Neurocomputing 469 (2022) 198–208

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

The capacity of the dense associative memory networks

Han Bao [a], Richong Zhang [a,*], Yongyi Mao [b]

[a] School of Computer Science and Engineering, Beihang University, Xueyuan Road No. 37, Haidian District, Beijing, China
[b] School of Electrical Engineering and Computer Science, Ottawa University, 75 Laurier Ave. East, Ottawa, Canada

ARTICLE INFO

Article history:
Received 28 October 2020
Revised 21 August 2021
Accepted 19 October 2021
Available online 21 October 2021
Communicated by Zidong Wang

Keywords:
Hopfield network
DAM networks
Capacity
Noise recovery

ABSTRACT

This paper revisits the dense associative memory (DAM) networks and studies rigorously the capacity of the DAM networks. We present the capacity theorem of the DAM networks with an attraction radius or a noise level from the messages and prove that the probe can converge to the targeted message just after the one-step update. Under this convergence, the capacity of DAM networks is between a lower bound and an upper bound. Although when the attraction radius is 0.0 away from the messages, i.e. noiseless, previous literature provides an approximate result. However, a rigorous proof is not given in this study. In addition, we consider a more general notion of capacity which allows the retrieval of messages from noisy probes (the attraction radius is not 0.0). We demonstrates that the convergence result can be acquired just after the one-step update when the probe is a corrupted version with a Gaussian noise from one message. We further provide simulated experiments to validate theorems herein.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

The conceptually powerful Hopfield network [1] is the classic framework for associative content-addressable memory [2,3]. Briefly, a Hopfield network consists of N neurons, which is intended for storing a set of messages, $\mathscr{M} = \{\xi^1, \xi^2, \ldots, \xi^M\}$, where each message ξ^μ is a binary (± 1) vector of length N. The connectivity of the neurons is defined via a $N \times N$ symmetric weight matrix W which encodes the message set $\mathscr{M}$ by

$$W_{ij} = \sum_{\mu=1}^{M} \xi_i^\mu \xi_j^\mu. \tag{1}$$

The network is associated with an *update rule* T defined as follows: for any configuration (or state) $\sigma \in \{\pm 1\}^N$ of the Hopfield network,

$$T(\sigma) = \text{sign}(W\sigma), \tag{2}$$

where the function sign : $\mathbb{R} \to \{\pm 1\}$ returns $+1$ for a positive input and -1 for a negative input. When acting on a vector, the function operates element-wise. It is possible to show that the network can be associated with an energy function

$$H = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sigma_i W_{ij}\sigma_j, \tag{3}$$

for which $H(T(\sigma)) \leqslant H(\sigma)$. Then the Hopfield network can be viewed alternatively as being characterized by the energy function H and an energy-minimizing update rule T.

To retrieve a message in $\mathscr{M}$, the network is presented with a probe $\mathbf{p} \in \{\pm 1\}^N$ similar to a message in $\mathscr{M}$; the network then initializes its state σ as $\mathbf{p}$ and iterates the update rule T. If the state can converge to the targeted message, then the retrieval is successful. The storage capacity of the Hopfield network, although having various definitions (see Section 5.1), refer to the maximal number of messages encoding the weight matrix (or equivalently the energy function) that can be successfully retrieved.

Since the seminal work of J. J. Hopfield [1], associative memory networks have become an attractive field in computer science [4,5], mathematics [6,7], neuroscience [8,9], and physics [10–12]. In 2016, D. Krotov and J. J. Hopfield extended the Hopfield networks to a wider family which they call dense associative memory (DAM) [13], which has an expanded storage capacity. Specifically, a DAM network is defined by a generalized notion of energy function and a corresponding energy-minimizing update rule. Remarkably, the works of [13,14] showed that there appears a "duality" relationship between the DAM networks and the feed-forward neural networks, a model architecture widely used in the machine learning field. This potentially enables new routes towards understanding deep neural networks in the modern deep learning paradigm.

* Corresponding author.
E-mail address: zhangrc@act.buaa.edu.cn (R. Zhang).

https://doi.org/10.1016/j.neucom.2021.10.058
0925-2312/© 2021 Elsevier B.V. All rights reserved.

Although in their paper [13], D. Krotov and J. J. Hopfield presented a formula for the storage capacity of DAM networks. Their work is limited in two ways. First, no rigorous proof is given; and second, only a restricted notion of capacity is considered. For the latter, we note that the notion of capacity considered in [13] is the maximum number of stable messages that can be encoded by the energy function, or equivalently, the maximum number of messages that can be successfully recovered when the probe $\mathbf{p}$ is noiseless, i.e., equal to the targeted message.

In this paper, we consider a more general notion of capacity which allows the retrieval of messages from noisy probes. As we will define such notion of capacity more precisely in Section 2.2, here we note that a noisy probe is a binary N-vectors perturbed from a message. It is worth noting that this notion of capacity reduces to the capacity considered in [13] in the noise-free limited and is thus a generalization of it. Additionally, this generalized notion of capacity has been studied for the classical Hopfield networks [15], but has not been investigated for DAM networks. In this paper, we analytically study this notion of capacity for DAM networks and present a capacity result in terms of an lower bound and an upper bound. We also conducted simulations to empirically validate these results.

Proofs of some results are given in Appendix.

2. Dense associative memory and capacity definition

2.1. DAM networks

A DAM network is defined via a modified energy function from the Eq. (3) and a correspondingly modified update rule from the Eq. (2). Specifically the energy function of the DAM is defined as [13],

$$H = -\sum_{\mu=1}^{M} F\left(\sum_{i=1}^{N} \xi_i^\mu \sigma_i\right), \tag{4}$$

where function $F : \mathbb{R} \to \mathbb{R}$ is defined as[1]

$$F(a) = a^n, \tag{5}$$

for a positive integer $n \geqslant 2$. Correspondingly the update rule T of the DAM networks is defined as a sequence $(T_1, T_2, \ldots, T_N)$ of functions. Specifically, for each $i \in \{1, \ldots, N\}, T_i : \{\pm 1\}^N \to \{\pm 1\}$ is defined as

$$T_i(\sigma) = \operatorname{sign}(H_{i^-}(\sigma) - H_{i^+}(\sigma)), \tag{6}$$

where

$$\begin{aligned} H_{i^-}(\sigma) &= -\sum_{\mu=1}^{M} F\left(-\xi_i^\mu + \sum_{j\neq i} \xi_j^\mu \sigma_j\right), \\ H_{i^+}(\sigma) &= -\sum_{\mu=1}^{M} F\left(\xi_i^\mu + \sum_{j\neq i} \xi_j^\mu \sigma_j\right). \end{aligned} \tag{7}$$

Let $T(\sigma)$ denote $(T_1(\sigma), T_2(\sigma), \ldots, T_N(\sigma))$. Then it is possible to show that $H(T(\sigma)) \leqslant H(\sigma)$ and thus updating with T reduces the energy of DAM networks (proof given in Appendix A). Additionally, when $n = 2$, it can be verified that the energy function H and the update rule T are identical to those of the Hopfield network.

[1] In [13], the function F is in fact defined as $F(a) = \max(a, 0)^n$.

But we found empirically that simplifying its form to (4) has no impact on the capacity of DAM.

For later use, we here a present a different characterization of the update rule T of DAM networks.

Lemma 1. *The update rule T of the above defined DAM networks can be expressed as*

$$T_i(\sigma) = \operatorname{sign}\left(2\sum_{\mu=1}^{M}\sum_{k=1}^{\frac{n}{2}} \binom{n}{2k-1} (\xi_i^\mu)^{n-2k+1} \left(\sum_{j\neq i} \xi_j^\mu \sigma_j\right)^{2k-1}\right), \tag{8}$$

for even n;

$$T_i(\sigma) = \operatorname{sign}\left(2\sum_{\mu=1}^{M}\sum_{k=0}^{\frac{n-1}{2}} \binom{n}{2k} (\xi_i^\mu)^{n-2k} \left(\sum_{j\neq i} \xi_j^\mu \sigma_j\right)^{2k}\right), \tag{9}$$

for odd n.

2.2. Definition of capacity

Consider the setting in which every message ξ^μ in $\mathscr{M}$ is a length-N i.i.d. sequence drawn from the uniform distribution on $\{\pm 1\}$. For any $\delta < 0.5$, a δ-perturbation of a message ξ^μ refers to randomly flipping the sign of each component ξ_i^μ of ξ^μ independently with a probability δ. We say that the random message set $\mathscr{M}$ is one-step (δ, ε)-retrievable if for each message $\xi^\mu \in \mathscr{M}$, when we apply a random δ-perturbation and obtain a probe $\bar{\xi}^\mu$, the probability that update under T does not converge after one-step update is less than ε, i.e. $P\left(\exists i : T_i\left(\bar{\xi}^\mu\right) \neq \xi_i^\mu\right) < \varepsilon$. We define the one-step δ-capacity C_δ of the DAM as the maximum size M of the random message set $\mathscr{M}$, for which $\mathscr{M}$ is one-step (δ, ε)-retrievable for any $\varepsilon > 0$ when N is made sufficiently large. Note that when $\delta = 0$, the one-step δ-capacity C_δ is in fact the notion of capacity in [13].

