Sparse Binary Matrices as Efficient Associative Memories

Vincent Gripon
Télécom Bretagne
vincent.gripon@telecom-bretagne.eu

Vitaly Skachek
University of Tartu
vitaly.skachek@gmail.com

Michael Rabbat
McGill University
michael.rabbat@mcgill.ca

***Abstract*—Associative memories are widely used devices which can be viewed as universal error-correcting decoders. Employing error-correcting code principles in these devices has allowed to greatly enhance their performance. In this paper we reintroduce a neural-based model using the formalism of linear algebra and extend its functionality, originally limited to erasure retrieval, to handle approximate inputs. In order to perform the retrieval, we use an iterative algorithm that provably converges. We then analyze the performance of the associative memory under the assumption of connection independence. We support our theoretical results with numerical simulations.**

I. Introduction

Associative memories can serve as an alternative to indexed memories, where the content is retrieved from a part of it. The associative memories behave similarly to error-correcting decoders. They are used in many domains, including network routers [1], intrusion detection systems [2], database engines [3], and CPU caches [4].

Recently, Gripon and Berrou [5], [6] introduced a novel architecture of associative memories based on the principles of modern error-correcting codes such as LDPC codes [7]. They proposed to store pieces of information in a multipartite binary graph by using cliques (a subset of fully interconnected nodes). The retrieving procedure instantiates a simple message-passing algorithm in two phases. First, the nodes sum their inputs. After that, an adaptive threshold is applied such that only those nodes which obtain the maximum score in their part remain activated. This is referred to as the "winner-takes-all" principle in the neuroscience literature.

Gripon and Berrou [5], [6] claim to provide near-optimal efficiency, along with limited computational complexity and low error probability. The recent extension [8] proposes novel decoding rules that both enhance performance and yield convergence guarantees. However, the model of [8] has only been studied for the case of erasure channels so far, where input messages have some missing symbols. In order to adapt these devices to other application domains, such as machine learning and noise cancellation, it is of interest to study the ability of these memories to recover a message from an approximate input.

While implementing these associative memories, it was shown in [9] that the original rule is a hard-thresholding instantiation, which uses a matrix product followed by a non-linear thresholding. In [10], the authors propose to use only binary operations to perform the retrieval process, using a smart rewriting of what was originally proposed in [8].

In this paper we introduce a linear-algebraic formalization of these associative memories and their functionality over the binary field $\mathbb{F}_2$. We extend their analysis to the case of blurred channels, where input symbols are transformed into a set of possible characters (including themselves). We derive error probabilities for this case which agree well with simulation results.

The paper is organized as follows. Section II presents related previous work. Section III introduces notation and operators. In Section IV we present the principles of these devices for storing and retrieving pieces of information. We also introduce the different problems we analyze in this paper. Section V introduces the main parameter to assess performance: density. Section VI presents the analytical developments to obtain messages retrieval error probabilities and finally Section VII concludes our work.

II. Previous work

The model that we study in this paper is close to the celebrated Willshaw model [11], [12], which appears in the neuroscience literature. In that model, pieces of information are stored as cliques in a neural network. The main difference between the Willshaw model and the one considered in this work is that the neurons are partitioned in the former model, thus allowing for more efficient decoding rules.

The celebrated Hopfield [13] model allows storing pieces of information and retrieving them using complete graphs with weighted connections. Contrary to the model considered in this paper, Hopfield's model provides zero efficiency [14] asymptotically, thus making it unsuitable for practical applications.

In [15], an associative memory that uses principles of error-correcting codes is proposed. However the devices of [15] require the use of the real-valued scalars to store the input messages, thus making it difficult to analyze their efficiency.

III. Notation and operators

Let Σ be the binary alphabet $\{0,1\}$, equipped with the Boolean operations OR and AND, denoted by $\vee$ and $\wedge$, respectively. For any $m, n \in \mathbb{N}$, denote by $\mathcal{M}_{m,n}(\Sigma)$ the set of $m \times n$ matrices over Σ, and by $\mathcal{M}_m(\Sigma)$ the set of $m \times m$ matrices over Σ.

Given two matrices $A \in \mathcal{M}_{m,n}(\Sigma)$ and $B \in \mathcal{M}_{n,p}(\Sigma)$, we define the matrix product $\otimes$ as:

$$(A \otimes B)_{i,j} = \bigvee_{k=1}^{n} A_{i,k} \wedge B_{k,j} \ ,$$

where $C_{i,j}$ denotes the entry at row i and column j in the matrix C.

We also define operations $\oplus$ and $\odot$ over $\mathcal{M}_{m,n}(\Sigma)$ as

$$(A \oplus B)_{i,j} = A_{i,j} \wedge B_{i,j} \ ,$$

and

$$(A \odot B)_{i,j} = A_{i,j} \vee B_{i,j} \ .$$

Note that $\odot$ is distributive over $\otimes$: for all A, B, and C of appropriate dimensions, $(A \odot B) \otimes C = (A \otimes C) \odot (B \otimes C)$.

