

Supplementary information of: “*The Autoregressive Neural Network Architecture of the Boltzmann Distribution of Pairwise Interacting Spin Systems*”

Indaco Biazzo

CONTENTS

I. ANRR architecture of the Curie-Weiss model	1
A. Exact expression of the conditional probability of the CW model	2
B. Thermodynamical limit of the conditional probability of the CW model	2
II. Analytic expression of the Curie-Weiss free energy at finite N	3
III. ARNN architectures of The Sherrington-Kirkpatrick model	4
A. Replica Symmetric solution (RS)	4
B. K-step Replica Symmetric Breaking (k-RSB) solution	6
C. Numerical evidence of integrals approximations	8
IV. Monte Carlo sampling of the SK model	8
References	9

I. ANRR ARCHITECTURE OF THE CURIE-WEISS MODEL

The Curie-Weiss model (CW) is a uniform, fully-connected Ising model. The Hamiltonian of the Curie-Weiss model (CW) with N spins is $H(\mathbf{x}) = -h \sum_{i=1}^N x_i - \frac{J}{N} \sum_{i<j} x_i x_j$. Defining $M_{<i}[\mathbf{x}_{<i}] = \sum_{s=1}^{i-1} x_s$ and $M_{>i} = \sum_{l=i+1}^N x_l$, the conditional probability of a spin i , see Eq. (11) in the main text, of the CW model, is:

$$P^{CW}(x_i = 1 | \mathbf{x}_{<i}) = \sigma \left(2\beta h + 2\beta \frac{J}{N} M_{<i}[\mathbf{x}_{<i}] + \log(\rho_i^+[\mathbf{x}_{<i}]) - \log(\rho_i^-[\mathbf{x}_{<i}]) \right), \quad (1)$$

where:

$$\rho_i^\pm[\mathbf{x}_{<i}] \propto \sum_{\mathbf{x}_{>i}} e^{\beta \left(h \pm \frac{J}{N} + \frac{J}{N} M_{<i}[\mathbf{x}_{<i}] \right) M_{>i} + \frac{\beta J}{2N} M_{>i}^2} \quad (2)$$

Defining $h_i^\pm[\mathbf{x}_{<i}] = h \pm \frac{J}{N} + \frac{J}{N} M_{<i}[\mathbf{x}_{<i}]$, at given $\mathbf{x}_{<i}$, the Eq. (2) is equivalent to the partition function of the CW model, with $N - i$ spins and external fields $h_i^\pm$, as written in the main text. The sums over the configurations of the spins l can be carried on easily:

$$\begin{aligned} \rho_i^\pm[\mathbf{x}_{<i}] &= \sum_{x_{i+1} \dots x_N} e^{\beta h_i^\pm[\mathbf{x}_{<i}] M_{>i} + \frac{\beta J}{2N} M_{>i}^2} = \sqrt{\frac{N}{2\pi\beta J}} \sum_{x_{i+1} \dots x_N} e^{\beta h_i^\pm[\mathbf{x}_{<i}] M_{>i}} \int e^{-\frac{N}{2J\beta} t^2 + t M_{>i}} dt \\ &= \sqrt{\frac{N}{2\pi\beta J}} \int dt e^{-\frac{N}{2J\beta} t^2} \sum_{x_{i+1} \dots x_N} e^{(\beta h_i^\pm[\mathbf{x}_{<i}] + t) M_{>i}} = \sqrt{\frac{N}{2\pi\beta J}} \int dt e^{-\frac{N}{2J\beta} t^2} \left(e^{\beta h_i^+[\mathbf{x}_{<i}] + t} + e^{(-\beta h_i^+[\mathbf{x}_{<i}] - t)} \right)^{N-i} \end{aligned} \quad (3)$$

where we used the Hubbard–Stratonovich (HS) transformation to obtain the second equality. The last equality is obtained just by summing over the variables x_l .

First, in the following, we derive the exact feed-forward representation of Eq. (1) at finite N number of variables, then in the limit of $N \rightarrow \infty$.

A. Exact expression of the conditional probability of the CW model

The integral in the equation 3 can be computed the following way:

$$\begin{aligned}
 \rho_i^\pm[\mathbf{x}_{<i}] &= \sqrt{\frac{N}{2\pi\beta J}} \int dt e^{-\frac{N}{2\beta} t^2} \sum_{k=0}^{N-i} \binom{N-i}{k} e^{(N-i-2k)*(\beta h_i^\pm[\mathbf{x}_{<i}]+t)} \\
 &= \sum_{k=0}^{N-i} \binom{N-i}{k} \sqrt{\frac{N}{2\pi\beta J}} \int dt e^{-\frac{N}{2\beta} t^2} e^{(N-i-2k)*(\beta h_i^\pm[\mathbf{x}_{<i}]+t)} \\
 &= \sum_{k=0}^{N-i} \binom{N-i}{k} e^{\frac{\beta J}{2N}(N-i-2k)^2 + (N-i-2k)(\beta h \pm \frac{\beta J}{N})} e^{\frac{\beta J}{N}(N-i-2k) \sum_s x_s} \\
 &= \sum_{k=0}^{N-i} e^{b_{i,k}^\pm + w_{i,k} \sum_s x_s}
 \end{aligned}$$

where:

$$b_{ik}^\pm = \log \binom{N-i}{k} + \frac{\beta J}{2N} (N-i-2k)^2 + (N-i-2k) \left(\beta h \pm \frac{\beta J}{N} \right) \quad (4)$$

$$\omega_{ik} = \frac{\beta J}{N} (N-i-2k), \quad (5)$$

proving Eq. (10) shown in the main text.

