Most animals have brains that can maintain a short-term memory. One theory is that brains create short-term memories from attractors created by circuits of neurons. Describe the key ideas of attractor theories of short-term memory, and discuss their strengths and weaknesses

January 20, 2020

Student ID: 20308798

Module Number: PSGY4061

Word Count (calculated by *Overleaf*, this line excluded): 1479

A Memory borrowed from Maths (*Introduction*)

The term attractor becomes highly intuitive when one thinks of a state space for a neural population: an attractor is a state in a network of neurons that can be easily driven to and not-so-easily escaped from. Authors like (Brody et al., 2003) very well describe them as states that minimize a *Lyapunov* (L-) function, in a way that very much resembles the stability of a particle under the influence of an energy potential. An attractor, thus, can define a particular combination of individual stable neuron states (which in turn make the network as a whole stable too). A natural extension of this concept comes when there exists a continuum of attractors in the network state space, forming a particular manifold; in this case one can call an attractor to this whole subspace, in the sense that nearby states will end up *falling* to one or another state within the attractor.

The question is, how can a mathematical construct as such be related to memory? The term memory has both many interpretations and classes, but if one thinks of a memory as a piece of information encoded in our brain, there must be a relatively robust relation between an environment state and a neural state codifying for its representation. Following this reasoning, from all the states a population of neurons can have, those by definition stable (attractors) might have a higher chance to be used in nature to *store* information.

On Attractor Theory and its Limitations (*Discussion*)

From Hopfield Networks to Population Dynamics

In this first section, the Hopfield Network (Hopfield, 1982) is reviewed as one of the simplest versions of a network able to store memory through attractors, and later extended to a continuous model.

Hopfield Networks The Hopfield Model proposes a discrete network where every unit S_i, can only be in either an active ($S_i = 1$) or inactive ($S_i = -1$) state, and is connected recurrently to each other through a weights matrix $W = (w_{ij})$. Given a state at a time t, $\vec{S}(t) = (S_1(t), ..., S_i(t), ...S_N(t))$, the population evolves according to:

$$S_i(t+1) = sgn\left[\sum_{j=1}^{N} w_{ij} S_j(t) - \theta_i(t)\right] \tag{1}$$

where θ is an external field and $sgn(x) = x/|x|$. Within these dynamics, one can define an *energy* term:

$$E[S, W] = -\frac{1}{2}\sum_{j,k} w_{jk} S_j S_k \tag{2}$$

so that one can prove that

$$\Delta E = E[S(t+\Delta t), W] - E[S(t), W] \geq 0 \;\; \forall \Delta t > 0 \tag{3}$$

It is thus a case where an explicit formulation of the L-function described in the introduction can be given. The shape of E[S] will make an initial state $\vec{S}_0 = \vec{S}(0)$ follow a trajectory towards one local minimum or the other. The set of $\vec{S}$ for which the system is stable ($\vec{S}(t+\Delta t) = \vec{S(t)}$, the attractor states) is called the *patterns* stored in the memory (network), and depends exclusively on W.

Continuous Attractors Continuous Network Dynamics can be naturally extended from Hopfield Networks:

$$\tau \dot{u}_i = -u_i + f\left(\sum_{j=1}^{N} w_{ij} y_j + h_i\right) \tag{4}$$

This equation is often used for modeling populations of neurons by setting u_i to be the membrane potential and y_j the Post Synaptic Current output from other neurons. h_i is again external input and can be understood in terms of a sensory input or noise.$f(x)$ is a monotonycally increasing function that generalizes $sgn(x)$. In recent studies (Rubin et al., 2015), *Stabilized Supralinear Networks* have been successful in describing not-so-well understood phenomena such as input normalization and surround suppression by setting

$$f(x) = [x]_+^n, \; n \geq 1 \tag{5}$$

where $[x]_+$ is equal to x if $x \geq 0$ and 0 otherwise

For continuous networks, W evolves as a dynamical system through *Hebbian learning* (Hebb, 1949). In a linear simplification, it does so by self-decaying to a value given by the coupling between the activities of pairs of units, constrained by a global synaptic capacity that projects learning to a particular subspace (Miller & MacKay, 1994):

$$\tau \dot{W} = CW - \gamma W \tag{6}$$

where $C = yy^{\mathrm{T}}$ is the covariance matrix of the activities of each unit y_i.

Hebbian learning, often summarized as "cells that fire together wire together" implies that the longer a state is visited the more stable it becomes: equation (6) links Short-Term Memory (STM) with Long-Term Memory (LTM) through modifications in W, as will be revised later in this essay.

Attractor models in Biology

Discrete Attractors In nature, the simplest kind of discrete attractors are bistable networks, which can take the form of single neurons, as very early observed by Fuster and Jervey (1981), or neuron populations. Recent studies have performed bifurcation analysis on population models in order to understand how attractors could guide social behaviour (Hurtado-López et al., 2017), where discrete attractors are proposed to drive male mice behaviour towards either exploration or mounting/fighting behaviour.

