Short-term memory both as working memory and as a basis for long-term memory

Emmanuel Guigon, Etienne Koechlin, Yves Burnod

INSERM U483, Université Pierre et Marie Curie
9, quai Saint-Bernard, 75005 Paris, France

January 26, 2002

Short title: Short-term memory

Correspondence to
Emmanuel Guigon
INSERM U483
U.P.M.C., Boîte 23
9, quai Saint-Bernard
75005 Paris, France
Fax 33 1 44 27 34 38
Phone 33 1 44 27 34 37
E-mail guigon@ccr.jussieu.fr

Submitted to: The Handbook of Brain Theory and Neural Networks, 2nd ed (Arbib MA, ed). Cambridge: The MIT Press.

Introduction

It is generally agreed that two temporally distinct neural processes contribute to the acquisition and expression of brain functions. Transient variations of membrane potential (neuronal activity), with a time scale of milliseconds, reflect the flow of information from neuron to neuron and define the function of neuronal networks. These variations can result in long-lasting (and maybe permanent) alterations in neuronal operations, for instance through activity-dependent changes in synaptic transmission.

There is now strong evidence for a complementary process, acting over an intermediate time scale. A wide variety of temporal patterns of activity are actively generated by neurons and local circuits of neurons, such as transforming transient inputs into long-lasting sustained or oscillatory activity. Experimental studies in invertebrates have demonstrated that such temporal patterns produce motor programs and are generated both by the molecular properties of each neuron, and the connectivity of the local network (Harris-Warrick and Marder 1991; Marder et al. 1996). In vertebrates, long-lasting activities are neural correlates of transient memory processes. These patterns of activity allow past events to be represented and behavioral reaction to future, predictable events to be prepared (see PREFRONTAL CORTEX IN TEMPORAL ORGANIZATION OF ACTION). The fact that they should also result from both intrinsic properties of single neurons and synaptic interactions between neurons is now well recognized (Llinás 1988; Marder et al. 1996; Durstewitz et al. 2000).

In the present chapter, we discuss cellular and neural network mechanisms which could be involved in the formation of short-term memory (STM) traces in the vertebrate brain. We address three issues: (1) What are the different types of STM traces? (2) How intrinsic and synaptic mechanisms contribute to the formation of STM traces? (3) How STM traces translate into long-term memory representation of temporal sequences? We note that we are concerned here neither with exact definitions and properties of psychological concepts nor with detailed biophysical or biochemical mechanisms involved in the characterization of short-term memory processes, but only with computational mechanisms underlying these processes. We also note that these mechanisms may well underlie a wide variety of seemingly different biological processes (emergence of orientation selectivity in visual cortex, dynamics of head-direction cells in the limbic system, directional tuning in motor cortex), and thus be relevant to understand brain functions.

Types of Neural Short-Term Memory Traces

There exist two broad types of STM traces (Fig. 1):

- Transient inputs are transformed into long-lasting activity patterns (Fig. 1*A*). The output traces could represent a membrane potential, a discharge frequency or any biophysical or biochemical variable (e.g. intracellular calcium concentration). Ideally, the level of maintained activity would be proportional to the intensity of the input stimulus (intensity memory; Fig. 1*A1*), or tuned about a preferred (spatial) stimulus (spatial memory or memory field; Fig. 1*A2*). Spatial and intensity memory mechanisms are relevant to working memory, i.e. the ability to hold a relevant information in memory for future utilization in the guidance of behavior. Neural correlates of working memory are found mainly in the anterior regions of the cerebral cortex (e.g. prefrontal cortex) as stimulus-selective sustained neuronal discharges (see PREFRONTAL CORTEX IN TEMPORAL ORGANIZATION OF ACTION). Complete references on the issue of working memory models can be found in Durstewitz et al. (2000).