We use the notation $\mathbb{K}_\ell$ to denote the triple $\langle \mathcal{M}_\ell(\Sigma), \oplus, \otimes \rangle$, where $\ell \in \mathbb{N}$. Note that $\mathbb{K}_\ell$ is not a field since $\otimes$ is not commutative, as shown in the following example.

Example 1. *Consider the following matrices in $\mathbb{K}_2$:*

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad \textit{and} \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \ .$$

It holds that

$$A \otimes B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \quad \textit{and} \quad B \otimes A = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \ .$$

Now consider the space $\mathcal{M}_c(\mathbb{K}_\ell)$ of $c\ell \times c\ell$ matrices; we adopt this specific notation explicitly to indicate that elements of $\mathcal{M}_c(\mathbb{K}_\ell)$ are treated as block matrices with each block being a $\ell \times \ell$ matrix. Let $W, W' \in \mathcal{M}_c(\mathbb{K}_\ell)$. We introduce the operation $W \cdot W'$ as: $\forall i, j = 1, \ldots, c$,

$$(W \cdot W')_{i,j} = \bigoplus_{k=1}^{c} W_{i,k} \otimes W'_{k,j} \ ,$$

where $W_{i,k}$ is the $\ell \times \ell$ block of W in position (i,k).

As a consequence of $\otimes$ being non-commutative, the operation $\cdot$ is non-associative.

Example 2. *To illustrate this property, consider, in addition to the matrices A and B defined in Example 1, two other matrices over $\mathbb{K}_2$:*

$$C = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \textit{and} \quad D = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

Then

$$\left(\begin{bmatrix} A & A \\ A & A \end{bmatrix} \cdot \begin{bmatrix} A & B \\ A & B \end{bmatrix} \right) \cdot \begin{bmatrix} A & A \\ A & A \end{bmatrix} = \begin{bmatrix} C & D \\ C & D \end{bmatrix}$$
$$\begin{bmatrix} A & A \\ A & A \end{bmatrix} \cdot \left(\begin{bmatrix} A & B \\ A & B \end{bmatrix} \cdot \begin{bmatrix} A & A \\ A & A \end{bmatrix} \right) = \begin{bmatrix} D & D \\ D & D \end{bmatrix} \ .$$

To avoid the above ambiguity, we adopt the convention that the operation $\otimes$ is ordered from right to left:

$$W \otimes W' \otimes W'' \triangleq W \otimes (W' \otimes W'') \ .$$

Note that this also removes ambiguity for the operation $W \cdot W$.

We also extend the transpose operator to $\mathcal{M}_c(\mathbb{K}_\ell)$ in the natural manner:

$$\left(W^T\right)_{i,j} = \left(W_{j,i}\right)^T \ .$$

IV. Proposed representation

Consider a set of messages, which is to be stored. These messages are assumed to be vectors in $\mathbb{Z}(\ell)^c$, where $\mathbb{Z}(\ell)$ denotes the set of integers $\{1, 2, \cdots, \ell\}$, and $c \geq 1$ is an integer. Denote by $\mathcal{P}(\mathbb{Z}(\ell))$ the set of all subsets of $\mathbb{Z}(\ell)$. We use the mapping π from the set of vectors with entries in $\mathbb{Z}(\ell)$ to the set of vectors with entries in $\mathcal{P}(\mathbb{Z}(\ell))$, defined as follows:

$$\begin{array}{rcl} \pi \ : \ \mathbb{Z}(\ell)^c & \rightarrow & (\mathcal{P}(\mathbb{Z}(\ell)))^c \\ \mathbf{m} & \mapsto & \pi(\mathbf{m}) \text{ where } \pi(\mathbf{m})_i = \{m_i\} \end{array} .$$

Let $\perp \notin \mathbb{Z}(\ell)$ be a special character that represents the ambiguity in the decoder, i.e., where more than one decoding result is possible. Note that any $\pi(\mathbf{m})$ can be associated with its preimage $\mathbf{m}$. We denote by ρ the corresponding operator that acts as an inverse of π when applied to a vector containing sets of cardinality 1 as its entries and extend it otherwise. More specifically,

$$\rho \ : \ (\mathcal{P}(\mathbb{Z}(\ell)))^c \longrightarrow (\mathbb{Z}(\ell) \cup \{\perp\})^c \ ,$$

where for any $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_c)^T \in (\mathcal{P}(\mathbb{Z}(\ell)))^c$,

$$\rho(\mu)_i = \begin{cases} j & \text{if } |\mu_i| = 1 \text{ and } \mu_i = \{j\} \\ \perp & \text{otherwise} \end{cases} .$$

