Solution space and storage capacity of fully connected two-layer neural networks with generic activation functions

Sota Nishiyama[1, *] and Masayuki Ohzeki[2, 3, 4, 5]

[1]*School of Engineering, Tohoku University, Sendai, 980-8579, Japan*
[2]*Graduate School of Information Sciences, Tohoku University, Sendai, 980-8579, Japan*
[3]*Department of Physics, Tokyo Institute of Technology, Tokyo, 152-8551, Japan*
[4]*International Research Frontier Initiative, Tokyo Institute of Technology, Tokyo, 108-0023*
[5]*Sigma-i Co., Ltd., Tokyo, 108-0075, Japan*
(Dated: April 23, 2024)

The storage capacity of a binary classification model is the maximum number of random input-output pairs per parameter that the model can learn. It is one of the indicators of the expressive power of machine learning models and is important for comparing the performance of various models. In this study, we analyze the structure of the solution space and the storage capacity of fully connected two-layer neural networks with general activation functions using the replica method from statistical physics. Our results demonstrate that the storage capacity per parameter remains finite even with infinite width and that the weights of the network exhibit negative correlations, leading to a *division of labor*. In addition, we find that increasing the dataset size triggers a phase transition at a certain transition point where the permutation symmetry of weights is broken, resulting in the solution space splitting into disjoint regions. We identify the dependence of this transition point and the storage capacity on the choice of activation function. These findings contribute to understanding the influence of activation functions and the number of parameters on the structure of the solution space, potentially offering insights for selecting appropriate architectures based on specific objectives.

I. INTRODUCTION

A. Background and Motivation

Deep Neural Networks (DNNs) demonstrate remarkable abilities across a wide range of tasks [1], yet a theoretical explanation for their high performance remains incomplete [2, 3]. The theory of deep learning is centered around three problem areas, namely *optimization*, *approximation*, and *generalization* [2]. In the problem of optimization, it is crucial to understand why DNNs can be successfully trained by gradient descent, despite the non-convex nature of their loss functions [4, 5]. Analyzing the structure of the loss function and the solution space can provide insights into characterizing the condition for successful learning [6–8].

The problem of approximation considers the ability of NNs to approximate a given function. A classical result in this problem is the universal approximation theorem for two-layer NNs [9]. This theorem claims that, under mild assumptions about activation functions, sufficiently wide two-layer neural networks can approximate any continuous function with arbitrary precision. From a practical standpoint, interest lies in determining precisely how many parameters are needed and in its dependence on the choice of network architectures.

Lastly, the issue of the generalization is about reconciling the bias-variance trade-off and the seemingly excessive number of parameters in modern NNs. Conventional statistical and learning theories predict a trade-off between the number of parameters in a machine learning model and its generalization performance [10]. Nonetheless, modern DNNs often contain a large number of parameters, to the extent that they can perfectly fit the training data [11]. In particular, the phenomenon of *double descent*, where increasing the number of parameters beyond the capacity to perfectly memorize the dataset further reduces generalization error, has been reported [12–14].

Statistical mechanics is a powerful theoretical framework for addressing these problems as a tool to study the typical behavior of systems with many degrees of freedom [3]. In particular, methods derived from spin glass theory such as replica and cavity methods [15] are useful for investigating the behavior of systems involving randomness and have been applied to evaluate the typical performance of various machine learning models, including NNs [16–18].

In the statistical mechanical analysis of NNs, there are mainly two settings considered. One is the problem of *storage capacity* [8, 19–25]. The storage capacity of a binary classifier refers to the maximum number of random input-output pairs that the model can learn. A higher storage capacity implies greater representational power and approximation capability of the model. The analysis of storage capacity examines the typical performance of learning when viewed as an optimization problem or a constraint satisfaction problem. Among the three problem areas of deep learning theory mentioned earlier, those directly relevant are optimization and approximation. In addition, one could argue a relationship with the problem of generalization through the phenomenon of double descent.

The other setting is known as the *teacher-student scenario*, which examines the typical performance of learning viewed as a statistical estimation problem [26–30]. In this setup, random input data and labels generated by a teacher model are provided to a student model for learning, aiming to estimate the parameters of the teacher model. Subsequently, the generalization ability of the student model to unknown data is evaluated. This setting is relevant to the problems of optimization and generalization.

These settings together provide complimentary insights for learning of NNs. In this study, we address the problem of storage capacity.

B. Related Works

The analysis of the storage capacity of NNs using the replica method was pioneered by Gardner and Derrida in their study of perceptrons [19, 20]. They used the replica method to determine the volume of solution space of perceptrons and showed that the storage capacity per parameter of perceptrons is two.

Gardner's framework was soon extended to two-layer NNs. Architectures considered in the study of two-layer NNs include tree-like committee machines (TCMs) and fully connected committee machines (FCMs) (see Figure 1). Storage capacities of TCMs and FCMs with sign activation functions (denoted sgn-TCM and sgn-FCM) were computed in the 1990s [21–24]. It was demonstrated that in the limit as the width of the hidden layer K tends to infinity, the storage capacity per parameter diverges to infinity as $O(\sqrt{\log K})$. Also, for sgn-FCMs, it was found that weights connected to the same input neuron have negative correlations, indicating a *division of labor* where different weights attempt to memorize different input-output pairs.

Recently, the storage capacity of TCMs with activation functions other than the sign function, such as Rectified Linear Units (ReLU), has been computed [8, 25]. The results showed that, unlike the case of the sign function, the storage capacity remains finite even in the limit as K tends to infinity.

These results are based on the non-rigorous replica method, prompting efforts to support them through more mathematically rigorous approaches. Stojnic recently published a series of preprints asserting proof that the storage capacities of TCMs derived from replica methods provide rigorous upper bounds on the actual storage capacities [31–34].

NNs commonly employed in practical applications often incorporate fully connected layers featuring various activation functions. Thus, investigating the storage capacity of FCMs with general activation functions, which is yet to be analyzed, would shed light on the performance of NNs used in real-world scenarios. In this study, we analyze the solution space and the storage capacity of FCMs with general activations, aiming to bridge the
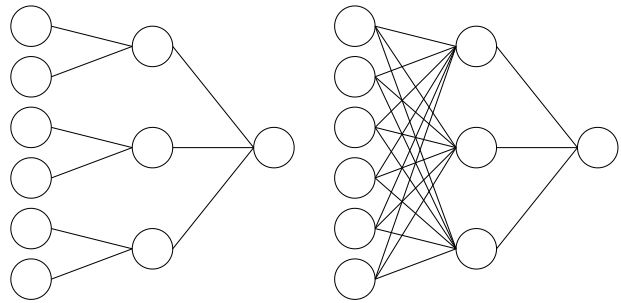

FIG. 1. Architectures of a tree-like committee machine (TCM, left) and a fully connected committee machine (FCM, right).

gap in previous studies.

C. Organization of This Paper

The remainder of this paper is organized as follows. In section II, we formulate the problem of the storage capacity of an FCM and present our analysis using the replica method. In section III, we conduct a numerical experiment where we train FCMs on randomly generated datasets to support our analysis. Finally, in section IV, we summarize the results and discuss future directions.

II. REPLICA ANALYSIS

A. Setup

The FCM architecture that we consider in this study is a fully connected two-layer network with N inputs, K hidden neurons, and one output. We make the following assumptions on the activation $g : \mathbb{R} \to \mathbb{R}$: the weak differentiability and the square-integrability of the weak derivative with respect to the Gaussian measure. Most activation functions used in practice satisfy these conditions, while the sign function does not.

Using the hidden weights $\{\boldsymbol{w}_k \in \mathbb{R}^N\}_{k=1,2,\dots,K}$ and the output weights $\{v_k \in \mathbb{R}\}_{k=1,2,\dots,K}$, the output $\hat{y}(\boldsymbol{x}) \in \{+1, -1\}$ of the model for an input $\boldsymbol{x} \in \mathbb{R}^N$ is given by

$$\hat{y}(\boldsymbol{x}) = \text{sgn}(s(\boldsymbol{x}; \{\boldsymbol{w}_k\}, \{v_k\})), \qquad (1)$$

$$s(\boldsymbol{x}; \{\boldsymbol{w}_k\}, \{v_k\}) = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} v_k g(\boldsymbol{w}_k \cdot \boldsymbol{x}), \qquad (2)$$

where sgn is the sign function.

We fix the output weights $\{v_k\}$ and learn only the hidden weights $\{\boldsymbol{w}_k\}$. To make the distribution of s symmetric even with an asymmetric activation function g, we set half of $\{v_k\}$ to $+1$ and the other half to -1. We enforce the norm constraint $\|\boldsymbol{w}_k\|_2^2 = N$ for $k = 1, 2, \dots, K$ on the hidden weights.

The dataset $\mathcal{D}$ is a list of P random input-output pairs $\mathcal{D} = \{(\boldsymbol{x}^\mu, y^\mu)\}_{\mu=1,2,\dots,P}$. Inputs $\boldsymbol{x}^\mu \in \mathbb{R}^N$ and outputs $y^\mu \in \{-1, +1\}$ are i.i.d. samples from $\mathcal{N}(0, N^{-1}I_N)$ and Uniform$(\{-1, +1\})$, respectively, where I_N is the $N \times N$ identity matrix. The number of samples per parameter is denoted as $\alpha := P/(NK)$.

To analyze the solution space structure of FCM, we consider the following partition function or the *Gardner volume*:

$$Z = \int \prod_{k=1}^K \mathrm{d}\boldsymbol{w}_k \delta(N - \boldsymbol{w}_k \cdot \boldsymbol{w}_k)$$
$$\times \prod_{\mu=1}^P \Theta(y^\mu s(\boldsymbol{x}^\mu; \{\boldsymbol{w}_k\}, \{v_k\}) - \kappa), \quad (3)$$

where Θ is the Heaviside step function. This is a volume of the space of weights $\{\boldsymbol{w}_k\}$ that classify the dataset $\mathcal{D}$ correctly with a margin of at least κ.

We expect $\log Z$ to be self-averaging; in the $N \to \infty$ limit, the distribution of $\log Z$ with respect to the realization of the disorder $\mathcal{D}$ concentrates sharply around its mean $[\log Z]_\mathcal{D}$, where $[\cdot]_\mathcal{D}$ is the expected value with respect to the distribution of $\mathcal{D}$. Under this assumption, we calculate the *quenched free entropy* per parameter

$$f := \lim_{K \to \infty} \lim_{N \to \infty} \frac{1}{NK}[\log Z]_\mathcal{D}. \quad (4)$$

Here, the infinite width limit $K \to \infty$ is taken to simplify the analysis.

B. Replica Calculation

To calculate the quenched free entropy (4), we employ the replica method, which is based on the identity $[\log Z]_\mathcal{D} = \lim_{n \to 0} n^{-1} \log[Z^n]_\mathcal{D}$. We evaluate $[Z^n]_\mathcal{D}$ for positive integer n and extrapolate the results to $n = 0$. We do not mathematically justify the analytic continuation from $n \in \mathbb{Z}_+$ to $n \in \mathbb{R}$ and the exchange of limits $n \to 0$ and $N \to \infty$. The replica method is thus a heuristic rather than a rigorous method.

Here, we sketch the outline of the calculation of $[Z^n]$ and state the main results. The calculation proceeds along the same line as sgn-FCMs [21, 22] and TCMs with generic activations [25]. We defer the details to the Supplemental Material [35].

