A Mathematical Analysis of Memory Lifetime in a simple Network Model of Memory

Pascal Helson*

October 14, 2019

Contents

1 Introduction 2

2 The model and the estimator 3
  2.1 The neural network and the protocol . . . . . . . . . . . . . . . . . . . . . . . . . 3
  2.2 Presentation of the estimator . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

3 Results 9
  3.1 Binomial mixture . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
  3.2 Main results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

4 Simulations 17

5 Discussion 20

A Proofs 23
  A.1 Proof of Proposition 3.5 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23
  A.2 Proof of Lemma 3.7 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23
  A.3 Proof of Lemmas 3.10 and 3.14 . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25

Abstract

We study the learning of an external signal by a neural network and the time to forget it when this network is submitted to other signals considered as noise. The presentation of an external stimulus changes the state of the synapses in a network of binary neurons. Multiple presentations of a unique signal leads to its learning. Then, the presentation of other signals also changes the synaptic weight (during the forgetting time). We study the number of external signals to which the network can be submitted until the initial signal is considered as forgotten. We construct an estimator of the initial signal thanks to the synaptic currents. In our model, these synaptic currents evolve as Markov chains. We study mathematically these Markov chains and obtain a lower bound on the number of external stimulus that the network can receive before the initial signal is forgotten. We finally present numerical illustrations of our results.

*pascal.helson@inria.fr
1 Introduction

Amit and Fusi proposed in [1] a model to study the memory capacity of neural networks. The main novelty of their work lies in the online learning and forgetting of an infinite sequence of random signals. The following experimental protocol was used. A neural network, with both binary synapses and binary neurons, receives and learns new random stimuli while forgetting the previous ones. Every signal may affect the synaptic weights. After a certain amount of time, the first stimulus is presented again (priming) and the ability of the network to recognize it is questioned: how many stimuli can be presented before it forgets the initial signal? To provide an answer, the authors of [1] performed a signal-to-noise ratio (SNR) analysis on the sum of the synaptic currents into a neuron when it receives the priming. They concluded that the stimuli need to be sparse in order to optimise the memory lifetime. They proposed a scaling of the coding level $f$ (probability a neuron has to be selective to a signal) as a function of the number $N$ of neurons in the network: the coding level optimising the memory lifetime according to their criterion is on the order of $f \sim \frac{\log(N)}{N}$. As $N$ is large, $f$ must be small which means sparse signals, and what they called optimal storage is then proportional to $\frac{1}{f^2}$.

This model has then been studied and extended in different articles [3–5,7,8,12,13]. They proposed different approaches to using SNR. First, in [5], Brunel et al. studied a different protocol, fixing the number of random stimuli and presenting them randomly multiple times. Their analysis relied on the comparison of two quantities: the mean potentiation (MP) and the intra-class potentiation (ICP). MP is the mean synaptic weight value and ICP is the mean synaptic weight among synapses candidate to potentiation when a stimulus is presented. Intuitively, when ICP is enough bigger than MP, the trace of this stimulus in the synaptic weights is still non negligible. They found two possible loading regimes, a low-loading (resp. high-loading) regime with a memory capacity of the order $f$ (resp. $\frac{1}{f}$). A deeper analysis of multiple [5] and one shot [1] learning models was done in [7] under the assumption of $N$ large and $f$ small. Then, in [8], the mean first passage time (MFPT) was considered. It corresponds to the mean number of signals presented before the synaptic current crosses a fixed threshold. More complex and biologically plausible models have been proposed in the following studies [2,12,17]. And finally to the best of our knowledge, the first article to present a precise way to retrieve stimuli is [3]. In this article, Amit and Huang insisted on the role played by the synaptic correlations and proposed a way to compute numerically an approximation of the distributions of the synaptic currents. It enables them to introduce a new retrieval criterion based on what they called retrieval probabilities.

Inspired by this article, we propose a statistical test based on the synaptic currents. Our study relies on how the initial stimulus can be estimated from this current. We extend previous analytical studies [1,3,8] on many points. First, we give properties of the synaptic current process such as the spectrum of its transition matrix, see Propositions 3.11 and 3.12. Secondly, we study the case of multiple presentations of the signal to be learnt. Finally, we obtain in our main result (Theorem 3.15) explicit bounds of the time spent under a certain probability to misevaluate the initial signal. In addition, this enables us to conclude on the importance of taking into account both homosynaptic and heterosynaptic depression.

The rest of the paper is organised as follows. We expose the model and the statistical test in Section 2. Section 3 is devoted to the study of the synaptic currents, the evaluation of the error of the test and the maximum number of stimuli one can present to reasonably remember the initial signal. The main result is presented and proved in this section. Then, we perform numerical simulations in Section 4. Finally, technical results are proved in the Appendix A.
2 The model and the estimator

First, we present the neural network and the protocol followed for learning and forgetting. Then, we define our estimator, we derive the equations describing the dynamics of the synaptic currents and we present our assumptions. Finally, we present typical numerical simulations at the end of this section.

2.1 The neural network and the protocol

In order to ease the introduction of the different variables, we associate to the model the following function of time. The synaptic weight can only take two values and we denote by $J = \{J^i_j, i \neq j\} \in \{J_-, J_+\}^{N(N+1)}$ the matrix of synaptic weights. We consider a plasticity rule which can be viewed as a classic Hebbian rule: the law of $J_{t+1}$ only depends on $J_t$ and $\xi_t$. The corresponding transition probabilities are

\[
\begin{align*}
\mathbb{P}\left(J_{t+1}^{ij} = J_+ | J_t^{ij} = J_-, (\xi_t^i, \xi_t^j) = (1, 1)\right) &= q^+, \\
\mathbb{P}\left(J_{t+1}^{ij} = J_- | J_t^{ij} = J_+, (\xi_t^i, \xi_t^j) = (0, 1)\right) &= q_{01}, \\
\mathbb{P}\left(J_{t+1}^{ij} = J_- | J_t^{ij} = J_+, (\xi_t^i, \xi_t^j) = (1, 0)\right) &= q_{10}, \\
\mathbb{P}\left(J_{t+1}^{ij} = J_+ | J_t^{ij} = J_+, (\xi_t^i, \xi_t^j) = (0, 0)\right) &= 1.
\end{align*}
\]

In order to simplify the notations and without loss of generality, we set:

\[J_- = 0 \text{ and } J_+ = 1.\]

Moreover, in order to avoid critical cases, we also assume that

\[f, q_{01}, q_{10}, q^+ \in [0, 1].\]  

(1)
The parameters \( q_{01} \) and \( q_{10} \) represent respectively the homosynaptic and heterosynaptic depressions.

We now give the protocol to learn and then forget a signal. The signal to learn is considered to be \( \xi_0 \). Before presenting it, we assume that the network has received a lot of random signals thereby driving the law of the synaptic weights matrix in its “stable” state at time \( t = -r + 1 \) (we prove in Proposition 2.7 that it exists and is unique). In order to learn \( \xi_0 \), we present it \( r \) times to the network. To be consistent with the previous description, the sequence of presented stimuli is then \((\cdots, \xi_{-r}, \xi_0, \cdots, \xi_0, \xi_1, \xi_2 \cdots)\) that is \( \xi_t = \xi_0 \) for \( t \in [-r + 1, 0] \). The presentation of the subsequent signals leads to the forgetting of \( \xi_0 \).

### 2.2 Presentation of the estimator

We study the consistency through time of the response of a neuron to the initial signal. To do so, we consider the previous protocol. After the repetitive presentation of \( \xi_0 \), it has left a certain footprint in the matrix \( J_1 \) which is erased gradually by the presentation of the following signals. How much is left from the learning at time \( t \)? As an answer, we define an error associated to a decision rule based on the projection of \( J_t \) on \( \xi_0 \). For neuron \( i \), such a projection at time \( t \) is given by \( \sum_{j \neq i} J_{ij}^t \xi_0^j \). In this framework, neurons are similar. Hence, in order to simplify the notations and without loss of generality, our study focuses on neuron 1. We denote by \( h_t \):

\[
h_t = \sum_{j=2}^{N+1} J_{1j}^t \xi_0^j, \tag{2}
\]

the synaptic current to neuron 1 when presenting again \( \xi_0 \) at time \( t \). In this framework, the initial signal is presented in a fictive way. This means that the synaptic weights do not change following this fictitious presentations. Note that \((h_t)_{t \geq 0}\) strongly depends on the initial number of active neurons that we denote by \( K \):

\[
K = \sum_{j=2}^{N+1} \xi_0^j.
\]

More particularly, conditioning on \( K \) ensures that the process \( h_{t,K} \) is Markovian (Proposition 2.2). Our estimator only depends on the postsynaptic current \( h_t \). We use a threshold \( \theta \) to define the estimator \( \hat{\xi}(t, \theta) : N^* \times [0, N] \rightarrow \{0, 1\} \):

\[
\hat{\xi}(t, \theta) = \mathbb{1}_{h_t > \theta}.
\]

This choice is justified in Section 2.2. The errors associated to the estimator are

\[
p_0^0(t, \theta) = \mathbb{P} \left( \hat{\xi}(t, \theta) = 1 \mid \xi_0^1 = 0 \right) = \mathbb{P} \left( h_t > \theta \mid \xi_0^1 = 0 \right),
\]

\[
p_0^1(t, \theta) = \mathbb{P} \left( \hat{\xi}(t, \theta) = 0 \mid \xi_0^1 = 1 \right) = \mathbb{P} \left( h_t \leq \theta \mid \xi_0^1 = 1 \right).
\]

\( p_0^0(t, \theta) \) (resp. \( p_0^1(t, \theta) \)) corresponds to the probability that the estimator responds positively (resp. negatively) to the priming presented at time \( t > 0 \) whereas the neuron was not activated (resp. activated) initially. In the following, we note

**Notation 2.1.**

\( h^y_t \triangleq \langle h_t \mid \xi_0^1 = y \rangle, \quad h_{t,K} \triangleq \langle h_t \mid K \rangle \quad \text{and} \quad h^y_{t,K} \triangleq \langle h_t \mid \xi_0^1 = y, K \rangle. \)
We aim at evaluating these errors: for fixed $\delta \in ]0,1[$, we estimate the largest time $t_*$ such that both $p^0$ and $p^y$ are smaller than $\delta$ up to time $t_*$.

$$t_*(\delta,r,N) := \max_{\theta \in [0,N]} \left( \inf \left\{ t \geq 1, p^0_t(t,\theta) \lor p^y_t(t,\theta) \geq \delta \right\} \right).$$

(3)

Our aim is to propose a method to find a lower bound on $t_*(\delta,r,N)$. Our study relies on the processes $(h_t^y)_{t \geq 0}$, $y \in \{0,1\}$.

