Selective Delay Activity in the Cortex: Phenomena and Interpretation

Daniel J. Amit1,2 and Gianluigi Mongillo1,3

1INFM, dip. di Fisica, Università di Roma, La Sapienza, Rome, Italy, 2Racah Institute of Physics, Hebrew University, Jerusalem, Israel and 3dip. di Fisiologia Umana, Università di Roma, La Sapienza, Rome, Italy

The present article does not intend to present technical progress nor recent successes in accounting for experiments, as this issue of the journal presents a rich inventory. Rather, the paper presents a retrospective reflection on the history of the subject; on the relation between the different aspects of the concepts and the phenomena involved; on its strengths and weaknesses; and on some future prospects. It is a tribute to an extremely rich and growing wealth of physiological phenomena and of interpretative concepts. Yet the extent of achievement is used to expose open questions, which appear to become ever deeper. It is also an attempt to make the subject a matter of discourse between biologists and modelers, without the distraction of technical details.

Introduction: the physiology of working memory states

We have explored in the monkey the electric activity of nerve cells of the prefrontal cortex and of its thalamic projection nucleus ... in the course of performance of delay response task. The basic hypothesis ... was that, if these structures are involved in ... transient memory function, their neurons should manifest distinct temporal variation of spike discharge related to the events ... in delay response trials (Fuster and Alexander, 1971).

This pioneering experiment has opened the way to three decades of study of working memory of ever increasing sophistication. A classic paradigm is the delay-match-to-sample (DMS) task (see e.g. Miyashita and Chang, 1988); a sample image is presented for a fraction of a second, followed by a delay interval of several seconds, then a test image is presented and the monkey is expected to indicate whether it is the same as the sample or different. In a different version, a cue is presented briefly in one of several (eight) directions on a screen. Following a delay, the monkey is required to make a saccade to the corresponding direction (Gnadt and Andersen, 1988; Funahashi et al., 1989). It must remember the direction in which a cue appeared (rather than the cue itself). While the monkey is performing such a task, the extracellular discharges of single neurons are recorded, to be analyzed offline. Usually, average discharge rates are estimated for each cell over each task period. The basic interpretative criterion is that any variation in spiking rate manifests the involvement of the corresponding cell in performing some aspect of the task.

Experiments of this type raise several, intricate questions. They engage the perceptive, the cognitive, the delayed mnemonic, the behavioral and the physiological. In the case of the DMS task, for instance, the monkey must be able to recognize images; it must maintain the memory of the image presented during the sample period, in order to bridge the delay; it must evaluate the similarity between the item in memory and the image presented during the test period, and then act accordingly. Finally, it must assimilate that there is a task to be performed. This entire syndrome is monitored, on the physiological side, by recording spikes from single cells.

There seems to be no reason to question the relation of strong variation in spike rates during the presentation of the cue stimuli and the perceptive element, at least in inferotemporal (IT) cortex. IT cortex is considered to be the final stage of the visual stream. Cells in IT show properties consistent with this hypothesized role: Most of them are visually responsive, i.e. they increase/decrease their spiking rates during the presentation of a stimulus. They also show selectivity in that they respond differently to distinct stimuli. Hence, the pattern of activity in IT could distinguish between different stimuli. Moreover, the enhanced selective activity persists following the removal of the stimulus along the entire delay interval, if the stimulus is familiar, i.e. repeatedly used in training. Figure 1 presents this phenomenon. Note, in particular, the reproducibility of the delay activity in different trials (raster by raster) with the same sample image, though these trials are interspersed with trials with other sample images; the rather sharp tuning curve of the cell, which has elevated delay activity only for 3–4 images out of 100 (Fig. 1c); and the absence of delay activity for the set of new images.

Selective delay activity, however, is not exclusive to IT cortex; it has also been observed in other regions as prefrontal (PF) cortex and entorhinal (ERh) cortex. The characteristics of neural responses and of the time course of the delay activity are quite different in different regions. This would indicate a functional dissociation among the corresponding areas. Until very recently single-cell recordings were carried out after training, i.e. after the monkey’s performance reaches a suitable high level, and hence it has not been possible to observe whether and how neural activity varies during training and to correlate the observed variations with the behavioral performance. Such information could furnish valuable insights into the functional role of the various regions expressing task-related neural activity. Because training increases substantially the performance level, it must be concluded that it modifies the neural circuitry involved. These modifications should be observable in new patterns of neural activity. As we discuss towards the end, such information has started to become available (see e.g. Erickson and Desimone, 1999; Mongillo et al., 2003).

Whether the improvement of behavioral performance is related to specific patterns of neural activity in a given cortical region or to the interaction of activity patterns in different regions is an open question. Logically, one is faced with several alternatives as to the scenario connecting training, performance and selective delay activity:

Cerebral Cortex November 2003;13:1139–1150; DOI: 10.1093/cercor/bhg103
1. that training serves to learn the representations of the discrete objects (as stable selective delay activity) as well as the task (long-term memory);
2. that long-term coding exists, but is innate or learned in natural experience of the animal, prior to the experiments, and training serves for the monkeys to learn the stimuli;
3. that there is no long-term coding of the stimuli. The representation expressed in selective delay activity is learned in a single-shot trial by trial (short-term memory), and training serves to learn the task.