1. Saddle-point equations

By evaluating $[Z^n]_\mathcal{D}$ for $n \in \mathbb{Z}_+$, we can formulate f as an extremization problem of a certain objective function as follows:

$$f = \lim_{n \to 0} \underset{Q, \hat{Q} \in \mathcal{S}_{nK}}{\mathrm{extr}} \left\{ G_s(Q, \hat{Q}) + \alpha G_e(Q) \right\}, \quad (5)$$

where we use $\mathcal{S}_{nK}$ to denote the set of $nK \times nK$ nonnegative-definite symmetric matrices and $\mathrm{extr}_{x \in S} f(x)$ to denote the saddle-point of f i.e. the value $f(x)$ for $x \in S$ such that $f'(x) = 0$. Here, the entropy term G_s and the energy term G_e are given as

$$G_s(Q, \hat{Q}) := \frac{1}{2nK}\Big(\mathrm{tr}(Q\hat{Q}) - \log \det \hat{Q}\Big), \quad (6)$$

$$G_e(Q) := \frac{1}{n} \log \left[\prod_{a=1}^n \Theta\left(y \frac{1}{\sqrt{K}} \sum_{k=1}^K v_k g(v_k h_k^a) - \kappa \right) \right]_\mathcal{D}, \quad (7)$$

where

$$h_k^a := v_k \boldsymbol{w}_k^a \cdot \boldsymbol{x}. \quad (8)$$

For replica indices $1 \le a, b \le n$ and hidden unit indices $1 \le k, l \le K$, components of the order parameters $Q, \hat{Q}$ are defined as

$$Q_{kl}^{ab} := \frac{1}{N} v_k v_l \boldsymbol{w}_k^a \cdot \boldsymbol{w}_l^b = \begin{cases} 1 & (k = l, a = b) \\ q^{ab} & (k = l, a \ne b) \\ c^a & (k \ne l, a = b) \\ d^{ab} & (k \ne l, a \ne b) \end{cases} \quad (9)$$

$$\hat{Q}_{kl}^{ab} := \begin{cases} \hat{E}^a & (k = l, a = b) \\ -\hat{q}^{ab} & (k = l, a \ne b) \\ \hat{c}^a & (k \ne l, a = b) \\ -\hat{d}^{ab} & (k \ne l, a \ne b) \end{cases} \quad (10)$$

where we omit the dependence on the hidden unit indices k, l due to the symmetry of the problem. q^{ab} represents the overlap of the same weights across different replicas, characterizing the size of *clusters* or the number of solutions. Here, a cluster refers to a single connected region in the solution space of the FCM. As q approaches one, the size of clusters decreases and the number of solutions diminishes. c^a denotes the overlap of different weights within a single replica, characterizing the degree of a division of labor. A negative value of c suggests the occurrence of a division of labor.

d^{ab} is associated with the concept of *permutation symmetry* (PS). For an FCM, permuting the indices of the hidden weights of a solution yields another solution. When this permuted solution belongs to the same cluster as the original solution, we say that PS holds. PS phase is characterized by $q^{ab} = d^{ab}$. If solutions belong to different clusters, we say that there is a permutation symmetry breaking (PSB), characterized by $q^{ab} \ne d^{ab}$. In sgn-FCM, it is known that PS holds for small α, but PSB occurs for large α [22].

Solving the extremization problem of Equation 5 for general $nK \times nK$ matrices $Q, \hat{Q}$ is intractable. Hence we need to make assumptions on the structure of $Q, \hat{Q}$ and parametrize them with a small number of order parameters. We consider two kinds of assumptions: *replica symmetry* (RS) and *one-step replica symmetry breaking* (1-RSB).

Under the RS ansatz, we assume that the components of Q and $\hat{Q}$ do not depend on the replica indices a, b and parametrize them with seven scalar order parameters $q, c, d, \hat{E}, \hat{q}, \hat{c}, \hat{d} \in \mathbb{R}$. This ansatz is equivalent to assuming that solutions of the FCM form a single connected cluster. Using these order parameters, the free entropy (5) is written as

$$f_{\mathrm{RS}} = \operatorname*{extr}_{q,\bar{c},\bar{d},\hat{E},\hat{q},\hat{c},\hat{d}} \left\{ \frac{1}{2} \left(\hat{E} + q\hat{q} + \bar{c}\hat{c} + \bar{d}\hat{d} + \frac{\hat{q} - \hat{d}}{\hat{E} + \hat{q} - \hat{c} - \hat{d}} - \log(\hat{E} + \hat{q} - \hat{c} - \hat{d}) \right) + \alpha \int Dz \log H\left(\frac{\kappa + \sqrt{\lambda}z}{\sqrt{\mu}} \right) \right\}, \quad (11)$$

where Dx is the standard Gaussian measure $\exp(-x^2/2)/\sqrt{2\pi}\,dx$ and $H(x)$ is the tail distribution of the standard normal distribution $H(x) = \int_x^\infty Dz$. In addition, we have defined the rescaled order parameters $\bar{c} := (K-1)c$, $\bar{d} := (K-1)d$ and defined λ, μ as

$$\lambda = q_{\mathrm{eff}} + \langle g' \rangle^2 \bar{d}, \quad (12)$$
$$\mu = \sigma^2 - q_{\mathrm{eff}} + \langle g' \rangle^2 (\bar{c} - \bar{d}), \quad (13)$$

where $\langle \cdot \rangle$ is the expected value with respect to the standard Gaussian measure Dx, $\sigma^2 = \langle g^2 \rangle - \langle g \rangle^2$ is the variance of g, and q_{eff} is the *effective order parameter* defined as

$$q_{\mathrm{eff}} := \operatorname{Cov}\left[g(x), g(y) \,\middle|\, \begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & q \\ q & 1 \end{pmatrix} \right) \right]. \quad (14)$$

By differentiating the RS free entropy (11) with each order parameter, we obtain saddle-point equations, a system of equations that gives a solution to the extremization problem (11), as follows:

$$\hat{q} = \alpha \frac{q'_{\mathrm{eff}}}{\mu} \int Dz \left(\frac{H'(u)}{H(u)} \right)^2 \quad (15)$$

$$\hat{c} = \alpha \frac{\langle g' \rangle^2}{\mu} \int Dz \frac{H'(u)}{H(u)} u \quad (16)$$

$$\hat{d} = \alpha \frac{\langle g' \rangle^2}{\mu} \int Dz \left(\frac{H'(u)}{H(u)} \right)^2 \quad (17)$$

$$\hat{E} + \hat{q} - \hat{c} - \hat{d} = \frac{1 + \sqrt{4(\hat{q} - \hat{d}) + 1}}{2} \quad (18)$$

$$q = \frac{\hat{q} - \hat{d}}{(\hat{E} + \hat{q} - \hat{c} - \hat{d})^2} \quad (19)$$

$$\bar{c} = -1 \quad (20)$$
$$\bar{d} = -q \quad (21)$$

where $u = (\kappa + \sqrt{\lambda}z)/\sqrt{\mu}$. We obtain the result $c = -1/(K-1), d = -q/(K-1)$ regardless of α, κ and g. The order parameter q can be determined by solving the saddle-point equations (15)–(21). We show in Figure 2
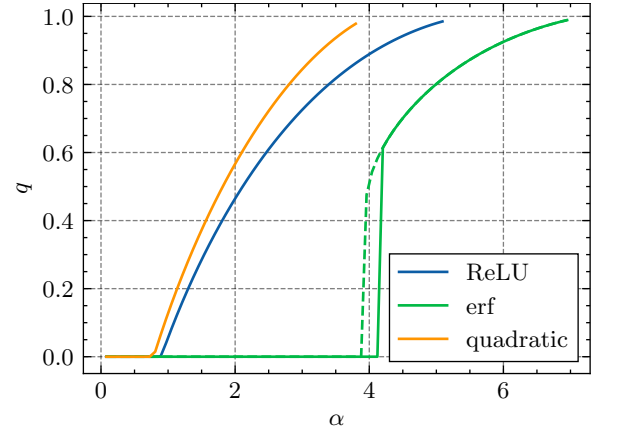

FIG. 2. The overlap between replicas q at $\kappa = 0$ for activations ReLU, erf, and quadratic x^2. We solved the saddle-point equations (15)–(21) iteratively starting from an appropriate set of initial values. For ReLU and quadratic, the transition from the PS phase with $q = 0$ to the PSB phase with $q > 0$ is continuous. The transition point α_{PS} is approximately 0.897 for ReLU and 0.785 for quadratic. On the other hand, in the case of erf activation, we observe a discontinuous transition to the PSB phase. At $\alpha_{\mathrm{spin}} \approx 3.949$, a local minimum of f_{RS} with $q > 0$ appears (dashed line). Above $\alpha_{\mathrm{PS}} \approx 4.142$, the $q > 0$ solution gives the global minimum of the f_{RS}, marking a discontinuous transition to the PSB phase (solid line).

the solutions for three activation functions: the rectified linear unit $\mathrm{ReLU}(x) = \max\{x, 0\}$, the error function $\mathrm{erf}(x) = 2/\sqrt{\pi} \int_0^x \exp(-t^2)\,dt$, and the quadratic function x^2. The overlap q increases monotonically as the number of samples per parameter α increases, showing a decrease in the number of solutions. For small α, we obtain a $q = 0$ solution, and for large α, we obtain a $q > 0$ solution. The transition from the $q = 0$ solution to the $q > 0$ solution is the permutation symmetry breaking and will be discussed in the next subsection. At some α, q reaches one and there exists a unique set of weights that correctly classifies the given dataset. This value of α is the RS storage capacity.

2. Permutation symmetry and its breaking

At the transition point α_{PS} which depends on the activation function, the system undergoes a phase transition from the PS phase with $q = d = 0$ to the PSB phase $q \neq d$. To examine the properties of this transition, we derive the condition for the PS solution $q = d = 0$ to be locally stable in the saddle-point equations (15)–(21). When $q \ll 1$ and $\hat{q} - \hat{d} \ll 1$, the leading terms in the saddle-point equations are given by

$$\hat{q} - \hat{d} \approx \alpha \frac{\langle g'' \rangle^2}{\mu} \left(\frac{H'(u)}{H(u)} \right)^2 q \tag{22}$$

$$q \approx \hat{q} - \hat{d}, \tag{23}$$

where $u = \kappa / \sqrt{\sigma^2 - \langle g' \rangle^2}$. Hence, the condition for $(q, \hat{q} - \hat{d}) = (0, 0)$ to be locally stable in the RS saddle-point equations can be expressed as follows:

$$\alpha < \frac{\sigma^2 - \langle g' \rangle^2}{\langle g'' \rangle^2} \left(\frac{H(u)}{H'(u)} \right)^2. \tag{24}$$

For activation functions satisfying $\langle g'' \rangle \neq 0$, the right-hand side of Equation 24 takes a finite value, denoted as α_{PS}. Such activation functions include ReLU and quadratic functions. When $\kappa = 0$, we have $\alpha_{\mathrm{PS}} =$

$(\pi/2 - 1)\pi/2 \approx 0.897$ for ReLU and $\alpha_{\mathrm{PS}} = \pi/4 \approx 0.785$ for quadratic functions. These values coincide with the point where q rises from 0 in Figure 2.

On the other hand, for activation functions satisfying $\langle g'' \rangle = 0$, the condition in Equation 24 is always satisfied, showing that the PS solution is always locally stable. This suggests that the transition to the PSB phase is a discontinuous first-order transition. This class of activation functions includes erf and, more generally, any odd function.

The transition to the PSB phase for $g = \mathrm{erf}$ occurs as follows. As we increase α, at the spinodal point $\alpha_{\mathrm{spin}} \approx 3.949$, a local solution $q \approx 0.41$ appears, distinct from the $q = 0$ solution. At $\alpha = \alpha_{\mathrm{spin}}$, the value of the free entropy f_{RS} for the $q > 0$ solution is greater than that of the $q = 0$ solution. Then, at $\alpha_{\mathrm{PS}} \approx 4.142$, $q \approx 0.59$ becomes the global minimum of f_{PS}, leading to a discontinuous transition from the PS solution with $q = 0$ to the PSB solution with $q > 0$.

3. Storage capacity

The value of α at which $q = 1$ corresponds to the storage capacity α_{RS}. α_{RS} can be obtained from the saddle-point equations and is expressed by the following equation:

$$\frac{1}{\alpha_{\mathrm{RS}}} = \frac{\sigma^2 - \langle g' \rangle^2}{\langle g'^2 \rangle - \langle g' \rangle^2} \int_{-\kappa/\sqrt{\sigma^2 - \langle g' \rangle^2}}^{\infty} Dz \left(\frac{\kappa}{\sqrt{\sigma^2 - \langle g' \rangle^2}} + z \right)^2. \tag{25}$$

In particular, at $\kappa = 0$,

$$\alpha_{\mathrm{RS}}(\kappa = 0) = \frac{2 \left(\langle g'^2 \rangle - \langle g' \rangle^2 \right)}{\sigma^2 - \langle g' \rangle^2}. \tag{26}$$

Omitting $\langle g' \rangle^2$ terms in these expressions gives the storage capacity of corresponding TCMs [25]. $\langle g' \rangle^2$ can be considered as a term characterizing the effect of division of labor. A comparison between the RS storage capacities of FCM and TCM is shown in Table I. For activation functions with $\langle g' \rangle \neq 0$, such as ReLU and erf, the RS storage capacities of FCMs are higher than those of the corresponding TCMs by the division of labor. However, for activation functions with $\langle g' \rangle = 0$, such as the quadratic function, the division of labor does not increase the storage capacities of FCMs.

