A flexible model of working memory

Flora Bouchacourt[1] **(floramb@princeton.edu)**
Timothy J. Buschman [1],[2] **(tbuschma@princeton.edu)**
[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ
[2]Department of Psychology, Princeton University, Princeton, NJ

Abstract

Working memory provides the workspace for holding and manipulating thoughts. It is flexible: we can hold anything in mind. However, typical models of persistent activity rely on tightly tuned attractors and do not allow for the flexibility observed in behavior. Here we present a novel model of working memory that maintains representations through random reciprocal connections between two layers of neurons: a selectively tuned layer and a randomly connected, untuned layer. As the recurrent interactions are unstructured, the network is flexible: it is able to maintain any input. However, adding multiple memories lead to interferences in the untuned layer, which result in a capacity limitation on the number of items that can be maintained. This is due to divisive-normalization-like reduction in neural responses coming from E/I balance in the network. Furthermore, it has been shown that time and load have a degrading effect on memory precision. Interferences in the network provide a possible mechanism for this psychophysical finding, as well as key neurophysiological results. Thus, we present a simple network model that allows for flexible representations while still capturing behavioral and neural hallmarks of working memory.

Keywords: Working memory; Neural networks

Working memory plays a critical role in cognition, decoupling behavior from the immediate sensory world. On the one hand, it is flexible. Indeed, one can hold anything in working memory and, more importantly, one can do it from the first experience. On the other hand, the capacity of working memory is limited to 4 items, a result found to be consistent in many paradigms in both humans and monkeys (Luck & Vogel, 1997; Buschman, Siegel, Roy, & Miller, 2011). Previous work has shown that the contents of working memory are reflected in the activity of neurons distributed across the brain, including prefrontal cortex, parietal cortex, and sensory cortex, as well as sub-cortical regions (Christophel, Klink, Spitzer, Roelfsema, & Haynes, 2017). These representations are thought to be encoded in the persistent activity of neurons (Barak & Tsodyks, 2014). Persistent activity is commonly modeled as a result of recurrent network interactions that generate stable fixed points in the network's dynamics (although some models rely on single cell biophysics). Other key neurophysiological properties are related to the dynamics of neural responses during working memory: e.g. the transient activity of a population of neurons, as well as "activity-silent" changes in the network (Stokes, Buschman, & Miller, 2017). Transient activity may reflect sequential activation of neurons and could co-exist with persistent activity either because the transient dynamics are orthogonal to the persistent representations (Murray et al., 2016) or they are balanced in a way that doesn't impact readout. Activity-silent representations may rely on short-term synaptic plasticity. However, many of these models fail to capture a defining characteristic of working memory – its flexibility. Models that rely on attractor or transient dynamics are inflexible; they must be finely tuned to remember specific items or the attractors must emerge through learning. Other models depend on synaptic plasticity or change in single-cell biophysics and show flexibility in representation, but do not account for the limited capacity of multi-item working memory seen in biology.

Here we propose a novel, flexible, model of working memory that relies on random reciprocal connections to generate persistent activity. As the connections are random, they are inherently untuned and do not need to be learned, allowing the network to maintain any representation. However, this flexibility comes at a cost - when multiple memories are stored in the network, they begin to interfere, imposing a capacity limit. Thus, our model provides a mechanistic explanation for the limited capacity of working memory; it is a necessary trade-off for working memory's flexibility.

The model

We model a simplified two-layer network of inhomogeneous Poisson spiking neurons (Fig. 1A). The firing rate of post-synaptic neuron i is a non-linear function of the weighted sum of all pre-synaptic inputs $r_i(t) = \Phi(\sum_j W_{ij} s_j(t))$ where W_{ij} is the synaptic strength from presynaptic neuron j, s_j is its synaptic activation, and Φ is a hyperbolic tangent. The first layer consists of 8 independent *sensory networks*, each of 512 neurons. These networks mimic simplified sensory networks, with neurons arranged topographically according to selectivity. Position around the ring corresponds to specific values of an encoded feature, such as orientation or color. Consistent with biological observations, connections within a sensory network have a center-surround structure: neurons with similar selectivity share excitatory connections while more dissimilar neurons have inhibitory connections (Fig. 1A, inset). Recurrent excitation in each of these networks is low such that they do not maintain memory by themselves. For simplicity, each sensory network is independent and can be thought of as reflecting stimuli at different locations in space. This allow us to vary working memory load. Note that, although independence of networks is not biological, we make this simplifying assumption in order to emphasize the impact of interactions

