# Auto-associative memory with two-stage dynamics of non-monotonic neurons*

Hiro-Fumi Yanai† and Shun-ichi Amari‡

**Abstract**

Dynamical properties of a neural auto-associative memory with two-stage neurons are investigated theoretically. The two-stage neuron is a model whose output is determined by a two-stage nonlinear function of the internal field of the neuron (internal field is a weighted sum of outputs of the other neurons). The model is general, including non-monotonic neurons as well as monotonic ones. Recent studies on associative memory revealed superiority of non-monotonic neurons to monotonic ones. The present paper supplies theoretical verification on the high performance of non-monotonic neurons, and proves that the capacity of the auto-associative memory with two-stage neurons is $O(n/\sqrt{\log n})$, in contrast to $O(n/\log n)$ of simple threshold neurons . There is also a discussion of recall processes, where the radius of basin of attraction of memorized patterns is clarified. Intuitive explanation on why the performance is improved by non-monotonic neurons is also provided by showing the correspondence of the recall processes of the two-stage-neuron net and orthogonal learning.

## 1 Introduction

The auto-associative memory model of the correlation type has been studied for long years [1], [3], [4], [8], [10], etc. and its performances have been analyzed theoretically (see [7]), although its exact analysis is still difficult. Various modifications of the prototype model have been proposed in order to increase the memory capacity. They are, for example, sparse encoding [2], [12], [17], introduction of hysteresis [22], [23], [24]. One interesting idea is introduction of non-momotonic neurons [9], [13], [14], [15].

It was Morita et al. (1990) [14] and Morita (1993) [13] who pointed out, by deep insight and computer simulations, that the capacity and attractivity of an analog model of neural auto-associative memory is enhanced by non-monotonic neurons. They also discussed a discrete version of the model, the "partial reverse method," a generalization of which is the two-stage neuron in this paper (This model uses a two-stage dynamics to realize the non-monotonic behavior of a single neuron). Stimulated by their work there appeared several theoretical papers on non-monotonic neural auto-associative memory. Yoshizawa et al. (1993) [26] showed by geometrical consideration on equilibrium point that a piecewise linear version of analog non-monotonic neuron net has a memory capacity of $0.4n$ (memory capacity is the maximum number of memory patterns stored as fixed points), where $n$ is the number of neurons. There is also an equilibrium point analysis of analog non-monotonic neuron net from physics background [20]. The capacity of a non-monotonic neuron is calculated by Kobayashi (1991) [9], concluding the capacity of over $10n$ (if the conventional neurons are used, the capacity is $2n$ [5]).

The partial reverse method by Morita (1993) [13] uses a two-stage dynamics to calculate the internal field, or stimulation potential, of each neuron, realizing the effect of non-monotonicity in the discrete time binary auto-associative model. It was pointed out that the autocorrelation matrix $W$ itself includes much more information capacity than the absolute capacity of $n/(2\log n)$ of McEliece et al. (1987) [11], or the relative capacity $0.14n$ of Amit et al. (1985) [4] or Amari & Maginu (1988) [3]. It was shown by computer simulation that the capacity becomes larger by changing the recall dynamics into the two-stage non-monotonic one. The present paper uses the statistical neurodynamical method [3], [15], [16], [22], [23], [25] to analyze the one-step recall dynamics of the two-stage dynamics model. We obtain the absolute capacity of one-step recall which is of the order of $n/\sqrt{\log n}$, fairly larger than $n/\log n$ of the conventional model. The basin of attraction is also analyzed, which is also drastically larger than the conventional model.

---

*To be published in *IEEE Transactions on Neural Networks*.

Part of this work has been presented at 1993 IEEE International Conference on Neural Networks [21]. The present work is supported in part by Grant-in-Aid for Scientific Research on Priority Areas on the Higher-Order Brain Functions from the Ministry of Education, Science and Culture of Japan, and by the Real World Computing Program, RWCP, Japan.

†H.-F. Yanai is with the Department of Information and Communication Engineering, Faculty of Engineering, Tamagawa University, Machida, Tokyo 194, Japan. E-mail: yanai@eng.tamagawa.ac.jp.

‡S. Amari is with the Department of Mathematical Engineering and Information Physics, Faculty of Engineering, University of Tokyo, Bunkyo-ku, Tokyo 113, Japan; The RIKEN Frontier Research Program, Brain Information Processing Group, Riken, Wako-city, Saitama 351-01, Japan.

# 2 Definition of the Model

## 2.1 Conventional Model of Associative Memory

The conventional auto-associative memory model is a neural net with recurrent connections. Let $W = (w_{ij})$ be its synaptic connection matrix, $w_{ij}$ being the synaptic weight from the $j$th to the $i$th neuron, where $w_{ij} = w_{ji}$ and $w_{ii} = 0$ are assumed. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ be a column vector representing the state of the network, where $x_i$ is the output of the $i$th neuron taking values on $\{-1, 1\}$. The net operates synchronously at discrete times, and the updated state $\boldsymbol{x}'$ is determined from $\boldsymbol{x}$ by

$$\begin{cases} \boldsymbol{u} = W\boldsymbol{x}, \\ \boldsymbol{x}' = \mathrm{sgn}(\boldsymbol{u}) \end{cases} \tag{1}$$

where the signum function sgn is operated componentwise, that is, $x'_i = 1$ when $u_i > 0$ and $x'_i = -1$ otherwise.

Let $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_\mu, \ldots, \boldsymbol{\xi}_m$ be $m$ state vectors which are to be memorized in the network as its equilibrium states and are to be recalled by the dynamics. The correlation type auto-associative memory model uses a Hebb-type rule to determine the connection matrix $W$ from the $m$ patterns $\boldsymbol{\xi}_\mu, \mu = 1, 2, \ldots, m$,

$$W = \frac{1}{n}\sum_{\mu=1}^{m} \boldsymbol{\xi}_\mu \boldsymbol{\xi}_\mu^T - \frac{m}{n} I$$

where $T$ denotes the transposition and $I$ is the unit matrix. In components, this is written as

$$w_{ij} = \frac{1}{n}\sum_{\mu=1}^{m} \xi_i^\mu \xi_j^\mu \ (i \neq j), \quad w_{ii} = 0$$

It is expected that $\boldsymbol{\xi}_\mu$'s are the equilibria of the dynamics,

$$\boldsymbol{\xi}_\mu = \mathrm{sgn}(W\boldsymbol{\xi}_\mu), \quad \mu = 1, 2, \ldots, m$$

or there exist a $\tilde{\boldsymbol{\xi}}_\mu$ in a small neighborhood of $\boldsymbol{\xi}_\mu$ such that

$$\tilde{\boldsymbol{\xi}}_\mu = \mathrm{sgn}(W\tilde{\boldsymbol{\xi}}_\mu),$$

and further that, given an initial state $\boldsymbol{x}(0)$ belonging to some neighborhood of $\boldsymbol{\xi}_\mu$, the memory pattern $\boldsymbol{\xi}_\mu$ itself or its neighbor $\tilde{\boldsymbol{\xi}}_\mu$ is recalled by the dynamics

$$\boldsymbol{x}(t+1) = \mathrm{sgn}(W\boldsymbol{x}(t)) \tag{2}$$

In order to analyze the characteristics of the model, it is assumed that patterns $\boldsymbol{\xi}_\mu$ are randomly generated subject to the probability distribution that all the $\xi_i^\mu$ are $+1$ or $-1$ independently with probability 0.5. There are a lot of statistical or statistical-physical researches on this model.

## 2.2 Two-Stage Recall Dynamics

It is known that the capacity of the conventional model is limited. It was suggested by Morita (1993) [13] that the dynamics Eq. (1) or (2) is not effective for making full use of information stored in the correlation matrix $W$. He proposed a non-monotonic dynamics or two-stage dynamics of state update in order to make efficient use of information in $W$. We generalize his idea in the following model. The updated state $\boldsymbol{x}'$ is determined from $\boldsymbol{x}$ by

$$\begin{cases} \tilde{\boldsymbol{u}} = W\{\boldsymbol{x} + f(W\boldsymbol{x})\}, \\ \boldsymbol{x}' = \mathrm{sgn}(\tilde{\boldsymbol{u}}) \end{cases} \tag{3}$$

where $f$ is, in general, a nonlinear function operating on $\boldsymbol{u} = W\boldsymbol{x}$ componentwise. In the following analysis, we assume that

$$f(-u) = -f(u)$$

to keep the symmetry between $+1$ and $-1$ of the components of $\boldsymbol{x}$. The original model of Morita (1993) [13], partial reverse method, is recovered by putting

$$f(u) = \begin{cases} c, & u < -h, \\ 0, & -h \le u \le h, \\ -c, & u > h \end{cases} \tag{4}$$

with $c \approx 2.7$ and $h \approx 1.9$. Eq. (4) represents the non-monotonic character of the dynamics such that, when $|u_i|$ is too large (larger than $h$ or smaller than $-h$), the effect of $x_i$ is reversed.

The Fundamental Lemma introduced in the next section is on the transformation of distance (i.e., noise reduction property), and is valid for any forms of the modification function $f$. However we specifically calculate the capacity and the basin of attraction of memorized patterns for an interesting class of piecewise linear functions: $f(u) = -au + c\,\mathrm{sgn}(u)$. The analytical results thus derived help us understand the importance of negatively sloped parts of the modification function. The negatively sloped part is the source of non-monotonic behavior of the two-stage neuron.

## 2.3 Reference to Analog-Type Models

There are papers which deal with dynamics of analog neural nets with non-monotonic neurons [20], [26]. They study models of the following type:

$$\begin{cases} \dfrac{d\boldsymbol{u}}{dt} = -\boldsymbol{u} + W g(\boldsymbol{u}) \\ \boldsymbol{x} = \mathrm{sgn}(\boldsymbol{u}) \end{cases} \tag{5}$$

where $g(\boldsymbol{u})$ is a non-monotonic response function. By a suitable discretization of the variables, the analog model has some correspondence with our two-stage model. But note that there is a (possibly essential) difference that in our two-stage model, states of the net in phase space are discrete (combinations of $+1$ and $-1$), so that there is no continuity of dynamical flows. On the other hand, in analog models, states are continuous, so that they have continuous flows. Another difference is the manner of identifying the outputs. In our model, internal field is always calculated from the 'outputs' of neurons. On the other hand, in the analog model, it is customary that a net evolves without using the formal outputs $x_i$ of neurons. That is, the first equation of Eq. (5) is used during evolution of the net, and outputs are converted to $\pm 1$ by some other observer based on the second equation of Eq. (5).

