

Mechanisms for the formation of neural representations of abstract rules

Emanuele Curti^a, Xiao-Jing Wang^b and Stefano Fusi^{a,c,1}

^a*Center for Theoretical Neuroscience, Columbia University, New York, USA*

^b*Volen Center for Complex Systems, Brandeis University, Waltham, USA*

^c*Institute of Neuroinformatics, ETH/Univ. of Zurich, Switzerland*

Abstract

In many circumstances primates' behavior cannot be described as a simple one-to-one mapping between a stimulus and a motor response. Every event or action can modify the animal disposition to behavior, and, for example, the same sensory stimulus might lead to different motor responses depending on the context, previously determined by a set of other cues. We propose a network model in which each inner mental state is represented by an attractor of the neural dynamics. Each state represents a rule encoding the prescriptions for going from one state to another or for expressing a decision about a motor response. We illustrate this theoretical framework with a simple example. We consider a task in which an animal has to respond with a saccadic movement (Left or Right) to two sensory stimuli (A and B). In one context A should lead to L and B to R. In another one the associations are reversed and A should lead to R and B to L. The two rules can be expressed in words as: 1) when A is associated with L, then B is associated with R 2) if A leads to R, then B leads to L. We address two issues: how are the rule representations built? How can the active representation of a rule lead to the decision about the motor response? We propose a network in which two populations of neurons compete to express a final decision about the motor response (L or R). These two populations are assumed to be highly structured due to heterogeneities. In particular we assume that within L there is a population AL which has a preference for A (i.e. before learning, A drives it to higher frequencies than other neurons within L). Analogously we can define populations BL, AR, BR. When the first rule (AL-BR) is in effect, the activation of AL is consistently followed by the activation of AL or BR. The rule representation then forms because of the temporal contiguity of AL-BR for rule 1, and AR-BL for rule 2. When rule 1 is active, then A favors L in the competition because AL receives a recurrent (from BR) and a sensory input, while BR receives only the recurrent input. Hence stimulus A steers the activity of the network towards a state which expresses the final decision about the motor response and it keeps in memory which rule is in effect.

Key words: Rule abstraction, Attractor Neural Networks, Context-dependent Learning

We illustrate a mechanism for rule abstraction and representation with a simple example in which visual stimuli are associated to one of two possible motor responses (say two saccadic movements, left and right). In one context the first stimulus (A) should generate Left and the second stimulus should lead to Right in order to receive a reward. In the second context the associations are reversed: A-Right and B-Left are the rewarded associations. The rules to get reward can be expressed in words as follows: Rule 1: "when A is associated to Left, then B is associated to Right" and Rule 2: "when A is associated to Right, then B is associated to Left". There are three fundamental questions that we address: 1) how are the rule representations built? 2) how can the active representation of a context lead to the decision about the motor response? 3) how can contextual cues indicate explicitly what rule is in effect? In order to answer these three questions we first need to describe the neural network which will implement the rules. We assume that there are populations of neurons which are selective to the intended motor response, similarly to [3]. In our case we will group together all the neurons with a preference for Left (population L) and those with a preference for Right (population R). The activation of one of the two groups would express the decision of the monkey to make a saccadic movement to a specific direction. The two populations of neurons compete through a population of inhibitory neurons, as in the decision making network introduced in [7]. Each pattern of neural activity in which one population is active (expressing the decision) and the others are inactive, is a global attractor of the neural dynamics [1]. We now consider the heterogeneity across cells. Within each decision population (L or R), we can identify and tag the neurons that have a preference for one specific sensory stimulus (see Figure 1. For example, we define neurons within population L with preference to A as those that exhibit the largest response when stimulus A is presented, and we tag them with the label AL. Analogously we can define population BL, again within population L, and AR, BR within population R. This kind of heterogeneity has been observed in prefrontal and in premotor cortex [2]. Rule representations are created by the temporal proximity of events in rewarded trials: for example when rule 1 is in effect, A-Left trials are followed by either A-Left or by B-Right trials. Previous experimental results have shown that neural representations of events that occur in a fixed order tend to merge into a single representation linking neighboring events in a sequence [4]. We then expect that if AL and BR are separated attractors, after long enough sequences of trials in which rule 1 is in effect, AL and BR will merge into a single attractor. Analogously AR and BL will fuse to represent rule 2. Can then the activation of one of these attractors affect the competition between L and R and express the decision which is dictated by the rule in effect? Intuitively this should be possible because the activation of the representation of a rule can bias the competition. Indeed, if rule 1 is in effect, upon the presentation of A, AL and BR receive the recurrent inputs from AL and BR due to the fact that the network