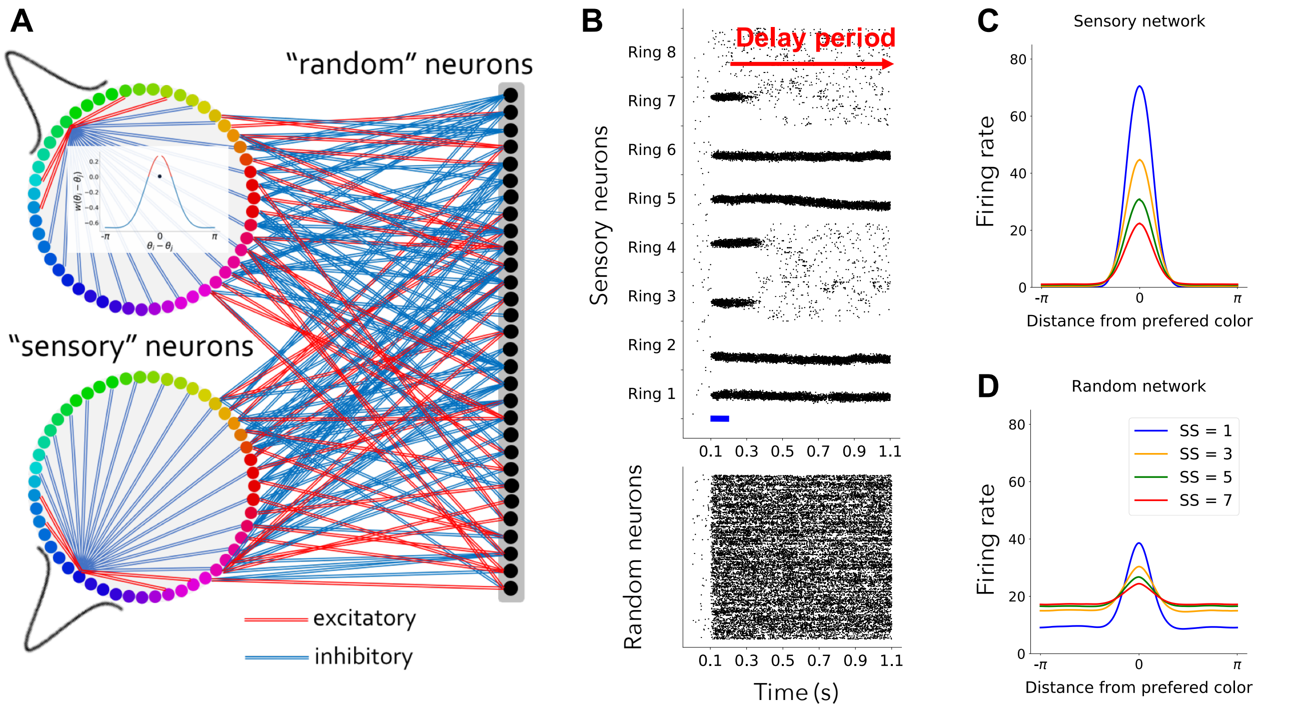

Figure 1: (A) Model layout. Only 2 (of 8) sensory networks are shown. The inset represents the center-surround connectivity structure of a sensory network. (B) Raster plot of a simulated trial. Seven sensory networks receive an initial Gaussian input from 0.1s to 0.2s (thick blue horizontal bar). Only four memories are maintained. (C) Average firing rate of a sensory network, at the end of the delay period, as a function of distance between the prefered color, and depending the set size (SS). (D) Average firing rate of the random network, at the end of the delay period, as a function of distance from the prefered color of a single sensory network, and depending on the set size.

in the second layer. The second layer is the *random network*, and is composed of 1024 neurons randomly and reciprocally connected to neurons in the sensory networks. In a presented version of the model, random neurons are not connected to each other (although we relax this constraint in future models). Each neuron in the random network has bi-directional excitatory connections with a random subset of neurons from the sensory networks (the default connectivity γ is 0.3). In between both networks, neurons receive balanced excitatory and inhibitory drive. To achieve this, all pairs of random and sensory neurons without excitatory connections have direct, weak, inhibitory connections, such that the sum total excitatory weights equals the sum total inhibitory weights for any given neuron. Note that both the bidirectionality and the balance constraints can be partially relaxed (not discussed here).

