Online Memorization of Random Firing Sequences by a Recurrent Neural Network

Patrick Murer and Hans-Andrea Loeliger
Dept. of Information Technology and Electrical Engineering (D-ITET)
ETH Zürich
Email: {murer, loeliger}@isi.ee.ethz.ch

***Abstract*—This paper studies the capability of a recurrent neural network model to memorize random dynamical firing patterns by a simple local learning rule. Two modes of learning/memorization are considered: The first mode is strictly online, with a single pass through the data, while the second mode uses multiple passes through the data. In both modes, the learning is strictly local (quasi-Hebbian): At any given time step, only the weights between the neurons firing (or supposed to be firing) at the previous time step and those firing (or supposed to be firing) at the present time step are modified.**

The main result of the paper is an upper bound on the probability that the single-pass memorization is not perfect. It follows that the memorization capacity in this mode asymptotically scales like that of the classical Hopfield model (which, in contrast, memorizes static patterns). However, multiple-rounds memorization is shown to achieve a higher capacity (with a nonvanishing number of bits per connection/synapse). These mathematical findings may be helpful for understanding the functions of short-term memory and long-term memory in neuroscience.

I. Introduction

In this paper, we study the capability of a simple recurrent neural network to memorize and to reproduce a random dynamical firing pattern.

The background of this paper are neural networks with spiking neurons [1] – [5]. Such networks may be studied either as models of biological neural networks, or as candidates for neuromorphic hardware, or as a mode of mathematical signal processing as in [6]. In any case, memorizing long sequences of firing patterns must be an elementary capability of such networks. (Think of whistling a tune after hearing it once, or a few times.)

The classic reference for memorization is the Hopfield network [7], [8, Chapter 42]. Recurrent networks with higher capacities have been proposed in [9] – [11]. However, all these networks memorize static vectors (as static attractors of a dynamical network). By contrast, in this paper, we study the memorization of dynamical firing sequences, which seems to have been somewhat neglected in the literature.

The present paper is not immediately related to the vast literature on (nonspiking) recurrent neural networks such as LSTM networks [12] and others [13] – [16].

We will consider two different modes of learning. The first mode is strictly online, with a single pass through the data; the second mode uses multiple passes through the data. In both modes, the learning is strictly local, or quasi Hebbian: At any given time n, only the weights between the neurons firing (or supposed to be firing) at time $n-1$ and those firing (or supposed to be firing) at time n are modified. The first mode may thus be viewed as a model for instantaneous learning in short-term memory.

The main result of this paper is an upper bound on the probability that the single-pass memorization is not perfect. From this bound, it follows that the asymptotic memorization capacity in the strict online mode is at least $O\big(L/\ln(L)\big)$ bits per neuron, which vanishes in terms of bits per connection (i.e., per synapse). By contrast, multiple-rounds memorization is easily seen to achieve a significantly higher capacity, with a nonvanishing number of bits per connection/synapse and exceeding the capacity of the Hopfield network. The (important) ability of single-pass online memorization thus appears to be bought at the expense of a smaller capacity, which may be of interest for understanding the functions of short-term memory and long-term memory in neuroscience [17] – [20].

The paper is structured as follows. The network model is defined in Section II. Section III introduces the considered learning rules. The main result—an upper bound on the probability of imperfect single-pass memorization—is stated in Section IV. The bulk of the paper is Section V, which proves the bound of Section IV. Section VI investigates multi-pass memorization via a least-squares approach. The asymptotic memorization capacity of both learning modes is addressed in Section VII, and Section VIII concludes the paper.

II. The Network Model

We consider a discrete-time network model with L neurons $\xi_1, \ldots, \xi_L$ as follows. Each neuron is a map $\xi_\ell \colon \mathbb{R}^L \to \{0,1\}$ defined as

$$\boldsymbol{y} \mapsto \xi_\ell(\boldsymbol{y}) := \begin{cases} 1, & \text{if } \langle \boldsymbol{y}, \boldsymbol{w}_\ell \rangle + \eta_\ell \geq \theta_\ell \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

which is characterized by a weight vector $\boldsymbol{w}_\ell \in \mathbb{R}^L$ and a threshold $\theta_\ell \in \mathbb{R}$ and where $\langle \boldsymbol{y}, \boldsymbol{w}_\ell \rangle := \boldsymbol{w}_\ell^\mathsf{T} \boldsymbol{y}$ is the standard inner product. The quantity η_ℓ is an arbitrary bounded disturbance (or error) with

$$-\eta \leq \eta_\ell \leq \eta, \tag{2}$$

which subsumes imprecise computations and freak firings. In our main result, η will be allowed to grow linearly with L, cf. (18) and (19) below.

These neurons are connected to form an autonomous recurrent network producing the signal (firing sequence) $\boldsymbol{y}[1], \boldsymbol{y}[2], \ldots \in \{0,1\}^L$ with

$$\boldsymbol{y}[k+1] \coloneqq \big(\xi_1(\boldsymbol{y}[k]), \ldots, \xi_L(\boldsymbol{y}[k])\big)^{\mathsf{T}} \tag{3}$$

beginning from some initial value $\boldsymbol{y}[0] \in \mathbb{R}^L$.