Let $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_c)^T \in (\mathcal{P}(\mathbb{Z}(\ell)))^c$. The one-to-one mapping $\phi \ : \ (\mathcal{P}(\mathbb{Z}(\ell)))^c \longrightarrow \mathcal{M}_{c,1}(\mathcal{M}_{\ell,1}(\Sigma))$ is defined as follows:

$$\phi(\boldsymbol{\mu}) = \begin{bmatrix} \phi(\mu_1) \\ \phi(\mu_2) \\ \vdots \\ \phi(\mu_c) \end{bmatrix} ,$$

where for all $i = 1, 2, \cdots, c$, and for all $j = 1, 2, \cdots, \ell$,

$$(\phi(\mu_i))_j = \begin{cases} 1 & \text{if } j \in \mu_i \\ 0 & \text{otherwise} \end{cases} .$$

Example 3. *Let $c = 2$ and $\ell = 3$. Then, the following holds:*

$$\begin{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \\ \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \end{bmatrix} = \phi\left(\begin{bmatrix} \{3\} \\ \{2\} \end{bmatrix} \right) = \phi\left(\pi\left(\begin{bmatrix} 3 \\ 2 \end{bmatrix} \right) \right) \ .$$

The crux of our approach is as follows. Rather than storing a message $\mathbf{m} \in \mathbb{Z}(\ell)^c$ in the memory directly, we propose to store $\phi(\pi(\mathbf{m}))$ instead. Observe that $\mathbf{m} = \rho(\phi^{-1}(\phi(\pi(\mathbf{m}))))$.

To store a set of M messages $\mathfrak{M}$ in $\mathbb{Z}(\ell)^c$, we use the following representation:

$$W(\mathfrak{M}) = \bigodot_{\mathbf{m}\in\mathfrak{M}} \phi(\pi(\mathbf{m})) \cdot \phi(\pi(\mathbf{m}))^T ,$$

where $\odot$ is extended to $\mathcal{M}_\ell(\mathbb{F}_2)$.

Definition 4. *A* blurred version *of* $\boldsymbol{m} \in \mathfrak{M}$ *is a vector* $\chi(\boldsymbol{m}) \in (\mathcal{P}(\mathbb{Z}(\ell)))^c$*, where* $\chi : \mathbb{Z}(\ell)^c \to (\mathcal{P}(\mathbb{Z}(\ell)))^c$ *is a random function such that for all* $i = 1, 2, \cdots, c \ : \ m_i \in (\chi(\boldsymbol{m}))_i$*.*

The intuition behind this definition is that typically a blurred version of an input could be obtained by using a convolution of its symbols with a low-pass filter. As a result, the input symbols are replaced by their support sets. In this way, the system aims to retrieve a previously stored message that matches the provided support.

In terms of associations, this allows the system to perform a query given an imprecise probe of some of its symbols. For example consider an associative memory that stores information about papers in the proceedings of a conference, and that can then be addressed given keywords, authors or dates. In this context a blurred input can be “retrieve some paper containing keywords k_1 or k_2 than has been written by author a_1 or a_2 in year y_1 or y_2”.

Sometimes $W(\mathfrak{M})$ contains sufficient information for retrieval of a message $\mathbf{m} \in \mathfrak{M}$ from the blurred version of $\mathbf{m}$. We usually take the set $\mathfrak{M}$ to be fixed in advance. To simplify the notation we write W instead of $W(\mathfrak{M})$.

We denote by ψ the operation of retrieving $\mathbf{m}$ given $\chi(\mathbf{m})$ when using $W(\mathfrak{M})$. To perform this operation, we first define the recurrent sequence $\{\chi^t\}_{t=0}^{\infty}$ as follows:

1) Initialization: $\chi^0 = \chi(\mathbf{m})$.
2) Step: for $t = 1, 2, \cdots$, let $\chi^{t+1} = W \cdot \chi^t$.

While having in mind the preceding observation on non-associativity of operation $(\cdot)$, we write $\chi^t = W^t \cdot \chi^0$.

Theorem 5. *The sequence* $\{\chi^t\}_{t=0}^{\infty}$ *converges.*

Proof: The proof of this result is straightforward. First, for all i, $1 \le i \le c$, $W(\mathfrak{M})_{i,i} = I_\ell$, where I_ℓ denotes $\ell \times \ell$ identity matrix over Σ.

It follows from the definition of the operation $\otimes$ that:

$$\forall i, j \ : \ \left[(\chi_i^t)_j = 0 \Rightarrow (\chi_i^{t+1})_j = 0\right] . \tag{1}$$

Note that the number of distinct possible vectors in $\mathcal{M}_{c,1}(\mathcal{M}_{\ell,1}(\Sigma))$ is finite, and therefore the sequence $(\chi^t)_{t=1}^{\infty}$ is periodic. Since, from (1), the entries of χ^t can only change from 1 to 0 (when t grows), for sufficiently large values of t all entries in χ^t do not change.