Importantly, all sensory networks converge onto the same random network. Neurons in the sensory networks show physiologically realistic tuning curves, due to their center-surround architecture (Fig. 1C). This tuning is effectively inherited by the random network (Fig. 1D). However, there is no consistency in the tuning of random neurons across inputs to different sensory networks. This is due to the fact that connectivity is random and therefore inconsistent across sensory networks. This leads to neurons in the random network showing “conjunctive” coding, preferentially responding to different inputs in different sensory networks (e.g. different colors at different locations, as observed in prefrontal neurons (Fusi, Miller, & Rigotti, 2016)).

Sensory stimuli are presented as Gaussian inputs during 100ms before the 900ms delay (Fig. 1B). Despite its simple architecture, and without involving any learning mechanism, the network is able to maintain stimulus inputs over an extended memory delay. This is due to the bidirectional connections between the sensory and random networks. Activity in the sensory network feeds-forward into the random network, activating a random subset of neurons. In turn, these random neurons feed-back into the sensory network, maintaining activity after the stimulus input is removed. In this way, the network is able to flexibly maintain the representation of any input into the sensory network. This relies on sufficient recurrent activity between the sensory and random networks. Given the connectivity $\gamma = 0.3$, we tune the feedforward excitatory strength such that one input to a sensory network is maintained (Fig. 2C, blue line, y-axis of the ROC plot), without creating an excessive top-down drive on the feedback pathway. Thus activity in the random network does not lead to sustained representations in other sensory networks (called “spontaneous” memories, Fig. 2C, blue line, x-axis of the ROC plot). As feedback connections are random and distributed over the sensory network, they “destructively interfere” at other locations, resulting in a slight increase in noise but no sustained representations. In other words, the feedback input from the random network to other sensory networks is orthogonal to the inherent dynamics of the sensory network and therefore is not maintained.

