Spatially heterogeneous learning by a deep student machine

Hajime Yoshino
Cybermedia Center, Osaka University, Toyonaka, Osaka 560-0043, Japan
and Graduate School of Science, Osaka University, Toyonaka, Osaka 560-0043, Japan

(Received 14 February 2023; revised 29 May 2023; accepted 12 July 2023; published 31 July 2023)

Despite spectacular successes, deep neural networks (DNNs) with a huge number of adjustable parameters remain largely black boxes. To shed light on the hidden layers of DNNs, we study supervised learning by a DNN of width N and depth L consisting of NL perceptrons with c inputs by a statistical mechanics approach called the teacher-student setting. We consider an ensemble of student machines that exactly reproduce M sets of N-dimensional input/output relations provided by a teacher machine. We show that the statistical mechanics problem becomes exactly solvable in a high-dimensional limit which we call a "dense limit": $N \gg c \gg 1$ and $M \gg 1$ with fixed $\alpha = M/c$ using the replica method developed by Yoshino [SciPost Phys. Core **2**, 005 (2020)] In conjunction with the theoretical study, we also study the model numerically performing simple greedy Monte Carlo simulations. Simulations reveal that learning by the DNN is quite heterogeneous in the network space: configurations of the teacher and the student machines are more correlated within the layers closer to the input/output boundaries, while the central region remains much less correlated due to the overparametrization in qualitative agreement with the theoretical prediction. We evaluate the generalization error of the DNN with various depths L both theoretically and numerically. Remarkably, both the theory and the simulation suggest that the generalization ability of the student machines, which are only weakly correlated with the teacher in the center, does not vanish even in the deep limit $L \gg 1$, where the system becomes heavily overparametrized. We also consider the impact of the effective dimension $D(\leqslant N)$ of data by incorporating the hidden manifold model [Goldt, Mézard, Krzakala, and Zdevorová, Phys. Rev. X **10**, 041044 (2020)] into our model. Replica theory implies that the loop corrections to the dense limit, which reflect correlations between different nodes in the network, become enhanced by either decreasing the width N or decreasing the effective dimension D of the data. Simulation suggests that both lead to significant improvements in generalization ability.

DOI: 10.1103/PhysRevResearch.5.033068

I. INTRODUCTION

The mechanism of machine learning by deep neural networks (DNNs) [1] remains largely unknown. One of the most puzzling points is the issue of overparametrization: supervised learning by DNNs can work even in the regime where the number of adjustable parameters is larger than the data size by orders of magnitudes. This conflicts sharply with the traditional wisdom of data modeling: for example, one should avoid fitting 10 data points by a fitting function with 100 adjustable parameters, which is just nonsense. However, empirically it has been found repeatedly that such overparametrized DNNs can somehow avoid overfitting and generalize well, i.e., they can successfully describe new data not used during training. Uncovering the reason for this peculiar phenomenon is a very interesting and challenging scientific problem [2,3]. An important point to be noted is that the effective dimension D of the data can be much smaller than the apparent dimension N of the data. It has been shown in studies of shallow networks that the generalization ability improves by increasing N/D due to a kind of self-averaging mechanism [4–6]. However, the generalization ability of the deeper system remains unexplained.

Statistical mechanics on neural networks has a long history that dates back to the 1980's [7–9]. Studies on the single perceptrons [8,9] and shallow networks [4,10] have provided many useful insights, and some progress has been made also on deeper networks [11–14]. However, what is going on in the hidden layers remains largely unknown. The first attempt to uncover the black box was made in [15] by the present author based on the replica method, and it predicted an unexpected phenomenology of DNNs: spatially heterogeneous learning. Unfortunately, the theory suffered from a serious problem due to an uncontrolled approximation, and the validity of the prediction remained elusive.

To understand the mechanism for the generalization ability of deep networks, we study supervised learning by DNNs considering the so-called teacher-student setting, which is a canonical setting to study statistical inference problems [16,17] by methods of statistical mechanics. We consider a prototypical DNN of a rectangular shape with width N and depth L consisting of NL perceptrons with c inputs, which defines a mapping between an N-dimensional input vector and an N-dimensional output vector (see Fig. 1). For the data,

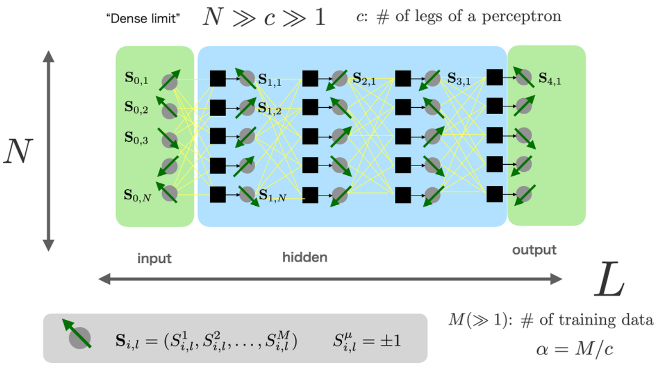

FIG. 1. Schematic picture of the multilayer perceptron network of depth L and width N. In this example, the depth is $L = 4$. Each arrow represents an M-component vector spin $\mathbf{S}_i = (S_i^1, S_i^2, \ldots, S_i^M)$ with its component $S_i^\mu = \pm 1$ representing the state of a "neuron" in the μth pattern.

we consider M pairs of input/output vectors provided by a teacher machine, and we consider an ensemble of student machines that exactly satisfy the same input/output relations as the teacher. The phase-space volume of such an ensemble is referred to as Gardner's volume [8,9], which should be very large for overparametrized DNNs. In fact, it is known that gradient descent dynamics find such a machine without going over barriers in the loss landscape [18–20]. In Fig. 2, we show a schematic picture of the phase space of the machines. If M is small, typically students will not find the teacher. This situation would be regarded as a *liquid phase*. If M is increased, a *crystalline phase* may emerge in which students find the (hidden) crystal, i.e., the teacher. We also wish to consider the impact of the effective dimension D of the data by incorporating the hidden manifold model [10,21,22] in our model. Using statistical mechanics methods, we wish to investigate how different machines that satisfy the same set of input/output boundary conditions become correlated with each other in the hidden layers, and to evaluate their generalization ability, i.e., the ability of the students to reproduce the teacher's output against new input data not used in training.

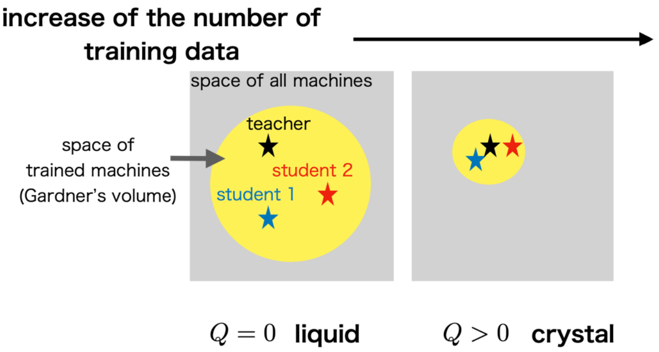
increase of the number of training data
space of all machines
teacher
student 2
student 1
space of trained machines (Gardner's volume)
$Q = 0$ liquid
$Q > 0$ crystal

FIG. 2. Schematic picture of the phase space of machines: the gray box represents the set of all machines which can be generated varying the parameters (e.g., synaptic weights) given a network structure. The yellow region represents the subspace in which machines agree with the teacher's machine for a given M set of training data. Liquid phase: if the number of training data M is small, the subspace is so large that the machines are typically widely separated and their mutual overlap Q is typically zero. Crystalline phase: M is large enough that machines have a finite overlap Q with respect to each other.

The first attempt to tackle the statistical mechanics problem of DNNs was made recently in [15] based on the replica method in a high-dimensional limit with $c = N = D \gg 1$ and $M \gg 1$ with fixed $\alpha = M/c$. Unfortunately, it suffered from an uncontrolled "tree" approximation, which is invalid for the global coupling case $c = N$. In the present paper, we show that the problem can be overcome in the limit $N \gg c \gg 1$, which we refer to as the *dense limit*. As a result, we establish an exactly solvable statistical mechanics model of DNNs that has long been anticipated.

Using the exact solution of the model, which can be obtained by the replica method developed in [15], we analyze the key question of the generalization ability of DNNs, including the overparametrized regime. We also show that the effect of the finiteness of the width (apparent dimension of the data) N and the effective dimension D similarly enhances loop corrections, which induce correlations between distant layers. In conjunction with the theoretical study, we also perform extensive numerical simulations to examine the theoretical predictions. We use the Monte Carlo method, which allows more efficient exploration of the solution space compared with the usual gradient descent algorithms.

The article is organized as follows. In Sec. II we summarize the main results of this paper. In Sec. III we introduce our model. We discuss the replica approach in Sec. IV and numerical simulations in Sec. V. In Sec. VI we conclude the paper with perspectives. In Appendix A we discuss a connection between some layered spin-glass models and DNNs, and in Appendix B we present some details of the replica theory.

II. SUMMARY OF RESULTS

Let us summarize the main results of this work. On the theoretical side, we find the following:

(i) We establish an exactly solvable statistical mechanics model of DNNs in the dense limit $N \gg c \gg 1$. The exact solution of the model is obtained using the replica approach. It is shown that the correction to the dense limit due to finiteness of the width N can be expressed by loop corrections. Fortunately, it turns out that the theoretical results presented in [15] are essentially valid in the dense limit $N \gg c \gg 1$, although they are unjustified for the global coupling $c = N$ assumed there.

(ii) We show that the smallness of the effective dimension $D(< N)$ of the hidden manifold model [21] enhances the loop corrections. Thus the finite dimension D effect is predicted to be similar to the finite width N effect.

(iii) The learning curve $\epsilon = \epsilon_L(\alpha)$ of the DNNs with various depths L is analyzed evaluating the generalization error $\epsilon_L(\alpha)$ in the case of a Bayes-optimal teacher-student setting where replica symmetry holds. It becomes independent of the depth L, i.e., $\epsilon_L(\alpha) = \epsilon_\infty(\alpha)$, as long as the network is deep enough such that the liquid phase, where students are decorrelated from the teacher, remains in the center reflecting strong overparametrization.

On the numerical side, we find the following:

(i) We simulated the model with finite connectivity c and width N in the Bayes-optimal teacher-student setting, and we found that a simple greedy Monte Carlo algorithm allows the student machines to equilibrate after sufficiently long times. Thus typical equilibrium states are accessible starting from typical random initial configurations without going over barriers in the loss landscape.

(ii) Observation of the overlap between the machines reveals spatially inhomogeneous learning in qualitative agreement with the theory. While the theory in the dense limit $N \gg c \gg 1$ predicts (in the case of strong overparametrization) crystalline regions with finite overlap close to the input/output layers separated by a liquid region with zero overlap in the center, the distinction between the crystalline and the liquid phases becomes blurred in systems with finite width N and finite connectivity c. Nonetheless, the presence of the liquid-like region in the center becomes clearer by making the width N large and the connectivity c large or the depth L large. We consider that the remnant overlap left in the center by the finite width N and the finite connectivity c effects play the role of a symmetry-breaking field, which connect the two crystalline regions attached to the boundaries.

(iii) Observation of the learning curve $\epsilon = \epsilon_{L,c,N,D}(\alpha)$ reveals that it becomes independent of the depth L in deep enough systems, in agreement with the theoretical prediction.

(iv) The observations reveal that the finite effective dimension D effect and the finite width N effect are indeed very similar, as suggested by consideration of the loop effects in the theory. The generalization error $\epsilon_{L,c,N,D}(\alpha)$ decreases significantly, decreasing either the width N or the effective dimension D.

III. MODEL

A. Multilayer perceptron network

We consider a simple multilayer neural network of a rectangular shape with width N and depth L (see Fig. 1). The input and output layers are located at the boundaries $l = 0$ and L, respectively, while $l = 1, 2, \ldots, L-1$ are hidden layers. On each layer $l = 0, 1, 2, \ldots, L$ there are N neurons labeled as (l, i) with $i = 1, 2, \ldots, N$. The state of the neuron (l, i) is represented by an Ising spin $S_{l,i}$: it is active if $S_{l,i} = 1$ and inactive if $S_{l,i} = -1$.

The network is constructed as follows. There are $N_{\blacksquare} = NL$ perceptrons. Consider a perceptron $\blacksquare = (l, i)$, which is the ith neuron in the lth layer. It receives c inputs from the outputs of the perceptrons $\blacksquare(k)$ $(k = 1, 2, \ldots, c)$ in the previous $(l-1)$th layer, weighted by $\mathbf{J}_{\blacksquare} = (J^1_{\blacksquare}, J^2_{\blacksquare}, \ldots, J^c_{\blacksquare})$. [For the special case $l = 1$, $\blacksquare(k)$ should be understood as one of the spins in the input layer.] The c perceptrons are selected randomly out of N possible perceptrons in the $(l-1)$th layer.

The output of the perceptron $\blacksquare$, which we denote as $S_{\blacksquare}$, is given by

$$S_{\blacksquare} = \mathrm{sgn}\left(\frac{1}{\sqrt{c}}\sum_{k=1}^{c} J^k_{\blacksquare} S_{\blacksquare(k)}\right), \tag{1}$$

where $\mathrm{sgn}(y) = y/|y|$ is our choice for the activation function. We assume that the synaptic weights $J^k_{\blacksquare}$ take real numbers normalized such that

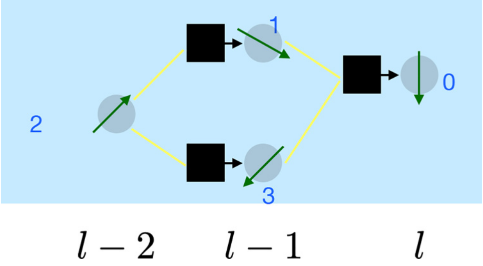

FIG. 3. A loop of interactions in a DNN extended over three layers, through three perceptrons and four bonds.

$$\sum_{k=1}^{c} (J^k_{\blacksquare})^2 = c. \tag{2}$$

For convenience, we refer to the state of the neurons $S_{l,i}$'s as "spins" and the synaptic weights $J^k_{\blacksquare}$s as "bonds" in the present paper. We denote the set of perceptrons in the lth layer as $\blacksquare \in l$ and the set of perceptrons whose outputs become input for $\blacksquare$ as $\partial\blacksquare$, i.e., $\partial\blacksquare = \{\blacksquare(1), \blacksquare(2), \ldots, \blacksquare(c)\}$. For convenience, we introduce also $\blacksquare \in 0$ so that we can write the set of spins in the input layer as $S_{\blacksquare\in 0}$.

B. Dense coupling

As stated above, c legs of a perceptron $\blacksquare$ at the lth layer are connected to c neurons $S_{\blacksquare(k)}$ $(k = 1, 2, \ldots, c)$ in the previous $(l-1)$th layer. The c neurons out of N possible neurons are selected randomly. Thus our graph becomes a sort of sparse (layered) random graph when c is finite. We will find in Sec. IV that this construction enables us to obtain an exactly solvable statistical mechanics model of DNN because of the following reasons:

(i) The graph becomes locally treelike as in the case of Bethe lattices so that contributions of "loops" can be neglected in the wide limit $N \to \infty$ with fixed c. This can be seen as follows. For instance, consider a loop $0 \to 1 \to 2 \to 3 \to 0$ shown in Fig. 3. Starting from 0, choose any 1 connected to 0. Then choose any 2 connected to 1. Then choose any 3 (different from 1) connected to 0. In the case of global coupling $c = N$, 2 is certainly connected to 3 completing a loop. However, in the case of dense coupling, in a given realization of the random graph, 2 is connected to 3 only with a probability $\sim c/N$. Thus in the limit $N \to \infty$ with fixed c, the probability to complete the loop vanishes. This argument can be generalized for 2-loops, 3-loops, etc., which happen with probability $O(c/N)^2$, $O(c/N)^3$, etc. Note that the loops cannot be neglected in the case of global coupling, $c = N$ (assumed in [15]).

(ii) In the case of the global coupling $c = N$, the system is symmetric under permutations of the perceptrons within each layer so that one has to consider whether this symmetry becomes broken spontaneously [23]. In the case of sparse coupling, $c < N$, we can eliminate this symmetry by choosing the connections in stochastic ways, i.e., a random graph.

(iii) In the setup of our theory, we finally consider $c \to \infty$ (and $M = \alpha c \to \infty$ [see Eq. (5)]) (*after* $N \to \infty$). This greatly simplifies the theory as it allows us to use the saddle-point method in theoretical analysis. We refer to such intermediately dense coupling with

$$N \gg c \gg 1 \tag{3}$$

as *dense coupling*.

C. Connection to spin glasses

The feed-forward network made of perceptrons is equivalent to the zero-temperature limit of the transfer matrix of a spin glass with a Hamiltonian

$$H = -\frac{1}{\sqrt{c}} \sum_{\blacksquare} \sum_{k=1}^{c} J_{\blacksquare}^{k} S_{\blacksquare} S_{\blacksquare(k)} \tag{4}$$

as shown in Appendix A. This is a spin-glass model put in a layered structure. Specifying the spin configuration on the boundary $l = 0$, spin configurations at layers $l = 1, 2, \dots, L$ become specified deterministically in the $T \to 0$ limit of the transfer matrix. The perceptrons Eq. (1) just do this operation.

An important point is that there are no direct interactions within each layer, much as the restricted Boltzmann machines (RBMs) [24] which make the operations of the $T = 0$ transfer matrices equivalent to the simple feed-forward nonlinear mappings Eq. (1). In Appendix A we also show that such representations are possible for generic activation functions including the function $\text{sgn}(y)$ which we employ in this paper just as a special case.

For a given set of interactions $J_{\blacksquare}^{k}$, the ground state of the system is unique if the boundaries are allowed to relax. But here we are considering ground states with different realizations of frozen boundaries. Specifying the boundary condition on one side, the configurations on the other side become fixed deterministically.

From this viewpoint, the exponential expressibility of DNNs [25] can be traced back to the chaotic sensitivity of a spin-glass ground state [26,27]. Even if a change of the configuration on the boundary $S_{\blacksquare\in 0}^{\mu} \to S_{\blacksquare\in 0}^{\nu(\neq\mu)}$ is small, the resultant changes of the spin configurations become larger going deeper into the system, $l = 1, 2, \dots$. This can be viewed as an avalanche process. In deeper layers $l = 1, 2, \dots$, larger number of nodes $i = 1, 2, \dots$ will be involved in a single avalanche event. In Sec. V B 4 we discuss a quantity that reflects the avalanche sizes, i.e., the number of nodes involved in a same avalanche caused by $S_{\blacksquare\in 0}^{\mu} \to S_{\blacksquare\in 0}^{\nu(\neq\mu)}$.

D. Teacher-student setting

As shown in Fig. 4, we consider a learning scenario by a teacher machine and a student machine. For simplicity, we assume that the teacher is a "quenched-random teacher": its synaptic weights $\{(J_{\blacksquare}^{k})_{\text{teacher}}\}$ are iid random variables, which take continuous values subjected to the normalization condition Eq. (2).

Training: we generate M sets of training data labeled as $\mu = 1, 2, \dots, M$ as follows. The values of the spins in the input layer $S_{\blacksquare\in 0}^{\mu} = \{S_{0,1}^{\mu}\}_{\text{teacher}}$ are set as iid random Ising numbers ± 1 ($i = 1, 2, \dots, N$, $\mu = 1, 2, \dots, M$) and the corresponding output of the teacher $\{S_{L,i}^{\mu}\}_{\text{teacher}}$ are obtained. The student does training by adjusting its own synaptic weights $\{(J_{\blacksquare}^{k})_{\text{student}}\}$ such that it reproduces perfectly the M sets of the input-output relations of the teacher. More precisely, we consider an idealized setting in which (i) the student has exactly the same architecture as the teacher, including the specific realization of the random network between adjacent layers, and (ii) the student knows exactly the M sets of the input/output relations of the teacher. In short, the student knows everything about the teacher except for its actual values of $\{(J_{\blacksquare}^{k})_{\text{teacher}}\}$. Within the framework of Bayesian inference, this is a so-called Bayes-optimal setting [17,28].

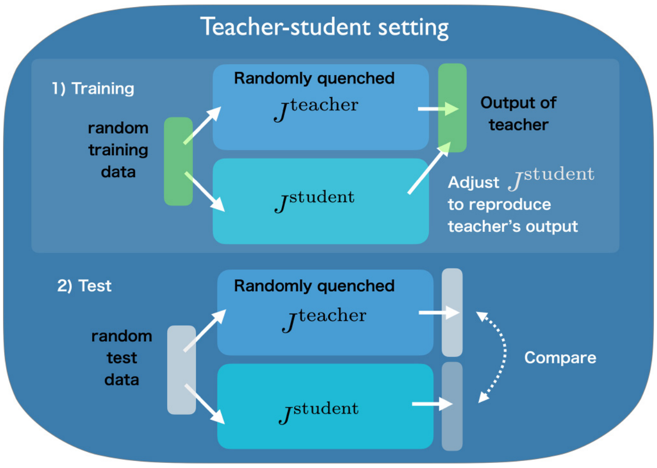

FIG. 4. Schematic pictures of the teacher-student setting.

The configurations of the spins associated with the M-patterns of the training data may be represented by M-component vectors $\mathbf{S}_{l,i} = (S_{l,i}^{1}, S_{l,i}^{2}, \dots, S_{l,i}^{M})$ (see Fig. 1). In the theory, we will consider the $M \to \infty$ limit with

$$\alpha \equiv \frac{M}{c} \tag{5}$$

fixed. Note that our network is parametrized by NcL variational bonds and the NM constrained spin components on the input and output boundaries. The ratio of the two scales as

$$r \equiv \frac{NcL}{NM} = \frac{L}{\alpha}. \tag{6}$$

Test (validation): the generalization ability of the student can be examined empirically using a set of test data. Preparing M' sets of test data as new iid random data $(S_{0,i}^{\mu})_{\text{teacher}}$ ($i = 1, 2, \dots, N$, $\mu = 1, 2, \dots, M'$) (not used for the training), we compare the output of the teacher and student machines. The probability that the student makes an error can be measured as

$$\epsilon = \frac{1}{2}\left(1 - \frac{1}{NM'} \sum_{\mu=1}^{M'} \sum_{i=1}^{N} (S_{L,i}^{\mu})_{\text{teacher}} (S_{L,i}^{\mu})_{\text{student}}\right). \tag{7}$$

If the student is just making random guesses, $\epsilon = 1/2$, while $\epsilon = 0$ if it perfectly reproduces the teacher's output.

E. Gardner's volume

Following the pioneering work by Gardner [8,9], we investigate the ensemble of all possible machines (choices of the synaptic weights $J_{\blacksquare}^{k}$s) of the student which are perfectly compatible with the M set of the input $\mathbf{S}_0$ and output data $\mathbf{S}_L$ provided by the teacher machine (see Figs. 2 and 4). As we noted in Sec. III C, each machine with the feed-forward propagation of signals can be viewed as a zero-temperature limit of the transfer matrix of a spin-glass with a set of $J^k_\blacksquare$'s. So the ensemble of machines is an ensemble of such transfer matrices, which are typically chaotic.

