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 defined as

\[ y \mapsto \xi_t(y) := \begin{cases} 1, & \text{if } \langle y, w_t \rangle + \eta_t \geq \theta_t \\ 0, & \text{otherwise}, \end{cases} \]

which is characterized by a weight vector \( w_t \in \mathbb{R}^L \) and a threshold \( \theta_t \in \mathbb{R} \) and where \( \langle y, w_t \rangle := w^T_t \cdot y \) is the standard inner product. The quantity \( \eta_t \) is an arbitrary bounded disturbance (or error) with

\[ -\eta \leq \eta_t \leq \eta, \]

which subsumes imprecise computations and freak firings. In our main result, \( \eta \) 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) \( y[1], y[2], \ldots \in \{0,1\}^L \) with

\[
y[k+1] = (\xi_1(y[k]), \ldots, \xi_L(y[k]))^T
\]

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

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 \( A = (a_1, \ldots, a_N) \in \{0,1\}^{L \times N} \) with columns \( a_1, \ldots, a_N \in \{0,1\}^L \), i.e., we want (3), when initialized with

\[
y[0] = a_N
\]

to yield

\[
y[k] = a_{(k \mod N)}
\]

for \( k = 1, 2, \ldots \), repeating the columns of \( A \) forever.

Such a network can be used as an associative memory as follows: When initialized with an arbitrary column of \( A \)

\[
y[0] = a_n
\]

the network will produce the sequence

\[
y[k] = a_{[(k+n) \mod N]}, \quad k = 1, 2, \ldots
\]

### III. LEARNING RULES

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

\[
w^{(n)}_\ell = w^{(n-1)}_\ell + \Delta w^{(n-1)}_\ell, \quad n = 1, \ldots, K,
\]

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

This mode of learning may be called quasi-Hebbian since the stated restrictions on \( \Delta w^{(n-1)}_\ell \) 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

\[
\Delta w^{(n)}_\ell := a_{\ell,n} (a_{n-1} - p1_L),
\]

where

\[
p := \Pr[a_{\ell,n} = 1],
\]

\[
\theta := \theta := \frac{1}{4} L p (1 - p), \quad \ell = 1, \ldots, L,
\]

and initialized with any column of \( A \) will reproduce a periodic extension of \( A \) with

\[
Pr[\xi]\leq 2LN e^{\frac{1}{2} (1-\eta)^2 p^2(1-p)^2} + LN e^{-D_{KL}(\frac{1}{2} \eta p || p) L},
\]

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

\[
N \leq \frac{1}{8} (1-\eta)^2 p^2(1-p)^2 \frac{L}{\ln(L^2)},
\]

and initialized with any column of \( A \) will reproduce a periodic extension of \( A \) with

\[
Pr[\xi] \leq 2LN e^{-\frac{1}{2} (1-\eta)^2 p^2(1-p)^2} + LN e^{-D_{KL}(\frac{1}{2} \eta p || p) L},
\]

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

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
\begin{equation}
LN \geq L^{2-\varepsilon}
\end{equation}
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 $(a_{n-1}, w_\ell)$.

The memorization is perfect if and only if $\xi(\theta) = a_{\ell,n}$ for all $\ell \in \{1, \ldots, L\}$ and for all $n \in \{1, \ldots, N\}$. By the union bound, we have
\begin{equation}
\Pr[\xi(\theta) = a_{\ell,n}] \leq \sum_{\ell=1}^{L} \sum_{n=1}^{N} \Pr[\xi(\theta) = a_{\ell,n}] = 0.
\end{equation}
Moreover, using the same threshold $\theta$ for each neuron and by the law of total probability, we have
\begin{equation}
\Pr[\xi(\theta) = a_{\ell,n}] = (1 - p) \Pr[\xi(\theta) = a_{\ell,n}] + \eta_\ell \geq \theta | a_{\ell,n} = 0]
+ p \Pr[\xi(\theta) = a_{\ell,n}] + \eta_\ell < \theta | a_{\ell,n} = 1].
\end{equation}
Now, let $\ell \in \{1, \ldots, L\}$ and let $n \in \{1, \ldots, N\}$ be fixed but arbitrary. Then
\begin{equation}
\langle a_{\ell,n-1}, w_\ell \rangle = \left( a_{\ell,n-1}, \sum_{j \in J_\ell} (a_{j-1} - pL) \right)
= \sum_{j \in J_\ell} \langle a_{\ell,n-1}, a_{j-1} - E[a_{j-1}] \rangle
+ \sum_{j=1}^{N} a_{\ell,j} \langle a_{\ell,n-1}, a_{j-1} \rangle
= a_{\ell,n} \langle a_{\ell,n-1}, a_{\ell,j} \rangle + S_{\ell,n},
\end{equation}
where
\begin{equation}
S_{\ell,n} := \sum_{j=1}^{N} a_{\ell,j} \langle a_{\ell,n-1}, a_{j-1} \rangle.
\end{equation}

