Neuron

Report

Synaptic Correlates of Working Memory Capacity

Highlights

- Multiple items can be brought to working memory via periodic brief reactivations
- Working memory capacity can be analytically estimated in the framework of this model
- Capacity scales as time of synaptic depression over that of synaptic current
- Stream of inputs can be segmented into chunks by modulating external excitation

Authors

Yuanyuan Mi, Mikhail Katkov,
Misha Tsodyks

Correspondence

misha@weizmann.ac.il

In Brief

Mi, Katkov, and Tsodyks derived an approximate analytical expression for working memory capacity in the framework of synaptic theory. This development predicts how manipulating the parameters of short-term synaptic plasticity and synaptic time constant will affect the capacity.

Mi et al., 2017, Neuron *93*, 323–330
January 18, 2017 © 2017 Elsevier Inc.
http://dx.doi.org/10.1016/j.neuron.2016.12.004

Synaptic Correlates of Working Memory Capacity

Yuanyuan Mi,[1,2,3] Mikhail Katkov,[2] and Misha Tsodyks[2,3,4,*]
[1]Brain Science Center, Institute of Basic Medical Sciences, Beijing 100850, China
[2]Department of Neurobiology, Weizmann Institute of Science, Rehovot 76100, Israel
[3]Department of Neuroscience, Columbia University, New York, NY 10032, USA
[4]Lead Contact
*Correspondence: misha@weizmann.ac.il

http://dx.doi.org/10.1016/j.neuron.2016.12.004

SUMMARY

Psychological studies indicate that human ability to keep information in readily accessible working memory is limited to four items for most people. This extremely low capacity severely limits execution of many cognitive tasks, but its neuronal underpinnings remain unclear. Here we show that in the framework of synaptic theory of working memory, capacity can be analytically estimated to scale with characteristic time of short-term synaptic depression relative to synaptic current time constant. The number of items in working memory can be regulated by external excitation, enabling the system to be tuned to the desired load and to clear the working memory of currently held items to make room for new ones.

INTRODUCTION

Working memory (WM) refers to short-term storage and manipulation of information (Miller et al., 1960; Baddeley and Hitch, 1974; Baddeley, 2003). There is hardly a cognitive task that does not involve WM, including visual processing, speech comprehension, and episodic memory (Cowan, 2001). Nevertheless, WM capacity is extremely limited, ranging between three and six items for most healthy human participants (Cowan, 2001; Fukuda et al., 2010; Luck and Vogel, 1997). It is often postulated that the brain possesses a specialized buffer, or "focus of attention," where memory items can be temporarily placed for short periods of time and removed when needed; hence, WM capacity corresponds to the size of this buffer (Cowan, 2001; Oberauer, 2002). The neuronal implementation of the focus of attention and its size, as well as the way memory items can be placed and removed from it, are not understood. The most popular hypothesis is that WM is mediated by persistent activity of neurons encoding the corresponding items in long-term memory (see, e.g., Compte et al., 2000; Wei et al., 2012; Edin et al., 2009). The maximal number of items simultaneously active depends on the characteristics of the network in a complex way, but there does not seem to be a fundamental upper limit on WM capacity in this model (Amit et al., 2003; Rolls et al., 2013). Another view posits that WM involves sequential activations of item representations (Cowan, 2010; Horn and Opher, 1996; Raffone and Wolters, 2001; Lundqvist et al., 2016, Lisman and Idiart, 1995). Lisman and Idiart (1995) suggested that WM involves periodic reactivation of memory representations at each gamma cycle within gamma-theta nested oscillations in hippocampus, mediated by slow after depolarization with time constant that should be matched to theta period. Assuming each memory is activated exactly once during a theta cycle and that the same memories are repeatedly reactivated over subsequent theta cycles (Siegel et al., 2009), the WM capacity is then estimated as a ratio of gamma and theta frequencies, which is compatible with earlier psychophysical estimates (Miller, 1956). The explicit dependence of WM capacity on the parameters of the model was not considered in this study, but it appears that the crucial factor limiting the capacity is the theta rhythm period, which could depend on several intrinsic cellular properties, e.g., nonselective cation channels (Colgin, 2013).

Recently, Mongillo et al. proposed a synaptic-based theory for short-term information storage in neural circuits (Mongillo et al., 2008; see also Lundqvist et al., 2011). In this model, memory is retained by item-specific pattern of synaptic facilitation. This mechanism does not require neurons to fire with elevated rate for the whole duration of the memory task, resulting in a robust and metabolically more efficient scheme. Several items can be maintained in the WM via consecutive brief reactivations of the corresponding neuronal groups. Here we aim to analytically estimate the maximal number of items that can be maintained in WM. The advantage of analytical expression is that it allows one to make predictions about how WM capacity depends on the various synaptic, neuronal, and circuit parameters that can potentially be tested by genetic manipulations.

A basic assumption of our model is that only one memory representation can be active at any single moment, which is guaranteed by strong reciprocal connections to a global nonspecific inhibitory pool (see Figure 1A below), consistent with experimental data (Fino and Yuste, 2011). If each memory representation had its own inhibition and hence was independent of the others, there would be no fundamental constraint to the capacity beyond the overlaps between the representations. We could think of several reasons against this idea: (1) the brain should then have an additional processing step that would disentangle the simultaneously active populations into a set of memories, and (2) each time a new memory is stored in the network, inhibition should adapt to this by generating a new population of inhibitory neurons that are specific to a new ensemble. This would make it much harder to add new representations, i.e., learning would suffer to improve WM capacity.

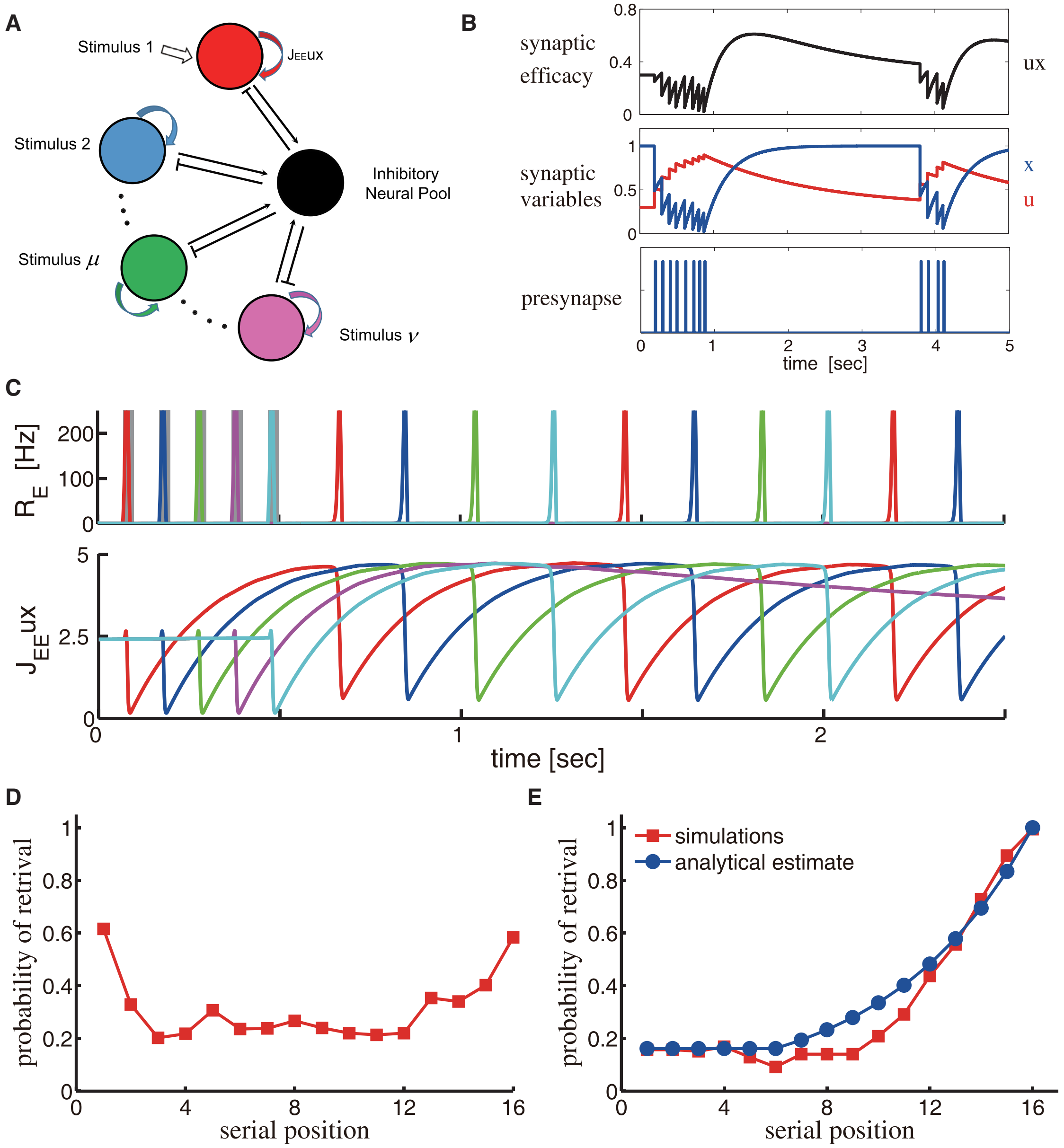

Figure 1. STP-Based WM Network Model

(A) Network architecture: a number of recurrent excitatory neural clusters, shown in different colors, reciprocally connected to an inhibitory neuron pool, shown in black.

(B) Model of a synaptic connection with STP. In response to a presynaptic spike train (lower panel), the neurotransmitter release probability *u* increases and the fraction of available neurotransmitter *x* decreases (middle panel), representing, synaptic facilitation and depression, respectively. The effective synaptic efficacy is proportional to *ux* (upper panel).

(C) Network simulation with five loaded memory items. Upper panel: firing rates of different clusters. Five clusters are sequentially stimulated by brief external excitation (shaded colored rectangles). Different colors correspond to different clusters as in (A). Following the stimulation, four clusters continue sequential activation in the form of PSs while the remaining item fades away. Lower panel: the instantaneous synaptic efficacy $J_{EE}ux$ for stimulated neural clusters during loading and subsequent reactivations.

(D) Probability of retaining an item in WM as a function of its serial position in a train of 16 loaded items, computed with 450 simulated trials with presentation frequencies between 16 and 66 items/s.

(E) Red: same as (D); computed with 450 simulated trials with presentation frequencies between 0.25 and 2 items/s. Blue: analytical calculation of the probability, as explained in the text and Supplemental Information. The parameters are as follows: $J_{EE}=8$, $\tau_f=1.5s$, $\tau_d=0.3s$, $U=0.3$, $\tau=8ms$, $J_{IE}=1.75$, $J_{EI}=1.1$, $\alpha=1.5$, $P=16$, and $I_b=3.0Hz$ in (C) and $I_b=8Hz$ in (D) and (E).