The phase-space volume, which is called Gardner's volume, can be expressed for the present DNN as [15]

$$\begin{aligned} V_M(\mathbf{S}_0, \mathbf{S}_L) &= e^{NM\mathcal{S}(\mathbf{S}_0,\mathbf{S}_l)} \\ &= \left(\prod_{\blacksquare} \mathrm{Tr}_{\mathbf{J}_\blacksquare}\right)\left(\prod_{\blacksquare\backslash\mathrm{output}} \mathrm{Tr}_{\mathbf{S}_\blacksquare}\right) \\ &\quad\times \prod_{\mu=1}^{M}\prod_{\blacksquare} e^{-\beta v(r^\mu_\blacksquare)}, \end{aligned} \tag{8}$$

where $v(r)$ is a hard-core potential,

$$e^{-\beta v(r)} = \theta(r), \tag{9}$$

with $\theta(r)$ being the Heaviside step function, and we introduced the "gap" variable,

$$r^\mu_\blacksquare \equiv S^\mu_\blacksquare \sum_{k=1}^{c} \frac{J^k_\blacksquare}{\sqrt{c}} S^\mu_{\blacksquare(k)}. \tag{10}$$

The trace over the spin and bond configurations can be written explicitly as

$$\mathrm{Tr}_{\mathbf{S}} = \prod_{\mu=1}^{M}\sum_{S^\mu=\pm1}, \quad \mathrm{Tr}_{\mathbf{J}} = \int_{-\infty}^{\infty}\prod_{j=1}^{c} dJ^j\,\delta\!\left(\sum_{k=1}^{c}(J^k)^2 - c\right). \tag{11}$$

In [Eq. (8)], $\blacksquare\backslash$ output means to exclude $\blacksquare$ in the output layer.

The key idea behind the expression Eq. (8) is the *internal representation* [29]: we are considering the spins (neurons) in hidden layers ($l = 1, 2, \ldots, L-1$) as dynamical variables in addition to the synaptic weights. This is allowed because the input-output relation of the perceptrons Eq. (1) is forced to be satisfied by requiring the gap to be positive $r^\mu_\blacksquare > 0$ for all perceptrons $\blacksquare = 1, 2, \ldots, NL$ in the network for all training data $\mu = 1, 2, \ldots, M$ in Eq. (8). As shown in Appendix A, the expression Eq. (8) can also be obtained considering the transfer-matrix representation.

The main quantity of our interest in the present paper is the generalization error ϵ [Eq. (7)]. Gardner's volume V_M provides a way to estimate the generalization ability of the network for the test data [30,31]. The probability that the network, which perfectly satisfies the constraint established by M sets of training data, happens to be compatible with one more unseen datum is given by the ratio V_{M+1}/V_M. Then the generalization error, namely the error probability ϵ, the probability that the configuration of one spin in the output layer $l = L$ of the student machine is wrong (different from the teacher) for a test datum, can be expressed as

$$\epsilon = 1 - \left(\frac{V_{M+1}}{V_M}\right)^{1/N}. \tag{12}$$

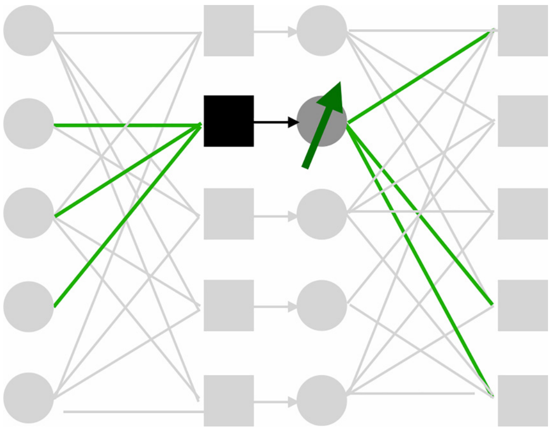

FIG. 5. Variables (green) associated with a perceptron $\blacksquare$ which changes sign by the flip of the gauge variable $\sigma_\blacksquare \to -\sigma_\blacksquare$.

F. Symmetries

Let us note here that there are some symmetries (besides the replica symmetry, which we discuss later) in the present problem. The following becomes important, especially in numerical simulations.

1. Gauge symmetry

For any $\blacksquare$, the system is invariant under gauge transformation,

$$S^\mu_\blacksquare \to \sigma_\blacksquare S^\mu_\blacksquare, \quad \mu = 1, 2, \ldots, M, \tag{13}$$

$$J^k_\blacksquare \to \sigma_\blacksquare J^k_\blacksquare \sigma_{\blacksquare(k)}, \quad k = 1, 2, \ldots, c \tag{14}$$

specified by gauge variables

$$\sigma_\blacksquare = \pm 1, \quad \blacksquare = 1, 2, \ldots, N(L-1). \tag{15}$$

Note that we do not have a gauge transformation in the output layer $l = L$ since the output layer is constrained. It can be easily seen that the gap variables $r^\mu_\blacksquare$ [see Eq. (10)] are invariant under the gauge transformation.

Thus for a given realization of a machine with a set of synaptic weights, there are $2^{(L-1)N}$ completely equivalent machines specified by $2^{(L-1)N}$ possible realizations of the gauge variables: all of them operate exactly in the same way yielding the same output for any input. If the synaptic weights only take Ising values $J^k_\blacksquare = \pm 1$, the number of possible configurations of the machines modulo the gauge symmetry is $2^{NLc^2 - N(L-1)}$.

The presence of the gauge invariance is natural given the connection to the spin glass as mentioned in Sec. III C. While the gauge variables are frozen in spin-glass problems with quenched bonds [32], here the bonds are dynamical variables so that the gauge variables also evolve in time during learning.

Importantly, this is a local symmetry in the sense that a change of any $\sigma_\blacksquare$ induce changes only in the neighborhood of $\blacksquare$ (see Fig. 5): $S^\mu_\blacksquare \to -S^\mu_\blacksquare$ for $\forall\mu$, $J^k_\blacksquare \to -J^k_\blacksquare$ for $\forall k$, $J^l_\square \to -J^l_\square$ for $\forall(\square, l)$ such that $\square(l) = \blacksquare$. This means that in sparse systems with finite connectivity c and $M(=\alpha c)$, the evolution of the machine from one to another connected by a local gauge transformation takes only a finite time in dynamics. Only in the limit $c \to \infty$ do such gauge transformations become frozen in time.

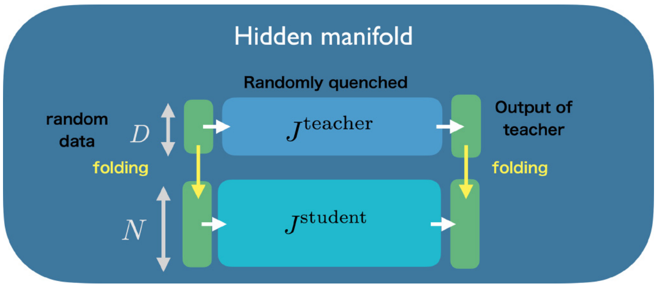

FIG. 6. Schematic picture of the hidden manifold model.

2. Permutation symmetry in globally coupled systems

As we already noted in Sec. III B, in globally coupled systems with $c = N$, the system is invariant under permutations of perceptrons $\blacksquare \in l$ within each layer, $l = 1, 2, \ldots, L$. This symmetry can be removed if the coupling is not global, $c < N$, since we can construct random networks (see Sec. III B).

G. Hidden manifold

We incorporate the hidden manifold model for the data [21] in our model as the following. We replace the original teacher machine of width N with a narrower teacher machine of width $D(\leqslant N)$ (see Fig. 6). The teacher is working entirely in D-dimensional space being subjected to D-dimensional input data and produces D-dimensional output. Student machines are provided N-dimensional input/output data, which are obtained from the D-dimensional input/output of the teacher via folding matrices $F_{i,k}$ of size $N \times D$,

$$(S_{\text{student}})^{\mu}_{0,i} = \text{sgn}\left(\sum_{k=1}^{D} F_{i,k}(S_{\text{teacher}})^{\mu}_{0,k}\right),$$
$$(S_{\text{student}})^{\mu}_{L,i} = \text{sgn}\left(\sum_{k=1}^{D} F_{i,k}(S_{\text{teacher}})^{\mu}_{L,k}\right) \quad (16)$$

for $i = 1, 2, \ldots, N$ and $\mu = 1, 2, \ldots, M$. For the folding matrix, we consider a simple model,

$$F_{i,k} = \begin{cases} 1, & k = \text{mod}(i-1, D) + 1, \\ 0 & \text{otherwise}. \end{cases} \quad (17)$$

In this model, N elements of the data for students are created simply by making N/D copies of the D elements of the teacher's data.

IV. REPLICA THEORY

Now we develop and analyze a replica theory for the statistical mechanics problem of DNNs in the dense limit $N \gg c \gg 1$ introduced in Sec. III B. We first show that it can be solved exactly, overcoming the issue of uncontrolled approximation made in [15]. This is the first main result of this paper. For clarity, we repeat the steps made in [15] and indicate how the problem is resolved. Then we revisit the replica-symmetric solution in the teacher-student setting presented in [15] and analyze it in more detail. Using the exact solution, we analyze the generalization ability of DNNs evaluating the generalization error ϵ via Eq. (12), which is the second main result of this paper. The technical details of the theory are presented in Appendix B.

A. Formalism

1. Order parameters

We are considering the dense coupling Eq. (3) in which (i) perceptrons have large connectivity $c \gg 1$, and (ii) the permutation symmetry of the perceptrons that exist in globally coupled systems $c = N$ is removed. We are also considering a large number of training patterns $M = \alpha c \gg 1$ [see Eq. (5)]. Then we can naturally introduce "local" order parameters associated with each perceptron $\blacksquare$,

$$Q_{ab,\blacksquare} = \frac{1}{c}\sum_{k=1}^{c}(J^{k}_{\blacksquare})^{a}(J^{k}_{\blacksquare})^{b}, \quad q_{ab,\blacksquare} = \frac{1}{M}\sum_{\mu=1}^{M}(S^{\mu}_{\blacksquare})^{a}(S^{\mu}_{\blacksquare})^{b}. \quad (18)$$

The overlaps between the teacher and student machines are represented by $Q_{0b,\blacksquare} = Q_{b0,\blacksquare}$ and $q_{0b,\blacksquare} = q_{b0,\blacksquare}$ $(b = 1, 2, \ldots, s)$ while those between the student machines are represented by $Q_{ab,\blacksquare} = Q_{ba,\blacksquare}$ and $q_{ab,\blacksquare} = q_{ba,\blacksquare}$ $(a, b = 1, 2, \ldots, s)$.

It is important to note that the order parameters $Q_{ab,\blacksquare}$ and $q_{ab,\blacksquare}$ defined above change sign under the change of the gauge variable $\sigma^{a}_{\blacksquare}$, which can be defined independently for each replica $(a = 1, 2, \ldots, n)$ (see Fig. 5). Thus they trivially vanish in thermal equilibrium in sparse systems with finite connectivity c. Only in the dense limit $c \to \infty$ can the gauge variables $\sigma^{a}_{\blacksquare}$ be considered as slow variables.

It is natural to expect that order parameters are homogeneous within each layer since we will take the average over realization of random connections between adjacent layers [see Eq. (23)]. Thus we assume they only depend on the index l of layers,

$$Q_{ab,\blacksquare} = Q_{ab}(l), \quad l = 1, 2, \ldots, L-1,$$
$$q_{ab,\blacksquare} = q_{ab}(l), \quad l = 0, 1, 2, \ldots, L-1, L. \quad (19)$$

Here we have included, for our convenience, the spin overlaps at the boundaries $l = 0, L$ where spins of all student replicas $a = 1, 2, \ldots, s$ are forced take the same values as the teacher $a = 0$,

$$q_{ab}(0) = 1, \quad q_{ab}(L) = 1. \quad (20)$$

Note also that the normalization condition for the bonds [Eq. (2)] and the spins (which take Ising values ± 1) implies $Q_{aa}(l) = q_{aa}(l) = 1$ for $\forall a$ and $\forall l$.

The order parameters also vanish in thermal equilibrium in globally coupled system with $c = N$ due to the permutation symmetry—the second symmetry mentioned in Sec. III F. This issue is removed by using the dense coupling by selecting connections between adjacent layers randomly.

2. Replicated Gardner volume, free energy

Let us introduce the replicated Gardner's volume, where the teacher machine is included as the zeroth replica,

$$\begin{aligned} V^{1+s}(\mathbf{S}_0, \mathbf{S}_L) &= e^{NM\mathcal{S}_{1+s}(\mathbf{S}_0,\mathbf{S}_L)} \\ &= \prod_{a=0}^{s} \left(\prod_{\blacksquare} \mathrm{Tr}_{\mathbf{J}^a_{\blacksquare}} \right) \left(\prod_{\blacksquare \backslash \mathrm{output}} \mathrm{Tr}_{\mathbf{S}^a_{\blacksquare}} \right) \\ &\quad \times \left\{ \prod_{\mu,\blacksquare,a} e^{-\beta v(r^{\mu}_{\blacksquare,a})} \right\} \end{aligned} \tag{21}$$

with

$$r^{\mu}_{\blacksquare,a} \equiv (S^{\mu}_{\blacksquare})^a \sum_{k=1}^{c} \frac{(J^k_{\blacksquare})^a}{\sqrt{c}} \left(S^{\mu}_{\blacksquare(k)}\right)^a. \tag{22}$$

Here the output $\mathbf{S}_L$ is the output of the teacher $a = 0$. The main object we are interested in is the free-energy functional (Franz-Parisi's potential [33]),

$$\begin{aligned} \frac{-\beta F[\{\hat{Q}(l), \hat{q}(l)\}]}{NM} &= \frac{\partial_s \overline{V^{1+s}(\mathbf{S}_0, \mathbf{S}_L(\mathbf{S}_0, (\mathcal{J}^k_{\blacksquare})^0)}|_{s=0}}{NM} \\ &= \partial_s s_{1+s}[\{\hat{Q}(l), \hat{q}(l)\}]\big|_{s=0}, \end{aligned} \tag{23}$$

where the overline denotes the average over (i) the random inputs $\mathbf{S}_0$ imposed commonly on all machines and (ii) realization of random connections between adjacent layers.

It turns out that the dense coupling [Eq. (3)] $N \gg c \gg 1$ allows us to obtain the exact expression for the replicated Gardner volume for $n = 1 + s$ replicas in terms of the order parameters [Eq. (19)],

$$\begin{aligned} s_n[\{\hat{Q}(l), \hat{q}(l)\}] = \frac{1}{\alpha} \sum_{l=1}^{L} s_{\mathrm{ent,bond}}[\hat{Q}(l)] + \sum_{l=1}^{L-1} s_{\mathrm{ent,spin}}[\hat{q}(l)] \\ - \sum_{l=1}^{L} \mathcal{F}_{\mathrm{int}}[\hat{\lambda}(l)] \end{aligned} \tag{24}$$

with

$$\lambda_{ab}(l) = q_{ab}(l-1) Q_{ab}(l) q_{ab}(l). \tag{25}$$

Here $s_{\mathrm{ent,bond}}[\hat{Q}(l)]$ and $s_{\mathrm{ent,spin}}[\hat{q}(l)]$ are the entropic part of the free-energy associated with bonds and spins, respectively, and $-\mathcal{F}_{\mathrm{int}}[\hat{q}(l-1), \hat{Q}(l), \hat{q}(l)]$ is the interaction part of the free energy. Fortunately, the free-energy functional obtained in [15] turns out to be valid in the dense limit $N \gg c \gg 1$, although it is unjustified for the global coupling $c = N$ assumed there. The details of the expressions and the derivation are presented in Appendix B 5.

The main reasons for the success, which allows us to overcome the problems in [15], are the three points discussed in Sec. III B. First, the sparseness of the network allows us to safely neglect the contribution of loops, as we explain in detail in Appendix B 4. Second, the random connections between adjacent layers eliminate the permutation symmetry that exists in globally coupled system $c = N$. Third, the limit $c \to \infty$ allows us to use the standard saddle-point method to evaluate thermodynamic quantities exactly.

3. Replica-symmetric ansatz

Since our current problem is a Bayes-optimal inference problem [17,28], we can safely assume a replica-symmetric (RS) solution,

$$\begin{aligned} (a, b = 1, \ldots, s) \quad & Q_{ab}(l) = [1 - Q(l)]\delta_{ab} + Q(l), \\ & q_{ab}(l) = [1 - q(l)]\delta_{ab} + q(l), \\ (a = 1, \ldots, s) \quad & Q_{0a}(l) = Q_{a0}(l) = R(l), \\ & q_{0a}(l) = q_{0a}(l) = r(l) \end{aligned} \tag{26}$$

for $l = 1, 2, \ldots, L$ and the Nishimori condition (see Sec. V B 3),

$$Q(l) = R(l), \quad q(l) = r(l), \tag{27}$$

which must hold in Bayes-optimal cases. The saddle-point equations, which extremize the replicated free energy, are obtained in [15]. It can be checked that the saddle-point equations can verify the relation Eq. (27).

4. Generalization error

Based on the above results, we can analyze the error probability Eq. (12), which is the main object of our interest in the present paper. Using the free energy (23) and (24), we readily find it as

$$\epsilon = 1 - \exp\left(\sum_{l=1}^{L-1} \partial_s s_{\mathrm{ent,spin}}[\hat{q}(l)]|_{s=0} - \sum_{l=1}^{L} \partial_s \mathcal{F}_{\mathrm{int}}[\hat{\lambda}(l)]|_{s=0} \right). \tag{28}$$

Explicit expressions of the free energy needed to evaluate the above quantity are given in Appendix B 6.

B. Analysis

1. Order parameters

We numerically solved the saddle-point equations (B41) to obtain the order parameters repeating the analysis in [15] but in a wider parameter space. In Fig. 7 we show the spatial profile of the order parameters. As already shown in [15], the theory predicts spatially heterogeneous learning. It can be seen that the "crystalline" phase with a finite order parameter (inference of the teacher's configuration is successful) grows with increasing α starting from the input/output boundaries. This is reminiscent of wetting transitions [34–36]. Details of the behavior of the order parameters are displayed in Fig. 8.

The central region remains in the liquid phase with a zero order parameter (where inference of the teacher's configuration is impossible) until the two crystalline phases meet in the center at a critical point, $\alpha_{c1}(L)$. Naturally, $\alpha_{c1}(L)$ increases with L. For $\alpha > \alpha_{c1}(L)$ the central liquid phase is absent. As far as $\alpha < \alpha_{c1}(L)$, we find that the crystalline parts attached to the two opposite boundaries grow with α, but the profiles of the order parameters remain independent of L.

Now for $\alpha > \alpha_{c1}(L)$, where the liquid phase is absent, the order parameters depend explicitly on the depth L. One may regard this as a "finite depth L effect." At some larger α it becomes difficult to follow the saddle-point solution numerically. Presumably this implies spinodal instability associated with a first-order transition to another solution $q = Q = 1$

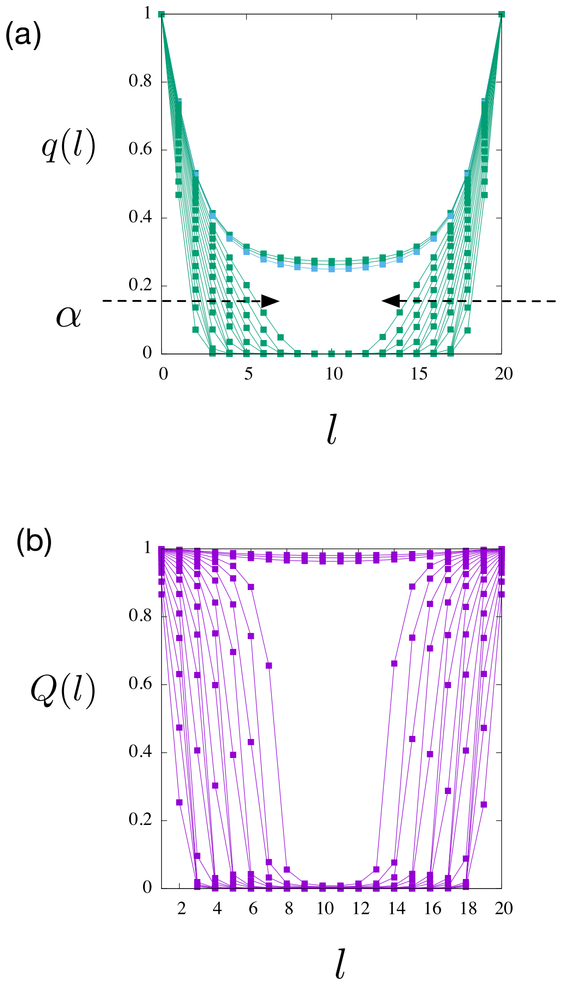

FIG. 7. Spatial profile of the order parameter obtained by solving the replica-symmetric saddle-point equations. Top: the overlap of spins (neurons), and bottom: the overlap of bonds (synaptic weights). Here $L = 20$. Different lines corresponds to $\alpha = 16 - 10^3$ with equal spacing in $\ln \alpha = 0.23 \ldots$.

(which is a saddle-point solution) at some $\alpha_{c2}(L)$ [$> \alpha_{c1}(L)$]. Such a discontinuous "perfect recovery" behavior has been found in the case of a single perceptron with binary couplings [37]. We skip a detailed analysis of the first-order transition and leave it for future works. Up to the discontinuous change, the evolution of the order parameters with increasing α is continuous.

2. *Generalization errors*

Now we turn to the generalization error ϵ (see [Eq. (28)]) evaluated by the replica symmetric solution obtained above, which is of our main interest in the present paper. It is obtained as shown in the bottom panels of Figs. 8 and 9. The relation ϵ versus α is referred to often as *learning curves*. Without learning $\alpha = 0$, $\epsilon = 1/2$ because the student just makes random guesses. The learning curves $\epsilon = \epsilon_L(\alpha)$ consist of two parts, as follows.

(i) Let us recall that for sufficiently small α, the two crystalline phases at the boundaries remain disconnected from each other separated by the liquid phase in the center and that the profiles of the order parameters are independent of L since the two crystalline regions do not meet, as we discussed in Sec. IV B 1. In this regime, the learning curve does not depend on the depth L, i.e., $\epsilon = \epsilon_\infty(\alpha)$. The reason is that the contribution from the liquid region where $q(l) = Q(l) = 0$ to ϵ [Eq. (28)] is just zero: it contributes neither positively nor negatively to ϵ. On the other hand, the crystalline region where $q(l), Q(l) > 0$ contributes negatively to ϵ [Eq. (28)], and it is independent of L as long as the two crystalline regions do not meet. It is remarkable that $\epsilon_\infty(\alpha) < 1/2$ and it decreases with increasing α: the system generalizes even though the central part is in the liquid phase due to overparametrization.

(ii) Increasing α, the crystalline phases meet at some critical point $\alpha_{c1}(L)$ and the central liquid phase disappears. Note again that $\alpha_{c1}(L)$ is larger for larger L. For sufficiently large $\alpha > \alpha_{c1}(L)$, where the central liquid gap is filled up by the crystalline phase, the learning curve depends on the depth L as the order parameters now depend on L. For even larger $\alpha > \alpha_{c2}(L)$, we speculate that ϵ jumps to 0 due to the first-order transition mentioned in Sec. IV B 1.

The L-independent behavior of the learning curve can be seen in Fig. 9 as follows. For instance, one can see that $\epsilon_L(\alpha)$ of $L = 10, 20$ are indistinguishable for $1/\alpha > 0.01$, and $L = 10, 20, 5$ are indistinguishable for $1/\alpha > 0.1$.

C. Finite-width N/dimension D effects and finite connectivity c effects

In reality, DNNs have some finite width N and finite connectivity c, while in the theory we assumed an idealized situation, namely the dense limit $N \gg c \gg 1$ and $M \gg 1$ with fixed $\alpha = M/c$. It is very important to consider the effects of finite width N and finite connectivity c (and M).