When $\langle g' \rangle > 0$, the RS storage capacity of FCM at $\kappa = 0$ exceeds that of TCM. This holds for cases including ReLU and erf. When $\langle g' \rangle = 0$, the RS storage capacities of FCM and TCM coincide regardless of κ. This holds for cases including quadratic functions. In general, for an even activation function g, we have $\langle g' \rangle = 0$.

TABLE I. Comparison of RS storage capacities of FCMs and TCMs at $\kappa = 0$. RS storage capacities of FCMs are given by Equation 26. RS storage capacities of TCMs are obtained by setting $\langle g' \rangle^2$ to zero in Equation 26 [25]. For ReLU and erf, FCM capacity is larger than TCM capacity due to the effect of division of labor. For the quadratic function, however, the FCM capacity matches the TCM capacity because $\langle g' \rangle = 0$.

Activation	FCM capacity	TCM capacity
ReLU	$\dfrac{2\pi}{\pi - 2} \approx 5.504$	$\dfrac{2\pi}{\pi - 1} \approx 2.934$
erf	$\dfrac{4(3/\sqrt{5} - 1)}{3 \arcsin(2/3) - 2} \approx 7.223$	$\dfrac{4}{\sqrt{5} \arcsin(2/3)} \approx 2.451$
quadratic	4	4

4. 1-RSB calculation and the stability of RS solution

One can evaluate the free entropy (5) under 1-RSB ansatz to obtain the 1-RSB storage capacity. 1-RSB capacities are shown in Figure 3 along with RS capacities. The explicit values of the 1-RSB capacities $\alpha_{\text{1-RSB}}$ for
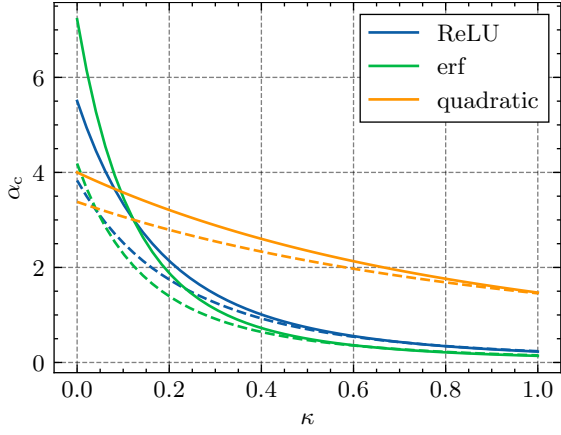
FIG. 3. Storage capacity of FCMs with ReLU, erf, and quadratic activations. Solid lines represent the RS storage capacity and dashed lines represent the 1-RSB storage capacity. For small κ, there is a large gap between RS and 1-RSB storage capacities, showing a strong replica symmetry-breaking effect.

several activation functions at $\kappa = 0$ are: 3.812 for ReLU, 4.186 for erf, and 3.375 for the quadratic.

When solving the 1-RSB saddle-point equations, the RS solution is restored for small α. Above a certain threshold α_{AT}, the RS solution becomes locally unstable and the 1-RSB solution emerges continuously. This instability of the RS solution is known as the de Almeida-Thouless (dAT) instability [36].

To determine α_{AT}, we examine the condition under which the 1-RSB saddle-point equations have $(\Delta q, \Delta \hat{q}, \Delta \bar{d}, \Delta \hat{d}) = (0, 0, 0, 0)$ as a stable fixed point, where $\Delta q := q_1 - q_0$, $\Delta \hat{q} := \hat{q}_1 - \hat{q}_0$, $\Delta \bar{d} := \bar{d}_1 - \bar{d}_0$, $\Delta \hat{d} := \hat{d}_1 - \hat{d}_0$. Thus, the condition for dAT stability for a given (α, κ) can be obtained as follows:

$$
\alpha(1-q)^2 \int Dz \left(\frac{q'_{\text{eff}} - \langle g' \rangle^2}{\lambda^2} \left(\frac{\partial^2}{\partial z^2} \log H(u) \right)^2 \right.
$$
$$
\left. + \frac{q''_{\text{eff}}}{\lambda} \left(\frac{\partial}{\partial z} \log H(u) \right)^2 \right) < 1, \quad (27)
$$

where the values of q and u are solutions to the RS saddle-point equations for the given (α, κ).

When we keep only the leading contributions by setting $q \ll 1$ in Equation 27, the condition coincides with Equation 24. Hence, for activation functions where PSB solutions appear continuously, α_{PS} and α_{AT} coincide. In other words, the PSB phase and the RSB phase coincide.

Moreover, when evaluating the stability condition (27) for the RS solution with $\kappa = 0$ and $g = \text{erf}$, we find that PS solutions for $\alpha < \alpha_{PS} \approx 4.142$ are stable, while PSB solutions for $\alpha > \alpha_{\text{spin}} \approx 3.949$ are unstable. Also for erf, RSB occurs simultaneously with PSB.

5. Comparison with teacher-student scenario

The PSB transition discussed in this section is analogous to a phenomenon known as *specialization transition* in the analysis of generalization errors of FCM in the teacher-student scenario. The specialization transition occurs when the weights of the student's FCM specialize to mimic specific weights of the teacher's FCM, breaking the symmetry of weight permutations. The specialization transition is known to be discontinuous for sgn and erf, while it is continuous for ReLU [26, 28–30]. It is noteworthy that the continuity of the PSB transition and that of the specialization transition match, at least for ReLU and erf.

III. NUMERICAL EXPERIMENT

The Gardner volume solely addresses the existence of solutions, without guaranteeing the presence of algorithms capable of efficiently discovering solutions in regions where they exist. Notably, training neural networks are known to be NP-hard in general [37, 38], suggesting the absence of polynomial-time algorithms for solution discovery. Consequently, empirically verifying the accuracy of the storage capacity derived through the replica method via numerical experiments presents a formidable challenge. Nonetheless, the discovery of solutions surpassing the theoretically derived storage capacity by an algorithm would serve to refute the theory. Thus, through experimental evidence demonstrating the inability of two-layer neural networks trained via gradient descent to discover solutions beyond the storage capacity for random datasets, we indirectly corroborate the validity of our calculation.

A. Settings of Experiments

We implement the FCM of Equation 1 using PyTorch [39] and train it using gradient descent on randomly generated datasets. We fix the number of parameters NK to 10000 and vary the ratio N/K over $\{0.5, 1, 5, 10, 50\}$. Since the replica analysis assumes $N \gg K$, higher values of N/K are closer to the settings of replica calculations. We examine the cases of ReLU and erf as activation functions g. The parameter α is varied from 0.2 to 4.8 in increments of 0.2.

We use hinge loss with $\kappa = 0$ and the Adam optimizer [40] for the learning algorithm. We terminate the training when the loss reaches below 10^{-5} or when the training reaches 5000 epochs.

For each combination of $(\alpha, N/K, g)$, we train the model on 10 different datasets. We compute the accuracy, loss, number of epochs until training completion, and the value of weight overlap c at the end of training for each trial and then average them over the 10 trials.

B. Results

The results for ReLU are shown in Figure 4. We observe abrupt increases in the loss values at around $\alpha \approx 2.5$, consistently across the values of N/K. Thus, the experimental storage capacity for ReLU is estimated to be around 2.5, which is significantly smaller than the 1-RSB storage capacity. This discrepancy is likely due to the breaking of replica symmetry and the emergence of numerous local minima at large α, causing gradient descent to become trapped in local minima. Indeed, the replica analysis predicts the breaking of replica symmetry in the PSB phase $\alpha > \alpha_{\mathrm{PS}} \approx 0.897$, suggesting the highly nonconvex loss landscape with numerous local minima. A similar phenomenon is observed in the numerical experiment for TCMs [25]. Hence, the gap in the experimental storage capacity and the theoretical 1-RSB storage capacity does not falsify the theory, as long as the former is upper-bounded by the latter.

In addition, the number of epochs required for training completion increases sharply as α grows larger. Also, the overlap $(K-1)c$ converges to -1 as α increases, which is consistent with the prediction of the replica analysis. The increase in c at small α is attributed to finite-size effects. Convergence to -1 is faster for larger values of K.

Next, the results for erf are presented in Figure 5. For erf, the loss does not seem to behave consistently across different N/K. For $N/K = 50$, which is the closest to the setting of replica analysis, the experimental storage capacity is estimated to be around 3.8, close to $\alpha_{\mathrm{spin}} \approx 3.9$ at which a PSB solution appears. Above α_{spin}, the replica symmetry is broken and the gradient descent algorithm is likely to get trapped in a local minimum. This explains the gap between the experimental storage capacity and the 1-RSB storage capacity $\alpha_{\text{1-RSB}} \approx 4.186$.

IV. DISCUSSION

A. Summary of Results

We have analyzed the structure of the solution space of fully connected two-layer neural networks (FCMs) with general activation functions using the replica method and determined their storage capacity. Under mild assumptions on the activation functions, we have shown that the storage capacity remains finite even in the infinite width limit $K \to \infty$. This result differs from that of sgn-TCM and sgn-FCM while resembling the results obtained for TCM with general activation functions. Also, the observation that the order parameter c becomes negative indicates the occurrence of a division of labor, similar to sgn-FCM.

In addition, the boundary between the PS and PSB phases, denoted as α_{PS}, is identified. The transition to the PSB phase is found to be continuous when ReLU

or quadratic functions are used as activation functions, while it is discontinuous when erf is used.

Finally, we have conducted a numerical experiment to train FCMs using gradient descent. In these experiments, the gradient descent algorithm is only able to find solutions up to significantly smaller α values compared to the 1-RSB storage capacity. This discrepancy is attributed to the non-convexity of the solution space, leading to a gap between the existence of solutions and learnability. However, consistent with our theory, the experimental estimate of the storage capacity of erf is found to be greater than that of ReLU. In addition, the value of c matched the theoretical prediction of $-1/(K-1)$.

The practical significance of this study can be summarized as follows. First, it enables the comparison of neural network performance based on the choice of activation functions. By observing the differences in storage capacity ($\alpha_{\mathrm{RS}}, \alpha_{\text{1-RSB}}$) and the transition points to PSB/RSB ($\alpha_{\mathrm{PS}}, \alpha_{\mathrm{AT}}$) across different activation functions, appropriate choices of activation functions depending on the task and computational resources can be made. Second, it aids in the design of architectures with appropriate parameter sizes. Understanding the impact of parameter sizes on the structure of the solution space allows for determining the necessary α to train models in the over-parameterized regime or the RS/PS regions.

B. Perspectives

Gardner's framework of storage capacity examines the uniform distribution on the solution space, which solutions obtained by learning algorithms like gradient descent do not generally follow. In particular, gradient descent is expected to perform some form of implicit regularization; that is, it converges to solutions with some desirable properties among the set of possible solutions, leading to solutions with high generalization performance when applied to over-parameterized neural networks [41]. Thus, the analysis of the dependency of solution space structure and storage capacity on learning algorithms is an important future task.

In addition, it is necessary to consider datasets with some structure, rather than random data, to understand practical machine learning. As an attempt to handle datasets with structure within the framework of statistical mechanics, Goldt et al. proposed the *hidden manifold model* [42]. It should be possible to calculate the storage capacity of FCM using datasets following the hidden manifold model.