B. Thermodynamical limit of the conditional probability of the CW model

In the thermodynamical limit, the Curie-Weiss model admits an analytical solution. The order parameter of the system is the magnetization, $m_\beta = \frac{1}{NZ} \sum_{\mathbf{x}} \sum_i x_i e^{-\beta H}$ with $Z = \sum_{\mathbf{x}} e^{-\beta H}$. At high temperatures, with zero external fields $h = 0$, the magnetization, m_β , is zero up to a critical temperature $\beta_c = \frac{1}{J}$, where a phase transition occurs, and a non-zero magnetization is observed [1]. Considering the following variables: $\eta_i = \frac{N-i}{N}$, $m_i = -\frac{N-i-2k}{N}$, and for simplicity, the $h = 0$ case, we can rewrite the expression, Eq. (3), as:

$$\rho_i^\pm[\mathbf{x}_{<i}] = \sqrt{\frac{N}{2\pi\beta J}} \int dt e^{-\frac{N}{2\beta} t^2} \sum_{k=0}^{N-i} \binom{N-i}{k} e^{(N-i-2k)*(\beta h_i^\pm[\mathbf{x}_{<i}]+t)} \quad (6)$$

$$= \sum_{k=0}^{N-i} \binom{N-i}{k} e^{\frac{\beta J}{2N}(N-i-2k)^2 + (N-i-2k)(\pm \frac{\beta J}{N})} e^{\frac{\beta J}{N}(N-i-2k) \sum_s x_s} \quad (7)$$

$$= \sum_{m_i = -\eta_i}^{\eta_i} \left(\frac{N\eta_i}{\frac{N(m_i + \eta_i)}{2}} \right) e^{\frac{N\beta J}{2} m_i^2 \mp \beta J m_i} e^{-N\beta J \frac{\sum_s x_s}{N}} \quad (8)$$

In the limit $N \gg 1$, we obtain:

$$\rho_i^\pm = \int_{-\eta_i}^{\eta_i} \left(\frac{N\eta_i}{N(m_i + \eta_i)} \right) e^{\frac{N\beta J}{2} m_i^2 \mp \beta J m_i} e^{N\beta J \frac{\sum_s x_s}{N}} dm_i \quad (9)$$

$$\approx \int_{-\eta_i}^{\eta_i} \exp \left\{ -N\eta_i \left(-\frac{1 + \frac{m_i}{\eta_i}}{2} \log \frac{1 + \frac{m_i}{\eta_i}}{2} - \frac{1 - \frac{m_i}{\eta_i}}{2} \log \frac{1 - \frac{m_i}{\eta_i}}{2} - \frac{\beta m_i^2}{2\eta_i} + \beta m_i \frac{\sum_s x_s}{N} \right) \right\} e^{\mp \beta J m_i} dm_i \quad (10)$$

$$\approx \exp \left\{ -N\eta_i \left(-\frac{1 + \frac{\tilde{m}_i}{\eta_i}}{2} \log \frac{1 + \frac{\tilde{m}_i}{\eta_i}}{2} - \frac{1 - \frac{\tilde{m}_i}{\eta_i}}{2} \log \frac{1 - \frac{\tilde{m}_i}{\eta_i}}{2} - \frac{\beta \tilde{m}_i^2}{2\eta_i} + \beta \tilde{m}_i \frac{\sum_s x_s}{N} \right) \right\} e^{\mp \beta J \tilde{m}_i} \quad (11)$$

$$\doteq e^{\mp \beta J \tilde{m}_i}$$

where in the second line I used the Stirling approximation for the binomial factor, in the third line I used the saddle point method, assuming $N \gg 1$, where $\tilde{m}_i$ is the extreme of the function inside the curly brackets. In the last line, I simplify all the elements that are equal between ρ_i^+ and ρ_i^- that cancel each other in the conditional probability function Eq. (1). Deriving by m_i the function inside the curly brackets, we obtain that the extreme $\tilde{m}_i$ should satisfy the following equation:

$$\frac{\tilde{m}_i}{\eta_i} = \tanh \left(\beta \left(\frac{\tilde{m}_i}{\eta_i} - \frac{\sum_s x_s}{N} \right) \right) \quad (12)$$