■

We thus define:

$$\chi^\infty = \lim_{t\to\infty} \chi^t ,$$

and apply $\rho \circ \phi^{-1}$ to χ^∞ to obtain $\psi(\chi(\mathbf{m}))$.

Therefore, the process to retrieve (i.e., decode) a message $\mathbf{m}$ given a blurred version $\chi(\mathbf{m})$ as input can be represented algebraically as

$$\psi(\chi(\mathbf{m})) = \rho(\phi^{-1}(W \cdot W \cdot \ldots \cdot W \cdot \chi(\mathbf{m}))) .$$

We now define three problems which use the proposed data structure. We analyze the error probability for each of the proposed problems. In all three problems, the network $W(\mathfrak{M})$ is known, and it represents $\mathfrak{M}$.

Problem 1. For a given encoded message $\chi(\mathbf{m}) \in \mathcal{M}_{c,1}(\mathcal{M}_{\ell,1}(\Sigma))$, we are interested in checking whether $\mathbf{m} \in \mathfrak{M}$ by using the test $\psi(\chi(\pi(\mathbf{m}))) = \mathbf{m}$. We denote this test $\mathcal{T}(\mathfrak{M}, \mathbf{m})$, and assume that it takes boolean values TRUE and FALSE. This use of $W(\mathfrak{M})$ essentially makes it a particular Bloom filter.

Problem 2. We restrict χ such that

$$\forall i \ : \ \chi(\mathbf{m})_i = \{m_i\} \quad \vee \quad \chi(\mathbf{m})_i = \mathbb{Z}(\ell) .$$

In particular, the case $\chi(\mathbf{m})_i = \mathbb{Z}(\ell)$ represents the situation, when the value of i-th coordinate in $\mathbf{m}$ is erased. We refer to such a symbol of $\chi(\mathbf{m})$ an *erased symbol*.

In this problem, for a given input $\chi(\mathbf{m}) \in \mathcal{M}_{c,1}(\mathcal{M}_{\ell,1}(\Sigma))$ with some symbols erased, we want to determine if $\chi(\mathbf{m})$ could have been obtained from an $\mathbf{m} \in \mathfrak{M}$ by erasing symbols, and if so, then which $\mathbf{m} \in \mathfrak{M}$ could correspond to $\chi(\mathbf{m})$.

The specific case with no erased symbol corresponds to Problem 1.

Problem 3. In this problem, we consider the b-regular blur on the messages, defined as follows. Consider the neighboorhood function $\nu(j) = \{j \mid |j - m_i| \mod \ell \le b\}$. Define $\forall i \in c \ : \ \chi(\mathbf{m})_i = \nu(j)$.

In this problem, for a given blurred input $\chi(\mathbf{m}) \in \mathcal{M}_{c,1}(\mathcal{M}_{\ell,1}(\Sigma))$, we are interested in first determining if there is any $\mathbf{m} \in \mathfrak{M}$ which could be transformed to $\chi(\mathbf{m})$ through a b-blurring operation, and if so, then determine the possible values of $\mathbf{m} \in \mathfrak{M}$.

Note that the case of 0-regular blurred messages corresponds to Problem 1.

V. DENSITY

First, we note that the stored messages (after applying π and ϕ) are fixed points of $W(\mathfrak{M})$. This assumption follows from the fact that the operation $\odot$ is distributive over $\otimes$ and from the definition of $W(\mathfrak{M})$. This property can also be regarded as a zero Type I error as far as Problem 1 is concerned, meaning that stored messages will always be recognized as such.

On the other hand, the number of possible matrices $W(\mathfrak{M})$ is smaller than $N = 2^{(c\ell)^2} = \#\left(\mathcal{M}_c(\mathcal{M}_\ell(\mathbb{K}))\right)$. This amount has to be compared to the number of possible sets of messages to store, $P = 2^{\ell^c}$. When $c \ge 3$ is fixed, $N = o(P)$, it is immediate that W is not a uniquely decipherable code and thus a non-zero Type II error should be expected: fixed points of $W(\mathfrak{M})$ are not necessarily in $\phi(\pi(\mathfrak{M}))$.

In order to assess the performance of the scheme under consideration, we restrict our study to a simple case of independent uniformly distributed messages. We denote by P the associated probability measure and we introduce a parameter $d((i,j),(i',j'))$, $i \neq i'$, that represents the probability that $W(\mathfrak{M})_{(i,j)(i',j')} = 1$ after storing a set $\mathfrak{M}$ of uniformly distributed messages.