1. *Finite width N effect*

The effects of finite width N can be attributed to the corrections due to geometrically closed loops in the network, which become non-negligible when the width N is finite, as we discussed in Sec. III B. The simplest is the one shown in Fig. 3, which connects three adjacent layers. More extended ones exist as shown in Fig. 20, which connect many layers. As we showed in Sec. III B, the probability to have such a geometrically closed single loop is proportional to c/N no matter how extended it is. The probability vanishes in $N \to \infty$ but exists as long as N is finite.

Most importantly, the loops connect different layers and different nodes within the same layer inducing correlations inside the network. Indeed, as discussed in detail Appendix B 4, the loops yield finite width N corrections to the interaction part of the free energy. We also note that the symmetry concerning the exchange of input/output sides present in the saddle-point solutions (see Fig. 7) becomes lost in the presence of such loop correction terms.

2. *Finite hidden dimension D effect*

It is interesting to discuss here the hidden manifold model [21] introduced in Sec. III G. Let us recall that our original model contains no correlations within the boundaries. We can consider the effect of the correlations put in the input/output

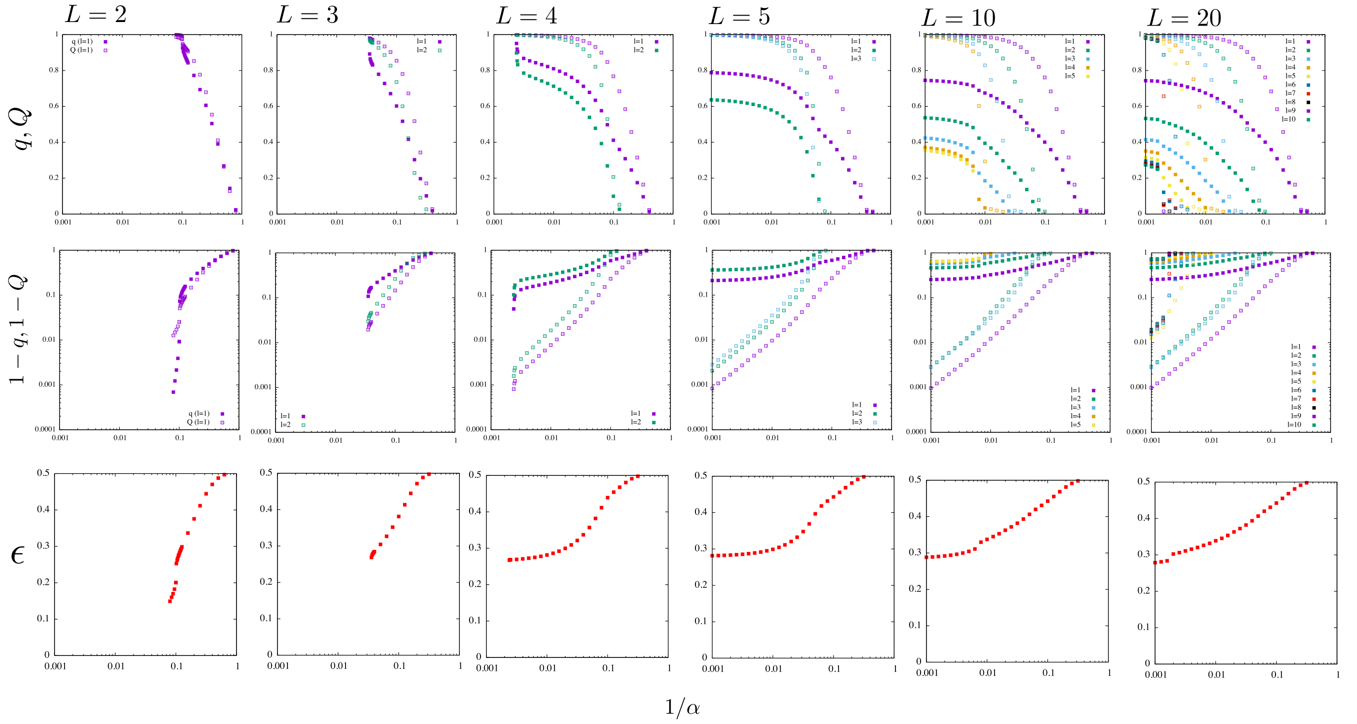

FIG. 8. Order parameters and generalization error obtained by solving the replica-symmetric saddle-point equations. In the panels on the first and second rows, overlaps of spins $q(l)$ (filled symbols) and $Q(l)$ bonds (open symbols) are shown. In the bottom row, the generalization error ϵ is shown.

boundaries by the hidden manifold model in a perturbative manner around the replica-symmetric saddle-point solution as the following.

Within the simplest model [Eq. (17)] for the folding matrix F, the same values are repeated in the input (output) data on different nodes i $(= 1, 2, \ldots, N)$. This induces additional closed loops. For example, the unclosed loop in panel (b) of Fig. 10 becomes closed if the input data at k_1 and k_2 are forced to take the same value by the simplest hidden manifold model. This means that finite width N effects become enhanced as the effective dimension D becomes smaller. This consideration implies that the finite width N effect and the finite hidden dimension D effect will be similar. Both will lead to an increase of correlations inside the network.

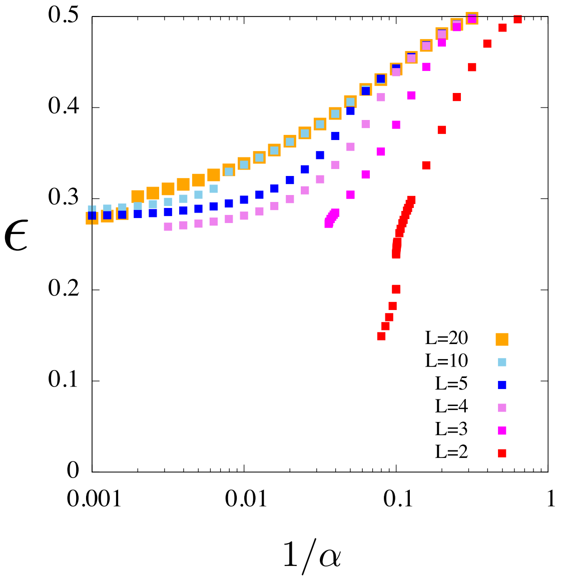

FIG. 9. Learning curves of DNN with various depths L obtained by solving the replica-symmetric saddle-point equations.

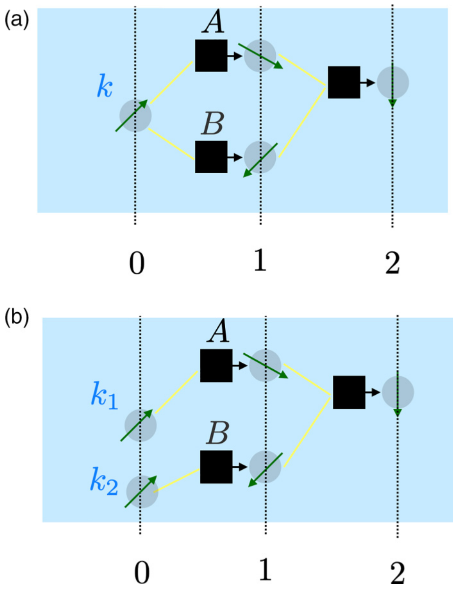

FIG. 10. Schematic picture of the closed and unclosed loop at the boundary.

Let us note that the teacher and students have different architectures in the hidden manifold model so that the inference by the students is not Bayes-optimal. In such a circumstance, the replica symmetry is not guaranteed. We leave the analysis beyond the perturbative analysis for future works.

3. Finite connectivity c effect

Finally, in the $N \to \infty$ limit, we will still be left with finite connectivity c effects. In our theoretical analysis, we assumed $c \to \infty$, which allowed us to perform the saddle-point computations. One can naturally consider $1/c$ corrections, taking into account contributions from fluctuations around the saddle point as sketched in Appendix B 7.

Naturally the fluctuating field around the saddle point induces correlations inside the network. Let us also note that the cubic term in the expansion breaks the symmetry concerning the exchange of input/output sides (see Appendix B 7 b).

4. Discussions

The corrections due to the loops (Sec. IV C 1 and Sec. IV C 2) and those due to the fluctuations around the saddle points (Sec. IV C 3) can be easily separated considering the dense limit $N \gg c \gg 1$, which is difficult for the case of global coupling $c = N$. Nonetheless, we found that the two corrections bring qualitatively similar effects: (i) correlations inside the network, and (ii) asymmetry with respect to the exchange of input/output sides.

We consider that the two effects, which disappear in the dense limit, are important in practice in the following respects:

(i) Remnant symmetry-breaking field. One would wonder: how a student machine can recognize the existence of the two crystalline regions (teacher's configuration) if the two are separated by the liquid region as in Fig. 7?" For an algorithm to work in this situation, some remnant symmetry breaking field should help the student. We consider the corrections due to the loops and the fluctuations around the saddle point play this role.

(ii) Input-output asymmetry. One would also wonder: how a DNN with the feed-forward propagation of information can have such spatial profile which is completely symmetric concerning the exchange of input/output sides as in Fig. 7? We consider the corrections due to the loops and the fluctuations around the saddle point are responsible for the breaking of this symmetry.

V. SIMULATION

Now let us discuss Monte Carlo simulations on the same model we analyzed theoretically. We first explain the simulation method in Sec. V A, we introduce the observables in Sec. V B, and then we present the results in Sec. V C.

A. Method

1. Learning scenarios

We simulate the teacher-student scenario (see Sec. III D) in the Bayes-optimal setting and the setting with the hidden manifold model (see Sec. III G).

(i) *Bayes-optimal Scenario.*

(a) *Network*: Teacher and student machines have the same rectangular network of width N and depth L (see Fig. 1). The rectangular network is created as a random graph as the following. Every $\blacksquare \in l$ is given c arms. Each of the arms is connected to a $\blacksquare \in l-1$ chosen randomly out of N possible ones.

(b) *Synaptic weights of teacher machine*: The teacher's synaptic weights $\{(J^k_{\blacksquare})_{\text{teacher}}\}$ for $\blacksquare \in 1, 2, \ldots, L$ and $k = 1, 2, \ldots, c$ are prepared as iid random numbers drawn from the Gaussian distribution with zero mean and unit variance.

(c) *Data*: M set of training data is prepared as follows. First the common input data for all machines including the teacher and students are prepared as iid random numbers $(S_{\text{teacher}})^{\mu}_{0,i} = \pm 1$ for $i = 1, 2, \ldots, N$ and $\mu = 1, 2, \ldots, M$. Then the output $(S_{\text{teacher}})^{\mu}_{L,i}$ for $i = 1, 2, \ldots, N$ and $\mu = 1, 2, \ldots, M$ are obtained by the feed-forward propagation of the signal using Eq. (1). These outputs are used as the target outputs $(S_*)^{\mu}_{L,i}$ to train the student machines (see below), i.e., $(S_*)^{\mu}_{L,i} = (S_{\text{teacher}})^{\mu}_{L,i}$. Another M' set of data for the test (validation) are created in the same way.

(ii) *Hidden Manifold Scenario* [21].

(a) *Network*: The networks of the teacher and student machines are the rectangular, random regular network as in the Bayes-optimal scenario but the teacher machine is narrower than the student machine, i.e., $D < N$ (see Fig. 6).

(b) *Synaptic weights of teacher machine*: The teacher's synaptic weights are prepared in the same manner as in the case of the Bayes-optimal scenario.

(c) *Data*: M sets of data for training and another M' sets of data for the test (validation) are created in the same way as the following. Pairs of input/output data of the teacher's machine is created just as in the case of Bayes-optimal scenario but with D replacing N. Then the N-dimensional inputs $(S_{\text{student}})_{0,i}$ for $i = 1, 2, \ldots, N$ to be given to the student machines are created using the simple folding matrix F given by Eq. (17). Similarly, the N-dimensional target output $(S_*)_{L,i}$ for $i = 1, 2, \ldots, N$ for the student machines is created using the same folding matrix F.

2. Learning algorithm: Greedy Monte Carlo method

For a set of temporal synaptic weights of a student machine $\{(J^k_{\blacksquare})_{\text{student}}\}$ for $\blacksquare \in 1, 2, \ldots, L$ and $k = 1, 2, \ldots, c$, we obtain the output data $(S_{\text{student}})_{L,i}$ $(i = 1, 2, \ldots, N)$ for a given input data $(S_{\text{student}})_{0,i}$ $(i = 1, 2, \ldots, N)$ using the feed-forward propagation based on Eq. (1).

To train the student machines, we use a simple zero-temperature or greedy Monte Carlo algorithm. We introduce the loss function defined as

$$E = \sum_{i=1}^{N} \sum_{\mu=1}^{M} |(S_{\text{student}})^{\mu}_{L,i} - (S_*)^{\mu}_{L,i}|, \tag{29}$$

where $(S_*)^{\mu}_{L,i}$ is the target output data defined above. Note that the loss function takes discrete values. In particular, we are interested with the ensemble of student machines in the $E = 0$ space, whose phase-space volume is just Gardner's volume.

Starting from a set of initial synaptic weights, the student machines are updated as follows:

(i) Select a perceptron $\blacksquare$ randomly out of the $N_{\blacksquare}$ possible ones and select a link k randomly out of the c possible ones,

$k = 1, 2, \ldots, c$. Then propose a new synaptic weight,

$$(J^k_{\blacksquare})^{\text{new}}_{\text{student}} = \frac{(J^k_{\blacksquare})_{\text{student}} + \delta x}{\sqrt{1+\delta^2}}, \tag{30}$$

where δ is a parameter and x is an iid random number drawn from the Gaussian distribution with zero mean and unit variance. Note that $(J^k_{\blacksquare})^{\text{new}}_{\text{student}}$ is normalized such that its variance remains to be 1.

(ii) Accept the proposed one if the resultant loss function *does not increase*. Otherwise reject it and go back to (i). Importantly, we accept updates by which the loss function remains unchanged. This is crucial to allow exploration of the $E = 0$ (SAT) space.

Within one Monte Carlo step (MCS), we repeat the above procedure $N_{\blacksquare}c$ times.

We simulate learning by two student machines “1” and “2” which are subjected to the same training data but evolve independently from each other using statistically independent random numbers for steps (i) and (ii) explained above.

3. Learning and unlearning

For the training, we consider the following two protocols:

Learning: the initial synaptic weights of the student machines $\{(J^k_{\blacksquare})_{\text{student}}\}$ are prepared just as iid Gaussian random numbers totally uncorrelated with the teacher's weights $\{(J^k_{\blacksquare})_{\text{teacher}}\}$.

To facilitate the training, we perform a sort of “annealing.” At a given time t (MCS), perform the greedy Monte Carlo update using a subset of the training data of size $M_{\text{batch}}(t)(< M)$. Starting from $M_{\text{batch}}(0) = 1$, increase $M_{\text{batch}}(t)$ logarithmically in time t progressively adding more data to the training data set such that $M_{\text{batch}}(t_{\max}) = M$ in the end of the simulation at $t_{\max}$ (MCS).

Unlearning (or planting): the initial synaptic weights of the student machines $\{(J^k_{\blacksquare})_{\text{student}}\}$ are set to be exactly the same as the teacher's weights $\{(J^k_{\blacksquare}\}_{\text{teacher}}\}$. The student machine explores the $E = 0$ (SAT) space.

If the greedy Monte Carlo method equilibrated the system, the two protocols should yield the same results for macroscopic observables, which we explain below, after averaging over time and/or initial configurations in the stationary state.

B. Observables

1. Simple overlaps

We are interested in the similarity between different machines in the hidden layers $l = 1, 2, \ldots, L-1$. To quantify this, we first introduce, between the two student machines “1,” “2” and the teacher machine “0,” the following “simple” overlaps:

$$q(l) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{\mu=1}^{M}(S_1)^{\mu}_{l,i}(S_2)^{\mu}_{l,j}, \tag{31}$$

$$r(l) = \frac{1}{2NM}\sum_{i=1}^{N}\sum_{\mu=1}^{M}(S_0)^{\mu}_{l,j}[(S_1)^{\mu}_{l,i} + (S_2)^{\mu}_{l,i}]. \tag{32}$$

Here $\mu = 1, 2, \ldots, M$ for the training data and $\mu = 1, 2, \ldots, M'$ for the test data (and replace the factor $1/M$ by $1/M'$ in the latter case).

These are the same as the order parameters for the spins used in the replica theory [see Eq. (18)]. However, as mentioned in Sec. IV A 1, the expectation value of the simplest overlaps defined above vanishes in thermal equilibrium because of the local gauge symmetry (and the permutation symmetry in the case $c = N$), as discussed in Sec. III F.

2. Squared overlaps

To overcome the above problem, we define the following order parameters, which we refer to as *squared overlaps*, and which are invariant under the symmetry operations. Let us first introduce

$$q_{ab,ij}(l) = \frac{1}{M}\sum_{\mu=1}^{M}(S_a)^{\mu}_{l,i}(S_b)^{\mu}_{l,j}. \tag{33}$$

$$\tag{34}$$

Here a and b are indices for machines: 0 for the teacher machine, 1 and 2 for the student machines. Then we introduce the squared overlaps as

$$q_{2,ab}(l) = \frac{1}{N}\sum_{i,j=1}^{N}[q_{ab,ij}(l)]^2 - \frac{N}{M}. \tag{35}$$

We note that this is analogous to the order parameter used in numerical simulations of vectorial spin-glass models which have rotational symmetry in spin space [38].

Finally, we introduce the normalized version of the squared overlap,

$$q_2(l) = \frac{q_{2,12}(l)}{\sqrt{q_{2,11}(l)}\sqrt{q_{2,22}(l)}},$$
$$r_2(l) = \frac{q_{2,01}(l) + q_{2,02}(l)}{\sqrt{q_{2,00}(l)}[\sqrt{q_{2,11}(l)} + \sqrt{q_{2,22}(l)}]}. \tag{36}$$

Interestingly, these are very similar to the measure proposed in [39] referred to as “centered kernel alignment.”

3. Nishimori condition

Since our teacher-student scenario is a Bayes-optimal inference, we have

$$q(l) = r(l), \quad q_2(l) = r_2(l), \quad l = 1, 2, \ldots, L. \tag{37}$$

This is a Nishimori condition which must hold in Bayes-optimal inferences [17,28,32]. These relations are useful to check the equilibration of the system.

4. Physical meaning of the squared overlaps

Let us discuss more closely the significance of the squared overlap defined in Eq. (35) and its normalized version Eq. (36). In the following, we denote the average over different realization of the inputs as $\overline{\cdots}^{\text{input}}$. From Eq. (34), we can write

$$\overline{[q_{ab,ij}(l)]^2}^{\text{input}} = \frac{1}{M} + \frac{1}{M}\sum_{\mu=1}^{M}\frac{1}{M}\sum_{\nu(\neq\mu)}\overline{r^{\mu\to\nu}_{a,i}r^{\mu\to\nu}_{b,j}}^{\text{input}}$$
$$\simeq \frac{1}{M} + \overline{r^{\mu\to\nu(\neq\mu)}_{a,i}r^{\mu\to\nu(\neq\mu)}_{b,j}}^{\text{input}} \tag{38}$$

with

$$r_{a,i}^{\mu\to\nu} = (S_a)_{li}^{\mu}(S_a)_{li}^{\nu}. \tag{39}$$

Here $r_{a,i}^{\mu\to\nu}$ can be regarded as a change of the sign of the spin (neuron) at the node (l,i) of student-a when the input pattern is changes from μ to ν. Using the above expression, we find that Eq. (35) with the subtraction term $-N/M$ becomes

$$q_{2,ab}(l) \simeq \frac{1}{N}\sum_{j=1}^{N} \overline{r_{a,i}^{\mu\to\nu(\neq\mu)} r_{b,j}^{\mu\to\nu(\neq\mu)}}^{\text{input}}. \tag{40}$$

This can be viewed as a kind of *correlation volume* within layer l in the following sense.

(i) *Totally uncorrelated random machines.* Suppose that student-a and student-b are totally uncorrelated (far beyond the trivial difference by the gauge transformation and the permutation) randomly generated machines. Then we naturally expect $\overline{r_{a,i}^{\mu\to\nu(\neq\mu)} r_{b,j}^{\mu\to\nu(\neq\mu)}}^{\text{input}} = 0$. This means that the minimum value of the squared overlap $q_{2,ab}(l)$ is 0.

(ii) *Same random machine modulo gauge transformation and permutation.* On the other hand, if the two machines are the same machine modulo the gauge transformation and permutation, we can write

$$\overline{r_{a,i}^{\mu\to\nu(\neq\mu)} r_{b,j}^{\mu\to\nu(\neq\mu)}}^{\text{input}} = \delta_{ij} + (1-\delta_{ij})\overline{r_{a,i}^{\mu\to\nu(\neq\mu)} r_{b,j}^{\mu\to\nu(\neq\mu)}}^{\text{input}}. \tag{41}$$

Thus in this case the squared overlap $q_{2,ab}(l)$ is at least 1 and can be larger.

In the case of the perceptrons with random synaptic weights and the highly nonlinear activation function [see Eq. (1)], we expect $\overline{r_{a,i}^{\mu\to\nu(\neq\mu)} r_{b,j}^{\mu\to\nu(\neq\mu)}}^{\text{input}}$ becomes significant also between different nodes $i \neq j$. This is because of the chaos effect, which we discussed in Sec. III C: it is known that in such a nonlinear random feed-forward network, a slight change of the input induces chaotic changes in the state of spins (neuron) as the signal propagates deeper into the network [25]. This is an avalanche-like process so that the correlation $\overline{r_{a,i}^{\mu\to\nu(\neq\mu)} r_{b,j}^{\mu\to\nu(\neq\mu)}}^{\text{input}}$ for $i \neq j$ becomes more significant with increasing l. In this case, the squared overlap $q_{2,ab}(l)$ can be viewed as a measure of avalanche size within layer l.

(iii) *General case.* Based on the above consideration, we naturally expect that in general $q_{2,ab}(l)$ quantifies the avalanche size and similarity of the avalanche patterns taking place in machines a and b through changes of inputs $\mu \to \nu(\neq \mu)$. Then it becomes clear that the normalized version [Eq. (36)] quantifies the similarity of the avalanche patterns in machines a and b.

5. Generalization error

To measure the generalization ability of the student machines (see Sec. III D), we measure

$$r_{\text{out}} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{\mu=1}^{M'}(S_{\text{student}})_{l,i}^{\mu}(S_{\text{teacher}})_{l,i}^{\mu}. \tag{42}$$

Here we use the M' sets of the outputs of the teacher and student machines for the test data (not used for training). The generalization error [see Eq. (12)] can be evaluated as

$$\epsilon = \tfrac{1}{2}(1 - r_{\text{out}}). \tag{43}$$

In the above expression, we used simple overlap defined on the output layer $l = L$ (see [Eq. (32)]). Note that there are no gauge transformations or permutations on the output layer.

C. Results

Now let us discuss the results of the simulations. First, we discuss the equilibration process through the learning and unlearning protocols (see Sec. V A 3). Next, we discuss the equilibrium properties of macroscopic observables.

In the simulations, we used $\delta = 0.1$ to generate new weights by Eq. (30). For the test, we used $M' = M$ data uncorrelated with the training data. In the following, observables are averaged over 240 statistically independent samples (different realizations of the teacher machine, initial configurations of student machines for learning, and realizations of random numbers used in Monte Carlo updates).

1. Learning

In Fig. 11, we present the relaxation of the loss function Eq. (29) in the learning protocol (see Sec. V A 3). It can be seen in panel (a) that relaxation of the loss function slows down by increasing the number of the training data $M = c\alpha$. On the other hand, it can be observed in panel (b) that relaxation becomes faster upon increasing the depth L of the network.