In the N large limit, and for a typical sample, I assume that: $\frac{\sum_s x_s}{N} = \frac{i}{N} \sum_s x_s \approx \frac{i}{N} |\tilde{m}_\beta| \text{sgn}(\sum_s x_s)$, where the m_β is the magnetization of the Curie-Weiss system at inverse temperature β and $\text{sgn}(x) = \frac{|x|}{x}$ is the sign function. We can distinguish two cases when the magnetization of the system is zero or not. In the first case, when $\beta \leq 1/J$, the solution of Eq. (12) is zero as well, and $\log(\rho_i^+) - \log(\rho_i^-) = 0$ because the only term that makes the two quantities different, $\mp \beta J m_i$, vanish. When instead the system acquires a magnetization m_β different from zero, Eq. (12) admit one maximum that depends on the two possible symmetric values of $\frac{i}{N} \sum_s x_s \approx \frac{i}{N} |\tilde{m}_\beta| \text{sign}(\sum_s x_s)$. The solution of Eq. (12), $\pm \tilde{m}_{\text{extrem}}$ depends again on $\text{sign}(\sum_s x_s)$, and we can write the maximum solution as $\tilde{m}_i = |\tilde{m}_i| \text{sgn}(\sum_s x_s)$. Easily we obtain that $\log(\rho_i^+) - \log(\rho_i^-) = -2\beta J |\tilde{m}_i| \text{sgn}(\sum_s x_s)$, demonstrating the formula in the main text.

II. ANALYTIC EXPRESSION OF THE CURIE-WEISS FREE ENERGY AT FINITE N

The free energy of a system is defined as:

$$F = -\frac{1}{\beta} \log Z \quad (13)$$

For the CW system, with similar computation that leads to Eq. (7), we find:

$$F = -\frac{1}{\beta} \log \sum_{k=0}^N \binom{N}{k} e^{\frac{\beta J}{2} (N-2k)^2 + (N-2k)h} + \frac{J}{2} \quad (14)$$

III. ARNN ARCHITECTURES OF THE SHERRINGTON-KIRKPATRICK MODEL

In order to derive the architecture for the SK model based on the replica method, I'll start from Eq. (17) presented in the main text:

$$\mathbb{E}_{\underline{J}}[(\rho_i^\pm[\mathbf{x}_{<i}])^n] = \int \prod_{l < l'} dP_{J_{ll'}} \left\{ \sum_{\{\underline{x}^a\}_{i+1}^N} \exp \left[\beta \left(\sum_{l,a} \left(\pm J_{il} + \sum_s J_{sl} x_s \right) x_l^a + \sum_{l,l',a} J_{ll'} x_l^a x_{l'}^a \right) \right] \right\} \quad (15)$$

where, the sums over (l, l') , s and a run respectively over $(i+1, N)$, $(1, i-1)$ and $(1, n)$, and $dP_{J_{ll'}} = P(J_{ll'}) dJ_{ll'}$. Here we assumed that $N - i \gg 1$, and the average over the disorder $\mathbb{E}_{\underline{J}}$ is on the parameters $J_{l,l'}$ with $l, l' > i$. Defining $h_l^\pm[\mathbf{x}_{<i}] = \pm J_{il} + \sum_{s=1}^{i-1} J_{sl} x_s$ as an external field, we can observe that the above quantity is the partition function of the SK model on $\mathbf{x}_{i>}$ variables, with external fields $h_l^\pm[\mathbf{x}_{<i}]$, at fixed $\mathbf{x}_{<i}$, and $J_{ll'}$ as coupling constants. Computing the integrals over the disorder variables $\{J_{ll'}\}$ yields [2]:

$$\mathbb{E}_{\underline{J}}[(\rho_i^\pm[\mathbf{x}_{<i}])^n] = \sum_{\{\underline{x}^a\}_{i+1}^N} \exp \left\{ \beta \left[\sum_l h_l^\pm[\mathbf{x}_{<i}] \sum_a x_l^a + \frac{\beta}{2N} \sum_{l,l'} \sum_{a,b} x_l^a x_l^b x_{l'}^a x_{l'}^b \right] \right\} \quad (16)$$

$$= e^{\frac{(N-i)\beta^2}{4N}((N-i)n-n^2)} \sum_{\{\underline{x}^a\}_{i+1}^N} \exp \left\{ \beta \left[\sum_l h_l^\pm[\mathbf{x}_{<i}] \sum_a x_l^a + \frac{\beta}{2N} \sum_{a < b} \left(\sum_l x_l^a x_l^b \right)^2 \right] \right\} \quad (17)$$

Using the Hubbard-Stratonovich transformation of the quadratic terms, we can write:

$$\mathbb{E}_{\underline{J}}[(\rho_i^\pm[\mathbf{x}_{<i}])^n] = c(n, N, i) \int \prod_{a < b} dQ_{ab} e^{-\frac{N}{2}\beta^2 Q_{a,b}^2} \sum_{\{\underline{x}^a\}_{i+1}^N} \exp \left\{ \beta \left[\sum_l h_l^\pm[\mathbf{x}_{<i}] \sum_a x_l^a + \beta \sum_{a < b} Q_{a,b} \sum_l x_l^a x_l^b \right] \right\} \quad (18)$$

$$= c(n, N, i) \int \prod_{a < b} dQ_{ab} e^{-\frac{N}{2}\beta^2 Q_{a,b}^2} \prod_l \left[\sum_{\{x_l^a\}} \exp \left\{ \beta \left[h_l^\pm[\mathbf{x}_{<i}] \sum_a x_l^a + \beta \sum_{a < b} Q_{a,b} x_l^a x_l^b \right] \right\} \right] \quad (19)$$

where we defined:

$$c(n, N, i) = e^{\frac{(N-i)\beta^2}{4N}((N-i)n-n^2)} \left(\frac{2\pi\beta^2}{N} \right)^{\frac{n(n-1)}{2}}.$$

The Parisi solutions of the SK model prescribe how to parametrize the matrix of the overlaps Q [3]. The following shows how to obtain neural network architectures based on the replica symmetric (RS) and k-step replica symmetric broken (k-RSB) solutions.