Line Attractors Line attractors were first proposed as a model for keeping eyes still (Seung, 1996), as they allowed for the brain to keep track of previous positions and avoid saccade movements to eventually drive eyes to an undesired position. Since, they have since been proposed to encode different types of simple linear information (parametric working memory), as can be the frequency of vibration in a vibrotactile experiment (Romo et al., 1999).

Ring Attractors Ring, *bump* or Amari (Amari, 1977) attractors encode periodic information, as can be the tuning orientation in a receptive field (Ben-Yishai et al., 1995) or the direction of appearance of a visual cue within a circumference (Funahashi et al., 1989). They are based on short excitation - long inhibition (also called *Mexican hat*) connections that generate attractor states in which solutions are bumps of activity that show translational symmetry.

Complex Attractors In the recent years, a double ring attractor has been proposed for encoding spacial information in the so-called *grid cells* in the hippocampus (Moser et al., 2008). If one extends Mexican hat connections to a 2-dimensional representation, each unit has a combined preference for two different radial pieces of information, finding itself inside a ring of rings (a *thorus*).

Electrophysiological recordings in rats moving freely show that grid cells have different firing tunings when placed in different environments they are familiar with (O'Keefe & Conway, 1978). Since (Battaglia & Treves, 1998), many authors as (Monasson & Rosay,

2015) have proposed an attractor of attractors model as a solution to encoding multiple representations in the same network: each attractor manifold is itself a stable state of a discrete attractor: the rat can be in the environment A, B or C, and within that environment it is also able to be in a particular place. This mechanism is a possible answer to one of the open-questions in attractor theory, dynamic encoding.

Attractors and Learning

This section will try to illustrate how attractor theories explain three complementary concepts: STM, LTM and learning.

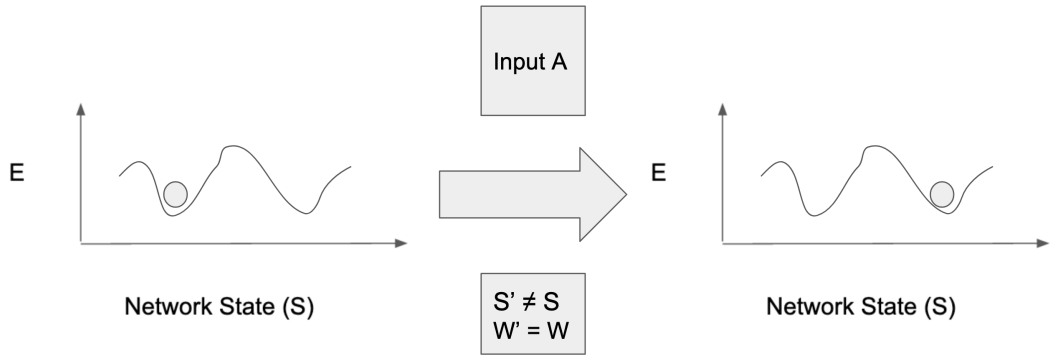

Figure 1: Attractors explain STM as the continuation in time of a stable state

STM In terms of attractors, STM can be explained if one thinks of a structure in the brain following the dynamics described by (4). After an input (h in (4)) gives momentum to a given state S, the synaptic weights will drive the network to a state S' that minimizes its energy, where it will be sustained until another input or noise shift the network to different state (figure 1).

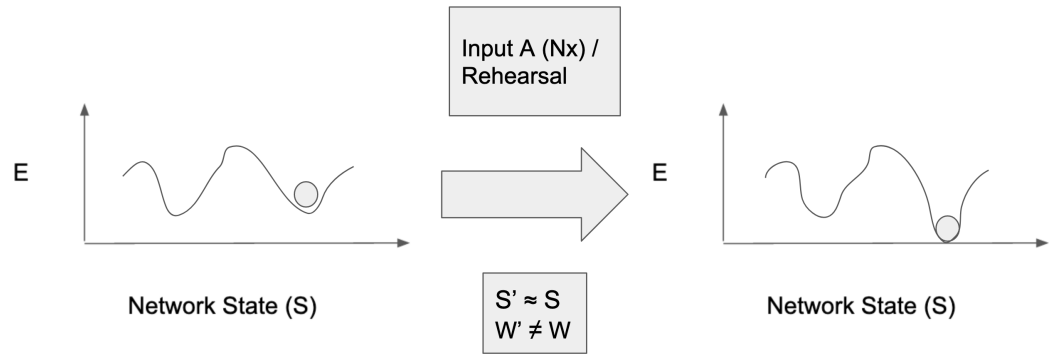

Figure 2: Although in a slow timescale, a network is always stabilizing the state it finds itself in through *learning*, which results in LTM.

