Information Sciences 694 (2025) 121697

Contents lists available at ScienceDirect

Information Sciences

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

Binary associative memory networks: A review of mathematical framework and capacity analysis

Han Bao [a,b,*], Zhongying Zhao [a]

[a] School of Computer Science and Engineering, SDUST, 579 Qianwangang Road, Economic and Technical Development Zone, Qingdao, Shandong Province, PR China

[b] Shandong Minde Chemical Co., Ltd. Mount Lu Industrial Park, Yishui, Shandong Province, PR China

ARTICLE INFO

Keywords:
Binary associative memory networks
Capacity
Mathematical framework
Hopfield network

ABSTRACT

In recent years, heightened interest has been ignited in associative memory networks, largely attributed to their perceived equivalence with the attention mechanism, a fundamental component of the Transformer architecture. The opaque nature of deep neural networks, often characterized as “black boxes”, has intensified the pursuit of explainability, positioning associative memory networks as promising candidates for illuminating the inherent complexities of deep learning models. Despite their increasing prominence, the mathematical analysis of their capacity remains a significant research gap, which constitutes the central focus of this paper. To address this gap, we commence with a review of the mathematical framework underpinning associative memory networks, with particular emphasis on their binary configurations, drawing insights from the derivation of the dense associative memory model. Additionally, we review a systematic methodology for analyzing the capacity of binary associative memory networks, building upon established studies of dense associative memory networks. Utilizing this analytical framework, we derive the capacity of several prominent associative memory networks, including binary modern Hopfield networks and binary spherical Hopfield networks. Through comprehensive discussions and rigorous deductions, we aim to elucidate the characteristics of binary associative memory networks, thereby providing valuable insights and practical guidance for their effective application in real-world scenarios.

1. Introduction

In 2020, Ramsauer et al. introduced a groundbreaking Hopfield model, called the modern Hopfield network, which integrated an exponential activation function into the update rule [1]. The authors presented compelling evidence regarding the convergence and capacity of the model, highlighting its potential implications for elucidating the effectiveness of attention mechanisms in deep learning. However, the mathematical theory articulated in their work is intricate and may present challenges in terms of comprehension, particularly concerning the capacity study. Subsequently, Krotov and Hopfield developed a concise mathematical model for continuous-time associative memory networks [2]. They demonstrated that, contingent on the choice of activation functions, this mathematical framework can be deduced to various models of associative memory networks, including dense associative memory (DAM) networks [3], modern Hopfield networks [1], and spherical Hopfield networks [2]. Despite these notable advancements, there

* Corresponding author.
E-mail address: baohan@sdust.edu.cn (H. Bao).

https://doi.org/10.1016/j.ins.2024.121697
Received 3 August 2023; Received in revised form 11 November 2024; Accepted 23 November 2024

Available online 28 November 2024

is a noticeable lack of mathematical studies focusing specifically on binary associative memory networks. Consequently, it is imperative to summarize a mathematical framework for binary associative memory networks and to develop a robust mathematical analysis method capable of effectively investigating their capacity. Such an approach would not only catalyze the creation of novel models but also deepen our understanding of the fundamental mechanisms governing these networks.

The primary application scenario for associative memory networks revolves around noise recovery. In this context, when the input pattern constitutes a noisy rendition of a memory message, the associative memory network is anticipated to generate an output pattern corresponding to the targeted memory message. This noise recovery problem is fundamental and critical. In this paper, we rigorously investigate the capacity of binary associative memory networks by addressing the noise recovery problem within a mathematical framework. The term "binary" connotes the binary nature of neuron states. Notably, the work of Bao et al. extensively delves into the capacity of DAM networks concerning noise recovery [4]. One of our contributions is to extend this methodology to encompass other binary associative memory networks. We undertake a comprehensive analysis of various binary associative memory networks, including DAM networks, binary modern Hopfield networks, and binary spherical Hopfield networks. The paper is designed to summarize a mathematical analysis method for examining the capacity of these networks under the noise recovery scenario.

In this study, our contributions unfold in two principal aspects. Firstly, leveraging the foundational framework delineated in [2], we review the mathematical framework for associative memory networks, specifically focusing on its binary circumstance, inspired by the deduction of the dense associative memory model from the associative memory network. We provide detailed deductions elucidating the structures of DAM networks, binary modern Hopfield networks, and binary spherical Hopfield networks derived from our summarized mathematical framework. Secondly, we apply the methodology delineated in [4] for assessing memory capacity in binary associative memory networks, such as binary modern Hopfield networks and binary spherical Hopfield networks, which is the main contribution of this paper. It is crucial to highlight that our capacity analysis takes into account the presence of noises—a factor of practical significance in real-world applications, given that authentic data often includes noises. The investigation into memory capacity and the summarized mathematical framework of binary associative memory networks constitute the principal contributions of this paper, diligently addressing a crucial gap in the theoretical understanding of binary associative memory networks.

The paper is organized as follows: Section 2 provides an extensive literature review encompassing associative memory networks and deep neural networks, offering a comprehensive overview of the existing knowledge. In Section 3, we delve into the mathematical framework of binary associative memory networks, providing a meticulous exposition of their fundamental components and underlying principles. Moving forward, Section 4 engages in an intricate discussion on the capacity of associative memory networks, delivering a clear declaration of capacity. Built upon this foundation, Section 5 consolidates a mathematical analysis method explicitly crafted for investigating the capacity of binary associative memory networks. We present the theoretical framework and analytical techniques employed to evaluate the DAM networks, the binary modern Hopfield networks, and the binary spherical Hopfield networks, elucidating their respective memory capacity. Finally, Section 6 succinctly encapsulates the principal contributions, providing valuable insights and practical guidance for the effective deployment of associative memory networks in real-world scenarios. Additionally, we conclude with recommendations for future research and development, thereby solidifying the scientific significance of the study.

2. Related works

The concept of associative memory networks has a rich history within the field of neural networks, with early developments dating back to 1961 when the learning matrix is introduced for implementing hetero-associative memories using ferromagnetic properties [5]. However, a significant milestone in the advancement of associative memory networks is the introduction of the Hopfield neural network in 1982, which is widely recognized as a pioneering work in associative memory network [6]. The spin glass model, as an associative memory network, has garnered significant attention at the intersection of computer science, physics and mathematics. Baldi and Venkatesh investigate the number of stable points for spin glasses and higher-order neural networks, introducing several upper bounds on the number of programmable stable states, according to different storage schemes [7]. Additionally, Bovier et al. delve into the capacity of sparse associative memory networks [8] and underscore the significance of investigating truly sparse cases within the realm of associative memory networks [9]. Then they investigate the statistical mechanics of the Hopfield neural network with pure p-spin (p is the power of polynomial activation function of the Hopfield networks) interactions with even $p \geq 4$, and find that the capacity of the Hopfield network grows polynomially with network size [10]. In 2013, Agliari et al. extend the statistical mechanical analysis of associative network models by exploring the medium load regime, where pattern-diluted associative networks are introduced as models for the immune system [11][12]. They also propose a general framework for evaluating the capacities of these p-spin models [13]. Barra et al. conduct a study in 2012 on a "hybrid" version of the Restricted Boltzmann Machine (RBM), where the thermodynamics of visible units are equivalent to the Hopfield network [14]. Their investigations into the paramagnetic-spin glass and spin glass-retrieval phase transitions reveal that the presence of a retrieval phase is robust and extends beyond the standard Hopfield model with Boolean patterns [15]. Next, they emphasize that the stability of pseudo-stable or metastable states can be significantly reduced [16]. Furthermore, Smart and Zilman establish an exact correspondence between Hopfield networks and RBM in the context of correlated pattern Hopfield networks, employing the QR-decomposition technique [17].

The original Hopfield network, though influential, suffers from limited capacity, impeding its broader adoption and practical applicability. In the pursuit of enhancing the memory capacity of Hopfield networks, various successful methodologies have been explored. In this paper, we consistently designate N as the number of neurons in the associative memory network. The classical Hopfield network, renowned for its memory storage capabilities, has exhibited the ability to store messages, of which the capacity is about $0.14N$ [6]. Next, McEliece et al. reveal that, for fixed stable patterns, it can effectively store approximately $N/(2\ln N)$ random memory messages when most of them can be recovered. In the scenario where every memory message is recoverable, the

storage capacity decreases to around $N/(4\ln N)$ [18]. Moving forward, Mazza has provided a lower bound for the capacity of associative memory networks by employing certain bounded synaptic functions [19]. Additionally, when the learning rule deviates from the Hebbian learning rule, the memory capacity attains its maximum value of N [20]. Furthermore, Chiueh et al. introduce the Exponential Correlation Associative Memory (ECAM), a high-capacity associative memory network based on recurrent correlation associative memories. The asymptotic capacity of the ECAM scales exponentially with the length of memory patterns. The study includes the development of a tangible CMOS chip prototype and its application in vector quantization [21].

Apart from models closely associated with the Hopfield network, other associative memory models, such as morphological associative memories and fuzzy associative memories, have been explored. In 1998, Ritter et al. propose morphological associative memories based on lattice algebra. These memories are proven to be robust in the presence of noise and exhibit unlimited storage capacity [22]. Additionally, fuzzy associative memory, capable of storing patterns with varying degrees of membership, offers enhanced flexibility in capturing nuanced relationships between patterns. This increased flexibility in dealing with partial matches and fuzzy associations may contribute to greater capacity [23]. Moreover, another approach is the Hopfield network with a nonzero diagonal matrix, of which besides the well-known region, $C << N$ (C is the storage capacity) region, there is another region, at $C >> N$, where the recovery is highly effective. This approach provides insights into the trade-off between the number of stored patterns and the network size [24]. These findings showcase the various techniques utilized to expand the memory capacity of Hopfield networks, each offering unique advantages and considerations, which reignite scientific interest, leading to a surge of research publications.