Under the assumption that the messages stored are independent and identically distributed,

$$\begin{aligned} d((i,j),(i',j')) &= P(W(\mathfrak{M})_{(i,j)(i',j')} = 1) \\ &= P\left(\bigvee_{\mathbf{m}\in\mathfrak{M}} (\phi(\pi(\mathbf{m}))_i)_j = 1 \wedge (\phi(\pi(\mathbf{m}))_{i'})_{j'} = 1\right) \\ &= P\left(\bigvee_{\mathbf{m}\in\mathfrak{M}} \mathbf{m}_i = j \wedge \mathbf{m}_{i'} = j'\right) \\ &= 1 - P\left(\bigwedge_{\mathbf{m}\in\mathfrak{M}} \mathbf{m}_i \neq j \vee \mathbf{m}_{i'} \neq j'\right) . \end{aligned}$$

Assume that the messages in $\mathfrak{M}$ are independent and uniformly distributed. We obtain that

$$\begin{aligned} d((i,j),(i',j')) &= 1 - (P(\mathbf{m}_i \neq j \vee \mathbf{m}_{i'} \neq j'))^M \\ &= 1 - (1 - P(\mathbf{m}_i = j)P(\mathbf{m}_{i'} = j'))^M \\ &= 1 - \left(1 - \frac{1}{\ell^2}\right)^M . \end{aligned}$$

If we consider the regime where both M and ℓ grow, then in order to keep d constant, M should be proportional to $-1/\log(1 - 1/\ell^2) \approx \ell^2$.

Since $W(\mathfrak{M})$ is symmetric, let us focus on its upper triangular matrix. It is clear that d is not independently distributed, since each stored message adds ones to multiple coordinates. However, simulation results support the assumption that each connection exists independently of other connections with probability d. This assumption leads to simulation performance, which is very close to the obtained performance evaluations. For that reason, and in order to simplify the analysis, in the next sections we assume that d is independent from cell to cell.

VI. Performance Analysis

A. Problem 1

As it was previously pointed out, there is no Type I error. Fix some message $\mathbf{m}$. It holds that:

$$\begin{aligned} P(\mathcal{T}(\mathfrak{M},\mathbf{m})) &= P(\mathcal{T}(\mathfrak{M},\mathbf{m}) \big| \mathbf{m} \notin \mathfrak{M}) P(\mathbf{m} \notin \mathfrak{M}) \; + \\ &\quad \underbrace{P(\mathcal{T}(\mathfrak{M},\mathbf{m}) \big| \mathbf{m} \in \mathfrak{M})}_{=1} P(\mathbf{m} \in \mathfrak{M}) \\ &\geq P(\mathcal{T}(\mathfrak{M},\mathbf{m}) \big| \mathbf{m} \notin \mathfrak{M}) P(\mathbf{m} \notin \mathfrak{M}) \; + \\ &\quad P(\mathcal{T}(\mathfrak{M},\mathbf{m}) \big| \mathbf{m} \notin \mathfrak{M}) P(\mathbf{m} \in \mathfrak{M}) \\ &= P(\mathcal{T}(\mathfrak{M},\mathbf{m}) \big| \mathbf{m} \notin \mathfrak{M}) . \end{aligned}$$

We thus use $P(\mathcal{T}(\mathfrak{M},\mathbf{m}))$ as an upper bound for Type II error.

To estimate this probability, we use the following theorem.

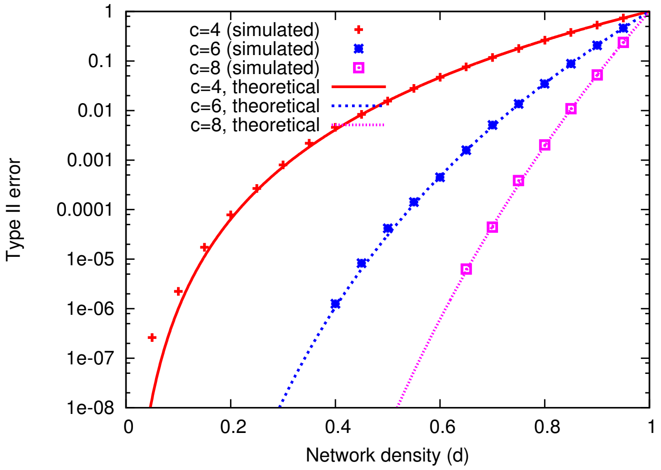

Figure 1. Evolution of the Type II error rate when stored messages are i.i.d. uniform, for various values of c and $\ell = 256$ and as a function of the number of stored messages. Both theoretical curves and simulation points obtained using Monte Carlo method are drawn.

Theorem 6. *The following condition holds:*

$$\begin{aligned} \forall \mathfrak{M}, \boldsymbol{m} \;:\quad & W(\mathfrak{M}) \cdot \chi(\pi(\boldsymbol{m})) = \chi(\pi(\boldsymbol{m})) \iff \\ & W(\mathfrak{M}) \odot (\chi(\pi(\boldsymbol{m})) \cdot \chi(\pi(\boldsymbol{m}))^T) = W(\mathfrak{M}) . \end{aligned}$$

Proof: One direction follows from the distributivity of the operation $\odot$ over the operation $\otimes$, and thus over the operation $(\cdot)$ as well.

