Sparsely Encoded Hopfield Model with Unit Replacement

SUMMARY We investigate a sparsely encoded Hopfield model with unit replacement by using a statistical mechanical method called self-consistent signal-to-noise analysis. We theoretically obtain a relation between the storage capacity and the number of replacement units for each sparseness $a$. Moreover, we compare the unit replacement model with the forgetting model in terms of the network storage capacity. The results show that the unit replacement model has a finite value of the optimal sparseness on an open interval $0 < a < 1$ (the limit of sparseness) to maximize the storage capacity for a large number of replacement units, although the forgetting model does not.

key words: neurogenesis, associative memory, unit replacement, SCSNA, sparse coding, forgetting process

1. Introduction

It has been reported that neurogenesis continues to take place in certain regions of the postnatal brain including the hippocampal regions [1]–[4]. Neurogenesis is crucial to the formation of neural networks, and thus it affects neural network functions, especially memory formation [4]. This raises questions about how it affects memories in the brain. Should neurogenesis, which breaks the memory circuit, be a negative factor against memories embedded in the network? Previous computational studies have demonstrated the possibility that neurogenesis plays a role in maintaining recent memories embedded in the network [5]–[7].

The Hopfield model [8], a typical abstract network model of associative memory, has a critical number of memory patterns that can be stably stored: $0.138N$, where $N$ is the system size. If the number of embedded memory patterns is larger than $0.138N$, the memory state of this network model disappears due to overloading [8]–[13].

On the other hand, it has been demonstrated in numerical simulations that the Hopfield model with unit replacement, i.e., mathematically-modeled neurogenesis, can stably retrieve recently embedded memory patterns while avoiding overloading [5]. The reason for this ability is that the unit replacement procedure, in which a few old units are replaced with new ones at each learning step for embedding a new pattern, is equivalent to erasing the old memory traces gradually.

A previous study using statistical mechanical analyses indicated that $3.21$ is the minimum number of replacement units at each learning step to avoid overloading [6]. Furthermore, the optimal number of replacement units at each learning step to maximize the number of retrievable patterns is $6.95$. These critical numbers of replacement units are independent of the system size $N$.

The characteristics of the Hopfield model largely depend on how items are encoded in the pattern vectors to be stored. When most of the components of the encoded patterns to be stored are inactivated and only a small share of the components are activated, the encoding scheme is said to be sparse [14]. Many reports have described how the storage capacity of the Hopfield model can be dramatically improved by storing sparse patterns instead of conventional patterns, in which the mean firing rate is $1/2$ [14], [15], [18]. Note that the stored pattern used in the unit replacement model stated above is $1/2$ coding.

In this paper, we investigate the sparsely encoded Hopfield model with unit replacement by using self-consistent signal-to-noise analysis (SCSNA) [10], [11] and numerical simulations. We derive order parameter equations for this model through the SCSNA. First, we show the relationship between the storage capacity and the number of replacement units at each sparseness level. Second, we show the relationship between sparseness and the minimum number of replacement units at each learning step to avoid overloading. Third, we show the relationship between sparseness and the optimal number of replacement units at each learning step to maximize the storage capacity. Finally, we compare the analytical results obtained from this model with those from the forgetting model [18], which also avoids overloading.

The results show that the unit replacement model has a finite value of the optimal sparseness on an interval $(0, 1)$ to maximize the storage capacity for a large number of replacement units, whereas the forgetting model does not.

This paper is organized as follows: Sect. 2 presents a formulation of the unit replacement procedure in the sparsely encoded Hopfield model. Section 3 presents the order parameter equations of the unit replacement model derived by the SCSNA, and Sect. 4 explains the numerical simulation methodology of the unit replacement procedure. Section 5 shows the phase diagrams obtained from the order parameter equations and the results of the numerical simulations, and Sect. 6 compares the unit replacement model with the forgetting model. Section 7 gives our conclusions.
2. Unit Replacement Procedure in the Sparsely Encoded Network

We begin by formulating a recurrent neural network with \( N \) units and \( N(N-1) \) synapses. We shall use a synchronous updating rule in the form of the following network dynamics:

\[
s_i(t+1) = F[h_i + \theta] = \text{sgn}[h_i + \theta] - b, \tag{1}
\]

\[
h_i = \sum_{j=1}^{N} J_{ij} s_j(t), \tag{2}
\]

where \( s_i(t) \) is the state of unit \( i \) at discrete time \( t \), and \( h_i \) is the local field of unit \( i \). The sparseness parameters \( b \) and \( \theta \) are called shift and threshold, respectively [15]. \( J_{ij} \) denotes the strength of the synapse connecting unit \( j \) to \( i \), and it is symmetrical, i.e. \( J_{ij} = J_{ji} \). Note that the network has no self-interaction, i.e., \( J_{ii} = 0 \). We take the sign function to be

\[
\text{sgn}[x] = \begin{cases} 
1 & x \geq 0, \\
-1 & \text{otherwise}. 
\end{cases} \tag{3}
\]