As shown in panel (c), the relaxation depends also on the width N but converges in large enough N with fixed c, and α suggests that relaxation time is finite in systems with finite connectivity c even in the $N \to \infty$ limit. For larger c, the relaxation curves converge to a slower curve as shown in panel (d), suggesting that the relaxation time becomes larger for larger connectivity c.

2. Unlearning

In Fig. 12, we show the simple overlaps defined in Eq. (32) observed in the unlearning protocol, which explores the $E = 0$ landscape (SAT phase) (see Sec. V A 3). Note that $q(l) = r(l) = 1$ at the beginning. The student machines become decorrelated from the teacher machine and also from each other as time t elapses. It is interesting to note that relaxation is inhomogeneous in space: relaxation is faster in the central part of the network and slower closer to the input/output boundaries.

It is important to note that the complete vanishing of the simple overlaps does not necessarily mean that the solution space is completely in a liquid state as the overlaps are not gauge-invariant. Because of the gauge symmetry (see Sec. III F 1), even machines that are completely the same as the teacher machine modulo the gauge transformation can have vanishing simple overlap with the teacher machine. Indeed, we will find below that normalized squared overlaps [Eq. (36)] (which are gauge-invariant) instead indicate correlations between different machines.

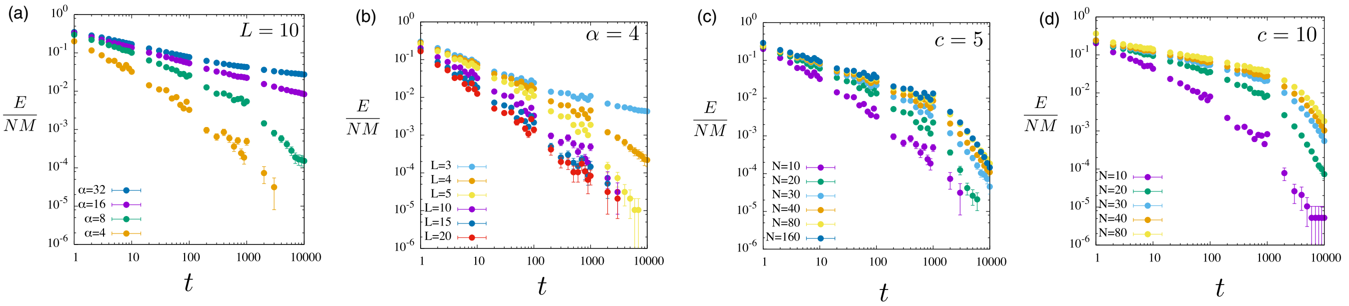

FIG. 11. Relaxation of the loss function in learning [annealing with $t_{\max} = 10^4$ (MCS)] observed by the MC simulation. In all cases, $N = 10$. (a) Various $\alpha = M/c$ with $L = 10$ and $c = 5$. (b) Various L with $\alpha = M/c = 4$ and $c = 5$. (c) Various N with $L = 10$, $\alpha = 4$, and $c = 5$. (d) The same as (c) but with $c = 10$. The unit of time t is 1 (MCS).

The inhomogeneity of the relaxation observed here suggests that the system is more constrained closer to the boundary while the center is freer. We have also observed that the deeper system relaxes faster, as shown in Fig. 11(b). These may be interpreted as an echo of the "crystal-liquid-crystal" sandwich structure predicted by the theory (Fig. 7).

3. Equilibration

In equilibrium, learning and unlearning protocols should give the same results for macroscopic observables after sufficiently long times. This is indeed verified, as shown in the top panels (a), (c), and (e) of Fig. 13. In panels (a) and (c) we show the normalized squared overlaps defined in Eq. (36), which are invariant under the gauge transformations. The normalized squared overlaps of unlearning and learning protocols agree, suggesting the establishment of equilibrium. Furthermore, it can be seen that the Nishimori condition $q_2(l) = r_2(l)$ [see Eq. (37)] expected for the Bayes-optimal inferences becomes satisfied after sufficiently long times. This is additional evidence of thermal equilibration. Equilibration can also be seen in panel (e), where we show the simple overlap between the teacher and student machines in the output layer $l = L$ for the test data.

As can be seen in Fig. 13, the spatial profiles of the normalized squared overlaps $q_2(l)$ and $r_2(l)$ are strongly inhomogeneous in space. As discussed in Sec. V B 4, we consider that the normalized squared overlaps quantify the similarity of the avalanche patterns taking place in different machines through changes of inputs $\mu \to \nu(\neq \mu)$. At the beginning of unlearning, which starts from the teacher's configuration, the normalized squared overlaps take high values. On the other hand, they are small at the beginning of learning, which is not surprising because teacher and student machines are totally uncorrelated at the beginning. In equilibrium, they converge to a nontrivial, spatially nonmonotonic function. This implies that the equilibrium phase is not just a liquid, as we might have thought based on the observation of the vanishing simple overlap (Fig. 12). On the contrary, the gauge-invariant quantity shows that the student machines are strongly correlated with each other and with the teacher machine in equilibrium. The spatial nonmonotonicity means that they become less correlated with each other in the center (beyond the trivial difference by the gauge transformations) while they are similar to each other (modulo the gauge transformation) closer to the input and output boundaries. This observation can be regarded as another echo of the spatial inhomogeneity predicted by the theory (Fig. 7). From the theoretical point of view, the strong asymmetry concerning the exchange of input/output sides, which is absent in the saddle-point solution, may be attributed to the finiteness of the width N and the connectivity c, as we discussed in Sec. IV C 4.

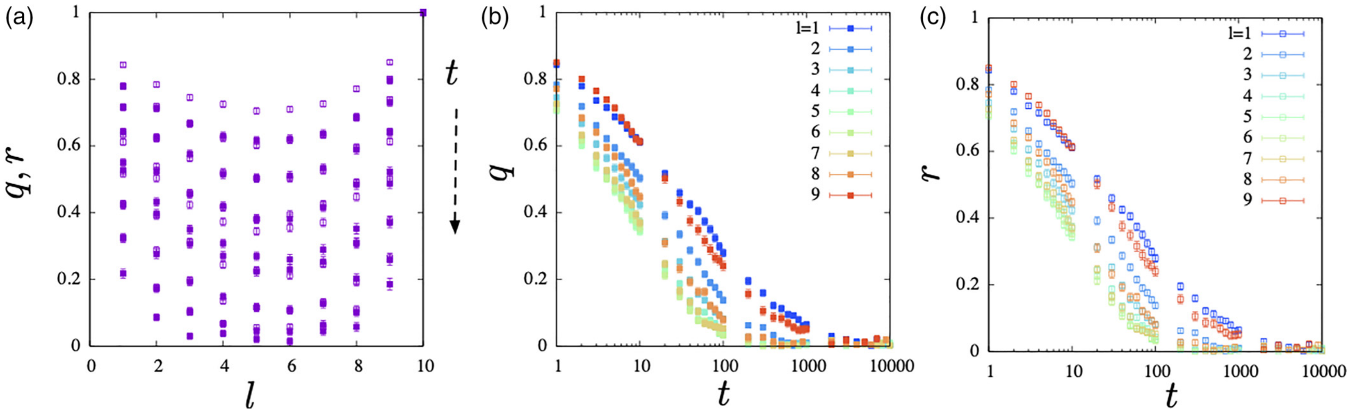

FIG. 12. Time evolution of simple student-student overlap $q(l)$ (filled symbols) and teacher-student overlap $r(l)$ (open symbols) observed in MC simulations of the unlearning protocol. Here $N = 10$, $\alpha = 4$, $c = 5$. Panel (a) shows data at $t = 1, 2, 4, 8, 10, 20, 40, 80$. Panels (b) and (c) show the simple student-student overlap $q(l)$ and simple teacher-student $r(l)$ overlap, respectively.

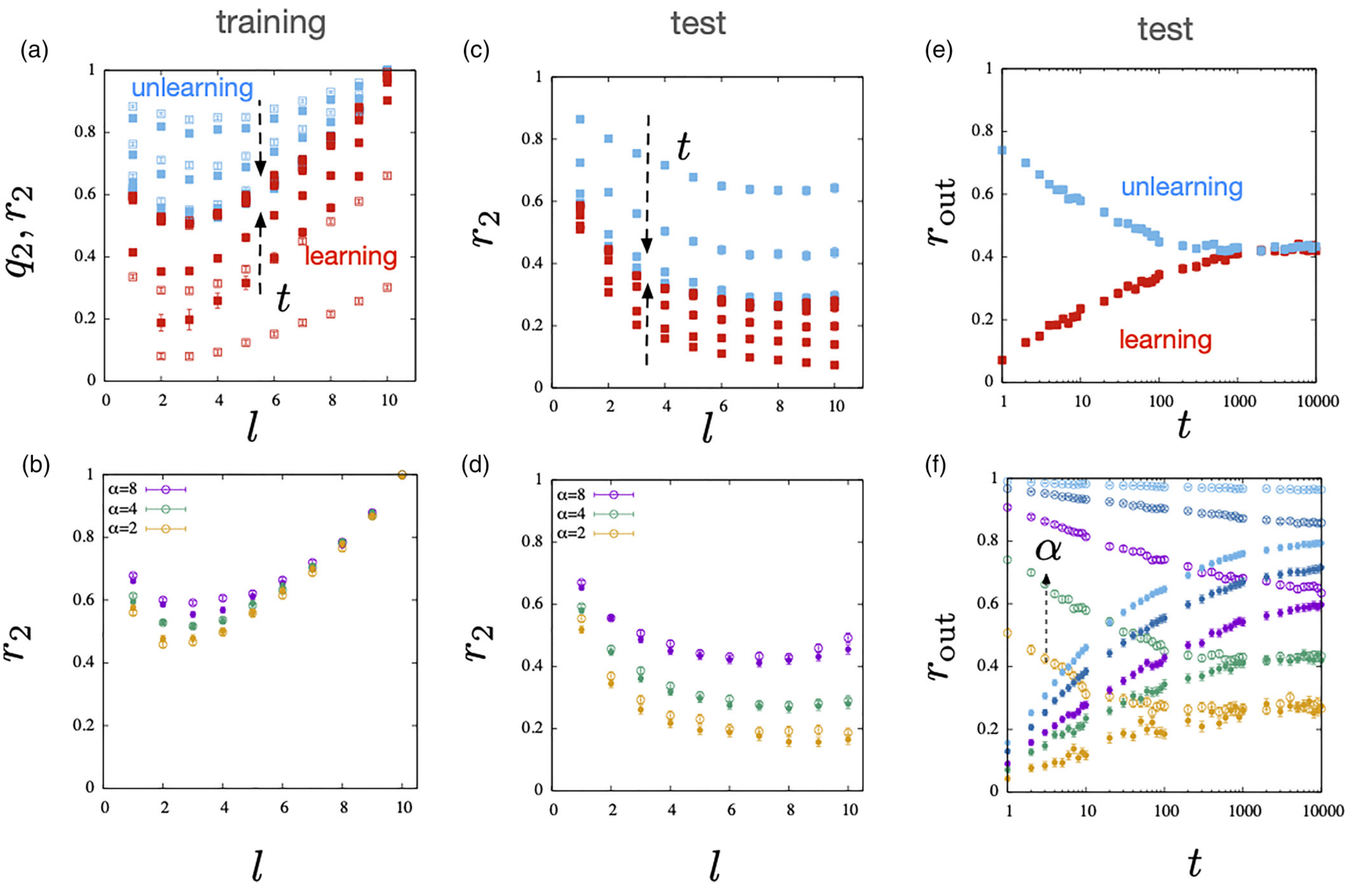

FIG. 13. Spatial profile of the normalized squared teacher-student overlaps $r_2(l)$ and student-student overlaps $q_2(l)$ [see Eq. (36)] for training (a),(b) and test (c),(d), and time evolution of the simple teacher-student overlap $r(L)$ in the output layer ($l = L$) for test (e),(f) [See Eq. (42)]. All data are obtained by the MC simulations. In panels (a), (c), and (e), data of $q_2(l)$ (open symbols) and $r_2(l)$ (filled symbols) of learning/unlearning are represented by red/blue points ($\alpha = 4$). Panels (a) and (c) show the normalized squared overlaps at various times $t = 1, 10, 100, 1000, 10\,000$ (increasing along the arrows) at each layer. Panel (e) shows the time evolution of the simple teacher-student overlap $r(L)$ at the output layer ($l = L$) for the test data. Panels (b) and (d) show the normalized squared teacher-student overlap for unlearning (open symbols)/learning (filled symbols) at $\alpha = 2, 4, 8$ at $t = 10^4$ (MCS). Panel (f) show the time evolution of the simple teacher-student overlap $r(L)$ at the output layer ($l = L$) for the test data, obtained by unlearning (open symbols)/learning (filled symbols) protocols with $\alpha = 2, 4, 8, 16, 32$. Here $N = 10$, $L = 10$, and $c = 5$ for all data.

As shown in the bottom panels (b), (d), and (f) in Fig. 13, the overlaps increase as α increases, as expected. In panel (f) it can be seen that the dynamics of both learning and unlearning slow down as α increases.

4. Typical student machines

Now let us examine further the equilibrium properties, i.e., properties of typical student machines sampled in the solution space. We show in Fig. 14 some data of the normalized squared overlaps $q_2(l)$. It can be seen again that the data obtained by both learning (filled symbols) and unlearning (open symbols) agree, confirming that the system is equilibrated. Quite remarkably, the equilibrium normalized squared overlap $q_2(l)$ evolves nonmonotonically in space for large enough N and c. It first decreases with l but finally increases with l. This means that avalanches taking place in different machines become decorrelated in the middle of the network but strongly correlated closer to the input and output boundaries. It appears that the situation has become closer to the "crystal-liquid-crystal" sandwich structure predicted by the theory (Fig. 7) increasing N and c.

In Sec. IV C 4 we discussed that the corrections due to the loops and fluctuations around the saddle point can be separated considering the dense limit $N \gg c \gg 1$, but they bring about similar effects: (i) a remnant symmetry-breaking field, and (ii) input-output asymmetry. Indeed, it can be seen in Fig. 14 that for fixed connectivity c, the normalized squared overlaps decrease around the center of the network, and the asymmetry with respect to the exchange of input/output becomes weaker as N increases. Furthermore, the data suggest convergence in the $N \to \infty$ limit with fixed c. Comparing panels (a) and (b), it can be seen that the remnant asymmetry becomes weaker as the connectivity c increases. Remarkably, decorrelation in the center also becomes clearer upon increasing N and c, suggesting the emergence of a liquidlike region in the center due to overparametrization, as suggested theoretically in the dense limit $N \gg c \gg 1$.

In Fig. 15 we show the normalized squared overlaps $q_2(l)$, $r_2(l)$ and the generalization error ϵ in systems with $\alpha = 4$,

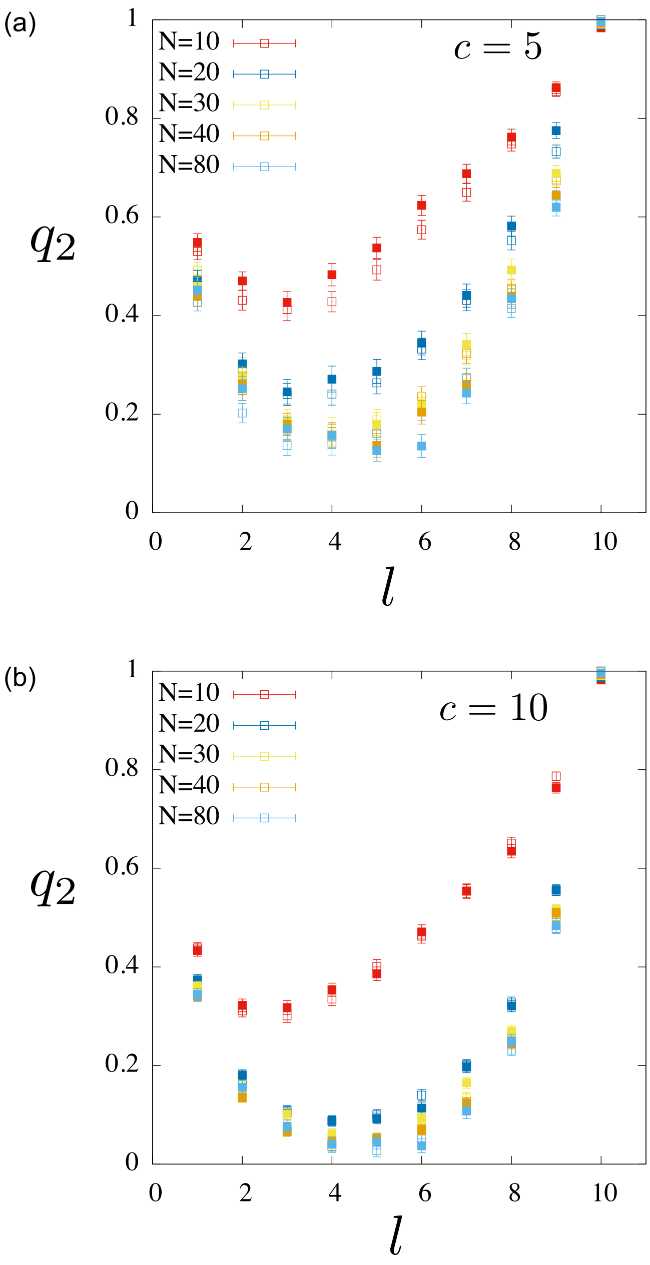

FIG. 14. Finite N effect and finite c effect: the asymmetry becomes smaller as N increases for unlearning (open symbols)/learning (filled symbols). $N = 10, 20, 30, 40, 80$ with $\alpha = 1.0$ and (left) $c = 5$, (right) $c = 10$. All data are obtained by MC simulations of $t = 10^4$ (MCS).

$c = 5$ observed after $t = 10^4$ (MCS) in systems with different depth $L = 5, 10, 20$. As shown in the top panels (a), (c), and (e), data obtained by both learning (filled symbols) and unlearning (open symbols) agree, proving again that the system is equilibrated. As shown in panels (a) and (c), we find again that the normalized squared overlaps are strongly inhomogeneous in space. The machines decorrelate more concerning each other in the central region in deeper systems, but correlations recover approaching the output layer. We also find again that normalized squared overlaps increase significantly and that the asymmetry concerning the exchange of the input and output sides becomes stronger upon decreasing the width N.

Now let us turn to the effect of finite-dimension D introduced by the hidden manifold model (see Sec. V A 1). The results of simulations on the hidden manifold model are displayed in panels (b) and (d) of Fig. 15. Here we used the simplest folding matrix F of the form Eq. (17), but we obtained qualitatively the same results (not shown) also in the case of random matrices. Comparing the panels (b) to (a) and (d) to (c), we immediately notice that the effect of hidden dimension D is quite similar to the effect of finite width N: decreasing D with fixed N is much like decreasing $N(= D)$. We conjecture that this is due to the enhancement of the loop corrections induced by the closing of the loops by the correlated inputs as discussed in Sec. IV C 2.

Finally, let us discuss the generalization error ϵ shown in panels (e) and (f) of Fig. 15. In panel (e) we also show the generalization error ϵ obtained by the theory in the dense limit $c \to \infty$ [see Eq. (12) and Fig. 9]. Remarkably, the effect of finite width N and hidden dimension D is very similar again. The generalization error improves significantly either by decreasing $N(= D)$ or D with fixed N. Presumably this is due to the increase of correlations inside the network induced by the loop corrections. Moreover, the generalization error becomes independent of the depth L at sufficiently deep systems, much like the theoretical prediction. The result implies that the generalization ability first decreases, making the system deeper, but it does not vanish even in the $L \to \infty$ limit. This is consistent with the L-independent learning curve $\epsilon = \epsilon_\infty(\alpha)$ predicted theoretically (see Sec. IV B 2).

VI. CONCLUSIONS

In the present paper, we obtained an exactly solvable statistical mechanics model of machine learning by a deep neural network (DNN) in the dense limit $N \gg c \gg 1$. Exact solutions are obtained using the replica method developed in [15]. We used the replica theory to analyze the generalization ability of the DNN in the Bayes-optimal teacher-student setting. The learning curve $\epsilon = \epsilon_L(\alpha)$ becomes independent of the depth L as long as the two crystalline phases attached to input/output boundaries are separated by the liquid phase in the center. Thus the system is predicted to generalize even in the limit $L \to \infty$ where the system becomes extremely overparametrized. We discussed the loop corrections to the dense limit and argued that finite width N and finite hidden dimension D effects appear similarly. Both should lead to an increase of correlations inside the network.

In simulations, the simple greedy Monte Carlo method turned out to work efficiently to enable sampling of typical machines in equilibrium, suggesting the simplicity of the loss landscape. The main obstacle in simulations is the gauge invariance of the system by which order parameters in the original simple form vanish. To overcome the difficulty, we measured the normalized squared overlap, which quantifies the correlation of avalanches concerning changes in input data between different machines. It is a gauge (and permutation) invariant quantity that reflects the similarity between machines modulo the gauge (and permutation) symmetries. The result is qualitatively consistent with the theoretical prediction that overparametrization in the DNN leads to spatially inhomogeneous learning: students become close to the teacher

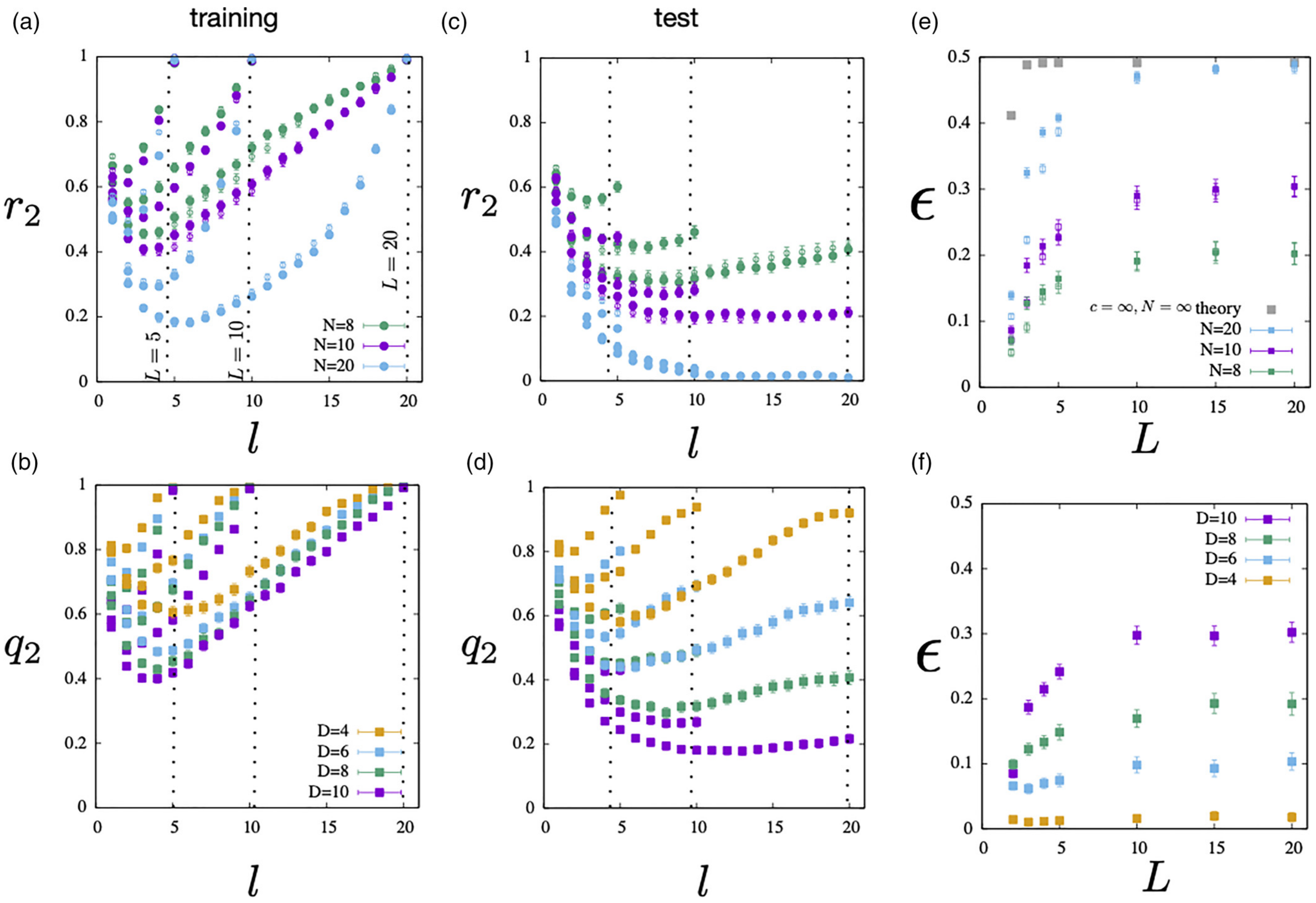

FIG. 15. Spatial profile of the normalized squared overlaps and the generalization error in systems with various widths N (top panels) and hidden dimension $D(\leqslant N)$ (bottom panels) obtained by MC simulations. In the top panels, data obtained by both learning (filled symbols) and unlearning (open symbols) are shown. Panels (a) and (b) show the normalized squared overlaps for training, while (c) and (d) show those for the test. Panels (e) and (f) show the generalization error ϵ [see Eq. (43)]. In panel (e) we also show the generalization error ϵ obtained by the theory in the dense limit: $N \to \infty$ followed by $N \to \infty$ [see Eq. (12) and Fig. 9]. In panels (a)–(d) data with $L = 5, 10, 20$ are shown. In panels (a), (c), and (e) data with $N = 8, 10, 20$ are shown. In panels (b), (d), and (f) data with $D = 4, 6, 8, 10$ are shown. In all cases, $\alpha = 4$, $c = 5$, and $t = 10^4$.

around the input/output boundaries, while they remain only weakly correlated in the center. We note that a liquidlike central region was also noticed in [40]. Furthermore, somewhat counterintuitively but in agreement with the theoretical prediction, the generalization error first increases the depth L but then becomes independent of it, suggesting that the generalization ability survives in the $L \to \infty$ limit. Simulations confirm that finite width N and finite hidden dimension D effects are quite similar and lead similarly to significant improvements in the generalization ability. Presumably this reflects the increase of correlations inside the network due to the loop corrections. As we noted in Sec. IV C 4, we consider that the corrections due to the loops and fluctuations around the saddle point play the role of a symmetry-breaking field, which allows the student to recognize the teacher again in spite of the liquidlike center.