## 2.4 Overlap, Distance, and Loading Rate

Now let us define characteristic variables and parameters used in the analysis. The first one is the overlap of the state $\boldsymbol{x}(t)$ at time $t$ and the memorized pattern $\boldsymbol{\xi}_\mu$ to be recalled,

$$l_t^\mu = \frac{\boldsymbol{\xi}_\mu^T \boldsymbol{x}(t)}{n} = \frac{1}{n}\sum_{i=1}^{n} \xi_i^\mu x_i$$

The overlap takes values between $-1$ and $1$, in other words, it is the direction cosine between the state and the pattern. In the following we consider only positive overlaps. This is because, by the symmetry of the dynamics, if $\boldsymbol{x}(0)$ evolves like $\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots$, then $-\boldsymbol{x}(0)$ does like $-\boldsymbol{x}(1), -\boldsymbol{x}(2), \ldots$. As a measure of overlap we also use the (normalized Hamming) distance $d_t^\mu$. The distance is related to the overlap by $l_t^\mu = 1 - 2d_t^\mu$. If there is no ambiguity, we shall omit the superscript $\mu$ and time $t$ concerning the overlap or distance.

Since we discuss statistical properties in the large $n$ limit, the loading rate is used as a characteristic parameter which is defined by the number of memorized patterns divided by the number of neurons,

$$r = \frac{m}{n}$$

## 2.5 The Memory Capacity and the Basin of Attraction

We analyze the one-step recall dynamics. A memorized pattern $\boldsymbol{\xi}_\mu$ is said to be correctly recalled from initial state $\boldsymbol{x}$ via one-step dynamics when the updated state $\boldsymbol{x}'$ of $\boldsymbol{x}$ is equal to $\boldsymbol{\xi}_\mu$, that is,

$$\boldsymbol{\xi}_\mu = \mathrm{sgn}\left\{W\boldsymbol{x} + Wf(W\boldsymbol{x})\right\}$$

holds. The above defined distance $d$ of two states $\boldsymbol{x}$ and $\boldsymbol{\xi}$ is written as

$$d(\boldsymbol{x},\boldsymbol{\xi}) = \frac{1}{2n}\sum_{i=1}^{n}|x_i - \xi_i|$$

Let $N_D$ be the $D$-neighborhood of $\boldsymbol{\xi}_\mu$,

$$N_D(\boldsymbol{\xi}_\mu) = \{\boldsymbol{x} | d(\boldsymbol{x},\boldsymbol{\xi}_\mu) \le D\}$$

When $\boldsymbol{\xi}_\mu$ is recalled correctly from any $\boldsymbol{x} \in N_D(\boldsymbol{\xi}_\mu)$ via one-step dynamics, $N_D$ is said to be included in the basin of attraction of $\boldsymbol{\xi}_\mu$. Since $\boldsymbol{\xi}_\mu$ are generated stochastically, we can study the probability of $N_D$ belonging to the basin of attraction.

Let $D(m)$ be the largest value such that, when $n$ is sufficiently large, the probability of $N_{D(m)}(\boldsymbol{\xi}_\mu)$ belonging to the basin of attraction of $\boldsymbol{\xi}_\mu$ tends to 1 for any $\boldsymbol{\xi}_\mu, \mu = 1, 2, \ldots, m$. We call $D(m)$ the radius of basin of attraction via one-step dynamics, when the number of memorized patterns is $m$. The capacity $m_c$ is defined by the largest number of memorized patterns such that the net has a positive radius of attraction $D(m) > 0$. It should be noted that $r_c = m_c/n \to 0$ as $n$ tends to infinity so that we need to scale the quantity $D, m_c$, or $r_c$ with $n$ adequately in order to state our results in the rigorous sense.

# 3 Theory of One-Step Update Dynamics: Main Results

Even in the case of the conventional monotonic neurons (simple threshold neurons or sigmoidal neurons), it is practically impossible to analyze the exact dynamics of a correlation-type auto-associative neural net. This is because that the number of necessary variables explodes as time goes on [6]. Therefore, the analysis of the long-range behavior or equilibrium relies on some approximation schemes. A famous one is the replica symmetry approximation at the equilibrium state [4]. Another interesting scheme is a "self-consistent signal-to-noise analysis" developed by Shiino & Fukai (1992) [19]. Unlike the replica method, this method is applicable to neural nets with asymmetric synapses or non-monotonic neurons. Another way of tackling the complicated problem is to analyze the short-range transient dynamics, in particular, the analysis of one-step update [11], [12]. Actually, the optimal set of parameters in the one-step update is closely related to the multiple-step ones. To deal with long-range behaviors, Amari & Maginu (1988) [3] has tried to renormalize infinite-range interactions into two macroscopic variables, that is the overlap and the variance of interference noise, and succeeded in explaining the recall processes within a certain approximation. This scheme is further refined in [16].

Here we rely on the analysis of the one-step update. The one-step analysis here, however, involves essentially two steps since the neuron here is a two-stage one. We first show how a pattern approaches a memorized one by the one-step update under a general modification function. We then apply the result to the case of a piecewise linear function $f$ to obtain the capacity and the radius of basin of attraction. The results show drastic improvements in both the capacity and the basin of attraction.

**Fundamental Lemma (one-step synchronous update)** *When an initial state is apart from a memorized pattern by distance $d$, the distance of the updated state is described by*

$$d' = (1-d)\,\phi\left(\frac{L+B}{\sigma}\right) + d\,\phi\left(\frac{L-B}{\sigma}\right)$$

*where $r$ is the loading rate, and*

$$\begin{aligned}
L &= l + F_1(l), \\
B &= r\dot{F}_1(l), \\
\sigma^2 &= r\Big[1 + F_2(l) + 2lF_1(l)\dot{F}_1(l) + \dot{F}_1(l)^2 \\
&\qquad + 2\left(lF_1(l) + \dot{F}_1(l)\right)\Big].
\end{aligned}$$

*The functions are defined by*

$$\phi(u) = \int_{-\infty}^{-u} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}}\,dt,$$

$$
\begin{aligned}
F_p(l) &= \int_{-\infty}^{\infty} f(l+\sqrt{r}t)^p \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}}\, dt, \quad p = 1 or 2, \\
\dot{F}_1(l) &= \frac{dF_1(l)}{dl}.
\end{aligned}
$$

We apply the Fundamental Lemma to the piecewise linear modification functions of the following form.

$$
f(u) = -au - (1-2a)\mathrm{sgn}(u), \quad a > 0.
$$

**Main Theorem** *In the case of the piecewise linear modification function, the absolute capacity scales with $n$, independently of $a$, as*

$$
m_c = \frac{n}{\sqrt{2\log n}}.
$$

*When the loading rate is $100\rho\%(0 < \rho \leq 1)$ of the capacity, that is, the number of memorized patterns is $m = \rho m_c$, the radius of basin of attraction is asymptotically described by*

$$
D(\rho m_c) = \begin{cases} \dfrac{(1-\rho^2)a^2}{4\rho(1-a)^2} \dfrac{1}{\sqrt{2\log n}} & (a \neq 1) \\ \dfrac{1}{2}\sqrt{\dfrac{1-\rho^2}{\rho}} \left(\dfrac{1}{2\log n}\right)^{\frac{1}{4}} & (a = 1). \end{cases}
$$

The capacity $m_c$ is uniformly larger than the capacity

$$
\frac{n}{2\log n}
$$

of the conventional model. It is also immediate to show that $D(\rho m_c)$ tends to 0 with $n$ tending to infinity for $\rho > 0$, in compensation with the increasing $m_c$, unless we let $\rho \to 0$. However, by letting $\rho \to 0$, we can show that the radius $D(m)$ is much larger than that of the conventional model.

This drastic increase takes place even when the modification function is linear.

**Corollary** *For the linear modification function of the form $f(u) = -au$, the absolute capacity $m_c$ scales like*

$$
m_c = \begin{cases} \left(\dfrac{1-a}{1-2a}\right)^2 \dfrac{n}{2\log n} & (a \neq 1/2) \\ \dfrac{n}{\sqrt{2\log n}} & (a = 1/2) \end{cases}
$$

*as $n$ tends to infinity. For $0 < a < 2/3$, this is larger than the capacity $m_c = n/(2\log n)$ of the conventional model ($a = 0$).*

It is known that, for the conventional model, when the loading rate is $100\kappa\%(0 < \kappa \leq 1)$ of its capacity $n/(2\log n)$, that is, $m = \kappa n/(2\log n)$, the radius of basin of attraction is

$$
D = (1-\sqrt{\kappa})/2
$$

On the other hand, when the same number $m$ is stored in the two-stage model, i.e., $\rho \to 0$, $D(m)$ is larger than the above $D$. In fact, in the extreme case of $a = 1$, the radius of basin of attraction is $D(m) = 1/2$ for any $\kappa$, showing the superiority of the two-stage recall dynamics.

The next section is devoted to the proof of the Fundamental Lemma by using the method of statistical neurodynamics.

# 4 Proof of the Fundamental Lemma

## 4.1 Method of Calculation of the Distance After the Update

Let $\boldsymbol{x}$ be a given initial state from which a memorized pattern, say $\boldsymbol{\xi}_1$, is to be recalled by the one-step update of the state. In the conventional model, internal field

$$
\begin{aligned}
\boldsymbol{u} &= W\boldsymbol{x} \\
&= \left(\frac{1}{n}\sum_{\mu=1}^{m} \boldsymbol{\xi}_\mu \boldsymbol{\xi}_\mu^T - \frac{m}{n} I\right) \boldsymbol{x}
\end{aligned}
$$

is decomposed as

$$u_i = \xi_i^1 L + B x_i + \sigma\tilde{\varepsilon}_i$$

where $L$ is the signal

$$L = \frac{1}{n}\sum_{j\neq i} \xi_j^1 x_j$$

$B$ is the bias which is equal to zero in this case, and $\sigma\tilde{\varepsilon}_i$ is the noise due to the crosstalk of other memorized patterns,

$$\sigma\tilde{\varepsilon}_i = \frac{1}{n}\sum_{\mu\neq 1}\sum_{j\neq i} \xi_i^\mu \xi_j^\mu x_j$$

where $\tilde{\varepsilon}_i$ is a unit Gaussian noise subject to $N(0,1)$. The updated state $\boldsymbol{x}'$ is given by

$$\begin{aligned} x_i' &= \mathrm{sgn}(u_i) \\ &= \xi_i^1 \mathrm{sgn}(L + B\xi_i^1 x_i + \sigma\varepsilon_i) \end{aligned}$$

where $\varepsilon_i = \xi_i^1 \tilde{\varepsilon}_i$.