A. Replica Symmetric solution (RS)

We assume that the overlaps between the replicas are symmetric under permutations, and the matrix of the overlaps between replicas is parametrized with only one variable q :

$$Q_{a,b} = \begin{cases} 0, & \text{if } a = b \\ q, & \text{otherwise} \end{cases},$$

obtaining:

$$\mathbb{E}_{\underline{J}}[(\rho_i^{\pm, sym}[\mathbf{x}_{<i}])^n] = c(n, N, i) \int dq e^{-\frac{n(n-1)N}{4}\beta^2 q^2} \prod_l \left[\sum_{\{\underline{x}_l^a\}} \exp \left\{ \beta \left[h_l^{\pm}[\mathbf{x}_{<i}] \sum_a x_l^a + \beta q \sum_{a < b} x_l^a x_l^b \right] \right\} \right] \quad (20)$$

$$= c(n, N, i) \int dq e^{-\frac{n(n-1)N}{4}\beta^2 q^2} e^{-\frac{nN\beta^2 q}{2}} \prod_l \left[\sum_{\{\underline{x}_l^a\}} e^{\beta \left[h_l^{\pm}[\mathbf{x}_{<i}] \sum_a x_l^a + \frac{\beta q}{2} (\sum_a x_l^a)^2 \right]} \right] \quad (21)$$

$$= c'(n, N, i) \int dq e^{-\frac{n(n-1)N}{4}\beta^2 q^2} e^{-\frac{nN\beta^2 q}{2}} \prod_l \left[\int \frac{dz_l}{\sqrt{2\pi q}} e^{-\frac{z_l^2}{q}} \sum_{\{\underline{x}_l^a\}} e^{\beta (h_l^{\pm}[\mathbf{x}_{<i}] + \beta z_l) \sum_a x_l^a} \right] \quad (22)$$

$$= c'(n, N, i) \int dq e^{-nN \left(\frac{(n-1)}{4}\beta^2 q^2 + \frac{\beta^2 q}{2} \right)} \prod_l \left[\int \frac{dz_l}{\sqrt{2\pi q}} e^{-\frac{z_l^2}{q}} 2^n \cosh^n (\beta (h_l^{\pm}[\mathbf{x}_{<i}] + \beta z_l)) \right]. \quad (23)$$

$$(24)$$

Using the limit that $n \rightarrow 0$ we can write:

$$\int \frac{dz_l}{\sqrt{2\pi q}} e^{-\frac{z_l^2}{q}} 2^n \cosh^n (\beta (h_l^{\pm}[\mathbf{x}_{<i}] + \beta z_l)) = e^{n \int \frac{dz_l}{\sqrt{2\pi q}} e^{-\frac{z_l^2}{q}} \log 2 \cosh (\beta (h_l^{\pm}[\mathbf{x}_{<i}] + \beta z_l))}. \quad (25)$$

obtaining:

$$\log(\rho_i^{\pm, sym}[\mathbf{x}_{<i}]) = \lim_{n \rightarrow 0} \frac{1}{n} \log \left(c'(n, N, i) \int dq e^{-\frac{n(n-1)N}{4}\beta^2 q^2} e^{-\frac{nN\beta^2 q}{2}} e^{n \sum_l \int \frac{dz_l}{\sqrt{2\pi q}} e^{-\frac{z_l^2}{q}} \log 2 \cosh (\beta (h_l^{\pm}[\mathbf{x}_{<i}] + \beta z_l))} \right) \quad (26)$$

$$= \log(c''(N, i)) + \left(+\frac{N}{4}\beta^2 q_0^2 - \frac{N\beta^2 q_0}{2} + \sum_l \int \frac{dz_l}{\sqrt{2\pi q_0}} e^{-\frac{z_l^2}{q_0}} \log 2 \cosh (\beta (h_l^{\pm}[\mathbf{x}_{<i}] + \beta z_l)) \right) \quad (27)$$

$$\doteq c(N, i, q_0) - \sum_l \int \frac{dz_l}{\sqrt{2\pi q_0}} e^{-\frac{z_l^2}{q_0}} \log \sigma (\beta (2h_l^{\pm}[\mathbf{x}_{<i}] + 2\beta z_l)) \quad (28)$$

In the second line, I use the saddle point methods to evaluate the integral over q , assuming that the single maximum value q_0 does not depend on the input values $\mathbf{x}_{<i}$ in the set of $h_l^{\pm}[\mathbf{x}_{<i}]$. It is a bold assumption to be verified *a posteriori* on the quality of the neural network architectures performances. In the third line, we use the identity $\log \cosh(x) = 2x - \log \sigma(2x)$ and the elements that are equals between $\log(\rho^+)$ and $\log(\rho^-)$ are simplified. We introduced the $\log \sigma$ non-linear operator for computational reasons.

