HAL open science

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

Pascal Helson

To cite this version:

Pascal Helson. A Mathematical Analysis of Memory Lifetime in a simple Network Model of Memory. 2020. hal-02308900v3

HAL Id: hal-02308900

https://hal.science/hal-02308900v3

Preprint submitted on 8 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Pascal Helson*

June 8, 2020

Contents

1 Introduction **2**

2 The model and the estimator **4**
2.1 The neural network and the protocol 4
2.2 Presentation of the estimator 5

3 Results **10**
3.1 Binomial mixture 10
3.2 Main results 12

4 Simulations **20**

5 Discussion **22**

A Proofs **26**
A.1 Proof of Proposition 3.4 26
A.2 Proof of Proposition 3.6 28
A.3 Proof of Lemma 3.14 29

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 noise. The presentation of an external stimulus to the recurrent network of binary neurons may change the state of the synapses. Multiple presentations of a unique signal leads to its learning. Then, during the forgetting time, the presentation of other signals (noise) may also modify the synaptic weights. We construct an estimator of the initial signal using the synaptic currents and define by this way a probability of error. In our model, these synaptic currents evolve as Markov

*pascal.helson@inria.fr

chains. We study the dynamics of these Markov chains and obtain a lower bound on the number of external stimuli that the network can receive before the initial signal is considered as forgotten (probability of error above a given threshold). Our results hold for finite size networks as well as in the large size asymptotic. Our results are based on a finite time analysis rather than large time asymptotic. We finally present numerical illustrations of our results.

1 Introduction

[Amit and Fusi, 1994] proposed a model to study the memory capacity of neural networks. The main novelty of their work was the online learning and forgetting of a sequence of random signals. Indeed, in previous models (e.g. [Willshaw et al., 1969] or [Hopfield, 1982]), signals are stored in a fixed weight matrix. This matrix is determined as a function of signals to learn. These models are called associative or attractor neural network (ANN) models: a stimulus is said to be stored if its neural representation is an attractor of the neural dynamics. The maximum storage capacity of ANN models have been widely studied. [Gardner and Derrida, 1988] computed this capacity for the optimal synaptic weight matrix. They showed that maximal storage is obtained for sparse coding. Moreover, there has been study of the robustness to noise in the synaptic weight matrix and in the initial input. [Sommer and Dayan, 1998] proposed Bayesian retrieval processes for a stochastic version of the Willshaw model. However, beyond the maximum number of stimuli learnt, blackout catastrophe (forgetting of all memories) appears in ANN models. This blackout can be avoided by allowing the plasticity of the synapses.

[Amit and Fusi, 1994] proposed the following experimental protocol: 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, Amit and Fusi performed a signal-to-noise ratio (SNR) analysis. The signal under consideration is the sum of the synaptic currents onto one neuron when the network receives the priming. As [Gardner and Derrida, 1988] found in the case of the ANN models, Amit and Fusi concluded that the coding of the stimuli needs to be sparse in order to optimise the storage capacity. They proposed a scaling of the coding level f (probability that a neuron is selective to a signal) as a function of the size N of the network. According to their retrieval criterion, the optimal coding level is on the order of $f \sim \frac{\log(N)}{N}$. In the large N asymptotic, what they called the storage capacity is then on the order of $\frac{1}{f^2}$ for depression probabilities proportional to f.

Extensions and approaches different from SNR have then been studied. First, [Brunel et al., 1998] studied a different protocol: they fixed the number of random stimuli and presented 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 of synaptic weights. ICP is the mean of synaptic weights among synapses involved in the learning of a stimulus. Intuitively, when ICP is much larger than MP,

the trace of a 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 on the order $\frac{1}{f}$ (resp. $\frac{1}{f^2}$). [Dubreuil et al., 2014] did a deeper analysis of the multiple presentations model of [Brunel et al., 1998] and the one shot learning model of [Amit and Fusi, 1994], under the assumption N large and f small. Then, [Elliott, 2014] considered the mean number of signals presented before the synaptic current crosses a fixed threshold: the mean first passage time (MFPT). More complex and biologically plausible models have been proposed and analysed numerically in the following studies: [Amit and Mongillo, 2003, Miller, 2012, Zenke, 2014]. Finally, to the best of our knowledge, the first article to present a precise way to retrieve stimuli is the one of [Amit and Huang, 2010]. In this article, they 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 last article, we study here a statistical test based on the synaptic currents. In particular, we study the probability of error associated to this test. Such an error has been studied before under some additive assumptions on the distribution of synaptic currents. [Amit and Huang, 2010] did a numerical analysis with a Gaussian approximation. [Dubreuil et al., 2014] gave an analytical result on the probability of no error in the large N asymptotic, assuming independence of synapses (which leads the synaptic currents to follow Binomial distributions). Here, we perform an analytical study of this error without such approximations and we manage to control it by extending previous analytical studies of [Amit and Fusi, 1994, Amit and Huang, 2010] on some points. First, we give properties of the synaptic current process such as the spectrum of its transition matrix (Propositions 3.9 and 3.10). Moreover, we study the case of multiple presentations of the signal to be learnt. Finally, we give in Remark 3.17 and Theorem 3.18 explicit bounds on the time during which a given signal is kept in memory (probability of error below a given threshold). These results deal with a broader range of depression probabilities than in the previous studies. We summarize our asymptotic results in Remark 3.19.

The rest of the paper is organised as follows. We expose the model and the statistical test in Section 2. After learning one specific signal, the network is submitted to random signals responsible for its forgetting. The statistical test consists in estimating the initial signal from the pre-synaptic inputs caused by priming (using a threshold estimator). We measure whether the signal is still in memory by computing the error associated to this test. After the formal definition of this error, the main results are presented. Then, Section 3 is devoted to deriving a lower bound on the maximum number of stimuli one can present while reasonably remembering the initial signal. This derivation relies on the fact that, asymptotically, as time goes to infinity, synaptic currents converge in law to a Binomial mixture (Corollary 3.7). We assume that, before learning, the synaptic currents follow their stationary distributions. Afterwards, the learning phase splits the network in two groups: the neurons activated by the signal and the others. Then, during the forgetting phase, the laws of the synaptic currents of these two groups are shown to remain Binomial mixtures with an explicit dynamics on their mixing distributions (Proposition 3.4). In the second part, we evaluate the probability of error of the test and the maximum number of stimuli one can present before the

test fails (Remark 3.17 and Theorem 3.18). The computations are based on estimates on the support and on the tail of the mixing distributions. 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 model and the protocol followed for learning and forgetting. Then, we define the estimator, derive the equations describing the dynamics of the synaptic currents and detail the main 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 suggest the reader to see the model as describing an experiment on a person's ability to learn a stimulus. In particular, we ask for how long a learnt signal can persist in memory when the person is presented some other signals which we termed loosely as noise.

Let us assume that we present a sequence of external stimuli to a network of $N+1$ neurons. Thus, we sum over N external synaptic currents to get the total synaptic current onto one neuron. We do not study the dynamics of the membrane potential nor the firing rate of neurons, but rather we consider their neural activities, $\xi \in \{0,1\}^{N+1}$. Hence, the neurons do not have their own dynamics but instead they follow the dynamics of the signals. We say that the neuron i is selective (resp. not selective) to a signal if its neural response is $\xi^i = 1$ (resp. $\xi^i = 0$). We assume that a given signal uniquely determines the neural response. Therefore, we refer in an equivalent way to stimulus/signal or neural response in the following. Signals are assumed to be random and we denote by $(\cdots, \xi_{-1}, \xi_0, \xi_1 \cdots)$ the corresponding sequence. We call t the time at which the tth signal after ξ_0 is shown. We assume that the ξ_ts are independent and identically distributed (*i.i.d.*) random variables (*r.v.*) in $\{0,1\}^{N+1}$. Moreover, for each t, the components $\xi_t^1, \cdots, \xi_t^{N+1}$ of ξ_t are themselves *i.i.d.* with Bernoulli distribution with parameter f:

$$\forall i \in [\![1, N+1]\!], \qquad \mathbb{P}\left(\xi_t^i = 1\right) = f = 1 - \mathbb{P}\left(\xi_t^i = 0\right).$$

The synaptic weight from neuron j to neuron i at time t is denoted by J_t^{ij}. It can only take two values $J_- < J_+$ and we denote by $J_t = \{J_t^{ij}, i \neq j\} \in \{J_-, J_+\}^{N(N+1)}$ 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 ξ_t. The corresponding transition probabilities are

- $\mathbb{P}\left(J_{t+1}^{ij} = J_+ | J_t^{ij} = J_-, (\xi_t^i, \xi_t^j) = (1,1)\right) = q^+,$
- $\mathbb{P}\left(J_{t+1}^{ij} = J_- | J_t^{ij} = J_+, (\xi_t^i, \xi_t^j) = (0,1)\right) = q_{01}^-,$
- $\mathbb{P}\left(J_{t+1}^{ij} = J_- | J_t^{ij} = J_+, (\xi_t^i, \xi_t^j) = (1,0)\right) = q_{10}^-.$

The transition probabilities not mentioned here and involving the change of state of a synaptic weight are set to zero. For example, $\mathbb{P}\left(J_{t+1}^{ij} = J_- | J_t^{ij} = J_+, (\xi_t^i, \xi_t^j) = (0,0)\right) = 0$. In order to simplify the notation and without loss of generality, we set:

$J_- = 0$ (weak synapse) and $J_+ = 1$ (strong synapse).

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

$$f, q_{01}^-, q_{10}^-, q^+ \in (0,1]. \tag{1}$$

The parameters q_{01}^- and q_{10}^- represent respectively the homosynaptic and heterosynaptic depressions, see [Brunel et al., 1998].

We now give the protocol to learn and then forget a signal. We denote by ξ_0 the signal to be learnt. 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 2.5 that there is a unique invariant measure). The learning phase consists in performing r presentations of ξ_0. In order to be consistent with the previous description, the sequence of presented stimuli is then $(\cdots, \xi_{-r}, \underbrace{\xi_0, \cdots, \xi_0}_{r \text{ times}}, \xi_1, \xi_2 \cdots)$ that is $\xi_t = \xi_0$ for $t \in [\![-r+1, 0]\!]$. The presentation of the subsequent signals leads to the forgetting of ξ_0.

2.2 Presentation of the estimator

We study the consistency through time of the response of one neuron to the initial signal. To do so, we consider the previous protocol. After the repetitive presentation of ξ_0, the signal has left a certain footprint in the matrix J_1. This trace is subsequently erased by the presentation of the following signals. How much information from a stimulus learnt is left at time t? As an answer, we define a probability of error. This error is associated to a decision rule based on the projection of J_t on ξ_0. For the neuron i, such a projection at time t is given by $\sum_{j \neq i} J_t^{ij} \xi_0^j$. In this framework, neurons are similar. Hence, in order to simplify the notation and without loss of generality, our study focuses on neuron $i = 1$. We denote by h_t,

$$h_t = \sum_{j=2}^{N+1} J_t^{1j} \xi_0^j, \tag{2}$$

the synaptic current onto neuron 1 when presenting again ξ_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 the process $(h_t)_{t \geq 0}$ strongly depends on the initial number K of active neurons

$$K = \sum_{j=2}^{N+1} \xi_0^j. \tag{3}$$

We denote by $h_{t,K}$ the process equal in law to h_t knowing K: $h_{t,K} \overset{\mathcal{L}}{=} (h_t | K)$. The process $h_{t,K}$ is Markovian, see Proposition 2.2.

We define a threshold estimator $\hat{\xi} : \mathbb{N}^* \times [\![0, N]\!] \to \{0, 1\}$ such that $\hat{\xi}(t, \theta) = \mathbb{1}_{h_t > \theta}$ with associated probability of errors:

$$p_e^0(t, \theta) = \mathbb{P}\left(\hat{\xi}(t, \theta) = 1 \mid \xi_0^1 = 0\right) = \mathbb{P}\left(h_t > \theta \mid \xi_0^1 = 0\right),$$
$$p_e^1(t, \theta) = \mathbb{P}\left(\hat{\xi}(t, \theta) = 0 \mid \xi_0^1 = 1\right) = \mathbb{P}\left(h_t \leq \theta \mid \xi_0^1 = 1\right).$$

Notation 2.1. *We denote by* $h_t^y \overset{\mathcal{L}}{=} \left(h_t \mid \xi_0^1 = y\right)$ *and* $h_{t,K}^y \overset{\mathcal{L}}{=} \left(h_t | \xi_0^1 = y, K\right)$.