One may similarly define notions of capacity for multi-step convergence. More precisely, a random message set $\mathscr{M}$ is said to be k-step (δ, ε)-retrievable if for each message $\xi^\mu \in \mathscr{M}$, when we apply a random δ-perturbation and obtain a probe $\bar{\xi}^\mu$, the probability that k-step update under T does not converge is less than ε, i.e. $P\left(\exists i : T_i^k\left(\bar{\xi}^\mu\right) \neq \xi_i^\mu\right) < \varepsilon$. Here we have used T^k to denote the k-fold composition $T \circ T \ldots \circ T$ of the updating function T. The k-step δ-capacity $C_\delta^{(k)}$ of the DAM is then the maximum size M of $\mathscr{M}$ for which $\mathscr{M}$ is k-step (δ, ε)-retrievable for any $\varepsilon > 0$ when N is made sufficiently large.

In this paper, we theoretically establish the one-step δ-capacity C_δ or $C_\delta^{(1)}$ for DAM networks. as a function of the network size N and the attraction radius or the noise level δ [16]. We also experimentally investigate the k-step δ-capacities $C_\delta^{(k)}$ of the DAM networks.

3. Capacity Theorem and Proof

We now state the main theoretical result of this paper concerning the one-step δ-capacity C_δ of the DAM networks, when N is a large number.

Theorem 1. *When N is a large number, the one-step δ-capacity C_δ of the DAM networks is bounded as*

$$\frac{(1-2\delta)^{2(n-1)} N^{n-1}}{-\ln 2\pi\varepsilon^2 (2n-3)!!} < C_\delta < \frac{(1-2\delta)^{2(n-1)} N^{n-1}}{2(2n-3)!!}, \tag{10}$$

where the lower bound is valid for $\varepsilon < \frac{1}{\sqrt{2\pi e^2}}$.

Subsequently, we will denote the lower bound and the upper bound in the theorem by C_{low} and C_{up}, respectively.

3.1. Proof of Theorem 1

We give a proof of Theorem 1 for the case when n is even. When n is odd, the proof is essentially the same, but with a slight difference about the update rule. Detailed will be referred in the proof process.

Assume that there are M independent messages, where each message consists of N i.i.d. binary RV's with $P(\xi_i^\mu = 1) = P(\xi_i^\mu = -1) = \frac{1}{2}, i = 1, \ldots, N$. In other words, the messages can be thought of as the columns of an $N \times M$ matrix of binary i.i.d. RV's. The noisy probe is a corrupted version of one of the messages, which without loss of generality we assume to be the first message ξ^1. Specifically, the noisy probe $\bar{\xi}^1$ is obtained from the message ξ^1 by independently flipping each component of the message with a probability δ.

Now, consider the RV's $\overset{\mu}{X}_i = \sum_{j\neq i}^{N}\left(\xi_j^\mu \bar{\xi}_j^1\right)$ for $i = 1, \ldots, N$ and for $\mu \neq 1$. Since ξ_j^μ and $\bar{\xi}_j^1$ are i.i.d then for a large N, the random variable $\overline{X}_i^\mu$ is Gaussian distributed with zero mean and variance as following, according to the central limit theory [17],

$$\begin{aligned}\mathrm{Var}\overline{X}_i^\mu &= E\left[\left(\sum_{j\neq i}^{N}\xi_j^\mu\bar{\xi}_j^1\right) - 0\right)^2\Bigg] \\ &= E\left[\sum_{j\neq i}^{N}\left(\xi_j^\mu\bar{\xi}_j^1\right)^2 + 2\sum_{j\neq i}^{N}\sum_{k\neq i j}^{N}\left(\xi_j^\mu\bar{\xi}_j^1\right)\left(\xi_k^\mu\bar{\xi}_k^1\right)\right] \\ &= (N-1) + 0 = N-1.\end{aligned} \tag{11}$$

Next, we calculate the probability $P\left(\exists i \leqslant N : T_i\left(\bar{\xi}^1\right) \neq \xi_i^1\right)$, denoted by P_{er}, where the noisy probe cannot converge to its targeted message in one-step update strategy. This is a strong convergence condition in that we still declare an error even if the probe can converge to the desired message in k-step update strategy (see [15] for more details), or if it converges to any of the other messages in k-step update strategy (in contrast to [13]). To obtain the desired bounds, we take a closer look at the values of M (the number of messages) for which

$$P_{\text{er}} < \varepsilon, \quad \forall \varepsilon > 0. \tag{12}$$

To this end, let $\overline{Y}_i^1 = \sum_{j\neq i}^{N}\xi_j^1\bar{\xi}_j^1$, then

$$\begin{aligned}E(\overline{Y}_i^1) &= \sum_{j\neq i}^{N}\left[\xi_j^1\left(-\xi_j^1\right)\delta + \xi_j^1\xi_j^1(1-\delta)\right] = (1-2\delta)(N-1), \\ E\left(\left(\overline{Y}_i^1\right)^n\right) &= E\left(\sum_{j\neq i}^{N}\left(\xi_j\bar{\xi}_j^1\right)\left(\overline{Y}_i^1\right)^{n-1}\right) = (1-2\delta)^n(N-1)^n.\end{aligned} \tag{13}$$

Since n is even, from the Eq. (8),

$$\begin{aligned}T_i\left(\bar{\xi}^1\right) &= \operatorname{sign}\left(2\sum_{\mu=1}^{M}\sum_{k=1}^{\frac{n}{2}}\binom{n}{2k-1}\left(\xi_i^\mu\right)^{n-2k+1}\left(\sum_{j\neq i}\xi_j^\mu\bar{\xi}_j^1\right)^{2k-1}\right) \\ &= \operatorname{sign}\left(2\sum_{k=1}^{\frac{n}{2}}\binom{n}{2k-1}\left(\xi_i^1\right)^{n-2k+1}\left(\sum_{j\neq i}\xi_j^1\bar{\xi}_j^1\right)^{2k-1} + \right. \\ &\quad \left. 2\sum_{\mu=2}^{M}\sum_{k=1}^{\frac{n}{2}}\binom{n}{2k-1}\left(\xi_i^\mu\right)^{n-2k+1}\left(\sum_{j\neq i}\xi_j^\mu\bar{\xi}_j^1\right)^{2k-1}\right) \\ &= \operatorname{sign}\Big(2\binom{n}{1}\left(\xi_i^1\right)^{n-1}\overline{Y}_i^1 + \cdots + 2\binom{n}{n-1}\left(\xi_i^1\right)\left(\overline{Y}_i^1\right)^{n-1} + \\ &\quad 2\sum_{\mu=2}^{M}\Big[\binom{n}{1}\left(\xi_i^\mu\right)^{n-1}\overline{X}_i^\mu + \cdots + \binom{n}{n-1}\left(\xi_i^\mu\right)\left(\overline{X}_i^\mu\right)^{n-1}\Big]\Big) \\ &= \operatorname{sign}\left(2n\xi_i^1\left[\left(\overline{Y}_i^1\right)^{n-1} + \sum_{\mu=2}^{M}\left(\xi_i^1\xi_i^\mu\right)\left(\overline{X}_i^\mu\right)^{n-1}\right] + o\left(N^{n-1}\right)\right) \\ &= \operatorname{sign}\left(2nK\xi_i^1 + o\left(N^{n-1}\right)\right),\end{aligned} \tag{14}$$

where

$$K = \left(\overline{Y}_i^1\right)^{n-1} + \sum_{\mu=2}^{M}\left(\xi_i^1\xi_i^\mu\right)\left(\overline{X}_i^\mu\right)^{n-1}. \tag{15}$$

We can obtain the mean of K as

$$\bar{K} = E(K) = (1-2\delta)^{n-1}(N-1)^{n-1}, \tag{16}$$

where the expectation of the second part of K is 0, because ξ_i^1 and ξ_i^μ are i.i.d of Bernoulli $(\frac{1}{2})$ distribution.