- Constant inputs are transformed into time-varying outputs (Fig. 1*B*), e.g. ramps with different slopes (Fig. 1*B1*) or oscillations at different frequencies (Fig. 1*B2*) or of different amplitudes (not shown). Again characteristics of the output patterns (slope, frequency, amplitude) should be related to the intensity of the input. Activity ramps are found as correlates of preparatory and anticipatory processes in sensorimotor and cognitive behaviors. They are ubiquitous in cortical parietal and frontal region, and could participate to muscular recruitment, preparation for response, and decision-making processes (Hanes and Schall 1996). Little attention has been paid to the formation of ramps. Oscillatory activities are mentioned here because they are typical STM traces which have been attributed a central role in many behavioral processes (see THALAMOCORTICAL OSCILLATIONS IN SLEEP AND WAKEFULNESS). Cellular and network mechanisms of oscillations are thoroughly dealt with in the literature (see OSCILLATORY AND BURSTING PROPERTIES OF NEURONS) and are not addressed here.

We note that many other STM traces could be built by combining these two types. Furthermore, in physiological recordings, additional time-varying components would be found in inputs and outputs corresponding to different types of variability in neural processing.

Cellular and Network Mechanisms of STM traces

Sustained Activity

The basic mechanism for the formation of maintained activities is a neuron with a recurrent excitatory connection

$$\tau \frac{dI}{dt} = -I + wf(I) + I_{in}, \tag{1}$$

where I is a variable which could represent the total synaptic current, $f(I) = 1/[1+\exp(s(0.5-I))]$ the firing rate of the neuron, w the weight of the recurrent connection, and I_{in} the input current. The bifurcation diagram of this equation was plotted with I_{in} as a parameter (Fig. 2*A*). For $I_{in} < I_1$ and $I_{in} > I_2$, the equation has a single stable fixed point. Elsewhere, there are three fixed points: two are close to minimum and maximum firing rate, respectively, and are stable (lower and upper branches in Fig. 2*A*). The third one is unstable (dashed middle branch). Transient inputs elicit transition between the stable states (Fig. 2*B*).

This model is illustrative of a general class of neural networks in which persistent states arise from reverberating activity through recurrent excitatory loops (see COMPUTING WITH ATTRACTORS). The main drawback of these models is that the level of maintained activities is close to the neural saturation level and the resting level is silent (Fig. 2*A*), in contradiction with physiological observations. More realistic persistent activities are found when excitation and inhibition are represented by different neuronal populations (Durstewitz et al. 2000). Furthermore low and high frequency persistent states can coexist when inhibition slightly dominates excitation.

The models discussed above describe neural processing in terms of firing rate or synaptic current whereas a physiologically realistic representation is defined by equations governing membrane potential [i.e. Hodgkin-Huxley equations and equations for synaptic inputs (see SYNAPTIC CURRENTS, NEUROMODULATION, AND KINETIC MODELS)]. An open question is thus whether similar properties would be found in a biophysically realistic model of a spiking neuron. Simulations show that the ability of a neural network to maintain robust delay activity at physiological rates (e.g. $15 - 20$ Hz) depends on the nature of synaptic transmission. In fact, the largest component of the synaptic transmission, mediated by AMPA receptors, has a fast decay which leads to persistent discharges at frequencies above 50 Hz (Wang 1999). Contribution of slow synaptic transmission through NMDA receptors could help bypass this effect and reduce the frequency of persistent activity to the required level (Wang 1999). However, it is unknown whether the density of NMDA receptors is large enough to play such a role.

We discussed how maintained activity can result from recurrent interactions within neuronal populations. Alternatively, maintained activity could correspond to the depolarized state of an intrinsically *bistable* neuron (Marder et al. 1996). Intrinsic bistability is characterized by the existence of two or more stable states (e.g. a hyperpolarized state and a more depolarized state in which the neuron discharges), and the transition from one state to the other by transient synaptic events (Marder et al. 1996). Numerous examples of bistability have been described in the literature, both in invertebrates (e.g. *Aplysia*, crab stomatogastric ganglion) and in vertebrates (spinal cord, cerebellum, thalamus). The cellular bases of bistability generally involve a low-threshold persistent inward conductance, i.e. a depolarizing conductance which activates in the subthreshold range and does not inactivate (see ION CHANNELS: KEYS TO NEURONAL SPECIALIZATION). If the neuron is endowed with a spiking mechanism, the depolarized state corresponds to the discharge of action potentials. Otherwise, it is a plateau potential.