In the following, we shall use the plural "distributions of $h_{t,K}^y$" to say distributions of $h_{t,K}^0$ and $h_{t,K}^1$. The probability of error $p_e^0(t, \theta) = \mathbb{P}\left(h_t^0 > \theta\right)$ (resp. $p_e^1(t, \theta) = \mathbb{P}\left(h_t^1 \leq \theta\right)$) 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. We aim at evaluating these errors: for fixed $\delta \in (0, 1)$, we estimate the largest time t_* such that both p_e^0 and p_e^1 are smaller than δ up to time t_*,

$$t_*(\delta, r, N) := \max_{\theta \in [\![0, N]\!]} \left(\inf\left\{t \geq 1,\, p_e^0(t, \theta) \vee p_e^1(t, \theta) \geq \delta\right\}\right), \tag{4}$$

where $x \vee y = \max(x, y)$ and $x \wedge y = \min(x, y)$.

Main Results (informal)

For any fixed error $\delta \in (0, 1)$, there is an unbounded set of couples $(N, r) \in \mathbb{N}^{*2}$ for which we show the existence of a threshold $\theta_{\delta,r,N} \in \{0, 1, \ldots, N\}$ ensuring

$$t_*(\delta, r, N) \geq \inf\left\{t \geq 1,\, p_e^0(t, \theta_{\delta,r,N}) \vee p_e^1(t, \theta_{\delta,r,N}) \geq \delta\right\} \geq \hat{t}(\delta, r, N),$$

where an explicit formula of $\hat{t}$ is given in Remark 3.17 for fixed potentiation and depression probabilities. Another formula of $\hat{t}$ is given in Theorem 3.18 for depression probabilities depending on N. In particular, assuming that the depression probabilities are proportional to the coding level f, we obtain that $\hat{t}(\delta, r, N)$ is on the order of $\frac{1}{f^2}$. The proofs of these results rely on the study of the Markov chains $(h_{t,K})_{t \geq 1}$ and $\left(h_{t,K}^y\right)_{t \geq 1}$.

Proposition 2.2. *The chains $(h_{t,K})_{t \geq 1}$ and $\left(h_{t,K}^y\right)_{t \geq 1}$ are Markovian. At the end of the learning phase, we have*

$$\begin{aligned} h_{1,K} \overset{\mathcal{L}}{=} h_{-r+1,K} &+ \xi_0^1 \mathrm{Bin}\left(K - h_{-r+1,K}, 1 - (1 - q^+)^r\right) \\ &- (1 - \xi_0^1) \mathrm{Bin}\left(h_{-r+1,K}, 1 - (1 - q_{01}^-)^r\right) \end{aligned} \tag{5}$$

where, conditionally on $h_{-r+1,K}$, the two Binomial random variables are independent. And during the forgetting phase, for all $t \geq 1$:

$$\begin{aligned} h_{t+1,K} \overset{\mathcal{L}}{=} h_{t,K} + \xi_t^1 &\left[\mathrm{Bin}\left(K - h_{t,K}, f q^+\right) - \mathrm{Bin}\left(h_{t,K}, (1 - f) q_{10}^-\right)\right] \\ &- (1 - \xi_t^1) \mathrm{Bin}\left(h_{t,K}, f q_{01}^-\right) \end{aligned} \tag{6}$$

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

The Markov chains $\left(h^y_{t,K}\right)_{t\geq 1}$ for $y \in \{0,1\}$, satisfy the equation (5) *with $\xi^1_0 = y$ and the equation* (6).

Proof. In order to study the jump from $h_{t,K}$ to $h_{t+1,K}$, we count the synapses that potentiate and the ones that depress upon presenting a signal ξ_t. From the definitions (2) and (3) of h_t and $h_{t,K}$, we only need to consider the K synapses J^{1j}_t with $j \geq 2$ such that $\xi^j_0 = 1$. At time t, there are $h_{t,K}$ strong synapses and $K - h_{t,K}$ weak synapses. Given ξ^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 $1-(1-q^-_{01})^r$ to depress. If $\xi^1_0 = 1$, every weak synapse is r times candidate to potentiation so it has probability $1-(1-q^+)^r$ to potentiate. Equation (5) follows.

From time $t \geq 1$ to $t+1$, if $\xi^1_t = 0$, the probability that a strong synapse depresses is fq^-_{01}. If $\xi^1_t = 1$, the probability that a weak synapse potentiates is fq^+ and the probability that a strong synapse depresses is $(1-f)q^-_{10}$. Equation (6) follows.

By definition of $h^y_{t,K}$, the chain satisfies the equations (5) with $\xi^1_0 = y$ and (6). □

Corollary 2.3. *Assume that* (1) *holds. Then, for all $K \in [\![0,N]\!]$, the Markov chain $(h_{t,K})_{t\geq 1}$ admits a unique invariant measure π_K with support in $[\![0,K]\!]$. Moreover, for any initial condition $h_{0,K}$, the Markov chain $(h_{t,K})_{t\geq 1}$ converges in law to π_K. In addition, the chain $(h_t)_{t\geq 1}$ converges in law to $\pi_\infty = \sum_{K=0}^N \mathbb{P}\left(\hat{K} = K\right)\pi_K$ where $\hat{K} = \sum_{j=2}^{N+1} \xi^j_0$.*

Proof. By (1), the Markov chain $(h_{t,K})_{t\geq 1}$ is irreducible and aperiodic on a finite state space. Thus, it admits a unique invariant measure towards which it converges.

Let $\hat{K} = \sum_{j=2}^{N+1} \xi^j_0$. From the Bayes' formula we get that for all $l \in [\![0,N]\!]$,

$$\lim_{t\to\infty} \mathbb{P}\left(h_t = l\right) = \lim_{t\to\infty}\left(\sum_{K=0}^N \mathbb{P}\left(\hat{K} = K\right)\mathbb{P}\left(h_{t,K} = l\right)\right) = \sum_{K=0}^N \mathbb{P}\left(\hat{K} = K\right)\pi_K(l).$$

□

Remark 2.4. *The Markov chains $\left(h^y_{t,K}\right)_{t\geq 1}$ have the same transition matrix as $(h_{t,K})_{t\geq 1}$. They differ by their distribution at time $t = 1$. Hence, they all converge to π_K. Moreover, both $\left(h^0_t\right)_{t\geq 1}$ and $\left(h^1_t\right)_{t\geq 1}$ converge in law to π_∞.*

Proposition 2.5. *Under the assumption* (1)*, the process $(\xi_t, J_t)_{t\geq 1}$ converges to its unique invariant measure. We denote it by ρ_∞.*

Proof. Same argument as for Corollary 2.3. □

We now give the main assumptions.

Assumption 2.6.

2.6.1 $(\xi_0, J_{-r+1}) \overset{\mathcal{L}}{=} \rho_\infty$ *and in particular* $h_{-r+1,K}, h^0_{-r+1,K}, h^1_{-r+1,K} \overset{\mathcal{L}}{=} \pi_K$.

2.6.2 Assume that f depends on N. Let us denote it by f_N such that $\lim_{N\infty} f_N = 0$ and $\lim_{N\infty} N f_N = +\infty$.

2.6.3 Let $q^-_{01,N} = a_N f_N$ and $q^-_{10,N} = b_N f_N$ with $a_N, b_N : \mathbb{N}^* \to \mathbb{R}$ such that $a_N,\ b_N$ both converge in $[0, +\infty)$. However, we assume that at least one of the two limits is not 0 and

$$\lim_{N\infty} q^-_{01,N} = \lim_{N\infty} q^-_{10,N} = \lim_{N\infty} \frac{b_N^2}{N f_N a_N} = \lim_{N\infty} \frac{b_N}{N f_N} = 0, \quad \lim_{N\infty} N f_N a_N = +\infty.$$

We consider a general paradigm in which before receiving the stimulus ξ_0, many stimuli have already been sent $(\cdots, \xi_{-r-2}, \xi_{-r-1}, \cdots)$. We assume that the process $(\xi_t, J_t)_{t \le -r+1}$ has reached its invariant measure at time $t = -r + 1$ by Assumption 2.6.1. Then, one key parameter is the coding level f. We assume that it depends on N in the analysis of the large N asymptotic: Assumption 2.6.2. This assumption refers to sparse coding as f_N tends to 0. An additional constraint put forward is that the mean number of selective neurons, $N f_N$, needs to be large enough: Assumption 2.6.2. In this context, we are interested to see how the dependence on N of the depression probabilities can affect the memory lifetime, see Assumption 2.6.3. This assumption gives conditions on the large N asymptotic behaviours of the depression probabilities.

First illustrations

In this subsection, we illustrate the dynamics of $\left(h^y_{t,K}\right)_{t\ge 0}$ and $(h_{t,K})_{t\ge 0}$. In particular, we are interested in the effects of the coding level f_N on these synaptic currents. Let us assume that the signal ξ_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 $k \le x < k+1$. Let us have a look at the expected size of jumps of $h_{t,K}$ from the formulas (5), (6).

$$\begin{aligned}
\text{For } t = 1, \quad & \mathbb{E}\left[h_{1,K} - h_{-r+1,K} | h_{-r+1,K}, \xi^1_0 = 0\right] = -h_{-r+1,K}(1 - (1 - q^-_{01})^r), \\
& \mathbb{E}\left[h_{1,K} - h_{-r+1,K} | h_{-r+1,K}, \xi^1_0 = 1\right] = (K - h_{-r+1,K})(1 - (1 - q^+)^r), \\
\forall t > 1, \quad & \mathbb{E}\left[h_{t+1,K} - h_{t,K} | h_{t,K}\right] = (K - h_{t,K}) f_N^2 q^+ - h_{t,K} f_N (1 - f_N)(q^-_{10} + q^-_{01}).
\end{aligned}$$

From these equations, we note that the average jump size strongly depends on f_N. When f_N is close to 1, the reception of ξ_0 has a large impact on the weight matrix, easy to detect. However, the following average jump size are close to the initial one. Thus, 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. Conversely, when f_N is close to 0, the average jump size is significantly different between the learning (relatively big jumps) and the forgetting (relatively small jumps) phases. As a consequence, the convergence to the stationary distribution, and thus forgetting, is slower. However, the learning still occurs: the initial jump is still big. In order to illustrate these phenomena, we plot simulation results obtained with a high coding level, $f_N = 0.8$ in Figure 1, and a low coding level, $f_N = 0.1$ in Figure 2.

Figure 1a shows that the size of jumps is effectively big for $f_N = 0.8$, just after learning as well as during forgetting time. Figure 1b illustrates the separation between

the initial distributions of $h^0_{t,K}$ and $h^1_{t,K}$. Indeed, at time $t = -r + 1 = 0$, both $h^0_{0,K}$ and $h^1_{0,K}$ follow the invariant measure plotted in black. Then, after the reception of ξ_0, the distribution of $h^0_{1,K}$ is shifted to the left and the distribution of $h^1_{1,K}$ to the right. Initially, the signal is learnt because the distributions are well separated, see Figure 1b. Figures 1c and 1d exhibit the fast overlapping of these two distributions. Indeed, following the learning phase, the reception of new stimuli makes the two distributions converge back quickly to the invariant distribution. At time $t = 5$, the signal is already forgotten. Figure 2 illustrates the advantages of a low coding level. Indeed, even at time