Following the introduction of the modern Hopfield network, there are more researchers devoting their passion on studying associative memory networks. This growing interest has led to numerous significant contributions in various areas related to associative memory networks. Salvatori et al. propose a novel neural model for associative memories, based on a hierarchical generative network trained using predictive coding. This model receives external stimuli via sensory neurons and exhibits improved retrieval accuracy and robustness compared to existing associative memory models, including autoencoders trained via backpropagation and modern Hopfield networks [25]. Then Gabbur et al. offer a probabilistic interpretation of attention mechanisms, specifically in the context of interactive segmentation problems. They demonstrate the relationship between the dot product attention mechanism used in Transformers and Maximum A Posterior (MAP) inference, shedding light on the probabilistic nature of attention [26]. Additionally, Tyulmankov et al. focus on a special case of the modern Hopfield network and propose a biologically plausible learning rule using local three-factor synaptic plasticity. Their work introduces an intriguing approach for incorporating biological principles into the learning process of associative memory networks [27]. Consequently, Widrich et al. employ modern Hopfield networks to address the challenge of delayed rewards in reinforcement learning, focusing on decomposing rewards separated from their causative actions by irrelevant actions. They aim to enhance learning efficiency in reinforcement learning scenarios [28]. Recently, Liang et al. investigate a neuro-biological network motif found in the fruit fly brain that can learn semantic representations of words and generate static and context-dependent word embeddings in natural language processing (NLP). They also explore the network embedding problem from the perspective of modern Hopfield networks, offering novel insights into the field [29,30]. Next, Yoo and Wood propose a Bayes predictive coding associative memory that enables continual one-shot memory writing without meta-learning. Their work contributes to the development of efficient memory systems within associative memory networks [31]. Furthermore, Schafl et al. introduce a novel deep learning architecture called “Hopular” designed for medium and small-sized tabular datasets. This architecture utilizes continuous modern Hopfield networks at each layer, offering a unique approach to handle tabular data using associative memory networks [32]. In summary, these works underscore the myriad research endeavors dedicated to exploring various aspects, applications, and enhancements of associative memory networks in recent years.

In the realm of deep learning, researchers have been actively investigating alternative methods for storing and retrieving information beyond the traditional use of recurrent neural networks (RNNs). Several approaches have emerged in this pursuit, aiming to enhance memory capabilities in neural networks. One such approach involves linear memory networks that utilizes linear automatic encoders as memory sequences. Carta et al. introduce this concept, emphasizing the encoding of messages for effective storage and retrieval [33]. Associative memory models have also been proposed to augment RNNs and improve their capabilities. For instance, Danihelka et al. propose an associative memory model based on holographic representation, integrated within a long short-term memory (LSTM) framework. By leveraging the associative memory, their model demonstrates improved performance in tasks requiring memory retention and retrieval [34]. Besides, Ba et al. introduce a classical associative memory model as an extension to RNNs, providing additional associative memories to reinforce the memory capacity and network’s capabilities. This augmentation enables the network to store and retrieve information more effectively [35]. Moreover, Zhang et al. extend the Hopfield neural network to a learning matrix, providing insights into overcoming the limitations of existing models and enhancing the ability of RNNs to effectively remember long sequences. Their work explores the potential of associative memory in enhancing the performance of deep neural networks [36]. These related works demonstrate the ongoing efforts to explore various memory-enhancement techniques in deep learning. Currently, the most advanced language processing technology is a transformer-based architecture known as Bidirectional Encoder Representations from Transformers (BERT) [37]. BERT has demonstrated exceptional performance in various NLP tasks, showcasing the power and effectiveness of transformer-based models in language understanding and processing. The widespread adoption of deep neural networks has heightened the demand for explainability, a challenge that associative memory networks, with their biological associative memory characteristics, may effectively address.

3. The mathematical framework of the binary associative memory networks

In this section, we undertake a comprehensive review of the mathematical framework outlined by Krotov and Hopfield [2] pertaining to associative memory networks. Following this, we systematically derive the mathematical framework specifically tailored

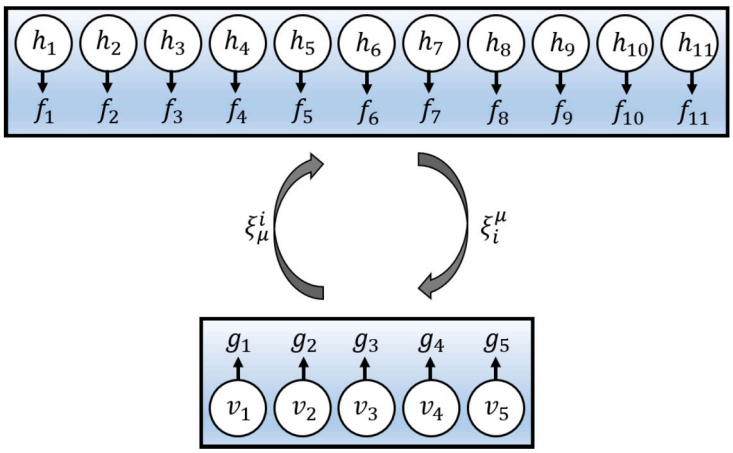

Fig. 1. An example of a continuous network with $N_f = 5$ feature neurons and $N_h = 11$ hidden neurons with symmetric synaptic connections between them.

for binary associative memory networks. Lastly, we present a detailed derivation of the mathematical structures that form the basis for DAM networks, binary modern Hopfield networks, and binary spherical Hopfield networks, all of which stem from our summarized framework. It is important to note that while the DAM network and the binary modern Hopfield network share similarities with their continuous formulations, the binary spherical Hopfield network departs significantly from its continuous counterpart as proposed by Krotov and Hopfield [2], as elaborated in Section 3.3.

In their study, Krotov and Hopfield propose a mathematical architecture derived from the DAM network [2]. This architectural configuration consists of N_f feature neurons and an additional set of N_h hidden neurons intricately coupled to feature neurons. It is noteworthy that the number of feature neurons corresponds to the network size N, i.e., $N_f = N$, while N_h equals the number of memory messages M intended for storage, i.e. $N_h = M$. Throughout the subsequent discussions, we will consistently refer to N_h to represent the number of memory messages, given that N_h is always equal to M. Their proposed associative memory network structure emphasizes "biological plausibility", characterized by synaptic connections limited to pairs of cells, specifically the i neuron in the feature layer and the j neuron in the hidden layer. The authors emphasize the importance of biological plausibility by noting the absence of many-body synapses. Importantly, there are synaptic connections between feature neurons and hidden neurons, but there are no connections among feature neurons and no connections among hidden neurons. Fig. 1 provides a visual representation of this network.

To establish clear and unambiguous notation, we provide comprehensive declarations in this paper. Initially, the associative memory network comprises N_f feature neurons coupled with an additional set of N_h hidden neurons. The value of the current in the feature neurons can be denoted by v_i, where i is represented by Latin indices. Similarly, the value of the current in the hidden neurons is denoted by h_μ, with μ using Greek indices. This notation ensures precision and clarity in describing the model's neurons. Additionally, the synaptic strength from feature neuron i to hidden neuron μ is expressed as ξ_μ^i, and the synaptic strength from hidden neuron μ to feature neuron i is denoted as ξ_i^μ. It is an important assumption that these synaptic strengths are symmetric, i.e., $\xi_i^\mu = \xi_\mu^i$. The outputs of feature neurons and hidden neurons are designated as g_i and f_μ, respectively, representing non-linear functions of their corresponding currents. Furthermore, the time constants for the two groups of neurons are symbolized by τ_f and τ_h. Thus, the continuous-time associative memory network model can be succinctly expressed as follows:

$$\begin{cases} \tau_f \dfrac{dv_i}{dt} = \displaystyle\sum_{\mu=1}^{N_h} \xi_i^\mu f_\mu - v_i + I_i, \\ \tau_h \dfrac{dh_\mu}{dt} = \displaystyle\sum_{i=1}^{N_f} \xi_\mu^i g_i - h_\mu, \end{cases} \tag{1}$$

where I_i denotes the input current into the feature neurons. The energy function for the network is given by

$$E(t) = \left[\sum_{i=1}^{N_f} (v_i - I_i) g_i - L_v\right] + \left(\sum_{\mu=1}^{N_h} h_\mu f_\mu - L_h\right) - \sum_{\mu,i} f_\mu \xi_\mu^i g_i, \tag{2}$$

where the $L_v(\{v_i\})$ and $L_h(\{h_\mu\})$ are Lagrangian functions for feature neurons and hidden neurons separately. In order to ensure that the derivatives of the Lagrangian functions correspond to the outputs of neurons, f_μ and g_i are such that Equation (3) holds,

$$f_\mu = \frac{\partial L_h}{\partial h_\mu}, \quad \text{and} \quad g_i = \frac{\partial L_v}{\partial v_i}. \tag{3}$$

First, we introduce the update time step notation, denoted by t, to facilitate our discussion of capacity. Specifically, in the context of discrete-time dynamics, this notation is straightforward. For continue-time dynamics, we can conceptualize the time as being divided into multiple discrete-time steps. It is consistent with the situation when the dynamics are fast, which mathematically corresponds to the limits $\tau_h \to 0$ and $\tau_f \to 0$. Additionally, we assume the absence of input currents to the feature neurons, expressed as $I_i = 0$. Consequently, the model specified by Equation (1) transforms into the standard associative memory networks:

$$\begin{cases} v_i^{(t+1)} = \sum_{\mu=1}^{N_h} \xi_i^\mu f_\mu^{(t)}, \\ h_\mu^{(t)} = \sum_{i=1}^{N_f} \xi_\mu^i g_i^{(t)}. \end{cases} \tag{4}$$

It is imperative to highlight that the computation of the values of the currents of the feature neurons at the time step $t+1$, involves aggregating the outputs from hidden neurons at the present time step t using the weight matrix. The value of the currents of the hidden neurons at the present time step t is determined by the outputs of the currents of the feature neurons at the same time step t. This distinction underscores that feature neurons are externally observable, while hidden neurons operate internally, contributing to the computation process.