Likewise, in the case of the two-stage model, we have

$$\begin{aligned} \tilde{\boldsymbol{u}} &= \boldsymbol{u} + W f(\boldsymbol{u}) \\ &= W\boldsymbol{x} + W f(W\boldsymbol{x}) \end{aligned}$$

We can again decompose $\tilde{u}_i$ in the three terms,

$$\xi_i^1 \tilde{u}_i = L + B\xi_i^1 x_i + \sigma\varepsilon_i$$

by assuming that the crosstalk term is asymptotically subject to the Gaussian distribution. In general, the bias $B$ is not equal to zero. Calculation of $\sigma$ is complicated because of the nonlinear correlation of $W$ in $Wf(W\boldsymbol{x})$ term.

Once $L$, $B$ and $\sigma$ are calculated, it is rather straightforward to obtain the overlap of the updated state $\boldsymbol{x}'$ and a pattern $\boldsymbol{\xi}_1$ to be recalled, or their distance $d'$,

$$\begin{aligned} d' &= \frac{1}{2n}\sum_{i=1}^{n} |\xi_i^1 - x_i'| \\ &= \frac{1}{2n}\sum_{i=1}^{n}(1 - \xi_i^1 x_i') \\ &= \frac{1}{2}\left(1 - \frac{1}{n}\sum_{i=1}^{n} \xi_i^1 x_i'\right) \end{aligned}$$

Here the last summation is equal to the overlap $l'$ of $\boldsymbol{x}'$ and $\boldsymbol{\xi}_1$, and is rewritten as

$$\begin{aligned} l' &= \frac{1}{n}\sum_{i=1}^{n} \xi_i^1 x_i' \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{sgn}(L + B\xi_i^1 x_i + \sigma\varepsilon_i) \\ &= \frac{1}{n}\sum_{i\in I_+} \mathrm{sgn}(L + B + \sigma\varepsilon_i) + \frac{1}{n}\sum_{i\in I_-} \mathrm{sgn}(L - B + \sigma\varepsilon_i) \end{aligned}$$

where $I_+$ and $I_-$ are, respectively, the sets of index $i$ that satisfy

$$\begin{aligned} I_+ &= \{i | \xi_i^1 x_i = 1\} \\ I_- &= \{i | \xi_i^1 x_i = -1\} \end{aligned}$$

That is, the neurons in the "right" states belong to $I_+$ and "wrong" states to $I_-$. There are $n(1-d)$ elements in $I_+$, and $nd$ in $I_-$, where $d$ is the distance between the initial state $\boldsymbol{x}$ and $\boldsymbol{\xi}_1$. So the expectation of the overlap $l'$ is

$$\begin{aligned}
E[l'] &= (1-d)E[\operatorname{sgn}(L+B+\sigma\varepsilon)] \\
&\quad +d\, E[\operatorname{sgn}(L-B+\sigma\varepsilon)] \\
&= (1-d)(1-2\ \Pr\{L+B+\sigma\varepsilon<0\}) \\
&\quad +d\ (1-2\ \Pr\{L-B+\sigma\varepsilon<0\}) \\
&= 1-2\left\{(1-d)\phi\left(\frac{L+B}{\sigma}\right)+d\,\phi\left(\frac{L-B}{\sigma}\right)\right\}
\end{aligned}$$

where $\phi(u)$ is the integral

$$\phi(u)=\int_{-\infty}^{-u}\gamma(t)dt \tag{6}$$

with

$$\gamma(t)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right) \tag{7}$$

Thus the expectation of the transformed distance is

$$\begin{aligned}
E[d'] &= \frac{1}{2}(1-E[l']) \\
&= (1-d)\phi\left(\frac{L+B}{\sigma}\right)+d\phi\left(\frac{L-B}{\sigma}\right)
\end{aligned} \tag{8}$$

When $n$ is sufficiently large, the actual distance tends to the expectation (according to the law of large numbers).

## 4.2 Details of the Calculation

Now let us proceed to the proof in detail. There remains calculation of $L, B$ and $\sigma$. The internal field $\tilde{\boldsymbol{u}}$ is the sum of $W\boldsymbol{x}$ and $Wf(W\boldsymbol{x})$.

Putting, without loss of generality, $\boldsymbol{\xi}_1=(1,1,1,...,1)^T$, it is easy to see that (see for example [3] or [25])

$$(W\boldsymbol{x})_i=\frac{1}{n}\sum_{j=1}^{n}w_{ij}x_j=l+N_i \tag{9}$$

with

$$l=\frac{1}{n}\sum_{j\neq i}\xi_j^1x_j=\frac{1}{n}\sum_{j\neq i}x_j$$

and

$$N_i=\frac{1}{n}\sum_{\mu\neq 1}\sum_{j\neq i}\xi_i^\mu\xi_j^\mu x_j$$

where $N_i$ is a Gaussian noise with mean 0 and variance $(m-1)/n$. This consequence is guaranteed by the law of large numbers and the central limit theorem. Here, the variance $(m-1)/n\simeq m/n=r$ is the loading rate, a key parameter in the dynamical behaviors.

For general two-stage neurons, if we put

$$(Wf(W\boldsymbol{x}))_i=M_i$$

the internal field of the $i$th neuron is written as

$$\tilde{u}_i=l+N_i+M_i$$

where the expectation of $M_i$ is composed of the signal term and the bias term. Here, $M_i$ is a Gaussian noise and is correlated with $N_i$. To calculate $L, B$ and $\sigma$, we need to evaluate the expectation of $M_i$ and the variance of the noise term

$$
\begin{aligned}
\sigma^2 &= V[N_i + M_i] \\
&= V[N_i] + V[M_i] + 2cov(N_i, M_i) \\
&= E[N_i^2] + V[M_i] + 2E[N_i M_i] \qquad (10)
\end{aligned}
$$

Calculation of $\sigma^2$ is rather complicated since $M_i$ is (in general) nonlinearly correlated with $N_i$ through the synaptic weight $W$, i.e., the randomly generated patterns $\boldsymbol{\xi}_\mu$.

Now let us proceed to the calculation of the expectation of $\tilde{u}_i$.

$$
\begin{aligned}
\tilde{u}_i &= (W\boldsymbol{x} + Wf(W\boldsymbol{x}))_i \\
&= l + N_i + \frac{1}{n}\sum_{\mu=1}^{m}\sum_{j\neq i}\xi_i^\mu \xi_j^\mu f(l + N_j) \\
&= l + \frac{1}{n}\sum_{j\neq i} f(l + N_i) + N_i \\
&\qquad + \frac{1}{n}\sum_{\mu\neq 1}\sum_{j\neq i}\xi_i^\mu \xi_j^\mu f(l + N_j) \qquad (11)
\end{aligned}
$$

Hence the expectation of $\tilde{u}_i$ is

$$
l + \frac{n-1}{n}\langle f(l + N_j)\rangle + \frac{(m-1)(n-1)}{n}\left\langle \xi_i^\mu \xi_j^\mu f(l + N_j)\right\rangle
$$

and this tends to

$$
l + \langle f(l + N_j)\rangle + (m-1)\left\langle \xi_i^\mu \xi_j^\mu f(l + N_j)\right\rangle
$$

with $n$ tending to infinity, where the bracket $\langle\bullet\rangle$ represents averaging with respect to random variables $\boldsymbol{\xi}_\mu$. Since we can put $N_j = \sqrt{r}\varepsilon_j$, where $\varepsilon_j$'s are mutually independent Gaussian noises $N(0, 1)$, the expectation $\langle f(l + N_j)\rangle$ is

$$
\begin{aligned}
F_1(l) &= \left\langle f(l + \sqrt{r}\varepsilon)\right\rangle \\
&= \int_{-\infty}^{\infty} f(l + \sqrt{r}t)\frac{1}{\sqrt{2\pi}}e^{\frac{-t^2}{2}}\,dt
\end{aligned}
$$

and, by the law of large numbers, we can regard this expectation as the actual value of $(1/n)\sum_{j\neq i} f(l + N_j)$, the second term of Eq. (11).

Now let us calculate the expectation $\left\langle \xi_i^\mu \xi_j^\mu f(l + N_j)\right\rangle$. In order to show the correlation of $N_j$ with the coefficients $\xi_i^\mu$ and $\xi_j^\mu$, we separate $N_j$ into two terms, one correlated with $\xi_i^\mu \xi_j^\mu$, and the other uncorrelated with it. We then write it as

$$
N_j = \frac{1}{n}\xi_j^\mu \xi_i^\mu x_i + N_j'
$$

Note that $N_j'$ can be written as $\sqrt{r}\varepsilon_j$ like $N_j$ because $\xi_j^\mu \xi_i^\mu x_i/n$ vanishes as $n$ tends to infinity. Also for simplicity's sake we adopt a notation $\alpha = \xi_j^\mu$ and $\beta = \xi_i^\mu$. Then the expectation is

$$
\langle \alpha\beta f(l + N_j)\rangle = \left\langle \alpha\beta f\left(l + \frac{\alpha\beta x_i}{n} + \sqrt{r}\varepsilon_j\right)\right\rangle
$$

We first take the expectation with respect to $\varepsilon_j$ where $\alpha$ and $\beta$ are fixed, giving

$$
\alpha\beta F_1\left(l + \frac{\alpha\beta x_i}{n}\right) \simeq \alpha\beta\left\{F_1(l) + \frac{\alpha\beta x_i}{n}\dot{F}_1(l)\right\}
$$

where the dot means differentiation by $l$. We then take the expectation with respect to $\alpha$ and $\beta$, yielding

$$
\frac{x_i}{n}\dot{F}_1(l)
$$

Thus, from Eq. (11), the bias in $\tilde{u}_i$, in the limit of large $n$, is

$$E[\tilde{u}_i] = l + F_1(l) + r\dot{F}_1(l)x_i \tag{12}$$

where we used the approximation $(m-1)/n \simeq m/n = r$.