Finally, what is the advantage of making the system deeper? One important advantage is that the learning dynamics become faster, increasing the depth, as we found numerically. This should be due to the presence of the central region, where the system is less constrained. We believe that this point will become more important as we move away from the idealized, Bayes-optimal teacher-student setting we considered in the present work. From a theoretical point of view, there is no guarantee that replica symmetry continues to hold as we move away from the Bayes-optimal situation toward the situations in the real world. For example, one can consider a noisy teacher-student scenario by adding noise to the training data provided by the teacher. Then the situation becomes closer to the random scenario considered in [15], where complex replica-symmetry breaking (RSB) was found in the DNN. In the latter case, RSB evolves in space such that the hierarchy of RSB becomes simplified layer-by-layer approaching the center so that the central region can remain in the replica-symmetric liquid phase if the network is made deep enough. This implies the deeper system will relax faster even in the presence of the RSB around the boundaries.

There are numerous directions in which to generalize and extend the present work. Let us mention a few of them here. It is straightforward to study the model exactly with the parameter α depending on space $\alpha = \alpha(l)$. This amounts to making the width N of the network vary in space $N = N(l)$.

It will be interesting to study how one can control the spatial heterogeneity of learning by changing $\alpha(l)$. The dense limit $N \gg c \gg 1$ will be useful not only for the replica theory but also for other theoretical approaches. For instance, it should be possible to develop cavity approaches in the dense limit. It will also be very interesting to generalize our theory considering more general activation functions, as we noted in Sec. III C.

ACKNOWLEDGMENTS

We thank G. Biroli, A. G. Cavaliere, Y. Kabashima, and R. Karakida for useful discussions. Numerical simulations presented in this work were done using the supercomputer systems SQUID at the Cybermedia Center, Osaka University. This work was supported by JSPS KAKENHI Grants No. 21K1846 and No. 22H05117.

APPENDIX A: TRANSFER-MATRIX REPRESENTATION OF FEED-FORWARD NETWORKS

Here we show that transfer-matrix representations of a family of spin-glass models put in the layered geometry (like in Fig. 1) become feed-forward DNNs in the zero-temperature limit.

In the following, we denote the configuration of the set of spins in the lth layer as $\mathbf{S}_l^\mu = \{S^\mu_{\blacksquare \in l}\}$, where μ is the label to specify a data set. Let us write the conditional probability to realize a spin configuration $\mathbf{S}_L^\mu$ on the output boundary given a spin configuration $\mathbf{S}_0^\mu$ on the input boundary as

$$P(\mathbf{S}_0^\mu \to \mathbf{S}_L^\mu) = \left(\prod_{l=1}^{L-1} \mathrm{Tr}_{\mathbf{S}_l^\mu} \right) \prod_{l=1}^{L} P(\mathbf{S}_{l-1}^\mu \to \mathbf{S}_l^\mu), \tag{A1}$$

where $P(\mathbf{S}_{l-1}^\mu \to \mathbf{S}_l^\mu)$ is the conditional probability to realize a spin configuration $\mathbf{S}_l^\mu$ on the lth layer given a spin configuration $\mathbf{S}_{l-1}^\mu$ on the $(l-1)$th layer,

$$P(\mathbf{S}_{l-1}^\mu \to \mathbf{S}_l^\mu) = \frac{\langle \mathbf{S}_{l-1}^\mu | \mathbf{T}_l | \mathbf{S}_l^\mu \rangle}{\mathrm{Tr}_{\mathbf{S}_l^\mu} \langle \mathbf{S}_{l-1}^\mu | \mathbf{T}_l | \mathbf{S}_l^\mu \rangle}, \tag{A2}$$

where $\mathbf{T}_l$ is the transfer matrix from the $(l-1)$th to the lth layer.

1. Layered Ising spin-glass model

Let us first consider the layered Ising spin-glass model Eq. (4) with the restricted Boltzmann machine (RBM) -like architecture [24],

$$H = -\frac{1}{\sqrt{c}} \sum_{\blacksquare} \sum_{k=1}^{c} J_\blacksquare^k S_\blacksquare S_{\blacksquare(k)}, \tag{A3}$$

where the spins are Ising variables $S^\mu = \pm 1$. The matrix elements of the transfer matrix are given by

$$\langle \mathbf{S}_{l-1}^\mu | \mathbf{T}_l | \mathbf{S}_l^\mu \rangle = e^{\sum_{\blacksquare \in l} \sum_{k=1}^{c} \frac{\beta J_\blacksquare^k}{\sqrt{c}} S_\blacksquare^\mu S_{\blacksquare(k)}^\mu}. \tag{A4}$$

Here $\beta = 1/T$ is the inverse of the temperature T. The size of the matrix is $2^N \times 2^N$. Then we find that the conditional probability Eq. (A2) becomes

$$P(\mathbf{S}_{l-1}^\mu \to \mathbf{S}_l^\mu) = \prod_{\blacksquare \in l} P(\mathbf{S}_{l-1}^\mu \to S_\blacksquare^\mu), \tag{A5}$$

where

$$\prod_{\blacksquare \in l} P(\mathbf{S}_{l-1}^\mu \to S_\blacksquare^\mu) = \prod_{\blacksquare \in l} \frac{e^{\beta r_\blacksquare^\mu}}{\mathrm{Tr}_{S_\blacksquare^\mu} e^{\beta r_\blacksquare^\mu}}, \tag{A6}$$

with $r_\blacksquare^\mu$ being the gap variable Eq. (10),

$$r_\blacksquare^\mu = S_\blacksquare^\mu \sum_{k=1}^{c} \frac{J_\blacksquare^k}{\sqrt{c}} S_{\blacksquare(k)}^\mu. \tag{A7}$$

It is important to notice the factorization of the conditional probability in Eq. (A5). It is a consequence of the RBM-type architecture: there are no direct interactions within each layer.

Taking the zero-temperature limit $T \to 0$ $(\beta \to \infty)$, we find

$$P(\mathbf{S}_{l-1}^\mu \to \mathbf{S}_l^\mu) \xrightarrow[T \to 0]{} \prod_{\blacksquare \in l} \theta(r_\blacksquare^\mu), \tag{A8}$$

where $\theta(r)$ is the Heaviside step function. Thus in this limit, the configuration $S_\blacksquare^\mu$ in $\blacksquare \in l$ is established deterministically given $\mathbf{S}_\square^\mu$ in $\square \in l-1$ by the perceptron's rule Eq. (1),

$$S_\blacksquare^\mu = \mathrm{sgn}\left(\sum_{k=1}^{c} \frac{J_\blacksquare^k}{\sqrt{c}} S_{\blacksquare(k)}^\mu \right). \tag{A9}$$

Note that the operation of the transfer matrix of size $2^N \times 2^N$, Eq. (A4), is now replaced in the $T \to 0$ limit by simple nonlinear mapping of a much lower computational cost of $O(Nc)$ thanks to (i) the RBM-like network structure, and (ii) the $T \to 0$ limit.

Similarly, we find

$$\lim_{T \to 0} P(\mathbf{S}_0^\mu \to \mathbf{S}_L^\mu) = \left(\prod_{\blacksquare \backslash \mathrm{output}} \mathrm{Tr}_{\mathbf{S}_\blacksquare^\mu} \right) \prod_{\blacksquare} \theta(r_\blacksquare^\mu), \tag{A10}$$

which means $S_{\blacksquare \in L}$ in the output layer becomes determined by the multilayer perceptron for a given $S_{\blacksquare \in 0}$ in the output layer. Note that Gardner's volume Eq. (8) can be written as

$$V_M(\mathbf{S}_0, \mathbf{S}_L) = \lim_{T \to 0} \left(\prod_{\blacksquare} \mathrm{Tr}_{\mathbf{J}_\blacksquare} \right) \prod_{\mu} P(\mathbf{S}_0^\mu \to \mathbf{S}_L^\mu). \tag{A11}$$

In this representation it becomes clear that the traces over the spins in the hidden layers used in Gardner's volume for DNN Eq. (8) (the internal representation [29]) are equivalent to computation of the products of transfer matrices in Eq. (A1).

2. Layered spin-glass models with continuous spins

Now let us consider a class of slightly more generalized models on the RBM-type network with the Hamiltonian

$$H = -\frac{1}{\sqrt{c}} \sum_{\blacksquare} \sum_{k=1}^{c} J_\blacksquare^k S_\blacksquare S_{\blacksquare(k)} + \sum_{\blacksquare} U(S_\blacksquare). \tag{A12}$$

Here we consider spins that take continuous values $-\infty < S < \infty$. The function $U(x)$ in the second term of the Hamiltonian Eq. (A12) represents a confining potential to regularize the spins. Then the conditional probability Eq. (A2) becomes

$$P(\mathbf{S}_{l-1}^\mu \to \mathbf{S}_l^\mu) = \prod_{\blacksquare \in l} \frac{e^{\beta(h_\blacksquare^\mu S_\blacksquare^\mu - U(S_\blacksquare^\mu))}}{\mathrm{Tr}_{S_\blacksquare^\mu} e^{\beta(h_\blacksquare^\mu S_\blacksquare^\mu - U(S_\blacksquare^\mu))}}, \tag{A13}$$

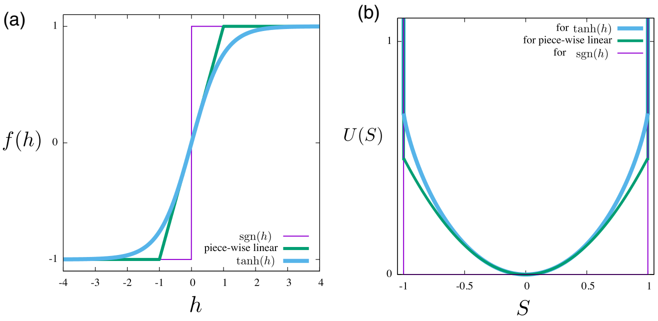

FIG. 16. Some examples of the activation function $f(h)$ and the associated confining potential $U(S)$. Here $u = 1$ for the piecewise linear function.

where

$$h^{\mu}_{\blacksquare} = \sum_{k=1}^{c} \frac{J^{k}_{\blacksquare}}{\sqrt{c}} S^{\mu}_{\blacksquare(k)}. \tag{A14}$$

In the $T \to 0$ limit, we find

$$P(\mathbf{S}^{\mu}_{l-1} \to \mathbf{S}^{\mu}_{l}) \to \prod_{\blacksquare \in l} \delta(\mathbf{S}^{\mu}_{\blacksquare} - f(h^{\mu}_{\blacksquare})), \tag{A15}$$

where the function $f(h)$ is determined such that

$$f(h) = \text{argmin}_S[-hS + U(S)]. \tag{A16}$$

Assuming that the potential $U(s)$ is differentiable, we find

$$f^{-1}(S) = \frac{dU(S)}{dS} \tag{A17}$$

or

$$U(S) = \int_{-\infty}^{S} f^{-1}(S) dS. \tag{A18}$$

Thus given a layered spin-glass model with a confining local potential $U(S)$, Eq. (A12), we find a corresponding feed-forward neural network with the activation function $f(h)$ in the $T \to 0$ limit.

In Fig. 16 we display some examples of activation functions $f(h)$ and the associated confining potentials $U(S)$. In Appendix A 1 we showed that the transfer matrix of the layered Ising spin-glass model becomes equivalent to the feed-forward network with the activation function $f(h) = \text{sgn}(h)$ in the $T \to 0$ limit. The same activation function can be obtained also from the continuous spin model using the confining potential

$$U(S) = \begin{cases} 0 & (-1 < u < 1), \\ \infty & (u < -1 \quad \text{or} \quad u > 1). \end{cases} \tag{A19}$$

Similarly, for a piecewise linear function parametrized by $u > 0$,

$$f(h) = \begin{cases} 1 & (h > u), \\ \frac{h}{u} & (-u < h < u), \\ -1 & (h < -u), \end{cases} \tag{A20}$$

we find

$$U(S) = \begin{cases} \frac{u}{2}\left(S - \frac{h}{u}\right)^2 + \frac{h^2}{u} & (-1 < u < 1), \\ \infty & (u < -1 \quad \text{or} \quad u > 1). \end{cases} \tag{A21}$$

Finally in the case of $f(h) = \tanh(h)$ one can easily find $U(S) = S \tanh^{-1}(S) + (1/2) \ln(1 - S^2)$.

APPENDIX B: DETAILS OF THE REPLICA THEORY

1. Replicated Gardner volume, Fourier transformation, and Legendre transformation

By introducing a Fourier representation of the Boltzmann factor,

$$e^{-\beta v(r)} = \int \frac{d\eta}{\sqrt{2\pi}} W_{\eta} e^{-i\eta r}, \tag{B1}$$

the replicated Gardner volume Eq. (21) can be rewritten as

$$\begin{aligned} V^{n}&(\mathbf{S}_0, \mathbf{S}_L) \\ &= e^{NM\mathcal{S}_n(\mathbf{S}_0, \mathbf{S}_l)} \\ &= \prod_{a} \left(\prod_{\blacksquare} \text{Tr}_{\mathbf{J}^{a}_{\blacksquare}} \right) \left(\prod_{\blacksquare \backslash \text{output}} \text{Tr}_{\mathbf{S}^{a}_{\blacksquare}} \right) \left\{ \prod_{\mu, \blacksquare, a} e^{-\beta v(r^{\mu}_{\blacksquare, a})} \right\} \\ &= \prod_{\mu, \blacksquare, a} \left\{ \int \frac{d\eta_{\mu, \blacksquare, a}}{\sqrt{2\pi}} W_{\eta_{\mu, \blacksquare, a}} \right\} \tilde{V}^{n}(\mathbf{S}_0, \mathbf{S}_L), \end{aligned} \tag{B2}$$

where we introduced the Fourier transform of the replicated Gardner volume

$$\tilde{V}^{n}(\mathbf{S}_0, \mathbf{S}_L) = \prod_{a} \left(\prod_{\blacksquare} \text{Tr}_{\mathbf{J}^{a}_{\blacksquare}} \right) \left(\prod_{\blacksquare \backslash \text{output}} \text{Tr}_{\mathbf{S}^{a}_{\blacksquare}} \right) \times \prod_{\mu, \blacksquare, a} e^{i\eta_{\mu, \blacksquare, a} r^{\mu}_{\blacksquare, a}} \tag{B3}$$

with the gap variable $r^{\mu}_{\blacksquare, a}$ defined as

$$r^{\mu}_{\blacksquare, a} \equiv (S^{\mu}_{\blacksquare})^{a} \sum_{k=1}^{c} \frac{(J^{k}_{\blacksquare})^{a}}{\sqrt{c}} (S^{\mu}_{\blacksquare(k)})^{a}. \tag{B4}$$

Introducing the identities

$$\begin{aligned} 1 &= \prod_{a<b} \int_{-\infty}^{\infty} \int_{-i\infty}^{i\infty} \left(\frac{c}{2\pi i} \right) dQ_{ab, \blacksquare} d\epsilon_{ab, \blacksquare} e^{c \sum_{a<b} \epsilon_{ab, \blacksquare} (Q_{ab, \blacksquare} - c^{-1} \sum_{k=1}^{c} (J^{k}_{\blacksquare})^{a} (J^{k}_{\blacksquare})^{b})}, \\ 1 &= \prod_{a<b} \int_{-\infty}^{\infty} \int_{-i\infty}^{i\infty} \left(\frac{M}{2\pi i} \right) dq_{ab, \blacksquare} d\varepsilon_{ab, \blacksquare} e^{M \sum_{a<b} \varepsilon_{ab, \blacksquare} (q_{ab, \blacksquare} - M^{-1} \sum_{\mu=1}^{M} (S^{\mu}_{\blacksquare})^{a} (S^{\mu}_{\blacksquare})^{b})}, \end{aligned} \tag{B5}$$

we can express the Fourier transformation of the replicated Gardner volume $\tilde{V}^{n}$ as

$$\tilde{V}^{n}(\mathbf{S}_0, \mathbf{S}_L) = \prod_{a<b, \blacksquare} \left\{ \int_{-\infty}^{\infty} dQ_{ab, \blacksquare} \right\} \prod_{a<b, \blacksquare \backslash \text{output}} \left\{ \int_{-\infty}^{\infty} dq_{ab, \blacksquare} \right\} e^{-\beta \tilde{F}_n[\hat{Q}, \hat{q}]}, \tag{B6}$$

where we introduced

$$e^{-\beta\tilde{F}_n[\hat{Q},\hat{q}]} = \prod_{a<b,\blacksquare}\left\{\int_{-i\infty}^{i\infty}\left(\frac{c}{2\pi i}\right)d\epsilon_{ab,\blacksquare}\right\}\prod_{a<b,\blacksquare\backslash\text{output}}\left\{\int_{-i\infty}^{i\infty}\left(\frac{M}{2\pi i}\right)d\varepsilon_{ab,\blacksquare}\right\} \\ \times e^{c\sum_{\blacksquare}\sum_{a<b}\epsilon_{ab,\blacksquare}Q_{ab,\blacksquare}+M\sum_{\blacksquare\backslash\text{output}}\sum_{a<b}\varepsilon_{ab,\blacksquare}q_{ab,\blacksquare}}e^{-\beta\tilde{G}_n[\hat{\epsilon},\hat{\varepsilon}]} \tag{B7}$$

with

$$-\beta\tilde{G}_n[\hat{\epsilon},\hat{\varepsilon}] = -\beta G_{n,0}^{\text{bond}}[\hat{\epsilon}] - \beta G_{n,0}^{\text{spin}}[\hat{\varepsilon}] + \ln\left(\left\langle\exp\left[i\sum_{\mu,\blacksquare,a}\eta_{\mu,\blacksquare,a}(S_{\blacksquare}^{\mu})^a\sum_{k=1}^{c}\frac{(J_{\blacksquare}^k)^a}{\sqrt{c}}\left(S_{\blacksquare(k)}^{\mu}\right)^a\right]\right\rangle_{\epsilon,\varepsilon}\right) \tag{B8}$$

and

$$-\beta G_{n,0}^{\text{bond}}[\hat{\epsilon}] = \sum_{\blacksquare}\ln\left(\prod_a \text{Tr}_{\mathbf{J}^a}\right)e^{-c\sum_{a<b}\epsilon_{ab,\blacksquare}J^aJ^b}, \quad -\beta G_{n,0}^{\text{spin}}[\hat{\varepsilon}] = \sum_{\blacksquare\backslash\text{output}}\ln\left(\prod_a \text{Tr}_{\mathbf{S}^a}\right)e^{-M\sum_{a<b}\varepsilon_{ab,\blacksquare}S^aS^b}. \tag{B9}$$

We also introduced

$$\langle\cdots\rangle_{\epsilon,\varepsilon} = \frac{\prod_{\blacksquare\backslash\text{output}}\text{Tr}_{\mathbf{S}_{\blacksquare}^a}e^{-\sum_\mu\sum_{a<b}\varepsilon_{ab}(S_{\blacksquare}^{\mu})^a(S_{\blacksquare}^{\mu})^b}\prod_{\blacksquare}\text{Tr}_{\mathbf{J}a}e^{-\sum_k\sum_{a<b}\epsilon_{ab}(J_{\blacksquare}^k)^a(J_{\blacksquare}^k)^b}\ldots}{\prod_{\blacksquare\backslash\text{output}}\text{Tr}_{\mathbf{S}_{\blacksquare}^a}e^{-\sum_\mu\sum_{a<b}\varepsilon_{ab}(S_{\blacksquare}^{\mu})^a(S_{\blacksquare}^{\mu})^b}\prod_{\blacksquare}\text{Tr}_{\mathbf{J}a}e^{-\sum_k\sum_{a<b}\epsilon_{ab}(J_{\blacksquare}^k)^a(J_{\blacksquare}^k)^b}}, \tag{B10}$$

which represents an averaging using a noninteracting system with polarizing field ϵ_{ab} and ε_{ab} conjugated to the order parameters Q_{ab} and q_{ab} [41].