We use a recursive formula conditioning on the initial number of active neurons, $K$. Therefore, we first study the Markov chain $h_{t,K} \overset{\text{law}}{=} (h_t|K)$ and then come back on $h_t$ at the end of the analysis.

**Proposition 2.2.** At the end of the learning phase, that is at time $t = 1$, we have

$$h_{1,K} = h_{-r+1,K} + \xi^1_0 \text{Bin} \left( K - h_{-r+1,K}, 1 - (1 - q^+)^r \right) - (1 - \xi_0^1) \text{Bin} \left( h_{-r+1,K}, 1 - (1 - q_{01})^r \right)$$

where, conditionally on $h_{-r+1,K}$, the two binomial random variables are independent.

And $\forall t \geq 1$:

$$h_{t+1,K} \overset{\text{law}}{=} h_{t,K} + \xi^1_t \left[ \text{Bin} \left( K - h_{t,K}, f q^+ \right) - \text{Bin} \left( h_{t,K}, (1 - f) q_{01} \right) \right] - (1 - \xi^1_t) \text{Bin} \left( h_{t,K}, f q_{01} \right)$$

(5)

where, conditionally on $h_{t,K}$, the three binomial random variables are independent.

**Proof.** In order to study the jump from $h_{t,K}$ to $h_{t+1,K}$, we count synapses that potentiate and the ones that depress upon presenting a signal $\xi_t$. From (2) and the definition (2.1) of $h_{t,K}$, we only need to consider the $K$ synapses $J_{t,j}^y$ with $j$ such that $\xi^1_0 = 1$. At time $t$, there are $h_{t,K}$ strong synapses and $K - h_{t,K}$ weak synapses. Given $\xi^1_t$ and $h_{t,K}$, every synapse evolves independently following a Bernoulli law. From time $-r + 1$ to $1$, if $\xi^1_0 = 0$ every strong synapse is $r$ times candidate to depression so it has probability $(1 - q_{01})^r$ to depress. If $\xi^1_0 = 1$ every weak synapse is $r$ times candidate to potentiation so it has probability $(1 - q^+)^r$ to potentiate. Equation (4) follows.

Let now consider time $t \geq 1$. From time $t$ to $t+1$, if $\xi^1_t = 0$, the probability that a strong synapse depresses is $f q_{01}$. If $\xi^1_t = 1$, the probability that a weak synapse potentiates is $f q^+$ and the probability that a strong synapse depresses is $(1 - f) q_{10}$. Equation (5) follows. \hfill \Box

**Corollary 2.3.** At time $t = 1$,

$$h^0_{1,K} \overset{\text{law}}{=} h_{-r+1,K} - \text{Bin} \left( h_{-r+1,K}, 1 - (1 - q_{01})^r \right)$$

$$h^1_{1,K} \overset{\text{law}}{=} h_{-r+1,K} + \text{Bin} \left( K - h_{-r+1,K}, 1 - (1 - q^+)^r \right).$$

(6)

(7)

where, conditionally on $h_{-r+1,K}$, the two binomial random variables are independent.

And $\forall t \geq 1$, $y \in \{0,1\}$,

$$h^y_{t+1,K} \overset{\text{law}}{=} h^y_{t,K} + \xi^1_t \left[ \text{Bin} \left( K - h^y_{t,K}, f q^+ \right) - \text{Bin} \left( h^y_{t,K}, (1 - f) q_{01} \right) \right] - (1 - \xi^1_t) \text{Bin} \left( h^y_{t,K}, f q_{01} \right)$$

(8)

where, conditionally on $h^y_{t,K}$, the three binomial random variables are independent.

**Proof.** Same as Proposition 2.2. \hfill \Box
Remark 2.4. From the previous Proposition and its Corollary, the chains \((h_{t,K})_{t \in \mathbb{N}}\) and \((h^y_{t,K})_{t \in \mathbb{N}}\) are Markovian.

We deduce the

Corollary 2.5. Assume \([1]\) holds. Then, for all \(0 \leq K \leq N\), the Markov chain \((h_{t,K})_{t \geq 0}\) admits a unique invariant measure \(\pi_K\) with support on \([0,K]\). Furthermore, for any initial condition \(h_{0,K}\), the Markov chain \((h_{t,K})_{t \geq 0}\) converges in law to \(\pi_K\).

Proof. The Markov chain \((h_{t,K})_{t \geq 0}\) is irreducible and aperiodic on a finite space. Thus, it admits a unique invariant measure towards which it converges.

Remark 2.6. The Markov chains \((h^0_{t,K})_{t \geq 1}\) and \((h^1_{t,K})_{t \geq 1}\) have the same transition matrix as \((h_{t,K})_{t \geq 1}\). They differ by their distribution at time \(t = 1\). Hence, they both converge to \(\pi_K\).

Proposition 2.7. Under the assumption \([1]\), the process \((\xi_t, J_t)_{t \geq 1}\) converges to its unique invariant measure. We denote it by \(\rho_\infty\).

Proof. The Markov chain \((\xi_t, J_t)_{t \geq 1}\) is irreducible and aperiodic on a finite space. Thus, it admits a unique invariant measure towards which it converges.

We now give the main assumptions.

Assumption 2.8.

1. \((\xi_0, J_{-r+1}) \equiv \rho_\infty\) and in particular \(h_{-r+1,K}, h^0_{-r+1,K}, h^1_{-r+1,K} \equiv \pi_K\).

2. Assume \(f\) depends on \(N\). Let us denote it by \(f_N\) such that \(\lim_{N \to \infty} f_N = 0\) and \(\lim_{N \to \infty} N f_N = +\infty\).

3. Let \(q^-_{01} = a_N f_N\) and \(q^-_{10} = b_N f_N\) with \(a_N, b_N \in \mathbb{R}^+\) such that:
   \[
   \lim_{N \to \infty} a_N = \lim_{N \to \infty} b_N = 0, \quad \lim_{N \to \infty} a_N f_N = \lim_{N \to \infty} b_N f_N = +\infty. \tag{9}
   \]

We consider a general paradigm in which before receiving the stimulus \(s_0\), many stimuli have already been sent \((\cdots, s_{-r-2}, s_{-r-1}, \cdots)\). We assume that the process \((\xi_t, J_t)_{t \leq -r+1}\) has reached its invariant measure at time \(t = -r + 1\) and Assumption 2.8.1. Then, one key parameter is the coding level \(f\). It is assumed to depend on \(N\) in the analysis of the large \(N\) asymptotic, Assumption 2.8.2. This assumption refers to sparse coding as \(f_N\) tends to 0. This assumption stresses the need of another one. Indeed, one needs to control the way \(f_N\) converges to 0. The constraint put forward is that the average information needs to be high enough: we assume that the mean number of selective neurons, \(N f_N\), is large with \(N\), Assumption 2.8.2. In this context, we are interested to see how the dependence in \(N\) of depressing probabilities can affect the memory lifetime, see Assumption 2.8.3. Moreover, this assumption prevents \(q^-_{01}\) and \(q^-_{10}\) from being too small or too big compared to \(f_N\).

Remark 2.9. From now on, all the considered processes depend on \(N\) through the parameters \(f_N, a_N\) and \(b_N\). In order to shorten notations, we do not add "\(N\)" in the notations.
First illustrations

In order to get some intuitions on the model, we plot the time evolution of the processes \((h^{y}_{t,K})_{t \geq 0}\). We assume that the signal \(\xi_0\) is of size \(K = \lfloor f_N N \rfloor\), where the floor function \(\lfloor x \rfloor\) is equal to \(k \in \mathbb{Z}\) if \(k \leq x < k + 1\). Let have a look at the expected size of jumps of \(h_{t,K}\) from the recursive formula (4), (5). At time \(t = 1\):

\[
\mathbb{E}[h_{1,K} - h_{-r+1,K} | h_{-r+1,K}, \xi_0^1 = 0] = -h_{-r+1,K}(1 - (1 - q_0^1)^r),
\]

(10) \[
\mathbb{E}[h_{1,K} - h_{-r+1,K} | h_{-r+1,K}, \xi_0^1 = 1] = (K - h_{-r+1,K})(1 - (1 - q^r)^r).
\]

(11)

At time \(t > 1\):

\[
\mathbb{E}[h_{t+1,K} - h_{t,K} | h_{t,K}] = (K - h_{t,K})f_N^2 q^+ - h_{t,K} f_N (1 - f_N)(q_{10}^- + q_0^-).
\]

(12)

The average jump size strongly depends on \(f_N\). We note that between time \(t = 1\) (equations (11) and (10)) and \(t > 1\) (equation (12)), the parameter \(f_N\) (and even \(f_N^2\) under assumption 2.8.3) appears in factor. Therefore, the major effect of \(f_N\) is on the forgetting phase.

Hence, when \(f_N\) is big, close to 1, the reception of \(\xi_0\) has a large impact on the weight matrix, easy to detect. However, the following average jump size are close to the initial one as soon as some other stimuli are presented, the initial signal is forgotten: the distributions of \(h^0_{t,K}\) and \(h^1_{t,K}\) quickly overlap. In order to illustrate this phenomenon, we plot one trajectory of \(h_{t,K}\) and the distributions of \(h^0_{t,K}\) and \(h^1_{t,K}\) after a learning phase with one presentation of the signal, \(r = 1\).
Figure 1: We used the parameters: \( r = 1, N = 1000, K = 800, f_N = 0.8, q^+ = 0.8, q_0^+ = 0.8 \) and \( q_0^- = 0.2 \). The first figure, 1a, shows a typical trajectory of \( h_{t,800} \). In 1b, the distributions of \( h_{t,800} \) and the invariant measure \( \pi_{800} \) are plotted. Then, the distributions of \( h_{t,800} \) are plotted at time \( t = 3 \) in 1c and time \( t = 5 \) in 1d.