In this paper, we want the network to reproduce a signal (i.e., a firing sequence) of length $N \geq 2$ that is given in the form of a matrix $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N) \in \{0,1\}^{L\times N}$ with columns $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N \in \{0,1\}^L$, i.e., we want (3), when initialized with

$$\boldsymbol{y}[0] = \boldsymbol{a}_0 \coloneqq \boldsymbol{a}_N \tag{4}$$

to yield

$$\boldsymbol{y}[k] = \boldsymbol{a}_{(k \bmod N)} \tag{5}$$

for $k = 1, 2, \ldots$, repeating the columns of $\boldsymbol{A}$ forever.

Such a network can be used as an associative memory as follows: When initialized with an arbitrary column of $\boldsymbol{A}$

$$\boldsymbol{y}[0] = \boldsymbol{a}_n, \tag{6}$$

the network will produce the sequence

$$\boldsymbol{y}[k] = \boldsymbol{a}_{((k+n) \bmod N)}, \quad k = 1, 2, \ldots \tag{7}$$

III. Learning Rules

Given the matrix $\boldsymbol{A} = (a_{\ell,n})$ (where $a_{\ell,n}$ is the entry in row ℓ and column n), we consider learning rules of the following form. Starting from some initial value $\boldsymbol{w}_\ell^{(0)} \in \mathbb{R}^L$ the weights are updated recursively by

$$\boldsymbol{w}_\ell^{(n)} = \boldsymbol{w}_\ell^{(n-1)} + \Delta\boldsymbol{w}_{\ell,n}, \quad n = 1, \ldots, K, \tag{8}$$

where the weight increment $\Delta\boldsymbol{w}_{\ell,n}$ of neuron ξ_ℓ at time n depends only on $a_{\ell,n}$ (the desired behavior of this neuron at this time) and on the preceding firing vector $\boldsymbol{a}_{n-1}$, and perhaps also on the previous weights $\boldsymbol{w}_\ell^{(n-1)}$ of this neuron.

This mode of learning may be called quasi-Hebbian since the stated restrictions on $\Delta\boldsymbol{w}_{\ell,n}$ essentially agree with those of Hebbian learning [21], except that the term “Hebbian” is normally reserved for unsupervised learning. The point of these restrictions is their suitability for hardware implementation, both biological and neuromorphic.

We will consider two versions of (8). In the first version (cf. Section IV), we pass through the data exactly once, i.e., $K = N$, and

$$\Delta\boldsymbol{w}_{\ell,n} \coloneqq a_{\ell,n}\big(\boldsymbol{a}_{n-1} - p\mathbf{1}_L\big), \tag{9}$$

where

$$\mathbf{1}_L \coloneqq \big(1, 1, \ldots, 1\big)^{\mathsf{T}} \in \mathbb{R}^L \tag{10}$$

and $0 < p < 1$ is defined in Section IV. In the second version (in Section VI), we allow multiple passes through the data, i.e., $K \gg N$, and

$$\Delta\boldsymbol{w}_{\ell,n} \coloneqq \beta^{(n)}\Big(a_{\ell,n} - \big\langle \boldsymbol{a}_{n-1}, \boldsymbol{w}_\ell^{(n-1)}\big\rangle\Big)\boldsymbol{a}_{n-1}, \tag{11}$$

for some step size $\beta^{(n)} > 0$.

IV. Single-pass Memorization – Main Result

For a network as in Section II, we now analyze the probability of perfect memorization for a random matrix $\boldsymbol{A} \in \{0,1\}^{L\times N}$ with i.i.d. entries $a_{\ell,n}$ parameterized by

$$p \coloneqq \Pr[a_{\ell,n} = 1], \tag{12}$$

which we denote by $\boldsymbol{A} \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p)^{L\times N}$.

The weight vectors are defined as

$$\boldsymbol{w}_\ell \coloneqq \boldsymbol{w}_\ell^{(N)} \tag{13}$$

where $\boldsymbol{w}_\ell^{(N)}$ is defined recursively as

$$\boldsymbol{w}_\ell^{(n)} \coloneqq \begin{cases} \boldsymbol{w}_\ell^{(n-1)}, & \text{if } a_{\ell,n} = 0 \\ \boldsymbol{w}_\ell^{(n-1)} + \boldsymbol{a}_{n-1} - p\mathbf{1}_L, & \text{if } a_{\ell,n} = 1, \end{cases} \tag{14}$$

and $\boldsymbol{w}_\ell^{(0)} \coloneqq \mathbf{0}$, for $n = 1, \ldots, N$, as in (9), resulting in

$$\boldsymbol{w}_\ell = \sum_{j\in J_\ell} \big(\boldsymbol{a}_{j-1} - p\mathbf{1}_L\big) = \sum_{j\in J_\ell} \boldsymbol{a}_{j-1} - |J_\ell| p\mathbf{1}_L, \tag{15}$$

where J_ℓ is the set

$$J_\ell \coloneqq \{n \in \{1, \ldots, N\} : a_{\ell,n} = 1\} \tag{16}$$

of desired firing positions of neuron ξ_ℓ and $|J_\ell|$ denotes its cardinality. It is easily verified that

$$\mathrm{E}[\boldsymbol{w}_\ell] = \mathbf{0}. \tag{17}$$

Let $\mathcal{E}_{\boldsymbol{A}}$ be the event that the memorization of $\boldsymbol{A}$ is not perfect. Our main result is the following theorem.

