A mathematical approach on memory capacity of a simple synapses model

Pascal Helson

ICMNS 2018 - ANTIBES

A MATHEMATICAL APPROACH ON MEMORY CAPACITY OF A SIMPLE SYNAPSES MODEL

PASCAL HELSON

INRIA TEAMS: TOSCA & MATHNEURO MAIL: PASCAL.HELSON@INRIA.FR

1. FRAMEWORK

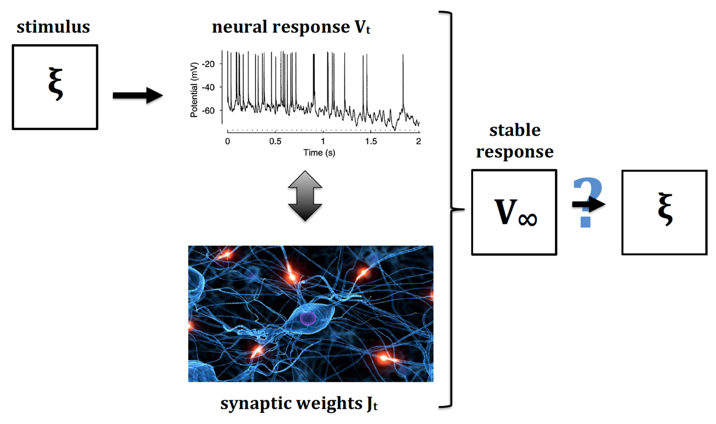
Network models of Memory: Capacity of neural networks in memorising external inputs is a complex problem which has given rise to numerous research. It is widely accepted that memory sits where communication between two neurons takes place, in synapses [1]. It involves a huge number of chemical reactions, cascades, ion flows, protein states and even more mechanisms, which makes it really complex. Such a complexity stresses the need of simplifying models: this is done in network models of memory.

Problem: Most of these models don't take into account both synaptic plasticity and neural dynamic. Adding dynamics on the weights makes the analysis more difficult which explains that most models consider either a neural [2, 3, 4] or a synaptic weight dynamic [5, 6, 7, 8]. We decided to study the binary synapses model of [9], model we wish to complete with a neural network afterwards in order to get closer to biology.

Purpose: Propose a rigorous mathematical approach of the model of [9] as part of a more ambitious aim which is to have a general mathematical framework adapted to many models of memory.

2. NETWORK MODELS OF MEMORY

Three main ingredients describes such models. A stimulus has direct effect on neurons which then modify the synaptic weight matrix leading to a stable response of the network possessing the information sent by the stimulus.



3. AMIT-FUSI MODEL [9]

Discrete time model with two coupled **binary** processes, stimuli $(\xi_t)_{t\geq 0} \in \{0,1\}^N$ and synaptic weight matrix $(\mathbf{J_t})_{t\geq 0} \in \{0,1\}^{N^2}$:

Stimuli: $(\xi_t)_{t\geq 0}$ i.i.d. random vectors $\sim Bernoulli(\mathbf{f})^{\otimes N}$

Synaptic weights dynamic:
$\forall i\ J_t^{ii} = 0$ and at each time step a new stimulus ξ_t is received by the network. The components J_t^{ij}, $i \neq j$, jump as follows:

- if $\left(\xi_t^i, \xi_t^j\right) = (1,1)$, $J_t^{ij} = 0 \to J_{t+1}^{ij} = 1$ with probability q^+,
- if $\left(\xi_t^i, \xi_t^j\right) = (0,1)$, $J_t^{ij} = 1 \to J_{t+1}^{ij} = 0$ with probability q_{01}^-,
- if $\left(\xi_t^i, \xi_t^j\right) = (1,0)$, $J_t^{ij} = 1 \to J_{t+1}^{ij} = 0$ with probability q_{10}^-,
- if $\left(\xi_t^i, \xi_t^j\right) = (0,0)$, $J_{t+1}^{ij} = J_t^{ij}$.

Remark: The initial condition is not defined on synaptic weights but on the synaptic input into neurons defined as follows.

Synaptic input into neurons:
$\left(h_t^i\right)_{t\geq 0}$ is the field induced by ξ_0 presented at time t, in neuron i:

$$h_t^i = \sum_{j\neq i} J_t^{ji}\xi_0^j$$

In the following, we are interested in the laws of $\left(h_t^i|\xi_0^i = 0\right)$ and $\left(h_t^i|\xi_0^i = 1\right)$, respectively called p_t^0 and p_t^1.