Each element \( \xi^i_\mu \) of the \( \mu \)-th memory pattern \( \xi^i \), which is stored in the present model, takes \( \pm 1 \), and is generated independently with the probability,

\[
\text{Prob}[\xi^i_\mu = \pm 1] = \frac{1 \pm a}{2}. \tag{4}
\]

The sparseness parameter \( a \), called bias, denotes the average of \( \xi^i_\mu \), i.e. \( E[\xi^i_\mu] = a \) [15]. The encoding scheme of the memory patterns is called sparse coding when the sparseness parameters \( a \neq 0, b \neq 0, \) and \( \theta \neq 0 \) [14, 15].

We will consider the following unit replacement procedure. \( R \) units are selected randomly and independently from \( N \) units and replaced with new ones at each learning step \( \mu \) for embedding a new memory pattern. Each of the new replacement units rebuilds the synapses connecting all other units. However, it is difficult to deal with the replacement process of the unit replacement procedure mathematically. For the purpose of mathematical tractability, instead of the unit replacement procedure, we use a backward unit elimination procedure in which \( R \) units are randomly and independently selected from \( N \) units and eliminated at each backward learning step. Note that there is no replacement process in the backward unit elimination procedure. The backward unit elimination procedure is effectively equivalent to the unit replacement procedure in the limit of infinite learning step as discussed below. We formulate the learning rule effectively incorporating the backward unit elimination procedure as the following,

\[
J_{ij} = \frac{1}{N(1-a^2)} \sum_{\mu=1}^{\infty} u[\tau_i - \mu](\xi^i_\mu - a) u[\tau_j - \mu](\xi^j_\mu - a), \tag{5}
\]

\[
u[x] = \begin{cases} 
1 & x \geq 0, \\
0 & \text{otherwise}. 
\end{cases} \tag{6}
\]

The index \( \mu \) of \( \xi^i_\mu \) denotes the learning step, which is a backward time running in a direction from present to past. Thus, a pattern with small \( \mu \) is newer than one with large \( \mu \). In Eq. (5), an infinite number of patterns are embedded, and \( \mu = 1 \) corresponds to the latest stored pattern. The variable \( \tau_i \) is the elimination time of unit \( i \), which is also the backward time.

It has been elucidated that the storage capacity of the sparsely encoded Hopfield model is improved by subtracting the bias \( a \) from the pattern \( \xi^i_\mu \) in covariance learning [14, 15, 18], as described by Eq. (5). Thus, we employed bias cancellation in this learning rule. Note that Eq. (5) is equivalent to the learning rule of the previous model [6] when \( a = 0 \).

Here, \( \tau_i \), the elimination time of unit \( i \), is assigned to a random and independent number with the following probability:

\[
P(\tau_i) = \frac{R}{N} \left( 1 - \frac{R}{N} \right)^{\tau_i-1}, \tag{7}
\]

where \( R \) denotes the number of elimination units at each backward learning step, as described above. If \( \mu \) is larger than \( \tau_i \), \( u[\tau_i - \mu] \) becomes zero. Therefore, memory patterns older than the elimination time of unit \( i \) or unit \( j \) are erased from the synaptic strengths \( J_{ij} \) and \( J_{ji} \). Note that the backward unit elimination procedure defined by Eqs. (5) and (7) is mathematically equivalent to the unit replacement procedure in the limit of infinite learning step. The elimination time, \( \tau_i \), of the backward unit elimination procedure corresponds to the latest elimination time of the unit replacement procedure, and a stationary distribution of the number of embedded memory patterns in the unit replacement procedure is equal to Eq. (7), because the unit replacement procedure erases memory patterns embedded before units are replaced and the random unit selection of this procedure is memoryless.