For the other direction, assume by contradiction that $\exists i, j, W(\mathfrak{M})_{(i,m_i)(j,m_j)} = 0$. Then, it follows that $((W(\mathfrak{M}) \cdot \chi(\pi(\mathbf{m})))_i)_{m_i} = 0$.

■

We thus obtain that:

$$\begin{aligned} P(\mathcal{T}(\mathfrak{M},\mathbf{m}) \big| \mathbf{m} \notin \mathfrak{M}) &\leq P(\forall i < j, W(\mathfrak{M})_{(i,m_i)(j,m_j)} = 1 \\ &\qquad \wedge \forall i, W(\mathfrak{M})_{(i,m_i)(i,m_i)} = 1) \\ &\leq P(\forall i < j, W(\mathfrak{M})_{(i,m_i)(j,m_j)} = 1) \\ &= d^{\binom{c}{2}} , \end{aligned} \tag{2}$$

by using the fact that $P(\forall i, W(\mathfrak{M})_{(i,m_i)(i,m_i)} = 1) \leq 1$.

Figure 1 depicts the evolution of Type II error using simulations and Equation (2) with $\ell = 256$. Curves for parameters $c = 4$, $c = 6$ and $c = 8$ are drawn. It shows that the bound is tight.

To obtain that if $W(\mathfrak{M})$ is an asymptotically lossless representation of $\mathfrak{M}$, then it is not sufficient to force $P(\mathcal{T}(\mathfrak{M},\mathbf{m}) \big| \mathbf{m} \notin \mathfrak{M}) \to 0$. Therefore, we require a stronger condition, namely that $P(\exists \mathbf{m} \notin \mathfrak{M}, \mathcal{T}(\mathfrak{M},\mathbf{m})) \to 0$. We then use $P(\exists \mathbf{m} \notin \mathfrak{M}, \mathcal{T}(\mathfrak{M},\mathbf{m})) = 1 - P(\forall \mathbf{m} \notin \mathfrak{M}, \neg\mathcal{T}(\mathfrak{M},\mathbf{m}))$. If we assume that the corresponding events are independent, we obtain the following sufficient condition: $1 - (1 - P(\mathcal{T}(\mathfrak{M},\mathbf{m}) \big| \mathbf{m} \notin \mathfrak{M}))^{\ell^c - M} \to 0$.

Fix a constant value d and $c \to \infty$, we obtain that $1 - (1 - P(\mathcal{T}(\mathfrak{M},\mathbf{m}) \big| \mathbf{m} \notin \mathfrak{M}))^{2^{c\ell} - M} \leq 1 - (1 - d^{\binom{c}{2}})^{\ell^c} \approx 1 - e^{-\ell^c d^{\binom{c}{2}}}$. We thus require that $\ell^c d^{\binom{c}{2}} \to 0$, which is equivalent to $d < 1/4$. Let us fix such a value $d < 1/4$, and take $\ell = 2^c$. We thus have $M \approx \log(1/(1-d))\ell^2$.

On one hand, it is possible to encode $W(\mathfrak{M})$ using $(c\ell)^2$ bits. On the other hand, there are $\binom{\ell^c}{M}$ distinct possible sets of messages of cardinality M. Assume that these sets are choosen uniformly at random. Denote by $\mathcal{Z}$ the random variable, which takes the value of the chosen set. There are $\binom{\ell^c}{M}$ different values that $\mathcal{Z}$ can take. The binary entropy of $\mathcal{Z}$ is $H_2(\mathcal{Z}) = \log_2 \binom{\ell^c}{M}$. We obtain

$$\begin{aligned} H_2(\mathcal{Z}) &= \log_2 \binom{\ell^c}{M} \\ &\approx \ell^c \log_2(\ell^c) - M\log_2(M) - (\ell^c - M)\log_2(\ell^c - M) \ . \end{aligned}$$

If M is small compared to ℓ^c, then we obtain

$$H_2(\mathcal{Z}) \approx Mc\log_2(\ell) \ ,$$

which is, in turn, is close to $(c\ell)^2$. Thus, under these conditions $W(\mathfrak{M})$ is an asymptotically lossless optimal representation of $\mathfrak{M}$.

B. Problem 2

The problem of retrieving a message in $\mathfrak{M}$ when exactly r of its symbols have been erased has already been studied in [8] using a similar retrieval algorithm. Using simple properties of the binomial random vairables, the corresponding message retrieval error rate when using χ^1 as an estimate is:

$$P_e = 1 - \left(1 - d^{c-r}\right)^{(l-1)r} \ .$$

The actual message retrieval error rate when using χ^∞ as the estimate is expected to be lower. However, we presently have no closed form expression for this error probability.