Exploring the relationship between FCM's storage capacity and the correspondence with the neural network-Gaussian process (NN-GP) is also significant. Zavatone-Veth and Pehlevan pointed out that the effective order parameter q_{eff} characterizing the storage capacity of TCM formally coincides with the kernel of a Gaussian process corresponding to a two-layer neural network [25]. The storage capacity of FCMs, however, is determined
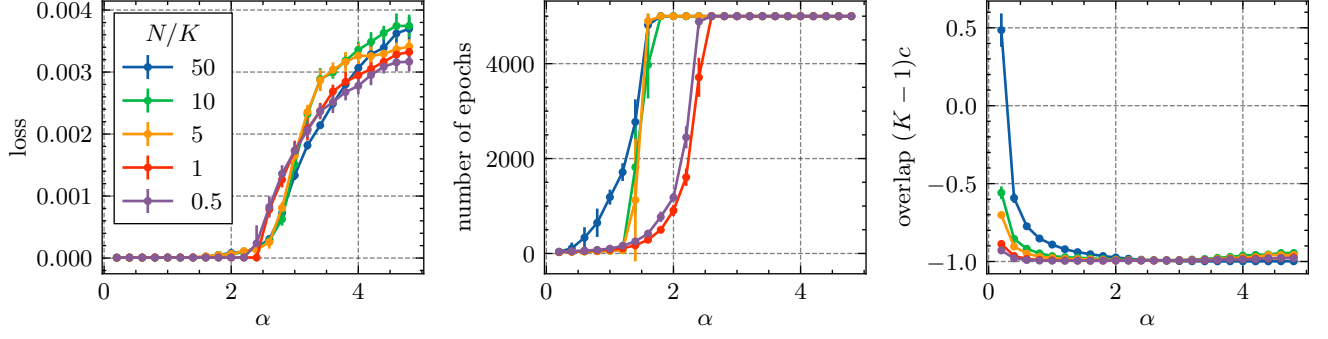
FIG. 4. Results for ReLU. We plot the value of the loss at the end of training (left), the number of epochs to the end of training (middle), and the overlap of the hidden weights (right). We observe a sudden increase in the loss at around $\alpha = 2.5$ and estimate it to be the experimental storage capacity of FCM with ReLU activations. The number of training epochs until termination increases rapidly as α grows. The overlap between weights c reaches $-1/(K-1)$ for sufficiently large α, consistent with the replica analysis.
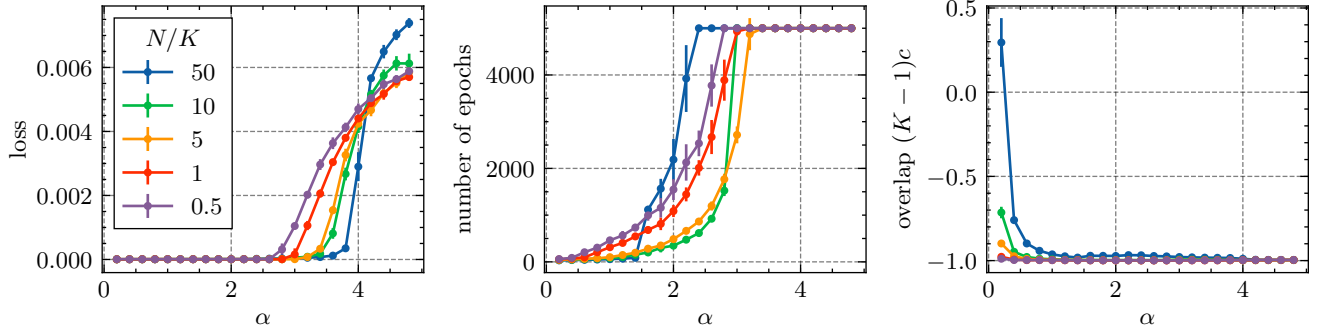

FIG. 5. Results for erf. The loss behaves differently for different input-hidden ratios N/K. For $N/K = 50$, the experimental storage capacity seems to be around 3.8 and this is well explained by our theory as stated in the main text. Since the replica analysis assumes infinite N/K, it fails to describe the behavior for small N/K. As with ReLU, the number of epochs until termination increases rapidly as α increases, and consistent with the theory, the order parameter c converges to $-1/(K-1)$.

from $\lambda = q_{\mathrm{eff}} - \langle g' \rangle^2 q$ in Equation 12, not just q_{eff}. One can argue a correspondence between λ and the kernel of a two-layer neural network with its weights initialized to have mutually negative correlations. The practical significance of such an initialization scheme awaits verification in future research.

Lastly, given that many practical neural networks are deep with three or more layers, identifying the structure of the solution space and storage capacity of DNNs is of high importance. In particular, the dependence of the storage capacity on the number of layers is one of the most interesting aspects. To this end, several results regarding the storage capacity of DNNs have been reported. Geiger et al. experimentally determined the storage capacity of DNNs trained on binary classification tasks using random and real datasets and found that the storage capacity does not depend on the number of layers [43]. In addition, Yoshino calculated the storage capacity of DNNs with all layers of width fixed at a constant value N using the replica method and found that the storage capacity increases exponentially with the number of lay-

ers [44]. We note that Yoshino's results should not apply to DNNs with a single output neuron; for the storage capacity per parameter α_c and the Vapnik-Chervonenkis dimension per parameter α_{VC} of a binary classifier, it is proved that $\alpha_c < 2\alpha_{\mathrm{VC}}$ holds [45] and that α_{VC} is of at most polynomial order of the number of layers [46]. Investigating Geiger et al.'s results about the storage capacity of DNNs with one output or a constant number of outputs theoretically is an important future task. We expect that Yoshino's method as well as a series of studies on the storage capacity of two-layer neural networks including this study will be useful.

---

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

[2] J. Berner, P. Grohs, G. Kutyniok, and P. Petersen, The modern mathematics of deep learning, arXiv preprint arXiv:2105.04026 , 86 (2021).

[3] L. Zdeborová, Understanding deep learning is also a job for physicists, Nature Physics **16**, 602 (2020).

[4] P. Auer, M. Herbster, and M. K. Warmuth, Exponentially many local minima for single neurons, Advances in neural information processing systems **8** (1995).

[5] I. Safran and O. Shamir, Spurious local minima are common in two-layer relu neural networks, in *International conference on machine learning* (PMLR, 2018) pp. 4433–4441.

[6] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, The loss surfaces of multilayer networks, in *Artificial intelligence and statistics* (PMLR, 2015) pp. 192–204.

[7] C. D. Freeman and J. Bruna, Topology and geometry of half-rectified network optimization, arXiv preprint arXiv:1611.01540 (2016).

[8] C. Baldassi, E. M. Malatesta, and R. Zecchina, Properties of the geometry of solutions and capacity of multilayer neural networks with rectified linear unit activations, Physical review letters **123**, 170602 (2019).

[9] G. Cybenko, Approximation by superpositions of a sigmoidal function, Mathematics of control, signals and systems **2**, 303 (1989).

[10] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2 (Springer, 2009).

[11] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, Understanding deep learning (still) requires rethinking generalization, Communications of the ACM **64**, 107 (2021).

[12] M. Belkin, D. Hsu, S. Ma, and S. Mandal, Reconciling modern machine-learning practice and the classical bias–variance trade-off, Proceedings of the National Academy of Sciences **116**, 15849 (2019).

[13] 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, Journal of Statistical Mechanics: Theory and Experiment **2020**, 023401 (2020).

[14] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, Deep double descent: Where bigger models and more data hurt, Journal of Statistical Mechanics: Theory and Experiment **2021**, 124003 (2021).

[15] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, Vol. 9 (World Scientific Publishing Company, 1987).

[16] H. Nishimori, *Statistical physics of spin glasses and information processing: an introduction*, 111 (Clarendon Press, 2001).

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

[18] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, Statistical mechanics of deep learning, Annual Review of Condensed Matter Physics **11**, 501 (2020).

[19] E. Gardner and B. Derrida, Optimal storage properties of neural network models, Journal of Physics A: Mathematical and general **21**, 271 (1988).

[20] E. Gardner, The space of interactions in neural network models, Journal of physics A: Mathematical and general **21**, 257 (1988).

[21] A. Engel, H. Köhler, F. Tschepke, H. Vollmayr, and A. Zippelius, Storage capacity and learning algorithms for two-layer neural networks, Physical Review A **45**, 7590 (1992).

[22] E. Barkai, D. Hansel, and H. Sompolinsky, Broken symmetries in multilayered perceptrons, Physical Review A **45**, 4146 (1992).

[23] R. Monasson and R. Zecchina, Weight space structure and internal representations: a direct approach to learning and generalization in multilayer neural networks, Physical review letters **75**, 2432 (1995).

[24] Y. Xiong, C. Kwon, and J.-H. Oh, The storage capacity of a fully-connected committee machine, Advances in Neural Information Processing Systems **10** (1997).

[25] J. A. Zavatone-Veth and C. Pehlevan, Activation function dependence of the storage capacity of treelike neural networks, Physical Review E **103**, L020301 (2021).

[26] H. S. Seung, H. Sompolinsky, and N. Tishby, Statistical mechanics of learning from examples, Physical review A **45**, 6056 (1992).

[27] T. L. Watkin, A. Rau, and M. Biehl, The statistical mechanics of learning a rule, Reviews of Modern Physics **65**, 499 (1993).

[28] M. Biehl, E. Schlösser, and M. Ahr, Phase transitions in soft-committee machines, Europhysics Letters **44**, 261 (1998).

[29] B. Aubin, A. Maillard, F. Krzakala, N. Macris, L. Zdeborová, *et al.*, The committee machine: Computational to statistical gaps in learning a two-layers neural network, Advances in Neural Information Processing Systems **31** (2018).

[30] E. Oostwal, M. Straat, and M. Biehl, Hidden unit specialization in layered neural networks: Relu vs. sigmoidal activation, Physica A: Statistical Mechanics and its Applications **564**, 125517 (2021).

[31] M. Stojnic, Capacity of the treelike sign perceptrons neural networks with one hidden layer–rdt based upper bounds, arXiv preprint arXiv:2312.08244 (2023).

[32] M. Stojnic, \emph {Lifted} rdt based capacity analysis of the 1-hidden layer treelike\emph {sign} perceptrons neural networks, arXiv preprint arXiv:2312.08257 (2023).

[33] M. Stojnic, Exact capacity of the\emph {wide} hidden layer treelike neural networks with generic activations, arXiv preprint arXiv:2402.05719 (2024).

[34] M. Stojnic, Fixed width treelike neural networks capacity analysis–generic activations, arXiv preprint

arXiv:2402.05696 (2024).

[35] See Supplemental Material for details of the calculation.

[36] J. R. de Almeida and D. J. Thouless, Stability of the sherrington-kirkpatrick solution of a spin glass model, Journal of Physics A: Mathematical and General **11**, 983 (1978).

[37] A. Blum and R. Rivest, Training a 3-node neural network is np-complete, Advances in neural information processing systems **1** (1988).

[38] J. Šíma, Training a single sigmoidal neuron is hard, Neural computation **14**, 2709 (2002).

[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems **32** (2019).

[40] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[41] B. Neyshabur, R. Tomioka, and N. Srebro, In search of the real inductive bias: On the role of implicit regularization in deep learning, arXiv preprint arXiv:1412.6614 (2014).

[42] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, Modeling the influence of data structure on learning in neural networks: The hidden manifold model, Physical Review X **10**, 041044 (2020).

[43] M. Geiger, S. Spigler, S. d'Ascoli, L. Sagun, M. Baity-Jesi, G. Biroli, and M. Wyart, Jamming transition as a paradigm to understand the loss landscape of deep neural networks, Physical Review E **100**, 012115 (2019).

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

[45] A. Engel, *Statistical mechanics of learning* (Cambridge University Press, 2001).

[46] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian, Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks, Journal of Machine Learning Research **20**, 1 (2019).