Next, we define a time step normalized to the system size \( N \): \( \tau^*_i = \tau_i/N \). The normalized time step \( \tau^*_i \) approaches a continuous time variable in the limit \( N \rightarrow \infty \). Furthermore, Eq. (7) becomes the probability density,

\[
P(\tau^*_i) = R \exp[-R\tau^*_i], \tag{8}
\]

We evaluate the average of \( u[\tau_i - \mu] \), which determines the contribution of each memory pattern to the synaptic strengths. This is the average of \( u[\tau_i - \mu] \) over the probability Eq. (7):

\[
E[u[\tau_i - \nu]] = \lim_{N \rightarrow \infty} \sum_{\tau_i=1}^{\infty} P(\tau_i) \cdot u[\tau_i - \nu] = \int_0^{\infty} d\tau^*_i R \exp[-R\tau^*_i] \cdot u[\tau^*_i - \nu^*] = \exp[-R\nu^*]. \tag{9}
\]

where \( \nu^* = \nu/N \), as defined above. The average of \( u[\tau_i - \mu] \) exponentially decays as the number \( \mu \) of learning steps increases, and the attenuation coefficient is equal to the number of replacement units, \( R \).
For the purpose of mathematical tractability and simplicity, we assume that the time-scale of the learning rule defined by Eq. (5) was far different from that of network dynamics described by Eq. (1). Under the assumption of time-scale separation, the synaptic weights are firstly determined in accordance with Eq. (5), and are fixed during the memory retrieval process. We analytically evaluate the quality of memory retrieval in the network with the fixed synaptic weights.

3. SCSNA

The order parameter \( m^\mu \), which denotes the overlap between the stationary system state \( s \) and the memory pattern \( \xi^\mu \), is defined as

\[
m^\mu = \frac{1}{N(1-a^2)} \sum_{j=1}^{N} u[\tau_j - \mu](\xi^\mu_j - a)s_j.
\]

(10)

The overlap \( m^\mu \) indicates how correctly the memory pattern \( \xi^\mu \) could be retrieved from the network.

We can rewrite the local field \( h_i \) defined in Eq. (2) with the overlap \( m^\mu \) in the following form:

\[
h_i = \sum_{\mu=1}^{\infty} u[\tau_i - \mu](\xi^\mu_i - a)m^\mu - \frac{1}{N(1-a^2)} \sum_{\mu=1}^{\infty} u[\tau_i - \mu](\xi^\mu_i - a)^2s_i,
\]

(11)

where the second term is the correction term for canceling the self-coupling \( J_{ii} \).

Henceforth we focus on the stationary system state \( s \) in which the configuration has appreciable overlap with the \( \mu \)-th (\( \mu = 1, 2, \cdots \)) stored pattern \( \xi^\mu \), i.e., \( m^\mu \sim O(1) \), and has little overlap with the other patterns \( \xi^\nu \) for \( \nu \neq \mu \), i.e., \( m^\nu \sim O(1/\sqrt{N}) \). In statistical mechanics, \( \xi^\mu \) (\( m^\mu \sim O(1) \)) is called the condensed pattern, and the others are called the uncondensed patterns. Accordingly, we can evaluate the contribution of the uncondensed patterns by using the SCSNA, and split the local field \( h_i \) into a pure effective local field and an Onsager reaction term (ORT) in a self-consistent manner. Finally, the order parameter equations \( m^\delta, \Gamma(\alpha), \) and \( q(\alpha) \) can be obtained from the pure effective local field (see the Appendix):

\[
m^\delta = \frac{1}{2} \int_{0}^{\infty} dtP(\tau)\left[ \text{erf}(\Phi_1) + \text{erf}(\Phi_2) \right],
\]

(12)

\[
\Gamma(\alpha) = \int_{0}^{\infty} dtP(\tau)\left[ \frac{1+a}{\sqrt{2\pi}\sigma(\tau)} \exp(-\Phi_1^2) + \frac{1-a}{\sqrt{2\pi}\sigma(\tau)} \exp(-\Phi_2^2) \right],
\]

(13)

\[
q(\alpha) = \int_{0}^{\infty} dtP(\tau)\left[ 1 + b^2 - b[(1+a)\text{erf}(\Phi_1) - (1-a)\text{erf}(\Phi_2)] \right],
\]

(14)

\[
\sigma^2(\tau) = \int_{0}^{\tau} da \frac{q(\alpha)}{(1-\Gamma(\alpha))^2},
\]

(15)

\[
\Lambda(\tau) = \int_{0}^{\tau} da \frac{\Gamma(\alpha)}{1-\Gamma(\alpha)},
\]

(16)

\[
\Phi_1 = \frac{(1-a)u[\tau - \psi]m^\delta - \Lambda(\tau)b + \theta}{\sqrt{2}\sigma(\tau)},
\]

(17)

\[
\Phi_2 = \frac{(1+a)u[\tau - \psi]m^\delta + \Lambda(\tau)b - \theta}{\sqrt{2}\sigma(\tau)},
\]

(18)

where \( R \) is the number of replacement units at each learning step. \( \psi, \alpha, \) and \( \tau \) denote time steps normalized to the system size \( N \), as defined above. \( P(\tau) \) is the probability density given by Eq. (8).