These alternatives may not be sharp. They could be more a question of emphasis on particular aspects. For instance, the neural coding for the stimuli could be task dependent. In this case, the distinction between learning the stimuli and learning the task becomes arbitrary. A case in point is the color sensitivity of representations in IT cortex. It is often assumed that at the level of delay activity, cells in IT are color blind (see e.g. Miyashita and Chang, 1988). Yet, Fuster and Jervey (1981) present evidence of rather significant color modulation of delay activity rates in this part of cortex, and speculate that what counts is what is ‘behaviorally relevant’, i.e. useful to correctly perform the task.

The experiments of Miyashita and Chang (1988) indicate that for new images there is no selective delay activity in IT cortex (Fig. 1e), while it is observed for familiar ones (Fig. 1c). Yet they find that monkeys perform well the DMS task for novel images, after they perform well with images on which they were thoroughly trained. This leads to the conclusion that the active trace of the visual objects can be impressed in a single presentation (somewhere, apparently not in IT), and if the task is internalized, it suffices for effective performance. In this case, long-term coding for the images may not be required for successful execution of the task. But selective delay activity does appear following extensive training.

One concludes that following extensive training, selective delay activity encoding for a set of stimuli appears automatically, independently of its behavioral relevance. Its appearance is likely to be related to synaptic modifications produced by repetitive presentation of stimuli.

Another issue concerns the locus of the observed neural activity. While the physiologist has various, precise means for establishing the cortical coordinates of the recording site in which the delay activity is observed, there is no incontrovertible evidence that the delay activity is actually sustained in that area, i.e. that the area recorded from is not just a readout of another area which strongly projects into it. For example, the delay activities in IT and PF cortices are coupled, as are the back and forth flows between the two (see e.g. Higuchi and Miyashita, 1996; Tomita et al., 1999). Our position on the matter is to assume that the delay activity is sustained where it is observed. Such a position may be corroborated or contrasted on the basis of anatomical and/or physiological evidence, as in the references mentioned above. It may also be possible to confirm or reject it analyzing features concerning stimulus selectivity, task...
dependence and functional dissociation. But at the present state of knowledge, the cart must be put before the horses.

**Selective Delay Activity as Attractor**

The basic microscopic modeling ideas were substantially proposed by Hebb (1949), who says:

> Let us assume that the persistence or repetition of a reverberatory activity (or 'trace') tends to induce lasting cellular changes that add to its stability ... When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.

Hebb, however, did not support these basic ideas by specific models of dynamical systems exhibiting the postulated properties. They can be summarized as follows:

1. perceptive experience activates sub-populations in neuronal assemblies (increasing/decreasing the emission rate of their component neural elements);
2. pairs of activated cells potentiate the excitatory efficacy between them, while synapses from activated to non-activated cells are depressed;
3. the resulting synaptic system, formed in the population of neurons affected by the perceptive experiences, may reverberate — i.e. may sustain a pattern of neural activity, similar to the perceptive one, after the removal of the stimulus.

The observed delay activity may reflect reverberations, sustained by the synaptic structure, in such a neural assembly. The assembly modeled is a column of 1 mm in diameter of cortical tissue, hence composed of \(10^5\) spiking cells, 80% excitatory and 20% inhibitory, connected at random, with recurrent connectivity of about 10%, see e.g. Braitenberg and Schütz (1991). Every neuron also receives excitatory inputs from outside the column. These external connections are supposed to carry noise as well as signaling from other regions. Within the network, there are sub-populations of visually responsive cells (in IT \( \sim 2-5\% \) per stimulus). ‘Visual response’ is used here in a rather abstract and generic sense. It refers mainly to the fact that this region is still in the visual path, and hence enhanced selective activity, upon presentation of an image on the screen, can be considered visual. It implies neither retinotopy nor image feature correspondence. For a detailed discussion of the different levels of representation of a visual scene, see e.g. Amit and Mascaro (2003). The neurons in such a population increase their emission rate when the corresponding stimulus is presented. Structuring, Hebbian learning, is envisaged as potentiation of excitatory-to-excitatory synapses among cells responding to the same stimulus, and depression of the excitatory-to-inhibitory synapses among cells which respond to different stimuli. The resulting synaptic structuring, the final outcome of the repeated presentation of a given set of stimuli, renders the network capable of sustaining a large variety of different stationary distributions of neural activity. Such a distribution is stationary in that it reproduces itself: a pattern of activity persists on a time scale much longer than the neural or the synaptic time scales. Furthermore, the stationary state is an attractor in that relatively small variations on this state lead it back to the same state.