Remark: The state space of h_t^i depends on the size of ξ_0: $K = \sum_{j\neq i}\xi_0^j$. Moreover, as neurons are similar we use notation $h_{t,K}^1$. Finally, it is easy to show $p_{t,K}^i$ converges to a unique $p_{\infty,K}$.

Initial condition:
Initially, synaptic input follows the stationary distribution:

$$h_{0,K}^1 \sim p_{\infty,K}$$

4. RETRIEVAL CRITERIA

Many methods have been used to study the storage capacity of network models. The more intuitive is maybe to see stimuli to be learned as attractors of a neural dynamic [3]: the maximal number of attractors would then be the memory capacity of the model. Signal to Noise Ratio (SNR) analysis [8, 9] and mean first passage time to a threshold [7] have also been proposed. The underlying idea of these methods is that the neural dynamic is ruled by a threshold on the synaptic input: a linear decision rule. In our case, we don't impose such a rule. Our retrieval criteria holds on the knowledge of the two distributions $p_{t,K}^0$ and $p_{t,K}^1$.

Decision rule [10]:
At fixed K, we aim at studying the minimal probability of error assuming the neuron 1 knows p_t^i and observes h_t^1. As $p_{t,K}^1$ and $p_{t,K}^0$ converges to $p_{\infty,K}$, the error increases with time as distributions get closer:

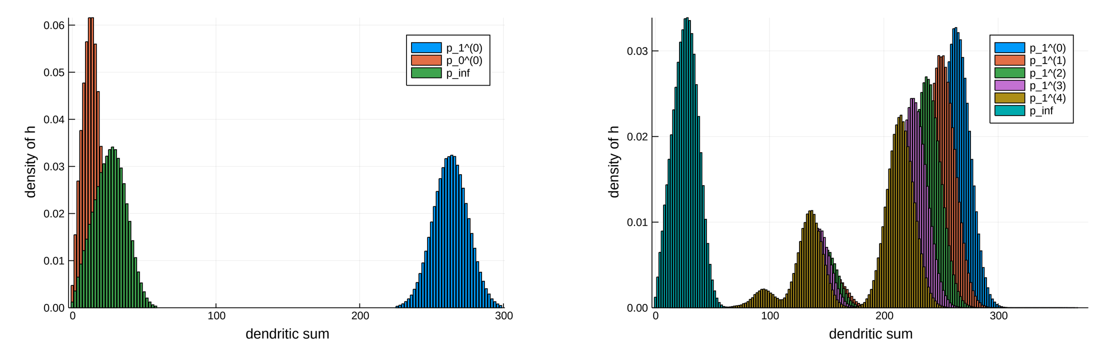


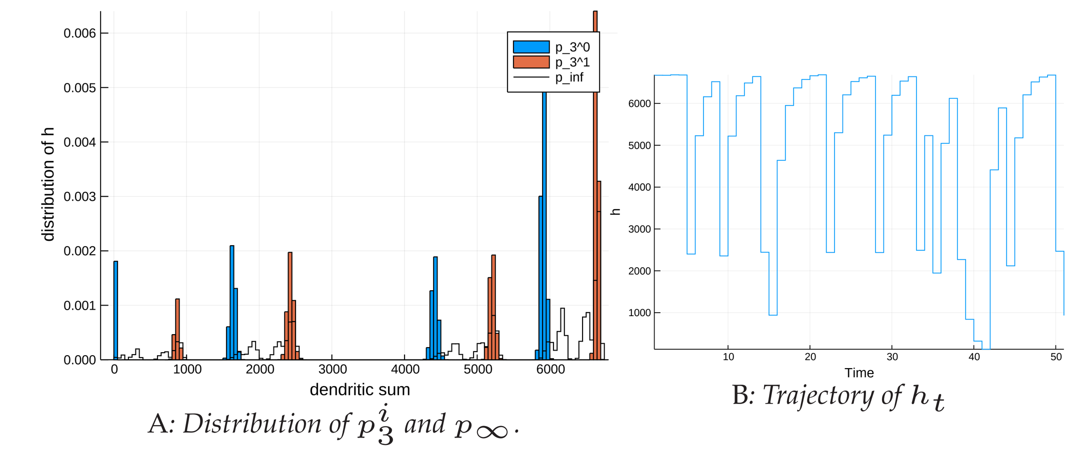
Fig: Distributions p_t^i when $f = 0.1$ and parameters $q_x^{+/-} = 0.5$. At time $t = 0$, the two distributions are well separated (left) and then get closer(right).