Theorem 1 (Upper Bound on $\Pr[\mathcal{E}_{\boldsymbol{A}}]$). *For all integers $L \geq 1$ and $N \geq 2$, $0 < p < 1$, $\boldsymbol{A} \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p)^{L\times N}$, the recurrent network with weight vectors (15), thresholds*

$$\theta_\ell \coloneqq \theta \coloneqq \frac{1}{4}Lp(1-p), \quad \ell = 1, \ldots, L, \tag{18}$$

disturbance bound

$$\eta \coloneqq \tilde{\eta}\cdot\theta, \quad 0 < \tilde{\eta} < 1, \tag{19}$$

and initialized with any column of $\boldsymbol{A}$ will reproduce a periodic extension of $\boldsymbol{A}$ with

$$\begin{aligned}&\Pr[\mathcal{E}_{\boldsymbol{A}}] \\ &\quad < 2LNe^{-\frac{1}{8}(1-\tilde{\eta})^2p^2(1-p)^2\frac{L}{N}} + LNe^{-D_{KL}\left(\frac{1+\tilde{\eta}}{2}p\|p\right)L},\end{aligned} \tag{20}$$

where $D_{KL}(p_1\|p_2)$ denotes the Kullback–Leibler divergence (as defined in (49) below) between two Bernoulli distributions with success probabilities $0 < p_1, p_2 < 1$. □

In consequence, a sufficient condition for the bound in (20) to vanish for $L \to \infty$ is

$$N \leq \frac{1}{8}(1-\tilde{\eta})^2p^2(1-p)^2\frac{L}{\ln(L^2)}. \tag{21}$$

Some numerical examples are given in Figure 1, which plots L vs. N for the right-hand side of (20) to achieve some desired level.

Clearly, for all $\varepsilon > 0$, there exists $L_\varepsilon \in \mathbb{N}$ such that $L^2/\ln(L) \geq L^{2-\varepsilon}$ for all $L \geq L_\varepsilon$. It follows that

$$LN \geq L^{2-\varepsilon} \tag{22}$$

for $N = L/\ln(L)$ and $L \to \infty$, i.e., asymptotically the network is able to memorize almost square matrices with instantaneous learning as in (13) – (15).

V. Proof of Theorem 1

We now prove Theorem 1, by using the union bound and by upper bounding the error probability for a single entry $a_{\ell,n}$ which amounts to bound the tails of $\langle \boldsymbol{a}_{n-1}, \boldsymbol{w}_\ell \rangle$.

The memorization is perfect if and only if $\xi_\ell(\boldsymbol{a}_{n-1}) = a_{\ell,n}$ for all $\ell \in \{1, \ldots, L\}$ and for all $n \in \{1, \ldots, N\}$. By the union bound, we have

$$\Pr[\mathcal{E}_{\boldsymbol{A}}] \leq \sum_{\ell=1}^{L} \sum_{n=1}^{N} \Pr[\xi_\ell(\boldsymbol{a}_{n-1}) \neq a_{\ell,n}] . \tag{23}$$

Moreover, using the same threshold θ for each neuron and by the law of total probability, we have

$$\begin{aligned} &\Pr[\xi_\ell(\boldsymbol{a}_{n-1}) \neq a_{\ell,n}] \\ &\quad = (1-p) \Pr[\langle \boldsymbol{a}_{n-1}, \boldsymbol{w}_\ell \rangle + \eta_\ell \geq \theta \,|\, a_{\ell,n} = 0] \\ &\qquad + p \Pr[\langle \boldsymbol{a}_{n-1}, \boldsymbol{w}_\ell \rangle + \eta_\ell < \theta \,|\, a_{\ell,n} = 1] . \end{aligned} \tag{24}$$

Now, let $\ell \in \{1, \ldots, L\}$ and let $n \in \{1, \ldots, N\}$ be fixed but arbitrary. Then

$$\begin{aligned} \langle \boldsymbol{a}_{n-1}, \boldsymbol{w}_\ell \rangle &= \left\langle \boldsymbol{a}_{n-1}, \sum_{j \in J_\ell} \left(\boldsymbol{a}_{j-1} - p\mathbf{1}_L\right) \right\rangle & (25) \\ &= \sum_{j \in J_\ell} \langle \boldsymbol{a}_{n-1}, \underbrace{\boldsymbol{a}_{j-1} - \mathrm{E}[\boldsymbol{a}_{j-1}]}_{=:\, \tilde{\boldsymbol{a}}_{j-1}} \rangle & (26) \\ &= \sum_{j=1}^{N} a_{\ell,j} \, \langle \boldsymbol{a}_{n-1}, \tilde{\boldsymbol{a}}_{j-1} \rangle & (27) \\ &= a_{\ell,n} \, \langle \boldsymbol{a}_{n-1}, \tilde{\boldsymbol{a}}_{n-1} \rangle + S_{\ell,n}, & (28) \end{aligned}$$

where

$$S_{\ell,n} \coloneqq \sum_{j \in \{1,\ldots,N\} \setminus \{n\}} a_{\ell,j} \, \langle \boldsymbol{a}_{n-1}, \tilde{\boldsymbol{a}}_{j-1} \rangle . \tag{29}$$