In particular, there are stationary states in which the set of (enhanced) delay activity cells is strongly correlated with the set of cells with visual response, and that delay activity is elicited by the corresponding image. The remaining cells are emitting at low rate. These states account for the selective delay activity, as observed by Miyashita and Chang (1988). As a matter of fact, each of these stationary states is ‘recalled’ from an entire set of initial conditions (basin of attraction), which are interpreted as noisy, incomplete or degraded versions of the perceived stimuli. This phenomenon, which is a strong hallmark of attractor dynamics, has been observed experimentally. This was done by varying continuously the visual stimuli (in a delay experiment), either by continuous noise superposition (Amit et al., 1997) or by morphing of images of different categories (Freedman et al., 2001). [We are grateful to Dr Saamah Abdallah for bringing this reference to our attention.] It was found that distribution of rates in delay activity remains unchanged, at the level of average emission rates. In other words, cells which have a given emission rate during the delay period following the ‘pure’ stimulus, have the same delay rate following the presentation of a degraded version of the image.

**Visualization of the Framework**

The conceptual framework is succinctly summarized in Figure 2 (we thank Dr Nicolas Brunel for this figure). The data in this figure are recorded in a simulation of a network of 6000 excitatory and 1200 inhibitory integrate-and-fire (spiking) neurons, randomly connected, each receiving external afferents which sustain spontaneous activity (see above). The parameters were chosen to ensure stable spontaneous activity and, following structuring, to sustain one of five different selective delay activity states. The figure presents spike rasters, average PST histograms and schematic system flow diagrams, in a fictitious network state space for an unstructured network and for a network with a set of imprinted working memory states. A small number of randomly selected cells has been chosen for exposition, as would be the case in reporting electrode recordings. At the center top it also presents a ‘bifurcation diagram’, which shows the rate in stationary states accessible to the network and the rate at the border between their basins of attraction, as function of the synaptic potentiation amplitude.

The column on the left represents life in the unstructured network, in which all synaptic efficacies are distributed at random with efficacies that ensure a given level of spontaneous activity. At the bottom is a flow diagram with a single basin of attraction, which expresses the fact that in this network, following the presentation and removal of any stimulus (provided it does not produce instant strong potentiation), the network relaxes to the same stationary state — spontaneous activity state — in which all neurons emit at relatively low fluctuating rates. The rasters at the top of the left column (panel marked I) the spike emission of 10, randomly selected, inhibitory neurons of the network, during 1 s (absissa). Below them are the rasters of 40 excitatory cells, all emitting at low rates. The five panels of four cells each (marked 1–5, to the left) are drawn from five sub-populations that will code for the working memory of five objects. At this stage the rates of these cells are the same as those of the others. Below the rasters is the PST histogram of the instantaneous rates in a moving bin, averaged over the recorded neurons in each of the seven cell groups. All six excitatory histograms overlap, fluctuating about at an...
average of 3 Hz, while the inhibitory cells have higher rates (as expected), ~10 Hz.

The bifurcation diagram shows that until a potentiation of 2.05 (relative to the average excitatory efficacy in the unstructured network) there is only one stationary rate in the network: spontaneous activity at 3 Hz for excitatory cells and 10 Hz for inhibitory cells. Then a bifurcation occurs and a new branch of states enters the picture, with rate starting at 12 Hz, represented by the rising full curve. These are the selective delay activity states. As seen in the diagram, these states coexist with spontaneous activity until a potentiation of ~2.3, at which point spontaneous activity becomes unstable. The higher, selective delay activity state exists for every sub-population in which potentiation has crossed the critical value (2.05). Which of these stationary states the system will relax to, depends on the stimulus. If it is similar enough to one of the imprinted populations, the network will have that population in elevated rate and the other neurons in spontaneous activity. If it is far from all populations, the network will relax to the global spontaneous activity state.

A Historical Narrative

The qualitative picture proposed by Hebb was progressively substantiated by formal models (Cragg and Temperley, 1954; Caianiello, 1961; Amari, 1972; Little and Shaw, 1974; Hopfield, 1982); particularly productive have been Amari (1972) and Hopfield (1982). The ideas originate in the physics of systems composed of a large number of simple interacting elements, where interactions yield collective phenomena, such as a stable
magnetic orientation, even in presence of significant levels of noise. The simplest of these is the Ising spin system, composed of a very large number of interacting binary threshold elements (spins), each of which can be in either an up or down state. From every initial global state (the combined states of all elements), the system flows, rather rapidly, to a ‘fixed point’. This is a self-reproducing collective state: The field felt by every spin, produced by all other spins, leaves it in its previous state. The physiological metaphor is: Neurons correspond to spins; the afferent current to a given neuron corresponds to the field, which is a weighted sum of the state values of all other neurons; the weight factors correspond to synaptic efficacies. If the afferent current is above a threshold, the neuron is spiking at high rate (the up state), otherwise it is spiking at low rate (the down state). A stationary state represents the active maintenance of a previously acquired memory (reverberating state, selective delay activity), passively stored in the pattern of interactions among the binary elements (synaptic structure).