The models described so far maintained persistent activities in an all-or-none fashion, at a level prescribed by their structure. They correspond to the notion of a subset of strongly and uniformly connected neurons representing a discrete attractor (see CORTICAL HEBBIAN MODULES). Below, we show that adequate choice of recurrent synaptic interactions allows memory encoding of continuously-valued variables (intensity or spatial memory; Fig. 1*A*).

Intensity memory (Fig. 1*A*1) can be addressed in the framework of linear recurrent neural networks, i.e.

$$\tau \frac{d\boldsymbol{I}}{dt} = -\boldsymbol{I} + \boldsymbol{W}\boldsymbol{I} + \boldsymbol{I}_{in}, \tag{2}$$

where $\boldsymbol{I}$ is the N-dimensional vector of output activities and $\boldsymbol{W}$ a symmetric synaptic matrix. This equation, which is a multidimensional linear generalization of Eq. (1), can be solved explicitly for $\boldsymbol{I}$,

$$\boldsymbol{I}(t) = \sum_{i=1}^{N} a_i(t)\boldsymbol{e}_i,$$

where $\{\boldsymbol{e}_i\}$ are the eigenvectors of $\boldsymbol{W}$. Persistent activity appears in the case where one eigenvalue (index k) of $\boldsymbol{W}$ is equal to 1 and all the other eigenvalues are smaller than 1. The solution becomes

$$\boldsymbol{I}(t) \approx \frac{\boldsymbol{e}_k}{\tau} \int_0^t \boldsymbol{I}_{in}(t') \cdot \boldsymbol{e}_k \, dt'. \tag{3}$$

This equation shows that the network can hold a faithful memory of the amplitude of a transient input. However, this property is lost when the synaptic matrix is even slightly perturbed. This would be also the case in a nonlinear version of this model.

This principle was used in the framework of conductance-based spiking models by Seung et al. (2000) to explore brain stem networks involved in the control of eye position. In this model, brain stem neurons are integrators which convert transient signals driving changes in eye position into a persistent memory of eye position (Fig. 3*A*).

The linear recurrent network described by Eq. (2) can also generate **spatial memory** profiles. This occurs when (1) each neuron i is identified by a periodic parameter θ_i (e.g. a preferred direction in $[0;2\pi]$); (2) the $\{\theta_i\}$ are uniformly distributed in $[0;2\pi]$; and (3) the synaptic weight between neurons i and j is $W_{ij} = \cos(\theta_i - \theta_j)$. In this case, $\boldsymbol{W}$ has only two nonzero eigenvalues equal to 1, and acts as a filter which suppresses all harmonics ≥ 2 in the input signal. Thus the network generates and maintains a cosine distribution of activity from any nonuniform transient input. The constraint on the eigenvalues of $\boldsymbol{W}$ can be relieved in a nonlinear version of Eq. (2)

$$\tau \frac{d\boldsymbol{I}}{dt} = -\boldsymbol{I} + \boldsymbol{W}[\boldsymbol{I}]_+ + \boldsymbol{I}_{in}, \tag{4}$$

where $[\;]_+$ ($[u]_+ = u$ if $u \geq 0$ otherwise $[u]_+ = 0$) translates current into firing rate. In this case, a spatially selective activity profile can persist in the presence of a constant background (Fig. 3*B*). This behavior appears in the case of strong, spatially modulated excitatory connections, and corresponds to the existence of a continuous line of stable states (Hansel and Sompolinsky 1998).