RESULTS

We consider a neural network model with memory items encoded by interconnected neuronal ensembles with short-term plasticity (STP) of recurrent connections (Mongillo et al., 2008). To achieve an analytical hold on capacity estimation, we drastically reduce the complexity of the network to leave only the most essential features that allow it to function as WM. We neglect the overlaps between different representations so each item is represented by a single excitatory unit (cluster) characterized by its activity rate, with self-excitation reflecting the strengthened connections between neurons encoding a given item in long-term memory.

Following the STP model developed in Markram et al. (1998), recurrent excitatory connections are characterized by fixed "absolute synaptic efficacy" and two dynamic variables: u, which stands for release probability, and x, the fraction of available neurotransmitters (Figure 1B). If J_{EE} is the absolute synaptic efficacy between two excitatory neurons, the instantaneous synaptic efficacy subject to STP is given by $J_{EE}ux$. Upon arrival of a spike, the release probability u temporarily increases, resulting in short-term facilitation. Meanwhile, the fraction of available neurotransmitters x decreases, resulting in short-term depression. After neuronal spiking, u returns to its baseline value U with a time constant τ_f, and x recovers to its maximum value $x = 1$ with a time constant τ_d.

In (Tsodyks et al., 1998), the STP model was used to derive the expression for the postsynaptic current resulting from the activity of a large, uncorrelated pre-synaptic population. The resulting network model has three differential equations for each of P excitatory clusters (synaptic current h_μ and two STP variables u_μ and x_μ for each cluster μ; $\mu = 1, \ldots, P$) and one additional equation for the inhibitory pool current h_I:

$$\tau \frac{dh_\mu}{dt} = -h_\mu + J_{EE} u_\mu x_\mu R_\mu - J_{EI} R_I + I_b + I_e(t), \qquad \text{(Equation 1)}$$

$$\frac{du_\mu}{dt} = \frac{U - u_\mu}{\tau_f} + U(1 - u_\mu) R_\mu, \qquad \text{(Equation 2)}$$

$$\frac{dx_\mu}{dt} = \frac{1 - x_\mu}{\tau_d} - u_\mu x_\mu R_\mu, \text{ and} \qquad \text{(Equation 3)}$$

$$\tau \frac{dh_I}{dt} = -h_I + J_{IE} \sum_\nu R_\nu, \qquad \text{(Equation 4)}$$

where τ is the neuronal time constant, for simplicity the same for excitation and inhibition; I_b is the constant background excitation that we assume to reflect the attentional state of the network (Zhang et al., 2014); and I_e is the external input used to load memory items into the network. As in Mongillo et al. (2008), we consider facilitating synapses with $\tau_f \gg \tau_d$. $R(h) = \alpha \ln(1 + \exp(h/\alpha))$ is neuronal gain chosen in the form of a smoothed threshold-linear function, also the same for excitatory and inhibitory neurons. The exact shape of the gain function is not important, but it should exhibit a tail for negative currents in order for the network to generate population spikes (see below; Tsodyks, 2004). The tail in the gain function could emerge either from noisy input that can cause some degree of firing for subthreshold levels of current, or when considering the effects of non-homogeneities in firing thresholds in neuronal clusters representing memory items.

To illustrate the proposed mechanism, we simulated the network with parameters that are compatible with experimental measurements of inter-pyramidal connections in the prefrontal cortex (Wang et al., 2006). We loaded five items into WM by applying transient external excitation inputs to the corresponding units and observed that four of them were maintained successfully in the form of brief reactivations called population spikes (PSs; Tsodyks et al., 2000), indicating that for this set of parameters, the capacity of WM is four (see Figure 1C). To further characterize the model, we subjected it to longer trains of stimuli at different presentation frequencies and computed the "retention curve," i.e., the probability that an item with a given serial position in the train is retained in WM at the end of the train. We observed a non-monotonic retention curve for high presentation frequencies: first and last items had a higher chance of being retained (primacy and recency effect, respectively; Figure 1D); for low frequencies, there was a pronounced recency effect that could be estimated accurately by assuming that each time a new item is presented, one of the previous items is erased from WM with equal probability (Figure 1E; see Supplemental Information, available online, for details).

As seen in Figure 1C, WM containing several items corresponds to the periodic activity state of the network ("limit cycle") where respective clusters emit PSs in the fixed order. As a dynamical system, the network exhibits multistability with many stable limit cycle solutions that evolve from different initial conditions. Rigorous analysis of coexisting limit cycle solutions is mathematically intractable, but this view allows us to address the issue of capacity numerically by simulating the network with random initializations until it converges to one of the periodic solutions (see Supplemental Information). In Table 1, we show the results of these simulations for different values of background input, with STP parameters that are compatible with experimental data from the prefrontal cortex (Wang et al., 2006). The table shows probabilities P_i that the network converges to a limit cycle with i sequentially activated clusters, computed with 200,000 random initializations of the network state at each background input value. P_0 corresponds to a baseline state with no PSs. The results show that (1) only when the background input is high enough can multiple items be kept in WM, and (2) with an increasing background input, WM capacity is gradually increasing. For network parameters chosen in these simulations, the maximal capacity is six, since further increase in background input leads to destabilization of the baseline spontaneous state of the network.

Even with the extreme simplifications described above, the model is still described by a relatively large number of parameters, namely STP synaptic parameters, neuronal gain functions, synaptic strengths, and time constant; hence, "brute force" numerical approach would be impractical and unrevealing. We now present the intuitive outline of the derivation of the final result (see Supplemental Information for more details).

The maximum number of items that can be maintained in WM is determined by the ratio of two factors: (1) the maximal period

Table 1. The Chances of Converging to a State with a Different Number of Items in WM for Different Arousal Levels

I_b	P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_7
2.4	1	0	0	0	0	0	0	0
2.45	0.9998	0.0002	0	0	0	0	0	0
2.5	0.9991	0.0008	0.0001	0	0	0	0	0
2.56	0.9950	0.0039	0.0010	0.0001	0	0	0	0
3.0	0.5668	0.1129	0.1420	0.1772	0.0011	0	0	0
3.7	0.0026	0.0151	0.0701	0.2872	0.6238	0.0012	0	0
5.5	0.0003	0.0008	0.0013	0.0242	0.2351	0.7015	0.0368	0
7	0.0002	0.0007	0.0008	0.0066	0.1187	0.6906	0.1824	0
14	0.0001	0.0004	0.0006	0.001	0.1387	0.8506	0.0086	0

The other parameters as in Figure 1.

$T_{\max}$ of the limit cycle of the network, i.e., the maximal time between subsequent reactivation of each cluster, and (2) the temporal separation between two consecutive PSs, referred to as t_s; the capacity of WM is given by the maximum number of PSs that can be accommodated in a single period of the limit cycle, i.e., by

$$N_C \approx T_{\max}/t_s. \qquad \text{(Equation 5)}$$

To estimate $T_{\max}$, we note that a PS is triggered by intrinsic instability due to recurrent excitation; hence, it is always produced by the cluster that has, at that moment, the largest effective recurrent strength $J_{EE}ux$. Since the time evolutions of the effective strengths for different clusters have identical shape triggered by corresponding PSs, and separated from each other by the time t_s, which is significantly shorter than the $T_{\max}$, we conclude that the longest time between activations of a cluster approximately equals the time it takes for the synaptic efficacy curve to reach a peak (see Figures 1C and S2). This is determined by the solution of Equations 2 and 3 above, which can be greatly simplified by neglecting the firing rate of a cluster between the PSs, turning them into linear equations, and resulting in the following expression:

$$T_{\max} \approx \tau_d \; \ln \frac{\tau_f/\tau_d}{1-U}, \qquad \text{(Equation 6)}$$

i.e., $T_{\max}$ is chiefly determined by the time constant of synaptic depression, depending weakly on other STP parameters.

The t_s has three components: the width of the PS of the previous item, the delay and the width of the inhibitory pulse triggered by this PS, and finally, the time it takes for a next cluster to recover from inhibition and initiate the new PS (Figures 2A and S3). It can be intuited that the first two components are proportional to the synaptic time constant τ (see Supplemental Information), and the third, dominant component should be found by solving Equations 1, 2, 3, and 4 above. To simplify these equations, we note that on the timescale of t_s, which is significantly shorter than $T_{\max}$, we can neglect the STP evolution of effective synaptic strength for a cluster that is about to emit a PS (Equations 2 and 3) and replace it with its maximal value, $J_{\max}$. If we also neglect the time evolution of inhibition between the PSs $(I_{\text{inh}} = J_{EI}R_I)$, we are left with the single dynamical equation for synaptic current of a cluster:

$$\tau \frac{dh}{dt} = F(h), \qquad \text{(Equation 7)}$$

$$F(h) = -h + J_{\max}R(h) + (I_b - I_{\text{inh}}). \qquad \text{(Equation 8)}$$

The function $F(h)$ that determines the flow of synaptic current is composed of two approximately linear branches with negative and positive slopes, respectively (Figure 2B). The strongly negative initial value h_0 is determined by the strength of inhibition triggered by the PS of the previously activated item. The time flow of the current due to Equation 7 will consist of two distinct segments: progressively slow recovery from inhibition until the minimum $h_{\min}$ of $F(h)$ is reached, followed by fast acceleration signaling the onset of the PS (see Figure 2B, red arrows). The time it takes for the cluster to emit the PS is thus given by the time for h to reach $h_{\min}$. Using the linear approximation for the negative branch, the analytical estimate for t_s can be computed as

$$t_s \approx \tau\left(\ln \frac{|h_0|}{I_b - I_{\text{crit}}} + C\right), \qquad \text{(Equation 9)}$$

where C reflects the contribution of PS width and inhibition duration, while $I_{\text{crit}} \approx I_{\text{inh}} - \alpha\ln(J_{\max} - 1)$ is the critical value for the background excitation for which t_s diverges logarithmically (see Supplemental Information). Combining the above analysis results in the following analytical estimate for WM capacity:

$$N_C \approx \frac{\tau_d}{\tau} \frac{\ln \frac{\tau_f/\tau_d}{1-U}}{\ln \frac{|h_0|}{I_b - I_{\text{crit}}} + C}. \qquad \text{(Equation 10)}$$

Two conclusions can be drawn from this result: (1) the WM capacity scales with the ratio of two time constants, one characterizing the synaptic depression and the other one synaptic current decay time, while also increasing with facilitating time constant in a weaker way via logarithmic term, and (2) the capacity is controlled by the background excitation that should be above the critical level below which no items can be maintained in WM. The second conclusion was already illustrated in the simulations presented above. To test the validity of the first