The size of the jumps can be an explanation of the multi-modal form of the distributions of \( h_{t,K}^p \) plotted in 1b, 1c and 1d. In 1b, one can see the changes from time \( t = -r + 1 \) to \( t = 1 \). Indeed, at time \( t = -r + 1 = 0 \), \( h_{0,K}^0 \) and \( h_{0,K}^1 \) both follow the invariant measure in black. Then, after the reception of \( \xi_0 \), the distribution of \( h_{t,K}^1 \) is drifted on left and the distribution of \( h_{t,K}^0 \) on the right. The first signal is learnt as distributions are well separated. After that, the reception of new stimuli makes them converge to the invariant distribution, time \( t = 3 \) in 1c and time \( t = 5 \) in 1d. After 5 presentations, the signal is already forgotten.

Conversely, when \( f_N \) is small, \( f_N \) close to 0, the difference between the learning and the forgetting phase is more significant. Indeed, the convergence to the stationary distribution, and thus forgetting, is slower. However, the learning still occurs: the initial jump is still big.
Even after 20 presentations, the two distributions do not overlap a lot and they stayed uni-modal. Thus, we see that when $f_N$ is small enough, the distributions are uni-modal. This makes the choice of a threshold estimator reasonable. Moreover, such an estimator allows a tractable analysis.

3 Results

In this section, we first give some properties satisfied by the distributions of $h^{N}_t,K$, see notation 2.1, and the invariant measure $\pi_K$. They enable us to prove Theorem 3.15, our main result, in the second part.

3.1 Binomial mixture

Denote by $F_{[0,1]}$ the set of cumulative distribution functions associated to $\mathcal{P}([0,1])$, the set of probability measures on $[0,1]$.

**Definition 3.1.** The distribution of $X$ is said to be a binomial mixture with mixing distribution $g \in \mathcal{P}([0,1])$ and size parameter $K$, denoted $\text{BinMix}(K,g)$, if

$$\forall j \in [0,K], \quad P(X = j) = \binom{K}{j} \int_0^1 u^j(1-u)^{K-j} g(du).$$

**Remark 3.2.**

- $X \overset{d}{=} \text{BinMix}(K,g)$ is equivalent to $X|Y \overset{d}{=} \text{Bin}(K,Y)$ where $Y$ is a random variable independent of the binomial and with law $g$. Indeed

$$P(X = j) = \int_0^1 P(X = j|Y = u) g(du) = \binom{K}{j} \int_0^1 u^j(1-u)^{K-j} g(du).$$

We use both notations $X \overset{d}{=} \text{BinMix}(K,g)$ and $X \overset{d}{=} \text{BinMix}(K,Y)$ in the following.

- The law of $X$ is fully characterized by the moments $E(Y), E(Y^2), \ldots, E(Y^K)$. Hence, if $\tilde{g} \in \mathcal{P}([0,1])$ is such that

$$\forall k \in [0,K], \quad \int_0^1 u^k \tilde{g}(du) = \int_0^1 u^k g(du),$$
then $\text{BinMix}(K, \tilde{g}) \equiv \text{BinMix}(K, g)$.

First, we show that the set of binomial mixtures is stable by the Markov chain $h_{t,K}$: assume that $h_{t,K} \equiv \text{BinMix}(K, g_t)$ for some $g_t \in \mathcal{P}([0,1])$, then there exists $g_{t+1} \in \mathcal{P}([0,1])$ such that $h_{t+1,K} \equiv \text{BinMix}(K, g_{t+1}) \equiv \text{BinMix}(K, Y_{t+1})$. Then, assuming that $h_{t,K}$ is a binomial mixture, we give an explicit expression of $G_{t+1}$, the cumulative distribution function associated to $g_{t+1}$, as function of the cumulative distribution function $G_t$. Indeed, for all $t \geq 2$, $G_{t+1}(x) = \mathcal{R}(G_t)(x)$, where

**Notation 3.3.** $\mathcal{R}$ is defined as

$$\forall \Gamma \in F_{[0,1]}, \ u \in \mathbb{R}, \ \mathcal{R}(\Gamma)(u) \overset{df}{=} f_N \Gamma \left( \frac{u - f_N q^+}{1 - (1 - f_N) q_{10} - f_N q^+} \right) + (1 - f_N) \Gamma \left( \frac{u}{1 - f_N q_{01}} \right).$$

**Proposition 3.4.** Assume that $h_{-r+1,K} \equiv \text{BinMix}(K, g_{-r+1})$, where $g_{-r+1} \in \mathcal{P}([0,1])$. Then for all $t \geq 1$, $g_t$, $g_0^b$, $g_1^b \in \mathcal{P}([0,1])$ such that $h_{t,K} \equiv \text{BinMix}(K, g_t)$ and $h_{t,K}^w \equiv \text{BinMix}(K, g_t^w)$ for $y = 0, 1$. Moreover, at time $t = 1$

$$G_1(u) = f_N G_{-r+1} \left( \frac{u + 1}{1 - q^+} - 1 \right) + (1 - f_N) G_{-r+1} \left( \frac{u}{1 - q_{01}^+} \right),$$

$$G_1^b(u) = G_{-r+1} \left( \frac{u + 1}{1 - q^+} - 1 \right) \quad \text{and} \quad G_1^w(u) = G_{-r+1} \left( \frac{u}{1 - q_{01}^+} \right),$$

and $\forall t \geq 1$,

$$G_{t+1}(u) = \mathcal{R}(G_t)(u) \quad \text{and} \quad G_{t+1}^w(u) = \mathcal{R}(G_t^w)(u).$$

Finally, we show that $\mathcal{R}$ is contracting and characterises $\pi_K$.

**Proposition 3.5.** The application $\mathcal{R}$ acting on $F_{[0,1]}$ is contracting for the norm $\| \cdot \|_{L^1(0,1)}$. Moreover, there exists a unique $G^* \in F_{[0,1]}$ invariant for $\mathcal{R}$.

**Corollary 3.6.** Let $G^*$ be the unique fixed point of $\mathcal{R}$ and $g^*$ its associated distribution. The invariant measure $\pi_K$ of the Markov chain $h_{t,K}$ satisfies

$$\pi_K = \text{BinMix}(K, g^*).$$

Moreover, the support of $g^*$ verifies

$$\text{Supp}(g^*) \subset \left[ 0, \frac{f_N q^+}{f_N q^+ + (1 - f_N) q_{10}} \right] := [0, w^*].$$

The proof of Proposition 3.4 relies on the following Lemma proved in the Appendix A.2

**Lemma 3.7.** Let $Z$ be a mixture of binomial $Z = \text{BinMix}(K, Y_Z)$. Let $0 \leq b < a < 1$. Conditionally on $K$, consider two independent binomial distributions Bin$(Z, a)$ and Bin$(K - Z, b)$ and define $X = \text{Bin}(Z, a) + \text{Bin}(K - Z, b)$. Then

$$X \equiv \text{BinMix}(K, Y_X) \ \text{with} \ \ Y_X = (a - b) Y_Z + b.$$
In the following, we use

**Notation 3.8.** Let $Z$ be a random variable in $[0, 1]$ with distribution $g_Z$ and cumulative distribution function $G_Z$. We denote by $g_{Z, (a, b)} \in P([0, 1])$ the distribution defined by

$$G_{Z, (a, b)}(u) = G_Z \left( \frac{u - b}{a - b} \right).$$

**Proof of Proposition 3.4**

**Proof.** We first show (13) and (15) for $h_{t, K}$, then the rest follows. At $t = 1$,

$$\mathbb{L} \left( h_{1, K} | \xi_0^1 = 1, h_{r-1+1, K} \right) = h_{r+1, K} + \text{Bin} \left( K - h_{r+1, K}, 1 - (1 - q^+)r \right)$$

$$\mathbb{L} \left( h_{1, K} | \xi_0^1 = 0, h_{r+1, K} \right) = \text{Bin} \left( h_{r+1, K}, (1 - q_0^+)r \right).$$

Applying twice Lemma 3.7 with $(a, b) = (1, 1 - (1 - q^+)r)$ and $(a, b) = (1 - (1 - q_0^+)r, 0)$, we obtain using notation 3.8,

$$\mathbb{L} \left( h_{1, K} | \xi_0^1 = 1, h_{r+1, K} \right) \leq \text{BinMix} \left( K, g_{r+1, (1, 1 - (1 - q^+)r)} \right)$$

$$\mathbb{L} \left( h_{1, K} | \xi_0^1 = 0, h_{r+1, K} \right) \leq \text{BinMix} \left( K, g_{r+1, (1 - (1 - q_0^+)r, 0)} \right).$$

Thus,

$$\mathbb{P} \left( h_{1, K} = j | h_{r+1, K} \right) = \mathbb{P} \left( \xi_0^1 = 1 \right) \mathbb{P} \left( h_{1, K} = j | \xi_0^1 = 1, h_{r+1, K} \right)$$

$$+ \mathbb{P} \left( \xi_0^1 = 0 \right) \mathbb{P} \left( h_{1, K} = j | \xi_0^1 = 0, h_{r+1, K} \right)$$

$$= f_N \left( \frac{K}{j} \right) \int_0^1 w^j (1 - u)^{K-j} g_{r+1, (1, 1 - (1 - q^+)r)} (du)$$

$$+ (1 - f_N) \left( \frac{K}{j} \right) \int_0^1 w^j (1 - u)^{K-j} g_{r+1, (1 - (1 - q_0^+)r, 0)} (du)$$

$$= \left( \frac{K}{j} \right) \int_0^1 w^j (1 - u)^{K-j} \left( f_N g_{r+1, (1, 1 - (1 - q^+)r)} \right) (du)$$

$$+ (1 - f_N) g_{r+1, (1 - (1 - q_0^+)r, 0)} (du),$$

which enables to get (13).