Calculations of the variance of $M_i$ and the covariance of $M_i$ and $N_i$ are carried out in a similar manner (see Appendix A). That is,

$$\begin{aligned} V[M_i] &= rF_2(l) + 2rlF_1(l)\dot{F}_1(l) + r\dot{F}_1(l)^2 && (13) \\ cov(N_i, M_i) &= rlF_1(l) + r\dot{F}_1(l) && (14) \end{aligned}$$

where $F_2(l)$ is

$$\begin{aligned} F_2(l) &= \left\langle f(l+\sqrt{r}\varepsilon)^2 \right\rangle \\ &= \int_{-\infty}^{\infty} f(l+\sqrt{r}t)^2 \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt \end{aligned}$$

The Fundamental Lemma is thus proved from Eqs. (8), (10), (12), (13) and (14).

# 5 Proof of the Main Theorem

Based on the Fundamental Lemma, we can derive the capacity and the radius of basin of attraction of memories. The capacity is obtained from the value of the loading rate at which radius of basin of attraction shrinks to zero. We first prove the linear case (the Corollary).

## 5.1 Linear Modification Functions

To contrast the behaviors of the two-stage neurons with the simple threshold neurons, let us illustrate, in the beginning, their noise reduction properties. A simplest example of the two-stage neuron would be the one with a linear modification function, i.e., $f(u) = -au$. Putting $a = 0$ the two-stage neuron is reduced to the simple threshold neuron. In this case, a simple addition of the modified internal field $Wf(\boldsymbol{u})$ to the original one $\boldsymbol{u}$ takes place. It is easy to get to the following noise reduction property (transformation of distance) from the Fundamental Lemma.

$$d' = (1-d)\,\phi\left(\frac{(1-a)l - ar}{\sigma}\right) + d\,\phi\left(\frac{(1-a)l + ar}{\sigma}\right)$$

where $l$ is the initial overlap and is related to the corresponding initial distance by $l = 1-2d$, and

$$\sigma^2 = r\left[1 + a^2(1+r+3l^2) - 2a(1+l^2)\right]$$

In particular for $l = 1$ (or equivalently $d = 0$), the above equation is reduced to

$$d' = \phi\left(\frac{1-a(1+r)}{\sigma}\right), \quad \sigma^2 = r\left[(1-2a)^2 + ra^2\right] \tag{15}$$

This equation describes how the pattern is disturbed starting from the memorized pattern itself. Fig. 1 shows this property. The stability is best at $a \approx 1/2$. Actually it is easy to find that the value of slope of the function $a$ that minimizes $d'$ is

$$a = (1-r)/(2-r)$$

Although most of the following discussion is devoted to the case of $r \ll 1$, this optimality condition holds in general. It would be worth noting that the two-stage neuron with linear modification function manifests its power dramatically for smaller $r$. This property is explained in Appendix B.

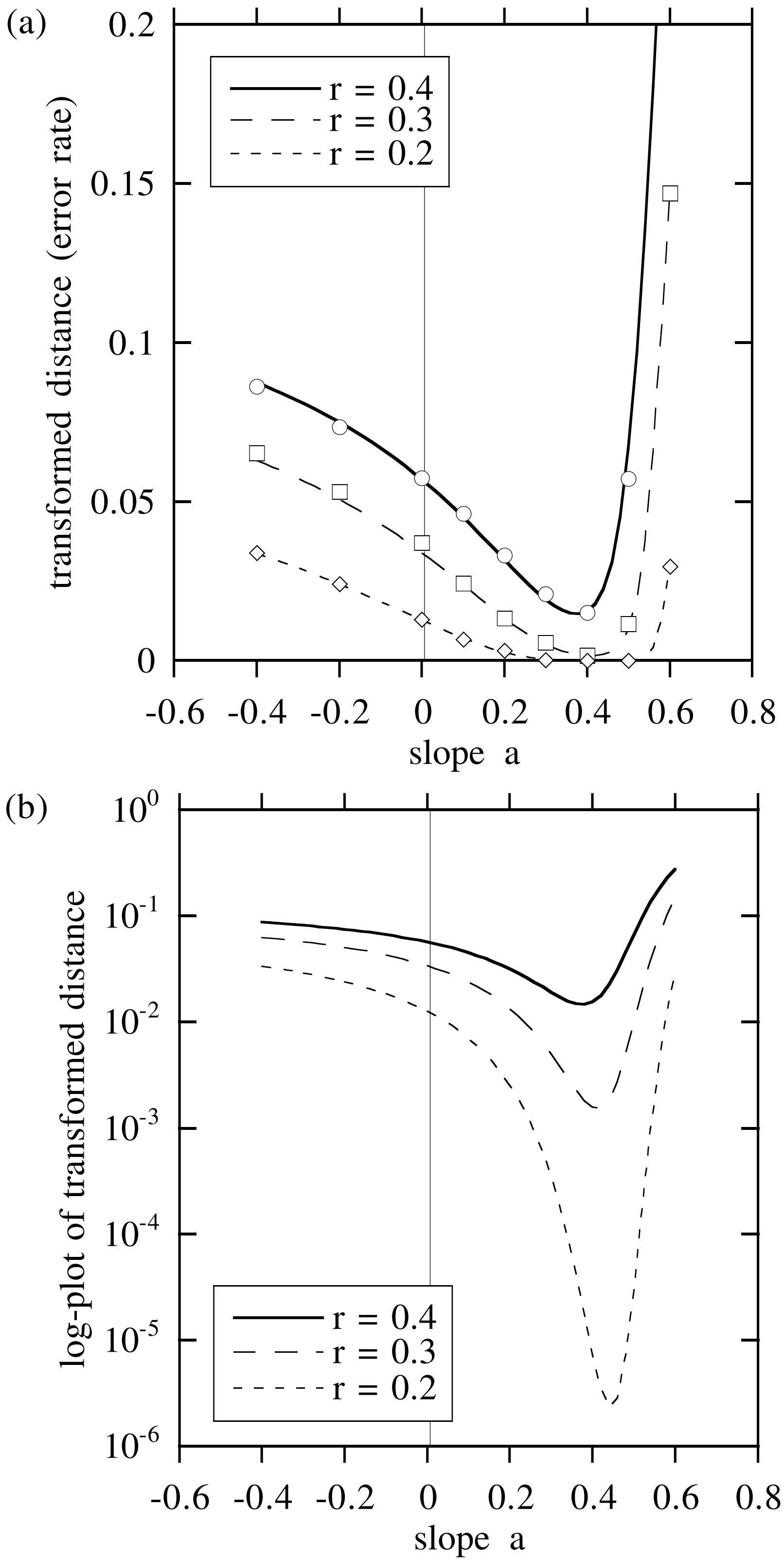


Figure 1: The relation between transformed distance $d'$ and slope "$a$" of the linear modification function $f(u) = au$, starting from one of the memorized patterns ($d = 0$). (a) Normal plot. Circles and squares are averages of ten trials of computer simulations with $n = 500$. (b) Semi-logarithm plot.

## 5.2 Absolute Memory Capacity for Linear Modification Functions

We will calculate the memory capacity for two-stage neuron nets with a linear modification function. The memory capacity is often called the storage capacity, or just the capacity. There are two usually accepted definitions of the capacity. One is the absolute capacity, which claims that the equilibrium state of the net to be strictly equal to a memory pattern, allowing no errors. The other is the relative capacity (or capacity in the sense of phase transition), which only claims that the net has equilibrium states close to memory patterns, allowing some errors in recollection. Since our analysis is statistical based on randomly generated memory patterns $\boldsymbol{\xi}_\mu$, it is required that the probability of the existence of the above equilibrium states tends to 1 as $n$ tends to infinity. It is known that the absolute capacity for the conventional neuron net is $n/(2\log n)$ [3], [11]. The well-known approximate value of $0.15n$ for auto-correlation matrix associative memory is an example of the relative capacity [3], [4], [8]. In the following, if not otherwise mentioned, we just use the word capacity for the absolute memory capacity.

Here we give a rough analysis. The exact analysis requires the large deviation theory or combinatorial arguments [11].

For the number of memory patterns to be within the capacity, the expected number of errors after one-step update has to be less than 1, that is, the condition

$$nd' = n\phi(U) < 1 \tag{16}$$

has to be satisfied for $l = 1$ (i.e., $d = 0$), where the expression for $U$ is seen in Eq. (15). For $n \gg 1$, $\phi(U)$ must satisfy $\phi(U) \ll 1$, then $U$ should be large, and therefore we can take the approximation $\phi(U) \approx \gamma(U)/U$ (cf. Eqs. (6) and (7)). Then the condition, Eq. (16), is approximated by

$$-\frac{U^2}{2} + \log n - \log \sqrt{2\pi}\, U < 0 \tag{17}$$

Assuming $U$ is of smaller order than $n$ (actually this assumption is consistent with the conclusion, Eq. (21)), we can take the most dominant terms into account.

$$-\frac{U^2}{2} + \log n < 0 \tag{18}$$

Note that the condition of Eq. (18) is stronger than that of Eq. (17), yielding the estimated capacity that is smaller than the corresponding one derived from Eq. (17). But also note that the difference vanishes as $n$ tends to infinity. Replacing the inequality in Eq. (18) by an equality, we obtain the critical condition for the capacity.

For the simple threshold neuron ($a = 0$), $U^2 = 1/r = n/m$ from Eq. (15). Thus from Eq. (18), the capacity is $m_c = n/(2\log n)$. Rewriting this capacity into the form of loading rate, we have $r_c = 1/(2\log n)$.

In the case of $a \neq 0$, the capacity is derived as follows. Since $U = \{1 - a(1+r)\}/\sigma$ with $\sigma^2 = r[(1-2a)^2 + ra^2]$, we have, from Eq. (18), the following equation to determine the capacity.

$$a^2(2\log n - 1)r^2 + 2[(1-2a)^2 \log n + a(1-a)]r - (1-a)^2 = 0$$

Putting $a = 0$, we have the results for the simple threshold neurons just mentioned above. For $a \neq 0$ and $a \neq 1/2$, this equation is approximated by

$$2a^2(\log n)r + 2(1-2a)^2(\log n)r - (1-a)^2 = 0 \tag{19}$$

hence the solution (the capacity in the form of loading rate) is

$$r_c = \left(\frac{1-2a}{a}\right)^2 \left[\sqrt{1 + \frac{a^2(1-a)^2}{(1-2a)^4 \log n}} - 1\right]$$

Supposing $O(a) = 1$, we can approximate the solution further as

$$r_c = \left(\frac{1-a}{1-2a}\right)^2 \frac{1}{2\log n} \tag{20}$$

If, in particular, $a = 1/2$, Eq. (19) is approximated by

$$(\log n)r^2 + r - 1/2 = 0$$

and the solution is

$$\begin{aligned} r_c &= \frac{1}{2\log n}\left[\sqrt{1+2\log n} - 1\right] \\ &\simeq \frac{1}{\sqrt{2\log n}} - \frac{1}{2\log n} \\ &\simeq \frac{1}{\sqrt{2\log n}} \end{aligned} \tag{21}$$

The capacity in the number of memorized patterns $m_c$ is $m_c = r_c n = n/\sqrt{2\log n}$. Thus the Corollary is proved.