Lemma 1. The random variable $S_{\ell,n}$ as defined in (29) has expectation zero, i.e.,
\begin{equation}
E[S_{\ell,n}] = 0,
\end{equation}
and its moment generating function is upper bounded by
\begin{equation}
E[e^{tS_{\ell,n}}] < e^{\frac{t^2}{2}LN}
\end{equation}
for all $t \in \mathbb{R}$. \hfill \Box

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

Let us define the event
\begin{equation}
\mathcal{E}_{\ell,n} := \{ \xi(\theta) \neq a_{\ell,n} \}.
\end{equation}

\begin{figure}[h]
\centering
\includegraphics[width=\columnwidth]{fig1.png}
\caption{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 $\bar{\eta} = 1/8$.}
\end{figure}

Then by (28) we can upper bound (24) as
\begin{equation}
\Pr[\mathcal{E}_{\ell,n}] 
\leq \Pr[\xi(\theta) = 0] 
+ p \Pr[\xi(\theta) = 1] + S_{\ell,n} < \theta - \eta | a_{\ell,n} = 1].
\end{equation}
As (global) threshold we choose
\begin{equation}
\theta := \frac{1}{4} \sum_{a \in \{0,1\}} E[\langle a_{\ell,n-1}, w_\ell \rangle | a_{\ell,n} = a],
\end{equation}
cf. Figure 2, and it can be shown that
\begin{equation}
\theta \geq 1 \geq E[\langle a_{\ell,n-1}, a_{\ell,n} \rangle].
\end{equation}
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., $\mathcal{E}_{\ell,n}$ implies either $|S_{\ell,n}| \geq \theta - \eta$, or $|S_{\ell,n}| < \theta - \eta$ and $a_{\ell,n-1}, a_{\ell,n} + S_{\ell,n} < \theta - \eta$, cf. Figure 2. Thus by the union bound, we obtain
\begin{equation}
\Pr[\mathcal{E}_{\ell,n}] \leq \Pr[|S_{\ell,n}| \geq \theta - \eta]
+ \Pr[|a_{\ell,n-1} - a_{\ell,n}| < 2(\theta - \eta)]
\leq \Pr[|S_{\ell,n}| \geq \theta - \eta]
+ \Pr[|a_{\ell,n-1} - a_{\ell,n}| < 2(\theta + \eta)]
\leq \Pr[|S_{\ell,n}| \geq \theta - (1 - \bar{\eta})]
+ \Pr[|a_{\ell,n-1} - a_{\ell,n}| < 2\theta(1 + \bar{\eta})],
\end{equation}
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{equation}
\Pr[|S_{\ell,n}| \geq \theta - (1 - \bar{\eta})] \leq \min_{t > 0} \frac{E[e^{tS_{\ell,n}}]}{e^{\theta(1 - \bar{\eta})}}
\end{equation}
\begin{equation}
< \min_{t > 0} \frac{e^{\frac{t^2}{2}LN}}{e^{\theta(1 - \bar{\eta})}}
= e^{-\frac{2\theta^2}{L}N^2}.
\end{equation}
The step from (39) to (40) follows from (31). The bound (40) is minimized by 
\( t_{\text{min}} = 4\theta(1 - \tilde{\eta})/(LN) \) which implies (41).

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

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

(42)

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

\[
(a_{n-1}, \tilde{a}_{n-1}) = \sum_{\ell = 1}^{L} a_{\ell,n-1}(a_{\ell,n-1} - p)
\]

\[
= (1 - p) \sum_{\ell = 1}^{L} a_{\ell,n-1}
\]

(43)

(44)

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

\[
\frac{1}{1 - p} (a_{n-1}, \tilde{a}_{n-1}) \sim \text{Bin}(L, p),
\]

(45)

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

\[
\theta = \frac{1}{4} Lp(1 - p).
\]

(46)

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

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

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

\[
\leq e^{-D_{\text{KL}}(1 + \tilde{\eta} / 2)pL},
\]

(47)

(48)

with Kullback–Leibler divergence (or relative entropy)

\[
D_{\text{KL}}(p_1 \| p_2) := p_1 \ln \left( \frac{p_1}{p_2} \right) + (1 - p_1) \ln \left( \frac{1 - p_1}{1 - p_2} \right),
\]

(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

\[
D_{\text{KL}}(p_1 \| p_2) \neq D_{\text{KL}}(p_2 \| p_1),
\]

(50)

and for all \( 0 < p_1, p_2 < 1 \)

\[
D_{\text{KL}}(p_1 \| p_2) \geq 0
\]

(51)

with equality if and only if \( p_1 = p_2 \).