Now, assume that $h_{t, K} \equiv \text{BinMix}(K, g_t)$, for some fixed $t \geq 1$. Then equation (6) gives

$$\mathbb{L} \left( h_{t+1, K} | \xi_0^1 = 1, h_{t, K} \right) = \text{Bin} \left( K - h_{t, K}, f_N q^+ \right) + \text{Bin} \left( h_{t, K}, 1 - (1 - f_N)q_{10}^t \right)$$

$$\mathbb{L} \left( h_{t+1, K} | \xi_0^1 = 0, h_{t, K} \right) = \text{Bin} \left( h_{t, K}, 1 - f_Nq_{10}^t \right),$$

where binomials are independent conditionally on $h_{t, K}$. Applying twice Lemma 3.7 with $(a, b) = (1 - (1 - f_N)q_{10}^t, f_N q^+)$ and $(a, b) = (1 - f_Nq_{10}^t, 0)$, we get

$$\mathbb{L} \left( h_{t+1, K} | \xi_0^1 = 1 \right) \leq \text{BinMix} \left( K, g_t, (1 - (1 - f_N)q_{10}^t, f_N q^+) \right)$$

$$\mathbb{L} \left( h_{t+1, K} | \xi_0^1 = 0 \right) \leq \text{BinMix} \left( K, g_t, (1 - f_Nq_{10}^t, 0) \right).$$

Hence, $h_{t+1, K} \leq \text{BinMix} \left( K, f_N g_t, (1 - (1 - f_N)q_{10}^t, f_N q^+) \right) + (1 - f_N) g_t, (1 - f_Nq_{10}^t, 0)$, and we deduce that $h_{t+1, K} \leq \text{BinMix}(K, g_{t+1})$ with $G_{t+1}(x) = R(G_t)(x)$. For the processes $\left( h_{t, K} \right)_{t \geq 0}$, we proceed exactly with the same method using Corollary 2.3.
Proposition 3.3 is proved in Appendix A.1. We now prove Corollary 3.6.

**Proof.** Assume that at time t, \( h_t \) is BinMix\((K, g^*)\) has the invariant distribution. Thanks to Proposition 3.4 and \( \mathcal{R}(G^*) = \Gamma^* \), we prove the first point. Now, let \([u_m, u^*]\) be the convex envelop of the support of \( g^* \), then \( \text{Supp}(\gamma^*) \subset [u_m, u^*] \subset [0, 1] \). Thus by \( \mathcal{R}(G^*) = \Gamma^* \), we get

\[
\begin{align*}
    u_m &= \min (u_m (1 - f_N q_{10}^0), u_m (1 - f_N q_{10}^0 - f_N q^+ + f_N q^+)), \\
    u^* &= \max (u^* (1 - f_N q_{10}^0), u^* (1 - f_N q_{10}^0 - f_N q^+ + f_N q^+)),
\end{align*}
\]

which implies that \( u_m = 0 \) and \( u^* = \frac{f_N q^+}{f_N q^+ + (1 - f_N q_{10}^0)} \). \( \square \)

**Remark 3.9.** The results of Lemma 3.7, Propositions 3.4, 3.5, and Corollary 3.6 are present in [3], but for completeness, we prove them again with a different method. A small difference with their result is that we consider a more general case: \( q_{10}^0 \) can be positive whereas \( q_{10}^0 = 0 \) in [3]. Moreover, we give more precision on the support of \( g^* \).

We give a last Lemma useful to describe the supports of \( g_t^y \) in what follows. It is proved in the Appendix A.3.

**Lemma 3.10.** Let \( \eta, a, b, c \) be in \([0, 1]\) and such that \( \frac{a}{1 - b}, \frac{a}{c - b} \in [0, 1] \). Let \( (\gamma_t)_{t \geq 1} \) be a sequence in \( \mathcal{P}([0, 1]) \) such that the corresponding cumulative distribution functions \( (\Gamma_t)_{t \geq 1} \) satisfy:

\[
\forall t \geq 1, \quad \Gamma_{t+1}(u) = \eta \Gamma_t \left( \frac{u - a}{b} \right) + (1 - \eta) \Gamma_t \left( \frac{u}{c} \right). \tag{18}
\]

Let \([u_1^m, u_1^M]\) be the smallest interval containing the support of \( \gamma_1 \), \( \text{Supp}(\gamma_1) \subset [u_1^m, u_1^M] \subset [0, 1] \). Assume that \( u_1^m > \frac{a}{1 - b} \) and \( u_1^M < \frac{a}{c - b} \).

Then,

\[
\begin{align*}
    \forall t \geq 1 & \quad u_t^M = \left( u_1^M - \frac{a}{1 - b} \right) b^{t-1} + \frac{a}{1 - b}, \tag{19} \\
    \forall 1 \leq t \leq [t_c] + 2 & \quad u_t^m = \left( u_1^m - \frac{a}{1 - b} \right) b^{t-1} + \frac{a}{1 - b}, \tag{20} \\
    \forall t \geq [t_c] + 2 & \quad u_t^m = u_{[t_c] + 2} b^{t-[t_c]-2}, \tag{21}
\end{align*}
\]

where

\[
    t_c = \frac{-\log \left( \frac{a(1-c)}{(1-b)(c-b)(u_1^m - \frac{a}{1-b})} \right)}{\log (b)} > 0.
\]

### 3.2 Main results

The learning phase and the forgetting phase are both described by Markov chains. We first give the spectrum of the transition matrices associated to these chains and then we give our main result on \( t_c \).

**Spectrum**

Let \( P_{y,K} \) be the transition matrix of the synaptic current after one presentation of \( \xi_0 \) knowing that \( \xi_0^1 = \gamma \). We can then write \( h_{1,K}^y = (P_{y,K})^y h_{-r+1,K}^y \).
Proposition 3.11. The spectrum of $P_{0,K}$ and $P_{1,K}$ are

$$\Sigma(P_{0,K}) = \left\{(1 - q_{01})^i, \ 0 \leq i \leq K \right\} \quad \text{and} \quad \Sigma(P_{1,K}) = \left\{(1 - q^+)^i, \ 0 \leq i \leq K \right\}. $$

Proof. $P_{0,K}^{ij} = \binom{j}{i}(q_{01})^i(1 - q_{01})^{j-i}$ and $P_{1,K}^{ij} = (K^{-1})(q^+)^{j-i}(1 - q^+)^{K-j}$. \hfill \qed

Proposition 3.12. The spectrum of the transition matrix $P_K$ of $(h_{t,K})_{t \geq 1}$ is

$$\Sigma(P_K) = \left\{\lambda_i = (1 - f_N)(1 - f_Nq_{01})^i + f_N(1 - f_Nq_{10} - f_Nq^+)^i, \ 0 \leq i \leq K \right\}. $$

Proof. We denote by

$$A_0 = 1 - f_Nq_{01} \quad \text{and} \quad A_1 = 1 - (1 - f_N)q_{10} - f_Nq^+, $$

and $Q_K, \tilde{P}_K$ two matrices in $\mathbb{R}^{(K+1) \times (K+1)}$ such that $\forall 0 \leq i,j \leq K:

$$Q_K^{ij} = \binom{K}{i} \binom{i}{j}(-1)^{i-j} \quad \text{and} \quad \tilde{P}_K^{ij} = f_N \binom{j}{i}A_1^i (f_Nq^+)^{j-i} + (1 - f_N)\delta_{ij}A_0^j. $$

First, a straightforward computation shows that $Q_KP_K = \tilde{P}_KQ_K$. The matrix $Q_K$ is triangular with positive diagonal coefficients, so it is invertible. Thus,

$$P_K = (Q_K)^{-1} \tilde{P}_KQ_K. $$

The matrices $P_K$ and $\tilde{P}_K$ have the same spectrum and $\tilde{P}_K$ is triangular. \hfill \qed

We deduce from Proposition 3.12 the rate of convergence of the law of $h_{t,K}$ to the invariant measure.

Corollary 3.13. For all $0 \leq K \leq N$, the law of $(h_{t,K})_{t \geq 1}$ converges exponentially fast to the unique invariant measure $\pi_K$ on $[0,K]$. In particular, $\exists \epsilon \in \mathbb{R}^+$ such that the distance in total variation between the distribution of $h_{t,K}$ and the invariant measure satisfies:

$$\forall t \geq 1, \ |\mathcal{L}(h_{t,K}) - \pi_K|_{TV} \leq c\lambda_1^t. $$

Therefore, $\lambda_1$ has to be as close as possible to one in hope of a slow forgetting. How does this eigenvalue appears in the two errors we aim to compute? Only one part of it plays a role in our main result, $\lambda_1$.

Memory lifetime

We study the projection of the stimulus $\xi_0$ of size $K$ on the synaptic matrix through time, $h_{t,K}$. Under Assumption 2.8.1, $h_{t+1,K}$ is under its invariant distribution $\pi_K$, a binomial mixture by Corollary 3.6. Thus, from Proposition 3.4, the conditional processes $(h_{t,K}^y)_{t \geq 0}$ associated to $(h_{t,K})_{t \geq 0}$ also are binomial mixtures. Combining inequalities on binomial tails and a control on the support of the mixing distributions of the binomial mixtures $h_{t,K}^y$ and $h_{t,K}$, we prove the Proposition 3.15.

The following Lemma is proved in Appendix A.3.

Lemma 3.14. Let $X \overset{d}{\leftarrow} \text{Bin}(n,p)$. Then,

$$\forall \epsilon \in [0,1[, \quad \mathbb{P}(X > np(1 + \epsilon)) \leq \exp \left(\frac{-np \epsilon^2}{2 + \epsilon}\right), $$

$$\mathbb{P}(X < np(1 - \epsilon)) \leq \exp \left(\frac{-np \epsilon^2}{2}\right). $$
Main Result

Theorem 3.15. Let \((h^0_t)_{t \geq 0}\) (resp. \((h^1_t)_{t \geq 0}\)) satisfying (6) and (5) (resp. (7) and (8)). Let assume that \(q_{01}, q_{10}, q^+\) are fixed in \([0,1]\) and Assumptions 2.8.1 and 2.8.2 hold. Then, for all \(0 < \delta < 1, r \in \mathbb{N}^+\) there exists \(N(\delta, r) \in \mathbb{N}\) such that for all \(N \geq N(\delta, r)\), there exists \(\theta \in [0, N]\) such that \(\max\{\mathbb{P}(h^0_t > \theta), \mathbb{P}(h^1_t \leq \theta)\} \leq \delta\) for all time \(t\) satisfying

\[
1 \leq t \leq t_0(\delta, r, N) \leq t_*(\delta, r, N).
\]

In particular, we give explicit conditions on \(N\) and \(r\) for a given \(\delta\) and an explicit formula of the lower bound \(t_0\) in Remark 3.16.