Figure 2 depicts the evolution of the message retrieval error rate when 4 symbols out of $c = 8$ are erased in messages with $\ell = 256$ and as a function of the number of stored messages. The theoretical curve is drawn when using χ^1 as the estimator and simulation points are given for χ^∞.

C. Problem 3

Using similar arguments, one can derive bounds on error probabilities in the case of b-blured messages when using χ^1 as estimator. Denote by m the initial message. As a matter of fact, we already know that “correct” 1’s will remain active after application of the operator W. Thus the message will be correctly retrieved if no spurious 1’s remain:

$$\begin{aligned} P_e &= P(\exists i, j, \left(\chi^0_i\right)_j = 1 \wedge \left(\chi^1_i\right)_j = 1 \wedge m_i \neq j) \\ &= P(\exists i, j, \left(\chi^0_i\right)_j = 1 \wedge m_i \neq j \wedge \forall i', \\ &\qquad \left(W(\mathfrak{M})_{(i,i')} \otimes \chi^0_{i'}\right)_j = 1) \ . \end{aligned}$$

To simplify the reasoning, let us assume that $\forall i, W(\mathfrak{M})_{(i,i)}$ is the identity matrix, which is not a very good approximation when $M = \omega(\ell)$. Recall that $\chi^0_{i'}$ contains exactly $2b+1$ ones, such that we obtain under the hypothesis of independence:

$$\begin{aligned} P_e &= 1 - P\Big(\exists i', \left(W(\mathfrak{M})_{(i,i')} \otimes \chi^0_{i'}\right)_j = 0 \\ &\qquad \Big| \left(\chi^0_i\right)_j = 1 \wedge m_i \neq j\Big)^{2bc} \\ &= 1 - (1 - P(\forall i', \left(W(\mathfrak{M})_{(i,i')} \otimes \chi^0_{i'}\right)_j = 1 \\ &\qquad \Big| \left(\chi^0_i\right)_j = 1 \wedge m_i \neq j))^{2bc} \\ &= 1 - (1 - P(\left(W(\mathfrak{M})_{(i,i')} \otimes \chi^0_{i'}\right)_j = 1 \\ &\qquad \Big| \left(\chi^0_i\right)_j = 1 \wedge m_i \neq j)^{c-1})^{2bc} \\ &= 1 - (1 - P(\exists j' \in \nu(m_i), \left(W(\mathfrak{M})_{(i,i')}\right)_{(j,j')} = 1 \\ &\qquad \Big| \left(\chi^0_i\right)_j = 1 \wedge m_i \neq j)^{c-1})^{2bc} \\ &= 1 - (1 - (1 - P(\forall j' \in \nu(m_i), \left(W(\mathfrak{M})_{(i,i')}\right)_{(j,j')} = 0 \\ &\qquad \Big| \left(\chi^0_i\right)_j = 1 \wedge m_i \neq j)^{c-1})^{2bc} \\ &= 1 - (1 - (1 - (1 - d)^{2b+1})^{c-1})^{2bc} \ . \end{aligned}$$

Figure 3 depicts the message retrieval error rate for the same network as in Figure 2 ($\ell = 256, c = 8$), for various values of b and as a function of the number of stored messages. Both the theoretical curve when using χ^1 as estimator and the simulation points when using χ^∞ as estimator are drawn.

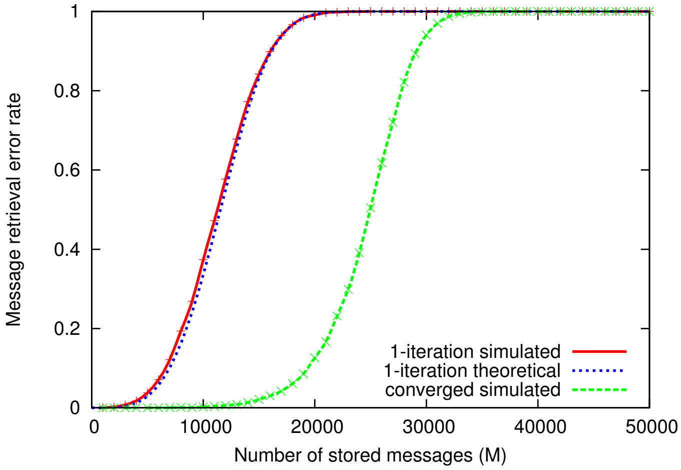

Figure 2. Evolution of the message retrieval error rate when stored messages are i.i.d. uniform and input messages contain 4 erased symbols as a function of the number of stored messages. Both theoretical curves and simulations are drawn when χ^1 is used as estimator and simulation are drawn when χ^∞ is used as estimator.

VII. Conclusion

We explored the properties of the associative memory, which was initially proposed in [5], [6], by using a fully linear algebraic formalism. We simplified the retrieving algorithm by making it fully binary. We then analyzed the error probabilities, when using χ^1 as an estimator for three problems: probabilistic data structure, erasure channel associative memory and blur channel associative memory.