Next we calculate the variance of K by the central moments [18,19]. Namely, when the mean of X is 0, and the variance is σ^2, for non-negative integer q, the plain central moments are: $E[X^q] = 0$ for q is odd and $E[X^q] = \sigma^q(q-1)!!$ for q is even, where σ is the standard deviation of X [18,19]. From this, with $2(n-1)$ even, we obtain the expectation of K^2,

$$\begin{aligned}E\left(K^2\right) &= E\Bigg(\left(\overline{Y}_i^1\right)^{2(n-1)} + 2\left(\overline{Y}_i^1\right)^{n-1}\sum_{\mu=2}^{M}\left(\xi_i^1\xi_i^\mu\right)\left(\overline{X}_i^\mu\right)^{n-1} \\ &\quad + \left[\sum_{\mu=2}^{M}\left(\xi_i^1\xi_i^\mu\right)\left(\overline{X}_i^\mu\right)^{n-1}\right]^2\Bigg) \\ &= E\Big((1-2\delta)^{2(n-1)}(N-1)^{2(n-1)} + 0 + \\ &\quad \left[\sum_{\mu=2}^{M}\left(\overline{X}_i^\mu\right)^{2(n-1)} + \sum_{\mu=2}^{M}\sum_{v=2,v\neq\mu}^{M}\left(\xi_i^1\xi_i^\mu\right)\left(\xi_i^1\xi_i^v\right)\left(\overline{X}_i^\mu\right)^{n-1}\left(\overline{X}_i^v\right)^{n-1}\right]\Big) \\ &= (1-2\delta)^{2(n-1)}(N-1)^{2(n-1)} + (M-1)(N-1)^{n-1}(2n-3)!!,\end{aligned} \tag{17}$$

where $\overline{X}_i^\mu$ has the mean 0 and the variance $N-1$. Then we can get the variance Σ^2 of K by $\Sigma^2 = E\left(K^2\right) - \bar{K}^2$,

$$\Sigma^2 = (M-1)(N-1)^{n-1}(2n-3)!!. \tag{18}$$

Since $2(n-1)$ is always even, regardless whether n is even or odd, the proof can be easily extended to the case when n is odd. This is the reason why when n is odd, the proof is essentially the same.

By the Eq. (14), we have $T_i\left(\bar{\xi}^1\right) = \xi_i^1$ if $K \geqslant 0$, and so, the probability P_{er} of the unstable state, defined above, is equal to the probability of the event $K < 0$. From this and the fact that for large N and M, the random variable K is Gaussian distributed, according to the central limit theorem [17].

$$\begin{aligned} P_{\text{er}} &= \int_{-\infty}^{0} \frac{\exp\left(-\frac{(x-\bar{K})^2}{2\Sigma^2}\right)}{\sqrt{2\pi\Sigma^2}}\, dx \\ &= \tfrac{1}{2}\left[1 - \operatorname{erf}\left(\frac{\bar{K}}{\sqrt{2\Sigma^2}}\right)\right] \\ &< \frac{\sqrt{2\Sigma^2}\exp\left(-\frac{\bar{K}^2}{2\Sigma^2}\right)}{2\sqrt{\pi}\bar{K}} \end{aligned} \tag{19}$$

for

$$\frac{\bar{K}^2}{\Sigma^2} > 2, \tag{20}$$

where

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-u^2}\, du \tag{21}$$

The inequality (19) follows by noting that the fact is

$$1 - \operatorname{erf}(x) = \frac{e^{-x^2}}{x\sqrt{\pi}}\sum_{n=0}^{\infty}(-1)^n\frac{(2n-1)!!}{(2x^2)^n}. \tag{22}$$

Additionally, $1 - \operatorname{erf}(x) < \frac{e^{-x^2}}{x\sqrt{\pi}}$ for $x > 1$ [20] [21].

Then we combine the Eq. (12) and the Eq. (19), and obtain the square:

$$\begin{aligned} \tfrac{1}{2\pi}\tfrac{\Sigma^2}{\bar{K}^2}\exp\left(-\tfrac{\bar{K}^2}{\Sigma^2}\right) &< \varepsilon^2 \\ \tfrac{1}{2\pi\varepsilon^2} &< \tfrac{\bar{K}^2}{\Sigma^2}\exp\left(\tfrac{\bar{K}^2}{\Sigma^2}\right) \\ -\ln 2\pi\varepsilon^2 - \tfrac{\bar{K}^2}{\Sigma^2} &< \ln\tfrac{\bar{K}^2}{\Sigma^2}. \end{aligned} \tag{23}$$

The intersection of the curve $y = \ln x$ and the curve $y = -x - \ln 2\pi\varepsilon^2$ is denoted by (x_0, y_0), The curve $y = \ln x$ and the x axis intersect at the point $(1, 0)$, and the curve $y = -x - \ln 2\pi\varepsilon^2$ and the x axis intersect at the point $(-\ln 2\pi\varepsilon^2, 0)$. When $-\ln 2\pi\varepsilon^2 > 1$, according to the image of two curves, we know $1 < x_0 < -\ln 2\pi\varepsilon^2$,

$$1 < \frac{\bar{K}^2}{\Sigma^2} < -\ln 2\pi\varepsilon^2. \tag{24}$$

Combining the Eq. (20) and the Eq. (24), we obtain

$$2 < \frac{\bar{K}^2}{\Sigma^2} < -\ln 2\pi\varepsilon^2, \tag{25}$$

Additionally, when $-\ln 2\pi\varepsilon^2 \leqslant 2$, i.e. $\varepsilon \geqslant \frac{1}{\sqrt{2\pi e^2}}$, according to the image of two curves, we know the intersection $x_0 \leqslant 1$. This means that when $\frac{\bar{K}^2}{\Sigma^2} > 2$, the curve $y = \ln x$ is over the curve $y = -x - \ln 2\pi\varepsilon^2$, i.e. the above inequality also holds true. Combining the Eq. (20), we obtain that when $\varepsilon \geqslant \frac{1}{\sqrt{2\pi e^2}}, 2 < \frac{\bar{K}^2}{\Sigma^2}$ holds true, which is sufficient.

When $-\ln 2\pi\varepsilon^2 > 2$, i.e. $\varepsilon < \frac{1}{\sqrt{2\pi e^2}}$, the above inequality holds true when $\frac{\bar{K}^2}{\Sigma^2} > x_0$. Next, we solve the above inequality (25) with the mean (16) and the variance (18) of K. We obtain

$$\frac{(1-2\delta)^{2(n-1)}N^{n-1}}{-\ln 2\pi\varepsilon^2(2n-3)!!} < C_\delta < \frac{(1-2\delta)^{2(n-1)}N^{n-1}}{2(2n-3)!!}. \tag{26}$$

It must be emphasized that when $\delta = 0, C_\delta = \frac{N^{n-1}}{(-\ln 2\pi + 2\ln N)(2n-3)!!}$, with $\varepsilon = \frac{1}{N}$. We can observe that $\frac{N^{n-1}}{(-\ln 2\pi + 2\ln N)(2n-3)!!}$ is similar with the capacity formulation $\frac{N^{n-1}}{(2\ln N)(2n-3)!!}$ in [13]. Especially when N is large, caused that $-\ln 2\pi + 2\ln N \approx 2\ln N$, so $C_\delta \approx \frac{N^{n-1}}{(2\ln N)(2n-3)!!}$. Finally, we obtain that C_δ of DAM networks is between C_{low} and C_{up}, of which

$$C_{\text{low}} = \frac{(1-2\delta)^{2(n-1)}N^{n-1}}{-\ln 2\pi\varepsilon^2(2n-3)!!}, \tag{27}$$

$$C_{\text{up}} = \frac{(1-2\delta)^{2(n-1)}N^{n-1}}{2(2n-3)!!}. \tag{28}$$

In summary, when $\varepsilon < \frac{1}{\sqrt{2\pi e^2}}, C_\delta$ of DAM networks is between C_{low} and C_{up}. And when $\varepsilon \geqslant \frac{1}{\sqrt{2\pi e^2}}, C_\delta$ of DAM networks is below C_{up}.

4. Experiments

We conduct some experiments to validate the capacity theorem of DAM networks.

4.1. Data generation

The data generation procedure is as follows:

1. For each experiment, we generate a set of M independent messages $\mathscr{M} = \{\xi^1, \ldots, \xi^M\}$, where each message a length N vector of i.i.d. Bernoulli (1/2) binary RVs. The weight matrix of the DAM network is determined by the memory message set $\mathscr{M}$.
2. The testing set $\widetilde{\mathscr{P}}_{\text{test}} = \{(\xi^1, \mathbf{x}^{1,j}), \ldots, (\xi^M, \mathbf{x}^{M,j}) | j = 1, \ldots, J\}$ is constructed by generating J noisy probes for each message $\xi^\mu \in \mathscr{M}, \mu = 1, \ldots, M$. Here, $\mathbf{x}^{\mu,j}$ represents the j-th noisy probe generated by the message ξ^μ, $\mu = 1, \ldots, M$, and so the size of $\widetilde{\mathscr{P}}_{\text{test}}$ is MJ. Specifically, each noisy probe $\mathbf{x}^{\mu,j}, j = 1, \ldots, J$ and $\mu = 1, \ldots, M$, is generated by flipping every component of ξ^μ independently and with a probability δ.