A realistic implementation of spatial memory was described by Compte et al. (2000). Their network involved (1) excitatory and inhibitory neurons modeled as leaky integrate and fire units (see SINGLE-CELL MODELS); (2) spatially structured connections between the excitatory neurons; (3) slow (NMDA) synaptic transmission (Wang 1999). The model displayed both low spontaneous activity and robust stimulus-evoked selective persistent discharges at physiological frequencies ($\sim$ 20-30 Hz). In the preceding model (Eq. 4), spatial pattern formation resulted from a continuous bifurcation. On the other hand, there is genuine network bistability in the present model, which authorizes transition between resting and activated states by transient excitatory inputs. Compte et al. (2000) observed a drift in time of persistent activity patterns in the presence of noise, which results in a degraded memory of encoded stimuli. In fact, in all models, the spatial patterns are only marginally stable (Hansel and Sompolinsky 1998).

An attractive hypothesis would be that both synaptic and intrinsic properties contribute to the formation of persistent activity. Lisman et al. (1998) proposed that NMDA-receptor-mediated bistability could participate to the maintenance of selective working memory activity. Camperi and Wang (1998) used a conditional bistability in a continuous attractor network (Eq. 4), and showed that it can contribute to the stability of maintained activities against perturbations although it was not involved in their maintenance *per se*.

On the whole, these mechanisms provide reasonable clues on how sustained activities can be maintained in neuronal populations. However, there remain several open questions: (1) All the models have a built-in instability and require finely tuned synaptic weights to work appropriately. Is this instability a characteristic feature of sustained discharges in the nervous system? How this instability could be removed? Which mechanisms allow the development and maintenance of exact synaptic structures? (2) The models have many features in common and apply to the emergence of orientation selectivity in visual cortex, dynamics of head-direction cells in the limbic system, directional tuning in motor cortex, persistent activities in prefrontal cortex. Are there definite differences in the neural substrate of these functions? For instance, is the putative role of slow synaptic transmission identified by some models a characteristic feature of prefrontal cortical circuits?

Activity Ramp

When a constant current is injected in a neuron, a time-varying pattern of activity is observed which depends on passive properties of the neuron (membrane time constant) and active membrane characteristics (voltage-gated ionic conductances). The former effect is illustrated by the voltage response of a passive membrane

$$\frac{dV}{dt} = -\frac{V}{\tau} + I,$$

for different values of τ and different I. The time to reach a threshold V_θ is given by

$$T_\theta = -\tau \ln \left(1 - \frac{V_\theta}{\tau I} \right) .$$

This relation is strongly nonlinear and shows that neither I nor τ can efficiently be used to specify a duration. At best it could be used for duration below 50 ms.

The same is approximately true in a Hodgkin-Huxley model (see AXONAL MODELING) because the sodium and potassium conductances of the action potential are weakly activated in the subthreshold range. The time to reach a given frequency is not yet an appropriate timing mechanism because steady-state discharge settles within a few time constants.

At the single neuron level, robust short-term memory properties arise from the presence of a slowly inactivating potassium (Ks) conductance (Marder et al. 1996; Delord et al. 2000). The functioning principle of the Ks conductance is the following. It creates a slowly decaying hyperpolarizing current whose initial level can be specified by prior conditioning of the neuron. For instance, prior hyperpolarization sets a large persistent outward current which slows down the rate of membrane potential changes in the subthreshold range during a subsequent depolarization. Figure 4*A* shows that the latency-to-the-first-spike can be up to 10 seconds in the presence of a Ks conductance with an inactivation time constant of 2 seconds, and the relationship between the injected current and the latency is close to linear for latency up to ≈7 s.

The Ks conductance also influences the suprathreshold behavior of the neuron. The discharge frequency gradually increases toward its steady state level as the Ks current decays (Fig. 4*B*). Both the initial and final frequency increased with the level of injected current which result in a modest change in the slope of the time-frequency curve with the injected current (Fig. 4*C*). Recruitment at variable rates as described in Fig. 1*B1*, can be approached by combining the effect of Ks conductance and synaptic interactions in a population of uniformly connected neurons (Delord et al. 2000). This is illustrated in Fig. 4*B*. In this case, due to recurrent excitation, a smaller amount of injected current is required to obtain a given steady state frequency. Thus the initial frequencies are lower and vary in a smaller range. Accordingly, the slope is more strongly modulated by the injected current than in the absence of synaptic interactions (Fig. 4*C*). The strength of this modulation is directly controlled by the strength of synaptic weights (Fig. 4*D*). Thus adaptive recruitment at variable rates is made possible by the simultaneous action of synaptic and intrinsic properties in a neural network. Interestingly, slowly inactivating potassium conductances are found in neurons of most regions of the central nervous system with a time constant ranging from hundreds of milliseconds to several tens of seconds (Llinás 1988).