Now we consider the following approximation of the Gaussian convolution:

$$\int dz e^{-z^2} \log \sigma (\beta (h + \beta \sqrt{q} z)) \sim b_0 + w_0 * \log \sigma(b_1 + w_1 h),$$

where (b_0, w_0, b_1, w_1) are free parameters to be determined. In the sec.III C of the SI a numerical analysis of the quality of this approximation is shown. Putting together all the pieces, we can parameterize the conditional probability as:

$$\begin{aligned} Q^{RS}(x_i = 1 | \mathbf{x}_{<i}) &= \sigma(x_i^1(\mathbf{x}_{<i}) + \log(\rho_i^+) - \log(\rho_i^-)) \\ &= \sigma \left(x_i^1(\mathbf{x}_{<i}) + \sum_{l^+=i+1}^N w_0^{i, l^+} \log \sigma(b_1^{i, l^+} + w_1^{i, l^+} x_{i, l^+}^1(\mathbf{x}_{<i})) + \sum_{l^-=i+1}^N w_0^{i, l^-} \log \sigma(b_1^{i, l^-} + w_1^{i, l^-} x_{i, l^-}^1(\mathbf{x}_{<i})) \right) \end{aligned} \quad (29)$$

where the set of $(\mathbf{b}, \mathbf{w})$ are free variational parameters to learn in the training process.

B. K-step Replica Symmetric Breaking (k-RSB) solution

Assuming that the replica symmetry is broken, we can use the ansatz called 1-step replica symmetric breaking (1RSB), where the overlaps between replicas are divided into m blocks:

$$Q_{a,b} = \begin{cases} q_1, & \text{if } I(a/m) = I(b/m) \\ q_0, & \text{if } I(a/m) \neq I(b/m). \end{cases} \quad (30)$$

With the above ansatz, we can compute the following quantities:

$$\sum_{ab} Q_{ab} x_a x_b = \frac{1}{2} \left[q_0 \left(\sum_a x_a \right)^2 + (q_1 - q_0) \sum_{\text{blocks}} \left(\sum_{a \in \text{block}} x_a \right)^2 - n q_1 \right] \quad (31)$$

$$\sum_{ab} Q_{ab}^2 = n^2 q_0^2 + n m (q_1^2 - q_0^2) - n q_1^2. \quad (32)$$

The equation 19 becomes:

$$\begin{aligned} \mathbb{E}_{\underline{J}} \left[(\rho_i^{\pm, 1RSB})^n \right] &= \\ &= c(n, N, i) \int dq_1 dq_0 e^{\frac{N}{2} \beta^2 [n^2 q_0^2 + n m (q_1^2 - q_0^2) - n q_1^2]} \prod_l \left[\sum_{\{x_l^a\}} e^{\beta [h_l^{\pm} \sum_a x_l^a + \beta q_0 (\sum_a x_p^a)^2 + \beta (q_1 - q_0) \sum_{\text{blocks}} (\sum_{a \in \text{block}} x_l^a)^2 - n q_1]} \right] \\ &= c(n, N, i) \int dq_1 dq_0 e^{\frac{N}{2} \beta^2 [n^2 q_0^2 + n m (q_1^2 - q_0^2) - n q_1^2 - n q_1]} \prod_l \left[\sum_{\{x_l^a\}} \int dP_{z_l} \prod_{k=1}^{n/m} \int dP_{y_{lk}} e^{\beta [h_l^{\pm} \sum_a x_l^a + \beta z_l \sum_a x_l^a + \beta \sum_{\text{blocks}} y_{lk} \sum_{a \in \text{block}} x_l^a]} \right] \\ &= c(n, N, i) \int dq_1 dq_0 e^{\frac{N}{2} \beta^2 [n^2 q_0^2 + n m (q_1^2 - q_0^2) - n q_1^2 - n q_1]} \prod_l \left[\int dP_{z_l} \prod_{k=1}^{n/m} \int dP_{y_{lk}} \cosh^m \left(\beta [h_l^{\pm} + \beta z_l + \beta y_{lk}] \right) \right] \\ &= c'(n, N, i) + c(n, N, i) \int dq_1 dq_0 e^{\frac{N}{2} \beta^2 [n^2 q_0^2 + n m (q_1^2 - q_0^2) - n q_1^2 - n q_1]} \prod_p \int dP_{z_l} \exp \left\{ \frac{n}{m} \log \left(\int dP_{y_l} \cosh^m \left(\beta [h_l^{\pm} + \beta z_l + \beta y_l] \right) \right) \right\}, \end{aligned}$$

where we defined:

$$dP_{z_l} = \frac{dz_l}{\sqrt{2\pi q_0}} e^{\frac{z_l^2}{2q_0}} \quad (33)$$

$$dP_{y_l} = \frac{dy_l}{\sqrt{2\pi(q_1 - q_0)}} e^{\frac{y_l^2}{2(q_1 - q_0)}}. \quad (34)$$