Note that Eq. (B7) defines $-\beta\tilde{F}_n[\hat{Q},\hat{q}]$ by a Legendre transformation of $-\beta\tilde{G}_n[\hat{\epsilon},\hat{\varepsilon}]$ defined by Eq. (B8). The integrations over ϵ and ε can be done by the saddle-point method for $c \gg 1$ and $M \gg 1$ yielding

$$-\beta\tilde{F}_n[\hat{Q},\hat{q}] = -\beta\tilde{G}_n[\hat{\epsilon}^*,\hat{\varepsilon}^*] + c\sum_{a<b,\blacksquare}\epsilon_{ab,\blacksquare}^*Q_{ab,\blacksquare} \\ +M\sum_{a<b,\blacksquare}\varepsilon_{ab,\blacksquare}^*q_{ab,\blacksquare}, \tag{B11}$$

where the saddle points $\epsilon^* = \epsilon^*[\hat{Q}]$ and $\varepsilon^* = \varepsilon^*[\hat{q}]$ satisfy

$$Q_{ab,\blacksquare} = -\frac{1}{c}\frac{\partial}{\partial\epsilon_{ab,\blacksquare}}(-\beta\tilde{G}_n[\hat{\epsilon},\hat{\varepsilon}])\bigg|_{\epsilon=\epsilon^*,\varepsilon=\varepsilon^*} \\ = \frac{1}{c}\sum_{k=1}^{c}\langle(J_{\blacksquare}^k)^a(J_{\blacksquare}^k)^b\rangle_{\epsilon^*,\varepsilon^*}, \\ ]q_{ab,\blacksquare} = -\frac{1}{M}\frac{\partial}{\partial\varepsilon_{ab,\blacksquare}}(-\beta\tilde{G}_n[\hat{\epsilon},\hat{\varepsilon}])\bigg|_{\epsilon=\epsilon^*,\varepsilon=\varepsilon^*} \\ = \frac{1}{M}\sum_{\mu=1}^{M}\langle(S_{\blacksquare}^{\mu})^a(S_{\blacksquare}^{\mu})^b\rangle_{\epsilon^*,\varepsilon^*}. \tag{B12}$$

The latter implies

$$\langle(J^k)^a(J^k)^b\rangle_\epsilon = Q_{ab} \quad \forall k, \quad \langle(S^\mu)^a(S^\mu)^b\rangle_\varepsilon = q_{ab} \quad \forall\mu \tag{B13}$$

since different components μ's and k's are equivalent and independent in the averaging Eq. (B10).

Note that $-\beta\tilde{G}_n[\hat{\epsilon},\hat{\varepsilon}]$ [Eq. (B8)] consists of a noninteracting part (entropic term) $-\beta G_{n,0}^{\text{bond}}[\hat{\epsilon}]$ and $-\beta G_{n,0}^{\text{spin}}[\hat{\varepsilon}]$ defined in Eq. (B9) and a contribution of interactions that involves an evaluation using the noninteracting system Eq. (B10). Certainly, the latter is the crucial one. Our strategy is to analyze the effect of interactions using a combination of the Plefka expansion (Appendix B 2) and the cumulant expansion (Appendix B 4).

2. Plefka expansion

Suppose that the effect of the interactions can be treated perturbatively, which enables the following decompositions [42]:

$$\tilde{F}_n = F_{n,0} + \lambda\tilde{F}_{n,1} + \frac{\lambda^2}{2}\tilde{F}_{n,2} + \cdots, \\ \tilde{G}_n = G_{n,0}^{\text{bond}} + G_{n,0}^{\text{spin}} + \lambda\tilde{G}_{n,1} + \frac{\lambda^2}{2}\tilde{G}_{n,2} + \cdots, \\ \epsilon_{ab} = (\epsilon_0)_{ab} + \lambda(\epsilon_1)_{ab} + \frac{\lambda^2}{2}(\epsilon_2)_{ab}\ldots, \\ \varepsilon_{ab} = (\varepsilon_0)_{ab} + \lambda(\varepsilon_1)_{ab} + \frac{\lambda^2}{2}(\varepsilon_2)_{ab}\ldots, \tag{B14}$$

where we introduced a parameter λ to keep track of the expansion. Here the quantities with the suffix 0 represent those that are present in the absence of interactions, and those with suffixes 1, 2, ... represent those due to interactions.

The Legendre transform Eq. (B11) becomes, at $O(\lambda^0)$,

$$-\beta F_{n,0}[\hat{Q},\hat{q}] = -\beta G_{n,0}^{\text{bond}}[\hat{\epsilon}_0^*] - \beta G_{n,0}^{\text{spin}}[\hat{\varepsilon}_0^*] \\ +c\sum_{a<b,\blacksquare}(\epsilon_0^*)_{ab,\blacksquare}Q_{ab,\blacksquare} \\ +M\sum_{a<b,\blacksquare}(\varepsilon_0^*)_{ab,\blacksquare}q_{ab,\blacksquare}, \tag{B15}$$

where $(\epsilon_0^*)_{ab}$ and $(\varepsilon_0^*)_{ab}$ are defined such that

$$Q_{ab} = -\frac{1}{c}\frac{\partial}{\partial\epsilon_{ab}}\left(-\beta G_{n,0}^{\text{bond}}[\hat{\epsilon}]\right)\bigg|_{\hat{\epsilon}=\hat{\epsilon}_0^*[\hat{Q}]}, \\ q_{ab} = -\frac{1}{M}\frac{\partial}{\partial\varepsilon_{ab}}\left(-\beta G_{n,0}^{\text{spin}}[\hat{\varepsilon}]\right)\bigg|_{\hat{\varepsilon}=\hat{\varepsilon}_0^*[\hat{q}]}. \tag{B16}$$

Then at $O(\lambda)$ we find

$$\begin{aligned}-\beta\tilde{F}_{n,1}[\hat{Q},\hat{q}] &= -\beta\tilde{G}_{n,1}[\hat{\epsilon}^*_0[\hat{Q}],\hat{\varepsilon}^*_0[\hat{q}]] + \sum_{a<b,\blacksquare} \left.\frac{\partial\left(-\beta G^{\rm bond}_{n,0}[\hat{\epsilon}]\right)}{\partial\epsilon_{ab,\blacksquare}}\right|_{\hat{\epsilon}=\hat{\epsilon}^*_0[\hat{Q}]} (\epsilon^*_1)_{ab,\blacksquare} + c\sum_{a<b,bs}(\epsilon^*_1)_{ab,\blacksquare}Q_{ab,\blacksquare} \\ &\quad + \sum_{a<b,\blacksquare}\left.\frac{\partial\left(-\beta G^{\rm spin}_{n,0}[\hat{\varepsilon}]\right)}{\partial\varepsilon_{ab,\blacksquare}}\right|_{\hat{\varepsilon}=\hat{\varepsilon}^*_0[\hat{q}]}(\varepsilon^*_1)_{ab,\blacksquare} + M\sum_{a<b}(\varepsilon^*_1)_{ab,\blacksquare}q_{ab,\blacksquare} = -\beta\tilde{G}_{n,1}[\hat{\epsilon}^*_0[\hat{Q}],\hat{\varepsilon}^*_0[\hat{q}]]. \qquad \text{(B17)}\end{aligned}$$

In the second equation, we used Eq. (B16).

Similarly, at $O(\lambda^2)$ we find

$$\begin{aligned}-\beta\tilde{F}_{n,2}[\hat{Q},\hat{q}] &= -\beta\tilde{G}_{n,2}[\epsilon^*_0,\varepsilon^*_0] + 2\sum_{a<b,\blacksquare}\left.\frac{\partial(-\beta\tilde{G}_{n,1}[\epsilon,\varepsilon])}{\partial\epsilon_{ab,\blacksquare}}\right|_{\epsilon=\epsilon^*_0,\varepsilon=\varepsilon^*_0}(\epsilon^*_1)_{ab,\blacksquare} + 2\sum_{a<b,\blacksquare}\left.\frac{\partial(-\beta\tilde{G}_{n,1}[\epsilon,\varepsilon])}{\partial\varepsilon_{ab,\blacksquare}}\right|_{\epsilon=\epsilon^*_0,\varepsilon=\varepsilon^*}(\varepsilon^*_1)_{ab,\blacksquare} \\ &\quad + \sum_{a<b,\blacksquare}\left.\frac{\partial\left(-\beta G^{\rm bond}_{n,0}[\epsilon]\right)}{\partial\epsilon_{ab,\blacksquare}}\right|_{\epsilon=\epsilon^*_0}(\epsilon^*_2)_{ab,\blacksquare} + \sum_{a<b,\blacksquare}\left.\frac{\partial\left(-\beta G^{\rm spin}_{n,0}[\varepsilon]\right)}{\partial\varepsilon_{ab,\blacksquare}}\right|_{\varepsilon=\varepsilon^*_0}(\varepsilon^*_2)_{ab,\blacksquare} \\ &\quad + \sum_{\blacksquare}\sum_{a<b,\blacksquare}\sum_{c<d}\left.\frac{\partial^2\left(-\beta G^{\rm bond}_{n,0}\right)[\epsilon,\varepsilon]}{\partial\epsilon_{ab,\blacksquare}\epsilon_{cd,\blacksquare}}\right|_{\epsilon=\epsilon^*_0}(\epsilon^*_1)_{ab,\blacksquare}(\epsilon^*_1)_{cd,\blacksquare} \\ &\quad + \sum_{\blacksquare}\sum_{a<b}\sum_{c<d}\left.\frac{\partial^2\left(-\beta G^{\rm spin}_{n,0}\right)[\epsilon,\varepsilon]}{\partial\varepsilon_{ab,\blacksquare}\varepsilon_{cd,\blacksquare}}\right|_{\varepsilon=\varepsilon^*_0}(\varepsilon^*_1)_{ab,\blacksquare}(\varepsilon^*_1)_{cd,\blacksquare} + c\sum_{a<b,\blacksquare}(\epsilon^*_2)_{ab,\blacksquare}Q_{ab,\blacksquare} + M\sum_{a<b,\blacksquare}(\varepsilon^*_2)_{ab,\blacksquare}q_{ab,\blacksquare} \\ &= -\beta\tilde{G}_{n,2}[\epsilon^*_0,\varepsilon^*_0] - \sum_{\blacksquare}\sum_{a<b}\sum_{c<d}\frac{\partial(-\beta\tilde{G}_{n,1})[\hat{\epsilon},\hat{\varepsilon}]}{\partial\epsilon_{ab,\blacksquare}}\left(\frac{\partial^2\left(-\beta G^{\rm bond}_{n,0}[\hat{\epsilon}]\right)}{\partial\epsilon_{ab,\blacksquare}\partial\epsilon_{cd,\blacksquare}}\right)^{-1}\frac{\partial(-\beta\tilde{G}_{n,1})[\hat{\epsilon},\hat{\varepsilon}]}{\partial\epsilon_{cd,\blacksquare}} \\ &\quad - \sum_{\blacksquare}\sum_{a<b}\sum_{c<d}\frac{\partial(-\beta\tilde{G}_{n,1})[\hat{\varepsilon},\hat{\varepsilon}]}{\partial\varepsilon_{ab,\blacksquare}}\left(\frac{\partial^2\left(-\beta G^{\rm spin}_{n,0}[\hat{\varepsilon}]\right)}{\partial\varepsilon_{ab,\blacksquare}\partial\varepsilon_{cd,\blacksquare}}\right)^{-1}\frac{\partial(-\beta\tilde{G}_{n,1})[\hat{\varepsilon},\hat{\varepsilon}]}{\partial\varepsilon_{cd,\blacksquare}}. \qquad \text{(B18)}\end{aligned}$$

To derive the last line, we used Eq. (B16) and

$$\begin{aligned}0 &= \left.\frac{\partial(-\beta\tilde{G}_{n,1}[\hat{\epsilon},\hat{\varepsilon}])}{\partial\epsilon_{ab,\blacksquare}}\right|_{\epsilon=\epsilon^*_0,\varepsilon=\varepsilon^*_0} + \sum_{c<d}\left.\frac{\partial^2\left(-\beta G^{\rm bond}_{n,0}[\hat{\epsilon}]\right)}{\partial\epsilon_{ab,\blacksquare}\partial\epsilon_{cd,\blacksquare}}\right|_{\epsilon=\epsilon^*_0}(\epsilon^*_1)_{cd,\blacksquare}, \\ 0 &= \left.\frac{\partial\left(-\beta\tilde{G}_{n,1}[\hat{\epsilon},\hat{\varepsilon}]\right)}{\partial\epsilon_{ab,\blacksquare}}\right|_{\epsilon=\epsilon^*_0,\varepsilon=\varepsilon^*_0} + \sum_{c<d}\left.\frac{\partial^2\left(-\beta G^{\rm spin}_{n,0}[\hat{\varepsilon}]\right)}{\partial\varepsilon_{ab,\blacksquare}\partial\varepsilon_{cd,\blacksquare}}\right|_{\varepsilon=\varepsilon^*_0}(\varepsilon^*_1)_{cd,\blacksquare}, \qquad \text{(B19)}\end{aligned}$$

which is obtained by expanding Eq. (B12) up to $O(\lambda)$ and then using Eq. (B16) for the zeroth-order terms.

If $O(\lambda)^2$ terms and higher-order terms vanish (as happens in the dense coupling), we can set $\lambda = 1$ and obtain

$$\begin{aligned}\tilde{F}_n[\hat{Q},\hat{q}] &= -\beta F_{n,0}[\hat{Q},\hat{q}] - \beta\tilde{F}_{n,1}[\hat{Q},\hat{q}] \\ &= -\beta G_{n,0}[\hat{\epsilon}^*,\hat{\varepsilon}^*] + c\sum_{a<b,\blacksquare}\epsilon^*_{ab,\blacksquare}Q_{ab,\blacksquare} + M\sum_{a<b,\blacksquare}\varepsilon^*_{ab,\blacksquare}q_{ab,\blacksquare} - \beta\tilde{G}_{n,1}[\hat{\epsilon}^*], \qquad \text{(B20)}\end{aligned}$$

where $\hat{\epsilon}^* = \hat{\epsilon}^*_0[\hat{q}]$ and $\hat{\varepsilon}^* = \hat{\varepsilon}^*_0[\hat{Q}]$ are those determined by Eq. (B16).

3. Summary 1

Here we can wrap up the above results to find the replicated Gardner volume [Eq. (B2)] expressed as

$$\begin{aligned}V^n(\mathbf{S}_0,\mathbf{S}_L) &= e^{NM\mathcal{S}_n(\mathbf{S}_0,\mathbf{S}_l)} \\ &= \prod_{a<b,\blacksquare}\left\{\int_{-\infty}^{\infty}dQ_{ab,\blacksquare}\right\}\prod_{a<b,\blacksquare\backslash\text{output}}\left\{\int_{-\infty}^{\infty}dq_{ab,\blacksquare}\right\}e^{-\beta F_n[\hat{Q},\hat{q}]}. \qquad \text{(B21)}\end{aligned}$$

The functional $-\beta F_n[\hat{Q},\hat{q}]$ may be regarded as replicated free-energy functional

$$-\beta F_n[\hat{Q},\hat{q}] = -\beta F_0[\hat{Q},\hat{q}] - \beta F_{\rm ex}[\hat{Q},\hat{q}], \qquad \text{(B22)}$$

where $-\beta F_0[\hat{Q},\hat{q}]$ given by Eq. (B15) may be regarded as the entropic part of the free-energy while $-\beta F_{\rm ex}$ is the interaction part of the free-energy,

$$e^{-\beta F_{\rm ex}[\hat{Q},\hat{q}]} = \prod_{\mu,\blacksquare,a}\left\{\int \frac{d\eta_{\mu,\blacksquare,a}}{\sqrt{2\pi}} W_{\eta_{\mu,\blacksquare,a}}\right\} e^{-\beta\tilde{F}_{\rm ex}[\hat{Q},\hat{q};\{i\eta_{\mu,\blacksquare,a}\}]} = e^{-\beta\tilde{F}_{\rm ex}[\hat{Q},\hat{q},\{\partial/\partial_{\mu,\blacksquare,a}\}]} \prod_{\mu,\blacksquare,a} e^{-\beta v(h_{\mu,\blacksquare,a})}\Bigg|_{\{h_{\mu,bs,a}\}=0} \tag{B23}$$

with

$$\tilde{F}_{\rm ex} = \tilde{F}_{n,1} + \tilde{F}_{n,2} + \cdots . \tag{B24}$$

In the first equation of Eq. (B23), we recalled that $\tilde{F}_{\rm ex}[\hat{Q},\hat{q};\{i\eta_{\mu,\blacksquare,a}\}]$ depends on $\{i\eta_{\mu,\blacksquare,a}\}$. In the second equation of Eq. (B23), $\tilde{F}_{\rm ex}[\hat{Q},\hat{q};\{i\eta_{\mu,\blacksquare,a}\}]$ is a differential operator.

4. Cumulant expansion

Now we turn to the explicit evaluation of the $-\beta\tilde{G}_n[\hat{\epsilon},\hat{\varepsilon}]$ defined in Eq. (B8) by a cumulant expansion, introducing the parameter λ,

$$\begin{aligned}
-\beta\tilde{G}_n[\hat{\epsilon},\hat{\varepsilon}] &= \ln\left\langle \exp\left[i\sum_{\mu,\blacksquare,a}\eta_{\mu,\blacksquare,a}(S^{\mu}_{\blacksquare})^a \sum_{k=1}^{c}\frac{\sqrt{\lambda}}{\sqrt{c}}(J^k_{\blacksquare})^a(S^{\mu}_{\blacksquare(k)})^a\right]\right\rangle_{\epsilon,\varepsilon}\\
&= \ln\Bigg\langle 1 + \sum_{\mu,\blacksquare,a} i\eta_{\mu,\blacksquare,a}(S^{\mu}_{\blacksquare})^a \sum_{k=1}^{c}\frac{\sqrt{\lambda}}{\sqrt{c}}(J^k_{\blacksquare})^a(S^{\mu}_{\blacksquare(k)})^a + \frac{1}{2!}\sum_{\mu,\blacksquare,a}\sum_{\nu,\square,b} i\eta_{\mu,\blacksquare,a} i\eta_{\nu,\square,b}(S^{\mu}_{\blacksquare})^a(S^{\nu}_{\square})^b\\
&\quad\times\sum_{k=1}^{c}\frac{\sqrt{\lambda}}{\sqrt{c}}(J^k_{\blacksquare})^a(S^{\mu}_{\blacksquare(k)})^a\sum_{k'=1}^{c}\frac{\sqrt{\lambda}}{\sqrt{c}}(J^{k'}_{\square})^b(S^{\nu}_{\square(k')})^b + \cdots\Bigg\rangle_{\epsilon,\varepsilon}.
\end{aligned} \tag{B25}$$

From Eq. (B10) we find that averages $\langle\cdots\rangle_{\epsilon,\varepsilon}$ of terms with odd numbers of spins $(S^{\mu}_{\blacksquare})^a$ and bonds $(J^k_{\blacksquare})^a$ vanish by symmetry. Consequently, we find nonvanishing terms at order $O(\lambda)$, $O(\lambda^2)$,...corresponding to the second- and fourth-order terms of the cumulant expansion which are represented by connected diagrams. They define $-\beta\tilde{G}_{n,1}, -\beta\tilde{G}_{n,2}\ldots$ in the Plefka expansion [Eq. (B14)] of $-\beta\tilde{G}_n$.

a. O(λ) term

We find that the second-order cumulant yields $O(\lambda)$, i.e., $-\beta\hat{G}_{n,1}$. Then by Eq. (B17) we find that this is also $-\beta\tilde{F}_{n,1}$,

$$\begin{aligned}
-\beta\tilde{F}_{n,1}[\hat{Q},\hat{q}] &= -\beta\tilde{G}_{n,1}[\hat{\epsilon}^*[\hat{Q}],\hat{\varepsilon}^*[\hat{q}]]\\
&= \left\langle \frac{1}{2!}\sum_{\mu,\blacksquare,a}\sum_{\nu,\square,b} i\eta_{\mu,\blacksquare,a} i\eta_{\nu,\square,b}(S^{\mu}_{\blacksquare})^a(S^{\nu}_{\square})^b \sum_{k=1}^{c}\frac{\sqrt{\lambda}}{\sqrt{c}}(J^k_{\blacksquare})^a(S^{\mu}_{\blacksquare(k)})^a\sum_{k'=1}^{c}\frac{\sqrt{\lambda}}{\sqrt{c}}(J^{k'}_{\square})^b(S^{\nu}_{\square(k')})^b\right\rangle_{\epsilon^*[\hat{Q}],\varepsilon^*[\hat{q}]}\\
&= \frac{\lambda}{2}\sum_{\mu,\blacksquare}\sum_{a,b} i\eta_{\mu,\blacksquare,a} i\eta_{\mu,\blacksquare,b} q_{ab,\blacksquare}Q_{ab,\blacksquare}\frac{1}{c}\sum_{k=1}^{c} q_{ab,\blacksquare(k)},
\end{aligned} \tag{B26}$$

where we have used Eq. (B13). Anticipating the homogeneous solution with each layer [Eq. (19)], we find $-\beta\hat{F}_{n,1}/(NM) \sim O(1)$. This term will become the dominant term that contributes to the interaction part of the free-energy $-\beta F_{\rm ex}$ in the dense limit, $N \gg c \gg 1$.

In Fig. 17 we show a graphical representation of the term. $\tilde{G}_1$ (and $\tilde{F}_1$) is obtained by associating two replicas to the diagram.

b. O(λ²) terms

At the fourth order of the cumulant expansion, we easily find a $O(\lambda^2)$ term that contributes to $-\beta G_{n,2}$ and thus $-\beta F_{n,2}$ via Eq. (B18) by associating four replicas to the same diagram shown in Fig. 17,

$$\begin{aligned}
-\beta\tilde{F}_{n,2}(\text{Fig. }17) = -\beta\tilde{G}_{n,2}(\text{Fig. }17) &= \frac{1}{c}\frac{\lambda^2}{4!}\sum_{\mu,\blacksquare}\sum_{a,b,c,d} i\eta_{\mu,\blacksquare,a} i\eta_{\mu,\blacksquare,b} i\eta_{\mu,\blacksquare,c} i\eta_{\mu,\blacksquare,d}\frac{1}{c}\sum_{k=1}^{c}[\langle(S^{\mu}_{\blacksquare})^a(S^{\mu}_{\blacksquare})^b(S^{\mu}_{\blacksquare})^c(S^{\mu}_{\blacksquare})^d(J^k_{\blacksquare})^a\\
&\quad\times(J^k_{\blacksquare})^b(J^k_{\blacksquare})^c(J^k_{\blacksquare})^d(S^{\mu}_{\blacksquare(k)})^a(S^{\mu}_{\blacksquare(k)})^b(S^{\mu}_{\blacksquare(k)})^c(S^{\mu}_{\blacksquare(k)})^d\rangle_{\hat{\epsilon},\hat{\varepsilon}} - q_{ab,\blacksquare}Q_{ab,\blacksquare}q_{ab,\blacksquare(k)}q_{cd,\blacksquare}Q_{cd,\blacksquare}q_{cd,\blacksquare(k)}\\
&\quad - q_{ac,\blacksquare}Q_{ac,\blacksquare}q_{ac,\blacksquare(k)}q_{bd,\blacksquare}Q_{bd,\blacksquare}q_{bd,\blacksquare(k)} - q_{ad,\blacksquare}Q_{ad,\blacksquare}q_{ad,\blacksquare(k)}q_{bc,\blacksquare}Q_{bc,\blacksquare}q_{bc,\blacksquare(k)}].
\end{aligned} \tag{B27}$$

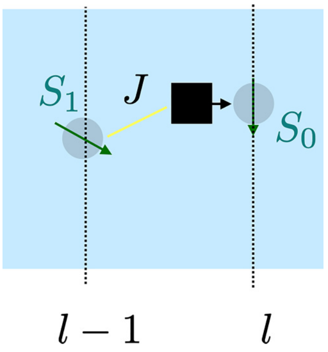

FIG. 17. Graphical representation of a contribution to $\tilde{G}_1$ (and $\tilde{F}_1$).

We see that $-\beta\tilde{F}_{n,2}$(Fig. 17)$/(NM) \propto 1/c$ and it vanishes in the $c \to \infty$ limit.