As long as the probability of error, $P_e(t)$ defined below, is less than a given ϵ, ξ_0^1 is considered to be retrievable from h_t^1 through the following decision rule. We are interested in the maximal time for which such a condition is achieved, and in particular we would like to know its dependence on the different parameters. We then define $P_e(t)$. Let $G = \{g : [\![0,N]\!] \to \{0,1\}\}$:

$$P_e(t) = \inf_{g\in G}\left\{\mathbb{P}\left(g(h_t^1) \neq \xi_0^1\right)\right\}$$

$$= \inf_{g\in G}\underbrace{\left\{f\mathbb{P}\left(g(h_t^1) = 0|\xi_0^1 = 1\right) + (1-f)\mathbb{P}\left(g(h_t^1) = 1|\xi_0^1 = 0\right)\right\}}_{L(t,g)}$$

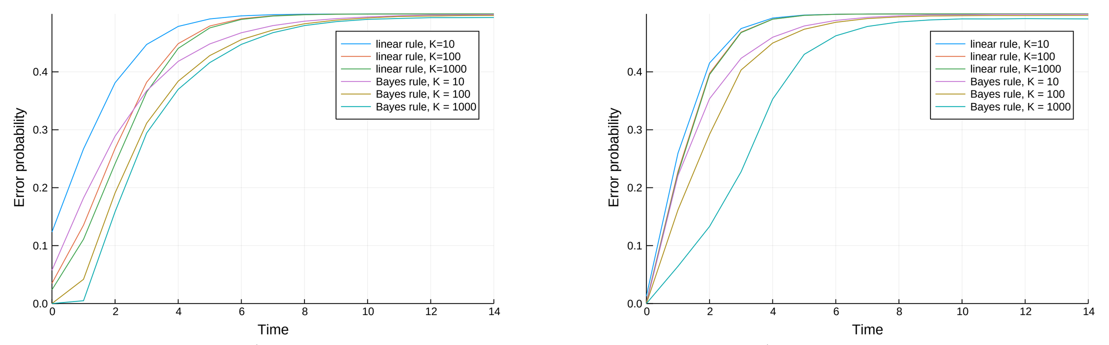
Unlike the SNR analysis which requires only the first two moments of the variable h_t^1, here we need the knowledge of both $p_{t,K}^1$ and $p_{t,K}^0$. Although it is more costly, it is always valid, which is not the case of SNR as it requires p_t^0 and p_t^1 to be approximately Gaussian and this is not the case for some parameters:



A: Distribution of p_3^1 and p_∞. B: Trajectory of h_t

Fig: Parameters are the following: $f = 0.5$, $q^+ = 0.9$, $q_{10}^- = 0.1$, $q_{01}^- = 0.9$

We see the difference between the two rules on the probability of error:



A: $f = 0.5 = q^+ = q_{10}^- = q_{01}^-$. B: $q^+ = q_{01}^- = 0.9$, $q_{10}^- = 0.1$, $f = 0.5$.

Fig: Probability of error in function of linear (SNR) and Bayes (the one we use) rules and parameters.

SNR analysis:
In [9], they propose an analysis based on $SNR_t = \frac{S_t^2}{R_t}$:

- Signal: $S_t = \mathbb{E}\left[h_t^1 | \xi_0^1 = 1\right] - \mathbb{E}\left[h_t^1 | \xi_0^1 = 0\right]$
- Noise: $R_t = \mathrm{Var}\left[h_t^1\right]$

Because of correlation between synapses dynamics, SNR is only computed in a specific case leading to capacity $P = \max\{t \in \mathbb{N}, SNR_t > \log(N)\}$:

$$f = \frac{\log(N)}{N},\ q_{01}^-, q_{10}^- \propto fq^+ \Rightarrow P \propto \left(\frac{N}{\log(N)}\right)^2 \quad (1)$$

5. FIRST RESULTS

Mathematical results:
As having the general forms of $p_{t,K}^1$ and $p_{t,K}^0$ is difficult, we first studied the spectrum of the transition matrix $M_{h,K}$ of $(h_t^1)_{t\geq 0}$ and got a first result:

Proposition 1 *The spectrum of the transition matrix $M_{h,K}$ and the one of $\left(\xi_t, \left(J_t^{j1}\right)_{1\leq j\leq K}\right)_{t\geq 0}$, $M_{\xi,J,K}$ is the following:*

$$\Sigma\left(M_{h,K}\right) = \{\mu_i = (1-f)\underbrace{(1 - fq_{01}^-)}_{\lambda_0}{}^i + f\underbrace{(1-(1-f)q_{10}^- - fq^+)}_{\lambda_1}{}^i,\ 0 \leq i \leq K\}$$

$\Sigma\left(M_{\xi,J,K}\right) = \Sigma\left(M_{h,K}\right) \cup \{0\}$, *multiplicity* $\binom{K-1}{i}$ *for* μ_i, 2^K *for* 0

In fact, the spectrum is linked to the speed at which stimuli are forgotten. However, the slower this speed is, the less plastic the network is. It is a classical compromise in optimising storage capacity.