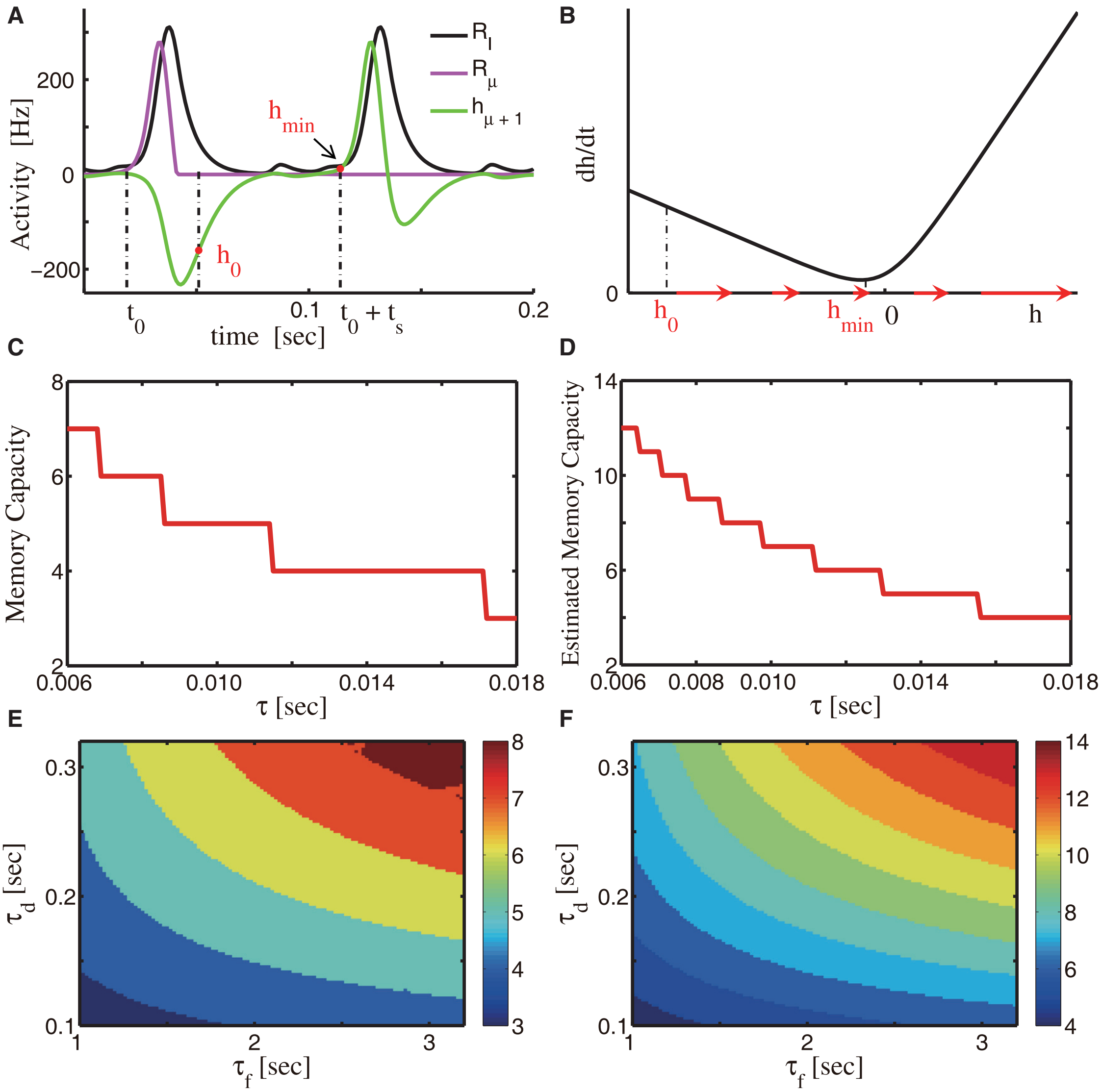

Figure 2. WM Capacity: Calculations and Numerical Results

(A) Activity of two consecutively activated excitatory clusters, μ and μ + 1, and the inhibitory cluster. The cluster μ elicits a PS at time t_0 (magenta trace), which triggers the response of the inhibitory neural pool (black trace), the latter generating negative synaptic current at the cluster μ + 1 (green trace). The cluster μ + 1 then recovers from inhibition until its current reaches the threshold point where next PS is emitted.

(B) The function $F(h)$ that determines the simplified one-dimensional dynamics of synaptic current according to Equation 7. Synaptic current recovers from initial hyperpolarized level h_0 to the slowest point of the flow at $h_{\min}$, after which the flow speed is increasing and a PS is generated. The speed of the synaptic current flow is illustrated with the red arrows.

(C) The WM capacity as a function of τ, obtained with numerical simulations of the model. τ_d and τ_f as in Figure 1.

(D) Same as (C); analytically estimated with Equations 5, 6, and 9. In Equation 9, we neglected the dependence of C, h_0, and I_{crit} on synaptic parameters of the model and replaced them by the constant values C = 4, $h_0 = -200Hz$, and $I_{\text{crit}} = 2.45Hz$. See Supplemental Information for details. τ_d and τ_f as in Figure 1.

(E) The WM capacity as a function of τ_f and τ_d obtained with numerical simulations of the model.

(F) Same as (E); analytically estimated as in (D). The parameters J_{EI}, J_{IE}, J_{EE}, U, α, P, and I_b are the same as Figure 1D.

conclusion, we simulated the model with various choices of STP timescales τ_f, τ_d, and synaptic time constant τ (see Supplemental Information for details of simulations). The results show that our analysis captured qualitatively the WM capacity: it decreases with increasing τ; as a function of STP timescales, capacity is increasing with both τ_d and τ_f in a way consistent with analytical estimate (see Figures 2C–2F). Simulation results are broadly compatible with analytical estimates obtained with Equation 10 (compare Figures 2C and 2E with Figures 2D and 2F), apart from a constant factor of about 2 by which analytical calculation overestimates the capacity. A closer inspection of Figure 2A reveals the origin of this discrepancy: when the synaptic current of the cluster that is about to emit the PS recovers from inhibition, the network exhibits an oscillatory activity epoch

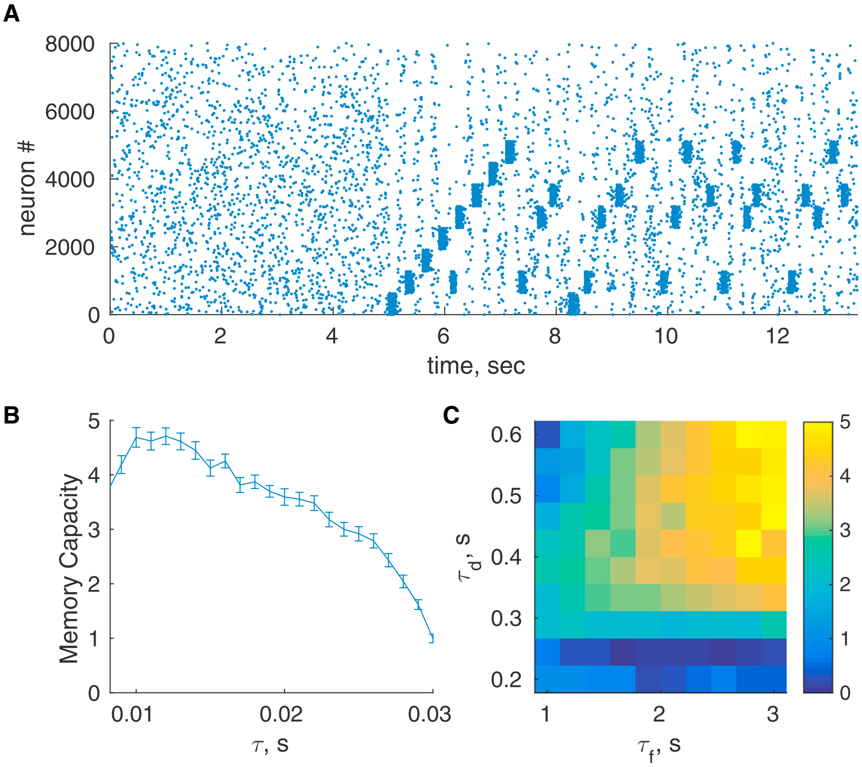

Figure 3. Spiking Neural Networks

(A) An example of network simulation, including spontaneous activity and WM triggered by loading ten stimuli. Spikes of 8,000 excitatory neurons are shown as dots; neurons are arranged in order such that the first 6,400 neurons are encoding ten patterns stored in the network. Out of those, eight items are loaded sequentially into the network after 5 s of spontaneous activity. Parameters are as follows: $\tau = 20ms$, $\tau_f = 3s$, $\tau_d = 0.6s$, and $U = 0.2$.

(B) The WM capacity as a function of τ, obtained with multiple simulations of the same network and averaging the results. Error bars: SEM.

(C) The WM capacity as a function of τ_f and τ_d. Parameters of the spiking network model are given in the Supplemental Information.

with transient increase of inhibition that delays the onset of the PS. This effect results from the nonlinearities in network dynamics and hence is not captured by our linear approximation used in estimating the inter-PS interval t_s (see Equation 9).

The model presented above was simplified greatly to allow for analytical estimates of WM capacity. It is therefore important to address the generality of obtained results. To this end, we simulated a more realistic spiking network of 20,000 integrate and fire neurons, 16,000 excitatory, and 4,000 inhibitory ones, with noisy inputs and probabilistic synaptic structure, considered in Mongillo et al. (2008) (see Supplemental Information for details of the network). Each memory is represented by a group of 640 excitatory neurons, 4% of the total number. An example simulation is shown in Figure 3A: the network maintains four items in WM; as opposed to the simplified model above, the corresponding populations are not activated with a constant frequency and the order of activations is variable. Overall, we found that the spiking network has a similar phase diagram with WM capacity increasing with external input up until the point where spontaneous steady state becomes unstable. By repeating simulations with different values of synaptic parameters, we observed similar trends of maximal WM capacity increasing with STP time constants τ_d and τ_f and decreasing with excitatory synaptic time constant τ (see Figures 3B and 3C).

Another unrealistic feature of the model concerns the absence of overlaps between representations of different memory items. One could conjecture that overlaps in representations have similar effects to direct excitatory connections between items in our simplified model. Our analysis presented in the Supplemental Information indeed confirmed this conjecture and showed that the resulting model has qualitatively similar dependence of WM capacity on synaptic parameters (see Figure S4).

The dependence of WM capacity on the background excitation enables the system to be efficiently "tuned" to the desired capacity; in particular, reducing the background below the critical value removes items from WM to make room for new inputs. This tuning is crucial for many cognitive tasks involving WM. In free recall experiments, it was shown that inserting pauses in the presentation of memory items allows human participants to encode and recall the incoming information in "chunks" of several items (Gilbert et al., 2014). We illustrate how a stream of inputs can be segmented into groups of simultaneously active items by a proper modulation of background input, if the pauses between the groups of inputs are long enough (Figure 4A). When the pauses are too short, the trace of previous items in the form of synaptic facilitation remains strong, so these items are reactivated when the background input is increased and the segmentation fails (Figure 4B).