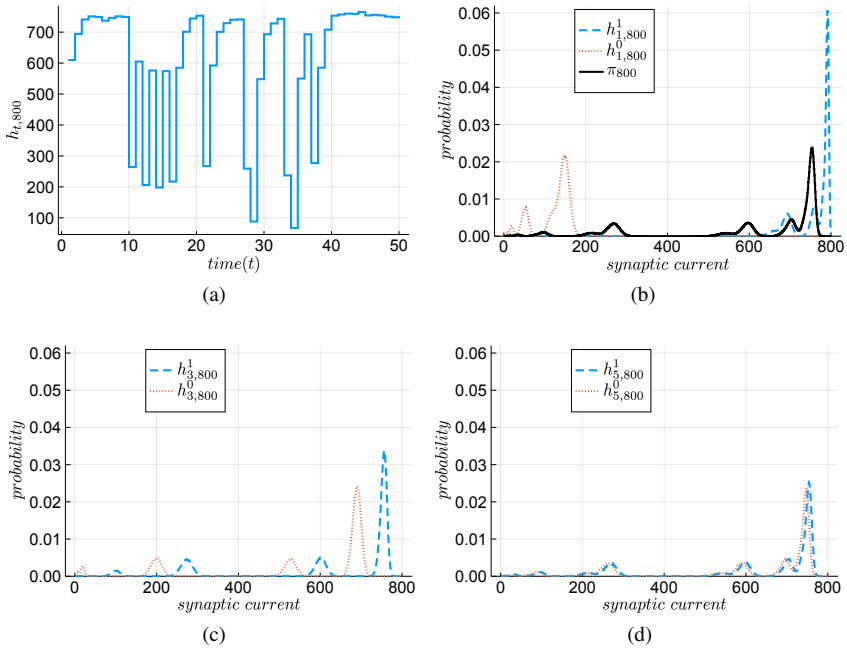
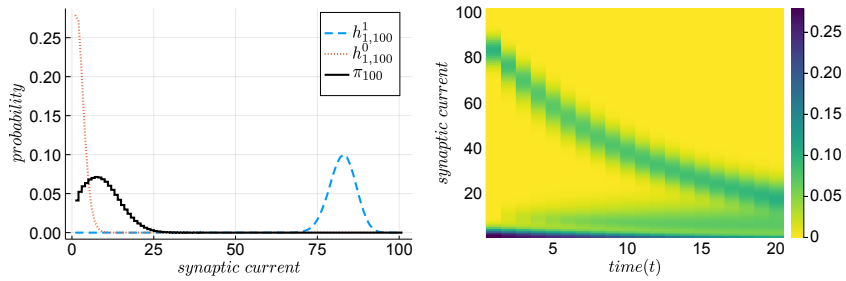


Figure 1: (1a) A typical trajectory of $h_{t,800}$. (1b) The distributions of $h^y_{1,800}$ and the invariant measure π_{800}. (1c),(1d) The distributions of $h^y_{t,800}$ at time $t = 3$ and $t = 5$. Parameters: $r = 1$, $N = 1000$, $K = 800$, $f_N = 0.8$, $q^+ = 0.8$, $q^-_{01} = 0.8$ and $q^-_{10} = 0.2$.

$t = 20$, the two distributions do not overlap a lot and they remained uni-modal. This makes the choice of a threshold estimator reasonable. Moreover, such an estimator allows a tractable analysis.



(a) (b)

Figure 2: (2a) The distributions of $h^y_{t,K}$ at time $t = 1$ and the invariant measure π_{100}. (2b) The sum of the two distributions $h^0_{t,K}$ and $h^1_{t,K}$ for $t \in [1, 20]$. The colour bar gives the probability values. Parameters: $r = 1$, $N = 1000$, $K = 100$, $f_N = 0.1$, $q^+ = 0.8$, $q^-_{01} = 0.8$, and $q^-_{10} = 0.2$.

3 Results

In this section, we first give some properties satisfied by the distributions of $h^y_{t,K}$, see Notation 2.1, and the invariant measure π_K. They enable us to prove Theorem 3.15, and then our main results, in the second part.

3.1 Binomial mixture

We denote by $F_{[0,1]}$ the set of cumulative distribution functions associated to the set $\mathcal{P}([0,1])$ 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 by* $\mathrm{BinMix}(K, g)$*, if*

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

Remark 3.2.

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

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

  We use both notations $X \overset{\mathcal{L}}{=} \mathrm{BinMix}(K, g)$ and $X \overset{\mathcal{L}}{=} \mathrm{BinMix}(K, Y)$.

- *The law of X is fully characterized by the moments $\mathbb{E}(Y), \mathbb{E}(Y^2), \cdots, \mathbb{E}(Y^K)$. Hence, if $\tilde{g} \in \mathcal{P}([0,1])$ is such that $\forall k \in [\![0, K]\!]$, $\int_0^1 u^k \tilde{g}(du) = \int_0^1 u^k g(du)$, then $\mathrm{BinMix}(K, \tilde{g}) \overset{\mathcal{L}}{=} \mathrm{BinMix}(K, g)$.*

First, we show that the set of Binomial mixtures is stable by the Markov chain $h_{t,K}$: assume that $h_{t,K} \overset{\mathcal{L}}{=} \mathrm{BinMix}(K, g_t)$ for some $g_t \in \mathcal{P}([0,1])$, then there exists a probability g_{t+1}, function of g_t, such that $h_{t+1,K} \overset{\mathcal{L}}{=} \mathrm{BinMix}(K, g_{t+1})$. Moreover, denoting by G_t the cumulative distribution function associated to g_t, we show that for all $t \geq 1$, $G_{t+1}(x) = \mathcal{R}(G_t)(x)$ where

Notation 3.3. *For all $\Gamma \in F_{[0,1]}$ and $u \in \mathbb{R}$, $\mathcal{R}$ is defined by*

$$\mathcal{R}(\Gamma)(u) \overset{def}{=} f_N \Gamma\left(\frac{u - f_N q^+}{1-(1-f_N)q_{10}^- - f_N q^+}\right) + (1-f_N)\Gamma\left(\frac{u}{1-f_N q_{01}^-}\right).$$

Proposition 3.4. *Let us assume that $h_{-r+1,K} \overset{\mathcal{L}}{=} \mathrm{BinMix}(K, g_{-r+1})$, for $g_{-r+1} \in \mathcal{P}([0,1])$. Then for all $t \geq 1$, $\exists g_t,\ g_t^0,\ g_t^1 \in \mathcal{P}([0,1])$ such that $h_{t,K} \overset{\mathcal{L}}{=} \mathrm{BinMix}(K, g_t)$ and $h_{t,K}^y \overset{\mathcal{L}}{=} \mathrm{BinMix}(K, g_t^y)$ for $y = 0, 1$. Moreover, at time $t = 1$,*

$$G_1(u) = f_N G_{-r+1}\left(\frac{u-1}{(1-q^+)^r} + 1\right) + (1-f_N) G_{-r+1}\left(\frac{u}{(1-q_{01}^-)^r}\right), \quad (7)$$

$$G_1^1(u) = G_{-r+1}\left(\frac{u-1}{(1-q^+)^r} + 1\right) \quad \textit{and} \quad G_1^0(u) = G_{-r+1}\left(\frac{u}{(1-q_{01}^-)^r}\right), \quad (8)$$

$$\textit{and} \ \forall t \geq 1, \quad G_{t+1}(u) = \mathcal{R}(G_t)(u) \quad \textit{and} \quad G_{t+1}^y(u) = \mathcal{R}(G_t^y)(u). \quad (9)$$

Remark 3.5. *First, we note that g_t does not depend on K. This is crucial for the proof of Theorems 3.15 and 3.18. Then, let the assumptions of the previous Proposition hold and denote by Y_t a random variable with distribution g_t. Knowing Y_t, we have $h_{t,K} \overset{\mathcal{L}}{=} \mathrm{Bin}(K, Y_t)$. In particular, the mean synaptic current is given by $\mathbb{E}(h_{t,K}) = K\mathbb{E}(Y_t) = K\mathbb{P}\left(J_t^{1j} = 1\right)$. Indeed, $\mathbb{E}(h_{t,K})$ is the mean number of strong synapses J_t^{1j} (with j such that $\xi_0^j = 1$) at time t.*

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

Proposition 3.6. *The application $\mathcal{R}$ acting on $F_{[0,1]}$ is contracting for the norm $L^1(0,1)$. Moreover, there exists a unique $G^* \in F_{[0,1]}$ invariant for $\mathcal{R}$.*

Propositions 3.4 and 3.6 are proved in the Appendices A.1 and A.2.

Corollary 3.7. *Let G^* be the unique fixed point of $\mathcal{R}$ and g^* its associated distribution. The invariant measure π_K of the Markov chain $h_{t,K}$ satisfies $\pi_K = \mathrm{BinMix}(K, g^*)$. In addition, the invariant measure π_∞ of h_t is given by $\pi_\infty = \mathrm{BinMix}\left(\hat{K}, g^*\right)$, where $\hat{K}$ has a Binomial law with parameters N and f_N, the two random variables being independent. Moreover, the smallest interval $[m_\infty, M_\infty]$ containing the support of g^* verifies*

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

Proof. Let $g^* \in \mathcal{P}([0,1])$ be a probability distribution such that its cumulative distribution function G^* satisfies $\mathcal{R}(G^*) = G^*$. Then, by Proposition 3.4, $\mathrm{BinMix}(K, g^*)$ is invariant for $(h_{t,K})_{t\geq 1}$. The result on π_∞ follows from Corollary 2.3.

Now, let $[m_\infty, M_\infty]$ be the convex envelop of the support of g^*, then $Supp(g^*) \subset [m_\infty, M_\infty] \subset [0,1]$. Thus by the equation $\mathcal{R}(G^*) = G^*$, we get

$$m_\infty = m_\infty(1 - f_N q_{01}^-) \wedge \left(m_\infty \left(1 - (1-f_N)q_{10}^- - f_N q^+\right) + f_N q^+\right),$$
$$M_\infty = M_\infty(1 - f_N q_{01}^-) \vee \left(M_\infty \left(1 - (1-f_N)q_{10}^- - f_N q^+\right) + f_N q^+\right).$$

As $(1 - f_N q_{01}^-) < 1$, the first equation implies that $0 \leq m_\infty \leq m_\infty(1 - f_N q_{01}^-)$ so $m_\infty = 0$, and the second equation implies that $M_\infty = M_\infty \left(1 - (1-f_N)q_{10}^- - f_N q^+\right) + f_N q^+$, thus $M_\infty = \frac{f_N q^+}{f_N q^+ + (1-f_N)q_{10}^-}$. □

Remark 3.8. *Propositions 3.4, 3.6, and the first part of Corollary 3.7 are in [Amit and Huang, 2010] with $q_{10}^- = 0$ and $r = 1$. We prove them here with a different method.*

3.2 Main results

The learning and the forgetting phases 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 results on t_*.

Spectrum

Let $P_{y,K}$ be the transition matrix of the synaptic current $\left(h_{t,K}^y\right)_{-r+1<t\leq 1}$. We denote by $\nu_{t,K}^y = \left[\nu_{t,K}^y(0),\ \nu_{t,K}^y(1),\ \ldots,\ \nu_{t,K}^y(K)\right]$ the distribution of $h_{t,K}^y$. We can then write $\nu_{1,K}^y = \nu_{0,K}^y P_{y,K} = \nu_{-r+1,K}^y \left(P_{y,K}\right)^r$.

Proposition 3.9. *The spectra of $P_{0,K}$ and $P_{1,K}$ are*

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

Proof. The dynamics give for all $j > i$, $P_{0,K}^{ij} = P_{1,K}^{ji} = 0$. So the matrices are triangular. Their spectra are given by the diagonal elements:

$$\forall i, \quad P_{0,K}^{ii} = (1 - q_{01}^-)^i \quad \text{and} \quad P_{1,K}^{ii} = (1 - q^+)^i.$$

□

Proposition 3.10. *The spectrum of the transition matrix P_K of $(h_{t,K})_{t\geq 1}$ is*

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

In the following, we denote by $\Lambda_0 = 1 - f_N q_{01}^-$, $\Lambda_1 = 1 - (1-f_N)q_{10}^- - f_N q^+$ and

$$\forall i \in [\![0, N]\!], \qquad \lambda_i = (1 - f_N)\Lambda_0^i + f_N \Lambda_1^i.$$

We prove the previous proposition using the

Lemma 3.11. *Let X and Y be two random variables in $[0,1]$ with cumulative distribution functions G_X and G_Y. We assume that there exist $\eta \in [0,1]$, $a, \bar{a} \in [0,1)$ and $b, \bar{b} \in (0,1]$ with $a+b \leq 1$, $\bar{a}+\bar{b} \leq 1$ such that*

$$G_Y(u) = \eta G_X\left(\frac{u-a}{b}\right) + (1-\eta) G_X\left(\frac{u-\bar{a}}{\bar{b}}\right). \tag{11}$$

Then $\forall k \in \mathbb{N},\ \mathbb{E}\left[Y^k\right] = \eta \mathbb{E}\left[(a+bX)^k\right] + (1-\eta)\mathbb{E}\left[(\bar{a}+\bar{b}X)^k\right]$.

