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 $t_{\rm dur} = 0.01s$. The presentation frequency is determined by the time interval between any two consecutive stimuli $t_{\rm int}$.

In Figure 1D, 450 different presentation frequencies were chosen, where $t_{\rm int}$ 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 $t_{\rm int}$ 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.

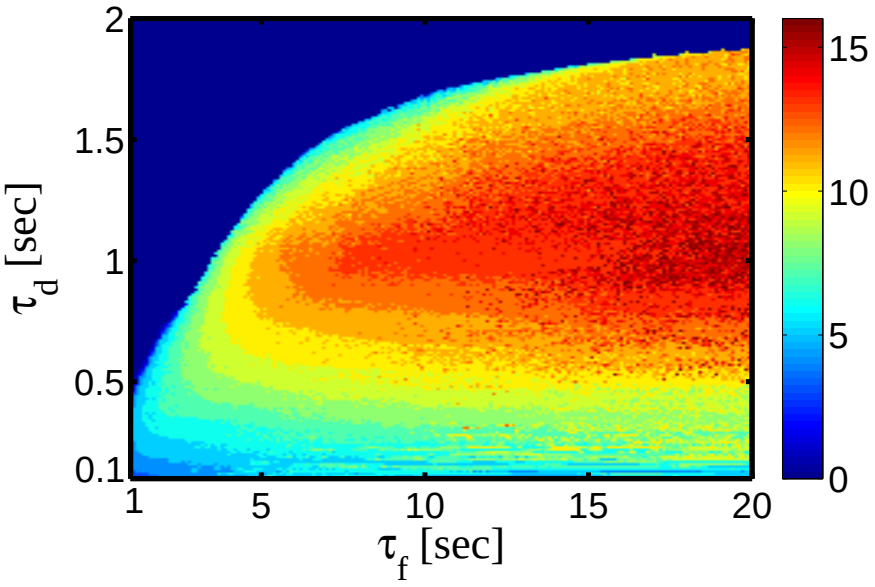

Figure S1: (Related to Figure 2) Working memory capacity dependence on neuronal and synaptic time constants τ_f and τ_d with numerical simulations of the model (Eqs. 1 - 4) of Main Text. In some regime, the network could not hold any memory items.

3 Theoretical Estimation of the WM Capacity

3.1 Maximum Time between Population Spikes of a Given Cluster ($T_{\max}$)(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:

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

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

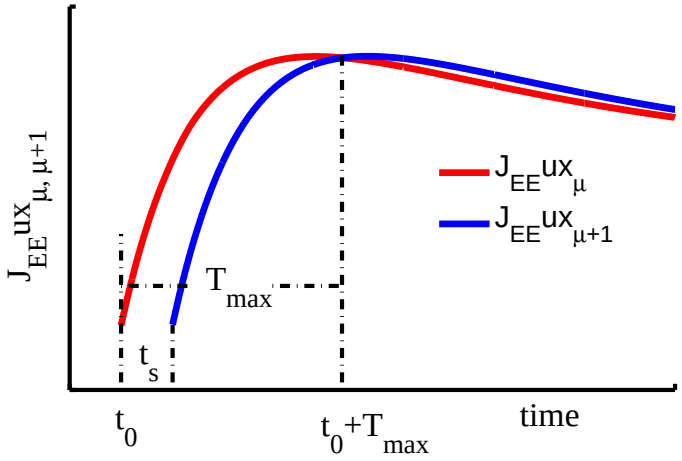

Figure S2: (Related to Figure 1) The dynamics of synaptic efficacy of two excitatory clusters which generate PSs consecutively with a time interval t_s. $t_0 + T_{\max}$ is the moment when the adjacent synaptic efficacy intersect with each other. The necessary condition for keeping the right retrieving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

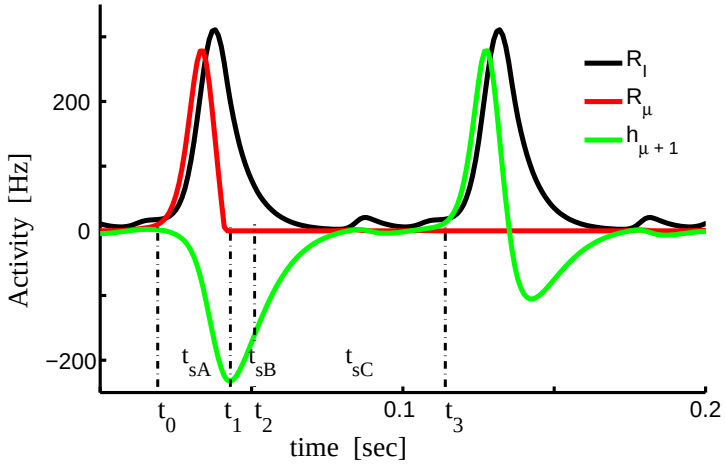

Figure S3: (Related to Figure 2) The inter-PS interval t_s can be divided into three different parts. Interval A: the cluster μ generates a PS; Interval B: the PS of the inhibitory cluster terminates; Interval C: the synaptic input to the cluster $\mu + 1$ keeps increasing to a threshold and starts to 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 $I_{\text{crit}} = I_{\text{inh}} - \alpha \ln(J_{\max} - 1)$.