Proof. Let \(\delta \in ]0, 1[, r \in \mathbb{N}^+\) and \(N \in \mathbb{N}\). We give conditions on \(N\) along the proof. Let \(\theta \in [0, N]\) a fixed threshold. We study the probabilities of the errors associated with the threshold estimator. These probabilities are given by \(p^0_c(t, \theta) = \mathbb{P}(h^0_t > \theta)\) and \(p^1_c(t, \theta) = \mathbb{P}(h^1_t \leq \theta)\). First, we bound \(p^0_c(t, \theta)\) and then give a \(\theta\) such that these bounds are less than \(\delta\) until a given time \(t \leq t_0(\delta, r, N)\).

Step 1:

Assumption 2.8.1 gives

\[
\mathbb{P}(h_{-r+1,K}^0 \leq h_{-r+1,K}^1 \leq h_{-r+1,K}^1) \leq \text{BinMix}(K, g^*) = \pi_K.
\]

Then from Lemma 3.7 there are \((g^0_t)_{t \geq 0}\) and \((g^1_t)_{t \geq 0}\) in \(\mathcal{P}([0,1]^N)\) such that \(h_r = \text{BinMix}(K, g^0_t)\). Let \(u^0_t\) (resp. \(u^1_t\)) denotes the supremum (resp. infimum) of the support of \(g^0_t\) (resp. \(g^1_t\)).

We want to apply Lemma 3.10 with:

\[
\eta = f_N, a = f_N q^+, b = 1 - f_N q^+ - (1 - f_N)q_{10}, c = 1 - f_N q_{01}, u^0_t = u^1_t, u^M_t = u^0_t.
\]

We check the assumptions of the Lemma. First, thanks to assumption 2.8.2 there exists \(N_0 \in \mathbb{N}\) such that for all \(N \geq N_0 \geq N_0\)

\[
a = f_N q^+ \geq \frac{f_N q^+}{f_N q^+ + (1 - f_N)q_{10} - f_N^2 q_{01}} < q^+.
\]

Thus, we get \(\alpha = f_N q^+ < \frac{\alpha}{q^+}\) and \(\gamma = \frac{\alpha}{c}\) belong to \([0,1]\). Then, from equation 16, we obtain \(u^0_{-r+1} = u^*, u^1_{-r+1} = 0\) and by equation 14, we get \(u^M_t = u^0_t = u^1_t = (1 - q_{10})^r < u^* = \frac{\alpha}{c}\) and \(u^1_t = 1 - (1 - q^+)r > q^+ > \frac{\alpha}{c}\). Finally, from 15, \(G^0_t\) satisfies equation 18 for all \(t \geq 1\). Therefore, we can apply Lemma 3.10 to \(h^0_t\) and \(h^1_t\) for all \(N \geq N_0\):

\[
t = 1, \ u^0_t = u^* (1 - q_{01})^r \quad \text{and} \quad u^1_t = 1 - (1 - q^+)r, \quad \forall t \geq 1, \ u^0_t = (u^1_t - u^*)^r + u^*, \quad \forall t \geq 1, \ u^1_t = (u^1_t - u^*)^r + u^*, \quad (24)
\]

where

\[
t = \frac{\log\left(f_N q^+ \log(1 - (1 - f_N)q_{10} - f_N q^+) - \log(u^1_t - u^*)\right)}{\log(1 - (1 - f_N)q_{10} - f_N q^+)} = \frac{C}{\log(1 - (1 - f_N)q_{10} - f_N q^+)} > 0.
\]

Step 2:

From Step 1, we can bound the error by splitting it in two terms.

\[p^0_c(t, \theta) = \mathbb{P}(h^0_t > \theta) = \sum_{k=0}^{N} \mathbb{P}(K = k) \mathbb{P}(h^0_{t_k} > \theta) \leq \inf_{K_t} \left\{ \mathbb{P}(K \geq K^0_t) + \sum_{k=0}^{K^0_t - 1} \mathbb{P}(K = k) \mathbb{P}(h^0_{t_k} > \theta) \right\}, \]

14.
\[
\begin{align*}
p^t_c(t, \theta) &= \mathbb{P}(h^1_t \leq \theta) = \sum_{k=0}^{N} \mathbb{P}(K = k) \mathbb{P}(h^1_{t,k} \leq \theta) \\
&\leq \inf_{K^*} \left\{ \mathbb{P}(K \leq K^*_1) + \sum_{k=K^*_1+1}^{N} \mathbb{P}(K = k) \mathbb{P}(h^1_{t,k} \leq \theta) \right\}.
\end{align*}
\]

But \( h^y_{t,k} \leq \text{BinMix}(k, q^y_t) \) so that \( \mathbb{P}(h^y_{t,k} > \theta) \geq \mathbb{P}(h^y_{t,k-1} > \theta) \) and \( \mathbb{P}(h^y_{t,k} > \theta) \geq \mathbb{P}(h^y_{t,k+1} \leq \theta) \).

Hence,
\[
\begin{align*}
p^0_c(t, \theta) &\leq \inf_{K^*_1} \left\{ \mathbb{P}(K \geq K^*_1) + \mathbb{P}(h^0_{t,K^*_1-1} > \theta) \right\} \\
&\leq \inf_{K^*_1} \left\{ \mathbb{P}(K \geq K^*_1) + \int_0^{u^*_1} \mathbb{P}(\text{Bin}(K^*_1 - 1, z) > \theta) g^0_t(dz) \right\} \\
&\leq \inf_{K^*_1} \left\{ \mathbb{P}(K \geq K^*_1) + \mathbb{P}(\text{Bin}(K^*_1 - 1, u^*_1) > \theta) \right\}.
\end{align*}
\]

\[
\begin{align*}
p^1_c(t, \theta) &\leq \inf_{K^*_1} \left\{ \mathbb{P}(K \leq K^*_1) + \mathbb{P}(h^1_{t,K^*_1+1} \leq \theta) \right\} \\
&\leq \inf_{K^*_1} \left\{ \mathbb{P}(K \leq K^*_1) + \int_{u^*_1}^{1} \mathbb{P}(\text{Bin}(K^*_1 + 1, z) \leq \theta) g^1_t(dz) \right\} \\
&\leq \inf_{K^*_1} \left\{ \mathbb{P}(K \leq K^*_1) + \mathbb{P}(\text{Bin}(K^*_1 + 1, u^*_1) \leq \theta) \right\}.
\end{align*}
\]

Using Lemma 3.14 we bound each term of the previous inequality by \( \frac{\delta}{2} \) for all time \( t \leq t_c \). To do so, we first fix \( K_0, K_1 \in \mathbb{N} \) such that \( \mathbb{P}(K \geq K_0) \leq \frac{\delta}{2} \) and \( \mathbb{P}(K \leq K_1) \leq \frac{\delta}{2} \). As \( K \overset{\text{d}}{=} \text{Bin}(N, f_N) \), we find by (22) (resp. (23)) with \( n = N, p = f_N \) and \( \epsilon = \frac{K_0}{N f_N} - 1 \) (resp. \( \epsilon = 1 - \frac{K_1}{N f_N} \)):
\[
\mathbb{P}(K \geq K_0) \leq \exp \left( -\frac{(K_0 - N f_N)^2}{K_0 + N f_N} \right) \quad \text{and} \quad \mathbb{P}(K \leq K_1) \leq \exp \left( -\frac{(N f_N - K_1)^2}{2 N f_N} \right). \quad (25)
\]

Using the inequality
\[
\forall x > 0, y > 0, \quad \sqrt{x + y} \geq \sqrt{x} + \sqrt{y}, \quad (26)
\]
one can show that:
\[
K_0 = \lfloor N f_N + \bar{K} \rfloor \quad \text{and} \quad K_1 = \lfloor N f_N + \bar{K} \rfloor \quad \Rightarrow \quad \mathbb{P}(K \geq K_0) \leq \frac{\delta}{2} \quad \text{and} \quad \mathbb{P}(K \leq K_1) \leq \frac{\delta}{2}
\]
where \( \bar{K} = \sqrt{-2 \log \left( \frac{3}{4} \right) N f_N - \log \left( \frac{3}{4} \right)} \). Now that we have fixed \( K_0 \) and \( K_1 \), we study the existence of a threshold \( \theta \) such that
\[
\mathbb{P}\left( \text{Bin}(K_0 - 1, u^*_1) > \theta \right) \leq \frac{\delta}{2} \quad \text{and} \quad \mathbb{P}\left( \text{Bin}(K_1 + 1, u^*_1) < \theta \right) \leq \frac{\delta}{2}. \quad (27)
\]
We conclude using (30) and defining \( \hat{P} \) denote

\[ P(\text{Bin}(K_0 - 1, u_0^i) > \theta) \leq \exp \left( -\frac{(\theta - (K_0 - 1)u_0^i)^2}{(K_0 - 1)u_0^i + \theta} \right) \leq \exp \left( -\frac{(\theta - (N_f + \hat{K})u_1^i)^2}{(N_f + \hat{K})u_1^i + \theta} \right) \]

Therefore, while

\[ P(\text{Bin}(K_1 + 1, u_1^i) < \theta) \leq \exp \left( -\frac{(K_1 + 1)u_1^i - \theta)^2}{2(K_1 + 1)u_1^i} \right) \leq \exp \left( -\frac{(N_f - \hat{K})u_1^i - \theta)^2}{2(N_f - \hat{K})u_1^i} \right) \]

there exists \( \theta \) satisfying inequalities (27). Using that

\[ u_1^i + u_0^i \leq 1 \quad \text{and} \quad \sqrt{(N_f + \hat{K})u_0^i + (N_f - \hat{K})u_1^i} \leq \sqrt{2(u_1^i + u_0^i)N_f} \leq \sqrt{2N_f}, \]

one can show that in order to satisfy (28) the inequality

\[ \sqrt{-2 \log \left( \frac{\delta}{2} \right) N_f (1 + \sqrt{2}) - 2 \log \left( \frac{\delta}{2} \right) \leq N_f(u_1^i - u_0^i) \] (29)

is sufficient. We deduce, thanks to Step 1, two sufficient conditions on \( t \):

\[ t \leq t_c \quad \text{and} \quad b^{t-1} \left( N_f(u_1^i - u_0^i) \right) \geq (1 + \sqrt{2}) \sqrt{-2 \log \left( \frac{\delta}{2} \right) N_f - 2 \log \left( \frac{\delta}{2} \right) } \] (30)