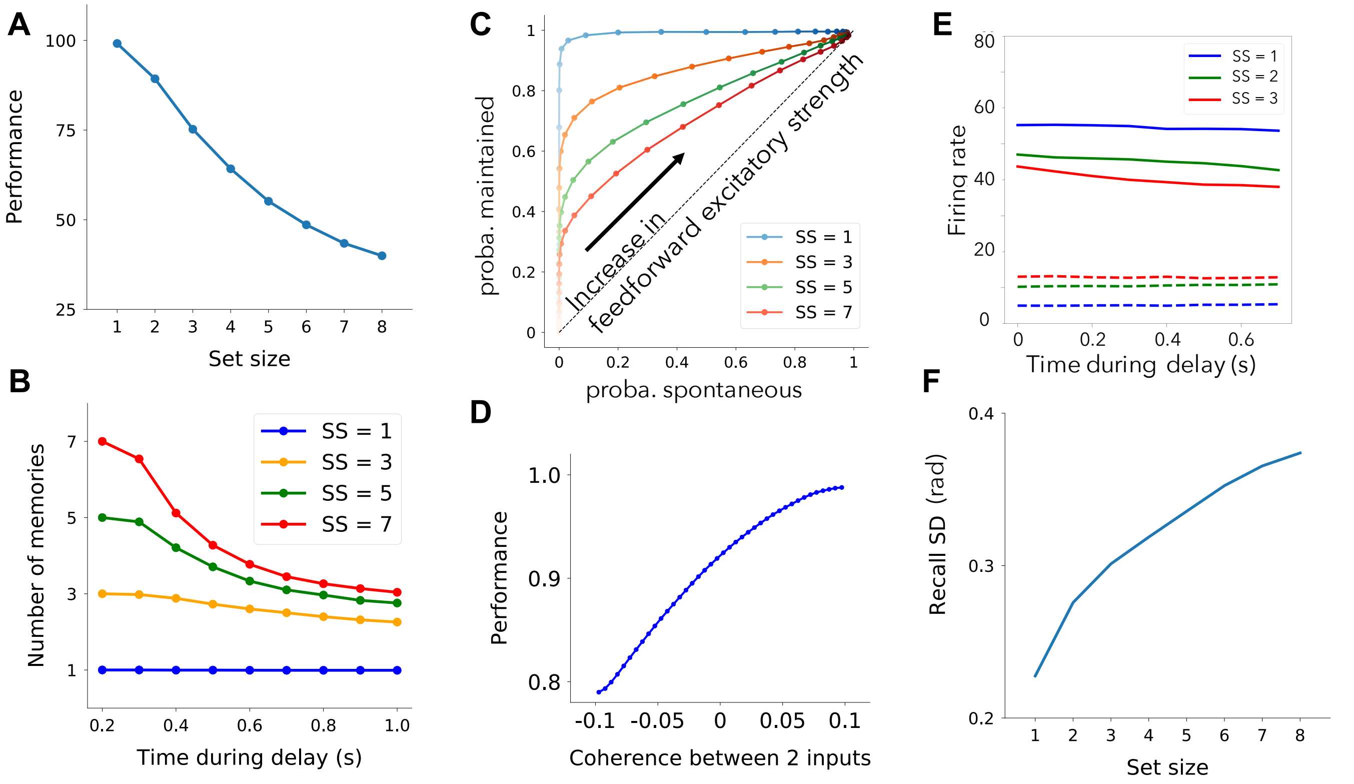

Figure 2: (A) Percentage of correct memories, after 1 second of network simulation, as a function of initial set size. (B) Speed of forgetting during the delay period as a function of the initial set size (SS). The speed is load-dependent. (C) ROC plot showing the probability of correctly maintaining a memory (hit rate) on the y-axis, as a function of the probability of creating a spontaneous memory in the other sensory networks (false alarm). Light to dark shades of colors stands for an increase of the net excitatory feedforward synaptic weight. Different colors represent different initial set sizes (SS). (D) Probability of correctly maintaining two memories in two sensory networks as a function of the coherence between the two initial inputs with respect to the feedforward weight matrix W to the random network. (E) Firing rate contrast for a subset of random neurons (about 10%), defined by a firing at a minimum difference of 40Hz between prefered (continuous line) and non-prefered (dashed line) inputs, and as a function of initial set size. Reproducing Fig. S2 of (Buschman et al., 2011)) (F) Standard deviation of the error distribution from maximum likelihood decoding after 1 second of network simulation, as a function of initial set size.

Interference between memory representations imposes a capacity limit

Multiple memories can be stored in the network simultaneously (Fig. 1B), but the capacity of the network is limited. Fig. 2A shows the percentage of correct memories, at the end of the delay period, as a function of load (initial set size). This closely matches behavioral results (Luck & Vogel, 1997; Buschman et al., 2011). Also consistent with behavior, the speed of forgetting during the delay period increases with load (Fig. 2B). Importantly, there is no feedforward excitatory weight that permits unlimited capacity for all possible initial set sizes without creating spontaneous, spurious memories (Fig. 2C).

For a few items, memories do not significantly overlap - as there is sufficient space in the high-dimensional random network for multiple patterns to be maintained. Some memories don't interfere. In general, a relatively small change in performance will be observed depending on the coherence of the inputs to different sensory networks with respect to their projection onto the random network (Fig. 1D). However, as the number of inputs to the sensory networks is increased, interference in the random network comes into play, causing memory failures. These are a result of the balance between excitation and inhibition into every neuron. As more inputs are presented in the sensory networks there is an increase in the effective inhibition on each activated neuron in the random network, suppressing its activity. This leads to a divisive-normalization-like reduction in neural responses as the number of to-be-remembered stimuli are increased (Fig. 2E), consistent with experimental observations (Buschman et al., 2011). As the level of activity is reduced, eventually it is insufficient to sustain the representation across both networks, and the memory is lost. Thus within the bounds of the random, reciprocal, and balanced connectivity implemented in the network, the limited capacity is a necessary trade-off for working memory flexibility.

Another key aspect of working memory capacity is the precision of analog recall as a function of load in human and monkeys. This measure have been important to assess whether working memory capacity is due to a limited amount of available slots, memorized with the same precision, or due to a shared ressource among items that compete for precision when the set size increases. It has been found that the error distribution variability increases steadily with the set size, arguing for the ressources model (Ma, Husain, & Bays, 2014). In our model, memory representations drift over time, due to the accumulation of noise from Poisson variability in neural spiking (Burak & Fiete, 2012). Greater interference in the random network leads to weaker feedback, increasing the effect of noise, thus causing greater drift in the representations. Memory representations degrade: the circular error increases as a function of memory load (not shown here). In addition the excitation/inhibition balance has the effect of reducing the signal to noise of neural activity, and thus the ability to accurately decode the memory representation in a finite amount of time (Fig. 2F). Together, these effects lead to an increase in memory error with increasing load and time, consistent with experimental results (Pertzov, Bays, Joseph, & Husain, 2013; Ma et al., 2014).