This result indicates drastic superiority of the two-stage neuron with modification function $f(u) = -u/2$. The number of memorized patterns that can be stably stored is $\sqrt{2\log n}$ times larger with $a = 1/2$ than with $a = 0$ (the simple threshold neuron). This reveals the emergence of a different phase in the dynamics of correlation-type associative memory.

## 5.3 Piecewise Linear Modification Functions

Another interesting class of modification functions is the piecewise linear functions of the following type. Analog neural associative memory discussed in [26] deals with this function as the response function of a neuron.

$$f(u) = -au + c\ \mathrm{sgn}(u) \tag{22}$$

Here we will analyze the associative memory with this class of two-stage neurons. To apply the Fundamental Lemma, first, the integrals involved in the Lemma are calculated.

$$\begin{aligned} F_1(l) &= \int_{-\infty}^{\infty} f(l+\sqrt{r}t)\frac{1}{\sqrt{2\pi}}e^{\frac{-t^2}{2}}dt \\ &= \int_{-\infty}^{-\frac{l}{\sqrt{r}}} \left[-a(l+\sqrt{r}t) - c\right]\frac{1}{\sqrt{2\pi}}e^{\frac{-t^2}{2}}dt \\ &\quad + \int_{-\frac{l}{\sqrt{r}}}^{\infty} \left[-a(l+\sqrt{r}t) + c\right]\frac{1}{\sqrt{2\pi}}e^{\frac{-t^2}{2}}dt \\ &= -al + c(1-2\phi), \\ \dot{F}_1(l) &= -a + \frac{2c}{\sqrt{r}}\gamma, \\ F_2(l) &= \int_{-\infty}^{\infty} f(l+\sqrt{r}t)^2\frac{1}{\sqrt{2\pi}}e^{\frac{-t^2}{2}}dt \\ &= a^2(l^2+r) - 2ac\left[(1-2\phi)l + 2\sqrt{r}\,\gamma\right] + c^2 \end{aligned}$$

where we simplified the notations as $\phi = \phi(l/\sqrt{r})$ and $\gamma = \gamma(l/\sqrt{r})$.

Now let us proceed to the calculation of the capacity. Since we are considering the situation $r \ll 1$, the terms of the form $\phi(l/\sqrt{r})$,$\sqrt{r}\gamma(l/\sqrt{r})$ and $\gamma(l/\sqrt{r})/\sqrt{r}$ involved in the above expression are neglegibly small in comparison to $r$. If we set $l = 1$, $L + B$ and $\sigma$ are approximated, respectively, by

$$L + B = 1 + c - a(1+r)$$

and

$$\sigma^2 = r(1+c-2a)^2 + a^2r^2$$

It is easy to see that the condition which maximizes the capacity, or equivalently minimizes the transformed distance, is

$$1 + c - a(2-r)/(1-r) = 0 \tag{23}$$

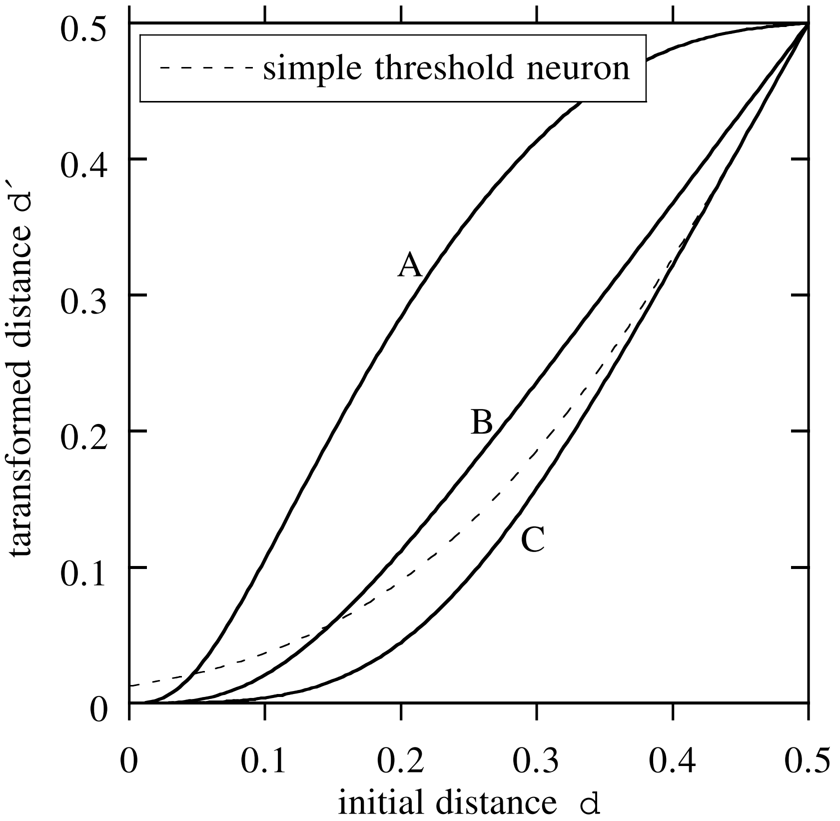


Figure 2: The relation between transformed distance and initial distance for piecewise linear modification functions of Eq. (22). A: $c = -0.25$, B: $c = 0$, C: $c = 1$, and $a = (1 + c)/2$ for all. Dashed line is for simple threshold neuron.

Since we suppose $r \ll 1$, this optimality condition is further approximated by

$$1 + c - 2a = 0 \tag{24}$$

Then we have

$$\frac{L+B}{\sigma} = \frac{1-r}{r}$$

under the condition $a > 0$. Note this is the same expression as in the case of $f(u) = -u/2$ (cf. Eq. (15)). The capacity, then, is the same as Eq. (21). Actually, if we put $c = 0$ into the above optimality condition, we have $a = 1/2$. The optimal choice of "$a$" for a linear modification function is a member of a more general class of piecewise linear functions with the restriction of Eq. (24).

We arrive at the fact that the capacity is uniform for a class of piecewise linear modification functions with the constraint $1 + c - 2a = 0$ and $a > 0$. The next question is whether the dynamical property is the same or not for different pairs of those $c$ and $a$. The answer is that the radius of basin of attraction of memorized patterns varies significantly for different choices of the pairs.

Before getting into concrete analysis of the radius of basin of attraction, we shall illustrate the noise reduction property for a macroscopic value of loading rate, $r = 0.2$. Fig. 2 shows theoretical results obtained from the equation in the Fundamental Lemma for several pairs of $c$ and $a$ which satisfies the above optimality condition (Eq. (23)). (Since the condition for optimal capacity is for $r \ll 1$, the above mentioned optimal parameter is not exactly optimal for the present case. But the difference is small.) To compare, a result for the simple threshold neuron ($c = a = 0$) is also shown. A linear modification function ($c = 0, a = 1/2$) is certainly good at $d = 0$, but if we take the error-correction property into account, strongly nonlinear functions with $a > 1/2$ are better. We shall evaluate the radius of basin of attraction in detail in the next section.

### 5.4 Radius of Basin of Attraction

For general values of $l$, by neglecting terms including $\phi(l/\sqrt{r})$,$\sqrt{r}\gamma(l/\sqrt{r})$, $\gamma(l/\sqrt{r})$ and $\gamma(l/\sqrt{r})/\sqrt{r}$, we obtain

$$L \pm B = l(1-a) + c \mp ra$$

and

$$\sigma^2 = r[1 + c^2 + 3a^2l^2 - 4acl + a^2 + 2cl - 2al^2 - 2a + a^2r]$$

We choose the parameters such that the capacity is maximized, i.e., $c = -1 + 2a$. Then we have

$$\sigma^2 = r[a^2r + (1-l)H(l,a)]$$

where

$$H(l,a) = 2 - 2(3-l)a + (5-3l)a^2$$

$H(l,a)$ can be zero if and only if $l = 1$, and is always positive otherwise. Minimum of $(1-l)H(l,a)$ is $(1-l)(1-l^2)/(5-3l)$ at $a = (3-l)/(5-3l)$.

Rewriting $L \pm B$ and $\sigma$ in terms of the initial distance $d$ and the slope of the function $a$,

$$\begin{aligned} L \pm B &= a - 2(1-a)d \mp ra, \\ \sigma^2 &= r[a^2r + 4(1-a)^2d - 4a(2-3a)d^2] \end{aligned}$$

The condition under which the initial state converges directly to the nearest memory pattern is obtained in a similar way to Eq. (18).

$$-\frac{(L \pm B)^2}{2\sigma^2} + \log n = 0$$

We estimate the maximum loading rate $R$ such that an initial state within a distance $D$ converges directly to the memorized pattern by one-step update. Scaling the loading rate as $R = \rho/\sqrt{2\log n}$, this condition is reduced to the following.

$$\begin{aligned} a^2\rho^2 + 4D\sqrt{2\log n}\{(1-a)^2 - a(2-3a)D\}\rho \\ +\{a - 2(1-a)D\}^2 = 0 \end{aligned}$$

The valid solution out of two is

$$\begin{aligned} \rho &= \frac{1}{a^2}\left[\sqrt{A^2 + a^2\{a - 2(1-a)D\}^2} - A\right], \\ A &= 2D\sqrt{2\log n}\left\{(1-a)^2 - a(2-3a)D\right\} \end{aligned} \tag{25}$$

By scaling $D$ as $D = \delta/\sqrt{2\log n}$, since $\delta$ is of order unity (in this case $D \ll 1$), Eq. (25) is approximated by

$$\rho = \sqrt{X^2+1} - X, \quad X = \frac{2\delta(1-a)^2}{a^2} \tag{26}$$

The relation between $\rho$ and $a$ is plotted in Fig. 3 for several values of $\delta$. For a desired radius of basin of attraction, the maximum loading rate $R$ is maximized at $a = 1$. Solving Eq. (26) with respect to $\delta$, we obtain the expression for the radius of basin of attraction given a loading rate.

$$D = \frac{\delta}{\sqrt{2\log n}} = \frac{1-\rho^2}{2\rho}\left(\frac{a}{1-a}\right)^2 \frac{1}{\sqrt{2\log n}} \tag{27}$$

In particular for $a = 1$, if the distance $d$ is of order $1/\sqrt{2\log n}$, an initial state converges directly to the nearest memory pattern by one-step with loading rate up to the capacity. In other words, the radius of basin of attraction is of order larger than $1/\sqrt{\log n}$ for $a = 1$.