Let denote \( N_1 \) such that for all \( N \geq N_1, \)

\[ 2 \exp \left( -\frac{N_f(u_1^i - u_0^i)}{(2 + \sqrt{2})\sqrt{N_f} + 2} \right) < \delta. \]

We conclude using (30) and defining \( N(\delta, r) = \max(N_0, N_1) \) such that for all \( N \geq N(\delta, r): \)

\[ \hat{t}(\delta, r, N) = \frac{\max \left( C, \log \left( \frac{(2 + \sqrt{2})\sqrt{-\log \left( \frac{\delta}{2} \right) N_f + 2 \log \left( \frac{\delta}{2} \right)} \} N_f(u_1^i - u_0^i) \right) \right)}{\log \left( 1 - f_Nq^+ - (1 - f_N)q_{10} \right) }, \]

with

\[ C = \log \left( \frac{f_N^2 q^+ q_{01}}{(f_N q^+ - (1 - f_N)q_{10})(f_N q^+ - (1 - f_N)q_{10} - f_N q_{01})(u_1^i - u^*)} \right) < 0. \]

\[ \square \]
Remark 3.16. \( \bullet \) Recall that \( u^* = \frac{f_Nq^+}{f_Nq^+ + (1 - f_N)q^-} \), \( u_1^* = u^*(1 - q^-) \) and \( u_1^* = 1 - (1 - q^+) \).

We proved that assuming \([2.8.1]\) and \([2.8.3]\) hold, then for all \( \delta \) and \( r \), there exists \( N(\delta, r) \) such that for all \( N \geq N(\delta, r) \):

\[
2\exp\left(-\frac{Nf_N(u_1 - u_1^*)}{(2 + \sqrt{2})\sqrt{Nf_N} + 2}\right) < \delta \quad \text{and} \quad \frac{f_Nq^+}{f_Nq^+ + (1 - f_N)q^-} < q^+, \quad (31)
\]

and then

\[
\hat{t}(\delta, r, N) = \max\left(C, \log\left(\frac{f_Nq^+ q_1^*}{f_Nq^+ + (1 - f_N)q^- q_1^* (u_1^* - u_1)}\right)\right),
\]

with

\[
C = \log\left(\frac{f_Nq^+ q_1^*}{(f_Nq^+ + (1 - f_N)q^-) (f_Nq^+ + (1 - f_N)q^- f_Nq^-)}\right) < 0.
\]

\( \bullet \) Moreover, assume that Assumption \([2.8.3]\) also holds. Then, \( \exists N_0, r_0 \in \mathbb{N} \) s.t. \( \forall N \geq N_0, r \geq r_0, (1 - f_N)q_{10} > f_Nq_{01}, \] \( 2\exp\left(-\frac{\sqrt{Nf_N}(u_1 - u_1^*)}{4}\right), 1 \right\} \neq \emptyset \)

and for all \( \delta \) in this interval, \( \exists c_0(N), c_1(N, r), c_2(N, r, \delta) \in \mathbb{R}_+^* \) s.t.

\[
t_*(\delta, r, N) \geq \frac{1}{c_0(N)} \min\left(\log\left(\frac{1}{c_1(N, r) f_N}\right), \log\left(\frac{\sqrt{Nf_N}}{c_2(N, r, \delta)}\right)\right),
\]

with

\[
c_0(N) \sim_N \infty f_N(q^+ + b_N),
\]

\[
c_1(N, r) \sim_N \infty \frac{q^+ a_N}{(q^+ + b_N)^2 (u_1^* - u_1)}
\]

\[
c_2(N, r, \delta) \sim_N \infty \frac{(2 + \sqrt{2})\sqrt{-\log(\delta)}}{(u_1^* - u_1)}.
\]

\( \bullet \) Note that we have also proved the following result: For every \( \delta > 0 \) and \( N \) large enough, there exists \( r_0 \) such that, if the initial signal is presented at least \( r_0 \) times, then it is well memorized after at least \( \hat{t}(\delta, r, N) \) presentations of noisy signals.

4 Simulations

We study how tight is our bound on \( t_* \). Indeed, we cannot rule out that the actual errors are much smaller than our estimates.

Our code follows these lines: we draw a signal size \( K \sim \text{Bin}(N, f_N) \), we simulate a trajectory of \( h_{t,K} \) long enough to be under the invariant measure. We present \( r \) times the signal to be learnt and then compute the trajectories of \( h_{t,K}^y \), \( y \in \{0, 1\} \). We reiterate this procedure \( N_{MC} = 10^7 \) times. This gives us an approximation of the distributions of \( h_{t,K}^y \) and \( h_{t,K}^1 \).

Our main result is interesting for large values of \( Nf_N \) combined with a small \( f_N \). Large \( Nf_N \) allows to have a satisfying interval for the error, and small \( f_N \) gives a non-negligible \( t \), see second
Unfortunately, this means large $N$ and simulations are not suited for too big $N$. Moreover, for this setting of parameters, the exact error is really small for short time. This means we need to compute many trajectories before getting the synaptic currents $h_y^{t}$ cross a reasonable threshold $\theta$. In this context, we use the following parameters:

$$\theta = 117, \quad N = 20,000, \quad f_N = 0.05, \quad q_{01} = 0.5, \quad q^+ = 0.5, \quad q_{10} = 0.05 \text{ and } r = 3.$$
too far from our theoretical result, around $10^{-6}$ \(\delta\). Moreover, in Proposition 3.15, the result is a maximum between two times. One of them does not depend on the error \(\delta\). This explains the plateau around \(\delta = 4.10^{-4}\) in Figure 3c. Indeed, for this set of parameters and \(\delta\) large enough, the time to consider changed of formula. Finally, we note in figure 3d that \(p_e(t, \theta)\) is above \(p_1(t, \theta)\) for short time. Then \(p_1(t, \theta)\) increases quickly until a value close to one whereas \(p_0(t, \theta)\) stays below \(10^{-2}\). This is because the majority of distribution of \(h_{t,K}\) stays on the good side of \(\theta\) for the corresponding error \(p_0(t, \theta)\). On the other hand, the distribution of \(h_{t,K}\) moves towards the wrong side of \(\theta\) for the corresponding error \(p_1(t, \theta)\). We now present the histograms of the distributions of the synaptic currents at certain.

\[
\text{Figure 4: Histograms of the distributions of } h_0 \text{ and } h_1 \text{ at different times. Distributions just after the learning phase and the invariant measure are plotted in 4c and distributions at time } t = 70 \text{ as well as the invariant measure are plotted in 4d.}
\]

We note that the invariant measure is concentrated around small values. This enables the post learning distribution of \(h_0\) to have a small variance as it is concentrated near 0, see figures 4a and 4c. However, the variance of this distribution increases quickly. It follows that the distribution of \(h_0\) has a multimodal shape with a high proportion staying near 0 for more than 50 presentations after learning. The distribution of \(h_1\) keeps a unimodal shape with a variance decreasing at the beginning and then increasing, see figure 4b. Distributions stays well separated.
approximately until time \( t = 70 \), see figure 4d.

In order to illustrate the role played by the parameter \( r \), we plot the distributions just after the learning phase for different values of \( r \) and for the following parameters:

\[
N_{MC} = 10^6, \quad N = 20\,000, \quad f_N = 0.1, \quad q_{01} = q_{10} = 0.01 \quad \text{and} \quad q^+ = 0.05.
\]

\[\text{(a)}\]

\[\text{(b)}\]

\[\text{(c)}\]

\[\text{(d)}\]

Figure 5: The distributions of \( h_1 \), just after learning, are plotted with the one of the invariant measure. Because of the parameters, the distributions of \( h_0 \) are really close to \( \pi_\infty \).

In this last set of parameters, the forgetting is really slow. However, if the signal to be learnt correctly with such a small \( q^+ \), then \( r \) has to be high enough. This shows the need of a large \( r \) in view of a slow forgetting.

5 Discussion

We provide a mathematical framework to study the memory retention of random signals by a recurrent neural network with binary neurons and binary synapses. We thus consider a paradigm linking synaptic plasticity and memory: a stimulus is remembered as long as its trace in the synaptic weights is strong enough. In order to measure the memory of a stimulus, we study the synaptic current on a neuron during the presentation of a stimulus. First, we computed
the spectrum of the transition matrix of the synaptic current which, to our knowledge, was not
done before. This enables us to conclude that the eigenvalues are strictly different whatever the
parameters are. We underline the interesting structure of the spectrum. Then, we completed the
previous work done in [3] on the invariant distribution and the synaptic current distributions.
This leads us to control the form of these distributions. The properties of these distributions
give enough information to find a lower bound on the time a neuron keeps a good estimate
on its response to the first stimulus. We measure the quality of this estimation performing a
statistical test based on the synaptic current onto a neuron. We define an error associated to
this test which depends on the two distributions knowing that the neuron was selective or not to
the initial signal. The time passed under a certain error is then lower bounded. In the large $N$
limit and small coding level asymptotic, this bound can be simplified and the minimum error is
obtained when $f_N$ is in the order of $\frac{1}{N^2}$. This put forward the need of having a coding level not
too sparse otherwise the learning is not good (high probability of error). Finally, unlike previous
studies, we take into account the possibility that heterosynaptic and homosynaptic depressions
can scale differently in $N$ and we consider the role of presenting several times a signal in the
learning phase.