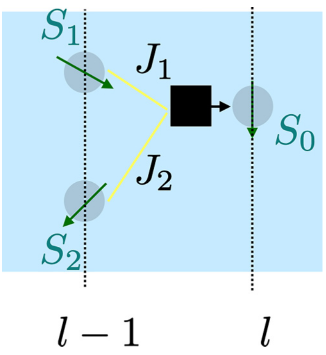

FIG. 18. A contribution to G_2 which is one-line reducible.

There is another contribution to $\tilde{G}_2$ which is associated with a diagram shown in Fig. 18. We associate two replicas a, b with branch "1" and replicas c, d with branch "2,"

$$\begin{aligned}
-\beta\tilde{G}_{n,2}(\text{Fig. 18}) \sim & \sum_{a<b}\sum_{c<d}[\langle (S_0)_a(J_1)_a(S_1)_a(S_0)_b(J_1)_b(S_1)_b(S_0)_c(J_2)_c(S_2)_c(S_0)_d(J_2)_d(S_2)_d\rangle_{\epsilon,\varepsilon} \\
& - \langle (S_0)_a(J_1)_a(S_1)_a(S_0)_b(J_1)_b(S_1)_b\rangle_{\epsilon,\varepsilon}\langle (S_0)_c(J_2)_c(S_2)_c(S_0)_d(J_2)_d(S_2)_d\rangle_{\epsilon,\varepsilon}] \\
= & \sum_{a<b}\sum_{c<d}\langle (S_0)_a(S_0)_b(S_0)_c(S_0)_d\rangle^{c}_{\epsilon,\varepsilon} Q_{ab,\blacksquare}q_{ab,\blacksquare}Q_{cd,\blacksquare}q_{cd,\blacksquare},
\end{aligned} \tag{B28}$$

where $\langle S^aS^bS^cS^d\rangle^c$'s are connected correlation functions defined as

$$\begin{aligned}
\langle S^aS^bS^cS^d\rangle^c &= \langle S^aS^bS^cS^d\rangle - \langle S^aS^b\rangle\langle S^cS^d\rangle \\
&= \langle S^aS^bS^cS^d\rangle - q_{ab}q_{cd}.
\end{aligned} \tag{B29}$$

Note that it involves four perceptrons associated with the four replicas so that we have a factor $(1/\sqrt{c})^4$ but there are $c(c-1)$ different ways to choose the end points of branch "1" and "2." Thus the contribution by this type of term survives in the $c \to \infty$ limit as the $O(1)$ contribution to $-\beta G_{n,2}/(NM)$.

However, this does not contribute to $-\beta F_{n,2}$ because it is exactly canceled by the second term in Eq. (B18). To see this, let us recall that $G_{n,1}$ is like

$$\tilde{G}_{n,1} \sim \sum_{a<b}\langle (S_0)_a(J_1)_a(S_1)_a(S_0)_a(J_1)_b(S_1)_b\rangle. \tag{B30}$$

Then we find

$$\begin{aligned}
& -\sum_{a<b}\sum_{c<d}\frac{\partial\tilde{G}_{n,1}[\hat{\varepsilon},\hat{\varepsilon}]}{\partial\varepsilon_{ab,\blacksquare}}\left(\frac{\partial^2\big(-\beta G^{\text{spin}}_{n,0}[\hat{\varepsilon}]\big)}{\partial\varepsilon_{ab,\blacksquare}\partial\varepsilon_{cd,\blacksquare}}\right)^{-1}\frac{\partial\tilde{G}_{n,1}[\hat{\varepsilon},\hat{\varepsilon}]}{\partial\varepsilon_{cd,\blacksquare}} \\
& \sim -\sum_{a<b}\sum_{c<d}\frac{\partial}{\partial\varepsilon_{ab,\blacksquare}}\left(\sum_{e<f}\langle (S_0)_e(J_1)_e(S_1)_e(S_0)_f(J_1)_f)(S_1)_f\rangle\right)\left(\frac{\partial^2\big(-\beta G^{\text{spin}}_{n,0}[\hat{\varepsilon}]\big)}{\partial\varepsilon_{ab,\blacksquare}\partial\varepsilon_{cd,\blacksquare}}\right)^{-1} \\
& \quad\times\frac{\partial}{\partial\varepsilon_{cd,\blacksquare}}\left(\sum_{g<h}\langle (S_0)_g(J_1)_g(S_1)_g(S_0)_h(J_1)_h)(S_1)_h\rangle\right) \\
& = -\sum_{e<f}\sum_{g<h}\langle (J_1)_e(S_1)_e(J_1)_f)(S_1)_f\rangle\langle (J_1)_g(S_1)_g(J_1)_h)(S_1)_h\rangle\sum_{a<b}\sum_{c<d}\frac{\partial q_{ef,\blacksquare}}{\partial\varepsilon_{ab,\blacksquare}}\left(\frac{\partial^2\big(-\beta G^{\text{spin}}_{n,0}[\hat{\varepsilon}]\big)}{\partial\varepsilon_{ab,\blacksquare}\partial\varepsilon_{cd,\blacksquare}}\right)^{-1}\sum_{g<h}\frac{\partial q_{gh,\blacksquare}}{\partial\varepsilon_{cd,\blacksquare}} \\
& = -\sum_{e<f}\sum_{g<h}Q_{ef,\blacksquare}q_{ef,\blacksquare}Q_{gh,\blacksquare}q_{gh,\blacksquare}\left(\frac{\partial^2\big(-\beta G^{\text{spin}}_{n,0}[\hat{\varepsilon}]\big)}{\partial\varepsilon_{ef,\blacksquare}\partial\varepsilon_{gh,\blacksquare}}\right) \\
& = -\sum_{e<f}\sum_{g<h}Q_{ef,\blacksquare}q_{ef,\blacksquare}Q_{gh,\blacksquare}q_{gh,\blacksquare}\langle (S_0)_e(S_0)_f(S_0)_g(S_0)_h\rangle^{c}_{\epsilon,\varepsilon}.
\end{aligned} \tag{B31}$$

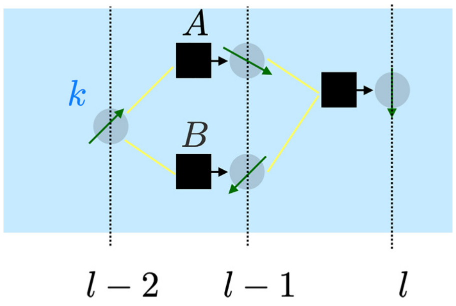

FIG. 19. A loop of interactions in a DNN extended over three layers, through three perceptrons and four bonds.

This exactly cancels $-\beta\tilde{G}_{n,2}$(Fig. 18). Thus the diagram shown in Fig. 18 does not contribute $-\beta\tilde{F}_{n,2}$.

Indeed, it is known in diagrammatic expansions that “one-line (or particle) reducible” diagrams like the one shown in Fig. 18 become canceled after Legendre transform from $-\beta\tilde{G}$ to $-\beta\tilde{F}$ [43,44] leaving only *loop diagrams*, which are *one-line irreducible*, i.e., diagrams that cannot be separated into two disconnected diagrams by cutting a line. At $O(\lambda^2)$ we do not have such a loop diagram.

To sum up, we find

$$-\beta\tilde{F}_{n,2}/(NM) = -\beta\tilde{F}_{n,2}(\text{Fig. 17})/(NM) \propto 1/c, \tag{B32}$$

which vanishes in the dense limit $c \to \infty$.

c. $O(\lambda^3)$ terms

At $O(\lambda^3)$ we will have a term that is obtained by associating six replicas to the diagram Fig. 17 whose contribution to $-\beta F_{n,3}/(NM)$ vanishes as $1/c^2$ in the $c \to \infty$ limit.

Apart from that, we find contributions of one-loop diagrams. As the simplest example, consider the loop shown in Fig. 19 (same one as shown in Fig. 3). Such a loop contributes to the form

$$\begin{aligned}-\beta\tilde{F}_{n,3}(\text{Fig. 19}) = {}& \frac{\lambda^3}{6!}\left(\frac{1}{\sqrt{c}}\right)^6 \sum_{\blacksquare,\mu}\sum_{a,b,c,d,e,f}\left(\sum_{\blacksquare_A,\blacksquare_B,k}\right)_{\text{loop}} i\eta_{\mu,\blacksquare,a}\, i\eta_{\mu,\blacksquare,b}\, i\eta_{\mu,\blacksquare_A,c}\, i\eta_{\mu,\blacksquare_A,d}\, i\eta_{\mu,\blacksquare_B,e}\, i\eta_{\mu,\blacksquare_B,f}\\ &\times Q_{ab,\blacksquare}Q_{cd,\blacksquare_A}Q_{ef,\blacksquare_B}\big[\langle S^a_{\blacksquare}S^b_{\blacksquare}S^c_{\blacksquare}S^d_{\blacksquare}\rangle_{\epsilon,\varepsilon}\big\langle S^a_{\blacksquare_A}S^b_{\blacksquare_A}S^c_{\blacksquare_A}S^d_{\blacksquare_A}\big\rangle_{\epsilon,\varepsilon}\big\langle S^a_{\blacksquare_B}S^b_{\blacksquare_B}S^e_{\blacksquare_B}S^f_{\blacksquare_B}\big\rangle_{\epsilon,\varepsilon}\big\langle S^c_kS^d_kS^e_kS^f_k\big\rangle_{\epsilon,\varepsilon}\\ &- q_{ab,\blacksquare}q_{cd,\blacksquare}q_{ab,\blacksquare_A}q_{cd,\blacksquare_A}q_{ab,\blacksquare_B}q_{ef,\blacksquare_B}q_{cd,k}q_{ef,k}\big].\end{aligned} \tag{B33}$$

Here the factor $(1/\sqrt{c})^6$ appears because six perceptrons (two replicas for each of the three perceptrons $\blacksquare$, $\blacksquare_A$, $\blacksquare_B$) are involved. The expression $(\sum_{\blacksquare_A,\blacksquare_B,k})_{\text{loop}}$ means to sum over $\blacksquare_A$, $\blacksquare_B$, and k conditioned that the loop $\blacksquare \to \blacksquare_A \to \blacksquare_B \to \blacksquare$ is closed.

Let us consider how many such loops exist for a given perceptron $\blacksquare$. Starting from 0, there are c choices for $\blacksquare_A$ connected to $\blacksquare$ and $c-1$ choices for $\blacksquare_B$ (different from $\blacksquare_A$) connected to $\blacksquare$. Similarly, there are c choices for k connected to $\blacksquare_A$. Finally, the probability (in a given realization of the random network) that k happens to be connected to $\blacksquare_B$ is $\sim c/N$. Thus

$$\left(\sum_{\blacksquare_A,\blacksquare_B,k}\right)_{\text{loop}} \sim c^2(c-1)\frac{c}{N}. \tag{B34}$$

Thus the net contribution of the one-loop terms scales as

$$\frac{-\beta\tilde{F}_{n,3}(\text{Fig. 19})}{NM} \propto \frac{c}{N}. \tag{B35}$$

Thus the contribution vanishes in the dense limit because the $N \to \infty$ limit is taken before the $c \to \infty$ limit. However, in the case of global coupling $c = N$ the contribution cannot be neglected.

d. Higher-order terms

Similarly to the $O(\lambda^3)$ terms, higher-order terms of $-\beta\tilde{F}_{\text{ex}}/(NM)$ can be classified into two cases.

(i) At $O(\lambda^p)$ ($p \geqslant 3$) we will have a term that is obtained by associating $2p$ replicas to the diagram Fig. 17 whose contribution to $-\beta F_{n,p}/(NM)$ vanishes as $1/c^{p-1}$ in the $c \to \infty$ limit.

(ii) All other terms are associated with loop diagrams. Similarly to the loop diagram considered at $O(\lambda^3)$, we can consider more extended one-loops such as the one shown in Fig. 20, which involves $2p$ perceptrons (two replicas for each of p perceptrons) extended over $(p-1)/2+2$ layers. It is easy to see that all such one-loops make $O(c/N)$ contributions to the higher-order terms of $-\beta\tilde{F}_{n,p}/(NM)$ for $p \geqslant 3$. It is interesting to note that the order of the correction term is order $O(c/N)$, which is independent of the size p of the loop.

Onto the same one-loop diagram, we can associate four replicas: two replicas along one path from the right to left and the other two replicas along the other path. This yields a contribution to $-\beta\tilde{F}_{n,2p}/(NM)$ of order $O(c^{-p}(c/N))$.

(iii) Contributions of two-loops, three-loops, etc., can be considered similarly. First one can see that the probability to close two-loops, three loops, etc., scales as

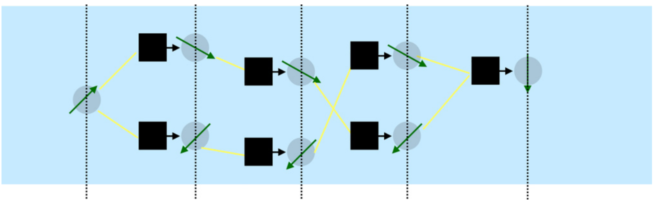

FIG. 20. More extended loop.

$O(c/N)^2, O(c/N)^3, \ldots$. By associating two replicas to such diagrams, we find contributions to $-\beta\tilde{F}_{n,p}/(NM)$ of order $O(c/N)^2, O(c/N)^3, \ldots$.

(iv) Note that loop corrections break the symmetry with respect to the exchange of input/output sides.

(v) In general, by associating more replicas to the same diagram, we find contributions that vanish more rapidly, increasing c.

5. Summary 2

Now we can collect the above results to obtain the free-energy functional $-\beta F_n[\hat{Q}, \hat{q}]$ defined in Appendix B 3,

$$\frac{-\beta F_n[\{\hat{Q}, \hat{q}\}]}{M} = \frac{1}{\alpha}\sum_{l=1,2,\ldots,L}\sum_{\blacksquare\in l} s_{\rm ent,bond}[\hat{Q}_\blacksquare] + \sum_{l=1,2,\ldots,L-1}\sum_{\blacksquare\in l} s_{\rm ent,spin}[\hat{q}_\blacksquare] + \frac{-\beta F_{\rm ex}[\{\hat{Q}_\blacksquare, \hat{q}_\blacksquare\}]}{M}, \tag{B36}$$

with the first two terms being the entropic part of the free energy due to bonds and spins [see Eqs. (B9) and (B15)],

$$c s_{\rm ent,bond}[\hat{Q}] = -\beta G^{\rm bond}_{n,0}[\hat{\epsilon}_0^*] + c\sum_{a<b}(\epsilon_0^*)_{ab}Q_{ab},$$
$$M s_{\rm ent,spin}[\hat{Q}] = -\beta G^{\rm spin}_{n,0}[\hat{\epsilon}_0^*] + M\sum_{a<b}(\varepsilon_0^*)_{ab}q_{ab}. \tag{B37}$$

The last term in Eq. (B36) is the interaction part of the free-energy $-\beta F_{\rm ex}[\{\hat{Q}_\blacksquare, \hat{q}_\blacksquare\}]$ [see Eq. (B23)].

In the dense limit $\lim_{c\to\infty}\lim_{N\to\infty}$, we have found that only the first-order term $-\beta\tilde{F}_{n,1}$ [see Eq. (B26)] in the Plefka expansion contributes to $-\beta\tilde{F}_{\rm ex}[\hat{Q}, \hat{q}, \{\partial/\partial_{\mu,\blacksquare,a}\}]$. Thus we find

$$\frac{-\beta F_{\rm ex}[\{\hat{Q}_\blacksquare, \hat{q}_\blacksquare\}]}{M} = \sum_{l=1,2,\ldots,L-1}\sum_{\blacksquare\in l}(-\mathcal{F}_{\rm int})[\hat{\lambda}_\blacksquare] \tag{B38}$$

with

$$-\mathcal{F}_{\rm int}[\hat{\Lambda}] = \ln \exp\left[\sum_{a,b}\Lambda_{ab}\frac{\partial^2}{\partial h_a\partial h_b}\right]\prod_a e^{-\beta v(h_a)}\Bigg|_{h=0} \tag{B39}$$

and

$$\lambda_{ab,\blacksquare} = Q_{ab,\blacksquare}q_{ab,\blacksquare}\frac{1}{c}\sum_{k=1}^{c}q_{ab,\blacksquare(k)}. \tag{B40}$$

On the boundaries we have $q_{ab,\blacksquare} = q_{ab,\blacksquare} = 1$ for $\blacksquare \in 0$ and $\blacksquare \in L$.

Finally, assuming that order parameters are homogeneous within the layers Eq. (19), we find the expression Eq. (24).

The saddle-point equations are

$$\begin{aligned}0 &= \frac{\partial}{\partial Q_{ab,\blacksquare}}(-\beta F_n[\{\hat{Q}, \hat{q}\}])\\ &= \frac{1}{\alpha}\frac{\partial}{\partial Q_{ab,\blacksquare}}s_{\rm ent,bond}[\hat{Q}_\blacksquare] + \sum_{\square}\frac{\partial\Lambda_\square}{\partial Q_{ab,\blacksquare}}(-\mathcal{F}_{\rm int})'[\Lambda_\square],\\ 0 &= \frac{\partial}{\partial q_{ab,\blacksquare}}(-\beta F_n[\{\hat{Q}, \hat{q}\}])\\ &= \frac{1}{\alpha}\frac{\partial}{\partial Q_{ab,\blacksquare}}s_{\rm ent,bond}[\hat{Q}_\blacksquare] + \sum_{\square}\frac{\partial\Lambda_\square}{\partial q_{ab,\blacksquare}}(-\mathcal{F}_{\rm int})'[\Lambda_\square].\end{aligned} \tag{B41}$$

6. Franz-Parisi's potential in the replica-symmetric ansatz

Here we display the expressions for the Franz-Parisi potential within the replica-symmetric ansatz needed to evaluate the generalization error.

From (128) of [15] we find

$$\begin{aligned}s_{\rm ent,spin}[\hat{\epsilon}^{1+s}, \hat{q}^{1+s}] &= s\epsilon_r r + \frac{1}{2}\epsilon_r + \frac{s}{2}\sum_{i=0}^{k}\epsilon_i q_i(m_i - m_{i+1}) + \frac{s}{2}\epsilon_k\\ &\quad + \ln \exp\left[\frac{\Lambda^{\rm Ising}_{\rm com}}{2}\sum_{a,b=0}^{s}\frac{\partial^2}{\partial h_a\partial h_b}\right]\prod_{i=0}^{k}\exp\left[\frac{\Lambda^{\rm Ising}_i}{2}\sum_{a,b=1}^{s}I^{m_i}_{ab}\frac{\partial^2}{\partial h_a\partial h_b}\right]\prod_{a=0}^{s}[2\cosh(h_a)]\Bigg|_{\{h_a=0\}}\\ &= s\epsilon_r r + \frac{1}{2}\epsilon_r + \frac{s}{2}\sum_{i=0}^{k}\epsilon_i q_i(m_i - m_{i+1}) + \frac{s}{2}\epsilon_k + \ln\gamma_{\Lambda_{\rm com}}\otimes\left[2\cosh(h)\gamma_{\Lambda_0^{\rm Ising}}\otimes e^{-sf^{\rm Ising}(m_1,h)}\right]\Big|_{h=0}.\end{aligned} \tag{B42}$$

Then we find

$$\partial_s s_{\rm ent,spin}[\hat{\epsilon}^{1+s}, \hat{q}^{1+s}]|_{s=0} = \epsilon_r r + \frac{1}{2}\sum_{i=0}^{k}\epsilon_i q_i(m_i - m_{i+1}) + \frac{1}{2}\epsilon_k + \frac{\int Dz_{\rm com}(2\cosh(\sqrt{\Lambda_{\rm com}}z_{\rm com}))\int Dz_0\big(-f^{\rm Ising}\big(m_1, \sqrt{\Lambda_{\rm com}}z_{\rm com} + \sqrt{\Lambda_0^{\rm Ising}}z_0\big)\big)}{\int Dz_{\rm com}(2\cosh(\sqrt{\Lambda_{\rm com}}z_{\rm com}))}. \tag{B43}$$

For the interaction part of the free energy, we find from (134) of [15],

$$
\begin{aligned}
&-\partial_s \mathcal{F}_{\text{int}}[\hat{q}^{1+s}(l-1), \hat{Q}^{1+s}(l), \hat{q}^{1+s}(l)]\big|_{s=0} = -\partial_s \ln \exp\left[\frac{\Lambda_{\text{com}}(l)}{2}\sum_{a,b=0}^{s}\frac{\partial^2}{\partial h_a \partial h_b}\right]\exp\left[\frac{\Lambda_{\text{teacher}}(l)}{2}\frac{\partial^2}{\partial h_0^2}\right],\\
&\prod_{i=0}^{k+1}\exp\left[\frac{\Lambda_i(l)}{2}\sum_{a,b=1}^{s} I_{ab}^{m_i}\frac{\partial^2}{\partial h_a \partial h_b}\right]\prod_{a=0}^{s} e^{-\beta v[r(h_a)]}\Bigg|_{\{h_a=0\}}\Bigg|_{s=0}\\
&\quad = -\partial_s \ln \int Dz_{\text{com}} \int Dz_{\text{teacher}} e^{-\beta v[\sqrt{\Lambda_{\text{com}}(l)}z_{\text{com}} + \sqrt{\Lambda_{\text{teacher}}(l)}z_{\text{teacher}}]}\int Dz_0 e^{-sf(m_1, \sqrt{\Lambda_{\text{com}}(l)}z_{\text{com}} + \sqrt{\Lambda_0(l)}z_0)}\Bigg|_{s=0}\\
&\quad = \frac{\int Dz_{\text{com}} g_{\text{teacher}}(\sqrt{\Lambda_{\text{com}}(l)}z_{\text{com}})\int Dz_0(-f(m_1, \sqrt{\Lambda_{\text{com}}(l)}z_{\text{com}} + \sqrt{\Lambda_0(l)}z_0))}{\int Dz_{\text{com}} g_{\text{teacher}}[\sqrt{\Lambda_{\text{com}}(l)}z_{\text{com}}]},
\end{aligned}
\tag{B44}
$$

where we introduced [see (147) of [15]]

$$
g_{\text{teacher}}(h) \equiv \int Dz_{\text{teacher}} e^{-\beta v(h - \sqrt{\Lambda_{\text{teacher}}}z_{\text{teacher}})}.
\tag{B45}
$$

7. Quadratic and cubic expansions of the free energy

Here we expand the free-energy functional given by Eq. (B36) supplemented by Eqs. (B38), (B39), and (B40) around the saddle point given by Eq. (B41). We can write

$$
Q_{ab,\blacksquare} = Q^*_{ab}(l) + \Delta Q_{ab,\blacksquare}, \quad q_{ab,\blacksquare} = q^*_{ab}(l) + \Delta q_{ab,\blacksquare},
\tag{B46}
$$

where $Q^*_{ab}(l)$ and $q^*_{ab}(l)$ are the saddle-point values of the order parameters, l is the label of the layer to which $\blacksquare$ belongs, and $\Delta Q_{ab,\blacksquare}$ and $\Delta q_{ab,\blacksquare}$ are fluctuations around the saddle point.