Proof. First, note that $G_X\left(\frac{u-a}{b}\right)$ is the cumulative distribution function of $a+bX$. Second, for all random variables U, V, W, we have

$$G_U(z) = \eta G_V(z) + (1-\eta) G_W(z) \Longrightarrow \mathbb{E}[U^k] = \eta \mathbb{E}[V^k] + (1-\eta)\mathbb{E}[W^k].$$

This last result is obtained by differentiation, multiplication by z^k and integration. It ends the proof of the lemma. □

In the proof below, we use the classical conventions $\binom{i}{j} = 0$ when $j > i$ or $j < 0$.

Proof of Proposition 3.10. We denote by $\nu_{t,K} = [\nu_{t,K}(0),\ \nu_{t,K}(1),\ \ldots,\ \nu_{t,K}(K)]$ the distribution of $h_{t,K}$. Its transition matrix $P_K = \left(P_K^{ij}\right)_{0 \leq i,j \leq K}$ can be derived from Proposition 2.2:

$$P_K^{ij} = (1-f_N)\binom{i}{i-j}(f_N q_{01}^-)^{i-j}(1-f_N q_{01}^-)^j$$
$$+ f_N \sum_{l=0}^{i} \binom{i}{l}((1-f_N)q_{10}^-)^l(\Lambda_1 + f_N q^+)^{i-l}\binom{K-i}{j-i+l}(f_N q^+)^{j-i+l}(1-f_N q^+)^{K-j-l}.$$

Let us define the two matrices $\tilde{P}_K$ and Q_K such that for all $0 \leq i, j \leq K$:

$$\tilde{P}_K^{ij} = f_N \binom{j}{i}\Lambda_1^i (f_N q^+)^{j-i} + (1-f_N)\delta_{ij}\Lambda_0^i \quad \text{and} \quad Q_K^{ij} = \binom{K}{i}\binom{i}{j}(-1)^{i-j}.$$

Then, assuming that $\nu_{t,K} \overset{\mathcal{L}}{=} \mathrm{BinMix}(K, g_t)$ and denoting by $U_t = \left[U_t^0,\ U_t^1,\ \ldots,\ U_t^K\right]$ with $U_t^k = \int u^k g_t(du)$, we get by definition 3.1: $\nu_{t,K} = U_t Q_K$. Moreover, by Lemma 3.11 we have $U_{t+1} = U_t \tilde{P}_K$. Finally, by definition we have $\nu_{t+1,K} = \nu_{t,K} P_K$, so we obtain:

$$\nu_{t+1,K} = U_{t+1} Q_K = U_t \tilde{P}_K Q_K = U_t Q_K Q_K^{-1} \tilde{P}_K Q_K = \nu_{t,K} Q_K^{-1} \tilde{P}_K Q_K = \nu_{t,K} P_K.$$

A straightforward computation shows that $Q_K P_K = \tilde{P}_K Q_K$. Thus P_K and $\tilde{P}_K$ have the same spectrum. Finally, $\tilde{P}_K$ is a triangular matrix with λ_i as diagonal elements. □

We deduce from Proposition 3.10 the rate of convergence of the law of $h_{t,K}$.

Corollary 3.12. *For all $0 \leq K \leq N$, the sequence of the distributions of the synaptic currents, $(\nu_{t,K})_{t\geq 1}$, converges exponentially fast to the unique invariant measure π_K. In particular, there exists $c \in \mathbb{R}^+$ such that the distance in total variation between $\nu_{t,K}$ and π_K satisfies:*

$$\forall t \geq 1, \ \ ||\nu_{t,K} - \pi_K||_{TV} := \frac{1}{2}\sum_{l=0}^{K} |\nu_{t,K}(l) - \pi_K(l)| \leq c\lambda_1^t.$$

We discuss, in the second paragraph of Section 5, the role played by this eigenvalue λ_1 in our main results.

Memory lifetime

Under Assumption 2.6.1, $h_{-r+1,K}$ follows its invariant distribution π_K, a Binomial mixture by Corollary 3.7. Thus, by Proposition 3.4, the processes $(h^y_{t,K})_{t\geq 1}$ follow also Binomial mixtures. Combining the inequality provided by Lemma 3.13, inequalities on Binomial tails (Lemma 3.14) and a control on the tail of the mixing distribution g^* and on the support of g^1_t, we prove Theorem 3.15.

Lemma 3.13. *Under Assumption 2.6.1, for all $\theta \in [\![0, N]\!]$, $\mathbb{P}\left(h^0_t > \theta\right) \leq \mathbb{P}\left(\pi_\infty > \theta\right)$.*

Proof. The proof is recursive and relies on the functional equation for the cumulative distribution of the synaptic currents (8) under Assumption 2.6.1. From (9), we have for all $x \in [0, 1]$, $G^0_1(x) = G^*\left(\frac{x}{\Lambda_0}\right) \geq G^*(x)$. Then,

$$\begin{aligned} G^0_2(x) &= f_N G^0_1\left(\frac{x - f_N q^+}{\Lambda_1}\right) + (1 - f_N) G^0_1\left(\frac{x}{\Lambda_0}\right) \\ &\geq f_N G^*\left(\frac{x - f_N q^+}{\Lambda_1}\right) + (1 - f_N) G^*\left(\frac{x}{\Lambda_0}\right) = G^*(x), \end{aligned}$$

and so forth so that for all $t \geq 1$ and x, $G^0_t(x) \geq G^*(x)$. It implies that for all $K,\ \theta \in \mathbb{N}$, $\mathbb{P}\left(\mathrm{BinMix}(K, g^0_t) > \theta\right) \leq \mathbb{P}\left(\mathrm{BinMix}(K, g^*) > \theta\right)$, which ends the proof. □

Lemma 3.14. *Let $S_N \overset{\mathcal{L}}{=} \mathrm{Bin}(N, p)$. Then, for all $\varepsilon \in (0, 1)$*

$$\mathbb{P}\left(S_N \geq Np(1 + \varepsilon)\right) \leq \exp\left(\frac{-Np\,\varepsilon^2}{2 + \varepsilon}\right), \tag{12}$$

$$\mathbb{P}\left(S_N \leq Np(1 - \varepsilon)\right) \leq \exp\left(\frac{-Np\,\varepsilon^2}{2}\right). \tag{13}$$

This Lemma is proved in A.3.

We now give our main results.

Theorem 3.15. *For $y \in \{0, 1\}$, let $(h^y_t)_{t\geq 1}$ be the solutions of* (5) *with $\xi^1_0 = y$ and* (6)*. Let us assume that Assumptions 2.6.1 and 2.6.2 hold and that q^-_{01} and q^+ are fixed in $(0, 1]$ and q^-_{10} in $[0, 1]$.*

Then, for all $0 < \delta < 1$ and $r \in \mathbb{N}^*$, there exists $N(\delta, r) \in \mathbb{N}$ such that for all $N \geq N(\delta, r)$, there exist $\theta_{\delta,N} \in [\![0, N]\!]$ and $\hat{t}(\delta, r, N)$ such that for all $1 \leq t \leq \hat{t}(\delta, r, N)$,

$$\mathbb{P}\left(h_t^0 > \theta_{\delta,N}\right) \vee \mathbb{P}\left(h_t^1 \leq \theta_{\delta,N}\right) \leq \delta.$$

In particular, we have $t_*(\delta, r, N) \geq \hat{t}(\delta, r, N)$. This result relies on the study of the mixing distributions g^* and g_t^1. Thanks to Lemma 3.13, we know that as long as g_t^1 is far enough from g^*, the probability of error,

$\mathbb{P}\left(h_t^0 > \theta\right) \vee \mathbb{P}\left(h_t^1 \leq \theta\right) \leq \mathbb{P}\left(\mathrm{BinMix}(K, g^*) > \theta\right) \vee \mathbb{P}\left(\mathrm{BinMix}(K, g_t^1) \leq \theta\right)$, is small enough. This condition appears as an inequality depending both on the time and the accepted error δ. As long as this inequality holds, there exists a threshold θ such that the probability of error is below δ for all previous times.

Example 3.16. *We give in Remark 3.17 an explicit formula for the lower bound $\hat{t}$ on t_* for any couple (δ, r). We give here a detailed result for a particular choice of parameters. Let $q^+ = q_{01}^- = 1$, q_{10}^- small enough, and $f_N = \frac{q_{10}^-}{3+q_{10}^-}$. Explicit computations give*

$$\hat{t}(\delta, r, N) = \left\lfloor \frac{\log\left(\frac{1}{9}\right) \vee \log\left(\frac{\sqrt{-2\log(\frac{\delta}{2})Nf_N} - 16\log(\frac{\delta}{2})}{3Nf_N}\right)}{\log\left(1 - 4f_N\right)} \right\rfloor.$$

For instance, for $q_{10}^- = 0.005$ we get $f_N = 0.00167$ and

$$\hat{t}(\delta = 0.001, r = 1, N = 2.10^5) = 246 \quad \textit{and} \quad \theta_{\delta,N} = 122.$$

We also give a formula when the depression probabilities depend on N in Theorem 3.18.

Proof of Theorem 3.15. The proof follows these lines: from Lemma 3.13 we have that $\mathbb{P}\left(h_t^0 > \theta\right) \leq \pi_\infty\left(]\theta, +\infty[\right)$. Hence, we propose a threshold θ based on the measure π_∞ such that $\pi_\infty\left(]\theta, +\infty[\right) \leq \delta$ and then we bound the time before which $\mathbb{P}\left(h_t^1 \leq \theta\right) \geq \delta$.

We split $\pi_\infty\left(]\theta, +\infty[\right)$ in two terms. We recall that $\pi_\infty = \mathrm{BinMix}\left(K, g^*\right)$ with $K \overset{\mathcal{L}}{=} \mathrm{Bin}\left(N, f_N\right)$ and $[0, M_\infty]$ is the smallest interval containing the support of g^*. So

$$\begin{aligned}\pi_\infty\left(]\theta, +\infty[\right) &= \int_0^{M_\infty} \mathbb{P}\left(\mathrm{Bin}\left(K, u\right) > \theta\right) g^*(du) = \int_0^{M_\infty} \mathbb{P}\left(\mathrm{Bin}\left(N, f_N u\right) > \theta\right) g^*(du) \\ &\leq \mathbb{P}\left(\mathrm{Bin}\left(N, f_N M_\delta\right) > \theta\right) + \int_{M_\delta}^{M_\infty} g^*(du).\end{aligned}$$

The second equality comes from the following property: assume $K \overset{\mathcal{L}}{=} \mathrm{Bin}\left(N, f_N\right)$ and conditionally on K, X is independent of K with law $\mathrm{Bin}\left(K, p\right)$, then $X \overset{\mathcal{L}}{=} \mathrm{Bin}\left(N, f_N p\right)$. Let Y^* be a random variable with distribution g^*. We propose a value for M_δ using the Bienaymé-Tchebytchev inequality:

$$M_\delta = \left(\sqrt{\frac{2\mathrm{Var}\left(Y_*\right)}{\delta}} + \mathbb{E}\left(Y_*\right)\right) \wedge M_\infty \quad \Rightarrow \quad \int_{M_\delta}^{M_\infty} g^*(du) \leq \frac{\delta}{2}.$$

We first fix $\theta_{\delta,N}$ such that $\mathbb{P}\left(\mathrm{Bin}\left(N, f_N M_\delta\right) \geq \theta_{\delta,N}+1\right) \leq \frac{\delta}{2}$. To do so we apply Lemma 3.14 with $\varepsilon = \frac{\theta_{\delta,N}+1}{N f_N M_\delta} - 1$ and obtain:

$$\theta_{\delta,N} = \left\lfloor N f_N M_\delta + \sqrt{-2\log\left(\frac{\delta}{2}\right) N f_N M_\delta} - \log\left(\frac{\delta}{2}\right) \right\rfloor.$$

We now bound the probability of error $\mathbb{P}\left(h_t^1 \leq \theta_{\delta,N}\right)$. Let $[m_t^1, M_t^1]$ be the smallest interval containing the support of g_t^1. Then, we get:

$$\mathbb{P}\left(h_t^1 \leq \theta_{\delta,N}\right) = \int_{m_t^1}^{M_t^1} \mathbb{P}\left(\mathrm{Bin}\left(K, u\right) \leq \theta_{\delta,N}\right) g_t^1(du) \leq \mathbb{P}\left(\mathrm{Bin}\left(N, f_N m_t^1\right) \leq \theta_{\delta,N}\right).$$