Sketch of the proof for $\Sigma\left(M_{\xi,J,K}\right)$

We can write $M_{\xi,J,K}$ as a matrix by block with $p_\xi = \mathbb{P}(\xi_t = \xi)$ and M_ξ the probability matrix of $(J_t^{j1})_j$ knowing that $\xi_t = \xi$:

$$M_{\xi,J,K} = \underbrace{\begin{bmatrix} p_{\xi_1}M_{\xi_1} & p_{\xi_2}M_{\xi_1} & \cdots & p_{\xi_{2^K}}M_{\xi_1} \\ p_{\xi_1}M_{\xi_2} & p_{\xi_2}M_{\xi_2} & \cdots & p_{\xi_{2^K}}M_{\xi_2} \\ \vdots & \vdots & & \vdots \\ p_{\xi_1}M_{\xi_{2^K}} & p_{\xi_2}M_{\xi_{2^K}} & \cdots & p_{\xi_{2^K}}M_{\xi_{2^K}} \end{bmatrix}}_{2^{2K-1}\times 2^{2K-1}\ matrix}$$

It is not difficult to show

$$\Sigma\left(M_{\xi,J,K}\right) = \Sigma\left(M_{J,K} = \sum_{k=1}^{2^K} p_{\xi_k}M_{\xi_k}\right) \cup \{0\}$$

In particular, if π is an invariant measure of the process with matrix transition $M_{J,K}$, $\pi M_{J,K} = \pi$, then $\pi_K = \left[p_{\xi_1}\pi\ p_{\xi_2}\pi \ldots p_{\xi_{2^K}}\pi\right]$ is an invariant measure for $\left(\xi_t, \left(J_t^{j1}\right)_{1\leq j\leq K}\right)_{t\geq 0}$. One can then compute M_ξ from the following 2×2 matrices:

$M_{00} = I_2$, $M_{01} = \begin{bmatrix} 1 & 0 \\ q_{01}^- & 1 - q_{01}^- \end{bmatrix}$, $M_{10} = \begin{bmatrix} 1 & 0 \\ q_{10}^- & 1 - q_{10}^- \end{bmatrix}$, $M_{11} = \begin{bmatrix} 1-q^+ & q^+ \\ 0 & 1 \end{bmatrix}$

Then, using Kronecker product properties, we have the lemma:

Lemma 1 *With the notation* $\otimes_N M = \underbrace{M \otimes M \otimes \ldots \otimes M}_{N\,times}$,

$$M_{J,K} = (1-f)\otimes_{K-1}\underbrace{((1-f)M_{00} + fM_{01})}_{M_0} + f\otimes_{K-1}\underbrace{((1-f)M_{10} + fM_{11})}_{M_1}$$

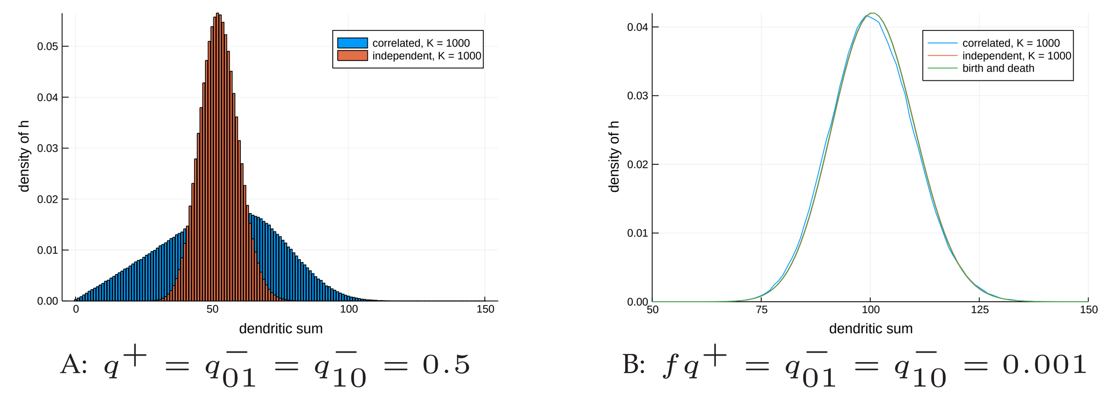
We conclude on $\Sigma\left(M_{\xi,J,K}\right)$ using $v_0 = [1\ 1]^T$ and $e_2 = [0\ 1]^T$:
$M_0v_0 = v_0 = M_1v_0$, $M_0e_2 = \lambda_0e_2$ and $M_1e_2 = \lambda_1e_2 + fq^+v_0$

