Supporting Derivations

Mean-field Equations

Below are the mean-field equations for an object WM network as in ref. (29), slightly modified. The equations describe how the rate of the cells storing memories in the network depends on the connectivity, external input and vsWM load. The modifications from ref. (29) were

- Each pyramidal cell in the network coded for a stimulus. There was no population of non-coding pyramidal cells
- Expressions for the number of encoded items were included.
- All cells were connected to all other cells.
- The proportion of cells in a single active population, *w*, was not negligible.

$$\mu_+ = J_{\mathrm{E}\to\mathrm{E}} N_\mathrm{E} \tau_\mathrm{E} \left[w g_+ r_+ + (p-1) w g_- r_+ + (1-pw) g_- r_0 \right] + J_{\mathrm{X}\to\mathrm{E}} \tau_\mathrm{E} r_{xe} - J_{\mathrm{I}\to\mathrm{E}} N_\mathrm{I} \tau_\mathrm{E} r_\mathrm{I}$$

$$\sigma_+^2 = J_{E\to E}^2 N_\mathrm{E} \tau_\mathrm{E} \left[w g_+^2 r_+ + (p-1) w g_-^2 r_+ + (1-pw) g_-^2 r_0 \right] + J_{X\to E}^2 \tau_\mathrm{E} r_{xe} + J_{I\to E}^2 N_\mathrm{I} \tau_\mathrm{E} r_\mathrm{I}$$

$$\mu_0 = J_{\mathrm{E}\to\mathrm{E}} N_\mathrm{E} \tau_\mathrm{E} \left[pw g_- r_+ + w g_+ r_0 + (1-(p+1)w) g_- r_0 \right] + J_{\mathrm{X}\to\mathrm{E}} \tau_\mathrm{E} r_{xe} - J_{\mathrm{I}\to\mathrm{E}} N_\mathrm{I} \tau_\mathrm{E} r_\mathrm{I}$$

$$\sigma_0^2 = J_{E\to E}^2 N_\mathrm{E} \tau_\mathrm{E} \left[p w g_-^2 r_+ + w g_+^2 r_0 + (1-(p+1)w) g_-^2 r_0 \right] + J_{X\to E}^2 \tau_\mathrm{E} r_{xe} + J_{I\to E}^2 N_\mathrm{I} \tau_\mathrm{E} r_\mathrm{I}$$

$$\mu_\mathrm{I} = J_{\mathrm{E}\to\mathrm{I}} N_\mathrm{E} \tau_\mathrm{I} \left[pw r_+ + (1-pw) r_0 \right] + J_{\mathrm{X}\to\mathrm{I}} \tau_\mathrm{I} r_{xi} - J_{\mathrm{I}\to\mathrm{I}} N_\mathrm{I} \tau_\mathrm{I} r_\mathrm{I}$$

$$\sigma_I^2 = J_{E\to I}^2 N_\mathrm{E} \tau_\mathrm{I} \left[pw r_+ + (1-pw) r_0 \right] + J_{X\to I}^2 \tau_\mathrm{I} r_{xi} + J_{I\to I}^2 N_\mathrm{I} \tau_\mathrm{I} r_\mathrm{I}$$

$\mu_{+,0,\mathrm{I}}$: Mean input potential to
+: excitatory pyramidal cells (E cells) holding memories
0: E cells not holding memories
I: inhibitory interneurons (I cells)

$\sigma^2_{+,0,\mathrm{I}}$: Standard deviation of input potential to the above populations.

$J_{a\to b}$: Mean connection strength from population *a* to *b* (*a*,*b*=E,I,X), relating presynaptic firing to postsynaptic potential.

$N_{\mathrm{E,I}}$: Number of cells.

$\tau_{\mathrm{E,I}}$: Membrane time constant.

w: Size of stimulus-encoding populations relative to total network size.

p: Number of memories encoded (memory load).

$g_{+,-}$: Relative connection strength (relative to mean connection strength, $J_{E\to E}$)

+: between E cells coding for similar stimuli.

-: between E cells coding for dissimilar stimuli. g_- is scaled so that $g_-=(1-g_+w) / (1 - w)$.