¹ Corresponding author, *E-mail address*: fusi@neurotheory.columbia.edu

is in the attractor corresponding to rule 1. In addition to these inputs, AL receives also a stronger activation from sensory stimulus A, which can then favor L in the competition. Finally how is the rule selected? Modification of the context can be determined by the feedback the monkey receives (e.g. when the monkey applies one rule and it is not rewarded any longer), or it can be explicitly signaled by one or more contextual cues (also called occasion setters in psychology literature[5]). We shall consider the second case below. We show that these mechanisms can be implemented in a simple rate model of a network of neurons. We illustrate the simulated network dynamics after learning in Figure 1. A trial starts from the attractor corresponding to the representation of rule 1 (AL-BR). After one second, a contextual cue is presented to indicate that the rule in effect will be rule 2 (AR-BL). Network's activity steers toward the attractor corresponding to rule 2 (between 1.5 and 2.5s). When stimulus B is shown, the competition between L and R starts, and the previous activation of rule 2 attractor favors L, as dictated by the rule in effect. Notice that the selection of L does not disrupt the information about the rule in effect. Indeed the activity of AR remains high, surviving the decision process and its reset following the execution of the motor response.

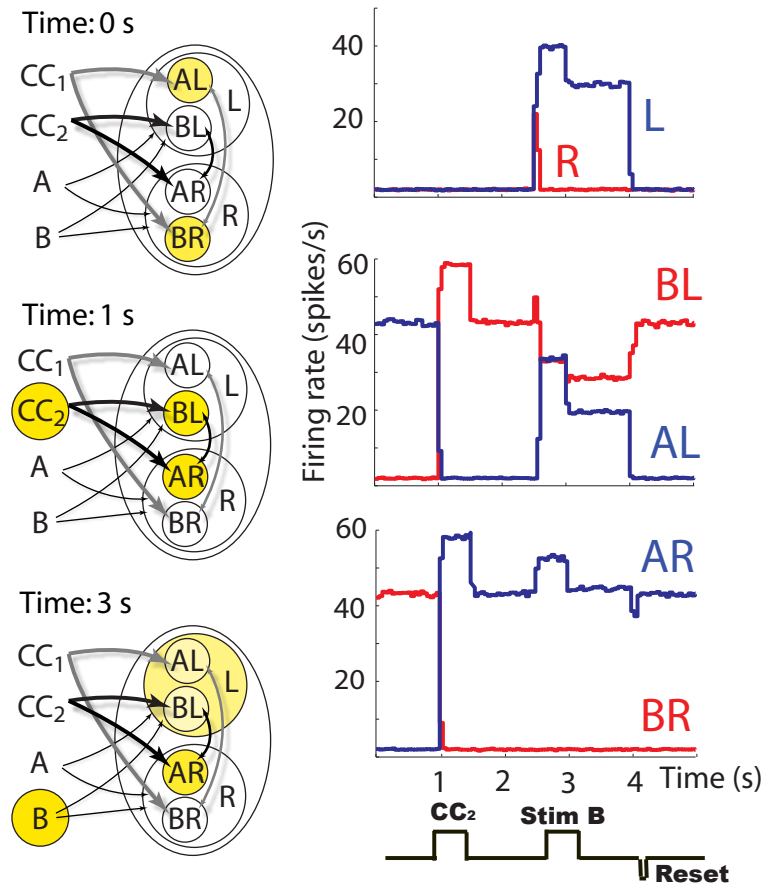


Fig. 1. Network dynamics in a typical trial. Left: three snapshots of the network patterns of activity at three different times. Right: average frequency as a function of time for all the network populations. See the text.

Acknowledgments

SF was supported by the SNF grant PP00A-106556, EC and SF by the grant NIH-2R01 MH58754.

References

- [1] Amit DJ (1989) Modeling brain function. Cambridge University Press.
- [2] Asaad WF, Rainer G, Miller EK (1998) Neural activity in the primate prefrontal cortex during associative learning. *Neuron* 21:1399-1407
- [3] Fusi S, Asaad WF, Miller EK and Wang X-J (2005) A microcircuit model of arbitrary sensori-motor mapping: learning and forgetting on multiple timescales. *Soc Neurosci Abstr* 813:10
- [4] Miyashita Y (1988) Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* 335: 817-20.
- [5] Schmajuk N and Holland PC (1998). *Occasion Setting: Associative learning and cognition in animals* (American Physiological Association, Washington, D.C., 1998).
- [6] Wallis JD and Miller EK (2003) From rule to response: neuronal processes in the premotor and prefrontal cortex. *J Neurophysiol* 90: 1790-1806.
- [7] Wang X-J (2002) Probabilistic decision making by slow reverberation in neocortical circuits. *Neuron* 36: 955-968.