To look at the case at $a = 1$ in detail, substitute $a = 1$ in Eq. (25). Then we obtain

$$X = 2D^2\sqrt{2\log n} \tag{28}$$

and conversely,

$$D = \frac{1}{2}\sqrt{\frac{1-\rho^2}{\rho}}\left(\frac{1}{2\log n}\right)^{\frac{1}{4}} \tag{29}$$

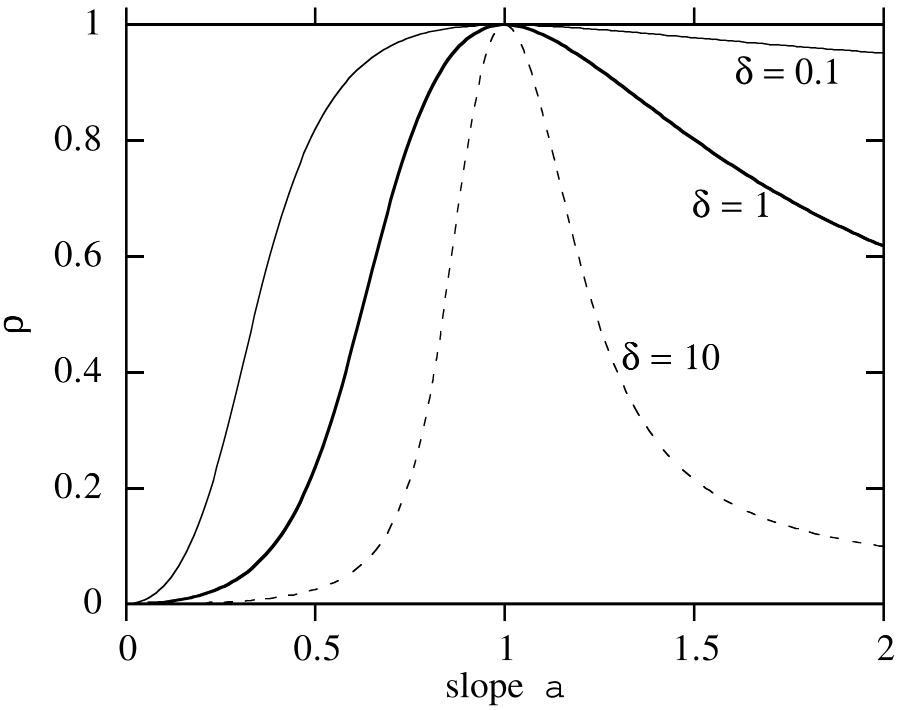


Figure 3: The relation between loading rate and the slope "$a$" for piecewise linear modification function of Eq. (22), for several values of radius of basin of attraction.

To see how the two-stage neuron overwhelms the simple threshold neuron, let us compare the radii of basins of attraction with the loading rate below the capacity of the simple threshold neuron, that is, $r < 1/(2\log n)$. Let us put $R = \kappa/(2\log n)$. By setting $a = c = 0$ in the above discussion or as we can find in the literature [11], the loading rate below which the radius of attraction basin is larger than $D$ is $\kappa = (1-2D)^2$, i.e., $R = (1-2D)^2/(2\log n)$, or conversely $D = (1-\sqrt{\kappa})/2$. Here $D$ is of order unity. For two-stage neuron, let us first choose the most powerful parameter $a = 1$. Since $\delta$ is of order $\sqrt{2\log n}$ now, the value $X$ in Eq. (28) is $X \gg 1$. Thus Eq. (25) is approximated by $\rho = 1/(2X)$. Substituting Eq. (28) into this equation, we obtain

$$\rho = \frac{1}{4D^2\sqrt{2\log n}}$$

Since $\kappa = \rho\sqrt{2\log n}$ by definition, we have

$$D = \frac{1}{2\sqrt{\kappa}}$$

Since radius of basin of attraction should satisfy $D \leq 1/2$, radius of the two-stage neuron net with $a = 1$ is the possible maximum for $\kappa \leq 1$. Similar calculations for $a = 1/2$ (optimal linear modification function) results in

$$\kappa = \frac{(1-2D)^2}{4D(1-D)}$$

and conversely,

$$D = \frac{1}{2}\left(1 - \sqrt{\frac{\kappa}{1+\kappa}}\right)$$

Radii of basins of attraction for $a = 1$ and $1/2$ is shown together with the one for the simple threshold neuron in Fig. 4.

# 6 Summary and Discussion

Capacity and radius of basin of attraction are calculated theoretically for neural auto-associative memory with two-stage discrete neurons. Those characteristic values are estimated based on the exact solution

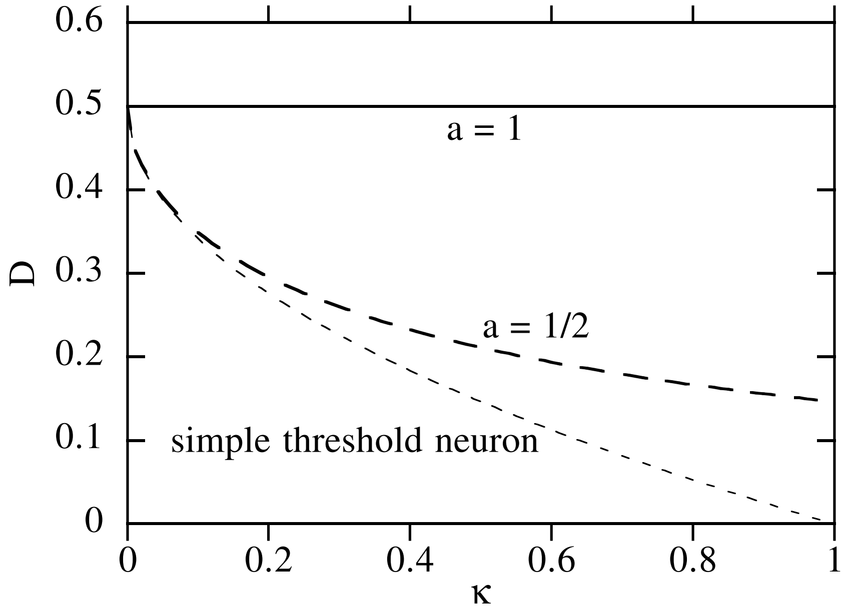


Figure 4: Radius of basin of attraction $D$ as a function of loading rate ($= \kappa/(2\log n)$).

to transformation of distance by one-step synchronous update of the net. The derived capacity is the absolute one, by which we mean there is no recall errors. Contrasted with the capacity $n/(2\log n)$ for the net with the simple threshold neurons, the net with the two-stage neurons has the capacity of $n/\sqrt{2\log n}$, when the modification function is

$$f(u) = -au - c\ \mathrm{sgn}(u)$$

with $1 + c - 2a = 0$ and $a > 0$. Radius of basin of attraction is also estimated for one-step direct recall of memory patterns. The maximum radius is realized at $(a, c) = (1, 1)$, decreasing as "$a$" goes apart from 1.

The Fundamental Lemma is for the number $n$ of neurons tending to infinity. However, as the computer simulated results show (Figure 1 (a)), it applies nicely to the net with hundreds of neurons.

The Main Theorem and the Corollary state the relative capacity of the net. As a natural consequence, the involved values of the loading rate are vanishingly small. Thus to see the relevance of those results with the relative capacity (macroscopic values of the loading rate) and the multi-step recall processes, we comment on computer simulations of the present model. Since theoretical estimation of the relative capacity is difficult, we just remark that, by computer simulations with hundreds of neurons, the relative capacity is about $0.3n$ for the above mentioned optimal combinations of $a$ and $c$. And for basin of attraction, the computer simulations for loading rate $r$ within $\sim 0.3$ exhibit corresponding behaviors to the theory for $r \ll 1$. That is, the radius of basin of attraction decreases as slope changes from $a \approx 1$ to 0. Moreover, faster convergence to the memorized patterns is observed for parameters which provide larger basin of attraction.

To compare our results with Morita's original two-stage model of Eq. (4), here we just show the results obtained from our Fundamental Lemma. The result is split into two cases depending on the scaling of parameters $c$ and $h$ according to the loading rate $r$. First, when we put $h = 1 + \alpha\sqrt{r}$ and $c = -\beta\sqrt{r}$ with $\alpha \simeq 0.6120$ and $\beta \simeq 3.023$, the capacity scales like $r_c \simeq 0.6363/\sqrt{2\log n}$. This is of the same order as the above discussed class of piecewise linear functions $f$. Second, if we use the constant parameter $c$ ($-2 \le c \le 0$), by scaling $h$ as the above, we obtain the capacity of $r_c \sim 1/(\log n)^{2/3}$

The improved performance by the two-stage model with $f(u) = -au$ can be understood by the correspondence between "correlation-matrix memory with two-stage neurons" and "orthogonal-projection-matrix memory with the simple threshold neurons." Define an $n \times m$ matrix $S(m < n)$ whose $\mu$th column is the $\mu$th memory pattern vector, that is, $S = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_m)$. The correlation matrix we defined is the $n \times n$ matrix $W = SS^T$ (times $1/n$ minus $rI$, precisely). The orthogonal-projection matrix for the above set of memory patterns is $P = SS^+ = S(S^TS)^{-1}S^T$, where $S^+$ is the generalized inverse matrix of $S$.

Using the von Neumann-type expansion for $S^+$ and truncating the series at the second term, we get

$$P \approx 2\alpha W\left[I - (\alpha/2)W\right]$$

with $0 < \alpha < 2/\lambda$, where $\lambda$ is the maximum eigen value of $SS^T$ (see e.g., [18]). Then we can see a good correspondence between the present two-stage model and the conventional model with the orthogonal projection matrix (see Eq. (3)).

