Supplementary Information

Contents

www.pnas.org/cgi/doi/10.1073/pnas.1918674117

1 Rate network

1.1 The network model

We consider a network of N randomly connected rate units, each of which make an average of K connections with other units. Unless otherwise noted, for all rate network simulations, $N = 40{,}000$ and $K = 200$. In analytical calculations, we take $N \to \infty$ and $K \to \infty$ while $K/N \to 0$. In all but Figure 7, units are governed by the following dynamics

$$\tau \frac{dr_i}{dt} = -r_i + \phi\left(h_i\right), \quad h_i = \sum_{j \neq i} J_{ij} r_j \tag{1}$$

where τ is the time constant of the rate dynamics, and h_i is the total synaptic input to neuron i. Synaptic input is transformed through a sigmoidal neural transfer function

$$\phi(x) = \frac{r_{\max}}{2}\left(1 + \operatorname{erf}\left(\frac{x - \theta}{\sigma\sqrt{2}}\right)\right) \tag{2}$$

where $r_{\max}$ is the maximal firing rate, θ is the input at which the firing rate is $r_{\max}/2$, and σ controls the gain of the transfer function. In particular, for $\sigma \to 0$ the transfer function becomes the Heaviside function, $\phi(x) = 1$ for $x > \theta$, and zero otherwise.

1.2 Learning rule

The temporally asymmetric Hebbian learning rule produces the following connectivity matrix J_{ij}:

$$J_{ij} = A \frac{c_{ij}}{K} \sum_{s=1}^{S} \sum_{\mu=1}^{P-1} f(\xi_i^{s,\mu+1}) g(\xi_j^{s,\mu}) \tag{3}$$

where A is a learning rate, c_{ij}s are i.i.d. Bernoulli random variables ($c_{ij} = 1$ with probability c and 0 otherwise) and $K = Nc$ represents the average in-degree of a neuron. This connectivity matrix stores S sequences of P patterns $\{\xi_i^{s,\mu}\}$, where $\xi_i^{s,\mu}$ can be thought as the input to neuron i ($i = 1, \ldots, N$) in pattern μ ($\mu = 1, \ldots, P$) of sequence s ($s = 1, \ldots, S$), using a temporally asymmetric Hebbian rule. With such a rule, synapses are modified for each pair of successive patterns in the sequence by an amount $f(\xi_i^{s,\mu+1})g(\xi_j^{s,\mu})$, where the function f describes the dependence of the learning rule on input to the post-synaptic neuron, and g describes the dependence on input to the pre-synaptic neuron. The patterns are identically and independently distributed (i.i.d.) as $\xi_i^{s,\mu} \overset{iid}{\sim} \mathcal{N}(0,1)$.

In the first part of the paper (Figures 1-3 and Supplementary Figures 1-6,10-11) we use a bilinear rule, $f(x) = g(x) = x$. In Figure 4-7 and Supplementary Figures 7-9,12 we use a non-linear learning rule, with

$$\mathrm{f}(x) = \begin{cases} q_f & \text{if } x > x_f \\ q_f - 1 & \text{if } x \leq x_f \end{cases} \tag{4}$$

$$\mathrm{g}(x) = \begin{cases} q_g & \text{if } x > x_g \\ q_g - 1 & \text{if } x \leq x_g \end{cases} \tag{5}$$

The variable x_f (x_g) defines the threshold separating potentiation and depression when post (pre) synaptic firing rates are varied, while q_f (q_g) controls the strength of plasticity at high post (pre) firing rate. In order to keep the average sum over incoming connection strengths zero, we set $q_g = F_z(x_f)$, where F_z is the cumulative distribution function of a standard Gaussian [1].