Considering $N \gg 1$ and $n \rightarrow 0$ to use the saddle point methods and the identity in Eq. (25), we can write:

$$\log(\rho_i^{\pm, 1RSB}) = \lim_{n \rightarrow 0} \frac{1}{n} \log \left(\mathbb{E}_{\underline{J}} \left[(\rho_i^{\pm, 1RSB})^n \right] \right) \quad (35)$$

$$= c_i + \text{Extr}_{q_0, q_1} \left[c'_i(N, n, q_0, q_1) + \frac{1}{m} \sum_l \int dP_{z_l} \log \left(\int dP_{y_l} \cosh^m \left(\beta [h_l^{\pm} + \beta z_l + \beta y_l] \right) \right) \right]. \quad (36)$$

The above integrals are rewritten as the following:

$$\int dP_{z_l} \log \left(\int dP_{y_l} \cosh^m \left(\beta [h_l^\pm + \beta z_l + \beta y_l] \right) \right) = \quad (37)$$

$$\int dP_{z_l} \log \left(\int dP_{y_l} e^{m \log \cosh(\beta [h_l^\pm + \beta z_l + \beta y_l])} \right) = \quad (38)$$

$$\beta h_l^\pm + \int dP_{z_l} \log \left(\int dP_{y_l} e^{\beta^2 m y_l - m \log \sigma(\beta [h_l^\pm + \beta z_l + \beta y_l])} \right) \quad (39)$$

We have two nested Gaussian convolutions.

The integrals over the variables (z_l, y_l) are approximated in the same way as the approach used previously for the RS case: the number of free parameters increases to have feed-forward functions without integrals. First, we consider the following maps:

$$\log \int dP_{y_l} e^{\beta^2 m y_l - m \log \sigma(\beta [h_l^\pm + \beta z_l + \beta y_l])} \approx \quad (40)$$

$$\log \left(\hat{b}_0^{l^\pm} + \hat{w}_0^{l^\pm} e^{\hat{b}_1^{l^\pm} + \hat{w}_1^{l^\pm} \log \sigma(\hat{b}_2^{l^\pm} + \hat{w}_2^{l^\pm} (h_l^\pm + \beta z_l))} \right) = \quad (41)$$

$$\log(\hat{b}_0^{l^\pm}) + \log \left(1 + e^{\log(\frac{\hat{w}_0^{l^\pm}}{\hat{b}_0^{l^\pm}}) + \hat{b}_1^{l^\pm} + \hat{w}_1^{l^\pm} \log \sigma(\hat{b}_2^{l^\pm} + \hat{w}_2^{l^\pm} (h_l^\pm + \beta z_l))} \right) = \quad (42)$$

$$\log(\hat{b}_0^{l^\pm}) - \log \sigma \left(\log \left(\frac{\hat{w}_0^{l^\pm}}{\hat{b}_0^{l^\pm}} \right) + \hat{b}_1^{l^\pm} + \hat{w}_1^{l^\pm} \log \sigma(\hat{b}_2^{l^\pm} + \hat{w}_2^{l^\pm} (h_l^\pm + \beta z_l)) \right) = \quad (43)$$

$$b_0^{l^\pm} + w_0^{l^\pm} \log \sigma(b_1^{l^\pm} + w_1^{l^\pm} \log \sigma(b_2^{l^\pm} + w_2^{l^\pm} (h_l^\pm + \beta z_l))) \quad (44)$$

Where in the last equation I renamed the variational parameters, for instance, $b_0^{l^\pm} = \log(\hat{b}_0^{l^\pm})$, $w_0^{l^\pm} = -1$, etc. Then once again:

$$\hat{b}_0^{l^\pm} + \hat{w}_0^{l^\pm} \int dP_{z_l} \log \sigma \left(\hat{b}_1^{l^\pm} + \hat{w}_1^{l^\pm} \log \sigma(\hat{b}_2^{l^\pm} + \hat{w}_2^{l^\pm} (h_l^\pm + \beta z_l)) \right) \approx \quad (45)$$

$$b_0^{l^\pm} + w_0^{l^\pm} \log \sigma(b_1^{l^\pm} + w_1^{l^\pm} \log \sigma(b_2^{l^\pm} + w_2^{l^\pm} (h_l^\pm))). \quad (46)$$

The set of parameters $(\mathbf{b}_0^\pm, \mathbf{w}_0^\pm, \mathbf{b}_1^\pm, \mathbf{w}_1^\pm, \mathbf{b}_2^\pm, \mathbf{w}_2^\pm)$ are free parameters to be determined by the learning procedures. The Gaussian convolution integrals are substituted by feed-forward non-linear operations, enlarging the space of parameters. In section III C numerical evidence for the goodness of the above approximations is shown. For the 1RSB case, we use the following architecture:

$$Q^{1RSB}(x_i = 1 | \mathbf{x}_{< i}) = \sigma \left(x_i^1(\mathbf{x}_{< i}) + \log(\rho_i^+) - \log(\rho_i^-) \right) = \sigma \left(x_i^1(\mathbf{x}_{< i}) + \sum_{l^+ = i+1}^N w_0^{i, l^+} \log \sigma(b_1^{i, l^+} + w_1^{i, l^+} \log \sigma(b_2^{i, l^+} + w_2^{i, l^+} x_{i, l^+}^1(\mathbf{x}_{< i}))) + \sum_{l^- = i+1}^N w_0^{i, l^-} \log \sigma(b_1^{i, l^-} + w_1^{i, l^-} \log \sigma(b_2^{i, l^-} + w_2^{i, l^-} x_{i, l^-}^1(\mathbf{x}_{< i}))) \right)$$