Supplemental Material for "Solution space and storage capacity of fully connected two-layer neural networks with generic activation functions"

Sota Nishiyama[1, *] and Masayuki Ohzeki[2, 3, 4, 5]

[1] *School of Engineering, Tohoku University, Sendai, 980-8579, Japan*
[2] *Graduate School of Information Sciences, Tohoku University, Sendai, 980-8579, Japan*
[3] *Department of Physics, Tokyo Institute of Technology, Tokyo, 152-8551, Japan*
[4] *International Research Frontier Initiative, Tokyo Institute of Technology, Tokyo, 108-0023*
[5] *Sigma-i Co., Ltd., Tokyo, 108-0075, Japan*

(Dated: April 23, 2024)

I. DETAILS OF REPLICA CALCULATION

A. Derivation of the Quenched Free Entropy

In this subsection, we calculate the replicated partition function

$$[Z^n] = \int \prod_{k=1}^{K} \prod_{a=1}^{n} \mathrm{d}\boldsymbol{w}_k^a \delta(N - \boldsymbol{w}_k^a \cdot \boldsymbol{w}_k^a) \times \left[\prod_{a=1}^{n} \Theta(ys(\boldsymbol{x}; \{\boldsymbol{w}_k^a\}, \{v_k\}) - \kappa) \right]_{\boldsymbol{x}, y}^{P}. \tag{1}$$

Let us define the local fields as $h_k^a = v_k \boldsymbol{w}_k^a \cdot \boldsymbol{x}$. Since the components of $\boldsymbol{x}$ are i.i.d. normal random variables, their weighted sums $\{h_k^a\}_{k=1,2,\ldots,K}^{a=1,2,\ldots,n}$ follow a multivariate normal distribution. Its mean and covariance are given by

$$\mathbb{E}[h_k^a] = 0, \tag{2}$$

$$\mathrm{Cov}[h_k^a, h_l^b] = \frac{1}{N} v_k v_l \boldsymbol{w}_k^a \cdot \boldsymbol{w}_l^b. \tag{3}$$

Introducing the order parameters as defined in the main text, the replicated partition function is written as follows:

$$
\begin{aligned}
[Z^n] &\propto \int \prod_{ka} (\mathrm{d}\boldsymbol{w}_k^a \delta(N - \boldsymbol{w}_k^a \cdot \boldsymbol{w}_k^a)) \prod_{a<b} \left(\mathrm{d}q^{ab} \prod_k \delta(Nq^{ab} - \boldsymbol{w}_k^a \cdot \boldsymbol{w}_k^b) \right) \\
&\quad \times \prod_a \left(\mathrm{d}c^a \prod_{k<l} \delta(Nc^a - v_k v_l \boldsymbol{w}_k^a \cdot \boldsymbol{w}_l^a) \right) \prod_{a<b} \left(\mathrm{d}d^{ab} \prod_{k<l} \delta(Nd^{ab} - v_k v_l \boldsymbol{w}_k^a \cdot \boldsymbol{w}_l^b) \right) \\
&\quad \times \left[\prod_a \Theta\left(y \frac{1}{\sqrt{K}} \sum_k v_k g(v_k h_k^a) - \kappa \right) \right]_{h,y}^{P} \\
&= \int \prod_{a<b} \mathrm{d}q^{ab} \prod_a \mathrm{d}c^a \prod_{a<b} \mathrm{d}d^{ab} \left[\prod_a \Theta\left(y \frac{1}{\sqrt{K}} \sum_k v_k g(v_k h_k^a) - \kappa \right) \right]_{h,y}^{P} \\
&\quad \times \left(\int \prod_{ka} \mathrm{d}\boldsymbol{w}_k^a \prod_{ka} \delta(N - \boldsymbol{w}_k^a \cdot \boldsymbol{w}_k^a) \prod_{a<b,k} \delta(Nq^{ab} - \boldsymbol{w}_k^a \cdot \boldsymbol{w}_k^b) \right. \\
&\quad \left. \times \prod_{a,k<l} \delta(Nc^a - v_k v_l \boldsymbol{w}_k^a \cdot \boldsymbol{w}_l^a) \prod_{a<b,k<l} \delta(Nd^{ab} - v_k v_l \boldsymbol{w}_k^a \cdot \boldsymbol{w}_l^b) \right).
\end{aligned}
$$

$$\tag{4}$$
$$\tag{5}$$

Let us define the energy term as

$$G_e(Q) \coloneqq \frac{1}{n} \log \left[\prod_a \Theta\left(y \frac{1}{\sqrt{K}} \sum_k v_k g(v_k h_k^a) - \kappa \right) \right]_{h,y}. \tag{6}$$

---

* snishiyama@g.ecc.u-tokyo.ac.jp

Then,

$$\left[\prod_a \Theta\left(y\frac{1}{\sqrt{K}}\sum_k v_k g(v_k h_k^a) - \kappa\right)\right]_{h,y}^P = \exp(nPG_e(Q)) = \exp(\alpha nNKG_e(Q)). \tag{7}$$

Let us evaluate the integral in the second line of Equation 5. Using the integral representation of the delta function

$$\delta(x) = \int_{-\infty}^{\infty} \frac{\mathrm{d}k}{2\pi} \exp(ikx) \tag{8}$$

and the Gaussian integral, in the $n \to 0$ limit, the integral is calculated as

$$\int \mathrm{d}\hat{Q} \exp\left(\frac{N}{2}\left(\mathrm{tr}(Q\hat{Q}) - \log\det\hat{Q}\right)\right). \tag{9}$$

using the conjugate order parameter $\hat{Q}$ introduced in the main text. Let us define the entropy term as

$$G_s(Q, \hat{Q}) := \frac{1}{2nK}\left(\mathrm{tr}(Q\hat{Q}) - \log\det\hat{Q}\right). \tag{10}$$

Thus, Equation 5 is

$$[Z^n] \propto \int \mathrm{d}Q\,\mathrm{d}\hat{Q} \exp\left(nNK\left(\alpha G_e(Q) + G_s(Q,\hat{Q})\right)\right). \tag{11}$$

In the $N \to \infty$ limit, applying the saddle-point method gives

$$[Z^n] \propto \exp\left(nNK \operatorname*{extr}_{Q,\hat{Q}}\left\{\alpha G_e(Q) + G_s(Q,\hat{Q})\right\}\right). \tag{12}$$

The quenched free entropy is

$$f = \operatorname*{extr}_{Q,\hat{Q}}\left\{G_s(Q,\hat{Q}) + \alpha G_e(Q)\right\}. \tag{13}$$

B.   The Free Entropy and the Storage Capacity under RS Ansatz

From now on, we assume c, d are of $O(1/K)$. It will be shown that this assumption is self-consistent. In addition, we define the rescaled order parameters $\bar{c}, \bar{d}$ as

$$\bar{c} := (K-1)c, \quad \bar{d} := (K-1)d. \tag{14}$$

Let us evaluate the entropy term $G_s(Q, \hat{Q})$. Using the formula

$$\det \underbrace{\begin{pmatrix} A & B & \cdots & B \\ B & A & & \vdots \\ \vdots & & \ddots & B \\ B & \cdots & B & A \end{pmatrix}}_{N} = \det(A + (N-1)B)(\det(A - B))^{N-1}, \tag{15}$$

we obtain

$$\det \hat{Q} = \left(1 - n\frac{\hat{q} + (K-1)\hat{d}}{\hat{E} + \hat{q} + (K-1)(\hat{c} + \hat{d})}\right) \times \left(\hat{E} + \hat{q} + (K-1)(\hat{c} + \hat{d})\right)^n$$
$$\times \left(1 - n\frac{\hat{q} - \hat{d}}{\hat{E} + \hat{q} - \hat{c} - \hat{d}}\right)^{K-1} \times \left(\hat{E} + \hat{q} - \hat{c} - \hat{d}\right)^{(K-1)n} \tag{16}$$

$$G_s(Q, \hat{Q}) = \frac{1}{2nK}\left(\text{tr}(Q\hat{Q}) - \log \det \hat{Q}\right) \tag{17}$$

$$\xrightarrow{n \to 0} \frac{1}{2}\left(\hat{E} + q\hat{q} + (K-1)(c\hat{c} + d\hat{d})\right)$$
$$- \frac{1}{2K}\left(\log(\hat{E} + \hat{q} + (K-1)(\hat{c} + \hat{d})) - \frac{\hat{q} + (K-1)\hat{d}}{\hat{E} + \hat{q} + (K-1)(\hat{c} + \hat{d})}\right) \tag{18}$$
$$- \frac{K-1}{2K}\left(\log(\hat{E} + \hat{q} - \hat{c} - \hat{d}) - \frac{\hat{q} - \hat{d}}{\hat{E} + \hat{q} - \hat{c} - \hat{d}}\right)$$

$$\xrightarrow{K \to \infty} \frac{1}{2}\left(\hat{E} + q\hat{q} + \bar{c}\hat{c} + \bar{d}\hat{d} - \left(\log(\hat{E} + \hat{q} - \hat{c} - \hat{d}) - \frac{\hat{q} - \hat{d}}{\hat{E} + \hat{q} - \hat{c} - \hat{d}}\right)\right). \tag{19}$$

Next, let us evaluate the energy term $G_e(Q)$. Let

$$s^a := \frac{1}{\sqrt{K}} \sum_k v_k g(v_k h_k^a). \tag{20}$$

Although s^a is a sum of a large number of random variables, $\{h_k\}_{k=1,\dots,K}$ have covariance c and thus are not independent. Therefore, the central limit theorem is not directly applicable. However, the covariance c is small (of $O(1/K)$), and one can still approximate s^a with a normal random variable. Therefore, $\{s^a\}_{a=1,\dots,n}$ follow a multivariate normal distribution with mean and covariance given by:

$$\mathbb{E}[s^a] = 0 \tag{21}$$

$$\text{Cov}[s^a, s^b] = \frac{1}{K} \sum_{kl} v_k v_l \text{Cov}[g(v_k h_k^a), g(v_l h_l^b)] \tag{22}$$

$$= \text{Cov}[g(h_k^a), g(h_k^b)] + (K/2 - 1)\text{Cov}[g(h_k^a), g(h_l^b)] - (K/2)\text{Cov}[g(h_k^a), g(-h_l^b)]. \tag{23}$$

Now, let us define the effective order parameters as follows:

$$\sigma^2 := \text{Var}[g(x) \mid x \sim \mathcal{N}(0,1)] \tag{24}$$

$$q_{\text{eff}} := \text{Cov}\left[g(x), g(y) \,\middle|\, \begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & q \\ q & 1 \end{pmatrix}\right)\right] \tag{25}$$

$$c_{\text{eff}}^{\pm} := \text{Cov}\left[g(x), g(y) \,\middle|\, \begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \pm c \\ \pm c & 1 \end{pmatrix}\right)\right] \tag{26}$$

$$d_{\text{eff}}^{\pm} := \text{Cov}\left[g(x), g(y) \,\middle|\, \begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \pm d \\ \pm d & 1 \end{pmatrix}\right)\right]. \tag{27}$$

Using the assumption that c, d are of $O(1/K)$ and expanding g to the first order, we obtain

$$c_{\text{eff}}^{\pm} = \pm\langle g'\rangle^2 c + O(1/K^2), \tag{28}$$
$$d_{\text{eff}}^{\pm} = \pm\langle g'\rangle^2 d + O(1/K^2). \tag{29}$$

Therefore,

$$\text{Cov}[s^a, s^b] = \delta_{ab}\left(\sigma^2 + \langle g'\rangle^2 \bar{c}\right) + (1 - \delta_{ab})\left(q_{\text{eff}} + \langle g'\rangle^2 \bar{d}\right) + O(1/K). \tag{30}$$

Introducing λ, μ as

$$\lambda := q_{\text{eff}} + \langle g'\rangle^2 \bar{d}, \tag{31}$$
$$\mu := \sigma^2 - q_{\text{eff}} + \langle g'\rangle^2 (\bar{c} - \bar{d}), \tag{32}$$

s^a can be written as follows:

$$s^a = \sqrt{\lambda}z + \sqrt{\mu}x^a, \tag{33}$$

where $x^1, \ldots, x^n, z \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$.