Let $u_{i,K} = (u_{i,K}^1, \ldots, u_{i,K}^{\binom{K}{i}})$ vectors which can be written as the Kronecker product of i vectors e_2 and $(K-i-1)$ v_0, $u_{i,K}^1 = \otimes_i e_2 \otimes_{K-1-i} v_0$, then:

$$M_{J,K}u_{i,K}^j = \left(\underbrace{(1-f)(\lambda_0)^i + f(\lambda_1)^i}_{\mu_i}\right)u_{i,K}^j + \sum_{k=0}^{i-1}\sum_l \alpha_{k,l}u_{k,K}^l$$

In this basis, $M_{J,K}$ is triangular superior with μ_i on the diagonal with the multiplicity $\binom{K-1}{i}$, it ends the proof on $\Sigma\left(M_{\xi,J,K}\right)$.

Simulation for shaping intuition:
Thanks to simulations, we have a look to the case f and q small:



A: $q^+ = q_{01}^- = q_{10}^- = 0.5$ B: $fq^+ = q_{01}^- = q_{10}^- = 0.001$

Fig: Invariant distributions when $f = 0.1$ and $K = 1000$.

The independent case is the model considering every J_t^{ij} evolves independently following the dynamic of one synapse in the model defined in 3. We can see in simulations that the behaviour of h_t^1 is similar to the one in the independent case when f and q are small enough. Moreover, this model of independent synapses leads to similar results as (1) for the SNR analysis. Finally, under the assumption that synapses evolve independently, π_t^i, the equivalent of p_t^i in the previous model, are binomial laws. In fact, J_t^{ij} converges in law to the invariant distribution $\pi = (\pi^-, \pi^+)$ with speed $\lambda^t = \left(1 - f^2q^+ - f(1-f)(q_{01}^- + q_{10}^-)\right)^t$. Hence, we get the following results thanks to the generating function:

$$\pi_{t,K}^0 \sim \mathcal{B}\left(K, \pi^+ - \pi^+q_{01}^-\lambda^t\right),\quad \pi_{t,K}^1 \sim \mathcal{B}\left(K, \pi^+ + \pi^-q^+\lambda^t\right)$$

Therefore, it seems to be possible to extend the result (1) for a larger range of parameters f and q thanks to the similarities between the independent model and the one presented in 3 when f and q are small. However, such a similarity would reduce the model to a too simple model loosing its initial interest.

6. CONCLUSION

When previous studies have considered small coding level f in order to get results on the storage capacity depending on N, such an assumption seems to make the model loose its initial interest sited in the correlations between synapses. Our approach aims at studying the synaptic input into a neuron taking correlations into account through a decision rule. It enabled us to get a first result on the speed of forgetting stimuli thanks to a spectral analysis.

7. PERSPECTIVES

- Show $p_{\infty,K}$ converges to $\pi_{\infty,K}$ when f, q are small
- Control the probability of error thanks to the study in the independent case: f, q small
- A study of the probability of error taking into account all the vector h_t
- Add neural dynamics as a feedback to maintain weights structure longer and enhance storage capacity

REFERENCES

[1] G. Mongillo, O. Barak, and M. Tsodyks. Longuet-Higgins. Synaptic Theory of Working Memory. In *Science '08*

[2] D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins. Non-Holographic Associative Memory. In *Nature '69*

[3] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the national academy of sciences '82*

[4] V. Gripon, J. Heusel, M. Lwe, and F. Vermet. A Comparative Study of Sparse Associative Memories. In *Journal of Statistical Physics '16*

[5] N. Brunel and F. Carusi. Slow stochastic Hebbian learning of classes in recurrent neural networks In *Network: Computation in Neural Systems '98*

[6] S. Fusi and L. F. Abbott. Limits on the memory storage capacity of bounded synapses. In *Nature Neuroscience '07*

[7] T. Elliott. Memory Nearly on a Spring: A Mean First Passage Time Approach to Memory Lifetimes. In *Neural Computation '14*

[8] M. K. Benna and S. Fusi. Computational principles of synaptic memory consolidation. In *Nature Neuroscience '16*

[9] D. J. Amit and S. Fusi. Learning in Neural Networks with Material Synapses. In *Neural Computation '94*

[10] L. Devroye, L. Györfi, and G. Lugosi. A probabilistic theory of pattern recognition. In *Springer '96*