Upon meticulous examination of the Lagrangian functions L_h and L_v, it becomes evident that L_h plays a pivotal role in influencing the memory capacity of the networks, while L_v is intricately linked to the network's output. Guided by this insight, we prescribe:

$$L_v = \sum_{i=1}^{N_f} |v_i|, \tag{5}$$

which ensures a binary output. Specifically, differentiate L_v with respect to v_i, we obtain:

$$g_i = \frac{\partial L_v}{\partial v_i} = \text{sign}[v_i]. \tag{6}$$

This choice facilitates a binary output, which is crucial in our analysis of the mathematical framework of binary associative memory networks. When the input v_i is binary, it follows directly that $g_i = \text{sign}[v_i] = v_i$, as the sign function preserves the binary nature of the input. With this configuration established, we proceed to derive the energy function of associative memory networks within the context of binary representations.

$$\begin{aligned} E(t) &= \left(\sum_{i=1}^{N_f} v_i g_i - L_v\right) + \left(\sum_{\mu=1}^{N_h} h_\mu f_\mu - L_h\right) - \sum_{\mu=1}^{N_h} f_\mu \sum_{i=1}^{N_f} \xi_\mu^i g_i \\ E(t) &= \left(\sum_{i=1}^{N_f} v_i \text{sign}[v_i] - \sum_{i=1}^{N_f} |v_i|\right) + \left(\sum_{\mu=1}^{N_h} h_\mu f_\mu - L_h\right) - \sum_{\mu=1}^{N_h} f_\mu h_\mu \\ E(t) &= -L_h. \end{aligned} \tag{7}$$

The synaptic connectivity pattern in the proposed network architecture adheres to the biological plausibility criterion by limiting the synaptic connections to pairs of cells: the feature neuron i and the hidden neuron μ. This constraint of two-cell synapses aligns with our understanding of biological neural networks, where pairs of individual neurons communicate through specific connections. Consequently, examining the dynamics of the energy function over time by taking its derivative with respect to time and utilizing the energy function (7), it becomes apparent that the energy consistently decreases as time progresses when the Hessian matrix of the Lagrangian functions L_h is positive semi-definite (see Appendix in [2] for details). On the one hand, this behavior implies that the network is driven towards a fixed point or a stable state, where the activities of the neurons settle into a steady configuration. On the other hand, it aligns with the concept of the associative memory, as the network's purpose is to store and retrieve specific stored memory messages reliably. In brief, the combination of limited synaptic connections, a balanced energy function, and the convergence to stable states highlight the biological plausibility and the functionality of associative memory networks. Last but not least, in the subsequent part of this section, we scrutinize three configurations of binary associative memory networks, distinguished by the utilization of different Lagrangian functions L_h. These comprise the DAM networks, the binary modern Hopfield networks, and the binary spherical Hopfield networks.

3.1. DAM networks

The DAM network is primarily characterized by the use of a polynomial function (n is the degree) in the hidden layer, corresponding to the choice of L_h as

$$L_h = \sum_{\mu=1}^{N_h} F(h_\mu) = \sum_{\mu=1}^{N_h} h_\mu^n. \tag{8}$$

Differentiate it as follows:

$$f_\mu = \frac{\partial L_h}{\partial h_\mu} = n h_\mu^{n-1}. \tag{9}$$

It is essential to emphasize that the preceding differentiation occurs within a single time step, denoted as t. Actually,

$$f_\mu^{(t)} = n\left(h_\mu^{(t)}\right)^{n-1}. \tag{10}$$

If there is no input current to feature neurons, i.e. $I_i = 0$, the energy function (2) reduces to

$$E(t) = -L_h = -\sum_{\mu=1}^{N_h} F(h_\mu^{(t)}) = -\sum_{\mu=1}^{N_h}\left(\sum_{i=1}^{N_f} \xi_\mu^i g_i^{(t)}\right)^n. \tag{11}$$

Derive the update rule from the first formula in (4), and substitute $f_\mu^{(t)}$ with the expression given in Equation (10):

$$v_i^{(t+1)} = \sum_{\mu=1}^{N_h} \xi_i^\mu f_\mu^{(t)} = \sum_{\mu=1}^{N_h} n\xi_i^\mu \left(h_\mu^{(t)}\right)^{n-1}. \tag{12}$$

Deduce the output of the feature neurons as follows:

$$g_i^{(t+1)} = \mathrm{sign}[v_i^{(t+1)}] = \mathrm{sign}\left[\sum_{\mu=1}^{N_h} n\xi_i^\mu \left(h_\mu^{(t)}\right)^{n-1}\right]. \tag{13}$$

This is the same with the update rule of DAM networks [3], affirming the correctness of our summarized mathematical framework for binary associative memory networks.

3.2. Binary modern Hopfield networks

In order to obtain the mathematical framework of binary modern Hopfield networks, we maintain the same form for L_h as outlined in [2], but change L_v as Equation (5).

$$L_h = \log\left[\sum_{v=1}^{N_h} \exp(h_v)\right]. \tag{14}$$

Obtain the expressions for f_μ in the time step t by differentiating the Lagrangian functions L_h as follows:

$$f_\mu^{(t)} = \frac{\partial L_h}{\partial h_\mu^{(t)}} = \frac{\exp\left(h_\mu^{(t)}\right)}{\sum_{v=1}^{N_h} \exp\left(h_v^{(t)}\right)} = \mathrm{softmax}\left(h_\mu^{(t)}\right). \tag{15}$$

In our binary modern Hopfield networks, there is also no input current to feature neurons, i.e. $I_i = 0$. Combining with Equation (6) and (15), the energy function (2) reduces to:

$$E(t) = -L_h = -\log\left[\sum_{v=1}^{N_h} \exp\left(h_v^{(t)}\right)\right]. \tag{16}$$

Deduce the update rule from the first formula in Equation (4), and substitute $f_\mu^{(t)}$ by the expression given in Equation (15),

$$v_i^{(t+1)} = \sum_{\mu=1}^{N_h} \xi_i^\mu f_\mu^{(t)} = \sum_{\mu=1}^{N_h} \xi_i^\mu \mathrm{softmax}\left(h_\mu^{(t)}\right). \tag{17}$$

Similar with the DAM model, the output of feature neurons in this binary modern Hopfield networks is as follows,

$$g_i^{(t+1)} = \mathrm{sign}[v_i^{(t+1)}] = \mathrm{sign}\left[\sum_{\mu=1}^{N_h} \xi_i^\mu \mathrm{softmax}\left(h_\mu^{(t)}\right)\right]. \tag{18}$$

It is consistent with the update rule of continue modern Hopfield networks in [1], confirming the naturalness from the continue state to the binary state.

3.3. Binary spherical Hopfield networks

The spherical Hopfield network is characterized by spherical normalization in the feature layer [2]. However, in our binary spherical Hopfield networks, the Lagrangian functions for feature neurons must be set to the sign function to ensure binary outputs. Consequently, it is naturally to incorporate spherical normalization into the Lagrangian functions for hidden neurons:

$$L_h = \sqrt{\sum_{v=1}^{N_h} \left(h_v^{(t)}\right)^2}. \tag{19}$$

Obtain the f_μ in the time step t by differentiating the above formulas as follows:

$$f_\mu^{(t)} = \frac{\partial L_h}{\partial h_\mu^{(t)}} = \frac{h_\mu^{(t)}}{\sqrt{\sum_{v=1}^{N_h} \left(h_v^{(t)}\right)^2}}. \tag{20}$$

Similar to the deduction of the binary modern Hopfield network, the energy function (2) simplifies to:

$$E(t) = -L_h = -\sqrt{\sum_{v=1}^{N_h} \left(h_v^{(t)}\right)^2}. \tag{21}$$

Then deduce the update rule from the first formula in Equation (4), and substitute $f_\mu^{(t)}$ by the expression given in Equation (20):

$$v_i^{(t+1)} = \sum_{\mu=1}^{N_h} \xi_i^\mu f_\mu^{(t)} = \sum_{\mu=1}^{N_h} \xi_i^\mu \left[\frac{h_\mu^{(t)}}{\sqrt{\sum_{v=1}^{N_h} \left(h_v^{(t)}\right)^2}} \right]. \tag{22}$$

Similar to the DAM networks, the output of feature neurons in binary spherical Hopfield networks is as follows,

$$g_i^{(t+1)} = \text{sign}\left[v_i^{(t+1)}\right] = \text{sign}\left\{ \sum_{\mu=1}^{N_h} \xi_i^\mu \left[\frac{h_\mu^{(t)}}{\sqrt{\sum_{v=1}^{N_h} \left(h_v^{(t)}\right)^2}} \right] \right\}. \tag{23}$$

We must emphasize that our deduced binary spherical Hopfield networks are markedly different from the spherical Hopfield networks proposed by [2]. This distinction underscores the necessity of our summarized mathematical framework for binary associative memory networks. By configuring the Lagrangian functions for hidden neurons to follow a binary output process, it naturally follows to incorporate spherical normalization into the Lagrangian functions for hidden neurons. However, we are currently unable to provide a detailed comparison of their advantages and disadvantages. A thorough exploration and analysis of these distinctions will be reserved for future research.

4. Discussion on the capacity of binary associative memory networks

In the process of retrieving a memory message, when confronted with a probe denoted as $\boldsymbol{v}$, the associative memory networks initialize their feature neurons as $\boldsymbol{v}$. The system then iterates through the update rule and the output function steps until convergence or until the maximum allowable number of iterations is reached. The final output of the feature neurons after this iteration process is declared as the retrieved pattern, which is expected to correspond to the targeted memory message. Considering Equation (4), we denote the single update of the feature neurons' output as a single step update, represented by T. We use $T^{(k)}$ to denote the k-fold composition $T \circ T \circ \cdots \circ T$ of the step update.

We provide a clear definition of convergence. A memory message, denoted as $\boldsymbol{\xi^\mu}$, is set as the synaptic strength, which is a length-N_f i.i.d. sequence drawn from the uniform distribution on $\{\pm 1\}$. The memory message is considered to be stored if and only if it remains stable under a finite number of retrieval step updates T. In other words, ξ_i^μ remains unchanged for all $i = 1, ..., N$ after the last step update T, $T(\xi_i^\mu) = \xi_i^\mu$. For any $\delta < 0.5$, a δ-perturbation of a message $\boldsymbol{\xi^\mu}$ refers to randomly flipping the sign of each component ξ_i^μ of $\boldsymbol{\xi^\mu}$ independently with a probability δ. Similarly, a probe $\bar{\boldsymbol{\xi}}^{\boldsymbol{\mu}}$ (which is a δ-perturbation of the message $\boldsymbol{\xi^\mu}$) is considered to have converged if and only if it remains stable under a finite number of step updates T. Specifically, after the last step update T, $T(\bar{\xi}_i^\mu) = \xi_i^\mu$ for all $i = 1, ..., N$.