Gathering these results, the energy term is

$$G_e(Q) = \frac{1}{n} \log \left[\prod_a \Theta(ys^a - \kappa) \right]_{s,y} \tag{34}$$

$$= \frac{1}{n} \log \left[\int Dz \left(\int Dx \Theta(y(\sqrt{\lambda}z + \sqrt{\mu}x) - \kappa) \right)^n \right]_y \tag{35}$$

$$= \frac{1}{n} \log \int Dz \left(H\left(\frac{\kappa + \sqrt{\lambda}z}{\sqrt{\mu}} \right) \right)^n \tag{36}$$

$$\xrightarrow{n \to 0} \int Dz \log H\left(\frac{\kappa + \sqrt{\lambda}z}{\sqrt{\mu}} \right). \tag{37}$$

Thus we obtain the RS free entropy as follows:

$$f_{\text{RS}} = \underset{q, \bar{c}, \bar{d}, \hat{E}, \hat{q}, \hat{c}, \hat{d}}{\text{extr}} \left\{ \frac{1}{2} \left(\hat{E} + q\hat{q} + \bar{c}\hat{c} + \bar{d}\hat{d} + \frac{\hat{q} - \hat{d}}{\hat{E} + \hat{q} - \hat{c} - \hat{d}} - \log(\hat{E} + \hat{q} - \hat{c} - \hat{d}) \right) \right.$$
$$\left. + \alpha \int Dz \log H\left(\frac{\kappa + \sqrt{\lambda}z}{\sqrt{\mu}} \right) \right\}. \tag{38}$$

To obtain the saddle-point equations, we differentiate the free entropy (38) with each order parameter. First, let us derive the equation for q. Define

$$\mathcal{Y} = H\left(\frac{\kappa + \sqrt{\lambda}z}{\sqrt{\mu}} \right) = \int Dx \Theta\left(\kappa + \sqrt{\lambda}z + \sqrt{\mu}x \right). \tag{39}$$

Then,

$$\frac{\partial f_{\text{RS}}}{\partial q} = \frac{1}{2}\hat{q} + \alpha \int Dz \frac{\partial \mathcal{Y}}{\partial q} \tag{40}$$

$$= \frac{1}{2}\hat{q} + \alpha \int Dz \frac{1}{\int Dx \Theta} \int Dx \Theta' \cdot \left(\frac{z}{2\sqrt{\lambda}} - \frac{x}{2\sqrt{\mu}} \right) q'_{\text{eff}} \tag{41}$$

$$= \frac{1}{2}\hat{q} + \alpha q'_{\text{eff}} \int Dz \left(\frac{z}{2\sqrt{\lambda}} \frac{\partial \mathcal{Y}}{\partial(\sqrt{\lambda}z)} - \frac{1}{2} \left(\frac{\partial^2 \mathcal{Y}}{\partial(\sqrt{\lambda}z)^2} + \left(\frac{\partial \mathcal{Y}}{\partial(\sqrt{\lambda}z)} \right)^2 \right) \right) \tag{42}$$

$$= \frac{1}{2}\hat{q} - \frac{1}{2}\alpha q'_{\text{eff}} \int Dz \left(\frac{\partial \mathcal{Y}}{\partial(\sqrt{\lambda}z)} \right)^2, \tag{43}$$

where $\frac{\partial}{\partial(\sqrt{\lambda}z)}$ means $\frac{1}{\sqrt{\lambda}} \frac{\partial}{\partial z}$. Thus, the saddle-point equation is

$$\hat{q} = \alpha q'_{\text{eff}} \int Dz \left(\frac{\partial \mathcal{Y}}{\partial(\sqrt{\lambda}z)} \right)^2 = \alpha \frac{q'_{\text{eff}}}{\mu} \int Dz \left(\frac{H'(u)}{H(u)} \right)^2, \tag{44}$$

where $u = \frac{\kappa + \sqrt{\lambda}z}{\sqrt{\mu}}$.

Similar calculation for $\bar{c}, \bar{d}$ gives

$$\hat{c} = \alpha \frac{\langle g' \rangle^2}{\mu} \int Dz \frac{H'(u)}{H(u)} u, \tag{45}$$

$$\hat{d} = \alpha \frac{\langle g' \rangle^2}{\mu} \int Dz \left(\frac{H'(u)}{H(u)} \right)^2. \tag{46}$$

The derivative with respect to $\hat{E}$ is

$$\frac{\partial f_{\text{RS}}}{\partial \hat{E}} = \frac{1}{2} - \frac{1}{2}\frac{\hat{E} - \hat{c} + 2(\hat{q} - \hat{d})}{(\hat{E} + \hat{q} - \hat{c} - \hat{d})^2}. \tag{47}$$

Solving this for $\hat{E} + \hat{q} - \hat{c} - \hat{d}$ gives

$$\hat{E} + \hat{q} - \hat{c} - \hat{d} = \frac{1 + \sqrt{4(\hat{q} - \hat{d}) + 1}}{2}. \tag{48}$$

Differentiating with respect to $\hat{q}, \hat{c}, \hat{d}$ gives

$$q = \frac{\hat{q} - \hat{d}}{(\hat{E} + \hat{q} - \hat{c} - \hat{d})^2} \tag{49}$$

$$\bar{c} = -\frac{\hat{E} - \hat{c} + 2(\hat{q} - \hat{d})}{(\hat{E} + \hat{q} - \hat{c} - \hat{d})^2} = -1 \tag{50}$$

$$\bar{d} = -\frac{\hat{q} - \hat{d}}{(\hat{E} + \hat{q} - \hat{c} - \hat{d})^2} = -q. \tag{51}$$

Next, let us evaluate the RS storage capacity α_{RS}. Let $q = 1 - \varepsilon$ and consider the limit $\varepsilon \to 0$. By the weak differentiability and the square-integrability of the weak derivative with respect to the Gaussian measure, the effective order parameter can be expanded as $q_{\text{eff}} = \sigma^2 - \langle g'^2 \rangle \varepsilon + O(\varepsilon^2)$ [1]. Thus,

$$\lambda \to \sigma^2 - \langle g' \rangle^2, \quad \mu = \left(\langle g'^2 \rangle - \langle g' \rangle^2 \right)\varepsilon + O(\varepsilon^2) =: \gamma\varepsilon + O(\varepsilon^2). \tag{52}$$

Also, for $u = \frac{\kappa + \sqrt{\lambda}z}{\sqrt{\mu}}$, we have $|u| \gg 1$ as $\varepsilon \to 0$. Using the asymptotic form of $H(x)$

$$H(x) = \begin{cases} \frac{1}{\sqrt{2\pi}x} \exp\left(-\frac{x^2}{2}\right)\left(1 + O(x^{-2})\right) & (x \gg 1) \\ 1 - \frac{1}{\sqrt{2\pi}x} \exp\left(-\frac{x^2}{2}\right)\left(1 + O(x^{-2})\right) & (x \ll -1) \end{cases} \tag{53}$$

and $H'(x) = -\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$, Equation 44 is

$$\hat{q} \to \alpha\frac{\langle g'^2 \rangle}{\mu} \int_{-\kappa/\sqrt{\lambda}}^{\infty} Dz \left(\frac{\kappa + \sqrt{\lambda}z}{\sqrt{\mu}}\right)^2 \tag{54}$$

$$= \alpha\frac{\langle g'^2 \rangle \lambda}{\gamma^2\varepsilon^2} \int_{-\kappa/\sqrt{\lambda}}^{\infty} \left(\frac{\kappa}{\sqrt{\lambda}} + z\right)^2. \tag{55}$$

Equation 46 is

$$\hat{d} \to \alpha\frac{\langle g' \rangle^2}{\mu} \int_{-\kappa/\sqrt{\lambda}}^{\infty} Dz \left(\frac{\kappa + \sqrt{\lambda}z}{\sqrt{\mu}}\right)^2 \tag{56}$$

$$= \alpha\frac{\langle g' \rangle^2 \lambda}{\gamma^2\varepsilon^2} \int_{-\kappa/\sqrt{\lambda}}^{\infty} \left(\frac{\kappa}{\sqrt{\lambda}} + z\right)^2. \tag{57}$$

Therefore, using $1 - q = \frac{1}{\hat{E} + \hat{q} - \hat{c} - \hat{d}}$, Equation 49 is written as follows:

$$q = \frac{\hat{q} - \hat{d}}{(\hat{E} + \hat{q} - \hat{c} - \hat{d})^2} \tag{58}$$

$$= \alpha\frac{\lambda}{\gamma} \int_{-\kappa/\sqrt{\lambda}}^{\infty} \left(\frac{\kappa}{\sqrt{\lambda}} + z\right)^2. \tag{59}$$

This gives the RS storage capacity as follows:

$$\frac{1}{\alpha_{\text{RS}}} = \frac{\sigma^2 - \langle g' \rangle^2}{\langle g'^2 \rangle - \langle g' \rangle^2} \int_{-\kappa/\sqrt{\sigma^2 - \langle g' \rangle^2}}^{\infty} Dz \left(\frac{\kappa}{\sqrt{\sigma^2 - \langle g' \rangle^2}} + z\right)^2. \tag{60}$$

C. The Free Entropy and the Storage Capacity under 1-RSB Ansatz

Under the 1-RSB ansatz, we parametrize the overlap matrix Q and its conjugate $\hat{Q}$ with twelve order parameters $m, q_1, q_0, c, d_1, d_0, \hat{E}, \hat{q}_1, \hat{q}_0, \hat{c}, \hat{d}_1, \hat{d}_0 \in \mathbb{R}$ as follows:

$$Q := \left.\begin{pmatrix} A & B & \cdots & B \\ B & A & & B \\ \vdots & & \ddots & \vdots \\ B & \cdots & B & A \end{pmatrix}\right\} K \qquad \hat{Q} := \left.\begin{pmatrix} \hat{A} & \hat{B} & \cdots & \hat{B} \\ \hat{B} & \hat{A} & & \hat{B} \\ \vdots & & \ddots & \vdots \\ \hat{B} & \cdots & \hat{B} & \hat{A} \end{pmatrix}\right\} K \tag{61}$$

$$\underbrace{\phantom{\begin{pmatrix} A & B & \cdots & B \end{pmatrix}}}_{K} \qquad \underbrace{\phantom{\begin{pmatrix} \hat{A} & \hat{B} & \cdots & \hat{B} \end{pmatrix}}}_{K}$$

$$A := \left.\begin{pmatrix} A_1 & A_0 & \cdots & A_0 \\ A_0 & A_1 & & A_0 \\ \vdots & & \ddots & \vdots \\ A_0 & \cdots & A_0 & A_1 \end{pmatrix}\right\} n/m \tag{62}$$

$$\underbrace{\phantom{\begin{pmatrix} A_1 & A_0 & \cdots & A_0 \end{pmatrix}}}_{n/m}$$

$$A_1 = q_1 \mathbf{1}_m \mathbf{1}_m^\top + (1-q_1) I_m, \quad A_0 = q_0 \mathbf{1}_m \mathbf{1}_m^\top \tag{63}$$

$$B := \left.\begin{pmatrix} B_1 & B_0 & \cdots & B_0 \\ B_0 & B_1 & & B_0 \\ \vdots & & \ddots & \vdots \\ B_0 & \cdots & B_0 & B_1 \end{pmatrix}\right\} n/m \tag{64}$$

$$\underbrace{\phantom{\begin{pmatrix} B_1 & B_0 & \cdots & B_0 \end{pmatrix}}}_{n/m}$$

$$B_1 = d_1 \mathbf{1}_m \mathbf{1}_m^\top + (c-d_1) I_m, \quad B_0 = d_0 \mathbf{1}_m \mathbf{1}_m^\top \tag{65}$$

$$\hat{A} := \left.\begin{pmatrix} \hat{A}_1 & \hat{A}_0 & \cdots & \hat{A}_0 \\ \hat{A}_0 & \hat{A}_1 & & \hat{A}_0 \\ \vdots & & \ddots & \vdots \\ \hat{A}_0 & \cdots & \hat{A}_0 & \hat{A}_1 \end{pmatrix}\right\} n/m \tag{66}$$

$$\underbrace{\phantom{\begin{pmatrix} \hat{A}_1 & \hat{A}_0 & \cdots & \hat{A}_0 \end{pmatrix}}}_{n/m}$$

$$\hat{A}_1 = -\hat{q}_1 \mathbf{1}_m \mathbf{1}_m^\top + (\hat{E} + \hat{q}_1) I_m, \quad \hat{A}_0 = -\hat{q}_0 \mathbf{1}_m \mathbf{1}_m^\top \tag{67}$$

$$\hat{B} := \left.\begin{pmatrix} \hat{B}_1 & \hat{B}_0 & \cdots & \hat{B}_0 \\ \hat{B}_0 & \hat{B}_1 & & \hat{B}_0 \\ \vdots & & \ddots & \vdots \\ \hat{B}_0 & \cdots & \hat{B}_0 & \hat{B}_1 \end{pmatrix}\right\} n/m \tag{68}$$

$$\underbrace{\phantom{\begin{pmatrix} \hat{B}_1 & \hat{B}_0 & \cdots & \hat{B}_0 \end{pmatrix}}}_{n/m}$$

$$\hat{B}_1 = -\hat{d}_1 \mathbf{1}_m \mathbf{1}_m^\top + (\hat{c} + \hat{d}_1) I_m, \quad \hat{B}_0 = -\hat{d}_0 \mathbf{1}_m \mathbf{1}_m^\top \tag{69}$$

where $\mathbf{1}_m$ is the m-dimensional vector with all components set to one.