We would like to draw reader's attention to another interpretation of the improvement. For the net with simple threshold neurons, although the absolute capacity of the auto-associative memory net and that of the layered or sequence-associative memory net are equal (i.e., $n/(2\log n)$), the relative capacities are different. This difference arises from the difference in correlation of consecutive internal fields of each neuron. Since layered or sequence-associative memory net has weaker such correlation, their relative capacity is as large as $0.27n$ (e.g., [7]). On the other hand, auto-associative memory net has stronger such correlation, hence it has larger variance of the noise term. However if we know the nature of that correlation, we can devise mechanisms that make positive use of the correlation. One of such mechanisms is our two-stage neuron.

Finally we point out some of remaining problems. That is, to discuss the following:

- The optimal shape of the modification function. If the class of functions investigated here is really optimal.
- The relationship of our results with analyses for analog models [20], [26]. In particular, whether or not our methodology contribute to verifying stability problem in equilibrium analyses there.
- Possible generalization of the present theorem to multiple-step behaviors, including the relative capacity. Or, behaviors of the net far apart from memorized patterns.
- Influence of bias in a group of memory patterns as discussed in [13].
- Applicability of two-stage neurons to other types of associative memory, e.g., sequence-associative memory, or neural nets other than associative memory.

# Appendix

# A Derivation of the Variance and the Covariance of Noise

By rewriting $M_i$ as

$$M_i^2 = \frac{1}{n^2}\sum_{j\neq i}\sum_{j'\neq i}\sum_{\mu\neq 1}\sum_{\mu'\neq 1}\xi_i^\mu\xi_j^\mu\xi_i^{\mu'}\xi_{j'}^{\mu'}f(l+N_j)f(l+N_{j'})$$

the expectation is

$$\begin{aligned}&E[M_i^2]\\ &= \frac{1}{n^2}\sum_{j\neq i}\sum_{j'\neq i}\sum_{\mu\neq 1}\sum_{\mu'\neq 1}\left\langle\xi_i^\mu\xi_j^\mu\xi_i^{\mu'}\xi_{j'}^{\mu'}f(l+N_j)f(l+N_{j'})\right\rangle\end{aligned}$$

In the following calculation we decompose $N_j$ and $N_{j'}$ according to correlation with $\xi_i^\mu, \xi_j^\mu, \xi_i^{\mu'}, \xi_{j'}^{\mu'}$, and/or combinations of them. The decomposition is, for $N_j$

$$\begin{aligned}N_j &= \frac{1}{n}\xi_j^\mu\xi_i^\mu x_i + N_j'\\ &= \frac{1}{n}\xi_j^\mu\xi_i^\mu x_i + \frac{\xi_j^\mu}{n}\sum_{k\neq j,i}\xi_k^\mu x_k + N_j''\end{aligned}$$

and likewise for $N_{j'}$. Choice of the expression depends on what correlation we have to consider. Note that

$$N_j = \sigma_0\varepsilon_j \simeq N_j' \simeq N_j''$$

in the large $n$ limit, where $\sigma_0 = \sqrt{r}$. Since $N_j$ is a sum of $m-1$ mutually uncorrelated terms, we have

$$\frac{1}{n}\sum_{k\neq j,i}\xi_k^\mu x_k \simeq \frac{1}{n}\sum_{k\neq j',i}\xi_k^\mu x_k \simeq \frac{\sigma_0}{\sqrt{m-1}}\delta \simeq \frac{\sigma_0}{\sqrt{m}}\delta$$

That is,

$$N_j = \sum_{\mu\neq 1}\xi_j^\mu\left(\frac{1}{n}\sum_{k\neq j}\xi_k^\mu x_k\right)$$

and likewise for $N_{j'}$. To summarize,

$$\begin{aligned} N_j &= \frac{\xi_j^\mu\xi_i^\mu x_i}{n} + \sigma_0\varepsilon_j = \frac{\xi_j^\mu\xi_i^\mu x_i}{n} + \frac{\sigma_0}{\sqrt{m}}\xi_j^\mu\delta + \sigma_0\varepsilon_j \\ N_{j'} &= \frac{\xi_{j'}^\mu\xi_i^\mu x_i}{n} + \sigma_0\varepsilon_{j'} = \frac{\xi_{j'}^\mu\xi_i^\mu x_i}{n} + \frac{\sigma_0}{\sqrt{m}}\xi_{j'}^\mu\delta + \sigma_0\varepsilon_{j'} \end{aligned}$$

where $\varepsilon_j, \varepsilon_{j'}$ and $\delta$ are mutually uncorrelated Gaussian noise with mean 0 and variance 1.

We divide the calculation into four parts according to whether $j = j'$ or not and whether $\mu = \mu'$ or not (see [25] for reference to these kind of statistical-neurodynamical calculations). That is,

$$\begin{aligned} E[M_i^2] &= E[M_i^2; j=j', \mu=\mu'] + E[M_i^2; j=j', \mu\neq\mu'] \\ &+ E[M_i^2; j\neq j', \mu=\mu'] + E[M_i^2; j\neq j', \mu\neq\mu'] \end{aligned} \tag{A.1}$$

There are

- $(n-1)(m-1) \simeq nm$ such terms that satisfy $j = j'$ and $\mu = \mu'$,
- $(n-1)(m-1)(m-2) \simeq nm^2$ for $j = j'$ and $\mu \neq \mu'$,
- $(n-1)(n-2)(m-1) \simeq n^2 m$ for $j \neq j'$ and $\mu = \mu'$, and
- $(n-1)(n-2)(m-1)(m-2) \simeq (nm)^2$ for $j \neq j'$ and $\mu \neq \mu'$.

The expectations are as follows. For $j = j'$ and $\mu = \mu'$,

$$\begin{aligned} E[M_i^2; j=j', \mu=\mu'] &= \frac{1}{n^2}nm\left\langle f(l+\sigma_0\varepsilon)^2\right\rangle \\ &= rF_2(l) \end{aligned} \tag{A.2}$$

and for $j = j'$ and $\mu \neq \mu'$,

$$\begin{aligned} E[M_i^2; j=j', \mu\neq\mu'] &= \frac{1}{n^2}nm^2\left\langle \xi_i^\mu\xi_j^\mu\xi_i^{\mu'}\xi_j^{\mu'} f(l+N_j)^2\right\rangle \\ &= \frac{m^2}{n}\left\langle \xi_i^\mu\xi_j^\mu\xi_i^{\mu'}\xi_j^{\mu'}\right\rangle\left\langle f(l+N_j)^2\right\rangle \\ &= 0 \end{aligned} \tag{A.3}$$

because asymptotically $N_j$ is uncorrelated with $\xi_i^\mu\xi_j^\mu\xi_i^{\mu'}\xi_j^{\mu'}$.

For $j \neq j'$ and $\mu = \mu'$,

$$\begin{aligned} &E[M_i^2; j\neq j', \mu=\mu'] \\ &= \frac{1}{n^2}n^2 m\left\langle \xi_j^\mu\xi_{j'}^\mu f\left(l + \frac{\xi_j^\mu\xi_{j'}^\mu x_{j'}}{n} + \frac{\sigma_0}{\sqrt{m}}\xi_j^\mu s + \sigma_0 t\right)\right. \\ &\qquad\left. \times f\left(l + \frac{\xi_{j'}^\mu\xi_j^\mu x_j}{n} + \frac{\sigma_0}{\sqrt{m}}\xi_{j'}^\mu s + \sigma_0 t'\right)\right\rangle \end{aligned} \tag{A.4}$$

The random variables $s$, $t$, and $t'$ are independently subject to unit Gaussian distribution. Calculation of this expectation is as follows.

By writing, for the sake of simplicity, $\xi = \xi_j^\mu$, $\xi' = \xi_{j'}^\mu$, $x = x_j$, and $x' = x_{j'}$,

$$
\begin{aligned}
&\langle \xi\xi' f(\bullet) f(\bullet)\rangle \\
&= \langle \xi\xi' \langle f(\bullet)\rangle_t \langle f(\bullet)\rangle_{t'}\rangle \\
&= \left\langle \xi\xi' \left\langle F_1\left(l + \tfrac{\xi\xi' x}{n} + \tfrac{\sigma_0}{\sqrt{m}}\xi s\right) F_1\left(l + \tfrac{\xi'\xi x'}{n} + \tfrac{\sigma_0}{\sqrt{m}}\xi' s\right)\right\rangle_s \right\rangle
\end{aligned}
$$

where the notation like $\langle\bullet\rangle_s$ is used to denote explicitly averaging with respect to $s$. By expanding $F_1$ into Taylor series around $l$ we have

$$
\begin{aligned}
&F_1\left(l + \tfrac{\xi\xi' x}{n} + \tfrac{\sigma_0}{\sqrt{m}}\xi s\right) \\
&\quad = F_1(l) + \left(\tfrac{\xi\xi' x}{n} + \tfrac{\sigma_0}{\sqrt{m}}\xi s\right)\dot{F}_1(l) + \cdots \\
&F_1\left(l + \tfrac{\xi'\xi x'}{n} + \tfrac{\sigma_0}{\sqrt{m}}\xi' s\right) \\
&\quad = F_1(l) + \left(\tfrac{\xi'\xi x}{n} + \tfrac{\sigma_0}{\sqrt{m}}\xi' s\right)\dot{F}_1(l) + \cdots
\end{aligned}
$$

so that their product is

$$
\begin{aligned}
F_1(l)^2 \quad &+ \quad \left(\frac{x + x'}{n}\xi\xi' + \frac{\xi + \xi'}{\sqrt{m}}\sigma_0 s\right) F_1(l)\dot{F}_1(l) \\
&+ \quad \left(\frac{xx'}{n} + \frac{\xi + \xi'}{\sqrt{m}\,n}\sigma_0 s + \frac{\xi\xi'}{m}\sigma_0^2 s^2\right)\dot{F}_1(l)^2 + \cdots
\end{aligned}
$$

Noting that $\langle s\rangle_s = 0$ and $\langle s^2\rangle_s = 1$, and that $\langle\xi\xi'\rangle = 0$ and $\langle(\xi\xi')^2\rangle = 1$, we can see

$$
\begin{aligned}
&\left\langle \xi\xi' \left[F_1(l)^2 + \frac{x + x'}{n}\xi\xi' F_1(l)\dot{F}_1(l) \right.\right. \\
&\qquad\qquad \left.\left. + \left(\frac{xx'}{n} + \frac{\xi\xi'}{m}\sigma_0^2\right)\dot{F}_1(l)^2\right]\right\rangle \\
&= \frac{x + x'}{n} F_1(l)\dot{F}_1(l) + \frac{\sigma_0^2}{m}\dot{F}_1(l)^2
\end{aligned}
$$

Multiplying this result by $m$ according to Eq. (A.4), we have the following result.