The convergence step is a crucial aspect of research in Hopfield networks. Specifically, it addresses the question of how many iterations of the step update rule are needed for the probe to converge to the targeted memory message. McEliece et al. provide valuable insights into this issue [18]. They delineate three distinct types of convergence for a probe: 1) The probe may be directly attracted to the memory message, i.e., $T^{(\infty)}(\boldsymbol{v}) = T^{(1)}(\boldsymbol{v})$, occurring in one synchronous update iteration (synchronous update implies that the probe vector updates all units simultaneously); 2) For synchronous updates, the probe converges to the memory message

with a high probability in two synchronous update iterations, i.e., $T^{(\infty)}(\boldsymbol{v}) = T^{(2)}(\boldsymbol{v})$; 3) The probe converges to the memory message in many synchronous update iterations. It is noted that probes within a convergence domain typically exhibit the first two types of convergence. In other words, most probes converge to their targeted memory messages after just two synchronous update iterations. In our provided mathematical analysis method, we employ the first scenario to ascertain the maximum number of memory messages, acknowledging that this value is a conservative estimate, likely lower than the actual storage capacity of the networks.

The concept of capacity, as delineated by McEliece et al. in their seminal work, is fundamentally rooted in the rate of growth with the network size and encompasses two distinct scenarios [18]. One posits that every message represents a fixed point with high probability, while the other, a weaker concept, stipulates that almost all messages manifest as fixed points. The pseudo-inverse protocol is introduced in [38], where the stable and metastable states of the networks are studied as a function of temperature using mean-field theory, yielding a maximum capacity of N. Addressing the limitations of sacrificing local (the weight of a synapse depends only on information available to the neurons it connects) and incremental (learning a new memory pattern can be done knowing only the old weight matrix and not the actual patterns stored) quality, Storkey proposes a new algorithm for calculating the weight matrix, which is both local and incremental, resulting in a capacity of $N/\sqrt{2\ln N}$ [39]. Employing an energy function based on $F(x) = \exp(x)$ enables the Hopfield neural network to achieve an exceptional memory capacity of $2^{N/2}$ [40]. In 2016, Krotov and Hopfield make a groundbreaking contribution with the proposed DAM network. This two-layer Hopfield network employs polynomial activation functions and exhibits unexpectedly high capacity, surpassing the limitations of its predecessors [3]. Moreover, in 2020, Ramsauer et al. propose the modern Hopfield network with the exponential activation function, achieving even higher capacity than previous models [1]. In summary, while these works provide valuable insights into capacity, they lack a clear and explicit declaration of capacity except that McEliece et al. define capacity as a rate of growth of N, but the result in [18] is not entirely clear, like "high probability" and "almost all messages". In this paper, we mathematically declare the capacity for the binary associative memory network in the following.

After the preceding preparations, we proceed to establish a rigorous mathematical definition for the capacity of associative memory networks. In alignment with McEliece et al. [18], our study similarly defines capacity as a rate of growth relative to the size N_f of the binary associative memory networks. However, we aim to elucidate the convergence process with greater precision, encompassing not only the convergence step but also the resultant convergence situation. Our formulation strives for mathematical rigor, which is also referred in [4]. Consider a scenario wherein every message $\boldsymbol{\xi^\mu}$ in $\mathcal{M}$ ($\mathcal{M}$ has N_h messages) is a length-N_f i.i.d. sequence drawn from the uniform distribution $\{\pm 1\}$. For any $\delta \leq 0.5$, a δ-perturbation of a message $\boldsymbol{\xi^\mu}$ entails the random inversion of the sign of each component ξ_i^μ independently, each with a probability δ. We define the message set $\mathcal{M}$ as one-step (δ, ε)-retrievable if for each message $\boldsymbol{\xi^\mu} \in \mathcal{M}$, subsequent to applying a random δ-perturbation yielding a probe $\bar{\boldsymbol{\xi^\mu}}$, the probability that the one-step update under T fails to converge, is less than ε, for any given $\varepsilon > 0$, i.e. $P(\exists i : T(\bar{\boldsymbol{\xi^\mu}}) \neq \xi_i^\mu) < \varepsilon$, for any given $\varepsilon > 0$. The one-step δ-capacity C_δ of the associative memory network is then defined as the maximum size of the random message set $\mathcal{M}$, i.e. C_δ is the maximum N_h, for which $\mathcal{M}$ is one-step (δ, ε)-retrievable where $\varepsilon > 0$ when N_f is made sufficiently large. It is imperative to note that when $\delta = 0$, the one-step δ-capacity C_δ aligns with the classical capacity concept as presented in [3].

5. Capacity analysis method and discussions

Considering the detailed capacity analysis of the DAM network presented in [4], we develop a comprehensive mathematical analysis method for studying the capacity of binary associative memory networks. It is essential to underscore that our capacity analysis focuses specifically on binary patterns. When the network size N_f is sufficiently large, determining the magnitude of the one-step δ-capacity C_δ, of the associative memory network becomes a pivotal question. To tackle this inquiry, we establish a systematic mathematical analysis method for evaluating the capacity of binary associative memory networks, as outlined below:

- **Message Generation:** Consider a set $\mathcal{M}$ consisting of N_h independent memory messages. This message set can be represented by a binary matrix of size $N_f \times N_h$, where the elements of this matrix are generated by independent random variables with a probability of 0.5 taking values from the set $\{\pm 1\}$.
- **Noise Memory Message:** Generate a noise message $\bar{\boldsymbol{\xi}^{\mathbf{1}}}$ (with the index 1 chosen without loss of generality) by corrupting the original memory message $\boldsymbol{\xi}^{\mathbf{1}}$. Each element of the noise message is created by flipping the corresponding element of the original memory message with a probability of δ, representing the noise level.
- **Error Probability:** As previous discussed, the one-step δ-capacity is the maximum N_h, which is related to P_{er}, the probability that the noise message $\bar{\boldsymbol{\xi}^{\mathbf{1}}}$ fails to converge to the targeted memory message $\boldsymbol{\xi}^{\mathbf{1}}$ after a single step update. Specifically, P_{er} is defined as the probability that at least one component of the configuration does not match the corresponding element in the targeted message after a single step update:

$$P_{\text{er}} = P\left(\exists i \leq N : T(\bar{\xi_i^1}) \neq \xi_i^1\right). \tag{24}$$

- **Capacity Calculation:** Determine C_δ, the maximum number of memory messages that can be reliably stored and retrieved, by finding the largest set of memory messages for which the error probability P_{er} is less than a small positive value ε, for any given $\varepsilon > 0$, where ε represents the desired level of error tolerance:

$$P_{\text{er}} < \varepsilon, \quad \text{any given} \quad \varepsilon > 0. \tag{25}$$

- **Inequality Solution:** The inequality $P_{\text{er}} < \varepsilon$ can be reformulated as an inequality with N_f as the independent variable and N_h as the dependent variable. Solving this inequality provides the maximum N_h, which constitutes the one-step δ-capacity C_δ. Consequently, the solution reveals the maximum number of memory messages that can be stored and retrieved with a specified error tolerance level.

It should be noted that the aforementioned convergence condition is strict, as it excludes two scenarios. One is when the noise memory message converges to the targeted memory message after many update steps T, and the other one is when the noise memory message converges to a different memory message, rather than the targeted memory message. By introducing this strict convergence condition, the capacity analysis aims to determine the maximum number of binary memory messages that can be reliably stored and retrieved, considering both the noise level and the strict convergence to the intended memory message. This analysis provides valuable insights into the capacity limitations of binary associative memory networks, thereby contributing to the enhancement of pattern retrieval accuracy. By comprehending the network's capacity, we can effectively evaluate its performance and facilitate the reliable retrieval of desired patterns in diverse applications.

5.1. Capacity discussion of the DAM network

Detailed discussions on the capacity analysis of the DAM network are presented in [4], where it is demonstrated that when the network size N_f is sufficiently large, the one-step δ-capacity of the DAM network is approximately $C_\delta \approx O\left((1-2\delta)^{2(n-1)} N_f^{n-1}\right)$. This finding reveals that, on one hand, the capacity C_δ increases polynomially with the N_f. On the other hand, when δ increases, C_δ decreases as a power function. These capacity results provide valuable insights into the performance and scalability of DAM networks, showcasing their capability to handle larger memory sizes and effectively retrieve information in noisy scenarios.

5.2. Capacity discussion of the binary modern Hopfield network

During our capacity analysis of the binary modern Hopfield network, we introduce a noise message $\bar{\boldsymbol{\xi}^1}$ as the input probe $\boldsymbol{\nu}$ with a noise level δ from the targeted memory message $\boldsymbol{\xi}^1$. Specifically, for each component of $\boldsymbol{\xi}^1$, we flip it with the probability δ. In our deduction, we ignore the number of the time step, because we just consider single update step. Let us revisit the output of feature neurons at location i from $T(\bar{\xi_i^1}) = g_i(\bar{\xi_i^1}) = \text{sign}\left(\sum_{\mu=1}^{N_h} \xi_i^\mu f_\mu\right)$ in the following form:

$$T_i(\bar{\xi_i^1}) = \text{sign}\left(\sum_{\mu=1}^{N_h} \xi_i^\mu f_\mu\right) = \text{sign}\left\{\xi_i^1\left[f_1 + \sum_{\mu=2}^{N_h}(\xi_i^\mu \xi_i^1) f_\mu\right]\right\}$$

$$T_i(\bar{\xi_i^1}) = \text{sign}\left\{\frac{\xi_i^1}{\sum_{\nu=1}^{N_h}\exp(h_\nu)}\left[\exp\left(\sum_{j=1}^{N_f}\xi_j^1\bar{\xi_j^1}\right) + \sum_{\mu=2}^{N_h}\left(\xi_i^\mu\xi_i^1\right)\exp\left(\sum_{j=1}^{N_f}\xi_j^\mu\bar{\xi_j^1}\right)\right]\right\}. \tag{26}$$

Given $Y_1 = \sum_{j=1}^{N_f}\xi_j^1\bar{\xi_j^1}$, we can determine the mean of Y_1 as follows,

$$\text{Expt}(Y_1) = \sum_{j=1}^{N_f}\xi_j^1\left[(-\xi_j^1)\delta + \xi_j^1(1-\delta)\right] = (1-2\delta)N_f. \tag{27}$$