4.2. Evaluation procedure

We consider two strategies, depending on the number of update steps, to study the capacity of DAM networks, which corresponds to the two capacity notions in the previous chapter. For the one-step update (OSU) strategy, we randomly update every component of the noisy probe once by the Eq. (8) (for n even) or the Eq. (9) (for n odd), which corresponds to the one-step δ-capacity C_δ. The k-step update (KSU) strategy repeatedly executes the one-step update strategy until the convergence or until a maximum number of the iteration k is reached, which corresponds to the k-step δ-capacity $C_\delta^{(k)}$. In all the experiments, k is chosen as 100, however, in the experiments the convergence was always attained in less than 10 step updates.

To experimentally demonstrate Theorem 1, we perform the following procedure:

1. **Stability.** According to the accuracy results under the OSU strategy, we first consider Theorem 1 when $\delta = 0.0$. The aim here is to verify that every message is stable, i.e., does not change its value under the KSU strategy, when the number of messages M is below the capacity in the theorem.
2. **Compute the accuracy of DAM networks on $\widetilde{\mathscr{P}}_{\text{test}}$ under the OSU strategy.** Here, for each probe, we apply the OSU strategy and compare the result to the message from which the probe

was originally constructed, where we declare an error if the two are not equal. We then compute the accuracy (fraction of successfully recovered messages) for different values of M.

3. **Compute the accuracy of DAM networks on $\widetilde{\mathscr{P}}_{\text{test}}$ under the KSU strategy.** Here we use a similar setup to Item 2 above, but with KSU instead of OSU.

In all experiments, the reported accuracy is the mean of 20 independent trials. Additionally, we change the number of messages for different experiments to obtain the capacity C_δ and $C_\delta^{(k)}$ of DAM networks. In all experiments, we choose $\varepsilon = \frac{1}{N}$ to identify if the random message set is (δ, ε)-retrievable, following the setup in [13]. For instance, when N is $100, \varepsilon = 0.01$, which is a very high accuracy. In other words, if the accuracy on $\widetilde{\mathscr{P}}_{\text{test}}$ is greater than $1 - \varepsilon$, we consider that the random message set is successfully stored, i.e. the DAM networks achieve (δ, ε)-retrieval. With the increase of M, we monitor the model accuracy. We denote the first $M + 1$ where the DAM networks do not successfully store messages. Thus the capacity of DAM networks equals M (We denote the first $M + 1$ when the accuracy first smaller than $1 - \varepsilon$. We choose M as our capacity of DAM networks).

4.3. The capacity evaluation

In this experiment, we demonstrate Theorem 1 for $N = 100, 200$ and 500. We also study the impact of various values of δ (the attraction radius) on the message set $\mathscr{M}$. We experiment using the following parameters setups:

- $N = 100, \varepsilon = 0.01, J = 100, n = 2$ or 3, and δ is chosen from $\{0.01, 0.02, 0.05, 0.1, 0.2\}$;
- $N = 200, \varepsilon = 0.005, J = 100, n = 2$ or 3, and δ is chosen from $\{0.01, 0.02, 0.05, 0.1, 0.2\}$;
- $N = 500, \varepsilon = 0.002$, trapped in the insufficient computing power, we adjust $J = 1, n = 2$ or 3, and δ is chosen from $\{0.01, 0.02, 0.05, 0.1, 0.2\}$.

We emphasize that the accuracy on the message set $\mathscr{M}$ is 1.0 under the OSU strategy for values of M that are below the capacity, as predicted by Theorem 1. In particular, for $\delta = 0.0$, we observe that all messages are stable when M is below C_δ. The DAM capacity for $\delta = 0.2$ is much smaller than the DAM capacity under other (smaller) values of δ that were tested. For the consistency of results, the curve on $\delta = 0.2$ is not included when we plot the image of the accuracy with the amount of messages. Due to the computational complexity we only perform experiments for $n \leqslant 3$ in this paper.

4.3.1. Behavior at $\boldsymbol{N = 100}$

Table 1 lists the lower bound of C_δ ("lower"), the upper bound of C_δ ("upper") calculated by Theorem 1 and the DAM capacity under the OSU strategy ("OSU"), the KSU strategy ("KSU") by experiments. All the capacity results are consistent with Theorem 1. Additionally, when $n = 2$ and $n = 3$, the capacities under the OSU strategy and the KSU strategy between the lower bound and the upper bound. It can be noted that when $n = 3$, under the KSU strategy, the capacity will increase, especially for big attraction radiuses, such as $\delta = 0.1$ or 0.2. But this increment is not as evident when $n = 2$.

Fig. 1 demonstrates the accuracy change with varying M for various attraction radiuses and under the KSU strategy when $n = 2$. It can be observed that when M is smaller than 12, the accuracy is high (greater than 0.99), which indicates that the noisy probe can be easily recovered. In addition, with the increase of M, the accuracy drops gradually. This demonstrates that when the amount of messages increases, the noise recovery becomes harder. For instance, when M is greater than 13, the accuracy significantly decreases and becomes unstable.

Table 1
The DAM capacity comparison under the lower bound of C_δ ("lower"), the upper bound of C_δ ("upper") calculated by Theorem 1 and the OSU strategy ("OSU"), the KSU strategy ("KSU") by experiments, when $N = 100, \delta \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$. When $n = 2$, the minimal interval of the DAM capacity is 1. While when $n = 3$, the minimal interval of the DAM capacity is 10.

n	type	0.01	0.02	0.05	0.1	0.2
2	lower	13	13	11	9	5
	OSU	13	13	13	10	6
	KSU	13	13	14	11	10
	upper	49	47	41	32	18
3	lower	418	385	297	186	59
	OSU	540	500	380	240	60
	KSU	570	570	500	500	190
	upper	1538	1416	1094	683	216

Fig. 2 shows the changing of accuracy by varying M for different attraction radiuses under the KSU strategy when $n = 3$. When M is smaller than 500, the accuracy is greater than 99.5 on all attraction radiuses. It should be noticed that when M is greater than 500, the accuracy drops significantly with the increase of M.

4.3.2. Behavior at $\boldsymbol{N = 200}$

Table 2 lists the lower bound of C_δ ("lower"), the upper bound of C_δ ("upper") calculated by Theorem 1 and the DAM capacity under the OSU strategy ("OSU"), the KSU strategy ("KSU") by experiments. Firstly, we ensure that all messages can converge to themselves, which agrees with Theorem 1 on $\delta = 0.0$. Then, we conduct experiments on $\widetilde{\mathscr{P}}_{\text{test}}$. When $n = 2$ and $n = 3$, the DAM capacity under OSU strategy is between C_{low} and C_{up}. Additionally, the DAM capacity under the OSU strategy and the DAM capacity under the KSU strategy are the same. This indicates that the KSU strategy cannot increase the accuracy when choosing $n = 2$. It is worth mentioning that the DAM capacity under the KSU strategy is greater than the DAM capacity under the OSU strategy, especially for bigger attraction radiuses, such as $\delta = 0.1$ or 0.2, which indicates that the KSU strategy make the noise recovering easily. Because under the OSU strategy, the noisy probe cannot directly converge to the targeted message on $\delta = 0.1$ or 0.2. But if we can adopt the KSU strategy, the probe also can converge to the targeted message. This is an intuitive understanding, and the theoretical analysis remains for the future work. It should be noted that when $n = 3$ under $\delta = 0.2$, the capacity under the KSU strategy (1500) is larger than the upper bound (864). It indicates that the KSU strategy can increase the DAM capacity, especially for a big attraction radius.

Fig. 3 demonstrates the accuracy curves with M for various attraction radiuses under the KSU strategy when $n = 2$. It can be observed that when M is smaller than 25, the accuracy is high (greater than 0.98), which indicates that the noisy probe is easily recovered. In addition, with the increase of M, the accuracy decreases. In other words, when the amount of messages increases gradually, the noise recovery becomes harder and harder. For instance, when M is greater than 25, the accuracy significantly decreases and becomes unstable.

Fig. 4 shows the accuracy curves with M for different attraction radiuses under the KSU strategy when $n = 3$. We observe that when M is smaller than $2,000$, the accuracy is more than 99.5 for all attraction radiuses. It can also be noticed that when M is greater than $2,000$, the accuracy drops significantly. Furthermore, the accuracy is not stable when M is greater than $2,000$.