$$
r(x_j + x_{j'})F_1(l)\dot{F}_1(l) + \sigma_0^2\dot{F}_1(l)^2
$$

By taking expectation over $x_j$, since $E[x_j] = l$, we obtain

$$
E[M_i^2; j \neq j', \mu = \mu'] = 2rlF_1(l)\dot{F}_1(l) + \sigma_0^2\dot{F}_1(l)^2 \tag{A.5}
$$

The case of $j \neq j'$ and $\mu \neq \mu'$ is easily calculated likewise, approximated asymptotically by

$$
E[M_i^2; j \neq j', \mu \neq \mu'] = E[M_i^2] \tag{A.6}
$$

Therefore from Eqs. (A.1), (A.2), (A.3), (A.5), and (A.6), the variance of $M_i$ is

$$
\begin{aligned}
V[M_i] &= E[M_i^2] - E[M_i]^2 \\
&= rF_2(l) + 2rlF_1(l)\dot{F}_1(l) + \sigma_0^2\dot{F}_1(l)^2
\end{aligned} \tag{A.7}
$$

Now let us calculate the covariance of $N_i$ and $M_i$. Since expectation of $N_i$ is zero, the covariance is equal to $E[N_iM_i]$.

$$
\begin{aligned}
&E[N_iM_i] \\
&= E\left[\left(\frac{1}{n}\sum_{j\neq i}\sum_{\mu\neq 1}\xi_i^\mu\xi_j^\mu x_j\right)\left(\frac{1}{n}\sum_{k\neq i}\sum_{\nu\neq 1}\xi_i^\nu\xi_k^\nu f(l + N_k)\right)\right] \\
&= \frac{1}{n}\sum_{j\neq i}\sum_{\mu\neq 1}\sum_{k\neq i}\sum_{\nu\neq 1}\langle\xi_i^\mu\xi_j^\mu\xi_i^\nu\xi_k^\nu x_j\langle f(l + N_k)\rangle\rangle
\end{aligned} \tag{A.8}
$$

Here again we decompose terms into four cases.

For $j = k$ and $\mu = \nu$,

$$\begin{aligned} E[N_i M_i; j = k, \mu = \nu] &= \frac{m-1}{n} \left( \frac{1}{n} \sum_{j \neq i} x_j \right) \langle f(l + N_j) \rangle \\ &= r l F_1(l) \end{aligned} \tag{A.9}$$

This is because $x_j$ is uncorrelated with $N_j$. For $j = k$ and $\mu \neq \nu$, as in the corresponding case in $E[M_i^2]$,

$$E[N_i M_i; j = k, \mu \neq \nu] = 0 \tag{A.10}$$

For $j \neq k$ and $\mu = \nu$,

$$\begin{aligned} & E[N_i M_i; j \neq k, \mu = \nu] \\ & = \frac{1}{n} \sum_{j \neq i} \sum_{k \neq i} \sum_{\mu \neq 1} \left\langle \xi_j^\mu \xi_k^\mu x_j \left\langle f(l + N_k) \right\rangle \right\rangle \\ & = \tfrac{(n-1)(n-2)(m-1)}{n^2} \left\langle \xi_j^\mu \xi_k^\mu x_j \left\langle f\left(l + \tfrac{\xi_j^\mu \xi_k^\mu x_j}{n} + \sigma_0 t\right) \right\rangle_t \right\rangle \\ & \simeq m \left\langle \xi_j^\mu \xi_k^\mu x_j F_1\left(l + \tfrac{\xi_j^\mu \xi_k^\mu x_j}{n}\right) \right\rangle \end{aligned} \tag{A.11}$$

Since we can put, by expanding the function $F_1$ around $l$,

$$F_1\left(l + \frac{\xi_j^\mu \xi_k^\mu x_j}{n}\right) \simeq F_1(l) + \frac{\xi_j^\mu \xi_k^\mu x_j}{n} \dot{F}_1(l)$$

the expectation is

$$E[N_i M_i; j \neq k, \mu = \nu] = r \dot{F}_1(l) \tag{A.12}$$

by a similar procedure as used above. And for $j \neq k$ and $\mu \neq \nu$,

$$E[N_i M_i; j \neq k, \mu \neq \nu] = E[N_i] E[M_i] = 0 \tag{A.13}$$

Therefore from Eqs. (A.8), (A.9), (A.10), (A.11), (A.12), and (A.13), the covariance is

$$cov(N_i, M_i) = r l F_1(l) + r \dot{F}_1(l) \tag{A.14}$$

Consequently the variance of the noise $\sigma^2$ is, from Eqs. (10), (A.7), and (A.14),

$$\begin{aligned} \sigma^2 &= r \Big[ 1 + F_2(l) + 2 l F_1(l) \dot{F}_1(l) \\ &\qquad + \frac{\sigma_0^2}{r} \dot{F}_1(l)^2 + 2 \left( l F_1(l) + \dot{F}_1(l) \right) \Big] \end{aligned}$$

# B Optimally Tuned Slope $a$ vs. $a = 0$ (the Simple Threshold Neuron) for Linear Modification Function

For $U$ large (i.e., $r$ small), we can approximate $\phi(U)$ as

$$\begin{aligned} \log \phi(U) &\approx \log \gamma(U) - \log U \\ &= -U^2/2 - \log U - \log \sqrt{2\pi} \\ &\approx -U^2/2 \end{aligned}$$

For optimal $a = (1-r)/(2-r)$, $U_{opt}^2 = -(1 - r + r^2)/(2r^2)$, and for $a = 0$, $U_{a=0}^2 = 1/r$. So we have the fraction of relative performance

$$\begin{aligned} \log \frac{\phi_{opt}}{\phi_{a=0}} &= \log \phi_{opt} - \log \phi_{a=0} \\ &\approx -\frac{1}{2} \left( U_{opt}^2 - U_{a=0}^2 \right) \\ &= -\frac{1}{2} \left( \frac{1-r}{r} \right)^2 \\ &\to -\infty \quad (\text{as } r \to \infty) \end{aligned}$$

# References


[1] Amari, S. (1972): Learning patterns and pattern sequence by self-organizing nets of threshold elements, *IEEE Trans. Computers*, vol. C-21, pp. 1197–1206

[2] Amari, S. (1989): Characteristics of sparsely encoded associative memories, *Neural Networks*, vol. 2, pp. 451–457

[3] Amari, S. & Maginu, K. (1988): Statistical neurodynamics of associative memory, *Neural Networks*, vol. 2, pp. 63–73

[4] Amit, D. J., Gutfreund, H. & Sompolinsky, H. (1985): Storing infinite number of patterns in a spin-glass model of neural networks, *Pys. Rev. Lett.*, vol. 55, pp. 1530–1533

[5] Gardner, E. (1988): The space of interactions in neural network models, *J. Phys. A: Math. Gen.*, vol. 21, pp. 257–270

[6] Gardner, E., Derrida, B. & Mottishaw, P. (1987): Zero temperature parallel dynamics for infinite range spin glasses and neural networks, *J. Physique*, vol. 48, pp. 741–755

[7] Hassoun, M. H. ed. (1993): "Associative Neural Memories — Theory and Implementation", Oxford University Press.

[8] Hopfield, J. J. (1982): Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 79, pp. 2554–2558

[9] Kobayashi, K. (1991): On the capacity of a neuron with a non-monotone output function, *Network*, vol. 2, pp. 237–243

[10] Kohonen, T. (1989): "Self-Organization and Associative Memory", Springer-Verlag

[11] McEliece, R. J., Posner, E. C., Rodemich, E. R. & Venkatesh, S. S. (1987): The capacity of the Hopfield associative memory, *IEEE Trans. Information Theory*, vol. IT-33, pp. 461–482

[12] Meunier, C., Yanai, H.-F. & Amari, S. (1991): Sparsely coded associative memories: capacity and dynamical properties, *Network*, vol. 2, pp. 469–487

[13] Morita, M. (1993): Associative memory with nonmonotone dynamics, *Neural Networks*, vol. 6, pp. 115–126

[14] Morita, M., Yoshizawa, S. & Nakano, K. (1990): Analysis and improvement of the dynamics of autocorrelation associative memory, *Transactions of the Institute of Electronics, Information and Communication Engineers of Japan*, vol. J73-D-II, pp. 232–242

[15] Nishimori, H. & Opris, I. (1993): Retrieval process of an associative memory with a general input-output function, *Neural Networks*, vol. 6, pp. 1061–1067

[16] Okada, M. (1995): A hierarchy of macrodynamical equations for associative memory, *Neural Networks*, to appear

[17] Palm, G. (1980): On associative memory, *Biological Cybernetics*, vol. 36, pp. 19–31

[18] Rao, C. R., & Mitra, S. K. (1971): "Generalized Inverse of Matrices and its Applications", John Wiley & Sons

[19] Shiino, M. & Fukai, T. (1992): Self-consistent signal-to-noise analysis and its application to analogue neural networks with asymmetric connections, *J. Phys. A: Math. Gen.*, vol. 25, pp. L375–L381

[20] Shiino, M. & Fukai, T. (1993): Self-consistent signal-to-noise analysis of the statistical behavior of analog neural networks and enhancement of the storage capacity, *Physical Review E*, vol. 48, pp. 867–897

[21] Yanai, H.-F. & Amari, S. (1993): A theory on a neural net with non-monotone neurons, *Proc. 1993 IEEE International Conference on Neural Networks*, pp. 1385–1390

[22] Yanai, H.-F. & Sawada, Y. (1990): Associative memory network composed of neurons with hysteretic property, *Neural Networks*, vol. 3, pp. 223–228

[23] Yanai, H.-F. & Sawada, Y. (1990): Effects of neuron property on the performance of associative memory networks, *Proceedings of the International Joint Conference on Neural Networks '90 (Washington)*, pp. I-489–I-452, Lawrence Erlbaum Associates

[24] Yanai, H.-F. & Sawada, Y. (1990): Integrator neurons for analogue neural networks, *IEEE Transactions on Circuits and Systems*, vol. 37, pp. 854–856

[25] Yanai, H.-F., Sawada, Y. & Yoshizawa, S. (1991): Dynamics of an auto-associative neural network model with arbitrary connectivity and noise in the threshold, *Network*, vol. 2, pp. 295–314

[26] Yoshizawa, S., Morita, M. & Amari, S. (1993): Capacity of associative memory using a nonmonotonic neuron model, *Neural Networks*, vol. 6, pp. 167–176