DISCUSSION

In this contribution, we considered the issue of short-term memory capacity in the framework of synaptic theory of WM proposed in Mongillo et al. (2008). We derived a simplified analytical expression for the capacity in terms of basic synaptic parameters of the network where memory items are encoded and stored in long-term memory. Surprisingly, even though the WM trace in the model is maintained by synaptic facilitation, the derived expression shows that WM capacity is chiefly increasing with the time constant of synaptic depression and only weakly increasing with the time constant of facilitation; the obtained expression also shows that capacity is inversely proportional to neuronal time constants in the network. These analytical predictions could be amenable to experimental verifications by genetic manipulations of these parameters. Our results also show that WM can be regulated by external excitation that can tune the capacity to the desired level. This regulation is crucially

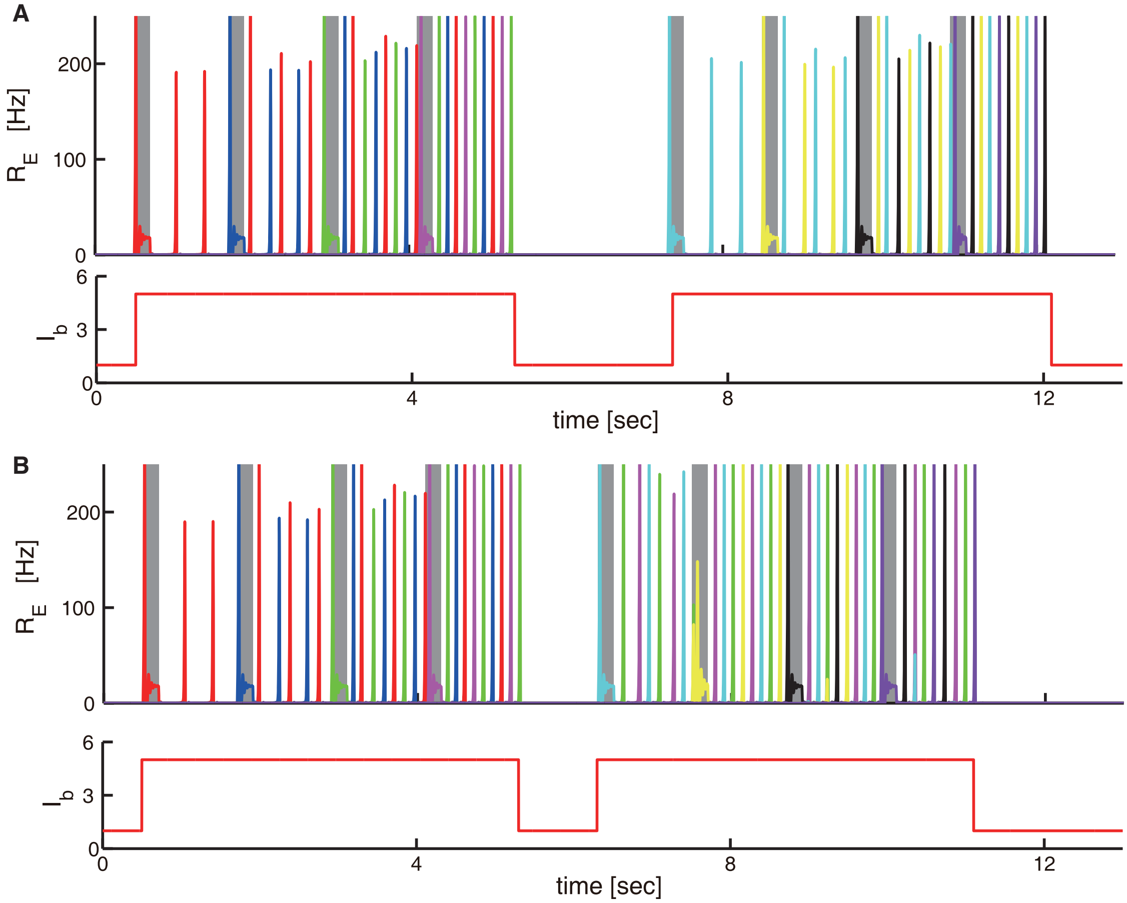

Figure 4. Segmenting a Stream of Inputs into Chunks via Modulation of Background Excitation

(A) Two groups of inputs of four items each are loaded sequentially with a pause between them (shaded color rectangles). Upper panel: activity of each cluster is shown in a different color. Lower panel: background input. After the fourth's input is presented, all four representations from the first chunk are active until the background input is reduced. When the background input is increased again and the next four items are presented, the second chunk is active after the eighth's memory is loaded.

(B) Same as (A), but with a shorter pause between two groups of inputs. The segmentation fails because when the background input is increased after the pause, the memories from the first chunk are reactivated and mix in with the second chunk. All other parameters as in Figure 1D.

important for WM functioning, which requires not only maintaining information in a readily accessible form but also clearing it out at the appropriate time to make room for future processing. One could reasonably assume that regulation of WM reflects attentionally driven inputs to the corresponding cortical areas. Indeed, imaging studies indicate that WM tasks trigger coordination between frontal and parietal cortical regions that are considered to be involved in WM and attention, respectively, and the degree of coordination is increasing with WM load (Honey et al., 2002). We simulated the network of integrate and fire neurons introduced in Mongillo et al. (2008) and observed qualitatively similar behavior of WM capacity on synaptic parameters, confirming the generality of obtained analytical estimates.

An interesting question raised by our results is what considerations could determine a particular set of synaptic parameters in the brain that result in the experimentally observed WM capacity of approximately four for most people. More detailed analysis of the model presented in the Supplemental Information shows that increasing the time constant of synaptic depression above a certain value brings the network to the regime where no PSs are possible and, hence, WM breaks down. The transition to this regime depends in a nontrivial manner on network parameters, but the maximal values of WM capacity tend to be substantially larger than four, both for simplified and spiking networks considered in this study. Another possibility is that WM capacity emerges from some yet unspecified functional tradeoffs. The exact nature of these tradeoffs would be difficult to determine, since networks that sustain WM could also be involved in many other cognitive processes. One potential tradeoff could be inferred from the results presented in Figure 4, where we show that in order to segment the stream of inputs into distinct chunks in WM, the pauses between the chunks should be long enough for the synaptic trace of the previous chunk to fade away, i.e., on the order of STP time constants that also determine the WM capacity. If those were made higher, segmentation of inputs would only be possible for slower presentation frequencies.

Since the WM capacity is defined by basic parameters of cortical networks, we predict that simple practice should not be enough to significantly improve the capacity beyond better tuning of attentional inputs to memory networks. In addition to having fundamental importance for understanding WM and its capacity, the analytical estimates obtained in this study might lead to new directions in clinical research of memory impairments associated with neurological disorders (Kenworthy et al., 2008; Levy and Farrow, 2001; Alloway, 2007).

EXPERIMENTAL PROCEDURES

Please see the Supplemental Information for full experimental procedures.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.neuron.2016.12.004.

AUTHOR CONTRIBUTIONS

Y.M. and M.T. designed and analyzed the model, and wrote the paper. Y.M. and M.K. performed numerical simulations.

ACKNOWLEDGMENTS

The study is supported by the EU FP7 (grant agreement 604102) and the Mortimer Zuckerman Mind Brain Behavior Institute. M.T. is also supported by Foundation Adelis. The authors thank L. Abbott and S. Fusi for helpful discussions.

Received: April 13, 2016
Revised: October 24, 2016
Accepted: November 11, 2016
Published: December 29, 2016

REFERENCES

Alloway, T.P. (2007). Working memory, reading, and mathematical skills in children with developmental coordination disorder. J. Exp. Child Psychol. *96*, 20–36.

Amit, D.J., Bernacchia, A., and Yakovlev, V. (2003). Multiple-object working memory—a model for behavioral performance. Cereb. Cortex *13*, 435–443.

Baddeley, A. (2003). Working memory: looking back and looking forward. Nat. Rev. Neurosci. *4*, 829–839.

Baddeley, A.D., and Hitch, G.J. (1974). Working memory. In The Psychology of Learning and Motivation, *Volume 8*, G.H. Bower, ed. (Academic Press), pp. 47–89.

Colgin, L.L. (2013). Mechanisms and functions of theta rhythms. Annu. Rev. Neurosci. *36*, 295–312.

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

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behav. Brain Sci. *24*, 87–114.

Cowan, N. (2010). The magical mystery four: how is working memory capacity limited, and why? Curr. Dir. Psychol. Sci. *19*, 51–57.

Edin, F., Klingberg, T., Johansson, P., McNab, F., Tegnér, J., and Compte, A. (2009). Mechanism for top-down control of working memory capacity. Proc. Natl. Acad. Sci. USA *106*, 6802–6807.

Fino, E., and Yuste, R. (2011). Dense inhibitory connectivity in neocortex. Neuron *69*, 1188–1203.

Fukuda, K., Awh, E., and Vogel, E.K. (2010). Discrete capacity limits in visual working memory. Curr. Opin. Neurobiol. *20*, 177–182.

Gilbert, A.C., Boucher, V.J., and Jemel, B. (2014). Perceptual chunking and its effect on memory in speech processing: ERP and behavioral evidence. Front. Psychol. *5*, 220.

Honey, G.D., Fu, C.H., Kim, J., Brammer, M.J., Croudace, T.J., Suckling, J., Pich, E.M., Williams, S.C., and Bullmore, E.T. (2002). Effects of verbal working memory load on corticocortical connectivity modeled by path analysis of functional magnetic resonance imaging data. Neuroimage *17*, 573–582.

Horn, D., and Opher, I. (1996). Temporal segmentation in a neural dynamic system. Neural Comput. *8*, 373–389.

Kenworthy, L., Yerys, B.E., Anthony, L.G., and Wallace, G.L. (2008). Understanding executive control in autism spectrum disorders in the lab and in the real world. Neuropsychol. Rev. *18*, 320–338.

Levy, F., and Farrow, M. (2001). Working memory in ADHD: prefrontal/parietal connections. Curr. Drug Targets *2*, 347–352.

Lisman, J.E., and Idiart, M.A. (1995). Storage of 7 ± 2 short-term memories in oscillatory subcycles. Science *267*, 1512–1515.

Luck, S.J., and Vogel, E.K. (1997). The capacity of visual working memory for features and conjunctions. Nature *390*, 279–281.

Lundqvist, M., Herman, P., and Lansner, A. (2011). Theta and gamma power increases and alpha/beta power decreases with memory load in an attractor network model. J. Cogn. Neurosci. *23*, 3008–3020.

Lundqvist, M., Rose, J., Herman, P., Brincat, S.L., Buschman, T.J., and Miller, E.K. (2016). Gamma and beta bursts underlie working memory. Neuron *90*, 152–164.