Using Lemma 3.14 with $\varepsilon = 1 - \frac{\theta_{\delta,N}}{N f_N m_t^1}$, we get

$$\mathbb{P}\left(\mathrm{Bin}\left(N, f_N m_t^1\right) \leq \theta_{\delta,N}\right) \leq \exp\left(-\frac{\left(N f_N m_t^1 - \theta_{\delta,N}\right)^2}{2 N f_N m_t^1}\right).$$

Using the inequality $\sqrt{x} + \sqrt{y} \geq \sqrt{x+y}$ for all $x, y > 0$, we obtain that if

$$N f_N m_t^1 \geq \theta_{\delta,N} + \sqrt{-2\log(\delta)\theta_{\delta,N}} - 2\log(\delta) \tag{14}$$

then $\mathbb{P}\left(h_t^1 \leq \theta_{\delta,N}\right) \leq \delta$. Let us define $m_{\delta,N} := \frac{1}{N f_N}\left(\theta_{\delta,N} + \sqrt{-2\log(\delta)\theta_{\delta,N}} - 2\log(\delta)\right)$. Using the bound $\theta_{\delta,N} \leq \left(\sqrt{N f_N M_\delta} + \frac{\sqrt{-2\log(\frac{\delta}{2})}}{2}\right)^2$ we get

$$N f_N m_{\delta,N} + \frac{3}{2}\log(\delta) = \left(\sqrt{\theta_{\delta,N}} + \frac{\sqrt{-2\log(\delta)}}{2}\right)^2 \leq \left(\sqrt{M_\delta N f_N} + \sqrt{-2\log(\frac{\delta}{2})}\right)^2. \tag{15}$$

We now find m_t^1. From equation (9) and the definition of $\mathcal{R}$ (see Notation 3.3), we have

$$\forall t \geq 1, \qquad m_{t+1}^1 = m_t^1 \Lambda_0 \ \wedge \ (m_t^1 \Lambda_1 + f_N q^+).$$

We note that for N large enough such that $m_t^1 > \frac{f_N q^+}{\Lambda_0 - \Lambda_1} \geq M_\infty$, we have

$$\frac{f_N q^+}{1-\Lambda_1} = M_\infty < m_t^1 \Lambda_1 + f_N q^+ < m_t^1 \Lambda_0 < m_t^1.$$

Denoting by $t_c = \inf\{t \in \mathbb{N}^*, m_t^1 \leq \frac{f_N q^+}{\Lambda_0 - \Lambda_1}\}$, we obtain

$$m_t^1 = \left(\left(m_1^1 - M_\infty\right)\Lambda_1^{(t\wedge t_c)-1} + M_\infty\right)\Lambda_0^{(t-t_c)\mathbb{1}_{t>t_c}}. \tag{16}$$

Let us now consider q^+, q_{01}^- and q_{10}^- fixed in $(0,1]$. By definition, $M_\delta \leq M_\infty$, hence

$$N f_N m_{\delta,N} \leq \left(\sqrt{M_\infty N f_N} + \sqrt{-2\log\left(\frac{\delta}{2}\right)}\right)^2 - \frac{3}{2}\log(\delta).$$

Therefore, the inequality (14) holds true as long as

$$t-1 \leq \left(\left\lfloor \frac{\log\left(\frac{2\sqrt{-2\log(\frac{\delta}{2})Nf_N M_\infty}-4\log(\frac{\delta}{2})}{Nf_N(m_1^1-M_\infty)}\right)}{\log(\Lambda_1)} \right\rfloor \wedge t_c \right)_+ \quad \text{with } (x)_+ = x\mathbb{1}_{x\geq 0}. \quad (17)$$

But $m_1^1 = 1-(1-q^+)^r \geq q^+$ and both $\frac{f_N q^+}{\Lambda_0-\Lambda_1}$ and $M_\infty = \frac{f_N q^+}{f_N q^+ + (1-f_N)(q_{01}^- + q_{10}^-)}$ tends to 0 with increasing N so there exists $N(\delta, r)$ such that for all $N \geq N(\delta, r)$ we can remove "$(\ )_+$" in the inequality (17): that is to say for all $N \geq N(\delta, r)$ such that,

$$\frac{2\sqrt{-2\log(\frac{\delta}{2})Nf_N M_\infty} - 4\log(\frac{\delta}{2})}{Nf_N(m_1^1 - M_\infty)} < 1 \quad \text{and} \quad m_1^1 > \frac{f_N q^+}{\Lambda_0 - \Lambda_1}.$$

Using the fact that for all $\delta \in (0,1)$, $\sqrt{-2\log(\frac{\delta}{2})} \leq -2\log(\frac{\delta}{2})$ and $m_1^1 \geq q^+$ we get the two following conditions on N:

$$2\exp\left(-\frac{Nf_N(m_1^1 - M_\infty)}{4(\sqrt{Nf_N M_\infty}+1)}\right) < \delta \ \text{ and } \ \frac{f_N q^+}{f_N q^+ + (1-f_N)q_{10}^- - f_N q_{01}^-} < q^+. \quad (18)$$

In the particular case $q_{10}^- = 0$, we have $M_\infty = 1$ so the dynamics of m_t^1 is simply $m_t^1 = m_1^1 \Lambda_0^{t-1}$. We compute $\mathbb{E}(Y_*)$ and $\text{Var}(Y_*)$ using Lemma 3.11 and equation (9):

$$\mathbb{E}(Y_*) = \frac{f_N^2 q^+}{1-\lambda_1} = \frac{f_N q^+}{f_N q^+ + (1-f_N)(q_{01}^- + q_{10}^-)}, \ \text{Var}(Y_*) = \frac{f_N^5 (1-f_N){q^+}^2 {q_{01}^-}^2}{(1-\lambda_1)^2(1-\lambda_2)}.$$

Hence,

$$M_\delta = \frac{f_N q^+ \left(1 + q_{01}^- \sqrt{\frac{2f_N}{\delta(1-\lambda_2)}}\right)}{f_N q^+ + (1-f_N)(q_{01}^- + q_{10}^-)}. \quad (19)$$

We note that $1-\lambda_2 \sim_{N\infty} 2f_N(q_{01}^- + q_{10}^-)$, so M_δ converges to 0 with increasing N. Thus, by inequality (15), there exists a $N(\delta, r)$ such that for all $N \geq N(\delta, r)$, $m_{\delta,N} < 1$. We conclude that for all $N \geq N(\delta, r)$, the inequality (14) holds true as long as

$$t-1 \leq \left\lfloor \frac{\log\left(\frac{\left(\sqrt{M_\delta N f_N} + \sqrt{-2\log(\frac{\delta}{2})}\right)^2 - \frac{3}{2}\log(\delta)}{Nf_N}\right)}{\log(\Lambda_0)} \right\rfloor > 0.$$

□

Remark 3.17. *Recall that* $M_\infty = \frac{f_N q^+}{1-\Lambda_1}$, $m_1^1 = 1-(1-q^+)^r$, $M_\delta = \frac{f_N q^+ \left(1+q_{01}^- \sqrt{\frac{2f_N}{\delta(1-\lambda_2)}}\right)}{f_N q^+ + (1-f_N)(q_{01}^- + q_{10}^-)}$, $\Lambda_0 = 1 - f_N q_{01}^-$, $\Lambda_1 = 1 - f_N q^+ - (1-f_N)q_{10}^-$ *and* $\lambda_2 = f_N \Lambda_1^2 + (1-f_N)\Lambda_0^2$.

We proved that under Assumptions 2.6.1 and 2.6.2, for all δ, r, $N \geq N(\delta, r)$ *(*N *for which the two conditions given by* (18) *are satisfied), there exists* $\theta_{\delta,N} \in [\![0, N]\!]$ *and* $\hat{t}$ *such that for all* $1 \leq t \leq \hat{t}(\delta, r, N)$, $\mathbb{P}\left(h_t^0 > \theta_{\delta,N}\right) \vee \mathbb{P}\left(h_t^1 \leq \theta_{\delta,N}\right) \leq \delta$.

In particular, if $q_{01}^-, q_{10}^-, q^+ \in (0,1]$

$$\hat{t}(\delta,r,N)-1 = \left\lfloor \frac{\log\left(\frac{2\sqrt{-2\log(\frac{\delta}{2})Nf_N M_\infty}-4\log(\frac{\delta}{2})}{Nf_N(m_1^1-M_\infty)}\right)}{\log(\Lambda_1)} \right\rfloor \wedge \left\lfloor \frac{\log\left(\frac{f_N^2 q^+ q_{01}^-}{(1-\Lambda_1)(\Lambda_0-\Lambda_1)(m_1^1-M_\infty)}\right)}{\log(\Lambda_1)} \right\rfloor,$$

and if $q_{10}^- = 0$, $\hat{t}(\delta,r,N)-1 = \left\lfloor \frac{\log\left(\frac{\left(\sqrt{M_\delta N f_N}+\sqrt{-2\log(\frac{\delta}{2})}\right)^2-\frac{3}{2}\log(\delta)}{Nf_N}\right)}{\log(\Lambda_0)} \right\rfloor$.

Theorem 3.18. *Assume Assumptions 2.6.1, 2.6.2 and 2.6.3 are satisfied. Then, for all* $\delta \in (0,1)$*,* r *large enough, there exists* $N(\delta,r) \in \mathbb{N}$ *such that for all* $N \geq N(\delta,r)$*,*

$$\hat{t}(\delta,r,N) = t_c + \left\lfloor \frac{\log\left(C(\delta,r,N)\right)}{\log(\Lambda_0)} \right\rfloor,$$

with t_c *defined in* (16) *and* $C(\delta,r,N) \in (0,1)$ *satisfies* $\frac{\log(C(\delta,r,N))}{f_N} \to +\infty$*. Moreover, if* $\lim a_N$*, and* $\lim b_N$ *exist and are finite,* $\frac{\log(C(\delta,r,N))}{\log(\Lambda_0)}$ *is on the order of* $\frac{1}{f_N^2}$*.*

We note that $\log(\Lambda_0) = \log(1-a_N f_N^2) \sim_{N\infty} -a_N f_N^2$. Concerning $C(\delta,r,N)$ (and then $\hat{t}(\delta,r,N)$), it mainly depends on the different large N asymptotic of a_N and b_N. We detail in Remark 3.19 the different large N asymptotic of $\hat{t}(\delta,r,N)$.

Proof. We use the results proved in the proof of Theorem 3.15. From the dynamics of m_t^1 given by the equation (16) and the bound $m_{t_c}^1 \geq m_1^1 \wedge M_\infty$, we obtain that the inequality (14) is satisfied as long as

$$t-1 \leq t_c + \left(\left\lfloor \frac{\log\left(\frac{m_{\delta,N}}{m_{t_c}^1}\right)}{\log(\Lambda_0)} \right\rfloor\right)_+ \leq t_c + \left(\left\lfloor \frac{\log\left(\frac{m_{\delta,N}}{(m_1^1\wedge M_\infty)}\right)}{\log(\Lambda_0)} \right\rfloor\right)_+. \tag{20}$$

We can remove "$(\ )_+$" in the last inequality if there exists N_0 such that

$$\forall N \geq N_0, \qquad \frac{m_{\delta,N}}{(m_1^1 \wedge M_\infty)} < 1.$$

Using the inequality (15), we deduce that this is the case if

$$C(\delta,r,N) = \sqrt{\frac{M_\delta}{m_1^1 \wedge M_\infty}} + 2\sqrt{\frac{-\log(\frac{\delta}{2})}{(m_1^1 \wedge M_\infty)\, N f_N}} < 1. \tag{21}$$

From the previous computation of M_δ, see equation (19), we obtain

$$\frac{M_\delta}{M_\infty} = \left(1 - \frac{(1-f_N)a_N}{q^+ + (1-f_N)(a_N+b_N)}\right)\left(1 + a_N f_N \sqrt{\frac{2f_N}{\delta(1-\lambda_2)}}\right).$$

Thus, we compare the three terms (recalling that $m_1^1 = 1-(1-q^+)^r$)

$$\frac{a_N}{q^+ + a_N + b_N}, \quad a_N f_N \sqrt{\frac{2f_N}{\delta(1-\lambda_2)}} \quad \text{and} \quad \frac{-\log(\frac{\delta}{2})}{((1-(1-q^+)^r)\wedge M_\infty)\, N f_N}.$$

First, $(1-\lambda_2) \sim_{N\infty} 2f_N^2(a_N+b_N+q^+)$. Then, we have to separate the different cases:

- If b_N tends to $+\infty$, both M_δ and M_∞ converge to 0. Hence, $(1-(1-q^+)^r) \wedge M_\infty = M_\infty$ and the Assumption 2.6.3 , in particular $\lim\limits_{N\infty} q^-_{01,N} = \lim\limits_{N\infty} q^-_{10,N} = \lim\limits_{N\infty} \frac{b_N^2}{Nf_N a_N} = \lim\limits_{N\infty} \frac{b_N}{Nf_N} = 0$, enables us to conclude that if $b_N = \mathcal{o}(a_N)$, $C(\delta,r,N) \sim_{N\infty} \sqrt{\frac{b_N}{a_N}} + 2\sqrt{\frac{-\log(\frac{\delta}{2})b_N}{q^+Nf_N}} \to 0$, else, $C(\delta,r,N) \sim_{N\infty} \left(1-\frac{a_N}{a_N+b_N}\right)$ so the inequality (21) holds true for any r and for a N large enough.