The generalization of the $\rho^\pm$ parametrization to the k-RSB case is straightforward. For instance, for 2-RSB we have:

$$\log \rho^{\pm, 2RSB}(x_i = 1 | \mathbf{x}_{< i}) = \sum_{l^\pm} w_0^{i, l^\pm} \log \sigma(b_1^{i, l^\pm} + w_1^{i, l^\pm} \log \sigma(b_2^{i, l^\pm} + w_2^{i, l^\pm} (\log \sigma(b_3^{i, l^\pm} + w_3^{i, l^\pm} x_{i, l^\pm}^1(\mathbf{x}_{< i}))))))$$

C. Numerical evidence of integrals approximations

In this section, I show the numerical evidence of some of the approximations used to represent the non-linear operations derived for the SK model. In the RS case, see Subsec. III A, the non-linear operation is approximated as in the following:

$$f(x) = \int \frac{dz}{\sqrt{2\pi q_0}} e^{-\frac{z^2}{q_0}} \log \sigma(\beta(x + \beta\sqrt{q}z)) \sim b_0 + w_0 * \log \sigma(b_1 + w_1 h). \quad (47)$$

The results in Fig. 1 of the fit, using the non-linear least squares method, of the $f(x)$ with the r.h.s of the above expression, with free parameters b_0, w_0, b_1, w_1 , are shown. The fits, for different values of β and q_0 , show the robustness of the approximation used. In the 1RSB case, see subsection III B, one of the non-linear operations is approximated as in the following:

$$f(x) = \log \int \frac{dz}{\sqrt{2\pi(q_1 - q_0)}} e^{-\frac{z^2}{(q_1 - q_0)}} e^{\beta^2 m z - m \log \sigma(\beta x + \beta^2 z)} \approx b_0 + w_0 \log \sigma(b_1 + w_1 \log \sigma(b_2 + w_2 x)) \quad (48)$$

As before, the function $f(x)$ is fitted with the r.h.s of the above expression, and the results in Fig. 2 show the robustness of the approximation used for different values of $(q_1 - q_0)$, m , and β .

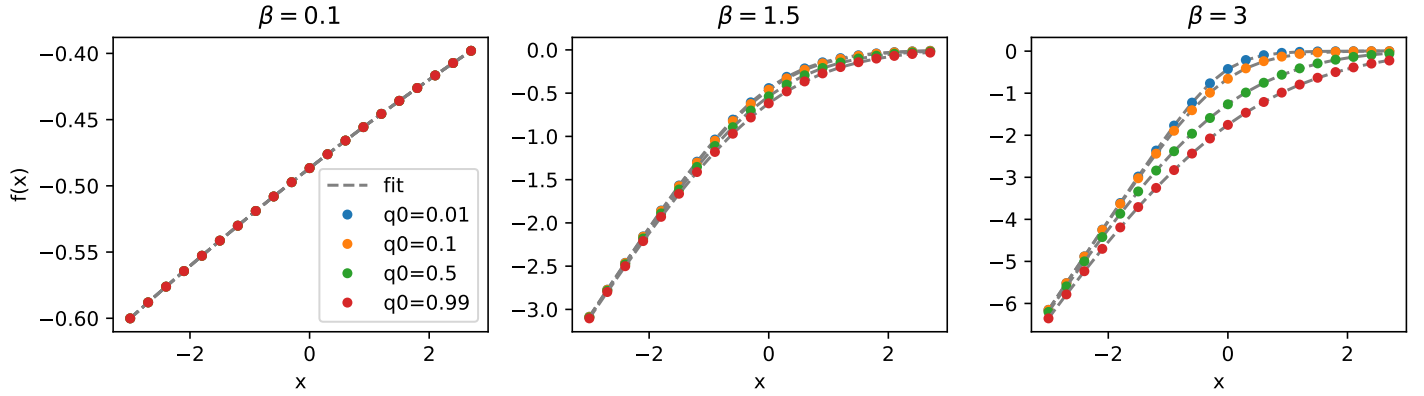


FIG. 1. Fit of Eq. (47) at different values of β and q_0 . The free parameters are b_0, w_0, b_1, w_1 . The points are numerical evaluations of the integrals, and the grey lines are the corresponding fit using the approximated functions. The results show the robustness of the approximation used.

