III-59
Cosyne14

Input-driven unsupervised learning in recurrent neural networks

Alireza Alemi-Neissi[1], Carlo Baldassi[1,2], Nicolas Brunel[3,4], Riccardo Zecchina[1,2]

[1]HuGeF, Torino, Italy [2]Politecnico di Torino, Torino, Italy [3]Department of Statistics & [4]Department of Neurobiology, University of Chicago, USA

Introduction

Understanding the theoretical foundations of how memories are encoded and retrieved in neural populations is a central challenge in neuroscience. A popular theoretical scenario for modeling memory function is the notion of attractors in a recurrent neural network [1,2].

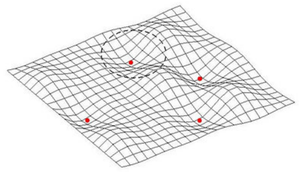


Figure : A cartoon of attractors and its basin of attraction in energy landscape.

The Hebbian learning is one candidate for learning attractors (memory patterns) in a recurrent network [1,2]. It is known [3] that the number of attractors (M) that can be stored with this learning rule in a binary recurrent network (of N binary neurons) is in the order of 0.138 times the number of neurons i.e. $\alpha = 0.138$ where $\alpha = \frac{M}{N}$, which is far from the maximum storage capacity that could be achieved by any learning rule in attractor neural networks (Gradner bound $\alpha = 2$) [4]. While there is no unsupervised learning rule for storing random patterns close to the maximum capacity in recurrent networks (except in the sparse coding limit [5]), the supervised perceptron learning rule achieves the maximum capacity.

Here, we propose an input-driven unsupervised learning rule for storing long-term memory in a recurrent neural network which is inspired by the perceptron learning rule that does not suffer from the drawbacks of Hebbian learning and reaches close to the maximum storage capacity.

Overview of the model

Our **goals**:

- To come up with **a learning rule** for a recurrent neural network
- It should be able to store memories close to the **maxial storage capacity**
- It should implement basic **biological constraints**
- To have **stable dynamics**.

Two crucial features from theoretical perspective to achieve the goals:

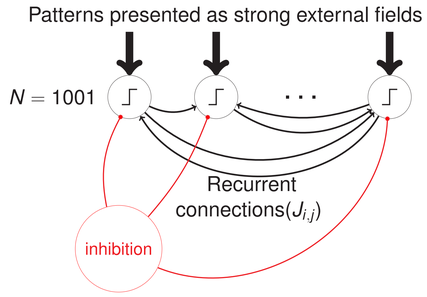
- **Strong external input (or external field) to each neuron**
- **Three learning thresholds for potentiation or depression**

Note: These features of the learning rule could be implemented in networks composed of any model neurons.

We address two questions:

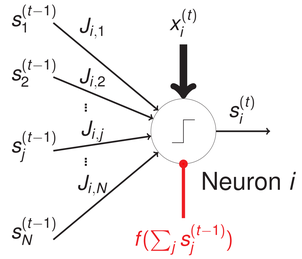
- **How many patterns** can be stored in a recurrent network of N neurons?
- **How strong the external field** should be to store the patterns?

The structure of the model



- Each neuron i receives an external binary input ($x_i \in \{0, X\}$) and N inputs from other neurons ($s_j \in \{0,1\}; j = 1, \ldots, N$).
- Synaptic strengths (J_{ij}) are continuous and non-negative ($J_{ij} \in [0, +\infty)$ & $J_{ii} = 0$).

The dynamics of the model



$$s_i^t = \Theta(v_i^t - \theta),$$
$$v_i^t = \sum_{j=1}^{N} (J_{ij} s_j^{(t-1)}) + x_i^t - I - \lambda(S^{(t-1)} - D_0),$$
$$S^{(t-1)} = \sum_{i=1}^{N} s_i^{(t-1)}$$

- s_i^t and v_i^t are the state and the local field of neuron i
- $x_i^t \in \{0, X\}$ the external field to neuron i at time t
- The strength of external field X was set to γN
- The parameters I, λ, D_0 are inhibition parameters (set such that the network remains stable)
- Θ is the Heaviside step function; θ is the threshold set to $0.35N$
- The dynamics was simulated with synchronous updating
- As a result the network avoided the two trivial dynamical fixed-points (all neurons zero or all neurons one)

The learning rule

The learning rule that is used to store M patterns ($M = \alpha N$) is implemented by comparing the value of local field with three learning thresholds:

– If $\theta_0 < v_i^t < \theta \Rightarrow$ depress active synapses ($J_{ij}^t = J_{ij}^{t-1} - \eta$)

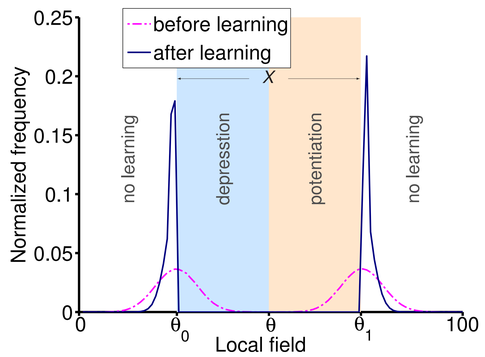
– If $\theta < v_i^t < \theta_1 \Rightarrow$ potentiate active synapses ($J_{ij}^t = J_{ij}^{t-1} + \eta$)