As a model of memory, the main limitation of the Ising system is that there are only two stable states, or only two memories. Amari (1972), in the most underestimated contribution in the field, first perceived the way to embed a multiplicity of stationary states: the synaptic matrix is constructed by choosing the efficacy connecting a pair of elements (cells) as the sum, over all stimuli to be memorized, of the product of the activities of the states of the two units (the Hebbian mechanism), as proposed previously (Caianiello, 1961). Amari observed that if the binary neural state variables assume values +1 and -1, the interaction between two units could end up either positive or negative. In other words, inhibition is introduced to compete with excitation, generating ‘frustration’, which leads to a multiplicity of stationary states.

The robustness of these properties in the simple models suggests that they would be present when more neurobiological details are added. This has indeed proved to be the case, as we discuss above. These networks share with the Amari–Hopfield model the multiplicity of steady states related to the learned stimuli, as well as their attractor property. The structure of the space of states of the Amari–Hopfield model is readily accessible to analysis, despite its significant complexity (Amit, 1989, and references therein). The insights gained from these studies have led to the modeling of the cortical tissue with more realistic functional and anatomical characteristics. Much less analytical progress has been made in the analysis of the dynamics of realistic networks of spiking neurons, except in the case of non-overlapping stimuli, in mean-field theory (Amit and Brunel, 1997; Brunel, 2003). Yet, experience has led to some general and less technical qualitative insights, to which we now turn.

Amari (1972) proposed to model neuronal activities by binary neural variables that take on the values \( s_i = +1, -1 \) (active, silent),

\[
s_i(t + \tau) = \text{Sign} \left( \sum_{j=1}^{N} w_{ij} s_j(t) \right)
\]

where the function Sign(\(x\)) produces the sign of its variable. The argument on the right hand side may be considered the afferent current arriving at neuron \( i \) due to the activities of all other neurons, relative to the neuron’s threshold. He also showed that if stimuli are represented by the neural activities they generate \( (x_i^s, i = 1, ..., N) \), and are presented in a sequence, a Hebbian dynamical process suggested by Caianiello (1961) for the synapses, leads to a synaptic matrix of the form:

\[
w_{ij} = \frac{1}{P} \sum_{\mu=1}^{P} x_i^\mu x_j^\mu
\]

This is a sum of the contributions of \( P \) sequentially presented stimuli (the \( N \)-bit words \( x_i^\mu, \mu = 1, P \)), each contributing the correlation of the activities of the two neurons connected by the synapse, as in Hopfield (1982). Upon presentation of one of the stimuli, synapses connecting two neurons that have the same activation in the stimulus presented are increased in efficacy by \( 1/P \), while those connecting two neurons in opposite states decrease by \( 1/P \). Consequently, following the presentation of \( P \) stimuli, synapses can assume values between +1 and -1.

The neural dynamics driven by equations (1) and (2) is equivalent to the downhill drifts on a landscape surface, ‘energy’ given by

\[
E = -\sum_{i,j} w_{ij} s_i s_j
\]

which has a form as schematically shown in Figure 3, with the states corresponding to \( s_i = x_i^s \) at the bottoms of the valleys.

**Workings of Realistic Models**

One of the crucial elements in the functioning of realistic model networks is that of inhibition. It plays several roles: it renders spontaneous activity stable (van Vreeswijk and Sompolinsky, 1996; Amit and Brunel, 1997); it prevents the continuous spreading of elevated delay activity beyond the stationary state of a specific object, rendering recall selective; it limits the capacity of memory; it limits the number of objects that can be present contemporaneously in working memory; and it limits the extension of context correlations (see below). The qualitative arguments underlying the role of inhibition expose much of the life in a network of spiking neurons.

**Spontaneous Activity**

When a cortical module of spiking neurons is unstructured, or is not expressing working memory, the neurons typically emit at low rates, 1–5 Hz. This activity is very stable both against freezing – decay of all emission, as well as against epilepsy – a positive feedback leading to an indefinite rise of the rates of all neurons in the network. The stability of this state implies that if
rates of neurons decrease on average, there exist a mechanism that brings them back up; but if the rates go above the spontaneous rate, some mechanism brings them back down.

If in the vicinity of spontaneous rates, the recurrent net current received by an excitatory cell is positive (depolarizing), a fluctuation augmenting the rates of excitatory cells would lead to positive feedback, which will keep increasing the rates, provoking an epileptic runaway. If the same net current were inhibitory (hyperpolarizing), then a drop in the rates of excitatory cells would lead to a depressive runaway and a cessation of activity. The way out is that the recurrent net current be negative, preventing activity runaway, and the decay of activity be prevented by the external, excitatory afferents, which are independent of the recurrent dynamics.

Note that for recurrent inhibition to overweigh recurrent excitation, inhibitory neurons (or synapses) must be privileged, because their number is four times lower than that of their excitatory fellows. This can be (and seems to be) achieved by inhibitory synapses having larger efficacies (in absolute value) and/or by inhibitory neurons emitting at higher rates (having shorter time constants).