The quantity \( m^\delta \) shows the overlap with the condensed pattern \( \xi^\mu \) in the retrieval phase. The quantity \( \Gamma(\alpha) \) corresponds to the susceptibility, which is a measure of the sensitivity to external fields. The quantity \( q(\alpha) \) is the so-called Edwards-Anderson order parameter. \( \sigma^2(\tau) \) denotes the variance of the pure effective local field, and \( \Lambda(\tau) \) represents the coefficient of the effective self-coupling term corresponding to the ORT. Note that \( \sigma^2(\tau) \) and \( \Lambda(\tau) \) are functions of the replacement time \( \tau \), and \( \Gamma(\alpha) \) and \( q(\alpha) \) are functions of the learning step \( \alpha \).

If the output function is assumed to have the form of Eq. (1), the variance of the cross-talk noise, which prevents retrieval of memory, is minimized at \( \theta = b = a \) [15, 18]. Therefore, we set the sparseness parameters to \( \theta = b = a \) in the following analyses.

4. Numerical Simulations

We verified the results obtained from the order parameter equations by using the numerical simulations. The stochastic unit replacement procedure defined in Sect. 2 was implemented in the numerical simulation as follows: We can write the average number \( \rho \) of units which have been replaced since the time step \( \mu \), using the probability of replacement units (Eq. (7)) as

\[
\rho = R \sum_{\nu=1}^{\mu} \left( \frac{1-R}{N} \right)^{\nu-1}, \quad (R \leq \rho < N).
\]

(19)

Every time the network memorizes a new pattern, \( \xi^\mu_i \) is assigned 0 in ascending order (\( 1 \leq i \leq \rho \)), among \( N \) elements of the memory pattern vector \( \xi^\mu \). The other (\( N - \rho \)) elements are assigned \( \pm 1 \) by the probability \( \text{Prob}[\xi^\mu_i = \pm 1] = \frac{1-a}{2} \) (Eq. (6)).

All the numerical simulations in this paper were carried out for \( N = 10000 \).

5. Results

First, let us investigate how many old patterns are retrievable in the network. Because the unit replacement procedure implies erasing the old memory traces gradually, the number of retrievable patterns is limited. Here, the storage capacity
$\psi_c$ is defined as the maximum number of retrievable patterns normalized to the system size $N$. Figure 1 (I), i.e., the analytical results obtained from the order parameter equations in Sect. 3, shows the storage capacity $\psi_c$ as a function of the number of replacement units, $R$, for each sparseness $a$ from 0 (i.e., 1/2 cording) to 0.8. If the normalized learning step $\psi$ is smaller than the storage capacity ($\psi < \psi_c$), the memory retrieval state ($m^w \neq 0, q \neq 0$) is stable. On the other hand, if $\psi > \psi_c$, the spin glass state ($m^w = 0, q \neq 0$) appears.

The minimum number of replacement units to keep the network in the retrieval phase is a point at the intersection of the abscissa with each curve in Fig. 1 (I). This point corresponds to the minimum number of replacement units, $R_c$, to avoid overloading. Figure 1 (II) shows the relation between sparseness $a$ and the minimum value $R_c$. As $a$ increases, $R_c$ decreases.

Moreover, there is an optimal number of replacement units, $R_{opt}$, to maximize the storage capacity $\psi_c$, as shown in Fig. 1 (I). Figure 1 (III) plots the optimal value $R_{opt}$ as a function of sparseness $a$. There is a peak to $R_{opt}$ ($\approx 7.6$) around $a \approx 0.35$ in Fig. 1 (III).