Could a similar property be obtained by purely synaptic effects? In fact, the linear recurrent network described by Eq. (2) has the required property (Fig. 3*B*).

The formation of persistent activities begins by a linear ramp with a slope proportional to the amplitude of the input current (Eq. 3). However, as mentioned before, exactly tuned synaptic weights are necessary to the proper functioning of the network. It is unclear whether the nervous system can reach the required degree of accuracy in the adjustment of synaptic weights (Seung et al. 2000).

From Short-Term to Long-Term Memory Traces

Synaptic plasticity is a central mechanism in models of learning and memory (see HEBBIAN SYNAPTIC PLASTICITY). The most popular approach involves shaping functions of neural networks by activity-dependent modification of synapses based on Hebbian learning rules. Accordingly, information stored into long-term memory reflects correlation (i.e. temporal contiguity) between transient neuronal activities on the timescale of 0-100 ms. Sutton and Barto (1981) recognized that learning rules based on temporal contiguity are inappropriate to represent temporal dependencies, e.g in the framework of classical conditioning. They proposed that STM traces (synapse-specific and output traces) are involved in the acquisition of new conditioned behaviors. This principle has led to the development of the temporal difference algorithm which provides a powerful way to learn predictions of future events (see REINFORCEMENT). In this framework, STM traces are translated into a long-term representation of the temporal structure of behavior. A related approach applied to the prefrontal cortex is described in Guigon et al. (1995).

Discussion

Short-term memory traces become essential as soon as the timescale of behavior extends beyond the duration of phasic signaling (e.g. $\sim$ 0-500 ms), i.e. in quite any behavioral situation. We discussed cellular and network mechanisms involved in the formation of STM traces. We described simplified network models which provide mathematical conditions on the formation and stability of STM traces, and more detailed realistic models which helped assess the biophysical basis of these phenomena. Despite impressive results, our understanding of these mechanisms is far to be complete. First, each neuron is endowed with a wealth of intrinsic properties (Llinás 1988), very few of which have been considered in models. The respective contribution of synaptic and intrinsic factors is unknown. Second, persistent discharges and ramps constitute a small subset of the rich dynamic repertoire of neural populations observed *in vivo*, and it is unclear how they can be combined to form more complex memory traces. Third, the great majority of models fail to be robust facing noise and inexact tuning of parameters (e.g. synaptic weights). A future challenge is also to relate these pieces of memory to cognitive functions which are very demanding of temporary storage and manipulation of information, e.g. planning, reasoning, language.

References

Camperi, M. and Wang, X.-J., 1998, A model of visuospatial working memory in prefrontal cortex: Recurrent network and cellular bistability. J. Comput. Neurosci., 5(4):383–405.

Compte, A., Brunel, N., Goldman-Rakic, P., and Wang, X.-J., 2000, Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. Cereb. Cortex, 10(9):910–923.

Delord, B., Baraduc, P., Costalat, R., Burnod, Y., and Guigon, E., 2000, A model study of cellular short-term memory produced by slowly inactivating potassium conductances. J. Comput. Neurosci., 8(3):251–273.

Durstewitz, D., Seamans, J., and Sejnowski, T., 2000, Neurocomputational models of working memory. Nat. Neurosci. Suppl., 3:1184–1191.

Guigon, E., Dorizzi, B., Burnod, Y., and Schultz, W., 1995, Neural correlates of learning in the prefrontal cortex of the monkey: A predictive model. Cereb. Cortex, 5(2):135–147.

Hanes, D. and Schall, J., 1996, Neural control of voluntary movement initiation. Science, 274:427–430.