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

$$\tau \frac{dh_i}{dt} = -h_i + \sum_{j=1}^{N} J_{ij} u_j x_j R_j - J_{EI} R_I + I_b + I_e(t), \tag{4}$$

$$\frac{du_i}{dt} = \frac{U - u_i}{\tau_f} + U(1 - u_i) R_i, \tag{5}$$

$$\frac{dx_i}{dt} = \frac{1 - x_i}{\tau_d} - u_i x_i R_i, \tag{6}$$

$$\tau \frac{dh_I}{dt} = -h_I + J_{IE} \sum_{j=1}^{N} R_j. \tag{7}$$

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:

$$\tau \frac{dh_\eta}{dt} = -h_\eta + \sum_\beta J_{\eta\beta} P S_\beta u_\beta x_\beta R_\beta - J_{EI} R_I + I_b + I_e(t), \quad (8)$$

$$\frac{du_\eta}{dt} = \frac{U - u_\eta}{\tau_f} + U(1 - u_\eta) R_\eta, \quad (9)$$

$$\frac{dx_\eta}{dt} = \frac{1 - x_\eta}{\tau_d} - u_\eta x_\nu R_\eta, \quad (10)$$

$$\tau \frac{dh_I}{dt} = -h_I + J_{IE} \sum_\beta R_\beta. \quad (11)$$

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, $f^2 N$ 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

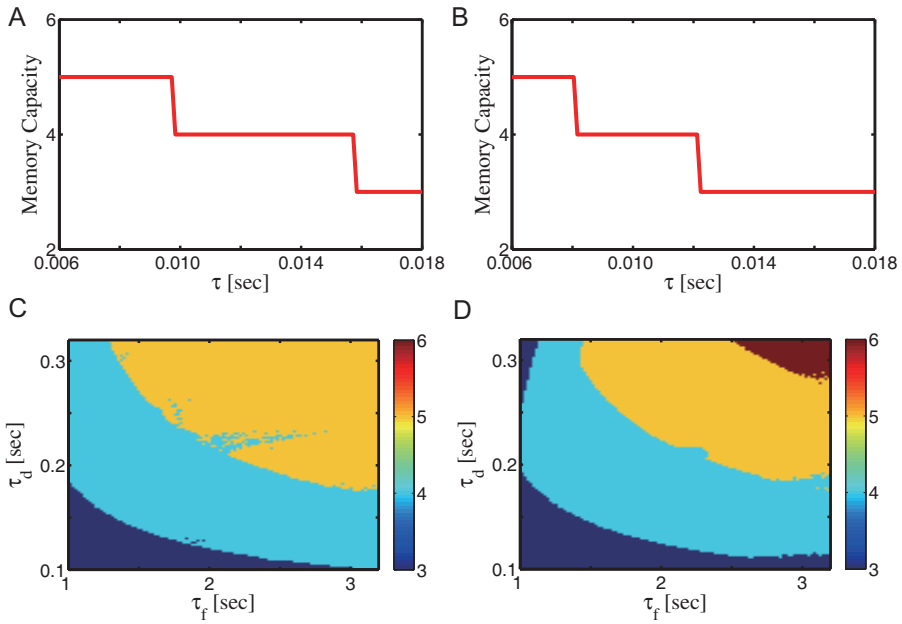

Figure S4: (Related to Figure 2) Working memory capacity dependence on neuronal and synaptic time constants with overlapping memory patterns. (A) The WM capacity as a function of τ, obtained with numerical simulations of the network model (Eqs. 8 - 11). The parameters are : $\tau_d = 0.3s$, $\tau_f =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