Lemma 1. *The random variable $S_{\ell,n}$ as defined in (29) has expectation zero, i.e.,*

$$\mathrm{E}[S_{\ell,n}] = 0, \tag{30}$$

and its moment generating function is upper bounded by

$$\mathrm{E}\big[e^{tS_{\ell,n}}\big] < e^{\frac{t^2}{8} LN} \tag{31}$$

for all $t \in \mathbb{R}$. □

A proof of Lemma 1 is given in [28, Appendix A].

Let us define the event

$$\mathcal{E}_{a_{\ell,n}} \coloneqq \big\{ \xi_\ell(\boldsymbol{a}_{n-1}) \neq a_{\ell,n} \big\} . \tag{32}$$

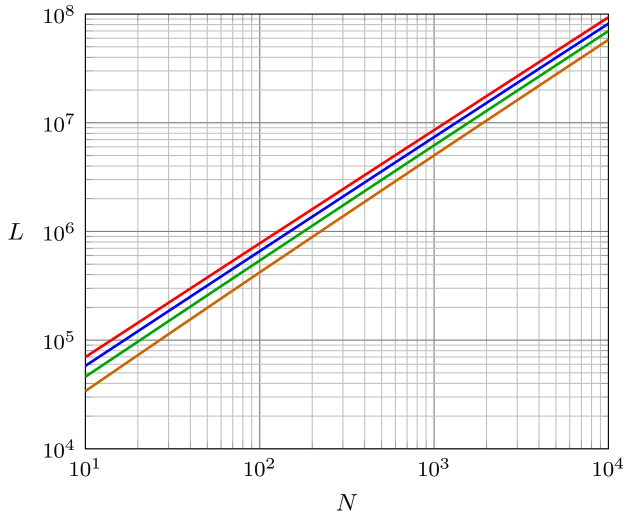

Fig. 1. Value of L required for the right-hand side of (20) to equal 10^{-3}, 10^{-6}, 10^{-9}, 10^{-12} (from bottom to top) for $p = 1/2$, and $\tilde{\eta} = 1/8$.

Then by (28) we can upper bound (24) as

$$\begin{aligned} &\Pr\big[\mathcal{E}_{a_{\ell,n}}\big] \\ &\quad \leq \Pr[S_{\ell,n} \geq \theta - \eta_\ell \,|\, a_{\ell,n} = 0] \\ &\qquad + \Pr[\langle \boldsymbol{a}_{n-1}, \tilde{\boldsymbol{a}}_{n-1} \rangle + S_{\ell,n} < \theta - \eta_\ell \,|\, a_{\ell,n} = 1] . \end{aligned} \tag{33}$$

As (global) threshold we choose

$$\theta \coloneqq \frac{1}{4} \sum_{a \in \{0,1\}} \mathrm{E}[\langle \boldsymbol{a}_{n-1}, \boldsymbol{w}_\ell \rangle \,|\, a_{\ell,n} = a] , \tag{34}$$

cf. Figure 2, and it can be shown that

$$\theta = \frac{1}{4} \mathrm{E}[\langle \boldsymbol{a}_{n-1}, \tilde{\boldsymbol{a}}_{n-1} \rangle] . \tag{35}$$

Note that $S_{\ell,n}$ depends on $a_{\ell,n}$. To get rid of the conditioning on $a_{\ell,n}$ in (33), we observe that an error, i.e., $\mathcal{E}_{a_{\ell,n}}$ implies either $|S_{\ell,n}| \geq \theta - \eta_\ell$, or $|S_{\ell,n}| < \theta - \eta_\ell$ and $\langle \boldsymbol{a}_{n-1}, \tilde{\boldsymbol{a}}_{n-1} \rangle + S_{\ell,n} < \theta - \eta_\ell$, cf. Figure 2. Thus by the union bound, we obtain

$$\begin{aligned} \Pr\big[\mathcal{E}_{a_{\ell,n}}\big] &\leq \Pr[|S_{\ell,n}| \geq \theta - \eta_\ell] \\ &\qquad + \Pr[\langle \boldsymbol{a}_{n-1}, \tilde{\boldsymbol{a}}_{n-1} \rangle < 2(\theta - \eta_\ell)] & (36) \\ &\leq \Pr[|S_{\ell,n}| \geq \theta - \eta] \\ &\qquad + \Pr[\langle \boldsymbol{a}_{n-1}, \tilde{\boldsymbol{a}}_{n-1} \rangle < 2(\theta + \eta)] & (37) \\ &= \Pr[|S_{\ell,n}| \geq \theta(1 - \tilde{\eta})] \\ &\qquad + \Pr[\langle \boldsymbol{a}_{n-1}, \tilde{\boldsymbol{a}}_{n-1} \rangle < 2\theta(1 + \tilde{\eta})] , & (38) \end{aligned}$$

where in (37) we applied (2), and (38) holds because of (19).

