.
- [32] Haiping Huang, K Y Michael Wong, and Yoshiyuki Kabashima. Entropy landscape of solutions in the binary perceptron problem. *Journal of Physics A: Mathematical and Theoretical*, 46:375002, 2013.