Object Delay Activity and Pattern Completion

Working memory (selective delay activity) of an object results from sufficient potentiation of the excitatory synapses among neurons with strong visual response to the object, as well as sufficient synaptic depression among neurons responding to different objects. Due to randomness in the recurrent connectivity and to stochastic emission activity in the network, the dynamics of the network is ‘noisy’. This noise could activate some cells which are not selective for the stimulus. However, since the synapses between neurons in the population in delay activity for the particular object and the neurons outside it are depressed, the inhibition produced by the rise in rate will overweigh the excitation, driving the invaded cells back to spontaneous activity. This is much as in the restoration of spontaneous activity described above. In other words, the inhibition suppresses the recruitment of cells belonging to a population coding for other objects, or for none [such spread would take place in purely excitatory models, such as those of Cragg and Temperley (1954) and Willshaw et al. (1969)]. Similarly, noise can sporadically silence cells that are part of the object’s reverberation. The collective contributions of the other cells in the population, will restore the activity of the cell.

During the presentation of a stimulus, the cells which respond may not coincide with the population that expresses enhanced delay activity after the stimulus is removed (the working memory prototype). Due to noise in the visual stimulus (intended or not) or fluctuations in connectivity, some cells may be visually responsive and not have elevated delay rate and others may not have visual response and participate in the reverberation with elevated rates. A large number of stimuli, relatively similar to each other, lead to delay activity in the same group of cells, following their removal. When the fraction of neurons excited in a sub-population with mutually potentiated synapses is sufficiently high, and those excited outside the population are not too numerous, the mutual collective contribution (current) to other neurons in the same population can overcome the resulting inhibition. Then, following the removal of the stimulus, the unactivated cells in the population get activated and those outside the population go back to spontaneous activity (Amit and Brunel, 1997). This goes also under the name of pattern completion or associative memory.

At the single-cell level one would observe neurons, as indeed one does, with strong visual response and no delay activity, and vice versa. Figure 4 shows all the possibilities of pattern completion, at the single-cell level, in the physiological situation.

The role that inhibition plays in the present framework is very similar to the role it plays in ‘winner-take-all’ networks (Ermentrout, 1992, and references therein). If the stimulus activates neurons corresponding to populations representing several familiar objects, the network’s dynamics will select the ‘closest’
one (in the sense of the basin of attraction), to settle in, eliminating activity corresponding to other objects. As we mention below, the winner may be more than one object, which finds its analogs in winner-take-all networks with multiple winners. What distinguishes the picture described here, from networks construed to implement this computational feature, is that here this property is a by-product of modeling, with realistic elements, the physiological phenomenon of selective delay activity.

Multiple Items in Working Memory

The above reasoning holds until, and unless, the number of invaded cells becomes a sufficient fraction of another population. In that case the network may find itself in a state with two or more populations in elevated delay activity. Large-scale invasion may take place due to either a rare fluctuation or the presentation of a second object while one is in active working memory. To render the multi-item attractor stable, the excitation must be such as to balance the extra inhibition induced by a double number of excitatory cells with elevated rates. Two factors may facilitate this process, one is higher potentiation, the other is a lowering of the delay rates. For sufficiently high potentiation within each population, a third population may also be drawn into working memory, etc. Eventually, the inhibitory feedback limits the maximal number of items that may coexist in working memory, because as delay rates decrease it becomes increasingly difficult to sustain their stability against a transition to spontaneous activity. This limitation of the depth of working memory, in terms of the number of concurrent objects that can be sustained, is an increase in selectivity of the network, as pointed out by Tanaka (2002a,b).

Multiple-item working memory states are the analog of the spurious states of the Amari-Hopfield network (Amit et al., 1985a; Amit, 1989). The joint activity of the neurons representing the working memory of several items, can provide the tool with which the module can transmit in time (across a delay) the engrams of several aroused items, required for future processing. They may be all required because it is not known, a priori, which item will lead to the successful execution of the task, as in e.g. Amit et al. (2003). Alternatively, they may be all required later, as when a sentence has to be formed from a group of pre-selected words.

Working Memory Capacity

In the Amari-Hopfield model the maximal number of items, (+1,-1)-words, orthogonal to each other that can be stored and recalled as attractors is given simply by a geometrical constraint, and is equal to the number of neurons (Amari, 1972). If the stored memories are randomly constructed, this number can be computed with high precision (Amit et al., 1985b, 1987; Amit, 1989). It is lower, but still proportional to $N$ (the number of neurons in the network). The limit is caused by the random overlaps between the stored memories, which give rise to noise in the retrieval currents of each memory. This noise increases with the number of stored memories, until the attractor corresponding to every pattern is destabilized.

In a network of spiking neurons, if populations do not share neurons, i.e. the tuning of each neuron is perfectly sharp, the maximal number of items that can be stored and retrieved individually, is the inverse of the fraction of neurons ($f$) with elevated rates in each population. In other words, it is the number of non-overlapping groups of $fN$ cells that the set of $N$ neurons in the network can be divided into, a purely geometrical constraint again.