LTM and learning According to (6), being in a particular state itself makes little changes in the synaptic weights, so that the energy associated to that state is lower. With time, when similar inputs appear repeatedly, or one actively tries to retain in STM a particular state (rehearsal), the energy profile changes and the state is more likely to be visited and sustained in the future. This is the process of learning and allows us to acquire Long-Term Memory (figure 2).

Strengths and Limitations of the model

Since (Wang, 2001), there has been a common sense that attractors help encoding STM in our brain through persistent activity. The model has been tested against in several occasions, as in (Wimmer et al., 2014), where electrophysiological recordings were correlated with behavioural variability. Results showed an extreme correspondence between observations and predictions from a Bump Attractor Model. Nevertheless, regardless the power of this theory, a series of open questions are still standing and are here briefly revised:

How are memories kept still without noise drifting them to close attractor states? Two possible answers are: (a) the energy profile will always be imperfect in nature, having natural wells that make encoding less accurate but resistant to noise, (b) mechanisms like synaptic facilitation (Itskov et al., 2011) can make temporal wholes in this profile to ensure an even more stable retention of short-term memories.

How can attractors store more than one item at a time? Time and space partitioning in the network have been proposed as an answer to this question (Barak & Tsodyks, 2014). Nevertheless, an alternative answer might arise if one considers the brain as a very complex attractor. For a network state of sufficient dimensions, attractor states could encode what we consider different elements of information at once.

If attractors are steady states, why can one observe non-persistent activity in single-cell recordings? This question is extensively discussed in (Lundqvist et al., 2018). A possible answer comes from the fluctuations inside an energy well a network can undergo, changing certain singe-cells activities while being in virtually the same network state. Another possibility is that this transient activity is actually part of the information of encoding mechanisms, as proposed for *dynamic attractors* (Laje & Buonomano, 2013).

Conclusion

In this essay we have seen how attractors provide a mathematical framework for understanding STM, LTM, and learning. In the author's opinion, future work will probably find answers to revised (and other) open questions by extending the model with more complex networks and attractor models rather than alternative theories: it seems certain that stability in our mental processes must come from stability in our brain, although the concept of stability might turn out to be more labyrinthine than we believe as of today.

References

Amari, S.-i. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological cybernetics*, *27*(2), 77–87.

Barak, O., & Tsodyks, M. (2014). Working models of working memory. *Current opinion in neurobiology*, *25*, 20–24.

Battaglia, F. P., & Treves, A. (1998). Attractor neural networks storing multiple space representations: a model for hippocampal place fields. *Physical Review E*, *58*(6), 7738.

Ben-Yishai, R., Bar-Or, R. L., & Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences*, *92*(9), 3844–3848.

Brody, C. D., Romo, R., & Kepecs, A. (2003). Basic mechanisms for graded persistent activity: discrete attractors, continuous attractors, and dynamic representations. *Current opinion in neurobiology*, *13*(2), 204–211.

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of neurophysiology*, *61*(2), 331–349.

Fuster, J. M., & Jervey, J. P. (1981). Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. *Science*, *212*(4497), 952–955.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. Psychology Press.

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.

Hurtado-López, J., Ramirez-Moreno, D. F., & Sejnowski, T. J. (2017). Decision-making neural circuits mediating social behaviors. *Journal of computational neuroscience*, *43*(2), 127–142.

Itskov, V., Hansel, D., & Tsodyks, M. (2011). Short-term facilitation may stabilize parametric working memory trace. *Frontiers in computational neuroscience*, *5*, 40.

Laje, R., & Buonomano, D. V. (2013). Robust timing and motor patterns by taming chaos in recurrent neural networks. *Nature neuroscience*, *16*(7), 925–933.

Lundqvist, M., Herman, P., & Miller, E. K. (2018). Working memory: delay activity, yes! persistent activity? maybe not. *Journal of Neuroscience*, *38*(32), 7013–7019.

Miller, K. D., & MacKay, D. J. (1994). The role of constraints in hebbian learning. *Neural computation*, *6*(1), 100–126.

Monasson, R., & Rosay, S. (2015). Transitions between spatial attractors in place-cell models. *Physical review letters*, *115*(9), 098101.

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, *31*, 69–89.

O'Keefe, J., & Conway, D. (1978). Hippocampal place units in the freely moving rat: why they fire where they fire. *Experimental brain research*, *31*(4), 573–590.

Romo, R., Brody, C. D., Hernández, A., & Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature*, *399*(6735), 470–473.

Rubin, D. B., Van Hooser, S. D., & Miller, K. D. (2015). The stabilized supralinear network: a unifying circuit motif underlying multi-input integration in sensory

cortex. *Neuron*, *85*(2), 402–417.

Seung, H. S. (1996). How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences*, *93*(23), 13339–13344.

Wang, X.-J. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends in neurosciences*, *24*(8), 455–463.

Wimmer, K., Nykamp, D. Q., Constantinidis, C., & Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature neuroscience*, *17*(3), 431–439.