$r_{+,0,I,xe,xi}$: Firing rates of the above populations as well as

xe: external input into E cell population

xi: external input into I cell population

The firing rates of integrate-and-fire neurons embedded in a noisy environment are given by the frequency-postsynaptic potential (*f-V*) equation (43):

$$\varphi(\mu,\sigma) = \left(\tau_{\mathrm{E,I}}^{ref} + \tau_{E,I}\int \sqrt{\pi} e^{u^2}\left[1+\mathrm{erf}(u)\right]du\right)^{-1}$$

$$\widetilde{\theta} = \frac{\theta-\mu}{\sigma}\left(1+0.5\frac{\tau_{\mathrm{E,I}}^{'}}{\tau_{\mathrm{E,I}}}\right)+1.03\sqrt{\frac{\tau_{\mathrm{E,I}}^{'}}{\tau_{\mathrm{E,I}}}}-0.5\frac{\tau_{\mathrm{E,I}}^{'}}{\tau_{\mathrm{E,I}}} \qquad \widetilde{V}_{res} = \frac{V_{res}-\mu}{\sigma}$$

φ: Firing rate

$\tau_{\mathrm{E,I}}^{ref}$: Refractory period

$\tau_{\mathrm{E,I}}^{'}$: Synaptic time constant

θ: Firing threshold

V_{res}: Reset potential

Simplification of Equations

To obtain an analytical solution to the WM network equations, a number of simplifications were needed. First, we observed that the mean and standard deviation of the external input can be varied independently by varying r_{xa} and $J_{\mathrm{X}\to a}$, a=I,E. Thus, we let $\mu_{xa}=r_{xa}\,J_{\mathrm{X}\to a}\,\tau_a$ and $\sigma_{xa}^2 = r_{xa}\,J_{X\to a}^2\,\tau_a$.

Next, we made the assumption that the input-output function of the inhibitory population is threshold linear, i.e. $r_{\mathrm{I}}=k[\mu_I - V_\theta]_+$, where V_θ is the firing threshold potential, and $[\cdot]_+$ is the thresholding operator, such that $[x]_+=x$ if $x>0$ and $[x]_+=0$ otherwise. This assumption basically says that the firing rate of the inhibitory cells is in the linear range. Since I cell rates were at least 10 Hz in simulations, and since inhibitory rates were always between 10 and 30 Hz in the ensuing theoretical analysis,

this is a reasonable approximation. We also noted that since I cells were always above the threshold, the thresholding operator could be removed. Solving for r_I, we obtained

$$\begin{aligned}
\mu_I &= J_{E\to I} N_E \tau_I \left[pwr_+ + (1-pw)r_0\right] + \mu_{xi} - J_{I\to I} N_I \tau_I r_I \\
r_I / k + V_\theta &= J_{E\to I} N_E \tau_I \left[pwr_+ + (1-pw)r_0\right] + \mu_{xi} - J_{I\to I} N_I \tau_I r_I \\
(1/k + J_{I\to I} N_I \tau_I)\, r_I &= J_{E\to I} N_E \tau_I \left[pwr_+ + (1-pw)r_0\right] + \mu_{xi} - V_\theta \\
r_I &= \frac{J_{E\to I} N_E \tau_I \left[pwr_+ + (1-pw)r_0\right] + \mu_{xi} - V_\theta}{1/k + J_{I\to I} N_I \tau_I}
\end{aligned} \qquad [S1]$$

Next, we assumed that the standard deviation of the input current into the excitatory population was independent of its firing rate (29), i.e. $\sigma \equiv \sigma_+ = \sigma_0$. Although this is questionable, the approximation is a good one when recurrence is dominated by NMDA and it simplified the analysis considerably.

After these simplifications, the equations for the standard deviations of the currents could be removed and the inhibitory firing rate could be replaced by equation [S1]. In order to express inputs as currents instead of potentials and to drop the standard deviation, we introduced $f(I) = \varphi(I/g_L, \sigma)$, where g_L = 25 nS was the pyramidal cell leak conductance and I was the mean input current. The network description could then be reduced to the following equations.

$$\begin{aligned}
r_+ &= f(g_L J_{E\to E} N_{E\to E} \tau_E (wg_+ r_+ + (p-1)wg_- r_+ + (1-pw)g_- r_0) + g_L \mu_{xe} - \ldots \\
&\quad g_L J_{I\to E} N_{I\to E} \tau_E \frac{J_{E\to I} N_{E\to I} \tau_I (pwr_+ + (1-pw)r_0) + \mu_{xi} - V_\theta}{1/k + J_{I\to I} N_{I\to I} \tau_I}) \\
r_0 &= f(g_L J_{E\to E} N_{E\to E} \tau_E (pwg_- r_+ + wg_+ r_0 + (1-(p+1)w)g_- r_0) + g_L \mu_{xe} - \ldots \\
&\quad g_L J_{I\to E} N_{I\to E} \tau_E \frac{J_{E\to I} N_{E\to I} \tau_I (pwr_+ + (1-pw)r_0) + \mu_{xi} - V_\theta}{1/k + J_{I\to I} N_{I\to I} \tau_I})
\end{aligned}$$