Again we self-consistently assume that c, d_1, d_0 are of $O(1/K)$, and define the rescaled order parameters as

$$\bar{c} := (K-1)c, \quad \bar{d}_1 := (K-1)d_1, \quad \bar{d}_0 := (K-1)d_0. \tag{70}$$

Let us evaluate the entropy term $G_s(Q, \hat{Q})$.

$$G_s \xrightarrow{n \to 0} \frac{1}{2}\Big(\hat{E} + (1-m)q_1\hat{q}_1 + mq_0\hat{q}_0 + (K-1)(c\hat{c} + (1-m)d_1\hat{d}_1 + md_0\hat{d}_0)\Big)$$

$$- \frac{1}{2K}\Bigg(\frac{1}{m}\log\Big(\hat{E} + (1-m)\hat{q}_1 + m\hat{q}_0 + (K-1)(\hat{c} + (1-m)\hat{d}_1 + m\hat{d}_0)\Big)$$

$$+ \Big(1 - \frac{1}{m}\Big)\log\Big(\hat{E} + \hat{q}_1 + (K-1)(c + \hat{d}_1)\Big)$$

$$- \frac{\hat{q}_0 + (K-1)\hat{d}_0}{\hat{E} + (1-m)\hat{q}_1 + m\hat{q}_0 + (K-1)(\hat{c} + (1-m)\hat{d}_1 + m\hat{d}_0)}\Bigg) \tag{71}$$

$$- \frac{K-1}{2K}\Bigg(\frac{1}{m}\log\Big(\hat{E} + (1-m)\hat{q}_1 + m\hat{q}_0 - \hat{c} - (1-m)\hat{d}_1 - m\hat{d}_0\Big) + \Big(1 - \frac{1}{m}\Big)\log\Big(\hat{E} + \hat{q}_1 - \hat{c} - \hat{d}_1\Big)$$

$$- \frac{\hat{q}_0 - \hat{d}_0}{\hat{E} + (1-m)\hat{q}_1 + m\hat{q}_0 - \hat{c} - (1-m)\hat{d}_1 - m\hat{d}_0}\Bigg)$$

$$\xrightarrow{K \to \infty} \frac{1}{2}\Big(\hat{E} + (1-m)q_1\hat{q}_1 + mq_0\hat{q}_0 + \bar{c}\hat{c} + (1-m)\bar{d}_1\hat{d}_1 + m\bar{d}_0\hat{d}_0\Big)$$

$$- \frac{1}{2}\Bigg(\frac{1}{m}\log\Big(\hat{E} + (1-m)\hat{q}_1 + m\hat{q}_0 - \hat{c} - (1-m)\hat{d}_1 - m\hat{d}_0\Big) + \Big(1 - \frac{1}{m}\Big)\log\Big(\hat{E} + \hat{q}_1 - \hat{c} - \hat{d}_1\Big) \tag{72}$$

$$- \frac{\hat{q}_0 - \hat{d}_0}{\hat{E} + (1-m)\hat{q}_1 + m\hat{q}_0 - \hat{c} - (1-m)\hat{d}_1 - m\hat{d}_0}\Bigg).$$

Let us evaluate the energy term $G_e(Q)$. Calculation along the same line as the RS ansatz, we obtain

$$G_e \xrightarrow{n \to 0} \frac{1}{m}\int Dz_0 \log \int Dz_1 \Bigg(H\Big(\frac{\kappa + \sqrt{\lambda_0}z_0 + \sqrt{\lambda_1}z_1}{\sqrt{\mu}}\Big)\Bigg)^m, \tag{73}$$

where

$$\lambda_0 = q_{0,\text{eff}} + \langle g'\rangle^2 \bar{d}_0 \tag{74}$$

$$\lambda_1 = q_{1,\text{eff}} - q_{0,\text{eff}} + \langle g'\rangle^2(\bar{d}_1 - \bar{d}_0) \tag{75}$$

$$\mu = \sigma^2 - q_{1,\text{eff}} + \langle g'\rangle^2(\bar{c} - \bar{d}_1). \tag{76}$$

Thus, the 1-RSB free entropy is given as follows:

$$f_{\text{1-RSB}} = \operatorname*{extr}_{\substack{m, q_1, q_0, \bar{c}, \bar{d}_1, \bar{d}_0, \\ \hat{E}, \hat{q}_1, \hat{q}_0, \hat{c}, \hat{d}_1, \hat{d}_0}} \Bigg\{\frac{1}{2}\Big(\hat{E} + (1-m)q_1\hat{q}_1 + mq_0\hat{q}_0 + \bar{c}\hat{c} + (1-m)\bar{d}_1\hat{d}_1 + m\bar{d}_0\hat{d}_0\Big)$$

$$- \frac{1}{2}\Bigg(\frac{1}{m}\log\Big(\hat{E} + (1-m)\hat{q}_1 + m\hat{q}_0 - \hat{c} - (1-m)\hat{d}_1 - m\hat{d}_0\Big)$$

$$+ \Big(1 - \frac{1}{m}\Big)\log\Big(\hat{E} + \hat{q}_1 - \hat{c} - \hat{d}_1\Big) - \frac{\hat{q}_0 - \hat{d}_0}{\hat{E} + (1-m)\hat{q}_1 + m\hat{q}_0 - \hat{c} - (1-m)\hat{d}_1 - m\hat{d}_0}\Bigg)$$

$$+ \frac{\alpha}{m}\int Dz_0 \log \int Dz_1 \Bigg(H\Big(\frac{\kappa + \sqrt{\lambda_0}z_0 + \sqrt{\lambda_1}z_1}{\sqrt{\mu}}\Big)\Bigg)^m\Bigg\}. \tag{77}$$

The saddle-point equations for $\bar{c}, \bar{d}_1, \bar{d}_0$ give $\bar{c} = -1, \bar{d}_1 = -q_1, \bar{d}_0 = -q_0$. Eliminating the order parameter

$\hat{E}, \hat{q}_1, \hat{q}_0, \hat{c}, \hat{d}_1, \hat{d}_0$ gives

$$f_{\text{1-RSB}} = \underset{m,q_1,q_0}{\text{extr}} \left\{ \frac{1}{2} \left(\frac{1}{m} \log(1 - (1-m)q_1 - mq_0) + \left(1 - \frac{1}{m}\right) \log(1 - q_1) \right. \right.$$

$$\left. \left. + \frac{q_0}{1 - (1-m)q_1 - mq_0} \right) + \frac{\alpha}{m} \int Dz_0 \log \int Dz_1 \left(H\left(\frac{\kappa + \sqrt{\lambda_0}z_0 + \sqrt{\lambda_1}z_1}{\sqrt{\mu}} \right) \right)^m \right\}, \quad (78)$$

where

$$\lambda_0 = q_{0,\text{eff}} - \langle g' \rangle^2 q_0 \tag{79}$$

$$\lambda_1 = q_{1,\text{eff}} - q_{0,\text{eff}} - \langle g' \rangle^2 (q_1 - q_0) \tag{80}$$

$$\mu = \sigma^2 - q_{1,\text{eff}} - \langle g' \rangle^2 (1 - q_1). \tag{81}$$

Next, let us calculate the 1-RSB storage capacity $\alpha_{\text{1-RSB}}$. As $\alpha \to \alpha_{\text{1-RSB}}$, we assume $q_1 \to 1$ and $m \to 0$. Here, we assume that $r := m/(1-q_1)$ approaches a finite value. Let $q_1 = 1 - \varepsilon$ and consider the limit $\varepsilon \to 0$.

Here, we will follow the method of [1, 2]. For the 1-RSB free entropy

$$f_{\text{1-RSB}} = \underset{r,\varepsilon,q_0}{\text{extr}} \frac{1}{\varepsilon} \{ \varepsilon G_s + \alpha \varepsilon G_e \} \tag{82}$$

to take a finite value as $\varepsilon \to 0$, we need

$$\min_{r,q_0} \lim_{\varepsilon \to 0} \{ \varepsilon G_s + \alpha_{\text{1-RSB}} \varepsilon G_e \} = 0. \tag{83}$$

Thus, it follows that

$$\alpha_{\text{1-RSB}} = \min_{r,q_0} \left\{ -\frac{\lim_{\varepsilon \to 0} \varepsilon G_s}{\lim_{\varepsilon \to 0} \varepsilon G_e} \right\}. \tag{84}$$

For the entropy term,

$$\lim_{\varepsilon \to 0} \varepsilon G_s = \lim_{\varepsilon \to 0} \frac{1}{2} \left(\frac{q_0}{1 + r(1 - q_0 - \varepsilon)} + \frac{1}{r} \log(\varepsilon(1 + r(1 - q_0 - \varepsilon))) + \frac{r\varepsilon - 1}{r} \log \varepsilon \right) \tag{85}$$

$$= \frac{1}{2} \left(\frac{q_0}{1 + r(1 - q_0)} + \frac{1}{r} \log(1 + r(1 - q_0)) \right). \tag{86}$$

For the energy term, defining $\lambda_1 \to \sigma^2 - q_{0,\text{eff}} - \langle g' \rangle^2 (1 - q_0), \mu \to (\langle g'^2 \rangle - \langle g' \rangle^2)\varepsilon =: \gamma\varepsilon$,

$$\lim_{\varepsilon \to 0} \varepsilon G_e = \lim_{\varepsilon \to 0} \frac{1}{r} \int Dz_0 \log \left(\int_{-\infty}^{-Q} Dz_1 \right.$$

$$\left. + \int_{-Q}^{\infty} Dz_1 \left(\frac{1}{\sqrt{2\pi}} \frac{\sqrt{\mu}}{\kappa + \sqrt{\lambda_0}z_0 + \sqrt{\lambda_1}z_1} \exp\left(-\frac{1}{2} \left(\frac{\kappa + \sqrt{\lambda_0}z_0 + \sqrt{\lambda_1}z_1}{\sqrt{\mu}} \right)^2 \right) \right)^{r\varepsilon} \right) \tag{87}$$

$$= \frac{1}{r} \int Dz_0 \log \left(\int_{-\infty}^{-Q} Dz_1 + \int_{-Q}^{\infty} Dz_1 \exp\left(-\frac{r(\kappa + \sqrt{\lambda_0}z_0 + \sqrt{\lambda_1}z_1)^2}{2\gamma} \right) \right) \tag{88}$$

$$= \frac{1}{r} \int Dz_0 \log \left(H(Q) + R \exp\left(-\frac{r(\kappa + \sqrt{\lambda_0}z_0)^2}{2\gamma + r\lambda_1} \right) H(-QR) \right) \tag{89}$$

where

$$Q := \frac{\kappa + \sqrt{\lambda_0}z_0}{\sqrt{\lambda_1}}, \quad R := \sqrt{\frac{\gamma}{\gamma + r\lambda_1}}. \tag{90}$$

Let

$$\psi(r, q_0; \kappa) = -\int Dz_0 \log(H(Q) + R \exp\left(-\frac{r(\kappa + \sqrt{\lambda_0}z_0)^2}{2\gamma + r\lambda_1} \right) H(-QR)) \tag{91}$$

and we obtain

$$\alpha_{\text{1-RSB}} = \min_{r,q_0} \left\{ \frac{1}{2\psi(r, q_0; \kappa)} \left(\frac{rq_0}{1 + r(1 - q_0)} + \log(1 + r(1 - q_0)) \right) \right\}. \tag{92}$$

D. Stability of the RS Solution

Let $\Delta q := q_1 - q_0$, $\Delta \hat{q} := \hat{q}_1 - \hat{q}_0$, $\Delta \bar{d} := \bar{d}_1 - \bar{d}_0$, $\Delta \hat{d} := \hat{d}_1 - \hat{d}_0$. Our goal is to obtain the condition for $(\Delta q, \Delta \hat{q}, \Delta \bar{d}, \Delta \hat{d}) = (0, 0, 0, 0)$ to be a stable solution of the 1-RSB saddle-point equations.