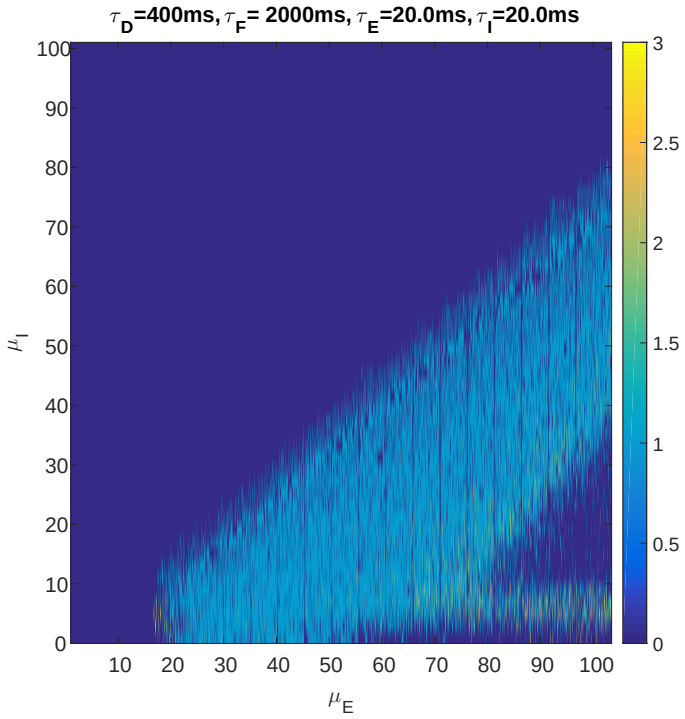

Figure S5: (Related to Figure 3) Bifurcation diagram. The number of simulatneously activated populations PS for different values of excitatory (μ_E) and inhibitory (μ_I) input currents.

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.

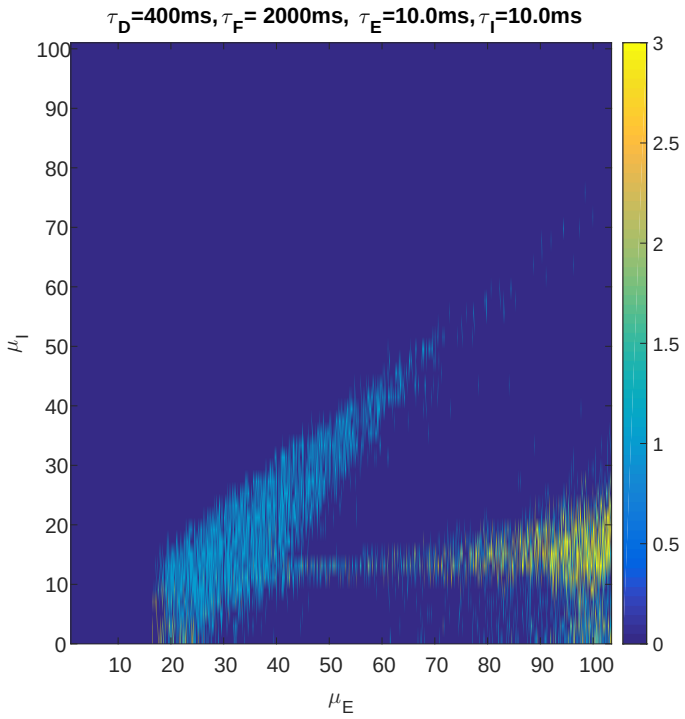

Figure S6: (Related to Figure 3) Another bifurcation diagram. Same as Fig. S5 but for different membrane time constants.

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 ($\mu_I = 20\text{mV}$).

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).

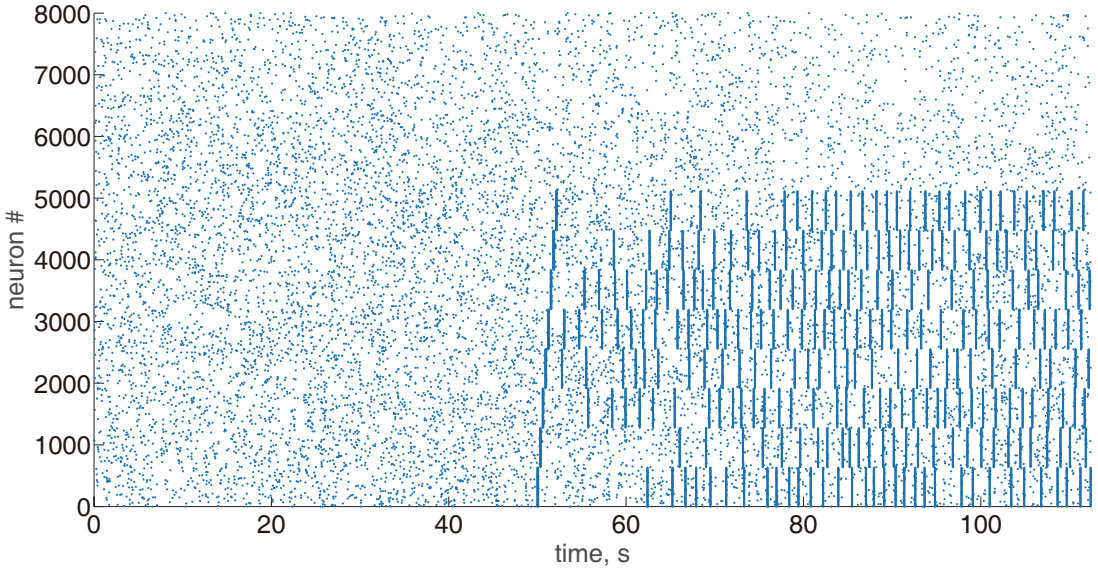

Figure S7: (Related to Figure 3) Raster plot of simulated network activity for large $\tau_f = 30$s, all other parameters as in Figure 3 of the main paper.