When the populations are no longer distinct, the relevant computations are only beginning to make progress (Curti et al., 2005). But the logic can be roughly discerned, combining the arguments about the role of inhibition with the interference arguments from the Amari-Hopfield model. If the neurons of the different populations are selected at random, and if the fraction of neurons belonging to each population is $f$, among the $fN$ neurons with elevated rates in the working memory of a given object, there will be a fraction $f$, on average, belonging to a population corresponding to a second object. The neurons in the overlap between two objects have potentiated synapses to the remaining neurons coding for the second object. This enhances the probability that the second population transit to its reverberating state at high rate, and in doing so infect more neurons in other populations. The probability of spreading increases with the number of, randomly assembled, populations with potentiated synapses. If two populations catch fire, the danger for a third one is approximately doubled. Hence, one can reasonably expect that as the number of populations increases, spreading of elevated rates eventually take place on a large scale, provoking a rise in inhibition that may destroy the delay activity of all.

Hebbian Plasticity in More Structured Delay Tasks

The basic framework in which elevated rate distributions create a synaptic structure which can maintain selective delay activity, leads naturally to the generation of complex neural activity correlates, as observed in more elaborate tasks. One such experimental situation is pair-associate matching (Sakai and Miyashita, 1991; Asaad et al., 1998; Erickson and Desimone, 1999; Rainer et al., 1999). In these experiments the images involved are divided into fixed pairs. In each trial the sample image (predictor) is one image of a pair and the monkey is rewarded for recognizing, following a delay, the presentation of the other image of that pair (choice) – its pair-associate (see Fig. 5, row PACS).

Note that information about the pair-associate image, required for the performance of the task, cannot be impressed in a single shot. Furthermore, it is not present in the presentation of the sample image and hence it must be recalled on demand during the delay interval. This would indicate the existence of long-
term coding (memory) of the set of stimuli, created during training (the pairings are completely arbitrary).

In these cases one observes, as a neural correlate, a growing similarity (correlation) of the visual responses and delay activities of images that are pair-associates (Sakai and Miyashita, 1991; Erickson and Desimone, 1999). This would be the result of a long-term embedding in the synaptic matrix of the frequent succession of an image by its pair associate. The generation of correlations between delay activity distributions is envisaged in the following manner: once selective delay activity for the single stimuli becomes stable, it persists along the delay interval until the presentation of the subsequent stimulus. There will be a short time window in which the neurons coding for the predictor and neurons coding for the choice are both active at high rate, and potentiation can take place in synapses between the two sub-populations. If pairs of stimuli are systematically seen in a fixed temporal order, Hebbian learning eventually leads to similar mnemonic representations (i.e. overlapping patterns of firing rates) for the two stimuli. In this way memories become associated: The delay activities, as well as the visual responses will become increasingly similar, independently of how similar they were to start with. Recent theoretical studies (Brunel, 2003; Mongillo et al., 2003) and experimental data (Sakai and Miyashita, 1991; Erickson and Desimone, 1999) suggest that this scenario is indeed plausible.

This argument can be extended to account for the generation of context correlations in the DMS paradigm, with sample images presented in a fixed sequence (Amit et al., 1994; Brunel, 1996). (Historically, this experiment and the subsequent theoretical account preceded the pair-associate paradigm.) In this paradigm the delay activity of a sample image is followed (in 50% of the trials) by the same image (‘same’ trials, Fig. 6) of a neighbor in the sequence on the match trials. If delay activity is elicited automatically, the test stimulus in the ‘same’ trials will give rise to inter-trial delay activity identical to the intra-trial delay activity elicited by the same image.

There is evidence for this type of inter-trial delay activity in IT cortex in Yakovlev et al. (1998) as well as in prefrontal cortex in Ó Scalaidhe (1997). Hence, the inter-trial delay activity will play the role of the intra-trial one in the pair-associate task, and upon meeting the visual response to the sample image of the next trial, will generate a synaptic structure leading to increased similarity of visual responses and delay activity distributions for neighboring images in the sequence. With the resulting synaptic structure, when image number $k$ is presented, and followed by its delay activity, neurons in the population $k + 1$ will have enhanced rates as well, due to the potentiated synapses between neurons of the two populations. Those will, in turn, enhance the rates of neurons corresponding to $k + 2$, etc. (see e.g. Brunel, 2003). What stops this chain is inhibition again, which rises as the number of excitatory neurons with enhanced rates increases. The view that correlation of the neural representations can be generated naturally (and automatically) where delay activity overlies systematically with afferent stimuli also provides a clue to experiments such as task switch in midst delay by color change (Naya et al., 1996). In this paradigm the monkey is trained on the pair-associate task and the DMS task, with the same set of stimuli. The images used are divided in pairs and all first members of a pair are of the same color, color 1, and all second members are of the same color, color 2 (see Fig. 5).