Now, we apply the Chernoff bound [22] to both terms on the right-hand side of (38). Thus, we have

$$\begin{aligned} \Pr[S_{\ell,n} \geq \theta(1 - \tilde{\eta})] &\leq \min_{t > 0} \frac{\mathrm{E}\big[e^{tS_{\ell,n}}\big]}{e^{t\theta(1-\tilde{\eta})}} & (39) \\ &< \min_{t > 0} \frac{e^{\frac{t^2}{8} LN}}{e^{t\theta(1-\tilde{\eta})}} & (40) \\ &= e^{-\frac{2\theta^2 (1-\tilde{\eta})^2}{LN}} . & (41) \end{aligned}$$

The step from (39) to (40) follows from (31). The bound (40) is minimized by $t_{\min} = 4\theta(1-\tilde{\eta})/(LN)$ which implies (41).

The lower tail of $S_{\ell,n}$, i.e., $\Pr[S_{\ell,n} \leq -(\theta-\eta)]$ can be upper bounded analogously. Thus by the union bound of both tails, we obtain

$$\Pr[|S_{\ell,n}| \geq \theta(1-\tilde{\eta})] < 2e^{-\frac{2\theta^2(1-\tilde{\eta})^2}{LN}}. \tag{42}$$

As for the other term on the right-hand side of (38), we note

$$\langle \boldsymbol{a}_{n-1}, \tilde{\boldsymbol{a}}_{n-1}\rangle = \sum_{\ell=1}^{L} a_{\ell,n-1}(a_{\ell,n-1}-p) \tag{43}$$

$$= (1-p)\sum_{\ell=1}^{L} a_{\ell,n-1} \tag{44}$$

since $a_{1,n-1},\dots,a_{L,n-1} \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p)$, thus

$$\frac{1}{1-p}\langle \boldsymbol{a}_{n-1}, \tilde{\boldsymbol{a}}_{n-1}\rangle \sim \mathrm{Bin}(L,p), \tag{45}$$

which together with (35) implies (cf. (18))

$$\theta = \frac{1}{4}Lp(1-p). \tag{46}$$

Then, inserting (46) into the right summand on the right-hand side of (38) yields

$$\Pr\left[\langle \boldsymbol{a}_{n-1}, \tilde{\boldsymbol{a}}_{n-1}\rangle < \frac{1+\tilde{\eta}}{2}Lp(1-p)\right]$$

$$= \Pr\left[\frac{1}{1-p}\langle \boldsymbol{a}_{n-1}, \tilde{\boldsymbol{a}}_{n-1}\rangle < \frac{1+\tilde{\eta}}{2}Lp\right] \tag{47}$$

$$\leq e^{-D_{\mathrm{KL}}\left(\frac{1+\tilde{\eta}}{2}p\|p\right)L}, \tag{48}$$

with Kullback–Leibler divergence (or relative entropy)

$$D_{\mathrm{KL}}(p_1\|p_2) := p_1\ln\left(\frac{p_1}{p_2}\right) + (1-p_1)\ln\left(\frac{1-p_1}{1-p_2}\right), \tag{49}$$

for $0 < p_1, p_2 < 1$, cf. [23]. From (47) to (48) we applied the bound stated in [28, Appendix B] with $1-\delta = (1+\tilde{\eta})/2$, $0 < \tilde{\eta} < 1$, because of (45). Note that in general

$$D_{\mathrm{KL}}(p_1\|p_2) \neq D_{\mathrm{KL}}(p_2\|p_1), \tag{50}$$

and for all $0 < p_1, p_2 < 1$

$$D_{\mathrm{KL}}(p_1\|p_2) \geq 0 \tag{51}$$

with equality if and only if $p_1 = p_2$.

Finally, we obtain

$$\Pr\left[\mathcal{E}_{a_{\ell,n}}\right] < 2e^{-\frac{2\theta^2(1-\tilde{\eta})^2}{LN}} + e^{-D_{\mathrm{KL}}\left(\frac{1+\tilde{\eta}}{2}p\|p\right)L} \tag{52}$$

$$= 2e^{-\frac{1}{8}(1-\tilde{\eta})^2p^2(1-p)^2\frac{L}{N}} + e^{-D_{\mathrm{KL}}\left(\frac{1+\tilde{\eta}}{2}p\|p\right)L}. \tag{53}$$

Inequality (52) follows from (38) together with the two upper bounds (42) and (48). In (53) we inserted (46).