1.3 Sequences in continuous time

The network defined above can also learn and retrieve sequences in continuous time. We simulated a network storing sequences $\eta_i^s(t)$ defined as realizations of Ornstein-Uhlenbeck (OU) processes, with zero mean, standard deviation σ_{OU} and correlation time constant τ_{OU}, using the connectivity matrix

$$J_{ij} = A\frac{c_{ij}}{K}\sum_{s=1}^{S}\int_{t=0}^{T-\Delta_t} f(\eta_i^s(t+\Delta_t))g(\eta_j^s(t)) \tag{6}$$

Note that Eq. 6 will produce connectivity with the same statistics as in Eq. 3 when $\tau_{\mathrm{OU}} \ll \Delta_t$, $\sigma_{\mathrm{OU}} = 1$, $T/\Delta_t = P$ and the OU process is discretized with a timestep of Δ_t.

In Figure 1 of the paper (left column), we use the continuous learning rule of Eq. 6 to store and retrieve OU processes generated using a time constant of $\tau_{\mathrm{OU}} = 4$ms and standard deviation $\sigma_{\mathrm{OU}} = 1$ with a discretization of 1ms. We use a temporal offset $\Delta_t = 10$ms and neuronal time constant $\tau = 10$ms.

2 Mean-field theory of sequential activity

2.1 Overlap with network activity

In this section we derive a mean field theory for a network where stored patterns are Gaussian and the learning rule is bilinear, i.e. $f(x) = x$ and $g(y) = y$. With this choice, the connectivity matrix is given by

$$J_{ij} = \frac{c_{ij}}{Nc}\sum_{s=1}^{S}\sum_{\mu=1}^{P-1}\xi_i^{s,\mu+1}\xi_j^{s,\mu}. \tag{7}$$

where the learning rate A in Eq. (3) has been absorbed in the parameters of the transfer function (Eq. 2), by rewriting $\sigma^* = \frac{\sigma}{A}$ and $\theta^* = \frac{\theta}{A}$. The input current to neuron i at time t is given by a weighted sum of the firing rates of all its pre-synaptic neurons:

$$h_i(t) = \sum_{j\neq i} J_{ij}r_j(t). \tag{8}$$

We start by assuming that the number of sequences S is large, $S \gg 1$, and the number of patterns per sequence is much smaller than the in-degree, $P \ll cN$. Assuming that the dynamics start from an initial condition that is correlated with the first pattern of sequence s, i.e. $\vec{\xi}^{s,1}$, the input current can be re-written as

$$h_i(t) = \sum_{\mu=1}^{P-1}\xi_i^{s,\mu+1}\frac{1}{Nc}\sum_{j\neq i}^{N} c_{ij}\xi_j^{s,\mu}r_j(t) + Y_i(t) \tag{9}$$

where Y_i describes the 'noise' term,

$$Y_i(t) = \frac{1}{Nc}\sum_{l\neq s}^{S}\sum_{\mu=1}^{P-1}\xi_i^{l,\mu+1}\sum_{j\neq i}^{N}c_{ij}\xi_j^{l,\mu}r_j(t). \tag{10}$$

In the large cN limit, due to the law of large numbers, and using the fact that $P \ll cN$, the first term in Eq. (9) converges in probability to

$$\sum_{\mu=1}^{P-1}\xi_i^{\mu+1}q_\mu^s(t), \tag{11}$$

where the q_μ^ss are given by

$$q_\mu^s(t) = \frac{1}{N}\sum_{j=1}^{N}\xi_j^{s,\mu}r_j(t). \tag{12}$$

The overlaps $\{q^s_\mu(t)\}^P_{\mu=1}$ are our first P order parameters. They describe how correlated the network state is with the stored patterns $\vec{\xi}^{s,1}, \vec{\xi}^{s,2}, \cdots, \vec{\xi}^{s,P}$. We assume that the network state is uncorrelated with the rest of stored patterns since $q^l_\mu(t) \sim O(1/\sqrt{N})$ for $l \neq s$. The 'noise term' Y_i then has mean zero, and variance