– Otherwise $\Rightarrow$ change nothing

- Applied only to the excitatory-to-excitatory recurrent connections.
- The patterns chosen at random at 0.5 coding probability.
- $\theta = 0.35N$, $\theta_0 = \frac{1}{2}\overline{J}N - I$ and $\theta_1 = \theta_0 + X$ where $\overline{J}$ is the initial average of connection weights and I is an inhibition constant.
- The learning rate was $\eta = 0.01$.
- To add robustness to the learning rule: $\theta_0' = \theta_0 - \phi$ and $\theta_1' = \theta_1 + \phi$. The auxiliary parameter ϕ can be tuned to set a trade-off between capacity and robustness.

Results

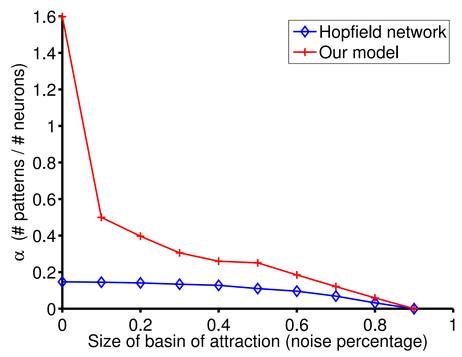
A



A The histogram of local field values of all of the neurons before and after learning. Local field values after learning are pushed away from the potentiation (light-orange) and depression (light-blue) regions due to the learning rule.

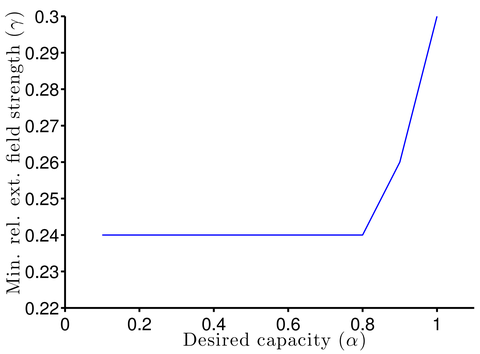
Results (Continued)

B



B The capacity of the network ($\alpha = M/N$) versus the size of basin of attraction for our model and a Hopfield network. The result of simulations with $N = 1001$, $\gamma = X/N = 0.5$, $\eta = 0.01$. ϕ was optimized such that for a given size of basin of attraction, the highest number of patterns could be stored. To obtain the critical storage capacity for each basin size, we initiated the network with patterns corrupted with a given noise level (this noise level was defined as the fraction of the units' states whose state was drawn randomly and independently from the pattern) and then determined the number of successful recalls of the patterns.

C



C Minimum γ (relative strength of external field) versus α. This curve is computed for a network with $N = 1001$ at a fixed $\phi = 0.05\theta_0$ and at zero basin of attraction. We simulated a range of values for gamma then we measured the storage capacity. Afterwards, we picked the minimum gamma for each storage capacity (10 different seeds). At $\gamma \sim 0.24$ the capacity is around $\alpha \sim 0.8$. To store more patterns one needs to increase gamma.

Discussion

- We proposed an unsupervised learning rule for storing long-term memory in recurrent neural networks.
- Our model can store close to the critical storage capacity (in our model $\alpha \sim 1.6$) whereas a Hopfield network with Hebbian learning rule is not able to go beyond $\alpha \sim 0.138$, therefore we achieve more than **11-fold improvement** at zero size basin of attraction.
- **Stable and robust** dynamics.
- The activity level of the network is **stable and robust**.
- The storage capacity reduces by lowering the strength of the external fields. Below $\gamma \sim 0.24$ the network cannot learn all the patterns perfectly.
- The learning rule can be implemented for any neuron models.
- In the **sparse coding limit**, the unsupervised, covariance rule reaches the Gardner bound [5]. Therefore, one expects that the benefit of these learning thresholds should decrease in that limit.
- Our proposed learning rule implements basic biophysical constraints: it uses only the local information available to a neuron and its synapses (i.e. **locality**), and it can store a new pattern independently of previously learned patterns (i.e. **incrementality**).
- The network contains separate **excitatory** or **inhibitory** units, i.e. synaptic strengths do not change sign (in contrast with the Hopfield model).
- **No explicit error signal**: neurons do not need to have access to an explicit error signal from their output, i.e. the difference between the desired output and the current output, but they can infer that information with high reliability by exploiting the statistical properties of the distribution of the local fields.
- The learning rule is **in agreement with experimental findings** [6]; it also predicts that when the firing rate (or post-synaptic membrane potential) goes above a certain threshold no potentiation should occur.

References

[1] J. J. Hopfield, Proc Natl Acad Sci USA 79, 2554 (1982).
[2] D.J. Amit. Cambridge University Press(1992).
[3] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev., A 35, 2293 (1987).
[4] E. Gardner, J. Phys. a: Math. Gen. 22, 1969 (1989).
[5] M. V. Tsodyks and M. V. Feigel'Man, EPL (Europhysics Letters) 6, 101 (1988).
[6] A. Ngezahayo, M. Schachner, and A. Artola, J Neurosci 20, 2451 (2000).