The upper bound in (53) is independent on ℓ and n, and thus (23) yields (20) which concludes the proof. ■

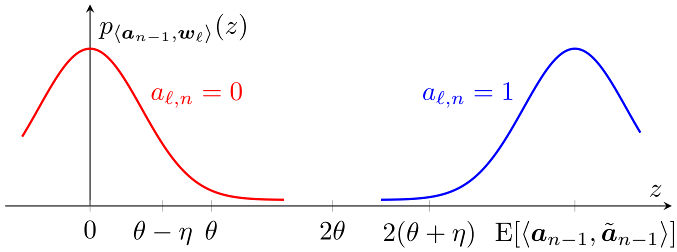

Fig. 2. Sketch of the probability distribution of (28) for the realization $\langle \boldsymbol{a}_{n-1},\tilde{\boldsymbol{a}}_{n-1}\rangle = \mathrm{E}[\langle \boldsymbol{a}_{n-1},\tilde{\boldsymbol{a}}_{n-1}\rangle]$ and the two cases $a_{\ell,n}=0$ (peak on the left) and $a_{\ell,n}=1$ (peak on the right).

VI. Multi-Pass Memorization

Perfect memorization can also be achieved via a certain least-squares problem, and solving this least-squares problem via stochastic gradient descent can be phrased as multi-pass learning according to (11).

Specifically, for fixed $\ell \in \{1,\dots,L\}$, consider the least-squares problem

$$\min_{\boldsymbol{w}_\ell}\sum_{n=1}^{N}\left|\langle \boldsymbol{a}_{n-1},\boldsymbol{w}_\ell\rangle - a_{\ell,n}\right|^2 = \min_{\boldsymbol{w}_\ell}\left\|\tilde{\boldsymbol{A}}\boldsymbol{w}_\ell - \tilde{\boldsymbol{a}}_\ell\right\|^2, \tag{54}$$

where

$$\tilde{\boldsymbol{A}} := \begin{pmatrix} \boldsymbol{a}_N^\mathsf{T} \\ \boldsymbol{a}_1^\mathsf{T} \\ \vdots \\ \boldsymbol{a}_{N-1}^\mathsf{T}\end{pmatrix} \in \mathbb{R}^{N\times L}, \quad \tilde{\boldsymbol{a}}_\ell := \begin{pmatrix} a_{\ell,1} \\ \vdots \\ a_{\ell,N}\end{pmatrix} \in \mathbb{R}^N. \tag{55}$$

Note that $\tilde{\boldsymbol{A}}$ is the transposed matrix of $(\boldsymbol{a}_N,\boldsymbol{a}_1,\dots,\boldsymbol{a}_{N-1}) \in \mathbb{R}^{L\times N}$, i.e., of the one time-step cyclic shifted version of $\boldsymbol{A}$, and $\tilde{\boldsymbol{a}}_\ell$ is the ℓ-th row of $\boldsymbol{A}$ turned into a column vector.

If $\mathrm{rank}(\tilde{\boldsymbol{A}}) = N$, then

$$\min_{\boldsymbol{w}_\ell\in\mathbb{R}^L}\left\|\tilde{\boldsymbol{A}}\boldsymbol{w}_\ell - \tilde{\boldsymbol{a}}_\ell\right\|^2 = 0, \tag{56}$$

which implies that $\boldsymbol{A}$ is (perfectly) memorizable, i.e., $\Pr[\mathcal{E}_{\boldsymbol{A}}] = 0$. For $L \geq N$, $0 < p \leq 1/2$, $\boldsymbol{A} \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p)^{L\times N}$, it follows from [24] that

$$\Pr\left[\mathrm{rank}(\tilde{\boldsymbol{A}}) = N\right] \geq 1 - \left(1-p+o_N(1)\right)^N, \tag{57}$$

where $o_N(1)$ denotes a sequence which converges to zero, i.e., $\lim_{N\to\infty} o_N(1) = 0$. Thus, any matrix $\boldsymbol{A} \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p)^{L\times N}$ with $L \geq N$, and in particular with

$$L = N \tag{58}$$

is memorizable as $N \to \infty$.

Clearly, the least-squares problem (54) could be solved by gradient descent as follows. Starting from some initial guess $\boldsymbol{w}_\ell^{(0)}$ we proceed by

$$\boldsymbol{w}_\ell^{(n)} = \boldsymbol{w}_\ell^{(n-1)} + \beta^{(n)}\tilde{\boldsymbol{A}}^\mathsf{T}\left(\tilde{\boldsymbol{a}}_\ell - \tilde{\boldsymbol{A}}\boldsymbol{w}_\ell^{(n-1)}\right) \tag{59}$$

for $n = 1, \dots, K$, $K \in \mathbb{N}$, and with step size $\beta^{(n)} > 0$. The recursion (59) with constant $\beta^{(n)} = \beta$ converges to a minimizer of (54) if

$$0 < \beta < \frac{2}{\lambda_{\max}(\tilde{\boldsymbol{A}}^{\mathsf{T}}\tilde{\boldsymbol{A}})}, \tag{60}$$

where $\lambda_{\max}(\tilde{\boldsymbol{A}}^{\mathsf{T}}\tilde{\boldsymbol{A}}) > 0$ is the largest eigenvalue of $\tilde{\boldsymbol{A}}^{\mathsf{T}}\tilde{\boldsymbol{A}}$.