$$\mathrm{Var}\left(Y_i(t)\right) = \alpha M(t) \tag{13}$$

where the sequential load is defined by

$$\alpha \equiv \frac{S(P-1)}{Nc} \tag{14}$$

while M, the mean of the squared firing rate, is an additional order parameter defined by

$$M(t) = \frac{1}{N}\sum_{j=1}^{N} r_j^2(t). \tag{15}$$

In this theory we assume that the total number of stored patterns is much larger than the number of patterns in a sequence, i.e. $S \gg 1$. We can now plug Eqs. (9-15) in Eq. (1) to yield

$$\tau\frac{dr_i}{dt} = -r_i + \phi\left(\sum_{\mu=1}^{P-1} \xi_i^{\mu+1} q_\mu(t) + Y_i(t)\right). \tag{16}$$

Since all sequences are statistically equivalent, we have dropped the index s corresponding to the particular sequence of concatenated patterns. The variable $Y_i(t)$ corresponds to the interference produced by stored patterns that do not belong to the sequence being retrieved (in this case, sequence s). By the central limit theorem, the variables Y_i are approximately i.i.d. Gaussian random variables across neurons, i.e. $Y_i(t) \overset{iid}{\sim} \mathcal{N}(0, \alpha M(t))$. Using equation (12) we get the following dynamical equations for the overlaps:

$$\tau\frac{dq_l}{dt} = -q_l + \int \mathcal{D}\vec{\xi}\mathcal{D}y\xi^l\phi\left(\sum_{\mu=1}^{P-1}\xi^{\mu+1}q_\mu(t) + \sqrt{\alpha M(t)}y\right) \tag{17}$$

where $\mathcal{D}y = e^{-y^2/2}/\sqrt{2\pi}$ and $\mathcal{D}\vec{\xi} = \prod_{i=2}^{P}\mathcal{D}\xi^i$.

The dynamical equation for the first overlap (i.e. q_1) simplifies to

$$\tau\frac{dq_1}{dt} = -q_1. \tag{18}$$

where we have used the fact that ξ^1 does not appear in the argument of ϕ in the r.h.s. of Eq. (17).

To simplify the equations for the other overlaps, we define

$$R_l^2(t) = \sum_{k\neq l}^{P-1} q_k^2(t) + \alpha M(t). \tag{19}$$

Since the stored patterns are Gaussian, we write Eq. (17) as

$$\tau\frac{dq_l}{dt} = -q_l + \int D\xi^l\mathcal{D}x\xi^l\phi\left(\xi^l q_{l-1}(t) + R_{l-1}(t)x\right) \quad l = 2,\ldots,P \tag{20}$$

where ξ^l and x are independent standard normal random variables. Using the change of coordinates

$$v = \frac{\xi^l q_{l-1} + xR_{l-1}}{\sqrt{q_{l-1}^2 + R_{l-1}^2}}$$

$$u = \frac{\xi^l R_{l-1} - xq_{l-1}}{\sqrt{q_{l-1}^2 + R_{l-1}^2}}$$

where u and v are also uncorrelated standard normal random variables, equation (20) becomes

$$\tau\frac{dq_l}{dt} = -q_l + q_{l-1}G\Bigg(\sum_{\mu}^{P-1} q_\mu^2(t) + \alpha M(t)\Bigg) \quad l = 2, \dots, P \tag{21}$$

where

$$G(x) \equiv \frac{\int \mathcal{D}vv\phi\left(v\sqrt{x}\right)}{\sqrt{x}}. \tag{22}$$

Using the neural transfer function of Eq. (2), we can use integration by parts to simplify:

$$G(x) \equiv \frac{1}{\sqrt{2\pi(\sigma^2 + x)}}\exp\left(-\frac{\theta^2}{2(\sigma^2 + x)}\right). \tag{23}$$

By defining the 'delay line' matrix as $L_{ij} = \delta_{i,j+1}$ we can also write equation (22) in vectorial form

$$\tau\frac{d\vec{q}}{dt} = -\vec{q} + G\Bigg(\sum_{\mu}^{P-1} q_\mu^2(t) + \alpha M(t)\Bigg)L\vec{q}. \tag{24}$$