- If a_N tends to $+\infty$ and not b_N, then M_δ converges to 0 and M_∞ converges to 1 (resp. $\frac{q^+}{q^++b}$) if b_N converges to 0 (resp. b). Thus, $C(\delta,r,N)$ converges to 0 with large N and for any r, the inequality (21) is satisfied.

- If a_N tends to 0 and b_N converges to $b>0$, then M_δ converges to M_∞. Then, there exists r_0 such that $(1-(1-q^+)^{r_0}) \wedge M_\infty = M_\infty$. Using the assumption $\lim\limits_{N\infty} a_N N f_N = +\infty$ we have for all $r \geq r_0$, $C(\delta,r,N) \sim_{N\infty} \left(1-\frac{a_N}{q^++b}\right)$.

- In all other cases, M_δ and M_∞ converges to a value in $(0,1)$. Moreover, $M_\delta < M_\infty$ so there exists r_0 such that for all $r \geq r_0$, $(1-(1-q^+)^{r_0}) \wedge M_\infty > M_\delta$, so $C(\delta,r,N)$ converges in $(0,1)$ with large N.

□

Remark 3.19. *In the large N asymptotic (under Assumptions 2.6.2 and 2.6.3), we can compute the terms equivalent to $\hat{t}$ in the different a_N, b_N cases ($a \in \mathbb{R}_*^+$ and $b \in \mathbb{R}^+$):*

conditions on $a_N,\ b_N$ and r	$\hat{t}(\delta,r,N)$ for large N
$b_N \to +\infty,\ b_N = \mathcal{o}(a_N)\,,\ \forall r$	$\frac{\log\left(\sqrt{\frac{b_N}{a_N}}+2\sqrt{\frac{-\log(\frac{\delta}{2})b_N}{q^+Nf_N}}\right)}{f_N^2 a_N}$
$a_N, b_N \to +\infty$ of same order, $\forall r$	$-\frac{\log\left(1-\frac{a_N}{a_N+b_N}\right)}{2f_N^2 a_N}$
$a_N = \mathcal{o}(b_N)\,,\ b_N \to +\infty,\ \forall r$	$\frac{1}{2f_N^2 b_N}$
$a_N \to +\infty,\ b_N \to b \in \mathbb{R}^+,\ \forall r$	$\frac{-\log\left(\sqrt{\frac{q^+}{\left((1-(1-q^+)^r)\wedge\frac{q^+}{q^++b}\right)a_N}}+2\sqrt{\frac{-\log(\frac{\delta}{2})(q^++b)}{q^+Nf_N}}\right)}{f_N^2 a_N}$
$a_N \to 0,\ b_N \to b > 0,\ \forall r > r_0$	$\frac{1}{2f_N^2(q^++b)}$
$a_N = a,\ b_N \to 0$ ou $b_N = 0,\ \forall r > r_0$	$-\frac{\log\left(\frac{q^+}{(1-(1-q^+)^r)(q^++a)}\right)}{2f_N^2 a}$
$a_N = a,\ b_N = b,\ \forall r > r_0$	$-\frac{\log\left(1-\frac{a}{q^++a+b}\right)}{2f_N^2 a}$

Table 1: The large N equivalent of $\hat{t}(\delta,r,N)$ in function of a_N and b_N.

Remark 3.20. *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. Moreover, in the large r asymptotic, $h^0_{1,K} \overset{\mathcal{L}}{=} \delta_0$ (dirac in 0) and $h^1_{1,K} \overset{\mathcal{L}}{=} \delta_K$ (dirac in K). Thus, the initial error is null. However, the $\hat{t}$ increases with r until reaching a threshold value which is given by the expression of Remarks 3.17 and 3.19 replacing the quantities m^1_1 by 1.

4 Simulations

Our code follows these lines. We draw ξ_0 and $K = \sum_{j=2}^{N+1} \xi_0^j$. We simulate a trajectory of $h_{t,K}$ long enough to be under the invariant measure. We perform r presentations of the signal to be learnt and then compute the trajectories of $h^y_{t,K}$, $y \in \{0, 1\}$. We reiterate this procedure $N_{MC} = 10^7$ times to get an approximation of the distributions of h^y_t.

The result of Theorem 3.15 is interesting for large values of Nf_N (small errors) combined with a small f_N (non-negligible $\hat{t}$). In this context, we need to compute many trajectories before the synaptic currents cross a reasonable threshold θ.

In Figure 3a, the top (resp. bottom) roughly represents the distribution of $h^1_{t,K}$ (resp. $h^0_{t,K}$). Before time $t = 50$, the distribution of h^0_t is highly concentrated in 0. Indeed, looking carefully to Figure 3a, we can observe a residue of this high probability (dark blue) for very weak synaptic currents until time $t = 65$, see also Figure 4a. This concentration drastically reduces the contrast of the plot. That is why the time axis starts at $t = 50$ in Figure 3a. This figure shows that a threshold θ around one hundred is a good choice: it seems to maximises the time for which the threshold estimation holds true. With this threshold, the numerical errors $p^0_e(t, \theta)$ and $p^1_e(t, \theta)$ does not exceed 10^{-4}, see Figure 3b, before time 15. It is coherent with $\hat{t}$ plotted in Figure 3c. Indeed, the time $\hat{t}$ is equal to 12 for errors on the order of 10^{-4}, see Figure 3c. Moreover, in Remark 3.19, the result $\hat{t}$ is a maximum between two times. The second one does not depend on the error δ (it is called t_c in the proof of Theorem 3.15, see equation (16)). This explains the plateau starting at an error just before 10^{-3} in Figure 3c. Indeed, for this set of parameters and δ large enough, the time $\hat{t}$ is equal to t_c. Finally, in Figure 3d, we note that p^0_e is above p^1_e for small values of t. Then, around time $t = 70$, p^1_e increases quickly until a value close to one whereas p^0_e stays below 10^{-2}. This is because the majority of the mass of the distribution of $h^0_{t,K}$ stays less than θ. On the other hand, most of the mass of the distribution of $h^1_{t,K}$ crosses θ around time $t = 70$. So, the error p^1_e becomes large. We present the histograms of the distributions of the synaptic currents at certain times.

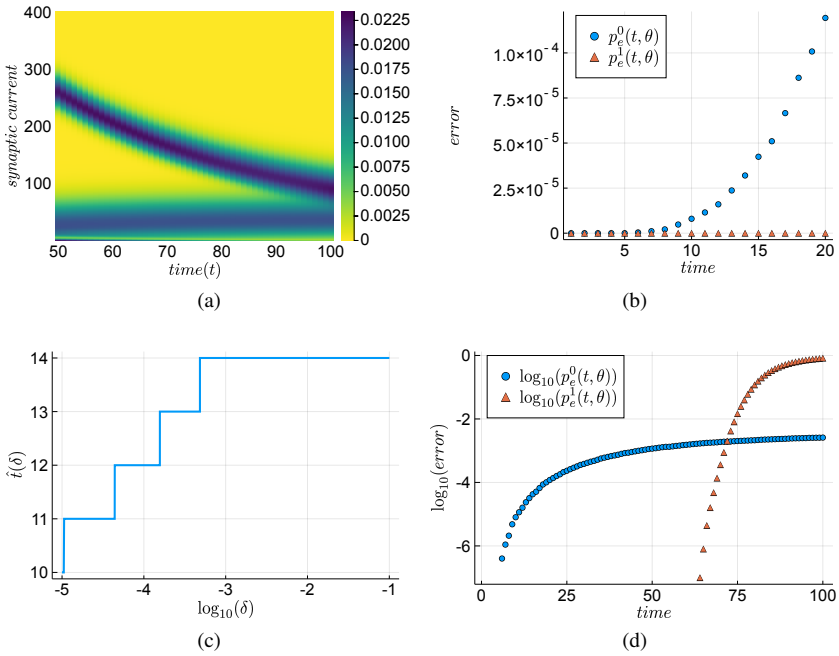


Figure 3: (3a) The sum of the distributions of $h^0_{t,K}$ and $h^1_{t,K}$. The colour bar gives the probability values. (3b) (resp. (3d)) The numerical errors $p^0_e(t,\theta)$ and $p^1_e(t,\theta)$ on a short (resp. large) timescale. (3c) $\hat{t}$ as a function of δ on the logarithmic to the base ten scale. Parameters: $\theta = 117,\ N = 20\,000,\ f_N = 0.05,\ q^+ = q^-_{01} = 0.5,\ q^-_{10} = 0.05,\ r = 3.$

We note again that the invariant measure is concentrated around small values. This enables the post learning distribution of h^0_1 to have a small variance, see Figures 4a and 4c. However, the variance of this distribution increases quickly. In particular, the distribution of h^0_t has a multimodal shape with a high proportion of the mass staying near 0 for more than 50 presentations after learning. On the other hand, the distribution of h^1_t keeps a unimodal shape with a variance decreasing at the beginning, then increasing before decreasing again, see Figure 4b. Distributions stay 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.

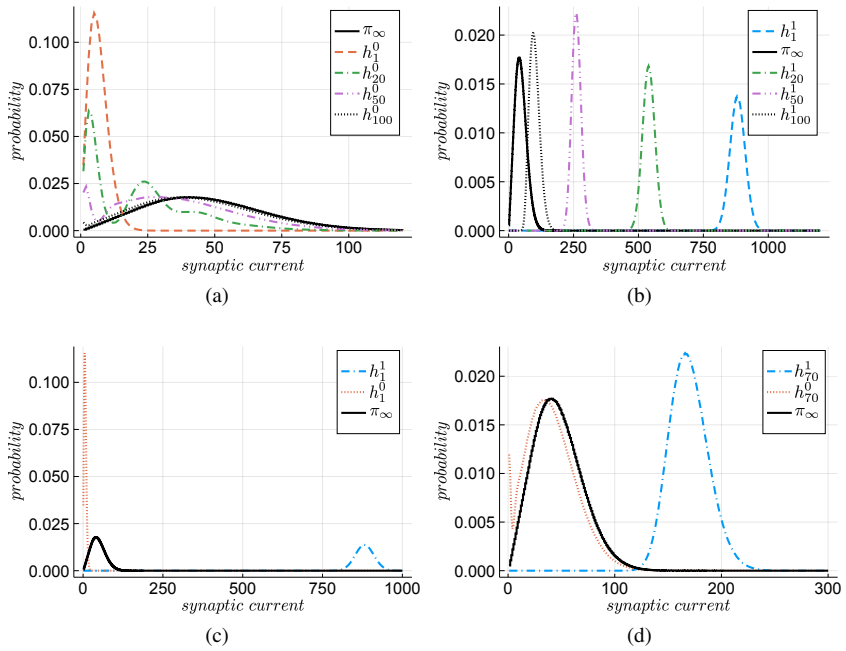


Figure 4: (4a) Histograms of the distributions of h_t^0 at different times. (4b) Histograms of the distributions of h_t^1 at different times. (4c) Distributions of h_1^y just after the learning phase and the invariant measure. (4d) Distributions of h_t^y at $t = 70$ and the invariant measure. Parameters: $N = 20\,000$, $f_N = 0.05$, $q^+ = q_{01}^- = 0.5$, $q_{10}^- = 0.05$, $r = 3$.