Hansel, D. and Sompolinsky, H., 1998, Modeling feature selectivity in local cortical circuits. In: *Methods in Neuronal Modeling: From Ions to Networks, 2nd ed* (Koch, C. and Segev, I., eds), pp 499–567. Cambridge: MIT Press.

Harris-Warrick, R. and Marder, E., 1991, Modulation of neural networks for behavior. Annu. Rev. Neurosci., 14:39–57.

Lisman, J., Fellous, J.-M., and Wang, X.-J., 1998, A role for NMDA-receptor channels in working memory. Nat. Neurosci., 1(4):273–275.

Llinás, R., 1988, The intrinsic electrophysiological properties of mammalian neurons: Insights into central nervous system function. Science, 242:1654–1663.

Marder, E., Abbott, L., Turrigiano, G., Liu, Z., and Golowasch, J., 1996, Memory from the dynamics of intrinsic membrane currents. Proc. Natl. Acad. Sci. U.S.A., 93(24):13481–13486.

Seung, H., Lee, D., Reis, B., and Tank, D., 2000, Stability of the memory of eye position in a recurrent network of conductance-based model neurons. Neuron, 26(1):259–271.

Sutton, R. and Barto, A., 1981, Toward a modern theory of adaptive networks: Expectation and prediction. Psychol. Rev., 88:135–170.

Wang, X.-J., 1999, Synaptic basis of cortical persistent activity: The importance of NMDA receptors to working memory. J. Neurosci., 19(21):9587–9603.

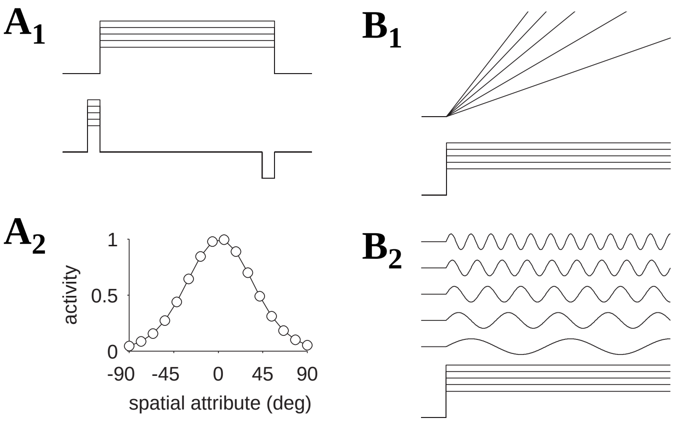

Figure 1: Types of STM traces. *A1*, Intensity memory. Intensity of a transient input (*lower trace*) is translated into a long-lasting activity (*upper trace*) of equal (or proportional) amplitude. Five activity traces are shown corresponding to five input intensities (in the same order). Horizontal time scale and vertical intensity scale are not specified (same for *B1* and *B2*). *A2*, Spatial memory. The memorized value of a spatial attribute (here 0°) is represented by a tuned activity distribution in a population of neurons selective to this attribute. *B1*, Activity ramp. A constant input (*lower trace*) is translated into a time-varying linearly increasing activity (*upper trace*). The slope of output activity is proportional to input intensity. *B2*, A constant input is translated into oscillations (the output traces are shown separately for clarity). Oscillatory frequency is proportional to input intensity.

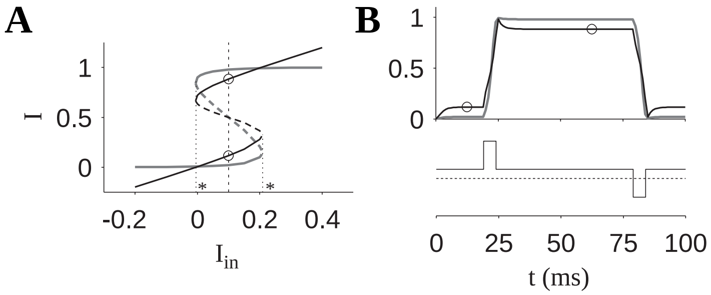

Figure 2: *A*, Bifurcation diagram of Eq. 1. The plot depicts the value of equilibrium state synaptic currents (black curve) and corresponding firing rates (gray curve). States belonging to plain (dashed) lines are stable (unstable). Vertical dotted lines delimit a region with three equilibrium states and define input currents I_1 and I_2 (stars). *B*, Activity profile (same conventions as in *A*) for $I_{in} = 0.1$ (vertical dashed line in *A*) and transient excitatory and inhibitory inputs (*lower trace*). Steady state are depicted by $\circ$. Parameters were $\tau = 2$, $s = 10$, $w = 0.8$.