And

$$Y_1^2 = \sum_{j=1}^{N_f}(\xi_j^1\bar{\xi_j^1})^2 + \sum_{j=1}^{N_f}\sum_{k\neq j}^{N_f}(\xi_j^1\bar{\xi_j^1})(\xi_k^1\bar{\xi_k^1}), \tag{28}$$

$$\text{Expt}(Y_1^2) = (1-2\delta)^2 N_f + (1-2\delta)^2 N_f^2. \tag{29}$$

Therefore, we obtain the variance of Y_1, $\text{Var}(Y_1)$, using the formula $\text{Var}(Y_1) = \text{Expt}(Y_1^2) - (\text{Expt}(Y_1))^2$.

$$\text{Var}(Y_1) = (1-2\delta)^2 N_f + (1-2\delta)^2 N_f^2 - (1-2\delta)^2 N_f^2 = (1-2\delta)^2 N_f. \tag{30}$$

In probability theory, a log-normal distribution is a continuous probability of a random variable whose logarithm is normally distributed. Thus, if the random variable X is log-normally distributed, then $\ln X$ has a normal distribution. If the mean and the variance of $\ln X$ are denoted by $\text{Expt}(\ln X)$ and $\text{Var}(\ln X)$, then the mean and the variance of X are $\exp\left[\text{Expt}(\ln X) + \frac{\text{Var}(\ln X)}{2}\right]$ and $\{\exp[\text{Var}(\ln X)] - 1\}\exp\left[2\text{Expt}(\ln X) + \text{Var}(\ln X)\right]$, detailed deductions in Appendix A.

Set

$$K = \exp\left(\sum_{j=1}^{N_f}\xi_j^1\bar{\xi_j^1}\right) + \sum_{\mu=2}^{N_h}\left(\xi_i^\mu\xi_i^1\right)\exp\left(\sum_{j=1}^{N_f}\xi_j^\mu\bar{\xi_j^1}\right). \tag{31}$$

Then we proceed to solve for the mean of K,

$$\mathrm{Expt}(K) = \exp\left[(1-2\delta)N_f + \frac{1}{2}(1-2\delta)^2 N_f\right] + C_1 N_h \exp\left(\frac{N_f}{2}\right), \tag{32}$$

where $C_1 = \frac{e^{2(1-2\delta)}-1}{2e^{1-2\delta}}$. Next, we derive the expression for K^2,

$$K^2 = \exp 2\left(\sum_{j=1}^{N_f} \xi_j^1 \bar{\xi}_j^1\right) + 2\exp\left(\sum_{j=1}^{N_f} \xi_j^1 \bar{\xi}_j^1\right) \sum_{\mu=2}^{N_h} (\xi_i^\mu \xi_i^1) \exp(\xi_i^\mu \bar{\xi}_i^1) \exp\left(\sum_{j\neq i}^{N_f} \xi_j^\mu \bar{\xi}_j^1\right) + \left[\sum_{\mu=2}^{N_h} (\xi_i^\mu \xi_i^1) \exp(\xi_i^\mu \bar{\xi}_i^1) \exp\left(\sum_{j\neq i}^{N_f} \xi_j^\mu \bar{\xi}_j^1\right)\right]^2. \tag{33}$$

Subsequently, we decompose the above formula into three terms and call i, ii and iii the mean value of each of the terms. Then we can obtain i and ii easily as follows,

$$\begin{aligned} \mathrm{i} &= \exp\left[2(1-2\delta)N_f + 2(1-2\delta)^2 N_f\right], \\ \mathrm{ii} &= 2\exp\left[(1-2\delta)N_f + \frac{1}{2}(1-2\delta)^2 N_f\right] N_h C_1 \exp\left(\frac{N_f}{2}\right). \end{aligned} \tag{34}$$

Additionally, expand the third term of Equation (33):

$$\sum_{\mu=2}^{N_h} (\xi_i^\mu \xi_i^1)^2 \exp\left(2\xi_i^\mu \bar{\xi}_i^1\right) \exp 2\left(\sum_{j=1}^{N_f} \xi_j^\mu \bar{\xi}_j^1\right) + \sum_{\mu=2}^{N_h} \sum_{\nu\neq\mu}^{N_h} (\xi_i^\mu \xi_i^1) \exp(\xi_i^\mu \bar{\xi}_i^1) \exp\left(\sum_{j\neq i}^{N_f} \xi_j^\mu \bar{\xi}_j^1\right) (\xi_i^\nu \xi_i^1) \exp(\xi_i^\nu \bar{\xi}_i^1) \exp\left(\sum_{j\neq i}^{N_f} \xi_j^\nu \bar{\xi}_j^1\right), \tag{35}$$

and obtain the expectation of the third term:

$$\mathrm{iii} = N_h \exp(2N_f) + N_h^2 C_1^2 \exp(N_f). \tag{36}$$

Given $\mathrm{Expt}(K^2) = \mathrm{i} + \mathrm{ii} + \mathrm{iii}$, the $[\mathrm{Expt}(K)]^2$ can be calculated as follows:

$$\left[\mathrm{Expt}(K)\right]^2 = \exp\left[2(1-2\delta)N_f + (1-2\delta)^2 N_f\right] + 2\exp\left[(1-2\delta)N_f + \frac{1}{2}(1-2\delta)^2 N_f\right] C_1 N_h \exp\left(\frac{N_f}{2}\right) + C_1^2 N_h^2 \exp(N_f). \tag{37}$$

Using the formula $\mathrm{Var}(K) = \mathrm{Expt}(K^2) - \left[\mathrm{Expt}(K)\right]^2$, we can calculate the variance of K,

$$\mathrm{Var}(K) = \exp\left[2(1-2\delta)N_f + (1-2\delta)^2 N_f\right]\left[\exp(1-2\delta)^2 N_f - 1\right] + N_h \exp(2N_f). \tag{38}$$

According to the large deviation central limit theorem [41] [18], when N_f is large, K is log-normally distributed with the mean $\mathrm{Expt}(K)$ and the variance $\mathrm{Var}(K)$. We have $T(\bar{\xi}_i^1) = \xi_i^1$ if $K \geq 0$, and so, the probability P_{er} of unstable state, is equal to the probability of the event $K < 0$. Thus the i-component, $\bar{\xi}_i^1$, is stable responding to $P_{\mathrm{er}} < \varepsilon$, for any given $\varepsilon > 0$, if $K \geq 0$.

$$\begin{aligned} P_{\mathrm{er}} &= \int_{-\infty}^{0} \frac{\exp\left[-\frac{(x-\mathrm{Expt}(K))^2}{2\mathrm{Var}(K)}\right]}{\sqrt{2\pi\mathrm{Var}(K)}} \mathrm{d}\,x \\ &= \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{\mathrm{Expt}(K)}{\sqrt{2\mathrm{Var}(K)}}\right)\right] \\ &< \frac{\sqrt{2\mathrm{Var}(K)}\exp\left[-\frac{(\mathrm{Expt}(K))^2}{2\mathrm{Var}(K)}\right]}{2\sqrt{\pi}\mathrm{Expt}(K)}. \end{aligned} \tag{39}$$

There we use the following facts,

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2} \mathrm{d}\,u. \tag{40}$$

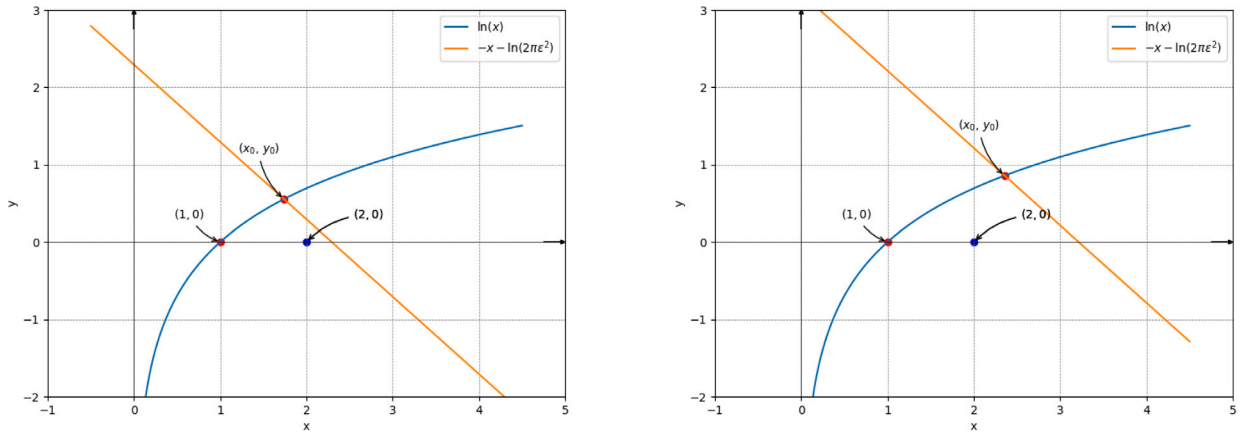

Fig. 2. The figure of $\ln x$ and $-x - \ln 2\pi\epsilon$. Left: $x_0 < 2$. Right: $x_0 \geq 2$.