The monkey is supposed to perform the DMS task, i.e. recognize the repetition of an image, if throughout the delay the empty square on the screen remains of color 1, and to switch to the pair-associate task if the color of the empty square changes to color 2. This, rather complex, situation falls naturally in the class of phenomena accounted for by the Hebbian structuring of attractor correlations. In this case, there would be cells coding for the color of the image. Cells coding for color 1 are common to all sample images (color 1) and cells coding for color 2 are common to all test (pair-associate) images. The color on the screen will nail down the delay activity of the sample image, if it remains fixed. And if the color switches to color 2, the cells responsive to color 2, together with the potentiated synapses between the sample population and its pair-associate one, will provoke the transition from the delay activity of the sample to that of its pair-associate (see Brunel, 2003).

### Perspectives on Working Memory

#### Transitions among Attractor States

Fluctuations in the neural spiking dynamics may produce transitions among different stationary patterns of activity, enriching significantly the phenomenology exhibited by attractor neural networks. Transitions among stationary patterns of activity could account for the time-varying profiles of delay activity, frequently observed in PF cortex. If two neural populations are consistently activated in a fixed temporal order, as for example in the pair-associate paradigm (see above), Hebbian learning leads to strengthening of the corresponding inter-population synapses. As a consequence, when one of the sub-populations is in its reverberating state, the neurons of the other will receive a stronger excitatory contribution, than without inter-population

![Figure 6. Context correlations in DMS with samples at fixed order. The task: top panel, a sequence of ‘fractal’ images, each given an SPN (serial position number) and trial after trial the sample stimulus is selected in the order dictated by this sequence, which closes circularly — after 6 comes 1. Second panel, the protocol of the DMS trials, trial number 2, sample (S) is image 2 in the sequence, followed by delay and then the test stimulus (T), same as sample in this case (match, 50% of trials). This is followed by an inter-trial interval and the presentation of image 3, as sample in the subsequent trial. The bottom panels are simulated PST rate histograms, for two cells, one with visual response to image 2 (top) and the other responsive to image 3 (see visual response, upon presentation of the two images). The inter-trial delay activity following the presentation of image 2 as test, in the top cell, overlaps with the visual response to image 3 of the bottom cell, leading to potentiation of the synapse.](https://academic.oup.com/cercor/article/13/11/1139/274140)
potentiation. This modifies the balance between excitation and inhibition in the associated population in the favor of excitation. In this situation, a smaller fraction of cells, belonging to the second population, which increase their emission rate by fluctuations, may provoke the activation of the entire population. The probability of occurrence of such a transition depends mainly on the level of noise in the system and on the strength of interpopulation connections. This mechanism would produce prospective delay activity, as observed, for example, by Erickson and Desimone (1999).

**Line Attractors Monitoring Gaze Direction**

An extension of the discrete delay activity distributions discussed here was introduced by Ben-Yishai et al. (1995) and Seung (1996). It accounts for the stability of gaze following a saccade, assuming a continuum of attractors, parameterized by the horizontal gaze angle. The passage from one attractor to another requires very slight stimulation force. What fixes the state of the network among the continuum of allowed ones, is a saccade command, followed by a stationary state in a new direction. This new direction can be maintained for quite some time, even in the dark. Then, slowly the network drifts effortlessly along the line of attractors, varying the angle. The slow drift corresponds nicely to the drift observed in eye position in the dark, following the saccade. Given that the natural noise in the neural dynamics causes slow transition between attractors (Rainer et al., 1999; Mongillo et al., 2003), the saccade convergence and drift could perhaps be subsumed in a network with correlated attractors for discrete eye directions (as in Funahashi et al., 1989). An interesting possibility is that the mechanism of inter-population potentiation associated with noise-driven transitions between discrete stationary states, may be a way in which the neural circuitry implements continuous line attractors (for continuous attractors, see e.g. Camperi and Wang, 1998; Compte et al., 2000; Renart et al., 2003). More importantly, the machinery of transitions might form the substrate of high-level cognitive operations (e.g. memory recall on demand, prediction of future stimuli) as well as of complex motor behaviors.

**The Role of NMDA**

Recently (Wang, 1999) emphasized the role of NMDA receptors in the functioning of selective delay activity, in particular in rendering it sufficiently stable. It appears that NMDA receptors have varied roles: They render state transitions smoother (see e.g. Mongillo et al., 2003); they render multi-attractors more stable (see e.g. Tanaka, 2002a,b) (Fig. 7); they are also instrumental in the preservation of active working memory against intervening stimuli (Brunel and Wang, 2001) as well as in the activation of multi-item working memory (Amit et al., 2003). Moreover, as reported by Williams and Goldman-Rakic (1995), they play a central role in determining the quality of working memory.

But NMDA receptors may have an even more crucial role in the process of learning, as the agents that facilitate the entry of Ca ions into the cell, to provide the essential post-synaptic variable relevant for plasticity (Shouval et al., 2002).