Finally, replacing gradient descent as in (59) by stochastic gradient descent yields

$$\boldsymbol{w}_\ell^{(n)} = \boldsymbol{w}_\ell^{(n-1)} + \beta^{(n)}\Big(a_{\ell,n} - \big\langle \boldsymbol{a}_{n-1}, \boldsymbol{w}_\ell^{(n-1)}\big\rangle\Big)\boldsymbol{a}_{n-1}, \tag{61}$$

which is (11). Again, as in (5) the column indices are taken modulo N and $\boldsymbol{a}_0 := \boldsymbol{a}_N$. It is shown in [25] that if at every iteration the column indices are chosen randomly, then (61) converges exponentially in expectation to a solution of (56).

VII. Memorization Capacity

Let $\mathcal{A}_{\text{typical}}$ be a typical set of matrices (in any standard sense of "typical sequences" [23]) for the random matrix $\boldsymbol{A} \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)^{L\times N}$ and $|\mathcal{A}_{\text{typical}}|$ denotes the cardinality of $\mathcal{A}_{\text{typical}}$. Then, we have

$$\lim_{L\to\infty} \frac{1}{L}\log_2 |\mathcal{A}_{\text{typical}}| = H_{\text{b}}(p)N, \tag{62}$$

with the binary entropy function

$$H_{\text{b}}(p) := -p\log_2(p) - (1-p)\log_2(1-p) \tag{63}$$

for $0 < p < 1$, cf. [23].

The absolute capacity of a network is equal to the total number of bits which can be memorized by the network, thus

$$C_{\text{absolute}} \geq \log_2 |\mathcal{A}_{\text{typical}}| \ [\,\text{bits}\,]. \tag{64}$$

A. Capacity per Neuron

From (62) and (64) it follows that the asymptotic memorization capacity in bits per neuron is lower bounded by

$$C_{\text{neuron}} \geq H_{\text{b}}(p)N \ [\,\text{bits per neuron}\,]. \tag{65}$$

For the single-pass memorization rule (14) we have (21) (which is a consequence of Theorem 1), and we thus obtain

$$C_{\text{single-pass}} \geq C_{p,\tilde{\eta}}\frac{L}{\ln(L)} \ [\,\text{bits per neuron}\,] \tag{66}$$

with constant

$$C_{p,\tilde{\eta}} := \frac{1}{16}(1-\tilde{\eta})^2 p^2 (1-p)^2 H_{\text{b}}(p) > 0, \tag{67}$$

for $0 < p, \tilde{\eta} < 1$. For the multi-pass memorization rule (59), we have (cf. (58))

$$C_{\text{multi-pass}} \geq H_{\text{b}}(p)L \ [\,\text{bits per neuron}\,]. \tag{68}$$

Both memorization capacities $C_{\text{single-pass}}$ and $C_{\text{multi-pass}}$ (in bits per neuron) are unbounded in L.

B. Capacity per Connection (Synapse)

The capacity per connection (i.e., per nonzero weight) is

$$C_{\text{connection}} = \frac{C_{\text{neuron}}}{L}, \tag{69}$$

because in both modes the network is fully connected, and we obtain

$$C_{\text{single-pass}} \geq C_{p,\tilde{\eta}}\frac{1}{\ln(L)} \ [\,\text{bits per connection}\,], \tag{70}$$

$$C_{\text{multi-pass}} \geq H_{\text{b}}(p) \ [\,\text{bits per connection}\,]. \tag{71}$$

Thus, in bits per connection the capacity $C_{\text{single-pass}}$ seems to vanish, whereas $C_{\text{multi-pass}}$ does not vanish as $L \to \infty$.

C. Comparison with the Hopfield Network

The capacity of the Hopfield model with L neurons is $L/(2\ln(L))$ vectors with the Hebbian learning rule [26] and $L/\sqrt{2\ln(L)}$ vectors with the Storkey learning rule [27]. However, for a fair comparison with the results of the present paper, it should be noted that each vector consists of L random bits, resulting in a capacity of $L/(2\ln(L))$ bits per neuron and $L/\sqrt{2\ln(L)}$ bits per neuron, respectively. It follows that the capacity (66) of single-pass memorization is (at least) of the same order as the Hopfield model with Hebbian learning while the capacity (68) of multi-pass learning exceeds the capacity of the Hopfield model.

VIII. Conclusion

We have studied the capability of a "spiking" dynamical neural network model to memorize random firing sequences by a form of quasi-Hebbian learning. Our main result was an upper bound on the probability that instantaneous memorization is not perfect. From this bound, the instantaneous-memorization capacity of a network with L neurons is (at least) $O\big(L/\ln(L)\big)$ bits per neuron. By contrast, iterative (i.e., multi-pass) learning is shown to achieve a capacity of $O(L)$ bits per neuron and $O(1)$ bits per connection/synapse. These results may be useful for understanding the functions of short-term memory and long-term memory in neuroscience and their potential analogs in neuromorphic hardware.