a. Quadratic expansion

The quadratic expansion of the replicated free-energy functional is specified in the Hessian matrix. It is obtained as

$$
\begin{aligned}
H^{QQ}_{ab,cd,\blacksquare_1,\blacksquare_2} &= \frac{\partial^2}{\partial Q_{ab,\blacksquare_1}\partial Q_{cd,\blacksquare_2}}\frac{(\beta F_n)[\{\hat{Q},\hat{q}\}]}{M} = -\delta_{\blacksquare_1,\blacksquare_2}\frac{1}{\alpha}\frac{\partial^2}{\partial Q^2_{ab,\blacksquare_1}} s_{\text{ent,bond}}[\hat{Q}_{\blacksquare_1}]\\
&\quad - \sum_{\square}\frac{\partial^2 \Lambda_\square}{\partial Q_{ab,\blacksquare_1}\partial Q_{ab,\blacksquare_2}}(-\mathcal{F}_{\text{int}})'[\Lambda_\square] - \sum_{\square}\frac{\partial \Lambda_\square}{\partial Q_{ab,\blacksquare_1}}\frac{\partial \Lambda_\square}{\partial Q_{ab,\blacksquare_2}}(-\mathcal{F}_{\text{int}})''[\Lambda_\square]\\
&= -\delta_{\blacksquare_1,\blacksquare_2}\left[\frac{1}{\alpha}\frac{\partial^2}{\partial Q^2_{ab,\blacksquare_1}} s_{\text{ent,bond}}[\hat{Q}_{\blacksquare_1}] + \frac{\partial^2 \Lambda_{\blacksquare_1}}{\partial Q^2_{ab,\blacksquare_1}}(-\mathcal{F}_{\text{int}})'[\Lambda_{\blacksquare_1}] + \left(\frac{\partial \Lambda_{\blacksquare_1}}{\partial Q_{ab,\blacksquare_1}}\right)^2(-\mathcal{F}_{\text{int}})''[\Lambda_{\blacksquare_1}]\right],\\
H^{Qq}_{ab,cd,\blacksquare_1,\blacksquare_2} &= \frac{\partial^2}{\partial Q_{ab,\blacksquare_1}\partial q_{cd,\blacksquare_2}}\frac{(\beta F)[\{\hat{Q},\hat{q}\}]}{M}\\
&= -\sum_{\square}\frac{\partial^2 \Lambda_\square}{\partial Q_{ab,\blacksquare_1}\partial q_{ab,\blacksquare_2}}(-\mathcal{F}_{\text{int}})'[\Lambda_\square] - \sum_{\square}\frac{\partial \Lambda_\square}{\partial Q_{ab,\blacksquare_1}}\frac{\partial \Lambda_\square}{\partial q_{ab,\blacksquare_2}}(-\mathcal{F}_{\text{int}})''[\Lambda_\square]\\
&= -\frac{\partial^2 \Lambda_{\blacksquare_1}}{\partial Q_{ab,\blacksquare_1}\partial q_{ab,\blacksquare_2}}(-\mathcal{F}_{\text{int}})'[\Lambda_{\blacksquare_1}] - \frac{\partial \Lambda_{\blacksquare_1}}{\partial Q_{ab,\blacksquare_1}}\frac{\partial \Lambda_{\blacksquare_1}}{\partial q_{ab,\blacksquare_2}}(-\mathcal{F}_{\text{int}})''[\Lambda_{\blacksquare_1}],\\
H^{qq}_{ab,cd,\blacksquare_1,\blacksquare_2} &= \frac{\partial^2}{\partial q_{ab,\blacksquare}\partial q_{cd,\square}}\frac{(\beta F)[\{\hat{Q},\hat{q}\}]}{M} = -\delta_{\blacksquare_1,\blacksquare_2}\frac{\partial^2}{\partial q^2_{ab,\blacksquare_1}} s_{\text{ent,spin}}[\hat{q}_{\blacksquare_1}]\\
&\quad - \sum_{\square}\frac{\partial^2 \Lambda_\square}{\partial q_{ab,\blacksquare_1}\partial q_{ab,\blacksquare_2}}(-\mathcal{F}_{\text{int}})'[\Lambda_\square] - \sum_{\square}\frac{\partial \Lambda_\square}{\partial q_{ab,\blacksquare_1}}\frac{\partial \Lambda_\square}{\partial q_{ab,\blacksquare_2}}(-\mathcal{F}_{\text{int}})''[\Lambda_\square],
\end{aligned}
\tag{B47}
$$

where

$$
\frac{\partial \Lambda_{\blacksquare_1}}{\partial Q_{ab,\blacksquare_1}} = q_{ab,\blacksquare_1}\frac{1}{c}\sum_{k=1}^{c} q_{ab,\blacksquare_1(k)}, \quad \frac{\partial^2 \Lambda_{\blacksquare_1}}{\partial Q^2_{ab,\blacksquare_1}} = 0,
\tag{B48}
$$

and

$$\begin{aligned}\frac{\partial^2 \Lambda_{\blacksquare_1}}{\partial Q_{ab,\blacksquare_1}\partial q_{ab,\blacksquare_2}} &= \delta_{\blacksquare_1,\blacksquare_2}\frac{1}{c}\sum_{k=1}^{c} q_{ab,\blacksquare_1(k)} + \frac{1}{c}q_{ab,\blacksquare_1}I_{\partial\blacksquare_1}(\blacksquare_2),\\ \frac{\partial \Lambda_{\blacksquare_1}}{\partial Q_{ab,\blacksquare_1}}\frac{\partial \Lambda_{\blacksquare_1}}{\partial q_{ab,\blacksquare_2}} &= \delta_{\blacksquare_1,\blacksquare_2}q_{ab,\blacksquare_1}Q_{ab,\blacksquare_1}\left(\frac{1}{c}\sum_{k=1}^{c} q_{ab,\blacksquare_1(k)}\right)^2 + \frac{1}{c}q^2_{ab,\blacksquare_1}Q_{ab,\blacksquare_1}\left(\frac{1}{c}\sum_{k=1}^{c} q_{ab,\blacksquare_1(k)}\right)I_{\partial\blacksquare_1}(\blacksquare_2),\end{aligned} \tag{B49}$$

and

$$\begin{aligned}\sum_{\square}\frac{\partial^2 \Lambda_{\square}}{\partial q_{ab,\blacksquare_1}\partial q_{ab,\blacksquare_2}} &= \frac{1}{c}[Q_{ab,\blacksquare_1}I_{\partial\blacksquare_1}(\blacksquare_2) + Q_{ab,\blacksquare_2}I_{\partial\blacksquare_2}(\blacksquare_1)],\\ \sum_{\square}\frac{\partial \Lambda_{\square}}{\partial q_{ab,\blacksquare_1}}\frac{\partial \Lambda_{\square}}{\partial q_{ab,\blacksquare_2}} &= \delta_{\blacksquare_1,\blacksquare_2}\left(Q_{ab,\blacksquare_1}\frac{1}{c}\sum_{k=1}^{c} q_{ab,\blacksquare_1(k)}\right)^2 + \frac{1}{c}Q^2_{ab,\blacksquare_1}q_{ab,\blacksquare_1}\frac{1}{c}\sum_{k=1}^{c} q_{ab,\blacksquare_1(k)}I_{\partial\blacksquare_1}(\blacksquare_2)\\ &\quad + \frac{1}{c}Q^2_{ab,\blacksquare_2}q_{ab,\blacksquare_2}\frac{1}{c}\sum_{k=1}^{c} q_{ab,\blacksquare_2(k)}I_{\partial\blacksquare_2}(\blacksquare_1) + \frac{1}{c^2}\sum_{\square}(q_{ab,\square}Q_{ab,\square})^2 I_{\partial\square}(\blacksquare_1)I_{\partial\square}(\blacksquare_2),\end{aligned} \tag{B50}$$

where $I_A(x)$ is the indicator function, i.e., $I_a(x) = 1$ if $x \in a$ and 0 otherwise.

Let us note that in the liquid phase where $Q_{ab} = q_{ab} = 0$ for $a \neq b$, the Hessian matrix become simplified as

$$\begin{aligned}H^{QQ}_{ab,cd,\blacksquare_1,\blacksquare_2} &= -\delta_{\blacksquare_1,\blacksquare_2}\frac{1}{\alpha}\frac{\partial^2}{\partial Q_{ab,\blacksquare_1}\partial Q_{cd,\blacksquare_1}}s_{\text{ent,bond}}[\hat{Q}_{\blacksquare_1}]\bigg|_{\hat{Q}_{\blacksquare_1}=0},\\ H^{Qq}_{ab,cd,\blacksquare_1,\blacksquare_2} &= 0,\\ H^{qq}_{ab,cd,\blacksquare_1,\blacksquare_2} &= -\delta_{\blacksquare_1,\blacksquare_2}\frac{\partial^2}{\partial q_{ab,\blacksquare_1}\partial q_{cd,\blacksquare_1}}s_{\text{ent,spin}}[\hat{q}_{\blacksquare_1}]\bigg|_{\hat{q}_{\blacksquare_1}=0}.\end{aligned} \tag{B51}$$

b. Cubic expansion

Here let us analyze the cubic expansion. For simplicity, let us only consider the liquid phase where $Q_{ab} = q_{ab} = 0$ for $a \neq b$. We find that the only nonvanishing contribution is due to

$$\begin{aligned}W^{qQq}_{ab,cd,ef,\blacksquare_1,\blacksquare_2,\blacksquare_3} &= \frac{\partial^3}{\partial q_{ab,\blacksquare_1}\partial Q_{cd,\blacksquare_2}\partial_{ef,\blacksquare_3}}\frac{(\beta F_n)[\{\hat{Q},\hat{q}\}]}{M}\bigg|_{\hat{Q}=\hat{q}=0}\\ &= -\frac{\partial^3\Lambda_{\blacksquare_2}}{\partial q_{ab,\blacksquare_1}\partial Q_{ab,\blacksquare_2}\partial q_{ab,\blacksquare_3}}(-\mathcal{F}_{\text{int}})'[\Lambda_{\blacksquare_2}]\bigg|_{\hat{Q}=\hat{q}=0}\delta_{(ab),(cd)}\delta_{(cd),(ef)}\\ &= -\frac{1}{c}\big[I_{\partial\blacksquare_2}(\blacksquare_1)\delta_{\blacksquare_2,\blacksquare_3} + \delta_{\blacksquare_2,\blacksquare_1}I_{\partial\blacksquare_2}(\blacksquare_3)\big](-\mathcal{F}_{\text{int}})'[\Lambda_{\blacksquare_2}]\big|_{\hat{Q}=\hat{q}=0}\delta_{(ab),(cd)}\delta_{(cd),(ef)}.\end{aligned} \tag{B52}$$

It is interesting to note that this cubic term breaks the symmetry concerning the exchange of input/output sides.

c. Correction to the saddle point

Now let us turn to corrections due to fluctuations around the saddle point. These give finite connectivity c or $M = c\alpha$ corrections (α is fixed).

We can write

$$Q_{ab,\blacksquare} = Q^*_{ab}(l) + \Delta Q_{ab,\blacksquare}, \quad q_{ab,\blacksquare} = q^*_{ab}(l) + \Delta Q_{ab,\blacksquare}, \tag{B53}$$

where $Q^*_{ab}(l)$ and $q^*_{ab}(l)$ are the saddle-point values of the order parameters, l is the label of the layer to which $\blacksquare$ belongs, and $\Delta Q_{ab,\blacksquare}$ and $\Delta q_{ab,\blacksquare}$ are fluctuations around the saddle point. Including the correction due to the fluctuations around the saddle point, the replicated Gardner volume Eq. (21) can be written as

$$\overline{V^{1+s}(\mathbf{S}_0, \mathbf{S}_L(\mathbf{S}_0, \mathcal{J}_{\text{teacher}}))}^{\mathbf{S}_0,\mathcal{J}_{\text{teacher}}} = e^{NMs_{1+s}[\{\hat{Q}^*,\hat{q}^*\}]}Z_{\text{fluctuation}}, \tag{B54}$$

where

$$\begin{aligned} Z_{\text{fluctuation}} = &\int \prod_{\blacksquare} \prod_{a<b} d\Delta Q_{ab,\blacksquare} d\Delta q_{ab,\blacksquare} \\ &\times \exp\Bigg[-\frac{M}{2}\sum_{a<b}\sum_{c<d}\sum_{\blacksquare,\square}\big[H^{QQ}_{ab,cd,\blacksquare,\square}\Delta Q_{ab,\blacksquare}\Delta Q_{cd,\square} + H^{Qq}_{ab,cd,\blacksquare,\square}\Delta Q_{ab,\blacksquare}\Delta q_{cd,\square} + H^{qq}_{ab,cd,\blacksquare,\square}\Delta q_{ab,\blacksquare}\Delta q_{cd,\square}\big] \\ &-\frac{M}{3!}\sum_{a<b}\sum_{c<d}\sum_{e<f}\sum_{\blacksquare_1,\blacksquare_2,\blacksquare_3}\big[W^{qQq}_{ab,cd,ef,\blacksquare_1,\blacksquare_2,\blacksquare_3}\Delta q_{ab,\blacksquare_1}\Delta Q_{cd,\blacksquare_2}\Delta q_{ef,\blacksquare_3} + \cdots\big]\Bigg], \end{aligned} \tag{B55}$$

where $H^{QQ}_{ab,cd,\blacksquare,\square}\ldots$ are the Hessian matrices given in Appendix B 7.

For the following discussion, we do not need to perform a complete analysis of the correction. We restrict ourselves in the liquid phase $Q = q = 0$. Then as shown in Appendix B 7, the Hessian matrices become completely local, i.e., $H^{QQ}_{\blacksquare,\square} = \delta_{\blacksquare,\square} H^{QQ}_{\blacksquare,\blacksquare}$ and $H^{qq}_{\blacksquare,\square} = \delta_{\blacksquare,\square} H^{qq}_{\blacksquare,\blacksquare}$ while $H^{Qq}_{\blacksquare,\square} = 0$ [see Eq. (B51)]. Thus at the quadratic level of fluctuations, there is no correlation between different layers in the liquid phase. In the cubic order, we find $W^{qQq}_{\blacksquare_1,\blacksquare_2,\blacksquare_3} \propto \frac{1}{c}[I_{\partial\blacksquare_2}(\blacksquare_1)\delta_{\blacksquare_2,\blacksquare_3} + \delta_{\blacksquare_2,\blacksquare_1} I_{\partial\blacksquare_2}(\blacksquare_3)]$ [see Eq. (B52)]. This will induce correlations between different layers even in the liquid phase. And this will be enhanced next to the frozen wall and by correlation in the frozen wall (the hidden manifold model).

To understand the key point, it is sufficient to consider a simplified model,

$$Z = \int \prod_{\blacksquare} dx_{\blacksquare} dy_{\blacksquare} \exp\left[-\frac{M}{2}\sum_{\blacksquare} h_{xx} x^2_{\blacksquare} - \frac{M}{2}\sum_{\blacksquare} h_{yy} y^2_{\blacksquare} - \alpha w \sum_{\blacksquare}\sum_{\square\in\partial\blacksquare} y_{\blacksquare} x_{\blacksquare} y_{\square}\right]. \tag{B56}$$

Then by introducing

$$Z_0 = \int \prod_{\blacksquare\in(1,2,\ldots,L)} dx_{\blacksquare} \exp\left[-\frac{M}{2}\sum_{\blacksquare} h_{xx} x^2_{\blacksquare}\right] \int \prod_{\blacksquare\in(1,2,\ldots,L-1)} dy_{\blacksquare} \exp\left[-\frac{M}{2}\sum_{\blacksquare} h_{yy} y^2_{\blacksquare}\right] = \left(\sqrt{\frac{2\pi}{Mh_{xx}}}\right)^{NL} \left(\sqrt{\frac{2\pi}{Mh_{yy}}}\right)^{N(L-1)} \tag{B57}$$

and

$$\langle \cdots \rangle_{x,y} = \frac{\int \prod_{\blacksquare} dx_{\blacksquare} \prod_{\blacksquare} dy_{\blacksquare} \exp\left[-M\frac{h_{xx}}{2}\sum_{\blacksquare} x^2_{\blacksquare} - M\frac{h_{yy}}{2}\sum_{\blacksquare} y^2_{\blacksquare}\right]\cdots}{\int \prod_{\blacksquare} dx_{\blacksquare} \prod_{\blacksquare} dy_{\blacksquare} \exp\left[-M\frac{h_{xx}}{2}\sum_{\blacksquare} x^2_{\blacksquare} - M\frac{h_{yy}}{2}\sum_{\blacksquare} y^2_{\blacksquare}\right]} \tag{B58}$$

we can write

$$\begin{aligned} \ln Z - \ln Z_0 &= \ln\left\langle \exp\left[-\alpha w \sum_{\blacksquare}\sum_{\square\in\partial\blacksquare} y_{\blacksquare} x_{\blacksquare} y_{\square}\right]\right\rangle_x \\ &= -\alpha w \sum_{\blacksquare}\sum_{\square\in\partial\blacksquare}\langle y_{\blacksquare} x_{\blacksquare} y_{\square}\rangle_{xy} + \frac{1}{2}(\alpha w)^2 \sum_{\blacksquare_1}\sum_{\square_1\in\partial\blacksquare_1}\sum_{\blacksquare_2}\sum_{\square_2\in\partial\blacksquare_2}\langle y_{\blacksquare_1} x_{\blacksquare_1} y_{\square_1} y_{\blacksquare_2} x_{\blacksquare_2} y_{\square_2}\rangle_{xy} + \cdots \\ &= \frac{1}{2}(\alpha w)^2 \sum_{\blacksquare}\sum_{\square\in\partial\blacksquare}\langle \underbrace{y^2_{\blacksquare} x^2_{\blacksquare} y^2_{\square}}_{M^3(h_{yy}h_{xx}h_{yy})^{-1}} \rangle_{xy} + \cdots . \end{aligned} \tag{B59}$$

[1] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature (London) **521**, 436 (2015).

[2] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, Rev. Mod. Phys. **91**, 045002 (2019).

[3] M. Geiger, A. Jacot, S. Spigler, F. Gabriel, L. Sagun, S. d'Ascoli, G. Biroli, C. Hongler, and M. Wyart, Scaling description of generalization with number of parameters in deep learning, J. Stat. Mech.: Theor. Expt. (2020) 023401.

[4] S. Mei and A. Montanari, The generalization error of random features regression: Precise asymptotics and the double descent curve, Commun. Pure Appl. Math. **75**, 667 (2022).

[5] B. Loureiro, C. Gerbelot, M. Refinetti, G. Sicuro, and F. Krzakala, Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension, in *International Conference on Machine Learning* (PMLR, 2022), Vol. 162, pp. 14283–14314.

[6] S. d'Ascoli, L. Sagun, and G. Biroli, Triple descent and the two kinds of overfitting: where and why do they appear? J. Stat. Mech.: Theor. Expt. (2021) 124002.

[7] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Spin-glass models of neural networks, Phys. Rev. A **32**, 1007 (1985).

[8] E. Gardner, The space of interactions in neural network models, J. Phys. A **21**, 257 (1988).

[9] E. Gardner and B. Derrida, Three unfinished works on the optimal storage capacity of networks, J. Phys. A **22**, 1983 (1989).

[10] F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová, Generalisation error in learning with random features and the hidden manifold model, in *International Conference on Machine Learning* (PMLR, 2020), Vol. 119, pp. 3452–3462.

[11] M. Gabrié, A. Manoel, C. Luneau, N. Macris, F. Krzakala, L. Zdeborová *et al.*, Entropy and mutual information in models of deep neural networks, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2018), Vo. 31.

[12] B. Aubin, B. Loureiro, A. Maillard, F. Krzakala, and L. Zdeborová, The spiked matrix model with generative priors, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2019), Vo. 32.

[13] D. Schröder, H. Cui, D. Dmitriev, and B. Loureiro, Deterministic equivalent and error universality of deep random features learning, arXiv:2302.00401.

[14] H. Cui, F. Krzakala, and L. Zdeborová, Optimal learning of deep random networks of extensive-width, arXiv:2302.00375.

[15] H. Yoshino, From complex to simple: hierarchical free-energy landscape renormalized in deep neural networks, SciPost Phys. Core **2**, 005 (2020).

[16] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2001).

[17] L. Zdeborová and F. Krzakala, Statistical physics of inference: Thresholds and algorithms, Adv. Phys. **65**, 453 (2016).

[18] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2018), Vo. 31.

[19] S. Mei, A. Montanari, and P.-M. Nguyen, A mean field view of the landscape of two-layer neural networks, Proc. Natl. Acad. Sci. (USA) **115**, E7665 (2018).

[20] L. Chizat and F. Bach, On the global convergence of gradient descent for over-parameterized models using optimal transport, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2018), Vo. 31.

[21] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, Modeling the Influence of Data Structure on Learning in Neural Networks: The Hidden Manifold Model, Phys. Rev. X **10**, 041044 (2020).

[22] S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, M. Mézard, and L. Zdeborová, The gaussian equivalence of generative models for learning with shallow neural networks, in *Mathematical and Scientific Machine Learning* (PMLR, 2022), pp. 426–471.

[23] E. Barkai, D. Hansel, and H. Sompolinsky, Broken symmetries in multilayered perceptrons, Phys. Rev. A **45**, 4146 (1992).

[24] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, A learning algorithm for boltzmann machines, Cognitive Sci. **9**, 147 (1985).

[25] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, Exponential expressivity in deep neural networks through transient chaos, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2016), Vol. 29.

[26] A. J. Bray and M. A. Moore, Chaotic Nature of the Spin-Glass Phase, Phys. Rev. Lett. **58**, 57 (1987).

[27] C. M. Newman and D. L. Stein, Multiple states and thermodynamic limits in short-ranged ising spin-glass models, Phys. Rev. B **46**, 973 (1992).

[28] Y. Iba, The nishimori line and bayesian statistics, J. Phys. A **32**, 3875 (1999).

[29] R. Monasson and R. Zecchina, Weight Space Structure and Internal Representations: A Direct Approach to Learning and Generalization in Multilayer Neural Networks, Phys. Rev. Lett. **75**, 2432 (1995).

[30] E. Levin, N. Tishby, and S. A. Solla, A statistical approach to learning and generalization in layered neural networks, Proc. IEEE **78**, 1568 (1990).

[31] M. Opper and W. Kinzel, Statistical mechanics of generalization, in *Models of Neural Networks III*, edited by E. Domany, J. L. van Hemmen, and K. Schulten (Springer, New York, 1996), pp. 151–209.

[32] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction*, 111 (Clarendon, Oxford (UK), 2001), p. 111.

[33] S. Franz and G. Parisi, Recipes for metastable states in spin glasses, J. Phys. I **5**, 1401 (1995).

[34] P.-G. De Gennes, Wetting: statics and dynamics, Rev. Mod. Phys. **57**, 827 (1985).

[35] F. Krzakala and L. Zdeborová, On melting dynamics and the glass transition. i. glassy aspects of melting dynamics, J. Chem. Phys. **134**, 034512 (2011).

[36] F. Krzakala and L. Zdeborová, On melting dynamics and the glass transition. ii. glassy dynamics as a melting process, J. Chem. Phys. **134**, 034513 (2011).

[37] G. Györgyi, First-order transition to perfect generalization in a neural network with binary synapses, Phys. Rev. A **41**, 7097 (1990).

[38] K. Hukushima and H. Kawamura, Chiral-glass transition and replica symmetry breaking of a three-dimensional heisenberg spin glass, Phys. Rev. E **61**, R1008(R) (2000).

[39] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, Similarity of neural network representations revisited, in *International Conference on Machine Learning* (PMLR, 2019), pp. 3519–3529.

[40] W. Zou and H. Huang, Data-driven effective model shows a liquid-like deep learning, Phys. Rev. Res. **3**, 033290 (2021).

[41] G. Parisi and M. A. Virasoro, On a mechanism for explicit replica symmetry breaking, J. Phys. **50**, 3317 (1989).

[42] T. Plefka, Convergence condition of the tap equation for the infinite-ranged ising spin glass model, J. Phys. A **15**, 1971 (1982).

[43] J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids* (Elsevier, Amsterdam, 1990).

[44] J. Zinn-Justin, *Quantum Field Theory and Critical Phenomena* (Oxford University Press, Oxford, 2021).