With w=0.1 and $r_+ \approx 50$ s^{-1}, the non-coding population made up at most 10% of all pyramidal cell activity, which is negligible. Therefore, we set r_0=0 and dropped the equation for r_0. This simplification is useful, as it allows the graphical analysis of the WM network that is reported in the paper.

We finally defined a number of auxiliary terms in order to make the resulting equations more compact. First, we normalized connection strengths with respect to the network, i.e. $G_{a\to b} = g_L J_{a\to b} N_a \tau_b$. Next, we introduced $h = k / g_L$, which is the gain of

the interneuronal input-output curve expressed as spikes per delivered current. Finally, we introduced

$$I_X \equiv g_L \mu_{xe} - \frac{G_{I\to E}}{1/h + G_{I\to I}} g_L (\mu_{xi} - V_\theta)$$

After these simplifications, the equations were reduced to one. So we collected the terms around r_+, applied the firing rate function, and dropped the now unnecessary subscript +

$$r = f\left(\left(G_{E\to E}(wg_+ + (p-1)wg_-) - \frac{G_{I\to E}G_{E\to I}pw}{1/h + G_{I\to I}}\right)r + I_X\right) \qquad [S2]$$

I_X should be set so that the neuron is just at the firing threshold.

Equation [S2] forms the basis of our analysis of the mechanisms governing working memory capacity. By introducing the terms

$$\begin{aligned} G^+ &= w(G_{E\to E}g_+ - G_{I\to E}G_{E\to I}/(1/h + G_{I\to I})) \\ G^- &= -w(G_{E\to E}g_- - G_{I\to E}G_{E\to I}/(1/h + G_{I\to I})) \end{aligned} \qquad [S3]$$

equation [S2] could also be rewritten into equation [1] in the Results section:

$$r = f((G^+ - G^-(p-1))\, r + I_X)$$

Explicit Capacity Formula

The above formula is an implicit equation for capacity. To obtain an explicit capacity formula, we used the fact that capacity is the value of p for which the straight line in Fig. 2A (the inverse of $I(r)$) is tangent to the input-output curve $f(I)$. At this point of intersection, $(I_{cap}, f(I_{cap}))$, we have the following equation for the straight line

$$r = \frac{f(I_{cap})}{I_{cap} - I_X}(I - I_X).$$

The condition that this line is tangent to $f(I)$ at I_{cap} is written as

$$\frac{f\left(I_{\text{cap}}\right)}{I_{\text{cap}}-I_X}=f'\left(I_{\text{cap}}\right) \qquad \text{[S4]}$$

From this equation one can derive (at least numerically, and maybe only for a range of parameter values) I_{cap} as a function of I_{X}. Then, we equated the slope of the straight line, $1/[G^{+}–G^{-}(p_{\text{cap}}–1)]$ (compare to $\Delta I/\Delta r$ in Fig. 2A), to the left hand side of [S4]

$$\frac{1}{G^{+}-G^{-}\left(p_{\text{cap}}-1\right)}=\frac{f\left(I_{\text{cap}}\left(I_X\right)\right)}{I_{\text{cap}}\left(I_X\right)-I_X}\equiv\frac{1}{H\left(I_X\right)}$$

Here, $H(I_X)$ represents the network effective connection strength at p_{cap}. We next isolated p_{cap}

$$p_{\text{cap}}=1+\frac{1}{G^{-}}\left(G^{+}-H\left(I_X\right)\right).$$

This is equal to [2] in the main text.

For some common models of the neuronal input-output function, e.g. the sigmoidal function $f(I) = (1+e^{-I})^{-1}$, $H(I_{\text{X}})$ can be written in closed form. To see this, enter $(1+e^{-I})^{-1}$ into equation [S4]. Further reorganization of the terms lead to the equation $I_{\text{cap}} – I_{\text{X}} – 1 – \exp(I_{\text{cap}}) = 0$. This equation has the solution

$$I_{\text{cap}}=1+I_X-W\left(-e^{1+I_X}\right),$$

with $W(x)$ being Lambert's W function. Therefore, for this case

$$H\left(I_X\right)=-\frac{\left[1-W\left(-e^{1+I_X}\right)\right]^2}{W\left(-e^{1+I_X}\right)}.$$

This function is graphically plotted in SI Fig. 1, showing its monotonously decreasing dependence on I_{X}.