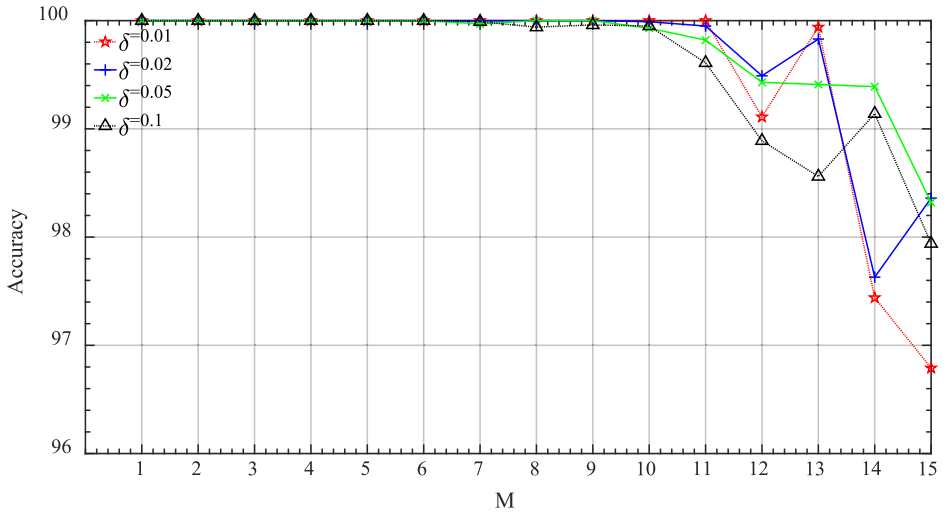


Fig. 1. The accuracy with the messages under the KSU strategy. $n = 2, N = 100, \delta \in \{0.01, 0.02, 0.05, 0.1\}$.

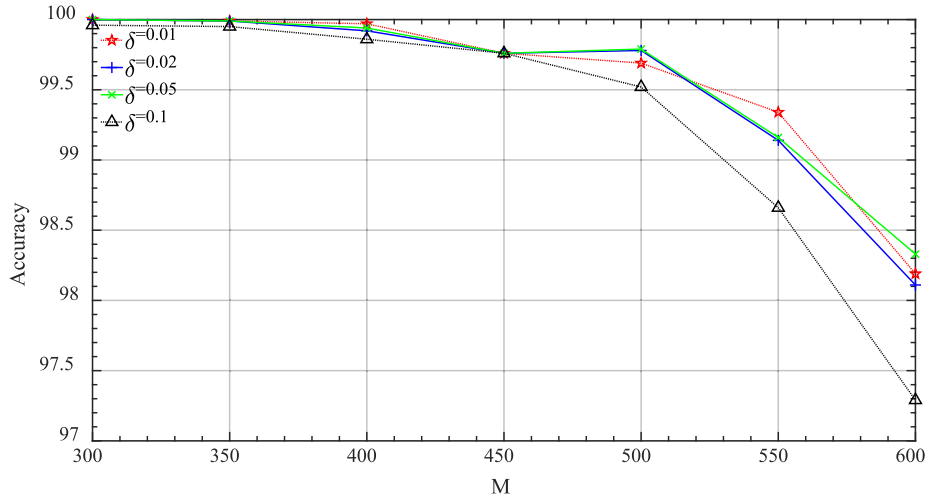


Fig. 2. The accuracy with the messages under the KSU strategy. $n = 3, N = 100, \delta = \{0.01, 0.02, 0.05, 0.1\}$.

Table 2
The DAM capacity comparison under the lower bound of C_δ ("lower"), the upper bound of C_δ ("upper") calculated by Theorem 1 and the OSU strategy ("OSU"), the KSU strategy ("KSU") by experiments, when $N = 200, \delta \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$. When $n = 2$, the minimal interval of the DAM capacity is 1. While when $n = 3$, the minimal interval of the DAM capacity is 100.

n	type	0.01	0.02	0.05	0.1	0.2
2	lower	22	22	19	15	9
	OSU	28	27	24	21	9
	KSU	28	27	24	23	20
	upper	97	93	81	64	36
3	lower	1405	1293	999	624	198
	OSU	1900	1800	1600	1300	500
	KSU	2000	2000	1900	1900	1500
	upper	6150	5663	4374	2731	864

4.3.3. Behavior at $N = 500$

Table 3 lists the lower bound of C_δ ("lower"), the upper bound of C_δ ("upper") calculated by Theorem 1 and the DAM capacity under the OSU strategy ("OSU"), the KSU strategy ("KSU") by experiments. All the capacity results are consistent with Theorem 1. Specifically, when $n = 2$ and $n = 3$, the capacities under the OSU strategy and the KSU strategy are between the lower bound and the upper bound. It should be noticed that the DAM capacity under the KSU strategy is the same with the DAM capacity under the OSU strategy, for $\delta \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$. The reason for this phenomenon may be that the accuracy demanded high under the large network.

Fig. 5 demonstrates the changing of accuracy by varying M for various attraction radiuses under the KSU strategy when $n = 2$. It can be observed that when M is smaller than 50, the accuracy is high (greater than 0.995), which indicates that the noisy probe can be easily recovered. In addition, with the increase of M, the accuracy drops gradually. This demonstrates that when the amount of messages increases, the noise recovery becomes difficult.

Fig. 6 shows the changing of accuracy by varying M for different attraction radiuses under the KSU strategy when $n = 3$. Except for the attraction radius $\delta = 0.1$, the accuracy decreases smoothly with the increase of M, and the accuracy is large than 0.98, which is a fairly high value. The phenomenon for the attraction radius $\delta = 0.1$ is that the accuracy fluctuates with the increase of M. The reason behind this phenomenon is that we generate just one sample for every message.

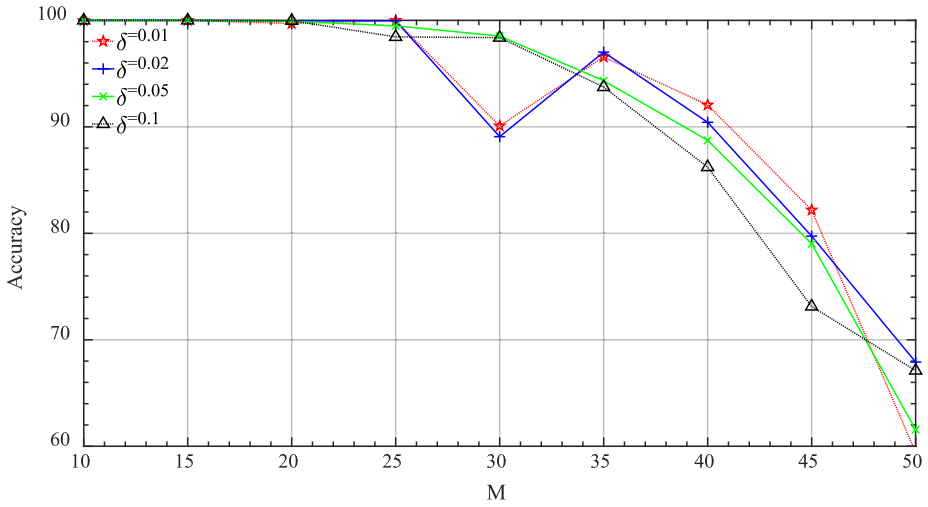


Fig. 3. The accuracy with the messages under the KSU strategy. $n = 2, N = 200, \delta \in \{0.01, 0.02, 0.05, 0.1\}$.

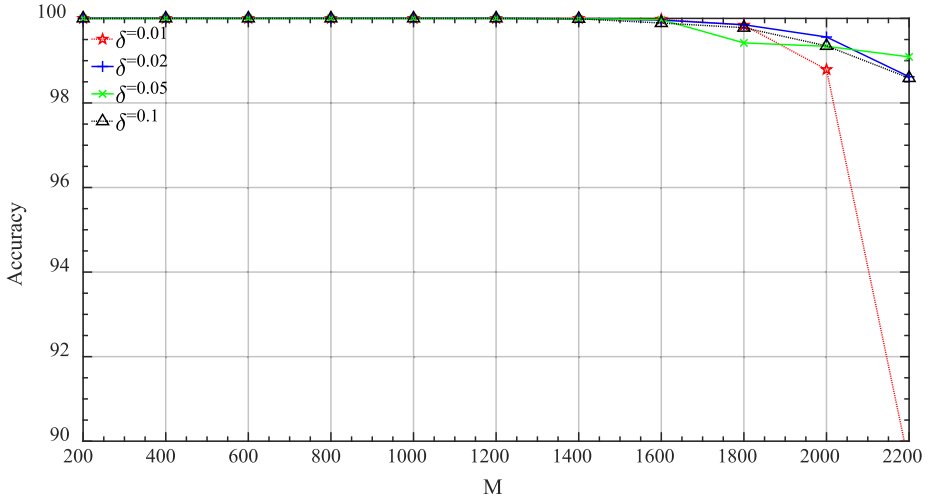


Fig. 4. The accuracy with the messages under the KSU strategy. $n = 3, N = 200, \delta \in \{0.01, 0.02, 0.05, 0.1\}$.