Because of the parameters choice, the distributions of h_1^0 are close to the invariant measure π_∞ whereas the distributions of h_1^1 are further from it, see Figures 5a and 5b. Moreover, the forgetting is really slow. However, if we want 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. Figures 5c and 5d show well the difference brought by a higher value of r: the separation between the two distributions is clearer.

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 onto one neuron during the presentation of this stimulus. First, we compute the spectrum of the transition matrix of the Markov chain associated to the synaptic current. This enables us to conclude that

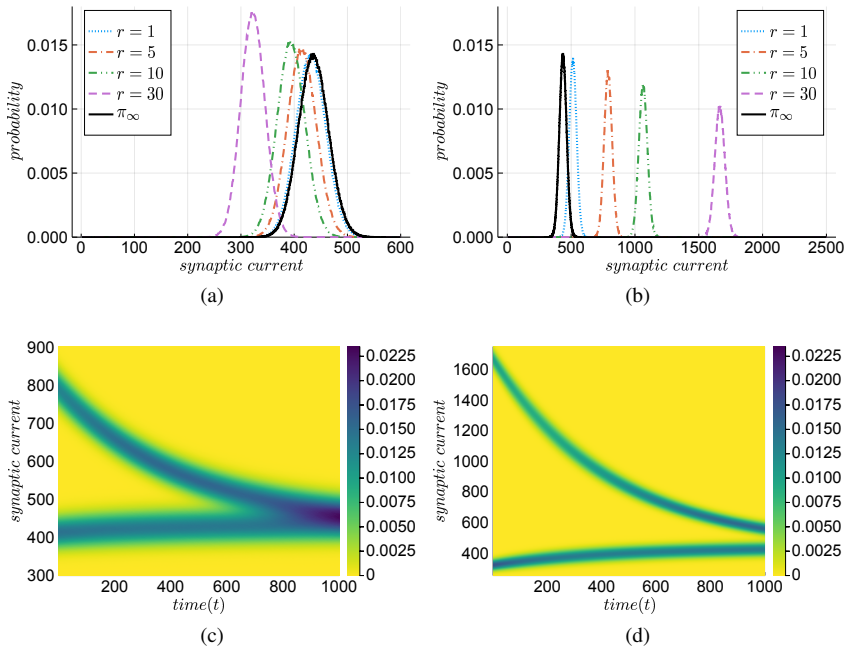


Figure 5: (5a) Distributions of h_1^0, just after learning, for different values of r and the invariant measure. (5b) Distributions of h_1^1, for different values of r and the invariant measure. (5c) The sum of the distributions of h_t^0 and h_t^1 for $r = 5$. (5d) The sum of the distributions of h_t^0 and h_t^1 for $r = 30$. The colour bar gives the probability values. Parameters: $N_{MC} = 10^6,\ N = 20\,000,\ f_N = 0.1,\ q_{01}^- = q_{10}^- = 0.01$ and $q^+ = 0.05$.

the eigenvalues are strictly different whatever the parameters are. In particular, we can compute the rate of convergence of the chain to its invariant measure, see Corollary 3.12. Then, we carry on the work done by [Amit and Huang, 2010] on the dynamics of the distributions of the synaptic current and their invariant distribution. This leads us to control the form of these distributions. Their properties give enough information to find a lower bound on the time a neuron keeps a good estimate on its response to the first stimulus and hence remembers it. We measure the quality of this estimation by performing a statistical test based on the observation of the synaptic current onto one neuron. We define an error associated to this test which depends on two distributions: the distribution of the synaptic current knowing that the neuron was selective to the initial signal and the distribution knowing that the neuron was not selective. Finally, unlike previous studies, we take into account the possibility that heterosynaptic and homosynaptic depressions scale differently in the network size N and we consider the role of presenting several times a signal in the learning phase.

We use the model presented by [Amit and Fusi, 1994] because of its relative simplicity and its consideration of synapse correlations. Their study focused 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). In particular, they studied the SNR in the large N asymptotic. They obtained a large SNR when the coding level f_N is low and the depression probabilities are proportional to f_N: $q^-_{01,N} \propto q^+ f_N$ and $q^-_{10,N} \propto q^+ f_N$. The lowest coding level possible f_N is on the order of $\frac{\log(N)}{N}$ and it gives a memory capacity on the order of $\frac{-1}{\log(\lambda_1)} \sim_{N\infty} \frac{1}{f_N^2}$. In [Romani et al., 2008, Amit and Huang, 2010], they assumed that $q^-_{10,N} = 0$ and showed the same result using a Gaussian approximation of the synaptic currents. Under the same assumption as in [Amit and Fusi, 1994] $\left(q^-_{01,N},\ q^-_{10,N} \propto q^+ f_N \text{ and } f_N \to 0\right)$, our result also predicts a forgetting time on the order of $\frac{-1}{\log(\Lambda_0)} \sim_{N\infty} \frac{1}{f_N^2}$, see Theorem 3.18. Moreover, we give a result for depression probabilities not depending on N and our result link the probability of error to the parameters. Note the presence of Λ_0 in our result rather than λ_1 in previous studies. This difference comes from our different measure of memory lifetime. The SNR analysis is based on the convergence of the means of the synaptic currents whereas our retrieval criterion requests the knowledge of their entire distributions. Indeed, we search for a memory lifetime obtained with a control on the errors p_e^0 and p_e^1. We conjecture that we could prove similar result as ours with λ_1 rather than Λ_0. Finally, our results do not necessarily need the large N asymptotic. Nevertheless, in this asymptotic, the expression of $\hat{t}$ simplifies, see Remark 3.19.

In this study, we assume that learning is generated by the divergence of the distributions of the synaptic currents h_t^0 and h_t^1 from their invariant distribution, see Figure 6. The main role of the number of signal presentations (r) is to separate these two distri-

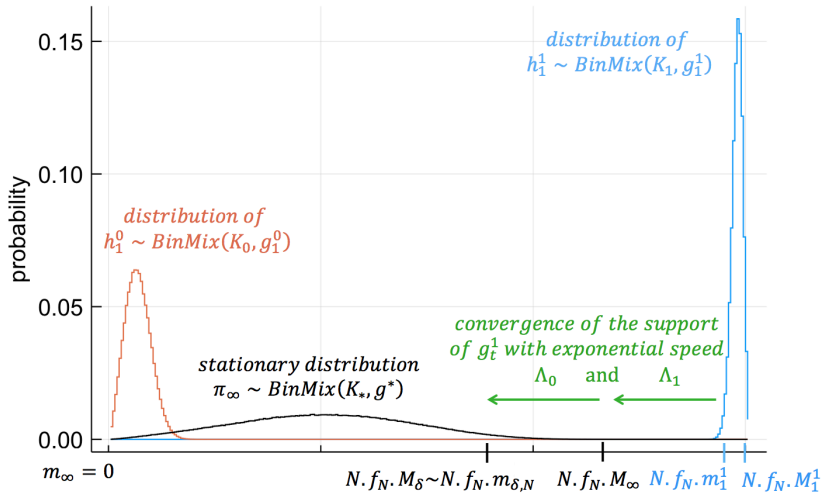


Figure 6: Illustration of the notations. The variables $K_0,\ K_1,\ K_*$ have Binomial laws with parameters N and f_N. They are respectively independent from $h_1^0,\ h_1^1,\ \pi_\infty$.

butions. Indeed, the larger the r, the more separated the support of the mixing distributions g_1^0 and g_1^1 are. In our proofs, we compare g_t^1 to g^* after showing that as long as g_t^1 is far enough from g^* it is far enough from g_t^0, see Lemma 3.13. As a consequence, the expression of $\hat{t}$ is an increasing function of $m_1^1 - M_\infty$, and so of r.

Let us now discuss the roles of the coding level, the potentiation and depression probabilities. They affect both learning and forgetting. The coding level directly affects the number of synapses candidate to depression and potentiation. Indeed, looking at an individual synapse, its probability to potentiate is $f_N^2 q^+$ and its probability to depress is $f_N(1 - f_N)(q_{10}^- + q_{01}^-)$. Thus, when the coding level is close to one, the fluctuations are important and seem to cause a fast forgetting as shown in the illustrations of Section 2. Therefore, we used a low coding level, see Assumption 2.6.2. This choice slows down the forgetting. However, f_N cannot be too small because it is detrimental to the learning phase as the distance between the two conditional distributions depends on f_N. More particularly, it depends on Nf_N which then need to be large enough, see Assumption 2.6.2. The last parameters we can tune are the potentiation and depression probabilities. As for f_N, there is a compromise between their role in learning and forgetting. Indeed, in order to promote learning, they need to be close enough to one but on the contrary, small probabilities reduce the forgetting rate. So we propose to take a potentiation probability (q^+) on the order of 1, to learn quickly, and small depression probabilities, to forget slowly. Potentiation increases the synaptic currents so it leads to a shift of the distribution of h_t^1 to the right and for the same reasons, depression results in a shift of the distribution of h_t^0 to the left. Therefore, smaller depression than potentiation implies that the distribution of h_t^1 is significantly shifted to the right whereas the distribution of h_t^0 is slightly shifted to the left. In view of learning, the initial separation between distributions can be limited if the invariant distribution π_∞ is already concentrated on high values of synaptic currents. As there are two depression probabilities, this situation can be avoided by choosing one probability big enough and the other one smaller. For example, when depression probabilities depend on N under Assumption 2.6.3, both $q_{10,N}^-$ and $q_{01,N}^-$ converge to 0. If they both converge too fast (a_N and b_N converge to 0), the invariant measure is concentrated around one and no learning is possible. However, if either a_N or b_N does not converge to 0, then the invariant measure is not concentrated around 0 and learning is possible. Then, depending on the different large N asymptotic of a_N and b_N, we computed the different memory lifetime summarized in Table 1. The best memory lifetimes are on the order of $\frac{1}{f_N^2}$ and are obtained when a_N (resp. b_N) converges to 0 and b_N (resp. a_N) converges to a constant in $\mathbb{R}^+$ (resp. $\mathbb{R}_*^+$) or (a_N, b_N) converges to constants in $\mathbb{R}_*^+ \times \mathbb{R}^+$. Thus, if one wants to increase the memory lifetime beyond this order, we seem to need a model more complex.

Our study is valid for a classic learning, which needs multiple stimulus presentations, but also 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 need to 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, 1973]. He showed that humans are able to recognize up to 10,000 images, presented only once, with 90 percent success rate.

Many perspectives can be studied as a follow-up to this study. First, the analysis carried out on the synaptic current onto one 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 completed in order to get closer to biology. Indeed, the formation of synaptic memory is far more complex than in our model. 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 [Fusi and Abbott, 2007, Huang and Amit, 2011], whereas meta-plastic transitions brought better SNR results [Fusi et al., 2005, Roxin and Fusi, 2013, Benna and Fusi, 2016]. Adding neural dynamics in such models would be a next challenging step. Nevertheless, the model analysed here illustrates well the trade-off 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. 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 3.4

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

$$\forall u \in \mathbb{R}, \qquad G_{Z,(a,b)}(u) = G_Z\left(\frac{u-b}{a-b}\right).$$

Proposition 3.4 relies on the following

Lemma A.2. *Let Z be a mixture of Binomial $Z = \mathrm{BinMix}(K, Y_Z)$. Let $0 \leq b < a < 1$. Conditionally on Z, consider two independent Binomial distributions $\mathrm{Bin}(Z,a)$ and $\mathrm{Bin}(K-Z,b)$ and define $X = \mathrm{Bin}(Z,a) + \mathrm{Bin}(K-Z,b)$. Then*

$$X \overset{\mathcal{L}}{=} \mathrm{BinMix}\,(K, Y_X) \quad \textit{with} \quad Y_X = (a-b)Y_Z + b. \tag{22}$$

In particular, $G_X(u) = G_{Z,(a,b)}(u)$.