Markram, H., Wang, Y., and Tsodyks, M. (1998). Differential signaling via the same axon of neocortical pyramidal neurons. Proc. Natl. Acad. Sci. USA *95*, 5323–5328.

Miller, G.A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. Psychol. Rev. *63*, 81–97.

Miller, G.A., Galanter, E., and Pribram, K.H. (1960). Plans and the Structure of Behavior (Holt, Rinehart and Winston, Inc).

Mongillo, G., Barak, O., and Tsodyks, M. (2008). Synaptic theory of working memory. Science *319*, 1543–1546.

Oberauer, K. (2002). Access to information in working memory: exploring the focus of attention. J. Exp. Psychol. Learn. Mem. Cogn. *28*, 411–421.

Raffone, A., and Wolters, G. (2001). A cortical mechanism for binding in visual working memory. J. Cogn. Neurosci. *13*, 766–785.

Rolls, E.T., Dempere-Marco, L., and Deco, G. (2013). Holding multiple items in short term memory: a neural mechanism. PLoS ONE *8*, e61078.

Siegel, M., Warden, M.R., and Miller, E.K. (2009). Phase-dependent neuronal coding of objects in short-term memory. Proc. Natl. Acad. Sci. USA *106*, 21341–21346.

Tsodyks, M. (2004). Activity dependent transmission in neocortical synapses. In Methods and Models in Neurophysics: Lecture Notes of the Les Houches Summer School 2003, C. Chow, B. Gutkin, D. Hansel, C. Meunier, and J. Dalibard, eds. (Elsevier), pp. 245–265.

Tsodyks, M., Pawelzik, K., and Markram, H. (1998). Neural networks with dynamic synapses. Neural Comput. *10*, 821–835.

Tsodyks, M., Uziel, A., and Markram, H. (2000). Synchrony generation in recurrent networks with frequency-dependent synapses. J. Neurosci. *20*, RC50.

Wang, Y., Markram, H., Goodman, P.H., Berger, T.K., Ma, J., and Goldman-Rakic, P.S. (2006). Heterogeneity in the pyramidal network of the medial prefrontal cortex. Nat. Neurosci. *9*, 534–542.

Wei, Z., Wang, X.J., and Wang, D.H. (2012). From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. J. Neurosci. *32*, 11228–11240.

Zhang, M., Wang, X., and Goldberg, M.E. (2014). A spatially nonselective baseline signal in parietal cortex reflects the probability of a monkey's success on the current trial. Proc. Natl. Acad. Sci. USA *111*, 8967–8972.

Neuron, Volume *93*

Supplemental Information

Synaptic Correlates of Working Memory Capacity

Yuanyuan Mi, Mikhail Katkov, and Misha Tsodyks

Supplemental Information for

Synaptic Correlates of Working Memory Capacity

Yuanyuan Mi, Mikhail Katkov, Misha Tsodyks

1 Primacy and Recency Effect

To further characterize the model, we loaded 16 external stimuli sequentially at different presentation frequencies in each trail and then count the probability that an item with a given serial position in the train is retained in working memory (WM) after the removal of the stimuli. The duration of loading each external stimulus is fixed at $t_{\text{dur}} = 0.01s$. The presentation frequency is determined by the time interval between any two consecutive stimuli t_{int}.

In Figure 1D, 450 different presentation frequencies were chosen, where t_{int} is equally spaced in the range of $[0.005s, 0.05s]$. The background input is $I_b = 8Hz$ and other parameters are the same as Figure 1C. In the case of high presentation frequency, we observed the primacy and recency effect, that is, the first and last loaded items could be retrieved with higher probability (see the red curve in Figure 1D).

In Figure 1E, 450 different presentation frequencies are chosen, where t_{int} is equally spaced in the range of $[0.5s, 4s]$. All the other parameters are the same as Figure 1D. In the case of low presentation frequency, we observed the pronounced recency effect, that is, only the last few loaded items have higher probability to be retrieved (see the red curve in Figure 1E).

For the case of low presentation frequency, the probability of each stimulus with a given serial position γ in the train could be estimated accurately in the following way. Assume that working memory capacity is N_C, and when the $(N_C + 1)$th memory item is loaded, one of the previous items can be erased from WM with equal probability; and then by inference when the following memory item is loaded at each time, one of previous memorized items could be lost with equal probability. Therefore, the analytical estimated probability of each memory item in the train could be retrieved is (see the blue curve in Figure 1E)

$$P_\gamma = \begin{cases} \left(1 - \frac{1}{N_C}\right)^{(N_S - N_C)}, & 0 < \gamma \leq N_C; \\ \\ \left(1 - \frac{1}{N_C}\right)^{(N_S - \gamma)}, & N_C < \gamma \leq N_S. \end{cases}$$

where N_S is the number of presented external stimuli. We found that the simulation results and analytical estimates coincide with each other very well (see Figure 1E).

2 Numerical Estimation of the WM Capacity

2.1 Numerical Calculation of WM Capacity with Different Background Inputs in Table 1

We carried out numerical simulations to compute the memory capacity of the model. For each value of the background input I_b, we initialize the network with $200,000$ random initial conditions by setting u_μ and x_μ ($\mu \in [1, P]$), randomly and uniformly distributed in the ranges of $u \in [U, 1]$ and $x \in [0, 1]$, respectively, independently for each excitatory cluster. All activities are initiated at zero. Thus, different sets of initial conditions (i.e., $\{u_\mu,\ x_\mu,\ h_\mu,\ R_\mu,\ h_I,\ R_I,\ \mu \in [1, P]\}$) fall into different attractive basins of the network. We then simulated the network started from the $200,000$ different sets of initial conditions by integrating Eqs.1 - 4 in Main Text and counted the probability P_i of converging to a state with i different number of items in WM (see Table 1).

2.2 Numerical Calculation of Memory Capacities with Different Model Parameters in Figure 2

In Figure 2C, 121 different values of τ equally spaced in the range of $[0.006s, 0.018s]$ were chosen. The background input is $I_b = 8Hz$ and the other parameters are the same as in Figure 1.

In Figure 2E, 111 different values of τ_d equally spaced in the range of $[0.10s, 0.32s]$ and 111 different values of τ_f equally spaced in the range of $[1.0s, 3.2s]$ were chosen. The background input $I_b = 8Hz$, other parameters are the same as Figure 1.

Limited by computer capacity, we did not use the intensive search method used in Table 1 to calculate the memory capacity in Figure 2 for all different parameter values, rather we employed a simplified approach. We first calculated the theoretically estimated capacity $N_{C\text{estimated}}$ by solving Eqs. 5,6,9 in Main Text. We then applied a strong transient external input ($I_e = 565Hz$ with duration $t_{\text{dur}} = 0.015s$) to activate $m = N_{C\text{estimated}}$ excitatory clusters sequentially (as illustrated in Figure 1C), with the separation between two stimulations t_{int} satisfying $m(t_{\text{dur}} + t_{\text{int}}) \leq T_{\max}$. This strong transient input triggers each of m clusters to be in the limit cycle state. After removing the stimulation, if these clusters keep generating PSs sequentially, we increase the value of m by one and repeat the above loading process; otherwise we decrease the value of m by one and repeat the above loading process. The memory capacity is given by the maximum value of m the system can sustain after removing the stimulation.

To confirm the above approach, we used the intensive search method in Table 1 to calculate the memory capacities at 20 different random example points, and found that the two approaches agree with each other.

2.3 Limits on STP time constants

To delineate the existence of WM regime of bistability of PSs and fixed point, we increase the time constants of short-term facilitation (τ_f) and shot-term depression (τ_d) to very large values. To save time, we employed a simplified approach to estimating WM capacity: we applied a very strong transient input ($I_e = 365Hz$ with duration $t_{\text{dur}} = 0.015s$) to activate all the excitatory clusters sequentially (as shown in Fig.1C of Main Text), with the time interval between two consecutive external input $t_{\text{int}} = 0.05s$ and then counted the number of memory items that could be retrieved after the removal of external inputs. The results are shown in the Figure S1. One can see that WM regime disappears beyond the line in the space of (τ_f, τ_d), in particular for high enough values of τ_d. In Fig. S1, 191 different values of τ_f equally spaced in the range of $[1s, 20s]$ and 191 different values of τ_d

equally spaced in the range of $[0.1s, 2s]$ were chosen. The other parameters are the same as Figure 2E in Main Text.

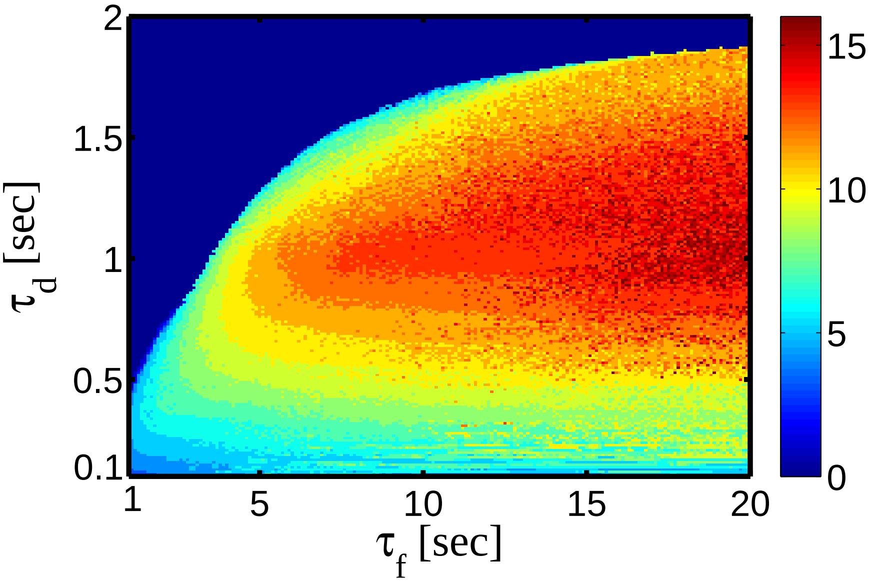

Figure S1: (Related to Figure 2) Working memory capacity dependence on neuronal and synaptic time constants τ_f and τ_d with numerical simulations of the model (Eqs. 1 - 4) of Main Text. In some regime, the network could not hold any memory items.

3 Theoretical Estimation of the WM Capacity

3.1 Maximum Time between Population Spikes of a Given Cluster ($T_{\mathbf{max}}$)(Related to Equation 6)

We analyze how the network dynamics, in particular the intrinsic properties of STP, restricts the WM capacity.

3.1.1 The stereo-typed dynamics of effective synaptic strength after a PS

After a strong transient PS in an excitatory neural cluster, its activity will remain silent until the synaptic efficacies recover and the cluster is ready to generate the second PS. During this period, since synaptic recovery of synapses follows its own intrinsic dynamics, the time evolution of the effective strength of a cluster displays a stereo-typed behavior.