Table 3
The DAM capacity comparison under the lower bound of C_δ ("lower"), the upper bound of C_δ ("upper") calculated by Theorem 1 and the OSU strategy ("OSU"), the KSU strategy ("KSU") by experiments, when $N = 500, \delta \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$. When $n = 2$, the minimal interval of the DAM capacity is 1. While when $n = 3$, the minimal interval of the DAM capacity is 100.

n	type	0.01	0.02	0.05	0.1	0.2
2	lower	45	44	39	31	17
	OSU	48	46	42	40	40
	KSU	48	46	42	40	40
	upper	241	231	203	160	90
3	lower	7258	6683	5163	3223	1020
	OSU	8300	8200	8000	6500	2000
	KSU	8300	8200	8000	6500	2000
	upper	38432	35389	27338	17067	5400

5. Related works

5.1. Capacities of Hopfield networks and DAM networks

The study on the Hopfield capacity is an important research problem in Hopfield networks. The capacity of Hopfield networks is referred to that the number of stable messages in [1]. Then some researches rectify the Hopfield network in order to enlarge the capacity of Hopfield networks. For instance, when the memory needs to be perfect retrieval without errors, I. Kanter and H. Sompolinsky improve the capacity to $C = N$ (C is the maximize number of messages which are stable in Hopfield networks) by removing second–order correlations between the stored memories [22]. Additionally, when $N \to \infty$, the probability that all the patterns are stable is 1, and C is the largest number of these stable patterns, called absolute capacity in [23]. The absolute capacity of Hopfield network is shown to be $C = \frac{N}{\sqrt{2 \ln N}}$ by a local, incremental, and immediate method in [23]. In 2016, D. Krotov and J. J. Hopfield propose a generalized Hopfield network, called dense associative memory (DAM) networks, with a rectified polynomial energy function and revised update rule [13]. They demonstrate that the capacity of DAM networks is $\frac{N^{n-1}}{2(2n-3)!! \ln N}$ (n is power of the energy function). However, their paper [13] do not give the exact capacity study of DAM networks. There are theoretical discussion about the capacity of DAM networks in M. Demircigil et al. paper [24], but their study is concentrated on the capacity of DAM networks with the update rule changed, and the proof is very complicated and tedious. In summary, these studies have a critical weakness of intolerance to noises.

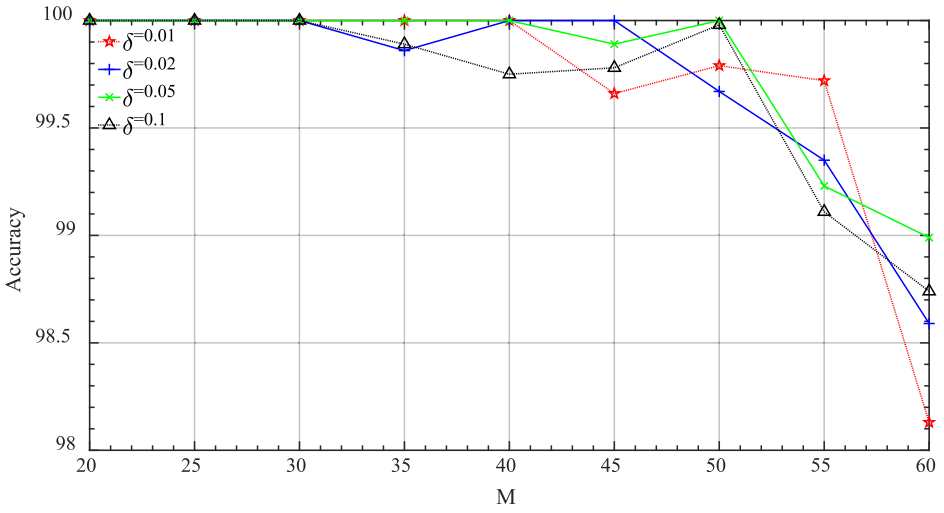


Fig. 5. The accuracy with the messages under the KSU strategy. $n = 2, N = 500, \delta \in \{0.01, 0.02, 0.05, 0.1\}$.

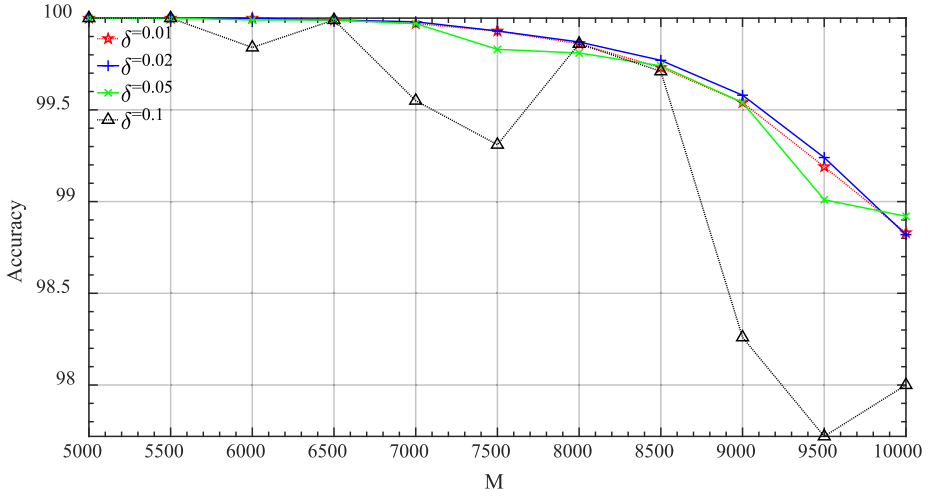


Fig. 6. The accuracy with the messages under the KSU strategy. $n = 3, N = 500, \delta \in \{0.01, 0.02, 0.05, 0.1\}$.

There are some references on the noise recovery of Hopfield networks. First, R. McEliece et al. [15] prove that the Hopfield network [1] can store $C = \frac{N(1-2\delta)^2}{2\ln N}$ if they require most of the messages are stable (a fixed but arbitrary pattern is stable), and $C = \frac{N(1-2\delta)^2}{4\ln N}$ if they ask for every one of the fundamental messages be recoverable exactly (all messages). Based on their paper [15], J. Komlos and R. Paturi show that each message has a non-trivial radius of attraction δ from a message by the large deviation theory [25] with a capacity result $C = \frac{N}{4\ln N}$. In their work, if the system is started at a state within a distance ρN from a fundamental message, it will reach a stable state within a distance of ϵN from the fundamental after a sequence of state transitions. Then in our paper $\epsilon = 0$, which is a very strong result. Lowe et al. study the storage capacity of the Hopfield model with correlated patterns in [26] with a new stable concept. In their paper, a pattern is called stable if it is close to a local minimum of the Hamiltonian or, in other words, if it is surrounded by a sufficiently high energy barrier. With this similar definition, J. Feng et al. estimate the critical capacity of the zero-temperature Hopfield model by using a novel approach based upon analysis of the Fourier transform of the joint distribution of the effective fields [6]. But their critical Hopfield capacity is small with $\alpha \leqslant 0.113$, which is a lower bound of the Hopfield capacity. There are other mathematical results about the lower bound of the Hopfield capacity in [27–29]. C. M. Newman obtains rigorous lower bounds of a certain simple neural network models of associative memory with N binary neurons and symmetric n-th order synaptic connections [27]. Based on his results [27], A. Bovier et al. bound the storage capacity of the dilute Hopfield network with a lower bound approximately 0.027, which generalizes the results of Newman for the standard Hopfield model [30]. Then, Loukianova generalizes the results for the standard Hopfield model with a lower bound approximately 0.05 [28]. Based on a crucial idea in [28], A. Bovier uses the negative association properties of some random variables to prove a sharp upper bound of the Hopfield network if the stored patterns are required to be fixed points [29]. In summary, we are more interested in the noise recovery in Hopfield network. Because it is more practical and easier to apply. So our paper is concentrated on the theoretical study of DAM networks in the noise recovery problem.