Our study is valid for a classic learning, which needs multiple stimulus presentations, and for
a one shot learning. This last one is possible only with a specific choice of parameters. Indeed,
when presenting a stimulus, the synaptic weights between selective neurons need to be potentiated
with a high probability (high $q^+$). When presenting other stimuli, these same weights will have
a very small probability of undergoing depression (low $q_{01}$ and $q_{10}$). As a result, following the
presentation of a stimulus, selective neurons develop strong links and then these connections take
time to disappear. Thus, the experiment associated with this model would focus on recognition
memory. A well-known experiment in this field was carried out by Standing [16]. It shows that
humans are able to recognize 10,000 images, presented only once, with 90 percent success rate.

The advantage of getting a theoretical result is the possibility to better understand the role
played by the parameters. First, $r$ is the number of times the signal to learn is initially presented.
Its role is to separate distributions. Our result confirms this role. Indeed, this can be seen in
the proof of our result. We first proved the distributions of the synaptic currents follow binomial
mixtures. Then in the proof of our main result, we need $r$ to be large enough such that the
supports of the distributions associated to these binomial mixtures do not overlap. However, it
is not a key parameter to optimize the memory lifetime because it appears only in a logarithm.
The most important parameter for maximising this time is the coding level: denominator of our result [3.16] proportional to $f_N$.
However, the smaller $f_N$, the bigger the error: numerator in $\log(\sqrt{N f_N})$, see [3.16]. The same
reasoning is valid for the depression rates. Let refine our analysis in order to put forward the
role played by homosynaptic and heterosynaptic plasticity. In our main result, the important
variables are $u^* = \frac{q^+}{q^+ + (1 - f_N) b_N}, f_N(q^+ + b_N)$ and $\frac{(q^+ + b_N)^2(u_1 - u^*)}{q^+ a_N f_N}$. $u^*$ has to be as small as
possible, which imposes $b_N$ to be large enough compared to $q^+$. Note that $b_N$ plays a crucial role
in our analysis: without it, the supports would be much bigger. $f_N(q^+ + b_N)$ as well has to be
as small as possible, which prevent $b_N$ from being too far from the value of $q^+$. $\frac{(q^+ + b_N)^2(u_1 - u^*)}{q^+ a_N f_N}$
has to be as big as possible, so $a_N$ should not be too big. Finally, we wonder what are realistic
values for these parameters and what are their biological interpretation? One can think of the
system as a group of $N \sim 10^4$ neurons encoding an external signal stimulating a proportion of
these neurons, $f_N \sim 10^{-2}$ according to experimental values [15]. A presentation of a signal can
be seen as long enough in order to stabilise the neural response and start to change synaptic
weights in the view of learning. Potentiation and depression rates are arbitrary in our model as
no experiment enables to measure them so we do not know them. We pick them in such a way

21
that the system does not learn in one step neither in too many steps (order of unity).

We use the model presented in [1] because of its relative simplicity and its consideration of synapse correlations. The study [1] focuses on the first two moments of the synaptic current. It leads to a result on the memory capacity of the network which depends on a global variable, the so-called signal-to-noise ratio (SNR). Here, we study more in detail the stochastic processes associated to the synaptic currents and propose precise indications on how to retrieve signals. We then manage to lower bound the time it takes to forget this signal, which depends on the accepted error. Our result predicts a forgetting time proportional to $\frac{1}{f_N}$ which is less than the one obtained in [1] which is proportional to $\frac{1}{f_N^2}$. This difference comes from our different measure of memory lifetime. The SNR analysis gives information on this time based on the convergence of the means of synaptic currents whereas our retrieval criterion requests the knowledge of their entire distributions. In particular, our result adds a significant condition: the accepted error must be greater than a certain value that depends on the parameters. In particular, this error tends to 1 when $f_N$ tends to 0. For a fixed $N$, the minimum error is obtained when $f_N$ is in the order of $\frac{1}{N}$. Moreover, our result holds under precise conditions we detailed and do not necessarily need the large $N$ asymptotic. Nevertheless, our result is greatly simplified and gives larger memory lifetime in this asymptotic.

Many perspectives can be studied as a follow-up to this study. First, the main result is based on the analysis of the support of distributions associated to Binomial mixtures, $g_y^t$. This support converges with exponential speed proportional to $1 - f_N$ whereas the distributions of synaptic currents converge with exponential speed proportional to $1 - f_N^2$. Hence, as the article of Amit and Fusi [1] suggests, we could expect that the memory lifetime be proportional to $\frac{1}{f_N^2}$. Then, the analysis performed for the synaptic current onto a neuron could be extended to the entire vector of synaptic currents, the correlations between synaptic weights would then play a major role. In addition, the model could be extended in order to get closer to biology. Indeed, the formation of synaptic memory is more complex than in the model studied. In particular, the link between the dynamics of the neurons and the synaptic weight is missing. Improving the model in this direction could be done by considering more structured and complex external signals, adding neural layers and a more realistic membrane potential neural dynamics. In the literature, adding synaptic states does not seem to be successful as the authors stated in [9,11], whereas meta-plastic transitions brought better SNR results [4,10,14]. Adding neural dynamics in such models would be a next challenging step. Nevertheless, the model analysed here illustrates well the compromise / tension between the plastic and the stable characteristics of memory. Indeed, learning implies changes of synaptic weights (plasticity) as well as mechanisms which maintain them (stability). In mathematical terms, stability is related to the minimal convergence rate and plasticity refers to the sensibility to disturbance. Intuitively, we see that there is a compromise: the more a dynamics is sensitive to disturbances, the less it is stable and vice-versa. 
Appendix

A Proofs

A.1 Proof of Proposition 3.5

1. The map $\mathcal{R}$ is a contraction

Let $\Gamma_1, \Gamma_2 \in F_{[0,1]}$, as $f_N q^+ < 1 - (1 - f_N)q_{i_0}$

$$
\|\mathcal{R}(\Gamma_2) - \mathcal{R}(\Gamma_1)\|_{L^1(0,1)} \\
\leq f_N \int_0^1 \left| \Gamma_2 \left( \frac{u - f_N q^+}{1 - (1 - f_N)q_{i_0} - f_N q^+} \right) - \Gamma_1 \left( \frac{u - f_N q^+}{1 - (1 - f_N)q_{i_0} - f_N q^+} \right) \right| du \\
+ (1 - f_N) \int_0^1 \left| \Gamma_2 \left( \frac{u}{1 - f_N q_{i_0}} \right) - \Gamma_1 \left( \frac{u}{1 - f_N q_{i_0}} \right) \right| du \\
= f_N \int_{f_N q^+}^{1-(1-f_N)q_{i_0}} \left| \Gamma_2 \left( \frac{u - f_N q^+}{1 - (1 - f_N)q_{i_0} - f_N q^+} \right) - \Gamma_1 \left( \frac{u - f_N q^+}{1 - (1 - f_N)q_{i_0} - f_N q^+} \right) \right| du \\
+ (1 - f_N) \int_0^{1-(1-f_N)q_{i_0}} \left| \Gamma_2 \left( \frac{u}{1 - f_N q_{i_0}} \right) - \Gamma_1 \left( \frac{u}{1 - f_N q_{i_0}} \right) \right| du \\
= f_N (1 - (1 - f_N)q_{i_0} - f_N q^+) \int_0^1 |\Gamma_2(u) - \Gamma_1(u)| du \\
+ (1 - f_N)(1 - f_N q_{i_0}) \int_0^1 |\Gamma_2(u) - \Gamma_1(u)| du \\
= \left( f_N (1 - (1 - f_N)q_{i_0} - f_N q^+) + (1 - f_N)(1 - f_N q_{i_0}) \right) \|\Gamma_2 - \Gamma_1\|_{L^1(0,1)}. \\
$$

As $\lambda_1 < 1$, the map $\mathcal{R}$ acting on $F_{[0,1]}$ is strictly contracting in $L^1(0,1)$.

2. Existence of a fixed point

We now prove the second point of the Lemma. For all $\Gamma_0 \in F_{[0,1]}$, by contraction of $\mathcal{R}$, $(\mathcal{R}^n(\Gamma_0))_{n \geq 0}$ is a Cauchy sequence for the $L^1(0,1)$ norm. By completeness of $L^1(0,1)$, this sequence converges to some $\Gamma \in L^1(0,1)$. It remains to prove that $\Gamma$ can be chosen in $F_{[0,1]}$.

First, any limit $\Gamma$ is non decreasing almost everywhere. Define $G_*(x) = \lim_{y \to x^+} \Gamma(y)$. The function $G_*$ is càdlàg and satisfies for every $x \leq 0$, $G_*(x) = 0$ and for every $x \geq 1$, $G_*(x) = 1$. Thus $G_* \in F_{[0,1]}$ and $\mathcal{R}(G_*) = G_*$.

3. Uniqueness of the fixed point

Assume there exists another $\hat{G} \in F_{[0,1]}$ such that $\mathcal{R}(\hat{G}) = \hat{G}$, we have

$$
\|\hat{G} - G_*\|_{L^1(0,1)} = \|\mathcal{R}(\hat{G}) - \mathcal{R}(G_*)\|_{L^1(0,1)} < \|\hat{G} - G_*\|_{L^1(0,1)}. \\
$$

So, $\hat{G} = G_*$.

A.2 Proof of Lemma 3.7

Let $\tilde{U}, (U_i)_{1 \leq i \leq K}, (\xi_i)_{1 \leq i \leq K}, (\eta_i)_{1 \leq i \leq K}$ and $(W_i)_{1 \leq i \leq K}$ be i.i.d. random variables following the uniform law on $[0,1]$. By the first point of the remark 3.2, $Z$ is the sum of $(Z_i)_{1 \leq i \leq K}$ i.i.d. Bernoulli of parameter $Y_Z = G_Z^{-1}(\tilde{U})$. 