Since

$$1 - \mathrm{erf}(x) = \frac{e^{-x^2}}{x\sqrt{\pi}} \sum_{n=0}^{\infty} (-1)^n \frac{(2n-1)!!}{(2x^2)^n}, \tag{41}$$

and when $x > 1$, $1 - \mathrm{erf}(x) < \frac{e^{-x^2}}{x\sqrt{\pi}}$ [42] [43], there is

$$\begin{aligned} &\frac{\mathrm{Expt}(K)}{\sqrt{2\mathrm{Var}(K)}} > 1 \\ &\frac{(\mathrm{Expt}(K))^2}{\mathrm{Var}(K)} > 2. \end{aligned} \tag{42}$$

Finally, combining Equation (39) and $P_{\mathrm{er}} < \epsilon$, for any given $\epsilon > 0$, we deduce the square of the inequality:

$$\begin{aligned} \frac{1}{2\pi} \frac{\mathrm{Var}(K)}{(\mathrm{Expt}(K))^2} \exp\left[-\frac{(\mathrm{Expt}(K))^2}{\mathrm{Var}(K)}\right] &< \epsilon^2 \\ \frac{1}{2\pi\epsilon^2} &< \frac{(\mathrm{Expt}(K))^2}{\mathrm{Var}(K)} \exp\left[\frac{(\mathrm{Expt}(K))^2}{\mathrm{Var}(K)}\right] \\ -\ln 2\pi\epsilon^2 - \frac{(\mathrm{Expt}(K))^2}{\mathrm{Var}(K)} &< \ln \frac{(\mathrm{Expt}(K))^2}{\mathrm{Var}(K)}. \end{aligned} \tag{43}$$

In the presented Fig. 2, the intersection point of the curves $y = \ln x$ and $y = -x - \ln 2\pi\epsilon^2$ is denoted as (x_0, y_0). The curve $y = \ln x$ intersects the x-axis at the point $(1, 0)$, while the curve $y = -x - \ln 2\pi\epsilon^2$ intersects the x-axis at $(-\ln 2\pi\epsilon^2, 0)$. Notably, the analysis is confined to the domain $x > 2$, according to Equation (42). Two scenarios are evident from the figure: one when $x_0 < 2$ and the other when $x_0 \geq 2$. Upon inspecting the left figure, it is observed that when $x_0 < 2$, the solution to the inequality (43) is $x > 2$. Conversely, from the examination of the right figure, when $x_0 \geq 2$, the solution to the inequality (43) is $x > x_0$. Consequently, to determine the maximum N_h, the inequalities $\frac{(\mathrm{Expt}(K))^2}{\mathrm{Var}(K)} > 2$ and $\frac{(\mathrm{Expt}(K))^2}{\mathrm{Var}(K)} > x_0$ need to be solved. Additionally, the value of x_0 can be ascertained through the Lambert W function.

The Lambert W function is a multivalued function that serves as the inverse of the function $f(y) = ye^y$. However, when restricting consideration to real numbers, attention is typically given to two specific branches of the function: W_0 and W_{-1}. In the range $-\frac{1}{e} \leq x < 0$, the solution for y can be obtained using $y = W_{-1}(x)$. Conversely, for $x \geq 0$, the solution for y is determined by $y = W_0(x)$. Through the appropriate branch of the Lambert W function, the value of x_0 in our specific context can be ascertained. Leveraging the increasing monotonicity of the Lambert W function, we derive:

$$x_0 = W_0\left(\frac{1}{2\pi\epsilon^2}\right). \tag{44}$$

Finally, the inequality $\frac{(\mathrm{Expt}(K))^2}{\mathrm{Var}(K)} > x_0$ can be reformulated as an inequality with N_f as the independent variable and N_h as the dependent variable (the solving process is straightforward, although the formulas involved are tedious, so they are omitted here). From this, we can derive that the N_h exhibits a growth rate relative to the size N_f. The analysis concludes that for a sufficiently

large network size N_f, the one-step δ-capacity C_δ of the binary modern Hopfield network exhibits an approximate exponential growth, expressed as $C_\delta \approx O\left(\exp\left[(1-2\delta)^2 N_f\right]\right)$, because C_δ is the maximum value of N_h. This finding is consistent with previous observations that C_δ grows exponentially with N_f [1] [40]. Furthermore, as δ increases, C_δ decreases exponentially, suggesting that the binary modern Hopfield network may perform relatively poorly at retrieval in practice with noisy queries, consistent with the conditions observed in continuous modern Hopfield networks [44]. We believe that our theoretical results regarding the capacity of the binary modern Hopfield network will be beneficial for practical applications.

5.3. Capacity discussion of the binary spherical Hopfield network

In our capacity analysis of the binary spherical Hopfield network, we focus on binary patterns. Specifically, we investigate the scenario where the input is a noise message denoted as $\bar{\boldsymbol{\xi}}^{\mathbf{1}}$. This probe is characterized by a noise level of δ from the targeted memory message $\boldsymbol{\xi}^{\mathbf{1}}$. By examining the update rule and considering the specific characteristics of binary patterns, we aim to gain insights into the network's capacity and its ability to handle noisy input scenarios. Similarly, we just consider single update step. Let us now revisit the output of feature neurons at location i from $T(\bar{\xi}_i^1) = g_i(\bar{\xi}_i^1) = \text{sign}\left(\sum_{\mu=1}^{N_h} \xi_i^\mu f_\mu\right)$ in the following form:

$$T(\bar{\xi}_i^1) = \text{sign}\left(\sum_{\mu=1}^{N_h} \xi_i^\mu f_\mu\right) = \text{sign}\left\{\xi_i^1\left[f_1 + \sum_{\mu=2}^{N_h}(\xi_i^\mu \xi_i^1) f_\mu\right]\right\}$$

$$T(\bar{\xi}_i^1) = \text{sign}\left\{\frac{\xi_i^1}{\sqrt{\sum_\nu^{N_h}(h_\nu)^2}}\left[\sum_{j=1}^{N_f}\xi_j^1\bar{\xi}_j^1 + \sum_{\mu=2}^{N_h}(\xi_i^\mu\xi_i^1)\left(\sum_{j=1}^{N_f}\xi_j^\mu\bar{\xi}_j^1\right)\right]\right\}. \tag{45}$$

Consider $Y_1 = \sum_{j=1}^{N_f}\xi_j^1\bar{\xi}_j^1$, the mean of Y_1 can be calculated as follows,

$$\text{Expt}(Y_1) = \sum_{j=1}^{N_f}\xi_j^1\left[(-\xi_j^1)\delta + \xi_j^1(1-\delta)\right] = (1-2\delta)N_f. \tag{46}$$

Since

$$Y_1^2 = \sum_{j=1}^{N_f}(\xi_j^1\bar{\xi}_j^1)^2 + \sum_{j=1}^{N_f}\sum_{k\neq j}^{N_f}(\xi_j^1\bar{\xi}_j^1)(\xi_k^1\bar{\xi}_k^1), \tag{47}$$

$$\text{Expt}(Y_1^2) = (1-2\delta)^2 N_f + (1-2\delta)^2 N_f^2. \tag{48}$$

Applying the formula $\text{Var}(Y_1) = \text{Expt}(Y_1^2) - (\text{Expt}(Y_1))^2$, we can obtain the variance of Y_1,

$$\text{Var}(Y_1) = (1-2\delta)^2 N_f + (1-2\delta)^2 N_f^2 - (1-2\delta)^2 N_f^2 = (1-2\delta)^2 N_f. \tag{49}$$

Set

$$K = \sum_{j=1}^{N_f}\xi_j^1\bar{\xi}_j^1 + \sum_{\mu=2}^{N_h}(\xi_i^\mu\xi_i^1)\left(\sum_{j=1}^{N_f}\xi_j^\mu\bar{\xi}_j^1\right). \tag{50}$$

By divide j into $j \neq i$ and $j = i$, it is easy to solve the mean of K,

$$\text{Expt}(K) = (1-2\delta)N_f + \text{Expt}\left[\sum_{\mu=2}^{N_h}(\xi_i^\mu\xi_i^1)\left(\xi_i^\mu\bar{\xi}_i^1 + \sum_{j\neq i}^{N_f}\xi_j^\mu\bar{\xi}_j^1\right)\right]$$

$$\text{Expt}(K) = (1-2\delta)N_f + \text{Expt}\left[\sum_{\mu=2}^{N_h}(\xi_i^\mu\xi_i^1)(\xi_i^\mu\bar{\xi}_i^1) + \sum_{\mu=2}^{N_h}(\xi_i^\mu\xi_i^1)\sum_{j\neq i}^{N_f}\xi_j^\mu\bar{\xi}_j^1\right]$$

$$\text{Expt}(K) = (1-2\delta)(N_f + N_h). \tag{51}$$

Next, we can calculate K^2 as follows:

$$K^2 = \left(\sum_{j=1}^{N_f}\xi_j^1\bar{\xi}_j^1\right)^2 + 2\left(\sum_{j=1}^{N_f}\xi_j^1\bar{\xi}_j^1\right)\left[\sum_{\mu=2}^{N_h}(\xi_i^\mu\xi_i^1)\left(\sum_{j=1}^{N_f}\xi_j^\mu\bar{\xi}_j^1\right)\right] + \left[\sum_{\mu=2}^{N_h}(\xi_i^\mu\xi_i^1)\left(\sum_{j=1}^{N_f}\xi_j^\mu\bar{\xi}_j^1\right)\right]^2. \tag{52}$$

Following that, we decompose the aforementioned formula into three parts and call i, ii and iii the mean value of each of the terms. Then we compute their respective expectations individually. Firstly, according to Equation (48), we obtain i:

$$\mathrm{i} = (1-2\delta)^2 N_f + (1-2\delta)^2 N_f^2. \tag{53}$$

Secondly, by dividing j into $j \neq i$ and $j = i$, we deduce ii:

$$\mathrm{ii} = \mathrm{Expt}\left\{2\left[\xi_i^1\bar{\xi}_i^1 + \left(\sum_{j\neq i}^{N_f}\xi_j^1\bar{\xi}_j^1\right)\right]\left[\sum_{\mu=2}^{N_h}(\xi_i^\mu\xi_i^1)\left(\xi_i^\mu\bar{\xi}_i^1 + \sum_{j\neq i}^{N_f}\xi_j^\mu\bar{\xi}_j^1\right)\right]\right\}$$

$$\begin{aligned}\mathrm{ii} = 2\mathrm{Expt}\Bigg\{&(\xi_i^1\bar{\xi}_i^1)\sum_{\mu=2}^{N_h}(\xi_i^\mu\xi_i^1)(\xi_i^\mu\bar{\xi}_i^1) + (\xi_i^1\bar{\xi}_i^1)\left[\sum_{\mu=2}^{N_h}(\xi_i^\mu\xi_i^1)\left(\sum_{j\neq i}^{N_f}\xi_j^\mu\bar{\xi}_j^1\right)\right]\\ &+\left(\sum_{j\neq i}^{N_f}\xi_j^1\bar{\xi}_j^1\right)\sum_{\mu=2}^{N_h}(\xi_i^\mu\xi_i^1)(\xi_i^\mu\bar{\xi}_i^1) + \left(\sum_{j\neq i}^{N_f}\xi_j^1\bar{\xi}_j^1\right)\left[\sum_{\mu=2}^{N_h}(\xi_i^\mu\xi_i^1)\left(\sum_{j\neq i}^{N_f}\xi_j^\mu\bar{\xi}_j^1\right)\right]\Bigg\}\end{aligned}$$

$$\mathrm{ii} = 2\left[N_h + 0 + (1-2\delta)^2 N_h N_f + 0\right] = 2\left[1 + (1-2\delta)^2 N_f\right] N_h. \tag{54}$$