**Provisos**

**Poverty of Synaptic Plasticity Description**

The main limitation of the framework presented is that the synaptic plasticity mechanism is still rudimentary. Here the situation is much more difficult, and much less satisfactory, than that of understanding spike dynamics, because in-vivo information of how synapses behave in well controlled situations is...
unavailable. One is reduced to indirect evidence, as in vitro studies (e.g. Markram et al., 1997; Bi and Poo, 1998) and the buildup of selective spike rates where available, as in Rainer and Miller (2002). In obtaining a realistic plasticity mechanism, a more promising approach may be the application of constraints of plausibility, such as the locality of learning (two neurons per synapse), the simplicity and implementability of the synaptic device, time scales (for synaptic state memory), the need for homeostasis in learning (so that over learning does not occur), to arrive at a model for a synapse (see e.g. Fusi et al., 2000; see also Amit et al., 1998; Brunel et al., 1998). The resulting model could then be evaluated by its function on the systemic level. In other words, given a synaptic model satisfying the constraints, it should be embedded in a large network, the network subjected to a flow of stimuli as in a given experiment and performance of the network compared to experiment.

Large Inter-spike Interval Fluctuations

It appears, in experiment, that the variability of inter-spike intervals (CV) is sometimes larger in delay activity than in spontaneous activity (Compte et al., 2003). It requires a special effort to account for this fact in the framework described here, in which it has been assumed throughout that synaptic structuring takes place exclusively in synapses between excitatory neurons. This assumption is related to the fact that often neurons that have non-selective delay activity have spike-form of inhibitory neurons. Such a feature would only imply that structuring does not take place from excitatory to inhibitory neurons (see Bi and Poo, 1998), but it does not exclude the possibility that learning take place in synapses from inhibitory to excitatory neurons. Plasticity of inhibitory synapses could be responsible for enhanced CV in elevated delay activity.

This could, however, also be due to the low number of neurons participating in the reverberating activity, giving rise to large fluctuations in the current and, consequently, to more irregular emission. Alternatively, it could be due to collective oscillations in the population activity around the stationary emission rate, as in Brunel and Hakim (1999), which would increase the variability in spike emission beyond that produced by noise.

Functional Interpretation of Delay Activity

Another fundamental issue is the functional interpretation of the observed neural activities. There is little direct evidence that the information observed in delay activity is relevant to the behavioral performance. It has been reported (Funahashi et al., 1989) that the level (average spike rate) of the delay activity, related to the cue stimulus, is correlated with success in performance. This is reproduced in Figure 8, where the average PST histograms during the delay period of delayed cue directed saccades are given on the left.

The trials are divided into trials of correct responses and trials of error. The first are assembled in the upper panel and the second in the lower one. The average rate is clearly higher for the correct responses. This last study is complemented by Williams and Goldman-Rakic (1995), who found that the rate in delay activity was selectively modulated by blockage of the dopamine D1 receptors, supposedly affecting as a result the performance level in saccades following an intervening delay (see Fig. 8, right).

It may be the case that the task dependence of working memory is part and parcel of the coding of the task’s behavioral...
cortex is not affected by successive stimuli (Miller expresses the last object viewed, for which stable delay activity program. It appears that in IT cortex, persistent delay activity preserves the delay activity of many stimuli in a sequence (Brunel and Wang, 2001). It may also be that PF cortex preserves the delay activity of many stimuli in a sequence (Amat et al., 2003). Whether varying the task paradigm influences such attributes awaits clarification.

Very stimulating is the study of Rao et al. (1997), in which monkeys are briefly presented an image (cue), at the center of the screen. Following a delay, the cue is shown together with a distractor, each at one of four (randomly chosen) positions on the screen. Following a second delay, no objects are shown on the screen but only cues in the four directions. The monkey is rewarded for making a saccade to 90° in the direction in which the first object was, following the first delay, see e.g. Figure 9. In other words, the working memory of the first object should allow the identification of that object in the second stimulus, where it appears in a given direction, following the first delay. The working memory of the direction, should guide the saccade following the second delay. Cells were found in prefrontal cortex which code selectively both for the object and for the direction and hence could translate their activity for the object to an object for the direction, following the second delay. In other words, information of the object representation in working memory is converted into information about a specific motor direction, yet neither physiological evidence nor modeling have been proposed for its conversion into the motor system.

Epilogue

The phenomena and ideas exposed above are a living testimony that the study of selective delay activity is an inexhaustible source of insights and puzzles, both empirical and theoretical, into the working of the brain in tasks that involve a delay between stimuli and response (input and output), which may be considered of essence in characteristically cognitive experience.

Notes

In memoriam to a precious person, a leading scientist and an inspiring editor.

This work was supported by the INFM Center of Excellence grant ‘Statistical Mechanics and Complexity’ and by a Center of Excellence Grant ‘Changing Your Mind’ of the Israel Academy of Science.

Address correspondence to Daniel J. Amit, INFM, dip. di Fisica, Università di Roma, La Sapienza, Rome, Italy. Email: daniel.amit@roma1.infn.it.

References