Proof. 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 Remark 3.2, Z is the sum of $(Z_i)_{1\leq i\leq K}$ *i.i.d.* Bernoulli of parameter $Y_Z = G_Z^{-1}(\tilde{U})$. Thus, we obtain that conditionally on Z,

$$X = \underbrace{\sum_{i=1}^{K} Z_i \mathbb{1}_{\{\xi_i \leq a\}}}_{\stackrel{\mathcal{L}}{=}\mathrm{Bin}(Z,a)} + \underbrace{\sum_{i=1}^{K} (1-Z_i) \mathbb{1}_{\{\eta_i \leq b\}}}_{\stackrel{\mathcal{L}}{=}\mathrm{Bin}(K-Z,b)}$$

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

$$X = \sum_{i=1}^{K} \mathbb{1}_{\left\{U_i \leq G_Z^{-1}(\tilde{U})\right\}} \mathbb{1}_{\{\xi_i \leq a\}} + \sum_{i=1}^{K} \mathbb{1}_{\left\{U_i > G_Z^{-1}(\tilde{U})\right\}} \mathbb{1}_{\{\eta_i \leq b\}}.$$

$$\text{So} \quad X = \sum_{i=1}^{K} \mathbb{1}_{\left\{U_i \leq G_Z^{-1}(\tilde{U}), \xi_i \leq a\right\} \bigcup \left\{U_i > G_Z^{-1}(\tilde{U}), \eta_i \leq b\right\}}. \tag{23}$$

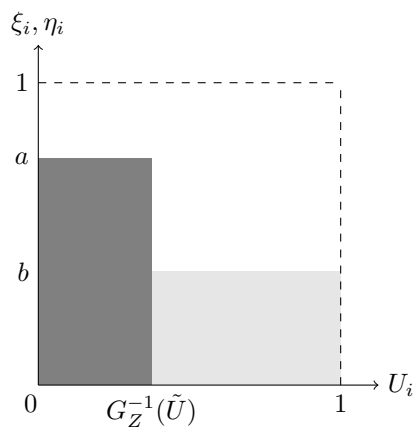


Figure 7: In gray, the domain to which the couple (U_i, ξ_i, η_i) needs to belong to from the equation (23).

For all Borel set $\mathcal{D} \subset [0,1]^3$, $\mathbb{P}\left((U_i, \xi_i, \eta_i) \in \mathcal{D}\right) = \mathcal{V}(\mathcal{D})$ where $\mathcal{V}(\mathcal{D})$ is the volume of $\mathcal{D}$. Thus, let $W_i \stackrel{\mathcal{L}}{=} \mathcal{U}([0,1])$, then $\mathbb{P}\left((U_i, \xi_i, \eta_i) \in \mathcal{D}\right) = \mathbb{P}\left(W_i \leq \mathcal{V}(\mathcal{D})\right)$. We put ξ_i and η_i on the same axis as they do not depend one on the other so that the volume $\mathcal{V}\left(\left\{U_i \leq G_Z^{-1}(\tilde{U}), \xi_i \leq a\right\} \bigcup \left\{U_i > G_Z^{-1}(\tilde{U}), \eta_i \leq b\right\}\right)$ is equal to the sum of the tow grey areas (see Figure 7). We deduce that

$$X = \sum_{i=1}^{K} \mathbb{1}_{\left\{W_i \leq b + (a-b) G_Z^{-1}(\tilde{U})\right\}} = \sum_{i=1}^{K} \mathbb{1}_{\left\{G_Z\left(\frac{W_i - b}{a-b}\right) \leq \tilde{U}\right\}} = \sum_{i=1}^{K} \mathbb{1}_{\left\{G_X(W_i) \leq \tilde{U}\right\}},$$

with $G_X(w) = G_{Z,(a,b)}(w)$. We conclude that (22) is satisfied. $\square$

Proof of Proposition 3.4.

Proof. We first show (7) and (9) for $h_{t,K}$, then the rest follows.

At $t = 1$, from equation (5) we get

$$\mathcal{L}\left(h_{1,K}|\xi_0^1 = 1, h_{-r+1,K}\right) = \text{Bin}\left(h_{-r+1,K}, 1\right) + \text{Bin}\left(K - h_{-r+1,K}, 1 - (1-q^+)^r\right)$$
$$\mathcal{L}\left(h_{1,K}|\xi_0^1 = 0, h_{-r+1,K}\right) = \text{Bin}\left(h_{-r+1,K}, (1-q_{01}^-)^r\right).$$

Applying twice Lemma A.2 with $(a,b) = (1, 1-(1-q^+)^r)$ and then $(a,b) = (1-(1-q_{01}^-)^r, 0)$, we obtain using notation A.1

$$\mathcal{L}\left(h_{1,K}|\xi_0^1 = 1, h_{-r+1,K}\right) \overset{\mathcal{L}}{=} \text{BinMix}\left(K, g_{-r+1,(1,1-(1-q^+)^r)}\right)$$
$$\mathcal{L}\left(h_{1,K}|\xi_0^1 = 0, h_{-r+1,K}\right) \overset{\mathcal{L}}{=} \text{BinMix}\left(K, g_{-r+1,(1-(1-q_{01}^-)^r,0)}\right).$$

Thus,

$$\begin{aligned}
&\mathbb{P}\left(h_{1,K} = j|h_{-r+1,K}\right)\\
=&\mathbb{P}\left(\xi_0^1 = 1\right)\mathbb{P}(h_{1,K} = j|\xi_0^1 = 1, h_{-r+1,K}) + \mathbb{P}\left(\xi_0^1 = 0\right)\mathbb{P}(h_{1,K} = j|\xi_0^1 = 0, h_{-r+1,K})\\
=&\binom{K}{j}\int_0^1 u^j(1-u)^{K-j}\left(f_N g_{-r+1,(1,1-(1-q^+)^r)}(du) + (1-f_N)g_{-r+1,(1-(1-q_{01}^-)^r,0)}(du)\right),
\end{aligned}$$

which enables to get (7).

Now, assume that $h_{t,K} \overset{\mathcal{L}}{=} \text{BinMix}(K, g_t)$, for some fixed $t \geq 1$. Then, by equation (6)

$$\mathcal{L}\left(h_{t+1,K}|\xi_t^1 = 1,\ h_{t,K}\right) = \text{Bin}\left(K - h_{t,K}, f_N q^+\right) + \text{Bin}\left(h_{t,K}, 1-(1-f_N)q_{10}^-\right),$$
$$\mathcal{L}\left(h_{t+1,K}|\xi_t^1 = 0,\ h_{t,K}\right) = \text{Bin}\left(h_{t,K}, 1 - f_N q_{01}^-\right),$$

where Binomials are independent conditionally on $h_{t,K}$. Applying twice Lemma A.2 with $(a,b) = (1-(1-f_N)q_{10}^-, f_N q^+)$ and $(a,b) = (1 - f_N q_{01}^-, 0)$, we get

$$\mathcal{L}\left(h_{t+1,K}|\xi_t^1 = 1\right) \overset{\mathcal{L}}{=} \text{BinMix}\left(K, g_{t,(1-(1-f_N)q_{10}^-, f_N q^+)}\right)$$
$$\mathcal{L}\left(h_{t+1,K}|\xi_t^1 = 0\right) \overset{\mathcal{L}}{=} \text{BinMix}\left(K, g_{t,(1-f_N q_{01}^-,0)}\right).$$

Hence, $h_{t+1,K} \overset{\mathcal{L}}{=} \text{BinMix}\left(K, f_N g_{t,(1-(1-f_N)q_{10}^-, f_N q^+)} + (1-f_N)g_{t,(1-f_N q_{01}^-,0)}\right)$, and we deduce that $h_{t+1,K} \overset{\mathcal{L}}{=} \text{BinMix}(K, g_{t+1})$ with $G_{t+1}(x) = \mathcal{R}(G_t)(x)$.

For the processes $\left(h_{t,K}^y\right)_{t\geq 0}$, we proceed exactly with the same method with the fact that $\xi_0^1 = y$ in Proposition 2.2. $\square$

A.2 Proof of Proposition 3.6

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

Let $\Gamma_1, \Gamma_2 \in F_{[0,1]}$. We recall that $\Lambda_1 = 1 - (1-f_N)q_{10}^- - f_N q^+$, $\Lambda_0 = 1 - f_N q_{01}^-$.

$$\begin{aligned}
&\|\mathcal{R}(\Gamma_2) - \mathcal{R}(\Gamma_1)\|_{L^1(0,1)} \\
&\leq \int_0^1 f_N \left|\Gamma_2\left(\frac{u - f_N q^+}{\Lambda_1}\right) - \Gamma_1\left(\frac{u - f_N q^+}{\Lambda_1}\right)\right| + (1-f_N)\left|\Gamma_2\left(\frac{u}{\Lambda_0}\right) - \Gamma_1\left(\frac{u}{\Lambda_0}\right)\right| du \\
&= f_N \int_{f_N q^+}^{1-(1-f_N)q_{10}^-} \left|\Gamma_2\left(\frac{(u - f_N q^+)}{\Lambda_1}\right) - \Gamma_1\left(\frac{(u - f_N q^+)}{\Lambda_1}\right)\right| du \\
&\qquad + (1-f_N)\int_0^{\Lambda_0} \left|\Gamma_2\left(\frac{u}{\Lambda_0}\right) - \Gamma_1\left(\frac{u}{\Lambda_0}\right)\right| du \\
&= f_N\Lambda_1 \int_0^1 |\Gamma_2(u) - \Gamma_1(u)|\, du + (1-f_N)\Lambda_0 \int_0^1 |\Gamma_2(u) - \Gamma_1(u)|\, du \\
&= \underbrace{(f_N\Lambda_1 + (1-f_N)\Lambda_0)}_{\lambda_1} \|\Gamma_2 - \Gamma_1\|_{L^1(0,1)}.
\end{aligned}$$

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

2. Existence and uniqueness 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 Γ can be chosen in $F_{[0,1]}$. First, any limit Γ 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 $\mathcal{R}(G^*) = G^*$. Finally, the uniqueness of G^* is deduced from the fact that $\mathcal{R}$ is strictly contracting. □

A.3 Proof of Lemma 3.14

Proof. We use the method of [Chernoff, 1952]. Let S_N be the sum of $X_1, X_2, \cdots, X_N$ which are independent Bernoulli random variables of parameter p.
For all $\varepsilon \in (0,1), u \in \mathbb{R}^+$,

$$\begin{aligned}
\mathbb{P}(S_N \geq Np(1+\varepsilon)) = \mathbb{P}\left(\mathrm{e}^{uS_N} \geq \mathrm{e}^{Np(1+\varepsilon)u}\right) &\leq \frac{\mathbb{E}\left(\mathrm{e}^{uS_N}\right)}{\mathrm{e}^{Np(1+\varepsilon)u}} = \frac{\prod_{i=1}^N \mathbb{E}\left(\mathrm{e}^{uX_i}\right)}{\mathrm{e}^{Np(1+\varepsilon)u}} \\
&\leq \frac{(1+p(\mathrm{e}^u - 1))^N}{\mathrm{e}^{Np(1+\varepsilon)u}} \leq \frac{\mathrm{e}^{Np(\mathrm{e}^u-1)}}{\mathrm{e}^{Np(1+\varepsilon)u}} = \mathrm{e}^{Np(\mathrm{e}^u - 1 - (1+\varepsilon)u)}.
\end{aligned}$$

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

$$\mathbb{P}(X > Np(1+\varepsilon)) \leq \left(\frac{\mathrm{e}^\varepsilon}{(1+\varepsilon)^{1+\varepsilon}}\right)^{Np} = \exp\left(Np\big(\varepsilon - (1+\varepsilon)\log(1+\varepsilon)\big)\right).$$

From the inequality, $\forall z > 0,\ \log(1+z) \geq \frac{2z}{2+z}$, we obtain (12). In order to show (13), we proceed with the same method and use the inequality $\log(1+z) \geq \frac{z}{2}\frac{2+z}{1+z}$ whenever $-1 < z \leq 0$. □

Acknowledgements

I am indebted to the help from Etienne Tanré and Romain Veltz.
I am very grateful to two anonymous referees for valuable comments and suggestions which really helped improving the paper.
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

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

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

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

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

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

[Chernoff, 1952] Chernoff, H. (1952). 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.

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

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

[Fusi and Abbott, 2007] Fusi, S. and Abbott, L. (2007). Limits on the memory storage capacity of bounded synapses. *Nature neuroscience*, 10(4):485–493.

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

[Gardner and Derrida, 1988] Gardner, E. and Derrida, B. (1988). Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and general*, 21(1):271.

[Hopfield, 1982] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558.

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

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

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

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

[Sommer and Dayan, 1998] Sommer, F. and Dayan, P. (1998). Bayesian retrieval in associative memories with storage errors. *IEEE Transactions on Neural Networks*, 9(4):705–713.

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

[Willshaw et al., 1969] Willshaw, D. J., Buneman, O. P., and Longuet-Higgins, H. C. (1969). Non-Holographic Associative Memory. *Nature*, 222:960.

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