Finally, we obtain

\[
\Pr[\tilde{E}_{a_{\ell,n}}] \leq 2e^{-\frac{2p^2(1-\eta)^2}{\eta L}} + e^{-D_{\text{KL}}(1 + \tilde{\eta} / 2)pL}.
\]

\[
= 2e^{-\frac{1}{2} (1 - \tilde{\eta})^2 p^2 (1 - p)^2} + e^{-D_{\text{KL}}(1 + \tilde{\eta} / 2)pL}.
\]

(52)

(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 \( \ell \) and \( n \), and thus (23) yields (20) which concludes the proof.

\[ \square \]

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

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 \( \ell \in \{1, \ldots, L\} \), consider the least-squares problem

\[
\min_{\omega_\ell} \sum_{n=1}^{N} |(a_{n-1}, \omega_\ell) - a_{\ell,n}|^2 = \min_{\omega_\ell} \| \tilde{A} \omega_\ell - \tilde{a}_\ell \|^2,
\]

(54)

where

\[
\tilde{A} := \begin{pmatrix} a_1^T \\ \vdots \\ a_N^T \end{pmatrix} \in \mathbb{R}^{N \times L}, \quad \tilde{a}_\ell := \begin{pmatrix} a_{\ell,1} \\ \vdots \\ a_{\ell,N} \end{pmatrix} \in \mathbb{R}^{N}.
\]

(55)

Note that \( \tilde{A} \) is the transposed matrix of \( (a_N, a_1, \ldots, a_{N-1}) \in \mathbb{R}^{L \times N} \), i.e., of the one time-step cyclic shifted version of \( A \), and \( \tilde{a}_\ell \) is the \( \ell \)-th row of \( A \) turned into a column vector.

If \( \text{rank}(A) = N \), then

\[
\min_{\omega_\ell \in \mathbb{R}^L} \| \tilde{A} \omega_\ell - \tilde{a}_\ell \|^2 = 0,
\]

(56)

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

\[
\Pr[\text{rank}(A) = N] \geq 1 - (1 - p + o_N(1))^N,
\]

(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 \( A \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)^{L \times N} \) with \( L \geq N \), and in particular with

\[
L = N
\]

(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 \( \omega_\ell^{(0)} \) we proceed by

\[
\omega_\ell^{(n)} = \omega_\ell^{(n-1)} + \beta^{(n)} \tilde{A}^T(\tilde{a}_\ell - \tilde{A} \omega_\ell^{(n-1)}).
\]

(59)
for \( n = 1, \ldots, K, \ K \in \mathbb{N}, \) and with step size \( \beta(n) > 0. \) The recursion (59) with constant \( \beta(n) = \beta \) converges to a minimizer of (54) if
\[
0 < \beta < \frac{2}{\lambda_{\text{max}}(\mathbf{A}^T \mathbf{A})},
\]
where \( \lambda_{\text{max}}(\mathbf{A}^T \mathbf{A}) > 0 \) is the largest eigenvalue of \( \mathbf{A}^T \mathbf{A}. \)

Finally, replacing gradient descent as in (59) by stochastic gradient descent yields
\[
\mathbf{w}_n = \mathbf{w}_n^{(n-1)} + \beta(n) \left( \mathbf{a}_{\ell,n} - \mathbf{a}_{n-1} \right) \mathbf{a}_{n-1},
\]
which is (11). Again, as in (5) the column indices are taken modulo \( N \) and \( a_0 \coloneqq a_N. \) 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 \( \mathbf{A} \overset{i.i.d.}{\sim} \text{Ber}(p)^{L \times N} \) and \( |\mathcal{A}_{\text{typical}}| \) denotes the cardinality of \( \mathcal{A}_{\text{typical}}. \) Then, we have
\[
\lim_{L \to \infty} \frac{1}{L} \log_2 |\mathcal{A}_{\text{typical}}| = H_b(p)N,
\]
with the binary entropy function
\[
H_b(p) \coloneqq -p \log_2(p) - (1 - p) \log_2(1 - p)
\]
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
\[
C_{\text{absolute}} \geq \log_2 |\mathcal{A}_{\text{typical}}| \quad [\text{bits}].
\]

A. Capacity per Neuron

From (62) and (64) it follows that the asymptotic memorization capacity in bits per neuron is lower bounded by
\[
C_{\text{neuron}} \geq H_b(p)N \quad [\text{bits per neuron}].
\]

For the single-pass memorization rule (14) we have (21) (which is a consequence of Theorem 1), and we thus obtain
\[
C_{\text{single-pass}} \geq C_{\text{p}, \tilde{\eta}} \frac{L}{\ln(L)} \quad [\text{bits per neuron}]
\]
with constant
\[
C_{\text{p}, \tilde{\eta}} \coloneqq \frac{1}{16} (1 - \tilde{\eta})^2 p^2 (1 - p)^2 H_b(p) > 0,
\]
for \( 0 < p, \tilde{\eta} < 1. \) For the multi-pass memorization rule (59), we have (cf. (58))
\[
C_{\text{multi-pass}} \geq H_b(p) L \quad [\text{bits per neuron}].
\]
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
\[
C_{\text{connection}} = \frac{C_{\text{neuron}}}{L},
\]
because in both modes the network is fully connected, and we obtain
\[
C_{\text{single-pass}} \geq C_{\text{p}, \tilde{\eta}} \frac{1}{\ln(L)} \quad [\text{bits per connection}],
\]
\[
C_{\text{multi-pass}} \geq H_b(p) \quad [\text{bits per connection}].
\]
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) \( O(L/\ln(L)) \) 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