IV. MONTE CARLO SAMPLING OF THE SK MODEL

In the main text, in the section results, in order to check the correlation between the weights of the first layer of the SK_{1RSB} architecture with the coupling of the Hamiltonian, 10,000 samples were generated through the Monte-Carlo Metropolis algorithm [4]. The samples are generated for a system with $N = 200$ variables at $\beta = 2$. Each configuration is sampled after $200 * N$ spin trail flips. I observe that, for the system under study, for this number of trail spin flips of the Metropolis algorithm the autocorrelation time between the configurations drops below 0.5. The training of the SK_{1RSB} was done minimizing, over the parameters of the SK_{1RSB} , the average of the negative log-likelihood (LL) of the generated m samples:

$$LL = -\frac{1}{m} \sum_m \log(Q_{1RSB}) \quad (49)$$

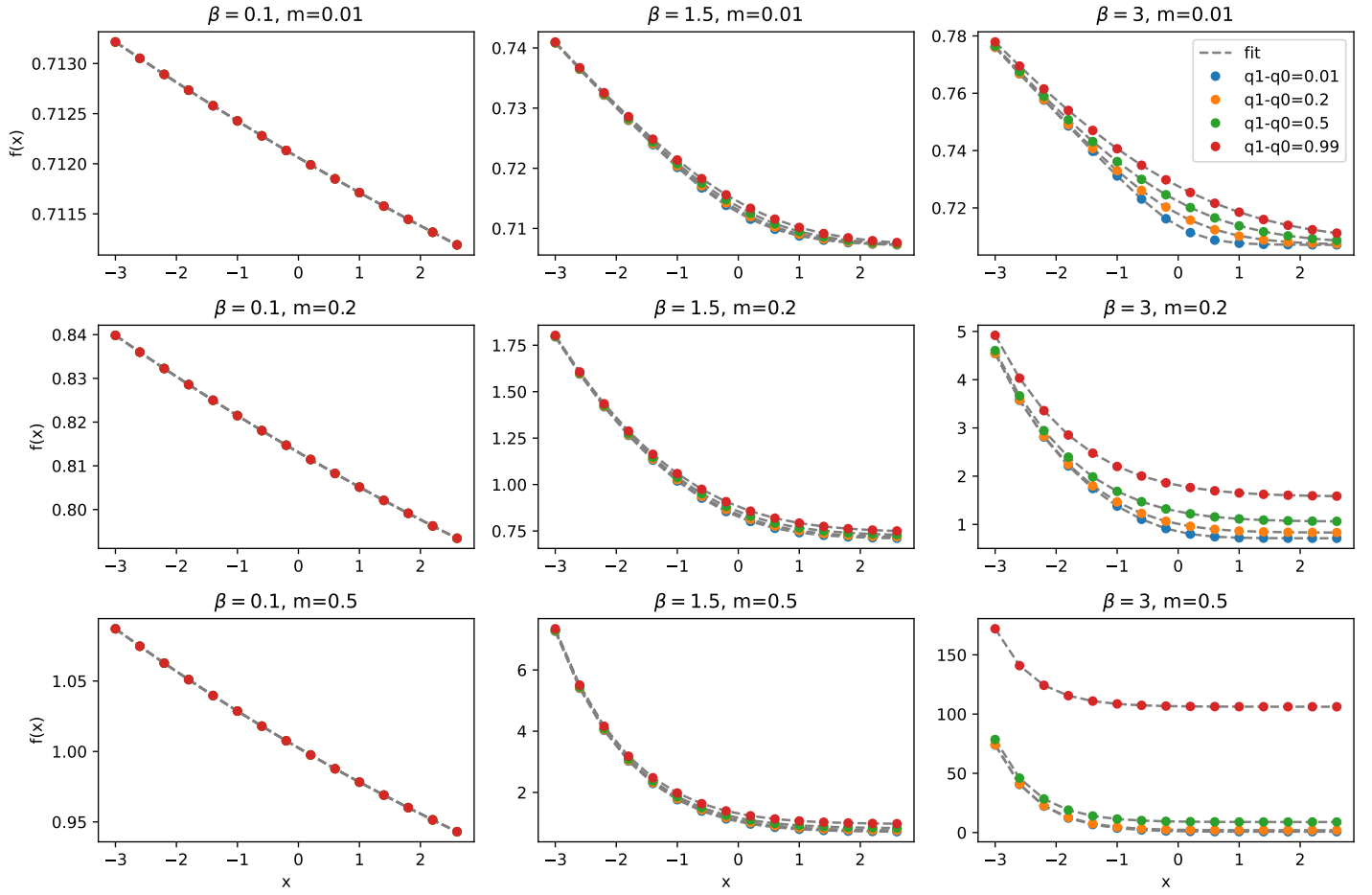


FIG. 2. Fit of Eq. (48) at different values of β , m , and $q_1 - q_0$. The free parameters are $b_0, w_0, b_1, w_1, b_2, w_2$. The points are numerical evaluations of the integrals, and the gray lines are the corresponding fit using the approximated functions. The results show the robustness of the approximation used.

-
- [1] L. P. Kadanoff, *Statistical physics: statics, dynamics and renormalization* (World Scientific, 2000).
 [2] H. Nishimori, *Statistical physics of spin glasses and information processing: an introduction* (Clarendon Press, 2001).
 [3] M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (1986).
 [4] J. E. Gubernatis, Marshall rosenbluth and the metropolis algorithm, *Physics of Plasmas* **12**, 057303 (2005), <https://doi.org/10.1063/1.1887186>.