5.2. Other related works

In recent years, the associative memory network, such as the spin glass model attracts more and more researchers, especially physicists and mathematicians. The Hopfield model is an inverse spin glass model in the sense that in the case of the Hopfield model the energy minima are given (the stored pattern) where in the case of the Sherrington and Kirkpatrick spin glass model (SK model) the energy minima is unknown [31]. P. Baldi and S. S. Venkatesh study the number of stable points for spin-glasses and neural networks of higher orders, i.e., with Hamiltonians given by an algebraic form [32]. Their results indicate an increased capacity with the size of networks and the power of Hamiltonians. The phase transition of the Hopfield network with pure p-spin interactions on different temperatures are studied in [33]. In addition, D. J. Amit et al. [34,35] study the long-time behaviour between the spin glass model and the Hopfield model. They also discuss the thermodynamic and dynamic properties of the system in the cases of more general distributions of random memories. In 1998, M. Talagrand performs a thorough investigation on the validity of the replica symmetric solution of the Hopfield model [36]. Moreover, A. Barra et al. study a "hybrid" version of Restricted Boltzmann Machine (RBM), which thermodynamics of visible units is equivalent to those of Hopfield networks [37]. Their analysis the para-magnetic spin glass and the spin glass-retrieval phase transitions show that the presence of a retrieval phase is robust and not peculiar to the standard Hopfield model with Boolean patterns [4]. In 2013, E. Agliari et al. develop further the statistical mechanical analysis of associative network models, by studying the medium load regime, of which pattern-diluted associative networks is introduced as models for the immune system [38,39]. Finally E. Agliari et al. work out a general setting for evaluating the capacities of these p-spin models in [7]. Additionally in 2011, M. Löwe et al. analyze the storage capacity of the Hopfield model on a sparse graph [40]. Later, A. H. Salavati et al. propose a network based on a sparse bipartite graph with integer valued neural activity showing exponentially many stable states [41]. In summary, all these works are not concentrated on the capacity of Hopfield network. Their works are more about the phase transition and application of the spin glass model or discussion of Hopfield network on a sparse graph.

6. Conclusion

After D. Krotov and J. J. Hopfield proposing the dense associative memory (DAM) network, we mathematically analyse the capacity of DAM networks, and prove that the capacity C_δ of DAM networks of length N of which the noisy probe can be one-step (δ, ε)-retrievable can be bounded. In other words, when it is to be presented with a number C_δ of random independent $\{\pm 1\}^N$ Bernoulli $(\frac{1}{2})$ messages to store and when probing with a noisy probe N-tuple at most δN $(0 \leqslant \delta < \frac{1}{2})$ away from a message just after the one-step update, we have seen that when $\varepsilon < \frac{1}{\sqrt{2\pi e^2}}$, the capacity C_δ of DAM networks is at least $\frac{(1-2\delta)^{2(n-1)}N^{n-1}}{-\ln 2\pi\varepsilon^2(2n-3)!!}$ and at most $\frac{(1-2\delta)^{2(n-1)}N^{n-1}}{2(2n-3)!!}$. In addition, when $\varepsilon \geqslant \frac{1}{\sqrt{2\pi e^2}}$, the capacity C_δ of DAM networks is below $\frac{(1-2\delta)^{2(n-1)}N^{n-1}}{2(2n-3)!!}$. Furthermore, it must be emphasized that our capacity of DAM networks is obtained under the one-step update strategy, which is a strong conclusion. Moreover, our empirical studies on simulated experiments are consistent with the presented theorem in this paper, which demonstrates that the DAM networks have a powerful noise recovery by the theoretical analysis.

Last but not least, we observe that the KSU strategy can endure more noises than the OSU strategy under small networks, such as when $N = 100$, or 200, but the results are same between the OSU strategy and the KSU strategy when $N = 500$. The theoretical reason for this phenomenon is not clear, which is reserved for our future work.

CRediT authorship contribution statement

Han Bao: Methodology, Software, Formal analysis, Investigation, Resources, Writing - original draft. **Richong Zhang:** Resources, Writing - review & editing, Visualization, Project administration, Funding acquisition. **Yongyi Mao:** Conceptualization, Validation, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported partly by the National Natural Science Foundation of China (No. 61772059), by the Beijing Advanced Innovation Center for Big Data and Brain Computing (BDBC), by the Fundamental Research Funds for the Central Universities, by the Beijing S&T Committee (No. Z191100008619007) and by the State Key Laboratory of Software Development Environment (No. SKLSDE-2020ZX-14).

Appendix A. Proof of monotonicity of energy

There we prove that the energy is not increasing under the update rule (6) with $F(a) = a^n$. We adopt the enumeration method to prove it. Consider the input σ is chosen from $\{\pm 1\}^N, H_{i^+}$, and H_{i^-} is defined as the Eq. (7). When $\sigma_i = +1$.

1. If $H_{i^-}(\sigma) - H_{i^+} > 0$, according to the update rule (6), then $T_i(\sigma) = +1$, i.e. σ_i is stable. At this moment, $H_{i^-} > H_{i^+}$, the energy keeps unchanged.
2. If $H_{i^-}(\sigma) - H_{i^+} < 0$, according to the update rule (6), then $T_i(\sigma) = -1$, i.e. σ_i is unstable. At this moment, $H_{i^-} < H_{i^+}$, the energy decreases.

When $\sigma_i = -1$, similarly,

1. If $H_{i^-}(\sigma) - H_{i^+} > 0$, according to the update rule (6), then $T_i(\sigma) = +1$, i.e. σ_i is unstable. At this moment, $H_{i^-} > H_{i^+}$, the energy decreases.
2. If $H_{i^-}(\sigma) - H_{i^+} < 0$, according to the update rule (6), then $T_i(\sigma) = -1$, i.e. σ_i is stable. At this moment, $H_{i^-} < H_{i^+}$, the energy keeps unchanged.

In summary, the change of $F(a)$ from $\max(a, 0)^n$ to a^n cannot change the monotonicity of the energy, of which the energy will not increase. Moreover, based on the above proof, if we don't change the expression of the energy which is a difference of two energies, the monotonicity of the energy will keep unchanged.

Appendix B. Proof of Lemma

In our paper, we introduce $F(a) = a^n$ as the Eq. (5). Consider the case when n is even, then we deduce two parts of the Eq. (7)

$$\begin{aligned} H_{i^-} &= -\sum_{\mu=1}^{M}\left(-\xi_i^\mu + \sum_{j\neq i}\xi_j^\mu\sigma_j\right)^n = -\sum_{\mu=1}^{M}\sum_{k=1}^{n}\binom{n}{k}\left(-\xi_i^\mu\right)^{n-k}\left(\sum_{j\neq i}\xi_j^\mu\sigma_j\right)^k \\ H_{i^+} &= -\sum_{\mu=1}^{M}\left(\xi_i^\mu + \sum_{j\neq i}\xi_j^\mu\sigma_j\right)^n = -\sum_{\mu=1}^{M}\sum_{k=1}^{n}\binom{n}{k}\left(\xi_i^\mu\right)^{n-k}\left(\sum_{j\neq i}\xi_j^\mu\sigma_j\right)^k, \end{aligned} \tag{B.1}$$

According to the Eq. (6), subtracting with two above equations, we have

$$\begin{aligned} H_{i^-} - H_{i^+} &= \sum_{\mu=1}^{M}\left(\xi_i^\mu + \sum_{j\neq i}\xi_j^\mu\sigma_j\right)^n - \sum_{\mu=1}^{M}\left(-\xi_i^\mu + \sum_{j\neq i}\xi_j^\mu\sigma_j\right)^n \\ H_{i^-} - H_{i^+} &= 2\sum_{\mu=1}^{M}\sum_{k=1}^{\frac{n}{2}}\binom{n}{2k-1}\left(\xi_i^\mu\right)^{n-2k+1}\left(\sum_{j\neq i}\xi_j^\mu\sigma_j\right)^{2k-1}. \end{aligned} \tag{B.2}$$

For odd n, a similar result holds. Above the difference of two energies becomes

$$H_{i^-} - H_{i^+} = 2\sum_{\mu=1}^{M}\sum_{k=0}^{\frac{n-1}{2}}\binom{n}{2k}\left(\xi_i^\mu\right)^{n-2k}\left(\sum_{j\neq i}\xi_j^\mu\sigma_j\right)^{2k}. \tag{B.3}$$

Hence, the update rule can be written as Lemma 1:

$$T_i(\sigma) = \operatorname{sign}\left(2\sum_{\mu=1}^{M}\sum_{k=1}^{\frac{n}{2}}\binom{n}{2k-1}\left(\xi_i^\mu\right)^{n-2k+1}\left(\sum_{j\neq i}\xi_j^\mu\sigma_j\right)^{2k-1}\right), \tag{B.4}$$

for even n;

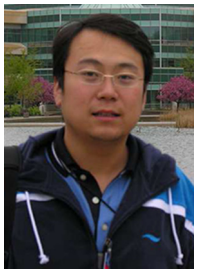
$$T_i(\sigma) = \operatorname{sign}\left(2\sum_{\mu=1}^{M}\sum_{k=0}^{\frac{n-1}{2}}\binom{n}{2k}\left(\xi_i^\mu\right)^{n-2k}\left(\sum_{j\neq i}\xi_j^\mu\sigma_j\right)^{2k}\right), \tag{B.5}$$

for odd n.