To evaluate the reliability of the approximate solutions of the order parameters obtained by the SCBSNA, we compare the values obtained from the order parameter equations with those from the numerical simulation. Figure 2 plots the overlap between stationary system state and memory pattern $\xi^1$ (ordinate) as a function of $R$ (abscissa) at each sparseness $a$. (I) $a = 0.2$. (II) $a = 0.4$. (III) $a = 0.6$. (IV) $a = 0.8$. Solid curves were obtained from the order parameter equations of the unit replacement model when $\psi = 0$. Dotted lines and error bars respectively show averages and standard deviations of 10 sampled overlaps of the numerical simulations ($N = 10000$).
Fig. 3  Overlap $m^\psi$ as a function of the normalized learning step $\psi$ ($= \mu/N$) (abscissa) at each sparseness $a$. We chose an integer number of replacement units, $R_{\text{int}}$, which is proximate to $R_{\text{opt}}$, shown in Figs. 1 (I) and (II). (I) $a = 0.2$, $R = 7$. (II) $a = 0.4$, $R = 7$. (III) $a = 0.6$, $R = 5$. (IV) $a = 0.8$, $R = 3$. Solid curves were obtained from the order parameter equations of the unit replacement model. Dotted lines and error bars respectively show averages and standard deviations of 10 sampled overlaps with the numerical simulations ($N = 10\,000$).

overlap between the most recently learned pattern $\xi^1$ and the stationary system state, as a function of $R$. For each sparseness $a$, the values of the overlap obtained by the numerical simulations are in good agreement with those of the order parameter equations in the memory retrieval state.

Figure 3 shows the overlap $m^\psi$ as a function of the normalized learning step $\psi$. For each sparseness $a$ and number of replacement units, $R$, the theoretical solutions in the memory retrieval state coincide with those of the numerical simulations. Moreover, in the numerical simulation results in Figs. 1 (II) and (III), the standard deviation of the sampled overlaps changes around the theoretically predicted transition point. Therefore, the transition point of the theoretical solution also coincides with those of the numerical simulation.

6. Comparison with the Forgetting Model

Reportedly, the Hopfield model with the forgetting process also avoids overloading and maintains recent memories, even if the number of embedded memory patterns is infinite [16]–[18]. The synaptic strength $J_{ij}$ of the sparsely encoded model with synaptic decay is determined from the following learning rule [18]:

$$J_{ij} = \frac{1}{N(1-a^2)} \sum_{\mu=1}^{\infty} \exp\left(-\frac{R}{N} \mu^2 \right) \left(\xi_{\mu i} - a\right) \left(\xi_{\mu j} - a\right),$$  (20)

where $R$ denotes the forgetting rate.

In this case, the equations relating the order parameters $m^\psi$, $\Gamma$, and $q$ can be obtained by using SCSNA [18].

$$m^\psi = \frac{1}{2} \left[ \text{erf}(\Phi_1) + \text{erf}(\Phi_2) \right],$$  (21)

$$\Gamma = \frac{1 + a}{\sqrt{2\pi}\sigma} \exp(-\Phi_1^2) + \frac{1 - a}{\sqrt{2\pi}\sigma} \exp(-\Phi_2^2),$$  (22)

$$q = 1 + b^2 - b \left[ (1+a)\text{erf}(\Phi_1) - (1-a)\text{erf}(\Phi_2) \right],$$  (23)

$$\sigma^2 = \int_0^{\infty} d\alpha \frac{\exp(-2R\alpha)q}{(1-\exp(-R\alpha))\Gamma^2},$$  (24)

$$\Lambda = \int_0^{\infty} d\alpha \frac{\exp(-2R\alpha)\Gamma}{1-\exp(-R\alpha)\Gamma},$$  (25)

$$\Phi_1 = \frac{(1-a)\exp(-R\psi)m^\psi - \Lambda b + \theta}{\sqrt{2}\sigma},$$  (26)

$$\Phi_2 = \frac{(1+a)\exp(-R\psi)m^\psi + \Lambda b - \theta}{\sqrt{2}\sigma}.$$  (27)

The quantity $m^\psi$ shows the overlap with the condensed pattern $\xi^\mu$ in the retrieval phase. The quantity $\Gamma$ corresponds to the susceptibility, which is a measure of the sensitivity to
because the average of should be similar to those of the unit replacement procedure, models are very di...show that the critical properties of these two (Eq. 9). However, the phase diagrams of the forgetting R...distribution of each memory pattern to the synaptic strengths, q external fields. The quantity q is the Edwards-Anderson order parameter. σ^2 denotes the variance of the pure effective local field, and Λ represents the coefficient of the ORT.

The macroscopic properties of the forgetting process should be similar to those of the unit replacement procedure, because the average of u[τ_i − ν], which determines the contribution of each memory pattern to the synaptic strengths, exponentially decays with the number of replacement units, R (Eq. 9). However, the phase diagrams of the forgetting model actually show that the critical properties of these two models are very different from each other.