Thirdly, expand the third term, by dividing j into $j \neq i$ and $j = i$, and applying the formula $\left[\sum_\mu (a_\mu + b_\mu)\right]^2 = \sum_\mu (a_\mu^2 + 2a_\mu b_\mu + b_\mu^2) + \sum_\mu \sum_{\nu\neq\mu}(a_\mu + b_\mu)(a_\nu + b_\nu)$, deduce its expectation:

$$\mathrm{iii} = \mathrm{Expt}\left(\left\{\sum_{\mu=2}^{N_h}\left[(\xi_i^\mu\xi_i^1)(\xi_i^\mu\bar{\xi}_i^1) + (\xi_i^\mu\xi_i^1)\left(\sum_{j\neq i}^{N_f}\xi_j^\mu\bar{\xi}_j^1\right)\right]\right\}^2\right)$$

$$\begin{aligned}\mathrm{iii} = \mathrm{Expt}\Bigg(\Bigg\{&\sum_{\mu=2}^{N_h}\left[(\xi_i^\mu\xi_i^1)^2(\xi_i^\mu\bar{\xi}_i^1)^2 + 2(\xi_i^\mu\xi_i^1)(\xi_i^\mu\bar{\xi}_i^1)(\xi_i^\mu\xi_i^1)\left(\sum_{j\neq i}^{N_f}\xi_j^\mu\bar{\xi}_j^1\right) + (\xi_i^\mu\xi_i^1)^2\left(\sum_{j\neq i}^{N_f}\xi_j^\mu\bar{\xi}_j^1\right)^2\right]\\ &+\sum_{\mu=2}^{N_h}\sum_{\nu\neq\mu}^{N_h}\Bigg[(\xi_i^\mu\xi_i^1)(\xi_i^\mu\bar{\xi}_i^1)(\xi_i^\nu\xi_i^1)(\xi_i^\nu\bar{\xi}_i^1) + (\xi_i^\mu\xi_i^1)(\xi_i^\mu\bar{\xi}_i^1)(\xi_i^\nu\xi_i^1)\left(\sum_{j\neq i}^{N_f}\xi_j^\nu\bar{\xi}_j^1\right)\\ &+ (\xi_i^\nu\xi_i^1)(\xi_i^\nu\bar{\xi}_i^1)(\xi_i^\mu\xi_i^1)\left(\sum_{j\neq i}^{N_f}\xi_j^\mu\bar{\xi}_j^1\right) + (\xi_i^\mu\xi_i^1)\left(\sum_{j\neq i}^{N_f}\xi_j^\mu\bar{\xi}_j^1\right)(\xi_i^\nu\xi_i^1)\left(\sum_{j\neq i}^{N_f}\xi_j^\nu\bar{\xi}_j^1\right)\Bigg]\Bigg\}\Bigg)\end{aligned}$$

$$\mathrm{iii} = (1-2\delta)^2 N_h + N_h N_f + (1-2\delta)^2 N_h^2. \tag{55}$$

Given $\mathrm{Expt}(K^2) = \mathrm{i} + \mathrm{ii} + \mathrm{iii}$, $\mathrm{Expt}(K^2)$ can be calculated,

$$\mathrm{Expt}(K^2) = (1-2\delta)^2(N_h + N_f)^2 + \left[2 + (1-2\delta)^2 + N_f\right] N_h + (1-2\delta)^2 N_f. \tag{56}$$

According to $\mathrm{Var}(K) = \mathrm{Expt}(K^2) - \left[\mathrm{Expt}(K)\right]^2$, the variance of K can be obtained:

$$\mathrm{Var}(K) = \left[2 + (1-2\delta)^2 + N_f\right] N_h + (1-2\delta)^2 N_f. \tag{57}$$

Finally, the calculation of the capacity for the binary spherical Hopfield networks follows a similar procedure, encompassing the deduction from Equation (39) to Equation (43) as in the case of the binary modern Hopfield networks. The analysis regarding the solution of Equation (43) remains consistent with the previous subsection. By solving the inequality $\frac{(\mathrm{Expt}(K))^2}{\mathrm{Var}(K)} > x_0$ and $\frac{(\mathrm{Expt}(K))^2}{\mathrm{Var}(K)} > 2$, we can conclude that for a sufficiently large N_f, the one-step δ-capacity C_δ of the binary spherical Hopfield network linearly increased as N_f, denoted as $C_\delta \approx O\left((1-2\delta)^2 N_f\right)$. This implies that the network possesses a capacity to store and retrieve patterns that scales linearly with the number of feature neurons. Additionally, the term $(1-2\delta)^2$ showcases that the area around the attraction state is precipitous. Further discussions and applications about the spherical Hopfield networks will be reserved for future work.

6. Summary

In this paper, we review a rigorous mathematical framework for binary associative memory networks and conduct an in-depth exploration of dense associative memory (DAM) networks, binary modern Hopfield networks, and binary spherical Hopfield networks. Our primary focus is on elucidating their mathematical derivations and fundamental principles within the realm of binary associative memory. We also highlight that our derived spherical Hopfield network differs from the formulation presented in Krotov's work [2], with a detailed comparative analysis to be addressed in future research. Finally, we summarize a comprehensive mathematical analysis

approach tailored for assessing the capacity of binary associative memory networks, building upon the methodologies established in [4]. By applying this framework, we evaluate the theoretical capacities of the aforementioned networks, thereby providing valuable insights for their practical applications.

Our future research endeavors will focus on exploring and leveraging the potential applications of dense associative memory (DAM) networks, binary modern Hopfield networks, and binary spherical Hopfield networks across diverse domains, including pattern recognition, image generation, and classification. Given the well-established theoretical capacities of DAM networks and binary modern Hopfield networks, we anticipate their significant impact in various practical applications. Additionally, we recognize the versatility of binary spherical Hopfield networks, which are particularly well-suited for scenarios involving memory messages with spherical distributions, as indicated by our mathematical framework. Notably, our analytical methodology can serve as a valuable guideline for capacity studies in continuous (non-binary) associative memory networks.

CRediT authorship contribution statement

Han Bao: Writing – original draft, Visualization, Validation, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Zhongying Zhao:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Data curation.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used GPT-3.5 in order to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

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 primarily supported by the Elite Foundation of Shandong University of Science and Technology (0104060540516) and the Youth Fund of Natural Science Foundation of Shandong Province (ZR2024QF269). Additionally, we sincerely appreciate the funding from Shandong Minde Chemical Co., Ltd. Their financial support has been invaluable in advancing our research efforts.

Appendix A. Expectation and variation of lognormal distribution

The lognormal distribution is a random variable with a continuous probability of which the logarithm is a normal distribution. If the random variable X is lognormal distribution, the $\ln X$ is a normal distribution. Additionally, if the mean and the variance of $\ln X$ is $\mathrm{Expt}(\ln X)$ and $\mathrm{Var}(\ln X)$, the mean and the variance of the random variable X is $\exp\left[\mathrm{Expt}(\ln X)+\frac{\mathrm{Var}(\ln X)}{2}\right]$ and $\{\exp[\mathrm{Var}(\ln X)]-1\}\exp\left[2\mathrm{Expt}(\ln X)+\mathrm{Var}(\ln X)\right]$. The detailed deduction is as follows:

Proof 1.

$$
\begin{aligned}
\mathrm{Expt}(X) &= \int_0^{+\infty} x\frac{1}{\sqrt{2\pi\mathrm{Var}(\ln X)x}}\exp\left[-\frac{(\ln x-\mathrm{Expt}(\ln X))^2}{2\mathrm{Var}(\ln X)}\right]\mathrm{d}x \\
&\xLeftrightarrow[\mathrm{d}x=\exp(t)\mathrm{d}t]{x=\exp(t)} \frac{1}{\sqrt{2\pi\mathrm{Var}(\ln X)}}\int_{-\infty}^{+\infty}\exp\left[-\frac{(t-\mathrm{Expt}(\ln X))^2}{2\mathrm{Var}(\ln X)}\right]\exp(t)\,\mathrm{d}t \\
&= \frac{1}{\sqrt{2\pi\mathrm{Var}(\ln X)}}\int_{-\infty}^{+\infty}\exp\left[-\frac{(t^2-2\mathrm{Expt}(\ln X)t+\mathrm{Expt}(\ln X)^2)}{2\mathrm{Var}(\ln X)}+t\right]\mathrm{d}t \\
&= \frac{1}{\sqrt{2\pi\mathrm{Var}(\ln X)}}\int_{-\infty}^{+\infty}\exp\left\{-\frac{\left[t-(\mathrm{Expt}(\ln X)+\mathrm{Var}(\ln X))\right]^2}{2\mathrm{Var}(\ln X)}+\left(\mathrm{Expt}(\ln X)+\frac{\mathrm{Var}(\ln X)}{2}\right)\right\}\mathrm{d}t \\
&= \frac{1}{\sqrt{2\pi\mathrm{Var}(\ln X)}}\int_{-\infty}^{+\infty}\exp\left\{-\frac{\left[t-(\mathrm{Expt}(\ln X)+\mathrm{Var}(\ln X))\right]^2}{2\mathrm{Var}(\ln X)}\right\}\mathrm{d}t\exp\left(\mathrm{Expt}(\ln X)+\frac{\mathrm{Var}(\ln X)}{2}\right) \\
&= \exp\left[\mathrm{Expt}(\ln X)+\frac{\mathrm{Var}(\ln X)}{2}\right]. \qquad (A.1)
\end{aligned}
$$