Let us denote $t = 0$ to be the moment when the first PS is terminated, and assume that after the PS, the parameters u and x are reset to be $u = u_0$ and $x = x_0$, while the cluster activity is $R = 0$. Up until the onset of the second PS, the dynamics of u and x are thus driven by two independent linear equations,

$$\begin{aligned}\frac{du_\mu}{dt} &= \frac{U - u_\mu}{\tau_f},\\ \frac{dx_\mu}{dt} &= \frac{1 - x_\mu}{\tau_d}.\end{aligned}$$

Solving the above equations, we obtain the change of the effective synaptic strength over time:

$$Jux(t) = J_{EE}\left[U + (u_0 - U)\exp\left(-t/\tau_f\right)\right]\left[1 - (1 - x_0)\exp\left(-t/\tau_d\right)\right], \qquad (1)$$

illustrated on Fig.S2. It only depends on the STP parameters.

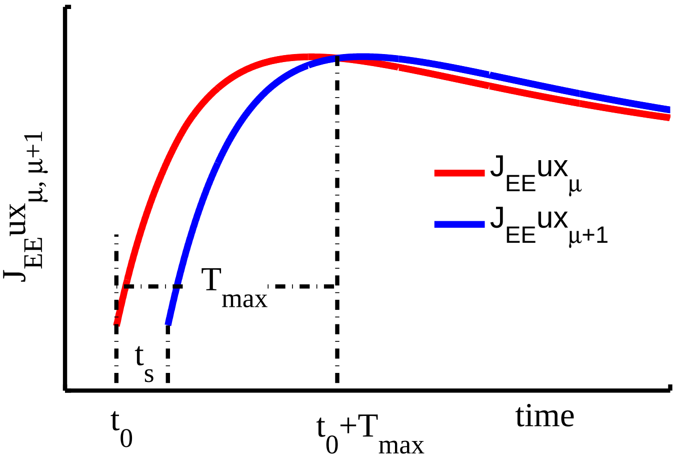

Figure S2: (Related to Figure 1) The dynamics of synaptic efficacy of two excitatory clusters which generate PSs consecutively with a time interval t_s. $t_0 + T_{\max}$ is the moment when the adjacent synaptic efficacy intersect with each other. The necessary condition for keeping the right retrieving order is that the synaptic efficacy of the cluster μ is always larger than that of the cluster $\mu + 1$ for the duration $t_0 + t_s < t < t_0 + T_{\max}$.

3.1.2 The maximum period of limit cycle

According to our model, each excitatory cluster generates PSs (retrieving a single memory item) sequentially with a period of T; all excitatory clusters are connected to an inhibitory neural pool, whose feedback induces competitions between excitatory clusters and regulates the timing of their PSs, so that multiple memory items can be recalled properly without interfering with each other.

Since all excitatory clusters receive the same background input and the same feedback from the inhibitory neural pool, when network emits a PS, it is from the cluster whose effective synaptic strength is the largest among all clusters at that moment. Hence the latest possible time for a cluster to emit the second PS is the moment when its effective synaptic strength declines to that of the cluster that is next in line (Fig.S2, see also Fig.1C of the Main Text). Since the time between consecutive PSs is significantly smaller than the period of the network, we can estimate the longest period $T_{\max}$ as the time from PS termination to the peak value of the effective synaptic strength. Using the Eq. 1 above, $T_{\max}$ can be found from the condition

$$\frac{dJux(t)}{dt}|_{t=T_{\max}} = 0.$$

Under the approximations of $\tau_d/\tau_f \ll 1$, the solution to this equation can be obtained in an analytical form:

$$T_{\max} \approx \tau_d \ln\left[\frac{\tau_f}{\tau_d}\frac{u_0(1-x_0)}{u_0-U}\right].$$

Initial values u_0 and x_0 depend on the precise form of the PS in a complicated way and hence cannot be computed analytically, however we can use the approximation $u_0 = 1$, $x_0 = 0$ that is satisfied if the PS is strong enough. This results in the Eq.6 of the Main Text.

3.2 The inter-PSs interval (t_s)(Related to Equation 9)

We analyze how the value of the inter-PSs interval t_s is determined by the network dynamics. Consider two clusters μ and $\mu+1$ that generate consecutive PSs. The inter-PSs interval t_s can be divided into three parts as shown in Fig. S3, which are

- **Interval A**: from the moment the cluster μ starts to generate a PS to the moment the PS terminates, i.e., the duration from t_0 to t_1 as shown in Fig. S3. During this period, the connected inhibitory neural cluster also exhibits PS-like activity burst due to receiving the strong excitatory input from the cluster μ, in turn suppressing the cluster $\mu+1$.

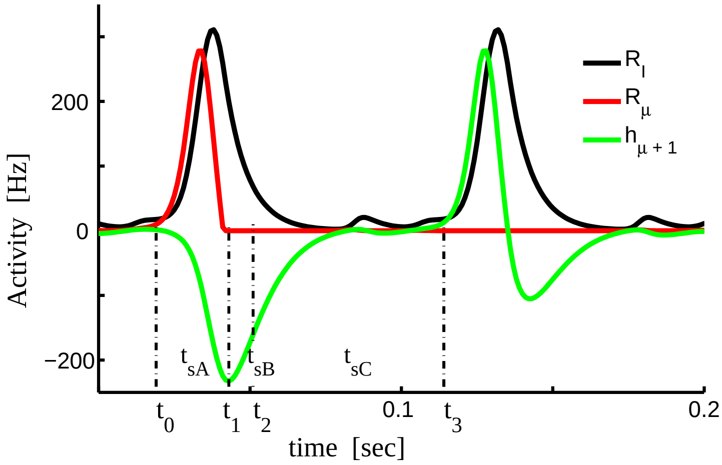

Figure S3: (Related to Figure 2) The inter-PS interval t_s can be divided into three different parts. Interval A: the cluster μ generates a PS; Interval B: the PS of the inhibitory cluster terminates; Interval C: the synaptic input to the cluster $\mu + 1$ keeps increasing to a threshold and starts to generate a PS.

- **Interval B**: from the moment t_1 to the moment t_2 when the activity of the inhibitory cluster terminates. During this period, the cluster $\mu + 1$ continuously receives a negative current from the inhibitory cluster.
- **Interval C**: from the moment t_2 to the moment t_3 when the synaptic input received by the cluster $\mu + 1$ increases until the new PS is emitted.

Below, we analyze the scaling of these intervals with network parameters.

3.2.1 Interval A

In this interval, the cluster μ generates a PS. Intuitively, its underlying mechanism is as follows: before generating a PS, the recurrent excitatory connection strength in the cluster reached the maximal value; because of receiving the homogenous background input, neurons fire mildly; however, since the reciprocal excitatory connections between neurons are strong, this weak neural activity is rapidly amplified by the positive feedback loop, leading to explosive responses of neurons; meanwhile, because of STD, the neuronal connections rapidly attenuate after explosive neuronal responses; and consequently, the neuronal responses terminate quickly, resulting in a transient PS of the neural population. It would naively appear that the PS duration should scale with synaptic depression time constant; however this intuition is wrong due to nonlinear dynamics of depression. Indeed, one can rewrite the Eq.3 of the Main Paper in the following form:

$$\frac{dx}{dt} = -\frac{x - x_\infty(R)}{\tau_d x_\infty(R)},$$

where

$$x_\infty(R) = \frac{1}{1 + u\tau_d R},$$

is the asymptotic value of depression variable that would be reached if R and u were constant. One concludes from this formulation that the *effective* time constant of synaptic depression, $\tau_d x_\infty(R)$ is much shorter than τ_d if the firing rate is very high and hence $x_\infty(R) \ll 1$. We conclude from this analysis that the duration of PS is mainly determined by the synaptic time constant τ, which is much smaller than the STP constants τ_f and τ_d.

3.2.2 Interval B

The duration of this interval is mainly determined by the delay for neural signals transmitted from an excitatory cluster to the inhibitory neuron pool, i.e., $t_{sB} \approx \tau$, and $t_2 \approx t_1 + \tau$.

3.2.3 Interval C

From t_0 until the cluster $\mu + 1$ emits a PS, its connection strength $J_{EE}u_{\mu+1}x_{\mu+1}$ is very close to the maximal value and does not change much, so we can approximate it as a constant, $J_{\max}$. Also the inhibitory current only changes mildly during interval. If we also neglect the time evolution of inhibition between PSs, i.e., $I_{\text{inh}} = J_{EI}R_I$, the dynamics of the cluster $\mu + 1$ can be approximated as

$$\tau\frac{dh}{dt} = -h + J_{\max}R(h) + (I_b - I_{\text{inh}}) \equiv F(h). \tag{2}$$

The function $F(h)$ has two branches: for negative h, it equals to $F(h) = -h + (I_b - I_{\text{inh}})$, and for positive h, it equals to $F(h) = (J_{\max} - 1)R(h) + (I_b - I_{\text{inh}})$. The whole function looks like a V shape with the minimum value at the point $h_{\min} = -\alpha \ln(J_{\max} - 1)$. When the synaptic current $h_{\mu+1}$ increases towards $h_{\min}$, it will slow down to pass through this point, and will then speed up to generate a new PS. The duration of this interval t_{sC} is then mainly determined by the time for $h_{\mu+1}$ reaching to the threshold $h_{\min}$ from an initial value h_0 at t_2, which is given by solving Eq. 2 and replacing $F(h)$ by its negative branch,

$$t_{sC} = t_3 - t_2 \approx \tau \ln\left(\frac{|h_0|}{I_b - I_{\text{crit}}}\right).$$

where $I_{\text{crit}} = I_{\text{inh}} - \alpha \ln\left(J_{\max} - 1\right)$.

In summary, the inter-PSs time t_s can be approximated as

$$t_s = t_{sA} + t_{sB} + t_{sC} \approx \tau \left(\ln \frac{|h_0|}{I_b - I_{\text{crit}}} + C\right). \tag{3}$$

In the above, the contribution of $t_{sA} + t_{sB}$ is absorbed in the constant $C\tau$.

In Eq. 3 the factors I_{crit}, h_0 and C depend on the parameters of the model, but this dependency is weak compared to the dominant dependency on τ. To generate Fig.2 of the Main Text, we therefore approximated these factors as constants and estimated the WM capacity as

$$N_C = \frac{T_{\max}}{t_s} = \frac{\tau_d}{\tau} \frac{\ln \frac{\tau_f/\tau_d}{1-U}}{\ln\left(\frac{|h_0|}{I_b - I_{\text{crit}}}\right) + C}$$

with constant values $C = 4$, $h_0 = -200Hz$, and $I_{\text{crit}} = 2.45Hz$.