Stable and dynamic encoding of memories

The model also captures other key electrophysiological findings related to working memory. First, as noted above, the random nature of connections in our model yields high-dimensional, mixed-selective, representations in the random network; as has been seen in prefrontal cortex (Fusi et al., 2016). Second, our model shows the same combination of stable and dynamic representations seen in neural data. Heterogeneous activity can be simply added in the random network either by intrinsic recurrence, noise, or as a consequence of a transient stimulation preceding the delay period. Building on methods described in (Murray et al., 2016), we apply dimensionality reduction to the high-dimensional state space of neural activity in the random network in response to the presentation of 8 different stimuli values into the first sensory network. Fig. 3A shows the projection of the time-dependent neural activities into the mnemonic subspace composed by two leading principal components capturing stimulus encoding (Stimulus PC1 and PC2). The random network exhibits temporal dynamics during working memory. However, the interactions with the sensory network stabilize these dynamics to a mnemonic null space. The high-dimensional state space of the random network contains a low-dimensional subspace in which each stimulus representations are separable and stable across time during the delay period. This allows for a stable memory representation over time and linear readout. Moreover, the two leading principal axes provide quasi-sinusoidal coding of stimuli (Fig. 3B), remarkably reproducing the recent monkey eletrophysiology results cited above. The model goes even further by establishing a prediction on the effect of load on the mnemonic subspace. This subspace is stable, and a discriminator built from the network activity when presented with a single item could be used to decode network activity when the load is increased (Fig. 3C). However the discriminability between memories in this subspace decreases with load, making them harder to decode.

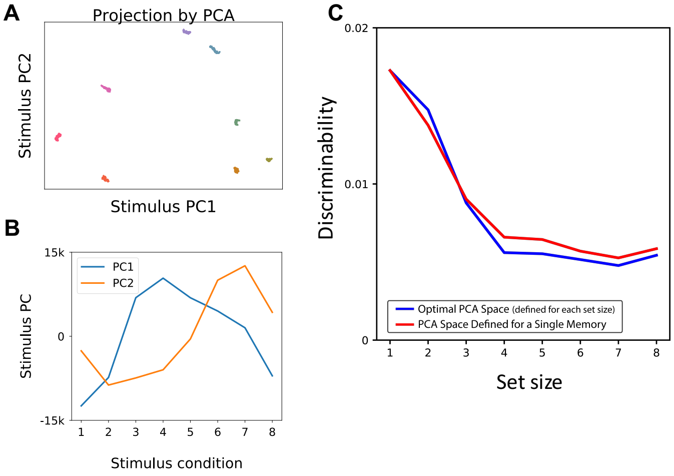

Figure 3: (A) Population trajectories during the delay period projected onto the mnemonic subspace. Each trace corresponds to a stimulus condition, and each point is a 100msec timestep. (B) Projections of the time-averaged delay activity along the leading principal axes. (C) Discriminability between memories as a function of load, computed by the euclidian distance between memories in the mnemonic subspace, corrected by the euclidian measure within each cluster for each memory over time.

This work was supported by ONR N000141410681 to TJB.

References

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

Burak, Y., & Fiete, I. R. (2012). Fundamental limits on persistent activity in networks of noisy neurons. *Proceedings of the National Academy of Sciences*.

Buschman, T. J., Siegel, M., Roy, J. E., & Miller, E. K. (2011). Neural substrates of cognitive capacity limitations. *Proceedings of the National Academy of Sciences*.

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J.-D. (2017). The distributed nature of working memory. *Trends in cognitive sciences*.

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current opinion in neurobiology*.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature neuroscience*.

Murray, J. D., Bernacchia, A., Roy, N. A., Constantinidis, C., Romo, R., & Wang, X.-J. (2016). Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proceedings of the National Academy of Sciences*.

Pertzov, Y., Bays, P. M., Joseph, S., & Husain, M. (2013). Rapid forgetting prevented by retrospective attention cues. *Journal of Experimental Psychology: Human Perception and Performance*.

Stokes, M. G., Buschman, T. J., & Miller, E. K. (2017). Cognitive control. *The Wiley Handbook of Cognitive Control*.