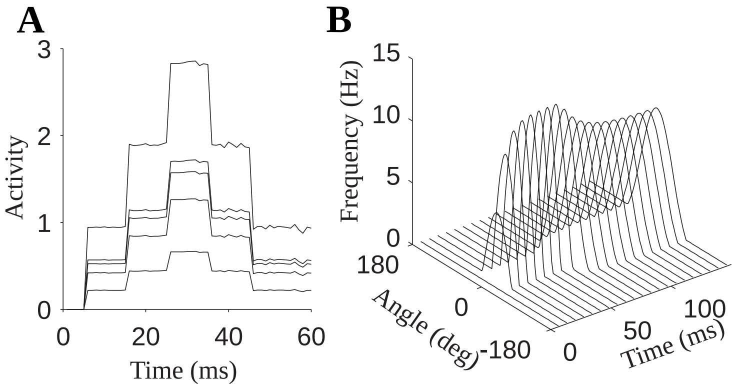

Figure 3: *A*, Intensity memory. Simulation of Eq. 3 for five neurons. The eigenvector e_k was [0 0.25 0.5 0.75 1]. Transient inputs (duration 2 ms) were delivered at times 5, 15, 25 (excitatory), and 35, 45 (inhibitory). *B*, Spatial memory. Simulation of Eq. 4. Matrix $\boldsymbol{W}$ was made of local Gaussian excitation and global inhibition. A tuned activity profile (width 40°) was presented at time 10 and replaced at time 50 by an untuned profile of the same amplitude.

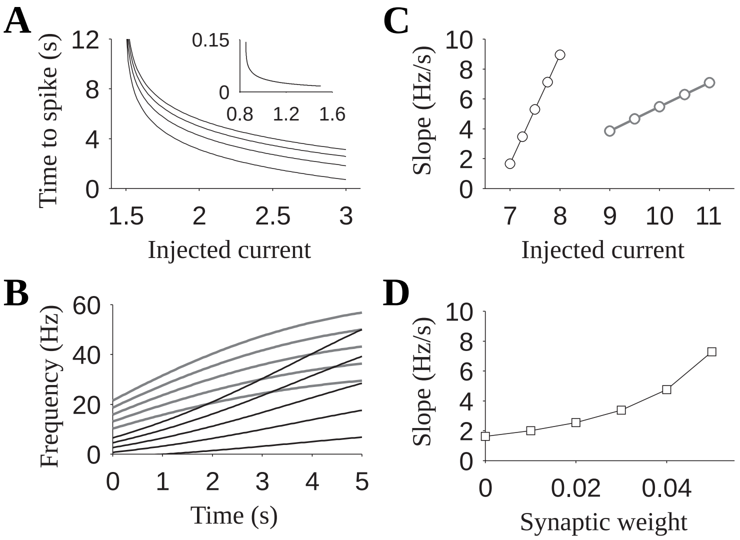

Figure 4: *A*, Time to the first spike as a function of the injected current in the presence of a Ks conductance in a Hodgkin-Huxley model. The curves correspond to different initial level of availability of the Ks conductance (maximal for the upper curve). The result in the absence of Ks conductance is shown in inset. *B*, Time course of frequency increase for different levels of injected current in a recurrent (dark lines) and nonrecurrent (gray lines) network. *C*, Slope of the frequency increase as a function of the injected current (linear regression on the results of *B*). *D*, Slope of the frequency increase as a function of the synaptic weight in the recurrent network. Details on the methods can be found in Delord *et al.* (2000) .