References

[1] J.J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proceedings of the National Academy of Sciences 79 (8) (1982) 2554–2558.
[2] W.A. Little, The existence of persistent states in the brain, in: From High-Temperature Superconductivity to Microminiature Refrigeration, Springer, 1974, pp. 145–164..
[3] S. Grossberg, Nonlinear neural networks: Principles, mechanisms, and architectures, Neural Networks 1 (1) (1988) 17–61.
[4] A. Barra, G. Genovese, P. Sollich, D. Tantari, Phase diagram of restricted boltzmann machines and generalized hopfield networks with arbitrary priors, Physical Review E 97 (2) (2018) 022310.
[5] J. Liu, M. Gong, H. He, Deep associative neural network for associative memory based on unsupervised representation learning, Neural Networks 113 (2019) 41–53.
[6] J. Feng, M. Shcherbina, B. Tirozzi, On the critical capacity of the hopfield model, Communications in Mathematical Physics 216 (1) (2001) 139–177.
[7] E. Agliari, F. Alemanno, A. Barra, A. Fachechi, Generalized guerra's interpolation schemes for dense associative neural networks, Neural Networks 128 (2020) 254–267.
[8] A. Fachechi, E. Agliari, A. Barra, Dreaming neural networks: forgetting spurious memories and reinforcing pure ones, Neural Networks 112 (2019) 24–40.
[9] D. Krotov, J. Hopfield, Dense associative memory is robust to adversarial inputs, Neural Computation 30 (12) (2018) 3151–3167.
[10] E. Agliari, F. Alemanno, A. Barra, M. Centonze, A. Fachechi, Neural networks with a redundant representation: Detecting the undetectable, Physical Review Letters 124 (2) (2020) 028301.
[11] E. Agliari, G. De Marzo, Tolerance versus synaptic noise in dense associative memories, The European Physical Journal Plus 135 (11) (2020) 1–22.
[12] F.M. de Paula Neto, A.J. da Silva, W.R. de Oliveira, T.B. Ludermir, Quantum probabilistic associative memory architecture, Neurocomputing 351 (2019) 101–110.
[13] D. Krotov, J.J. Hopfield, Dense associative memory for pattern recognition, in: Advances in Neural Information Processing Systems, 2016, pp. 1172–1180.
[14] H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, L. Gruber, M. Holzleitner, M. Pavlović, G.K. Sandve, V. Greiff, et al., Hopfield networks is all you need, arXiv preprint arXiv:2008.02217..
[15] R. McEliece, E. Posner, E. Rodemich, S. Venkatesh, The capacity of the hopfield associative memory, IEEE Transactions on Information Theory 33 (4) (1987) 461–482.
[16] H. Bao, R. Zhang, Y. Mao, J. Huai, Writing to the hopfield memory via training a recurrent network, in: Pacific Rim International Conference on Artificial Intelligence, Springer, 2019, pp. 241–254.
[17] R. Durrett, Probability: Theory and Examples, vol. 49, Cambridge University Press, 2019.
[18] G. Casella, R.L. Berger, Statistical Inference, vol. 2, Duxbury Pacific Grove, CA, 2002.
[19] H.H. Nguyen, E. Shwedyk, A First Course in Digital Communications, Cambridge University Press, 2009.
[20] A.A. Cuyt, V. Petersen, B. Verdonk, H. Waadeland, W.B. Jones, Handbook of Continued Fractions for Special Functions, Springer Science & Business Media, 2008.
[21] J.W. Craig, A new, simple and exact result for calculating the probability of error for two-dimensional signal constellations, in: MILCOM 91-Conference record, IEEE, 1991, pp. 571–575..
[22] I. Kanter, H. Sompolinsky, Associative recall of memory without errors, Physical Review A 35 (1) (1987) 380.
[23] A. Storkey, Increasing the capacity of a hopfield network without sacrificing functionality, in: Artificial Neural Networks–ICANN'97, 1997, pp. 451–456..
[24] M. Demircigil, J. Heusel, M. Löwe, S. Upgang, F. Vermet, On a model of associative memory with huge storage capacity, Journal of Statistical Physics 168 (2) (2017) 288–299.
[25] J. Komlós, R. Paturi, Convergence results in an associative memory model, Neural Networks 1 (3) (1988) 239–250.
[26] M. Löwe et al., On the storage capacity of hopfield models with correlated patterns, The Annals of Applied Probability 8 (4) (1998) 1216–1250.
[27] C.M. Newman, Memory capacity in neural network models: Rigorous lower bounds, Neural Networks 1 (3) (1988) 223–238.
[28] D. Loukianova, Lower bounds on the restitution error in the hopfield model, Probability Theory and Related Fields 107 (2) (1997) 161–176.
[29] A. Bovier, Sharp upper bounds on perfect retrieval in the hopfield model, Journal of Applied Probability 36 (3) (1999) 941–950.
[30] A. Bovier, V. Gayrard, Rigorous bounds on the storage capacity of the dilute hopfield model, Journal of Statistical Physics 69 (3–4) (1992) 597–627.
[31] A. Loettgers, The hopfield model and its role in the development of synthetic biology, in: 2007 International Joint Conference on Neural Networks, IEEE, 2007, pp. 1470–1475.
[32] P. Baldi, S.S. Venkatesh, Number of stable points for spin-glasses and neural networks of higher orders, Physical Review Letters 58 (9) (1987) 913.
[33] A. Bovier, B. Niederhauser, The spin-glass phase-transition in the hopfield model with p-spin interactions, arXiv preprint cond-mat/0108235..
[34] D.J. Amit, H. Gutfreund, H. Sompolinsky, Spin-glass models of neural networks, Physical Review A 32 (2) (1985) 1007.
[35] D.J. Amit, H. Gutfreund, H. Sompolinsky, Storing infinite numbers of patterns in a spin-glass model of neural networks, Physical Review Letters 55 (14) (1985) 1530.
[36] M. Talagrand, Rigorous results for the hopfield model with many patterns, Probability Theory and Related Fields 110 (2) (1998) 177–275.
[37] A. Barra, A. Bernacchia, E. Santucci, P. Contucci, On the equivalence of hopfield networks and boltzmann machines, Neural Networks 34 (2012) 1–9.
[38] E. Agliari, A. Annibale, A. Barra, A. Coolen, D. Tantari, Immune networks: multi-tasking capabilities at medium load, Journal of Physics A: Mathematical and Theoretical 46 (33) (2013) 335101.
[39] E. Agliari, A. Annibale, A. Barra, A. Coolen, D. Tantari, Immune networks: multitasking capabilities near saturation, Journal of Physics A: Mathematical and Theoretical 46 (41) (2013) 415003.
[40] M. Löwe, F. Vermet, The hopfield model on a sparse erdös-renyi graph, Journal of Statistical Physics 143 (1) (2011) 205–214.
[41] A.H. Salavati, K.R. Kumar, A. Shokrollahi, Nonbinary associative memory with exponential pattern retrieval capacity and iterative learning, IEEE Transactions on Neural Networks and Learning Systems 25 (3) (2014) 557–570.

Bao Han received his Bachelor Degree of Science from Beihang University, Beijing, China, in 2014. He is currently studying for a Ph.D in Computer Science and Technology in the Institute of Advanced Computing Technology, Beihang University. His research interests include Hopfield Network, Machine Learning and Deep Learning.



Richong Zhang received the B.Sc. and M.A.Sc. degrees from Jilin University, Changchun, China, in 2001 and 2004, respectively, the M.Sc. degree from Dalhousie University in 2006, and the Ph.D. from the School of Information Technology and Engineering, University of Ottawa, in 2011. He is currently an Associate Professor with the School of Computer Science and Engineering, Beihang University. His research interests include machine learing, natural language processing and recommender systems.

Yongyi Mao received the B.E. degree from Southeast University, in 1992, the medical degree from Nanjing Medical University, in 1995, the M.Sc. degree from the Department of Medical Biophysics, University of Toronto, in 1998, and the Ph.D. degree in electrical engineering from the University of Toronto, in 2003. He joined the Faculty of the School of Information Technology and Engineering, University of Ottawa, as an Assistant Professor. He was promoted to Associate Professor, in 2008, and then to Full Professor, in 2012. His main research interests include communications and machine learning.