Capacity is limited to a few items

This section derives the formula [3] in the main text for the upper bound of capacity, p_{cap}^{UL}. The formula for capacity p_{cap} ([2] in main text) comes with a series of constraints on the parameters that limit the possible values of p_{cap}. These constraints are:

1. effective positive feedback within a local active population has to be strong enough so that the network can sustain persistent activity for a single item. This is mathematically expressed as $G^{+} > H(I_X) > 0$.
2. in order to maintain selectivity during multiple-item persistent activity, effective interactions between different active populations have to be inhibitory. This is mathematically expressed as $G^{-} > 0$.

Using the expressions [S3] for G^{+} and G^{-} as a function of the physiological parameters, we obtain:

$$G_{E\to E}g_{+} > G_{I\to E}G_{E\to I}\left(\frac{1}{h}+G_{I\to I}\right)^{-1}+\frac{H(I_X)}{w}$$
$$G_{E\to E}g_{-} < G_{I\to E}G_{E\to I}\left(\frac{1}{h}+G_{I\to I}\right)^{-1},$$

from which we conclude that there is a real number γ, constrained by $g_{-} < \gamma < g_{+} - \frac{H(I_X)}{wG_{E\to E}} < g_{+}$, such that $\gamma G_{E\to E} = G_{I\to E}G_{E\to I}\left(\frac{1}{h}+G_{I\to I}\right)^{-1}$. Since g_{+} cannot be too large without making g_{-} negative, the non-dimensional parameter γ lies close to unity. The equation shows that if $\gamma \approx 1$, excitatory and inhibitory feedback are approximately balanced during the spontaneous state of network activity, when none of the selective populations are active. The case $\gamma > 1$ corresponds to the situation where inhibitory feedback exceeds excitatory feedback during baseline activity. Such a regime has been shown to be a requirement of baseline activity stability in attractor networks of working memory (29). Using this new parameter γ reflecting network balance, we can now write eq. [S3] as:

$$G^{+} = wG_{E\to E}\left(g_{+} - \gamma\right)$$
$$G^{-} = wG_{E\to E}\left(\gamma - g_{-}\right) = G^{+}\left(\frac{\gamma - g_{-}}{g_{+} - \gamma}\right). \qquad \text{[S5]}$$

This shows that effective interactions within and between active populations (G^{+} and G^{-}, respectively) are not independent from each other, but are linked through network balance and connectivity tuning. Therefore, the equation for the network capacity becomes:

$$p_{cap} = 1 + \frac{g_{+} - \gamma}{\gamma - g_{-}}\left(1 - \frac{H(I_X)}{G^{+}}\right).$$

Mathematically, p_{cap} is in general not limited. For instance, eq. [S5] shows that when $\gamma = g_{-}$, $G^{-} = 0$ and p_{cap} diverges. Obviously, a lack of lateral inhibition, i.e. a complete disconnection between subpopulations, leads to unlimited capacity. However, this is not biologically realistic. In particular, it has been shown that the stability of network activity in the absence of memorized items (spontaneous activity) can only be achieved when inhibitory feedback is at least as strong as excitatory feedback (29): $\gamma > 1 > g_{-}$. In the balanced case of $\gamma = 1$, an upper limit for p_{cap} (denoted p_{cap}^{UL}), can be determined. In this case, the expression for p_{cap} takes a simple form, considering the relation between the parameters g_{+} and g_{-} ($g_{-} = (1 - g_{+}w) / (1 - w)$):

$$p_{cap} = \frac{1}{w}\left(1 - (1-w)\frac{H(I_X)}{G^{+}}\right) < p_{\max}\left(1 - (1-w)\frac{H(I_X)}{G^{+}_{\max}}\right) = p_{cap}^{UL} \qquad \text{[S6]}$$

,

with $1/H(I_X)$ being bounded above by the maximal slope of the neuronal *f-I* curve $f(I)$ (see lower bound for $H(I_{\mathrm{X}})$ in Fig. S1). $G^{+}_{\max}$ is the maximum value of G^{+} for which the spontaneous activity state is still stable. The equation shows that, assuming that the network operates in a balanced regime in the spontaneous state, inhibition-limited capacity is determined by two factors: $p_{\max}=1/w$ and $G^{+}_{\max}/H(I_X)$, the ratio of the maximum effective feedback excitation from the local population that still preserves

the quiescent memory state and the minimal effective feedback that can sustain persistent activity.