Acknowledgement

Section VI developed from a suggestion by an anonymous reviewer of an earlier version of this paper.

REFERENCES

[1] R. Van Rullen, R. Guyonneau, and S.J. Thorpe, "Spike times make sense," *Trends in Neurosciences,* vol. 28, no. 1, Jan. 2005.

[2] J. Anumula et al., "Feature representations for neuromorphic audio spike streams," *Frontiers in Neuroscience,* Feb. 2018.

[3] J.C. Tapson et al., "Synthesis of neural networks for spatio-temporal spike pattern recognition and processing," *Frontiers in Neuroscience,* Aug. 2013.

[4] E.M. Izhikevich, "Polychronization: computation with spikes," *Neural Computation,* vol. 18, no. 2, pp. 245–282, Feb. 2006.

[5] Z. Marzi, J. Hespanha, and U. Madhow, "On the information in spike timing: neural codes derived from polychronous groups," *2018 Information Theory and Applications Workshop (ITA),* Feb. 2018.

[6] H.-A. Loeliger and S. Neff, "Pulse-domain signal parsing and neural computation," *2015 IEEE Int. Symp. on Information Theory (ISIT),* Hong Kong, China, June 2015.

[7] J.J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. Nat. Academy of Sciences of the USA,* vol. 79 no. 8 pp. 2554–2558, April 1982.

[8] D. MacKay, *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003.

[9] A. Karbasi, A.H. Salavati, and A. Shokrollahi, "Iterative learning and denoising in convolutional neural associative memories," *Proc. 30th Int. Conf. on Machine Learning,* Atlanta, Georgia, USA, June 2013.

[10] A.H. Salavati, *Coding Theory and Neural Associative Memories with Exponential Pattern Retrieval Capacity.* PhD thesis 6089 at EPFL, 2014.

[11] R. Chaudhuri and I. Fiete, "Bipartite expander Hopfield networks as self-decoding high-capacity error correcting codes," *Adv. Neural Inf. Proc. Systems (NeurIPS),* Dec. 2019.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation,* vol. 9 no. 8 pp. 1735–1780, Dec. 1997.

[13] E. Gardner, "The space of interactions in neural network models," *J. Phys. A: Math. Gen.,* vol. 21 no. 1 pp. 257–270, July 1988.

[14] T.D. Chiueh and R.M. Goodman, "Recurrent correlation associative memories," *IEEE Trans. on Neural Networks,* vol. 2 no. 2 pp. 275–284, March 1991.

[15] S. Jankowski, A. Lozowski, and J.M. Zurada, "Complex-valued multistate neural associative memory," *IEEE Trans. on Neural Networks,* vol. 7 no. 6 pp. 1491–1496, Nov. 1996.

[16] S. Miyoshi, H. Yanai, and M. Okada, "Associative memory by recurrent neural networks with delay elements," *Neural Networks,* vol. 17 pp. 55–63, 2004.

[17] G.A. Miller, "The magical number seven, plus or minus two: some limits on our capacity for processing information," *Psychological Review,* vol. 63 no. 2 pp. 81–97, 1956.

[18] R.C. Atkinson and R.M. Shiffrin, "Human memory: a proposed system and its control processes," *The psychology of learning and motivation,* vol. 2 pp. 89–195, 1968.

[19] A.D. Baddeley and G. Hitch, "Working memory," *The psychology of learning and motivation,* vol. 8 pp. 47–89, 1974.

[20] N. Cowan, "What are the differences between long-term, short-term, and working memory?," *Prog. Brain Res.,* vol. 169 pp. 323–338, 2008.

[21] D.O. Hebb, *The Organization of Behavior.* Wiley & Sons, 1949.

[22] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Statist.,* vol. 23 no. 4 pp. 493–507, 1952.

[23] T.M. Cover and J.A. Thomas, *Elements of Information Theory.* Wiley & Sons, 2006.

[24] K. Tikhomirov, "Singularity of random bernoulli matrices," arXiv:1812.09016v4 [math.PR], Aug. 2019

[25] T. Strohmer and R. Vershnin, "A randomized Kaczmarz algorithm with exponential convergence," *Journal of Fourier Analysis and Applications,* vol. 15 no. 2 pp. 262–278, 2009.

[26] R.J. McEliece et al., "The capacity of the Hopfield associative memory," *IEEE Trans. on Inform. Theory,* vol. IT-33 no. 4 pp. 461–482, July 1987.

[27] A.J. Storkey and R. Valabregue, "The basins of attraction of a new Hopfield learning rule," *Neural Networks,* vol. 12, no. 6, pp. 869–876, July 1999.

[28] P. Murer and H.-A. Loeliger, "Online memorization of random firing sequences by a recurrent neural network," arXiv:2001.02920v1 [cs.IT], Jan. 2020.