Figure 4 shows the critical properties of the forgetting model, whose graphs correspond to those of Fig. 1. According to Figs. 1 (I) and 4 (I), the forgetting model has a larger storage capacity than the unit replacement model has. Moreover, in Figs. 1 (II) and 4 (II), the peak of R_{opt} at a ≈ 0.35 that had been confirmed to exist in the unit replacement model does not exist in the forgetting model.

Furthermore, as shown in Figs. 1 (I) and 4 (I), the unit replacement model has the different relationship between sparseness a and storage capacity ψ_c from that of the forgetting model when R > 10. Figure 5 illustrates the storage capacities of both models as a function of sparseness a at R = 20. The unit replacement model has a finite optimal sparseness (a ≈ 0.4) on an open interval 0 < a < 1 to maximize the storage capacity ψ_c, whereas, ψ_c of the forgetting model at R = 20 decreases monotonically as a increases to approximately 0.8. Note that in both models ψ_c increases as a approaches 1, reflecting that the sparsely encoded networks have an extremely large storage capacity, which diverges (ψ_c → ∞) in the limit a → 1 [14],[15].

7. Conclusions

We investigated the properties of the sparsely encoded Hopfield model with unit replacement by performing SCSNA and numerical simulations. In the case of a = 0, the obtained SCSNA order-parameter equation is completely equal to that of the previous work [6]. The SCSNA and numerical simulations revealed that the Hopfield model with unit replacement can stably retrieve recently embedded memory patterns while avoiding overloading.

We analyzed the relationship between the storage capacity ψ_c and the number of replacement units, R, at each sparseness. Second, we showed the relation between sparseness a and the minimum number of replacement units, R_c, at each learning step to avoid overloading. As sparseness a increases, the minimum number R_c decreases. Third, we showed the relation between sparseness a and the optimal number of replacement units, R_{opt}, at each learning step to maximize the storage capacity ψ_c. R_{opt} has a peak (≈ 7.6) around a ≈ 0.35. In other words, the sparsely encoded model around a ≈ 0.4 requires a slightly larger R than at a = 0 (i.e., 1/2 coding) to maximize the storage capacity. Finally, we compared the storage capacity ψ_c of the forgetting model with that of the unit replacement model. On the whole, the unit replacement model has a lower storage capacity ψ_c than that of the forgetting model. Furthermore, the unit replacement model has a finite value of the optimal sparseness on an open interval 0 < a < 1 to maximize the storage capacity at large values of R, while the forgetting model does not. The existence of the finite value of the optimal sparseness to maximize the storage capacity on the open interval (0, 1) is a nontrivial phenomenon that was found by numerically solving the order parameter Eqs. (12)-(18).

We surmise that these different properties come from the difference in the local field distributions in these two models, as previously reported [6]. Whereas the local field of the forgetting model obeys a single Gaussian, the local field distribution of the unit replacement model is a mixture...
of Gaussians (MOG). Each Gaussian of the MOG corresponds to a subgroup consisting of units replaced at the same time, and thus, the order parameter Eqs. (12)-(14) have a multiscale structure reflecting the MOG local field [6]. This is a notable difference between these two models.

Acknowledgements

This work was supported by KAKENHI 23500375.

References


Appendix

Derivation of the Order Parameter Equations by the SCSNA

Here, we estimate the pure effective local field in the framework of the SCSNA. In this analysis, we will focus on the retrieval phase in which the system state has appreciable overlap with the condensed pattern $\xi^\mu$, i.e., $m^\mu = O(1)$, and has little overlap with the uncondensed pattern $\xi^\nu$, for $\nu \neq \mu$, i.e., $m^\nu = O(1/\sqrt{N})$. Note that the solutions of the SCSNA based on a Gaussian ansatz for the local field are consistent with those of the replica method based on a replica symmetric ansatz [9]–[11].

In the first step of the SCSNA, we split the local field into a signal part (the first term) and a cross-talk noise part (the second term):

$$h_i = u \left[ \tau_i - \mu \right] (\xi_i^\mu - \alpha) m^\mu + \sum_{\nu \neq \mu} u \left[ \tau_j - \nu \right] (\xi_j^\nu - \alpha) m^\nu$$

$$- \frac{1}{N(1-a^2)} \sum_{\nu=1}^{\infty} u \left[ \tau_j - \nu \right] (\xi_j^\nu - \alpha)^2 s_i. \quad (A-1)$$

The third term is for canceling the self-coupling.