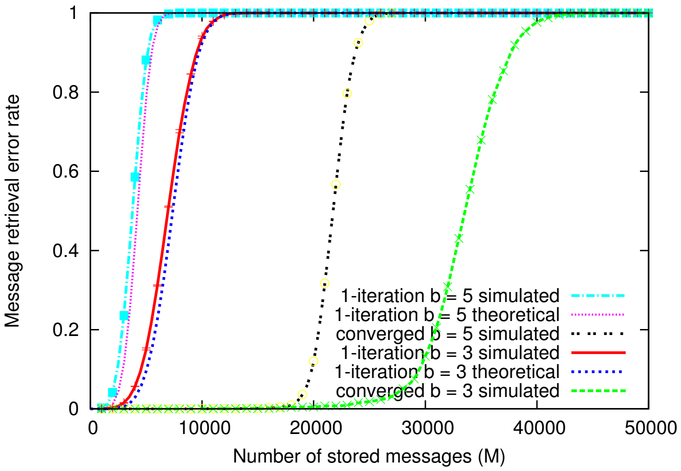

Figure 3. Evolution of the message retrieval error rate when stored messages are i.i.d. uniform and input messages are b-blurred as a function of the number of stored messages. Both theoretical curves and simulations are drawn when χ^1 is used as estimator and simulation points are drawn when χ^∞ is used as estimator.

The latter associative memory is of a particular interest as it combines interesting retrieval capacities along with asymptotically optimal memory consumption and simple iterative retrieving principles.

Future work includes deriving message retrieval error rate, when using χ^∞ as an estimator, as well as proposing extensions to solve classification problems and to handle larger cardinalities of the set of messages by using multi-dimensional matrices. Deriving bounds for retrieval error probability without using the independent connections hypothesis is another work in progress.

Acknowledgements

The work of V. Gripon is supported by the European Research Council (ERC-AdG2011 290901 NEUCOD). The work of V. Skachek is supported in part by the research grants PUT405 and IUT2-1 from the Estonian Research Council.

References

[1] H. J. Chao, "Next generation routers," in *Proceedings of the IEEE*, 2002, pp. 1518–1558.

[2] A. Papadogiannakis, M. Polychronakis, and E. P. Markatos, "Improving the accuracy of network intrusion detection systems under load using selective packet discarding," in *Proceedings of the Third European Workshop on System Security*, ser. EUROSEC '10, 2010, pp. 15–21.

[3] C. S. Lin, D. C. P. Smith, and J. M. Smith, "The design of a rotating associative memory for relational database applications," *ACM Transactions on Database Systems*, vol. 1, pp. 53–65, Mar. 1976.

[4] N. P. Jouppi, "Improving direct-mapped cache performance by the addition of a small fully-associative cache and prefetch buffers," in *Proceedings of the 17th annual International Symposium on Computer Architecture*, 1990, pp. 364–373.

[5] V. Gripon and C. Berrou, "Sparse neural networks with large learning diversity," *IEEE Transactions on Neural Networks*, vol. 22, no. 7, pp. 1087–1096, July 2011.

[6] ——, "A simple and efficient way to store many messages using neural cliques," in *Proceedings of IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain*, Paris, France, April 2011, pp. 54–58.

[7] R. G. Gallager, "Low-density parity-check codes," *IRE Transactions on Information Theory*, vol. 8, no. 1, pp. 21–28, 1962.

[8] V. Gripon and C. Berrou, "Nearly-optimal associative memories based on distributed constant weight codes," in *Proceedings of Information Theory and Applications Workshop*, San Diego, CA, USA, February 2012, pp. 269–273.

[9] Z. Yao, V. Gripon, and M. G. Rabbat, "A massively parallel associative memory based on sparse neural networks," in *Arxiv preprint 1303.7032*, March 2013.

[10] H. Jarollahi, N. Onizawa, V. Gripon, and W. J. Gross, "Reduced-complexity binary-weight-coded associative memories," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2013, to appear.

[11] D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins, "Non-holographic associative memory." *Nature*, 1969.

[12] F. Schwenker, F. T. Sommer, and G. Palm, "Iterative retrieval of sparsely coded associative memory patterns," *Neural Networks*, vol. 9, pp. 445–455, 1995.

[13] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational properties," *Proceedings of National Academia of Science, Biophysics*, vol. 79, pp. 2554–2558, 1982.

[14] R. J. McEliece, E. C. Posner, E. R. Rodemich, and S. S. Venkatesh, "The capacity of the Hopfield associative memory," *IEEE Transactions on Information Theory*, vol. 33, no. 4, pp. 461–482, 1987.

[15] A. H. Salavati and A. Karbasi, "Multi-level error-resilient neural networks," in *International Symposium on Information Theory Proceedings*. IEEE, 2012, pp. 1064–1068.