According to $\mathrm{Var}(X) = \mathrm{Expt}(X^2) - \left[\mathrm{Expt}(X)\right]^2$ and $\mathrm{Expt}(X) = \exp\left[\mathrm{Expt}(\ln X) + \frac{\mathrm{Var}(\ln X)}{2}\right]$, we deduce $\mathrm{Expt}(X^2)$:

$$\begin{aligned}
\mathrm{Expt}(X^2) &= \int_0^{+\infty} x^2 \frac{1}{\sqrt{2\pi}\sigma x} \exp\left[-\frac{(\ln x - \mathrm{Expt}(\ln X))^2}{2\mathrm{Var}(\ln X)}\right] \mathrm{d}x \\
&\xLeftrightarrow[\mathrm{d}x=\exp(t)\mathrm{d}t]{x=\exp(t)} \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \exp(t) \exp\left[-\frac{(t - \mathrm{Expt}(\ln X))^2}{2\mathrm{Var}(\ln X)}\right] \exp(t)\, \mathrm{d}t \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \exp\left[-\frac{(t^2 - 2\mathrm{Expt}(\ln X)t + \mathrm{Expt}(\ln X)^2)}{2\mathrm{Var}(\ln X)} + 2t\right] \mathrm{d}t \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \exp\left\{-\frac{\left[t - (\mathrm{Expt}(\ln X) + 2\mathrm{Var}(\ln X))\right]^2}{2\mathrm{Var}(\ln X)} + (2\mathrm{Expt}(\ln X) + 2\mathrm{Var}(\ln X))\right\} \mathrm{d}t \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \exp\left\{-\frac{\left[t - (\mathrm{Expt}(\ln X) + 2\mathrm{Var}(\ln X))\right]^2}{2\mathrm{Var}(\ln X)}\right\} \mathrm{d}t \exp\left(2\mathrm{Expt}(\ln X) + 2\mathrm{Var}(\ln X)\right) \\
&= \exp\left[2\mathrm{Expt}(\ln X) + 2\mathrm{Var}(\ln X)\right]. \qquad \text{(A.2)}
\end{aligned}$$

Considering $\mathrm{Var}(X) = \mathrm{Expt}(X^2) - \left[\mathrm{Expt}(X)\right]^2$, we obtain the variance of X:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{Expt}(X^2) - \left[\mathrm{Expt}(X)\right]^2 \\
&= \exp(2\mathrm{Expt}(\ln X) + 2\mathrm{Var}(\ln X)) - \exp\left[2\left(\mathrm{Expt}(\ln X) + \frac{\mathrm{Var}(\ln X)}{2}\right)\right] \\
&= \exp\left[2\mathrm{Expt}(\ln X) + \mathrm{Var}(\ln X)\right] \left\{\exp\left[\mathrm{Var}(\ln X)\right] - 1\right\}. \qquad \text{(A.3)}
\end{aligned}$$

Data availability

No data was used for the research described in the article.

References

[1] H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, L. Gruber, M. Holzleitner, T. Adler, D. Kreil, M.K. Kopp, et al., Hopfield networks is all you need, in: International Conference on Learning Representations, 2021.
[2] D. Krotov, J.J. Hopfield, Large associative memory problem in neurobiology and machine learning, in: International Conference on Learning Representations, 2021.
[3] D. Krotov, J.J. Hopfield, Dense associative memory for pattern recognition, in: Advances in Neural Information Processing Systems, 2016, pp. 1172–1180.
[4] H. Bao, R. Zhang, Y. Mao, The capacity of the dense associative memory networks, Neurocomputing 469 (2022) 198–208.
[5] K. Steinbuch, Die Lernmatrix-the beginning of associative memories, in: Advanced Neural Computers, Elsevier, 1990, pp. 21–29.
[6] J.J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proc. Natl. Acad. Sci. 79 (8) (1982) 2554–2558.
[7] P. Baldi, S.S. Venkatesh, Number of stable points for spin-glasses and neural networks of higher orders, Phys. Rev. Lett. 58 (9) (1987) 913.
[8] A. Bovier, V. Gayrard, Rigorous bounds on the storage capacity of the dilute Hopfield model, J. Stat. Phys. 69 (3–4) (1992) 597–627.
[9] M. Löwe, et al., On the storage capacity of Hopfield models with correlated patterns, Ann. Appl. Probab. 8 (4) (1998) 1216–1250.
[10] A. Bovier, B. Niederhauser, The spin-glass phase-transition in the Hopfield model with p-spin interactions, Adv. Theor. Math. Phys. 5 (2001) 1001–1046.
[11] E. Agliari, A. Annibale, A. Barra, A. Coolen, D. Tantari, Immune networks: multi-tasking capabilities at medium load, J. Phys. A, Math. Theor. 46 (33) (2013) 335101.
[12] E. Agliari, A. Annibale, A. Barra, A. Coolen, D. Tantari, Immune networks: multitasking capabilities near saturation, J. Phys. A, Math. Theor. 46 (41) (2013) 415003.
[13] E. Agliari, F. Alemanno, A. Barra, A. Fachechi, Generalized Guerra's interpolation schemes for dense associative neural networks, Neural Netw. 128 (2020) 254–267.
[14] A. Barra, A. Bernacchia, E. Santucci, P. Contucci, On the equivalence of Hopfield networks and Boltzmann machines, Neural Netw. 34 (2012) 1–9.
[15] A. Barra, G. Genovese, P. Sollich, D. Tantari, Phase diagram of restricted Boltzmann machines and generalized Hopfield networks with arbitrary priors, Phys. Rev. E 97 (2) (2018) 022310.
[16] A. Barra, M. Beccaria, A. Fachechi, A new mechanical approach to handle generalized Hopfield neural networks, Neural Netw. 106 (2018) 205–222.
[17] M. Smart, A. Zilman, On the mapping between Hopfield networks and restricted Boltzmann machines, in: International Conference on Learning Representations, 2021.
[18] R. McEliece, E. Posner, E. Rodemich, S. Venkatesh, The capacity of the Hopfield associative memory, IEEE Trans. Inf. Theory 33 (4) (1987) 461–482.
[19] C. Mazza, On the storage capacity of nonlinear neural networks, Neural Netw. 10 (4) (1997) 593–597.
[20] Y. Abu-Mostafa, J.S. Jacques, Information capacity of the Hopfield model, IEEE Trans. Inf. Theory 31 (4) (1985) 461–464.
[21] T.-D. Chiueh, R.M. Goodman, Recurrent correlation associative memories, IEEE Trans. Neural Netw. 2 (2) (1991) 275–284.
[22] G.X. Ritter, P. Sussner, J. Diza-de Leon, Morphological associative memories, IEEE Trans. Neural Netw. 9 (2) (1998) 281–293.
[23] B. Kosko, Fuzzy associative memories, in: NASA, Lyndon B. Johnson Space Center, Proceedings of the 2nd Joint Technology Workshop on Neural Networks and Fuzzy Logic, vol. 1, 1991.
[24] V. Folli, M. Leonetti, G. Ruocco, On the maximum storage capacity of the Hopfield model, Front. Comput. Neurosci. 10 (2017) 144.

[25] T. Salvatori, Y. Song, Y. Hong, L. Sha, S. Frieder, Z. Xu, R. Bogacz, T. Lukasiewicz, Associative memories via predictive coding, Adv. Neural Inf. Process. Syst. 34 (2021) 3874–3886.
[26] P. Gabbur, M. Bilkhu, J. Movellan, Probabilistic attention for interactive segmentation, Adv. Neural Inf. Process. Syst. 34 (2021) 4448–4460.
[27] D. Tyulmankov, C. Fang, A. Vadaparty, G.R. Yang, Biological learning in key-value memory networks, Adv. Neural Inf. Process. Syst. 34 (2021) 22247–22258.
[28] M. Widrich, B. Schäfl, M. Pavlović, H. Ramsauer, L. Gruber, M. Holzleitner, J. Brandstetter, G.K. Sandve, V. Greiff, S. Hochreiter, et al., Modern Hopfield networks and attention for immune repertoire classification, Adv. Neural Inf. Process. Syst. 33 (2020) 18832–18845.
[29] Y. Liang, C.K. Ryali, B. Hoover, L. Grinberg, S. Navlakha, D. Krotov, M.J. Zaki, Can a fruit fly learn word embeddings?, in: International Conference on Learning Representations, 2021.
[30] Y. Liang, D. Krotov, M.J. Zaki, Associative learning for network embedding, CoRR, arXiv:2208.14376, 2022.
[31] J. Yoo, F. Wood, Bayespcn: a continually learnable predictive coding associative memory, Adv. Neural Inf. Process. Syst. 35 (2022) 29903–29914.
[32] B. Schäfl, L. Gruber, A. Bitto-Nemling, S. Hochreiter, Hopular: Modern Hopfield networks for tabular data, in: NeurIPS 2023 Second Table Representation Learning Workshop, 2023.
[33] A. Carta, A. Sperduti, D. Bacciu, Encoding-based memory modules for recurrent neural networks, CoRR, arXiv:2001.11771, 2020.
[34] I. Danihelka, G. Wayne, B. Uria, N. Kalchbrenner, A. Graves, Associative long short-term memory, in: International Conference on Machine Learning, PMLR, 2016, pp. 1986–1994.
[35] J. Ba, G.E. Hinton, V. Mnih, J.Z. Leibo, C. Ionescu, Using fast weights to attend to the recent past, Adv. Neural Inf. Process. Syst. 29 (2016).
[36] W. Zhang, B. Zhou, Learning to update auto-associative memory in recurrent neural networks for improving sequence memorization, CoRR, arXiv:1709.06493, 2017.
[37] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186.
[38] I. Kanter, H. Sompolinsky, Associative recall of memory without errors, Phys. Rev. A 35 (1) (1987) 380.
[39] A. Storkey, Increasing the capacity of a Hopfield network without sacrificing functionality, in: Artificial Neural Networks—ICANN'97, 1997, pp. 451–456.
[40] M. Demircigil, J. Heusel, M. Löwe, S. Upgang, F. Vermet, On a model of associative memory with huge storage capacity, J. Stat. Phys. 168 (2) (2017) 288–299.
[41] R. Durrett, Probability: Theory and Examples, vol. 49, Cambridge University Press, 2019.
[42] A.A. Cuyt, V. Petersen, B. Verdonk, H. Waadeland, W.B. Jones, Handbook of Continued Fractions for Special Functions, Springer Science & Business Media, 2008.
[43] 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.
[44] B. Millidge, T. Salvatori, Y. Song, T. Lukasiewicz, R. Bogacz, Universal Hopfield networks: a general framework for single-shot associative memory models, in: International Conference on Machine Learning, PMLR, 2022, pp. 15561–15583.