In the next step, we assume that the second term in Eq. (A-1) can be split into a Gaussian random variable and an Onsager reaction term (ORT). The overlap $m^\nu (\nu \neq \mu)$ defined in Eq. (10) can be rewritten in the following form by performing a Taylor series expansion around $m^\nu = O(1/\sqrt{N})$,

$$m^\nu = \frac{1}{N(1-a^2)} \sum_{j=1}^{N} u \left[ \tau_j - \nu \right] (\xi_j^\nu - \alpha) F(h_i^{(\nu)} + \theta)$$

$$- \frac{1}{N(1-a^2)} \sum_{j=1}^{N} u \left[ \tau_j - \nu \right] (\xi_j^\nu - \alpha)^2 F'(h_i^{(\nu)} + \theta), \quad (A-2)$$

where $h_i^{(\nu)}$ is the cavity field [19] defined by

$$h_i^{(\nu)} = u \left[ \tau_i - \mu \right] (\xi_i^\mu - \alpha) m^\mu + \sum_{\eta \neq \mu, \nu} u \left[ \tau_i - \eta \right] (\xi_i^\eta - \alpha) m^\eta$$

$$- \frac{1}{N(1-a^2)} \sum_{j=1}^{\infty} u \left[ \tau_i - \eta \right] (\xi_j^\eta - \alpha)^2 s_i, \quad (A-3)$$

and we define another order parameter $\Gamma(\nu)$ corresponding to the susceptibility,

$$\Gamma(\nu) = \frac{1}{N(1-a^2)} \sum_{j=1}^{N} u \left[ \tau_j - \nu \right] (\xi_j^\nu - \alpha)^2 F'(h_i^{(\nu)} + \theta). \quad (A-4)$$

The susceptibility $\Gamma(\nu)$ measures the sensitivity of the unit output to the external input. By substituting Eqs. (A-2) and (A-4) into the cross-talk noise of Eq. (A-1), we split the cross-talk noise into a Gaussian random variable and the ORT, as follows:
The average and variance of the ORT, and the coefficient $\Lambda_i(\tau_i)$ of unit $i$ is defined as

$$\Lambda_i(\tau_i) = \frac{1}{N(1-a^2)} \sum_{v=1}^\infty \sum_{j=1}^N \frac{u[\tau_j - \nu]\xi_v}{1 - \Gamma(\nu)} F[h_j^{(y)} + \theta],$$

where $z_i$ in the second term of Eq. (A.5) is a Gaussian random variable and the third term of Eq. (A.5) is the ORT. Note that the cavity field $h_j^{(y)}$ is not correlated with the memory pattern $\xi_v$ because $h_j^{(y)}$ does not contain $\xi_v$. Therefore, $z_i$ is a pure Gaussian random variable in the limit $N \to \infty$. The average and variance of $z_i$ become

$$E[z_i] = 0,$$

$$E[z_i^2] = \sigma^2(\tau_i) = \frac{1}{N} \sum_{v=\mu}^{\infty} \sum_{j=1}^N \frac{u[\tau_j - \nu]q(\nu)}{(1 - \Gamma(\nu))^2},$$

where $q(\nu)$ is the Edwards-Anderson order parameter,

$$q(\nu) = \frac{1}{N} \sum_{j=1}^N u[\tau_j - \nu] F^2[h_j^{(y)} + \theta].$$

Here, we can use a graphical method to evaluate the pure effective local field $\tilde{h}_i$, which renormalizes the ORT. Figure A.1 plots the following functions on the XY plane: $X = F(Y)$ and $Y = \theta + u[\tau - \mu]\xi_v \pm \sigma z + \Delta X$. The average and variance of $z_i$ become

$$E[z_i] = 0,$$

$$E[z_i^2] = \sigma^2(\tau_i) = \frac{1}{N} \sum_{v=\mu}^{\infty} \sum_{j=1}^N \frac{u[\tau_j - \nu]q(\nu)}{(1 - \Gamma(\nu))^2},$$

where $q(\nu)$ is the Edwards-Anderson order parameter,

$$q(\nu) = \frac{1}{N} \sum_{j=1}^N u[\tau_j - \nu] F^2[h_j^{(y)} + \theta].$$

Note that $\Lambda \geq 0$ because $F'[\xi] \geq 0$ and $\Gamma(\nu) < 1$. The three crossing points of the functions are fixed points of the state transition in a unit in a unit. One point is unstable, while the other two are stable. The size of each enclosed area (the gray areas in the figure) corresponds to the depth of the microscopic energy of each stable fixed point. According to the Maxwell rule, we select a stable fixed point with the largest enclosed area. In this case, the magnitude relation of two enclosed areas only depends on the cross point of the function $Y = \theta + u[\tau - \mu]\xi_v \pm \sigma z + \Delta X$ and the Y-axis. Therefore, we can determine the pure effective local field:

$$\tilde{h}_i = u[\tau_i - \mu]\xi_v \pm \sigma z_i - \Lambda_i(\tau_i) \theta.$$  (A.11)