23
Thus we obtain that

\[ X = \sum_{i=1}^{K} Z_i \mathbb{1}_{\{\xi_i \leq a\}} + \sum_{i=1}^{K} (1 - Z_i) \mathbb{1}_{\{\eta_i \leq b\}} \]

\[ \overset{\text{Bin}(Z,a)}{\text{Bin}(Z,a)} \]

\[ \overset{\text{Bin}(K-Z,b)}{\text{Bin}(K-Z,b)} \]

where the Binomials are independent. Then, let consider \( \forall i, Z_i = \mathbb{1}_{\{U_i \leq G^{-1}_Z(U)\}} \). Thus,

\[ X = \sum_{i=1}^{K} \mathbb{1}_{\{U_i \leq G^{-1}_Z(U)\}} \mathbb{1}_{\{\xi_i \leq a\}} + \sum_{i=1}^{K} \mathbb{1}_{\{U_i > G^{-1}_Z(U)\}} \mathbb{1}_{\{\eta_i \leq b\}}. \]

So

\[ X = \sum_{i=1}^{K} \mathbb{1}_{\{U_i \leq G^{-1}_Z(U), \xi_i \leq a\}} \cup \{U_i > G^{-1}_Z(U), \eta_i \leq b\}. \] (32)

\[ \xi_i, \eta_i \]

\[ 1 \]

\[ 0 \]

\[ G^{-1}_Z(U) \]

\[ 1 \]

\[ U_i \]

\[ a \]

\[ b \]

\[ \xi_i, \eta_i \]

\[ 1 \]

\[ 0 \]

\[ G^{-1}_Z(U) \]

\[ 1 \]

\[ U_i \]

\[ a \]

\[ b \]

\[ \xi_i, \eta_i \]

\[ 1 \]

\[ 0 \]

\[ G^{-1}_Z(U) \]

\[ 1 \]

\[ U_i \]

\[ a \]

\[ b \]

\[ \xi_i, \eta_i \]

\[ 1 \]

\[ 0 \]

\[ G^{-1}_Z(U) \]

\[ 1 \]

\[ U_i \]

\[ a \]

\[ b \]

\[ \xi_i, \eta_i \]

\[ 1 \]

\[ 0 \]

\[ G^{-1}_Z(U) \]

\[ 1 \]

\[ U_i \]

\[ a \]

\[ b \]

\[ \xi_i, \eta_i \]

\[ 1 \]

\[ 0 \]

\[ G^{-1}_Z(U) \]

\[ 1 \]

\[ U_i \]

\[ a \]

\[ b \]

\[ \xi_i, \eta_i \]

\[ 1 \]

\[ 0 \]

\[ G^{-1}_Z(U) \]

\[ 1 \]

\[ U_i \]

\[ a \]

\[ b \]

\[ \xi_i, \eta_i \]

\[ 1 \]

\[ 0 \]

\[ G^{-1}_Z(U) \]

\[ 1 \]

\[ U_i \]

\[ a \]

\[ b \]

\[ \xi_i, \eta_i \]

\[ 1 \]

\[ 0 \]

\[ G^{-1}_Z(U) \]

\[ 1 \]

\[ U_i \]

\[ a \]

\[ b \]

\[ \xi_i, \eta_i \]

\[ 1 \]

\[ 0 \]

\[ G^{-1}_Z(U) \]

\[ 1 \]

\[ U_i \]

\[ a \]

\[ b \]

\[ \xi_i, \eta_i \]

\[ 1 \]

\[ 0 \]

\[ G^{-1}_Z(U) \]

\[ 1 \]

\[ U_i \]

\[ a \]

\[ b \]

\[ \xi_i, \eta_i \]

\[ 1 \]

\[ 0 \]

\[ G^{-1}_Z(U) \]

\[ 1 \]

\[ U_i \]

\[ a \]

\[ b \]

\[ \xi_i, \eta_i \]

\[ 1 \]

\[ 0 \]

\[ G^{-1}_Z(U) \]

\[ 1 \]

\[ U_i \]

\[ a \]

\[ b \]

\[ \xi_i, \eta_i \]

\[ 1 \]

\[ 0 \]

\[ G^{-1}_Z(U) \]

\[ 1 \]

\[ U_i \]

\[ a \]

\[ b \]

\[ \xi_i, \eta_i \]

\[ 1 \]

\[ 0 \]

\[ G^{-1}_Z(U) \]

\[ 1 \]

\[ U_i \]

\[ a \]

\[ b \]

\[ \xi_i, \eta_i \]

\[ 1 \]

\[ 0 \]

\[ G^{-1}_Z(U) \]

\[ 1 \]

\[ U_i \]

\[ a \]

\[ b \]

\[ \xi_i, \eta_i \]

\[ 1 \]
A.3 Proof of Lemmas 3.10 and 3.14

Proof of Lemma 3.10

Proof. If \( 0 < x < \frac{a}{1-b} \), then \( \frac{a}{1-b} > xb + a > x \), so from (18), \( \forall t \geq 1 \),

\[
u_{t+1}^{M} = \max (\nu_{t}^{M}c, \nu_{t}^{M}b + a) = \left( \nu_{1}^{M} - \frac{a}{1-b} \right) b^{t} + \frac{a}{1-b}.
\]

On the other hand, for all \( x > \frac{a}{c-b} \), then \( xc > xb + a \), so from (18), while \( u_{t}^{m} > \frac{a}{c-b} \) we have

\[
u_{t+1}^{m} = \max (u_{t}^{m}c, u_{t}^{m}b + a) = u_{t}^{m}b + a = \left( \nu_{1}^{m} - \frac{a}{1-b} \right) b^{t} + \frac{a}{1-b}.
\]

The last equation is true for all \( t \leq \lceil t_{c} \rceil + 1 \) such that \( t_{c} \geq 0 \) verifies

\[
\left( \nu_{1}^{m} - \frac{a}{1-b} \right) b^{t_{c}} + \frac{a}{c-b} = \frac{a}{c-b}.
\]

We conclude with, for all \( t \geq \lceil t_{c} \rceil + 2 \) then \( u_{t}^{m} \leq \frac{a}{c-b} \) and so

\[
u_{t+1}^{m} = \max (u_{t}^{m}c, u_{t}^{m}b + a) = u_{t}^{m}c = u_{\lceil t_{c} \rceil + 2}^{m} c^{\lceil t_{c} \rceil - 1}.
\]

\( \square \)

Proof of Lemma 3.14

Proof. We use the method of Chernoff [6]. Let \( X \) be the sum of \( X_{1}, X_{2}, \ldots, X_{n} \) which are independent Bernoulli random variables of parameter \( p \). \( \forall \epsilon \in [0,1], s \in \mathbb{R}^{+} \)

\[
P(X > np(1 + \epsilon)) \leq \mathbb{P}(e^{sX} \leq e^{np(1 + \epsilon)s}) \leq \frac{\mathbb{E}(e^{sX})}{e^{np(1 + \epsilon)s}} \leq \prod_{i=1}^{n} \frac{\mathbb{E}(e^{sX})}{e^{np(1 + \epsilon)s}} = \frac{(1 + p(e^{s} - 1))^{n}}{e^{np(1 + \epsilon)s}}.
\]

The minimum of the last term is reached for \( s = \log(1 + \delta) \) so

\[
P(X > np(1 + \epsilon)) \leq \left( \frac{e^{\delta}}{(1 + \epsilon)^{1+\epsilon}} \right)^{np} = \exp(np(\epsilon - (1 + \epsilon) \log(1 + \epsilon))).
\]

From the inequality, \( \forall x > 0, \log(1 + x) \geq \frac{x}{1+x} \), we obtain (22). In order to show (23), we proceed with the same method and use the inequality \( \log(1 + x) \geq \frac{2x}{1+x} \) whenever \(-1 < x \leq 0\). \( \square \)

Acknowledgements

This research has received funding from the European Union’s Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 785907 (Human Brain Project SGA2).
References

[1] D. J. Amit and S. Fusi. Learning in Neural Networks with Material Synapses. *Neural Computation*, 6(5):957–982, Sept. 1994.

[2] D. J. Amit and G. Mongillo. Spike-Driven Synaptic Dynamics Generating Working Memory States. *Neural Computation*, 15(3):565–596, Mar. 2003.

[3] Y. Amit and Y. Huang. Precise capacity analysis in binary networks with multiple coding level inputs. *Neural computation*, 22(3):660–688, 2010.

[4] M. K. Benna and S. Fusi. Computational principles of synaptic memory consolidation. *Nature Neuroscience*, 19(12):1697–1706, Oct. 2016.

[5] N. Brunel, F. Carusi, and S. Fusi. Slow stochastic hebbian learning of classes of stimuli in a recurrent neural network. *Network: Computation in Neural Systems*, 9(1):123–152, 1998.

[6] H. Chernoff. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. *The Annals of Mathematical Statistics*, 23(4):493–507, Dec. 1952.

[7] A. M. Dubreuil, Y. Amit, and N. Brunel. Memory capacity of networks with stochastic binary synapses. *PLoS computational biology*, 10(8):e1003727, 2014.

[8] T. Elliott. Memory Nearly on a Spring: A Mean First Passage Time Approach to Memory Lifetimes. *Neural Computation*, 26(9):1873–1923, Sept. 2014.

[9] S. Fusi and L. F. Abbott. Limits on the memory storage capacity of bounded synapses. *Nature Neuroscience*, Mar. 2007.

[10] S. Fusi, P. J. Drew, and L. Abbott. Cascade Models of Synaptically Stored Memories. *Neuron*, 45(4):599–611, Feb. 2005.

[11] Y. Huang and Y. Amit. Capacity analysis in multi-state synaptic models: a retrieval probability perspective. *Journal of Computational Neuroscience*, 30(3):699–720, June 2011.

[12] A. Miller. *Neural network models for Recognition Memory*. PhD thesis, Hebrew University of Jerusalem, 2012.

[13] S. Romani, D. J. Amit, and Y. Amit. Optimizing one-shot learning with binary synapses. *Neural computation*, 20(8):1928–1950, 2008.

[14] A. Roxin and S. Fusi. Efficient Partitioning of Memory Systems and Its Importance for Memory Consolidation. *PLoS Computational Biology*, 9(7):e1003146, July 2013.

[15] A. Spanne and H. Jörntell. Questioning the role of sparse coding in the brain. *Trends in Neurosciences*, 38(7):417–427, July 2015.

[16] L. Standing. Learning 10000 pictures. *Quarterly Journal of Experimental Psychology*, 25(2):207–222, May 1973.

[17] F. Zenke. *Memory formation and recall in recurrent spiking neural networks*. PhD thesis, Ecole Polytechnique Federale de Lausanne, 2014.