3.3 Network with overlapping representations of memory items

To address the generality of obtained results, here we consider a neural network model with distributed overlapping representations of memory items, which consists of N excitatory neurons with STP effect and global inhibitory neurons. Each memory is represented by a randomly selected sparse group of neurons, mathematically described by a binary N-dimensional vector of zeros and ones:

$$\eta^{\mu\in[1,P]} = \underbrace{01100101110\ldots001}_{N\text{neurons}}, \quad \eta_i^\mu = 1 \text{ iff neuron i encodes memory } \mu,$$

where for each bit a value of 1 is chosen with the probability f and 0 with probability $1-f$, independently for each neuron and each memory. The sparseness parameter f determines the average number of neurons in each memory representation (fN) and the average size of the overlaps between different memory patterns (f^2N). For the connectivity, we assume that each pair of neurons that jointly represent a certain pattern are connected with the connection strength J_{EE}, while all other pairs are unconnected:

$$J_{ij} = \begin{cases} J_{EE}, & \sum_{\mu=1}^{P} \eta_i^\mu \eta_j^\mu \geq 1; \\ \\ 0, & \sum_{\mu=1}^{P} \eta_i^\mu \eta_j^\mu = 0. \end{cases}$$

The local dynamics of sngle neurons is given by

$$\begin{aligned} \tau \frac{dh_i}{dt} &= -h_i + \sum_{j=1}^{N} J_{ij} u_j x_j R_j - J_{EI} R_I + I_b + I_e(t), & (4) \\ \frac{du_i}{dt} &= \frac{U - u_i}{\tau_f} + U(1 - u_i) R_i, & (5) \\ \frac{dx_i}{dt} &= \frac{1 - x_i}{\tau_d} - u_i x_i R_i, & (6) \\ \tau \frac{dh_I}{dt} &= -h_I + J_{IE} \sum_{j=1}^{N} R_j. & (7) \end{aligned}$$

where τ is the time constant of excitatory and inhibitory neurons. h, R are the synaptic current and firing rate of excitatory and inhibitory neurons, respectively. $R(h) = \alpha \ln(1 + \exp(h/\alpha))$ is neuronal gain chosen in the form of a smoothed threshold-linear function. I_b is the background excitation input and I_e is the external input. u and x refer to the short-term facilitation and depression effect, respectively.

Instead of directly simulating the above equations, we employ a dimensionality reduction scheme that significantly reduces the simulation time for large networks. We notice that all the neurons that participate in encoding of the same set of memories, i.e. have same sets of P binary components ($\eta_i^\mu \equiv \eta^\mu, \mu = 1, ..., P$), have idential connections to all the neurons in the network (see Eqs. 3.3). Hence their synaptic inputs are all equal to each others, as can be seen from the Eqs. 4, and thus their STP parameters u and x also converge to the same alues. The network can thus be divided into groups specified by binary P-dimensional vectors η. Instead of simulating Eqs. 4 for all neurons, we define the currents and STP variables for each population η as h_η, u_η and x_η, respectively. These variables satisfy the following set of dynamic equations:

$$\begin{aligned}
\tau\frac{dh_\eta}{dt} &= -h_\eta + \sum_\beta J_{\eta\beta} P S_\beta u_\beta x_\beta R_\beta - J_{EI} R_I + I_b + I_e(t), && (8)\\
\frac{du_\eta}{dt} &= \frac{U - u_\eta}{\tau_f} + U(1 - u_\eta) R_\eta, && (9)\\
\frac{dx_\eta}{dt} &= \frac{1 - x_\eta}{\tau_d} - u_\eta x_\nu R_\eta, && (10)\\
\tau\frac{dh_I}{dt} &= -h_I + J_{IE} \sum_\beta R_\beta. && (11)
\end{aligned}$$

Here S_η is the fraction of neurons in a given population η.

The effective connection matrix between populations is given by

$$J_{\eta\beta} = \begin{cases} J_{EE}, & \sum_{\eta=1}^{P} \eta^\mu \beta^\mu \geq 1; \\ \\ 0, & \sum_{\mu=1}^{P} \eta^\mu \beta^\mu = 0. \end{cases}$$

The system of Eqs. 8 - 11 is the reduction of the original system of Eqs. 4 - 7, it has $2^P - 1$ equations instead of N. We can further reduce the dimensionality of the system by only considering the populations η with one or two nonzero components η^μ only, i.e. neglecting neurons that encode more than two memories simultaneously, which is a good approximation in the limit of very sparse conding $f << 1$. This results in the system of $P + \frac{P(P-1)}{2} = 136$ equations.

To test the generality of the dependence of WM capacity on synaptic parameters, we simulate the network model of Eqs. 8 - 11 with various choices of STP time scales τ_f, τ_d and synaptic time constant τ, and the results are shown in Figure S4 AC. It is found that capacity in the network with overlapping representation of memory patterns has similar dependence on the neuronal and synaptic parameters.

To better understand the obtained results, we conjecture that overlaps among different memory patterns have similar effects to adding excitatory connections among excitatory clusters in the simplified model of Main Text with the strength proportional to f. The rationale for this conjecture is that out of fN neurons encoding a given population μ, f^2N neurons, i.e. fracion f, also encode another population β, and thus, according to the connectivity scheme proposed above, these neurons are connected with the strength J_{EE} to neurons from both populations. The reduced system is thus described by the following

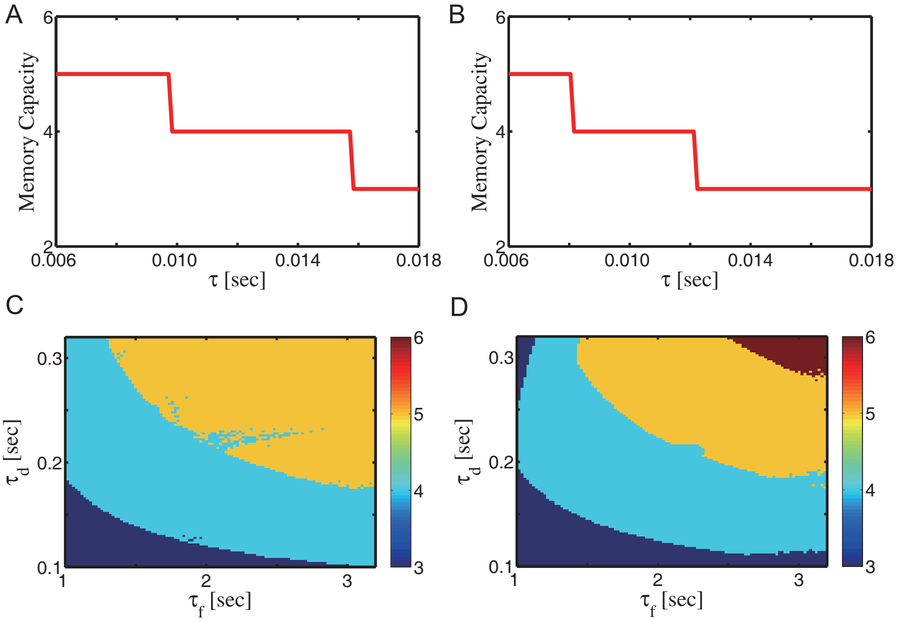

Figure S4: (Related to Figure 2) Working memory capacity dependence on neuronal and synaptic time constants with overlapping memory patterns. (A) The WM capacity as a function of τ, obtained with numerical simulations of the network model (Eqs. 8 - 11). The parameters are : $\tau_d = 0.3s$, $\tau_f = 1.5s$, $J_{EE} = 7.5$, $J_{IE} = 2.2$, $J_{EI} = 1.1$, $U = 0.3$, $\alpha = 1.5$, $I_b = 8Hz$, $P = 16$, $f = 0.05$. (B) The WM capacity as a function of τ, obtained with numerical simulations of the network model (Eqs. 12 - 15). $f = 0.05$, the other parameters are the same as (A). (C) The WM capacity as a function of τ_f and τ_d obtained with numerical simulations of the model (Eqs. 8 - 11). $\tau = 0.008s$, the other parameters are the same as (A). (D) The WM capacity as a function of τ_f and τ_d obtained with numerical simulations of the model (Eqs. 12 - 15). $\tau = 0.008s$, the other parameters are the same as (B).

system of $3P + 1$ equations:

$$\tau \frac{dh_\mu}{dt} = -h_\mu + \sum_\beta J_{\mu\beta} u_\beta x_\beta R_\beta - J_{EI} R_I + I_b + I_e(t), \tag{12}$$

$$\frac{du_\mu}{dt} = \frac{U - u_\mu}{\tau_f} + U(1 - u_\mu) R_\mu, \tag{13}$$

$$\frac{dx_\mu}{dt} = \frac{1 - x_\mu}{\tau_d} - u_\mu x_\mu R_\mu, \tag{14}$$

$$\tau \frac{dh_I}{dt} = -h_I + J_{IE} \sum_\beta R_\beta. \tag{15}$$

for $\mu, \beta = 1, ..., P$, with the connectivity matrix J in the form of

$$J_{\mu\beta} = \begin{cases} J_{EE}, & \mu = \beta; \\ \\ fJ_{EE}, & \mu \neq \beta. \end{cases}$$

We stimulate the network model of Eqs. 12 - 15 with variable τ_f, τ_d and τ, the results are shown in Fig. S4BD. These results confirmed our conjecture, that is, the overlaps in memory pattern representationss have similar effect to considering the excitatory connections between different memory clusters in the model of the Main Text.

4 Integrate and fire network

The spiking network was described in detail in our previous publication (Mongillo et al, 2008). Briefly, we considered a recurrent network of N_E excitatory and N_I inhibitory current-based integrate and fire neurons, driven by noisy background inputs with mean μ_b and variance σ_b. Each neuron, after emitting a spike, becomes refractory for time τ_{arp}. Recurrent connections exhibit transmission delay uniformly distributed between 0.1 and 1 ms. For simplicity, we neglect rise and decay times of the postsynaptic currents. Excitatory-to-excitatory synapses display short-term plasticity according to the scheme illustrated in Fig. 1B; remaining synaptic populations, inhibitory and excitatory-to-inhibitory, exhibit linear synaptic transmission.

4.1 Long-term synaptic structuring

There are P items to be memorized, each of them encoded by a subset of excitatory cells (selective population). Every selective population is formed by randomly selected fN_E neurons, where f is the coding level, enforcing the constraint that a given neuron belongs to at most one selective population (non-overlapping memories). Network connectivity is generated in the following way. Each cell receives $c(N_E + N_I)$ presynaptic connections, where c is the connectivity level, partitioned as follows: cfN_E randomly selected connections from each of the selective populations, $c(1 - fP)N_E$ randomly selected connections from the non-selective excitatory population, and cN_I randomly selected connections from the inhibitory population. The values of the efficacy for the various synaptic populations are reported in Table 1. Excitatory-to-excitatory synapses can take on two possible absolute efficacies: baseline, J_b, and potentiated, J_p. Synapses connecting two neurons within the same selective population have potentiated efficacy; Synapses connecting a selective neuron to a neuron from another selective population or to a non-selective neuron, have baseline efficacy; The remaining synapses (i.e. non-selective to selective and non-selective to non-selective) have potentiated efficacy with probability θ.