The extremum condition of the 1-RSB free entropy (77) with respect to $\hat{q}_0, \hat{q}_1, \hat{d}_0, \hat{d}_1$ gives

$$q_0 = \frac{\hat{q}_0 - \hat{d}_0}{(\hat{E} + (1-m)\hat{q}_1 + m\hat{q}_0 - \hat{c} - (1-m)\hat{d}_1 - m\hat{d}_0)^2} \tag{93}$$

$$q_1 = \frac{1}{m}\left(\frac{1}{\hat{E} + (1-m)\hat{q}_1 + m\hat{q}_0 - \hat{c} - (1-m)\hat{d}_1 - m\hat{d}_0} - \frac{1}{\hat{E} + \hat{q}_1 - \hat{c} - \hat{d}_1} \right)$$
$$+ \frac{\hat{q}_0 - \hat{d}_0}{(\hat{E} + (1-m)\hat{q}_1 + m\hat{q}_0 - \hat{c} - (1-m)\hat{d}_1 - m\hat{d}_0)^2} \tag{94}$$

$$\bar{d}_0 = -q_0, \quad \bar{d}_1 = -q_1. \tag{95}$$

When $\Delta \hat{q}, \Delta \hat{d} \ll 1$,

$$\Delta q \approx \frac{1}{(\hat{E} + \hat{q}_1 - \hat{c} - \hat{d}_1)^2}(\Delta \hat{q} - \Delta \hat{d}) \tag{96}$$

$$\Delta \bar{d} = -\Delta q. \tag{97}$$

Next, we consider the extremum condition with respect to q_0, q_1, d_0, d_1. Express the energy term as

$$G_e = \frac{1}{m} \int Dz_0 \log \int Dz_1 \exp(-m\mathcal{Y}) \tag{98}$$

$$\mathcal{Y} = -\log H\left(\frac{\kappa + \sqrt{\lambda_0}z_0 + \sqrt{\lambda_1}z_1}{\sqrt{\mu}} \right) \tag{99}$$

$$= -\log \int Dx\,\Theta\left(\kappa + \sqrt{\lambda_0}z_0 + \sqrt{\lambda_1}z_1 + \sqrt{\mu}x \right). \tag{100}$$

Then we have

$$\frac{\partial \mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)} = -\frac{1}{\int Dx\Theta} \int Dx\,\Theta' \tag{101}$$

$$\frac{\partial^2 \mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)^2} = -\frac{1}{\int Dx\Theta} \int Dx\,\Theta'' + \left(\frac{\partial \mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)} \right)^2 \tag{102}$$

$$\frac{\partial \mathcal{Y}}{\partial z_1} = \sqrt{\lambda_1} \frac{\partial \mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)} \tag{103}$$

$$\frac{\partial \mathcal{Y}}{\partial q_0} = q'_{0,\mathrm{eff}}\left(\frac{z_0}{2\sqrt{\lambda_0}} - \frac{z_1}{2\sqrt{\lambda_1}} \right) \frac{\partial \mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)} \tag{104}$$

$$\frac{\partial \mathcal{Y}}{\partial q_1} = q'_{1,\mathrm{eff}}\left(\frac{z_1}{2\sqrt{\lambda_1}} \frac{\partial \mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)} - \frac{1}{2}\left(\frac{\partial^2 \mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)^2} - \left(\frac{\partial \mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)} \right)^2 \right) \right). \tag{105}$$

Taking the derivative of $f_{\text{1-RSB}}$ with respect to q_0 gives

$$\frac{\partial f}{\partial q_0} = \frac{m}{2}\hat{q}_0 + \frac{\alpha}{m} \int Dz_0 \frac{1}{\int Dz_1 \exp(-m\mathcal{Y})} \int Dz_1 \exp(-m\mathcal{Y})(-m)\frac{\partial \mathcal{Y}}{\partial q_0} \tag{106}$$

$$= \frac{m}{2}\hat{q}_0 - \alpha q'_{0,\mathrm{eff}} \int Dz_0 \frac{1}{\int Dz_1 \exp(-m\mathcal{Y})}$$
$$\int Dz_1 \exp(-m\mathcal{Y})\left(\frac{z_0}{2\sqrt{\lambda_0}} - \frac{z_1}{2\sqrt{\lambda_1}} \right) \frac{\partial \mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)}. \tag{107}$$

Integration by parts with respect to z_1 gives

$$\int Dz_1 \exp(-m\mathcal{Y})\left(-\frac{z_1}{2\sqrt{\lambda_1}}\right)\frac{\partial\mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)} \tag{108}$$

$$= -\frac{1}{2\sqrt{\lambda_1}}\int Dz_1\frac{\partial}{\partial z_1}\left(\exp(-m\mathcal{Y})\frac{\partial\mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)}\right)$$

$$= -\frac{1}{2\sqrt{\lambda_1}}\sqrt{\lambda_1}\frac{\partial}{\partial(\sqrt{\lambda_0}z_0)}\int Dz_1\exp(-m\mathcal{Y})\frac{\partial\mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)} \tag{109}$$

$$= -\frac{1}{2}\frac{\partial}{\partial(\sqrt{\lambda_0}z_0)}\int Dz_1\exp(-m\mathcal{Y})\frac{\partial\mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)}. \tag{110}$$

Integration by parts with respect to z_0 gives

$$\int Dz_0\frac{1}{\int Dz_1\exp(-m\mathcal{Y})}\int Dz_1\exp(-m\mathcal{Y})\frac{z_0}{2\sqrt{\lambda_0}}\frac{\partial\mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)} \tag{111}$$

$$= \frac{1}{2}\int Dz_0\frac{\partial}{\partial(\sqrt{\lambda_0}z_0)}\left(\frac{1}{\int Dz_1\exp(-m\mathcal{Y})}\int Dz_1\exp(-m\mathcal{Y})\frac{\partial\mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)}\right)$$

$$= \frac{1}{2}\int Dz_0\frac{1}{\int Dz_1\exp(-m\mathcal{Y})}\frac{\partial}{\partial(\sqrt{\lambda_0}z_0)}\int Dz_1\exp(-m\mathcal{Y})\frac{\partial\mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)}$$

$$+ \frac{m}{2}\int Dz_0\left(\frac{1}{\int Dz_1\exp(-m\mathcal{Y})}\int Dz_1\exp(-m\mathcal{Y})\frac{\partial\mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)}\right)^2. \tag{112}$$

Thus we have

$$\frac{\partial f}{\partial q_0} = \frac{m}{2}\hat{q}_0 - \alpha\frac{m}{2}q'_{0,\text{eff}}\int Dz_0\left(\frac{1}{\int Dz_1\exp(-m\mathcal{Y})}\int Dz_1\exp(-m\mathcal{Y})\frac{\partial\mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)}\right)^2. \tag{113}$$

Next, taking the derivative of $f_{\text{1-RSB}}$ with respect to q_1 gives

$$\frac{\partial f}{\partial q_1} = \frac{1-m}{2}\hat{q}_1 + \frac{\alpha}{m}\int Dz_0\frac{1}{\int Dz_1\exp(-m\mathcal{Y})}\int Dz_1\exp(-m\mathcal{Y})(-m)\frac{\partial\mathcal{Y}}{\partial q_1} \tag{114}$$

$$= \frac{1-m}{2}\hat{q}_1 - \alpha q'_{1,\text{eff}}\int Dz_0\frac{1}{\int Dz_1\exp(-m\mathcal{Y})}\int Dz_1\exp(-m\mathcal{Y})$$

$$\times\left(\frac{z_1}{2\sqrt{\lambda_1}}\frac{\partial\mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)} - \frac{1}{2}\left(\frac{\partial^2\mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)^2} - \left(\frac{\partial\mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)}\right)^2\right)\right). \tag{115}$$

Integration by parts with respect to z_1 gives

$$\int Dz_1\exp(-m\mathcal{Y})\left(\frac{z_1}{2\sqrt{\lambda_1}}\right)\frac{\partial\mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)} \tag{116}$$

$$= \frac{1}{2}\frac{\partial}{\partial(\sqrt{\lambda_0}z_0)}\int Dz_1\exp(-m\mathcal{Y})\frac{\partial\mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)}$$

$$= \frac{1}{2}\int Dz_1\exp(-m\mathcal{Y})\left(-m\left(\frac{\partial\mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)}\right)^2 + \frac{\partial^2\mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)^2}\right). \tag{117}$$

Thus

$$\frac{\partial f}{\partial q_1} = \frac{1-m}{2}\hat{q}_1 - \alpha\frac{1-m}{2}\int Dz_0\frac{1}{\int Dz_1\exp(-m\mathcal{Y})}\int Dz_1\exp(-m\mathcal{Y})\left(\frac{\partial\mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)}\right)^2 \tag{118}$$

and

$$\Delta\hat{q} = \alpha \int Dz_0 \left(q'_{1,\text{eff}} \frac{1}{\int Dz_1 \exp(-m\mathcal{Y})} \int Dz_1 \exp(-m\mathcal{Y}) \left(\frac{\partial \mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)} \right)^2 \right.$$
$$\left. - q'_{0,\text{eff}} \left(\frac{1}{\int Dz_1 \exp(-m\mathcal{Y})} \int Dz_1 \exp(-m\mathcal{Y}) \frac{\partial \mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)} \right)^2 \right) \tag{119}$$

$$= \alpha \int Dz_0 \left(q'_{0,\text{eff}} \text{Var} \left[\frac{\partial \mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)} \right] + (q'_{1,\text{eff}} - q'_{0,\text{eff}}) \mathbb{E} \left[\left(\frac{\partial \mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)} \right)^2 \right] \right). \tag{120}$$

Here, the expectation and variance are taken with respect to the measure $\frac{1}{\int Dz_1 \exp(-m\mathcal{Y})} Dz_1 \exp(-m\mathcal{Y})$. Expanding $\frac{\partial \mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)}$ around $\Delta q = 0$, we get

$$\frac{\partial \mathcal{Y}}{\partial(\sqrt{\lambda_0}z_0)} \approx \frac{\partial \mathcal{Y}_{\text{RS}}}{\partial(\sqrt{\lambda}z_0)} + \frac{\partial^2 \mathcal{Y}_{\text{RS}}}{\partial \left(\sqrt{\lambda}z_0 \right)^2} \sqrt{\Delta q} z_1 \tag{121}$$

$$\mathcal{Y}_{\text{RS}} = -\log H \left(\frac{\kappa + \sqrt{\lambda}z_0}{\sqrt{\mu}} \right). \tag{122}$$

Also, when $\Delta q \ll 1$, $\frac{1}{\int Dz_1 \exp(-m\mathcal{Y})} Dz_1 \exp(-m\mathcal{Y}) \approx Dz_1$ holds and thus $\text{Var}[z_1] \approx 1$. Therefore,

$$\Delta\hat{q} \approx \alpha \int Dz_0 \left(q'_{\text{eff}} \left(\frac{\partial^2 \mathcal{Y}_{\text{RS}}}{\partial \left(\sqrt{\lambda}z_0 \right)^2} \right)^2 + q''_{\text{eff}} \left(\frac{\partial \mathcal{Y}_{\text{RS}}}{\partial(\sqrt{\lambda}z_0)} \right)^2 \right) \Delta q. \tag{123}$$

Similarly,

$$\Delta\hat{d} \approx \alpha \int Dz_0 \langle g' \rangle^2 \left(\frac{\partial^2 \mathcal{Y}_{\text{RS}}}{\partial \left(\sqrt{\lambda}z_0 \right)^2} \right)^2 \Delta q. \tag{124}$$

The condition that $(0,0,0,0)$ is stable in the system of equations (96),(97),(123),(124) is

$$\frac{1}{(\hat{E} + \hat{q} - \hat{c} - \hat{d})^2} \cdot \alpha \int Dz_0 \left((q'_{\text{eff}} - \langle g' \rangle^2) \left(\frac{\partial^2 \mathcal{Y}_{\text{RS}}}{\partial \left(\sqrt{\lambda}z_0 \right)^2} \right)^2 + q''_{\text{eff}} \left(\frac{\partial \mathcal{Y}_{\text{RS}}}{\partial(\sqrt{\lambda}z_0)} \right)^2 \right) < 1. \tag{125}$$

Reorganizing, we obtain

$$\alpha(1-q)^2 \int Dz \left(\frac{q'_{\text{eff}} - \langle g' \rangle^2}{\lambda^2} \left(\frac{\partial^2}{\partial z^2} \log H(u) \right)^2 + \frac{q''_{\text{eff}}}{\lambda} \left(\frac{\partial}{\partial z} \log H(u) \right)^2 \right) < 1. \tag{126}$$

[1] J. A. Zavatone-Veth and C. Pehlevan, Activation function dependence of the storage capacity of treelike neural networks, Physical Review E **103**, L020301 (2021).
[2] A. Engel, H. Köhler, F. Tschepke, H. Vollmayr, and A. Zippelius, Storage capacity and learning algorithms for two-layer neural networks, Physical Review A **45**, 7590 (1992).