Using the pure effective local field, we can safely rewrite the order parameters as

$$m^b = \frac{1}{N(1-a^2)} \sum_{j=1}^\infty u[\tau_j - \mu](\xi_v^b - \alpha) F[\tilde{h}_j + \theta],$$

$$\Gamma(\nu) = \frac{1}{N(1-a^2)} \sum_{j=1}^\infty u[\tau_j - \nu](\xi_v^b - \alpha)^2 F'[\tilde{h}_j + \theta],$$

$$q(\nu) = \frac{1}{N} \sum_{j=1}^\infty u[\tau_j - \nu](\xi_v^b - \alpha) F^2[\tilde{h}_j + \theta].$$

As mentioned in Sect. 2, we define the learning step normalized to the system size $N$ as $\psi = \mu/N$. In the limit $N \to \infty$, the normalized time step $\psi$ approaches a continuous time variable. Accordingly, we obtain

$$m^b = \frac{1}{1-a^2} \int d\tau \int d\psi P(\tau) \exp\left(-\frac{\psi^2}{2}\right) u[\tau - \psi]$$

$$\times \left\{ F'[\xi - \alpha] u[\tau - \psi] m^b + \sigma(\tau) z - \Lambda(\tau) b + \theta \right\}_{\xi},$$

$$\Gamma(\alpha) = \int d\tau \int d\psi P(\tau) \exp\left(-\frac{\psi^2}{2}\right) u[\tau - \alpha]$$

$$\times \left\{ F'[\xi - \alpha] u[\tau - \psi] m^b + \sigma(\tau) z - \Lambda(\tau) b + \theta \right\}_{\xi},$$

$$q(\alpha) = \int d\tau \int d\psi P(\tau) \exp\left(-\frac{\psi^2}{2}\right) u[\tau - \alpha]$$

$$\times \left\{ F'[\xi - \alpha] u[\tau - \psi] m^b + \sigma(\tau) z - \Lambda(\tau) b + \theta \right\}_{\xi},$$

$$\sigma^2(\tau) = \int_0^\infty da \frac{u[\tau - \alpha]}{1 - \Gamma(\alpha)} g(\alpha).$$

Here, we can safely replace the ensemble averaging $\frac{1}{N} \sum_{j=1}^\infty$ with Gaussian averaging $\int_0^\infty d\psi \int_0^{\infty} dz \exp\left(-\frac{\psi^2}{2}\right)$ and pattern averaging $\langle \cdot \rangle_{\xi}$, because $m^b$, $z_i$, and $\Lambda_i(\tau_i)$ in $\tilde{h}_i$ is independent of any microscopic configuration of the unit $i$. In conclusion, we can obtain the order parameter equations. In this way, our deviation process is complete in a self-consistent manner.
Ryota Miyata received the B.E. and M.E. degrees in control engineering from University of the Ryukyus, Okinawa, Japan, in 2009 and 2011, respectively. He is currently a doctoral student at Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology. His research interests include theoretical neuroscience and machine learning.

Koji Kurata received the B.S., M.S., and Ph.D. degrees in mathematical engineering from the University of Tokyo, Japan, in 1981, 1983, and 1990, respectively. From 1984 to 1990, he was a research associate at the University of Tokyo. From 1990 to 2000, he was an assistant professor at Osaka University. He is currently a professor at Faculty of Engineering, University of the Ryukyus. His research interests include computational aspects of neural networks, especially those of topological mapping organization. He is also interested in pattern formations in homogeneous fields.

Toru Aonishi received the B.E. degree in control engineering in 1993 from Kyusyu Institute of Technology, Fukuoka, Japan, and the M.E. and D.E. degrees in biophysical engineering from Osaka University, Osaka, Japan, in 1995 and 1998, respectively. He was a Researcher at RIKEN Brain Science Institute in Japan from 1998 to 2004. He is currently an associate professor at the graduate school of Tokyo Institute of Technology. His research focuses on non-linear dynamics, statistical mechanics and theoretical neuroscience.