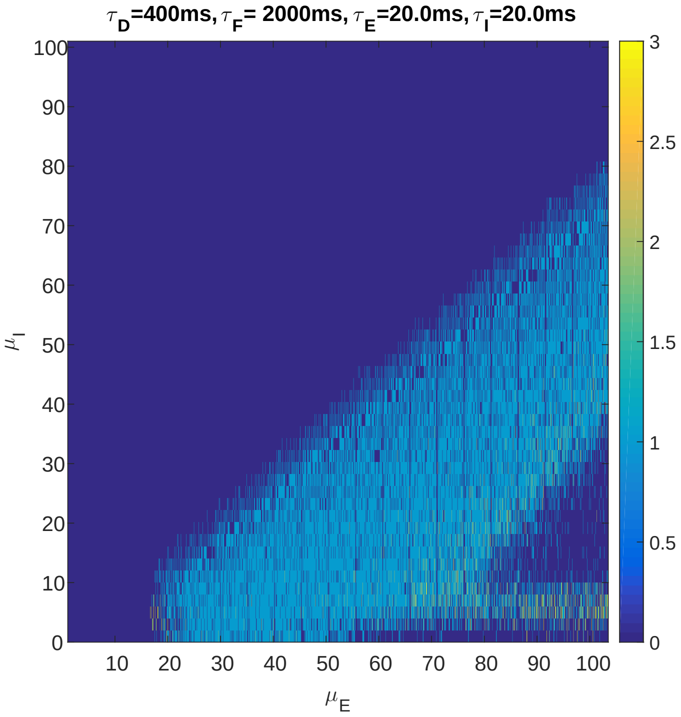

Figure S5: (Related to Figure 3) Bifurcation diagram. The number of simulatneously activated populations PS for different values of excitatory (μ_E) and inhibitory (μ_I) input currents.

4.2 Bifurcation diagram

It has been previously shown (Mongillo et al., 2008) that the network exhibits several regimes depending on the background excitatory current. For low background excitatory currents the network is in the low activity state, for high background excitatory current the network in high activity state. For intermediate background excitatory current there is a region where stable limit cycle solution exists. Moreover, in part of this region low activity state and limit cycle regime can coexist. Slowly increasing background excitatory current,

Single-cell parameters	**E**	**I**
Θ -Spike emission threshold	20 mV	20mV
V^r - Reset potential	16mV	13mV
τ - Membrane time constant	7-30 ms	7-30 ms
τ^{arp} - Absolute refractory period	2ms	2ms
Network parameters	**Values**	
f - Coding level	0.04	
p - Number of memories	10	
c - Probability of synaptic contact	0.2	
N - Number of excitatory/inhibitory cells	16000	4000
μ - Mean external current	variable	20.0mV
σ - Standard deviation of external current	1.0mV	1.0mV
Synaptic parameters	**Values**	
J_{IE} - Synaptic efficacy $E \to I$	0.135mV	
J_{EI} - Synaptic efficacy $I \to E$	0.25mV	
J_{II} - Synaptic efficacy $I \to I$	0.2mV	
J_b - Baseline level of $E \to E$ synapses	0.05mV	
J_p - Potentiated level of $E \to E$ synapses	0.45mV	
γ_0 - Fraction of potenciated synapses	0.1	
δ - Synaptic delays	0.1 - 1 ms	
Short-term synaptic dynamic parameters	**Values**	
U - Baseline utilization factor	0.2	
τ_F - Recovery time of utilization factor	1000-3000ms	
τ_D - Recovery time of synaptic resources	200-600ms	
Selective stimulation	**Values**	
T_{cue} - Stimulus duration	100ms	
A_{cue} - Contrast factor	1.15	
T_{isi} - Inter-stimulus interval	200ms	
Number of sequentially activated memories	8	

Table S1: (Related to Figure 3) Network parameters.

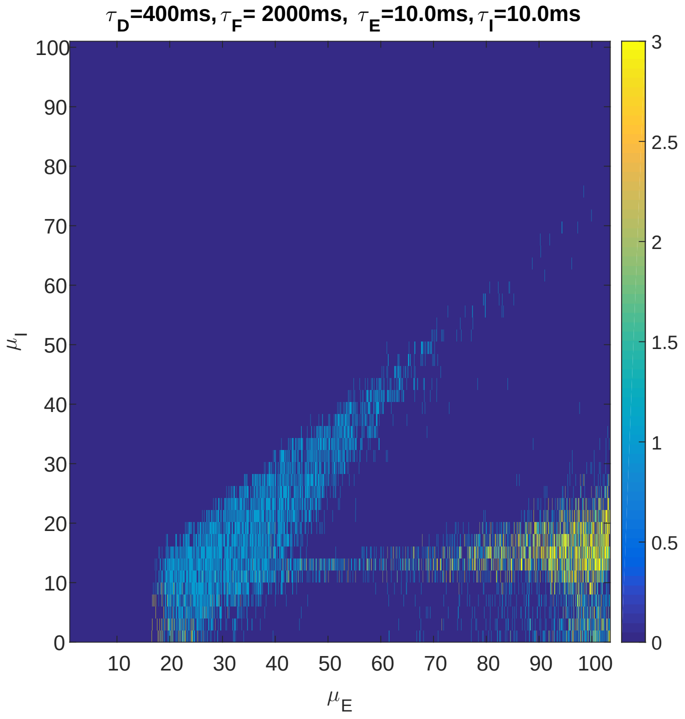

Figure S6: (Related to Figure 3) Another bifurcation diagram. Same as Fig. S5 but for different membrane time constants.

it is possible to find the value of the input where low activity state becomes unstable and the network enters into the limit cycle regime. Therefore, we computed a bifurcation diagram (Fig. S5), where for each level of inhibitory background current we have gradually increased excitatory background current and recorded for each instance of time how many selective populations, corresponding to different encoded items, were generating PS at any given moment. It can be seen that for each inhibitory input there is a range of excitatory background currents where the network is in the limit cycle regime, except for a range of low inhibitory background currents where the network exhibits PSs of more than one population. We carried out numerical bifurcation analysis for several values of time constants and found that for time constants for which higher WM capacity is expected according to theoretical predictions, the range of excitatory background currents where the network exhibits limit cycle behavior shrinks. For illustration we present another bifurcation diagram (Fig. S6) where synaptic time constants are smaller. For this set of parameters the network does not exhibit a limit cycle behavior for large values of inhibitory background current. Therefore, to compute WM capacity we performed simulations for inhibitory background current where the network exhibits limit cycle behavior for large range of time constants ($\mu_I = 20$mV).

4.3 WM capacity estimation

The bifurcation analysis presented above shows the approximate values of excitatory background current where low activity state of the network becomes unstable. It does not show the level of excitatory background current where stable limit cycle solution appears. Moreover, the integrate and fire network is by construction stochastic and a slightly different results is expect for each simulation even for the same set of parameters and initial conditions. Therefore, we repeated each simulation with the same parameters and the same initial conditions 10 times. Each simulated trial consisted of initial period with fixed input currents for 5 seconds, followed by the train of excitatory pulses sequentially presented to 8 out of 10 stored populations, followed by 6 seconds of retention period where the input currents were constant and equal to the those before activation. The network was deemed to be in the spontaneous limit cycle behavior when it generated at least one PS during the period before activation or it generated PS during retention for the selective population that was not activated. If the network did not show the spontaneous limit cycle behavior, the number of retained memories for a given simulation was computed as the number of selective populations exhibiting at least one PS in the last 5 seconds of retention period, i.e. we did not count PS during the first 1 second after the end of activation sequence to remove the effects of transient activity. The WM capacity was computed as the average number of retained memories for 10 repetition if none of them was producing spontaneous limit cycle activity. We performed simulations using 5 equidistantly placed levels of excitatory current for each set of time constants, adjusting the range by the following criteria: (1) only the largest excitatory background current produced spontaneous limit cycle activity; (2) the smallest excitatory background current produced WM capacity greater than 0 (in at least one of the repetition at least one memory was retained); (3) the ratio of highest to lowest WM capacities for 4 lower excitatory levels is between 2 and 3, or the range of excitatory currents used for simulation is less than $10^{-2}mV$. The WM capacity for a given set of time constants was then defined as the largest WM capacity among lower 4 input excitatory levels in the range.

4.4 Limits on WM capacity in spiking networks

To confirm that there are no theoretical limits on WM capacity arising from stability issues, we simulated an unrealistic network with $\tau_f = 30$s (all other parameters as in Figure 3 of the main paper) and showed that the network maintains all 8 loaded items in working memory(see Figure S7).

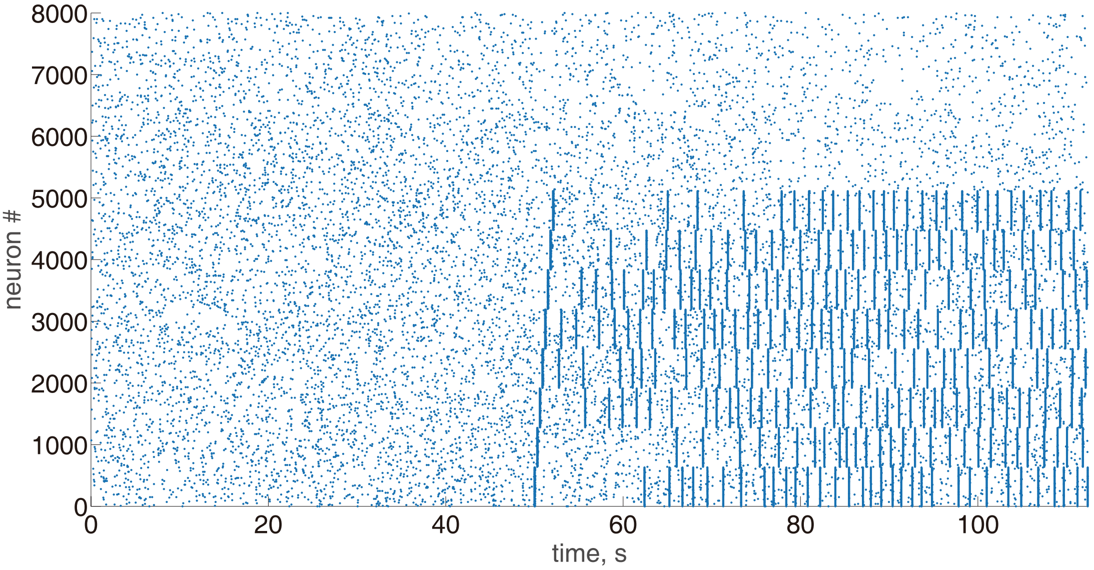

Figure S7: (Related to Figure 3) Raster plot of simulated network activity for large $\tau_f = 30$s, all other parameters as in Figure 3 of the main paper.