To find an approximate value for p_{cap}^{UL}, we must determine these two quantities. Unfortunately, the magnitude of *w* is not known, but recent estimates from delayed match-to-sample tasks in monkeys lie between 0.05 and 0.25 in the inferior temporal and prefrontal cortex (33). $G_{\max}^{+}/H(I_X)$, on the other hand, must be low enough so that the spontaneous activity remains stable. Since the spontaneous activity was assumed to be zero in the derivation of the capacity formula, $G_{\max}^{+}$ cannot be obtained from the formulas derived above. However, based on the analysis by Brunel (28), $1/G_{\max}^{+}$ is the effective connection strength (inverse of the slope of the line) making the lower and middle fixed points in Fig. 2A and SI Fig. 3A fuse ($1/H(I_X)$ is the strength making the upper and middle fixed points fuse). Thus, the right and left borders of the golden area in SI Fig. 3A are the positions of the connectivity line corresponding to $G_{\max}^{+}$ and $H(I_X)$, respectively. In a network with maximal capacity, when only one item is stored, the effective connection strength is $G_{\max}^{+}$. At capacity, it is $H(I_X)$. As seen in eq. [6], capacity is related to the ratio $G_{\max}^{+}/H(I_X)$. Both $G_{\max}^{+}$ and I_X depend on $r_{sp,E}$, the spontaneous rate of excitatory neurons. SI Figure 3 shows how $G_{\max}^{+}/H(I_X)$ and p_{cap}^{UL} vary with $r_{sp,E}$ when taking the *f-I* curve of a leaky integrate-and-fire neuron. For this neuronal model, with physiologically realistic values of $r_{sp,E} > 1$ Hz, $G_{\max}^{+}/H(I_X) < 1.5$ and p_{cap}^{UL} comes close to experimentally observed value. For example, taking $G_{\max}^{+}/H(I_X) = 1.5$ in the above example we find $p_{cap}^{UL} = 4$. Naturally, in addition to $r_{sp,E}$, $G_{\max}^{+}$ varies with the shape of the *f-I* curve. The more linear the *f-I* curve is, the smaller is the ratio between $G_{\max}^{+}$ and $H(I_X)$. For integrate-and-fire neurons, the *f-I* curve is generally very linear for frequencies between a few Hz and up to 50 Hz (SI Fig. 3A), and it becomes more linear if neuronal input is noisy, as is the case in the brain. Experimentally measured *f-I* curves for neurons subject to noisy inputs also have a mostly linear rise at moderate rates (45,46), indicating that $1 < G_{\max}^{+}/H(I_X) < 1.5$ is indeed reasonable in real neurons. In

addition, physiological mechanisms present in cortical neurons such as spike-frequency adaptation linearize non-linear input-output functions (47).

Notice that our finding that the stability of the spontaneous and the persistent state imposes a limit on the capacity of the network through the inequality $1 < G^{+}_{\max} / H(I_X) < 1.5$, setting it to about 40% of its maximal capacity, was obtained by positioning ourselves in the limit cases: perfect input balance (γ=1), and stability of fixed points, assuming a fluctuation-free environment. In the general cases of γ>1 and fluctuation-rich background, the constraints on $G^{+}_{\max} / H(I_X)$ will be more severe and will result in a sharper reduction of capacity. Thus, although our analysis does not unequivocally limit the theoretical upper limit of WM capacity to what is found experimentally, it shows that such a low limit exists and is due to the physiological properties of cortical neurons (relatively linear *f-I* curves) and cortical networks (excitation-inhibition balance).

Supporting References

45. Arsiero M, Luscher H, Lundstrom BN, Giugliano M (2007) The Impact of input fluctuations on the frequency-current relationships of layer 5 pyramidal neurons in the rat medial prefrontal cortex. *J Neurosci* 27:3274-3284.

46. Rauch A, La Camera G, Luscher H, Senn W, Fusi S (2003) Neocortical pyramidal cells respond as integrate-and-fire neurons to in vivo-like input currents. *J Neurophysiol* 90:1598-612.

47. Ermentrout B (1998) Linearization of F-I curves by adaptation. *Neural Comput* 10:1721-9.