When S is of order 1, and P is of order Nc, the variance of the two terms in the r.h.s. of Eq. (9) are of the same order, and in particular the variance of the first term no longer vanishes. In this scenario, we assume that at any given time during retrieval of a sequence, the network state has a finite overlap with only a small fraction of the patterns in the retrieved sequence. The 'signal' term in Eq. (9) then needs to include only those patterns, while the noise term $Y_i(t)$ includes all the patterns that are far from the patterns that have currently a finite overlap with the network state. All the resulting equations are therefore the same as the ones derived above in the case $S \gg 1$. Note that when S is of order 1, and $P \ll cN$, the dynamics of Eq. 24 are driven mainly by the "signal" term, the "noise" term becomes negligible, and so incorrect assumptions about the statistics of the noise have only minimal consequences for the dynamics of the overlaps.

2.2 Average squared firing rate

The system of equations describing the dynamics of the overlaps, Eq. (24), depends on $M(t)$. The next step is therefore to find a self-consistent mean-field equation for M:

$$M(t) = \frac{1}{N}\sum_{i}^{N} r_i^2(t). \tag{25}$$

Taking the derivative with respect to time, we find

$$\tau\frac{dM}{dt} = \tau\frac{2}{N}\sum_{i}^{N} r_i(t)\frac{dr_i(t)}{dt} \tag{26}$$

$$= -2M + \frac{2}{N}\sum_{i} r_i(t)\phi(h_i(t)) \tag{27}$$

$$= -2M + e^{-t/\tau}\frac{2}{N}\left[\sum_{i} r_i(0)\phi(h_i(t)) + \sum_{i}\int_0^t \frac{du}{\tau}\; e^{u/\tau}\phi(h_i(u))\phi(h_i(t))\right] \tag{28}$$

where we have used the general solution of the ODE for r_i:

$$r_i(t) = e^{-t/\tau}\Big(r_i(0) + \int_0^t \frac{du}{\tau}\; e^{u/\tau}\phi\left[h_i(u)\right]\Big). \tag{29}$$

Recalling Eqs. (9-12), the field $h_i(t)$ in the first sum in the r.h.s. of Eq. (28) can be described as a random Gaussian variable, whose variance $\sigma_h^2(t)$ is the sum of the norm of the vector of overlaps, plus the noise term due to interference with the other stored sequences,

$$\sigma_h(t)^2 = \sum_{\mu}^{P-1} q_\mu^2(t) + \alpha M(t). \tag{30}$$

In the large N limit, this sum can be replaced by an integral over the Gaussian distribution of $h_i(t)$.
We turn now to the other sum in the r.h.s. of Eq. (28). This sum can also be replaced in the large N limit by an integral over the joint distribution of $h(u)$ and $h(t)$, which is a correlated bivariate Gaussian distribution. Here we write $h(u)$ and $h(t)$ in terms of 3 uncorrelated Gaussian variables, where each are the sum of an independent and shared variable. Specifically, we write $h_i(u) = \sigma_h(u)(a(u,t)x + b(u,t)z)$, and $h_i(t) = \sigma_h(t)(a(t,u)y + b(t,u)z)$. It remains now to find the time-dependent parameters a and b in terms of the original order parameters. By computing the variance and covariance of the fields using both the original mean-field description and our Gaussian formulation, we can solve for a and b:

$$\langle h_i(t)^2 \rangle = \sum_{k}^{P-1} q_k^2(t) + \alpha M(t) = \sigma_h(t)^2(a^2(u,t) + b^2(u,t)) \tag{31}$$

$$\langle h_i(t)h_i(u) \rangle = \sum_{k}^{P-1} q_k(t)q_k(u) + \alpha C(t,u) = \sigma_h(u)\sigma_h(t)b^2(u,t) \tag{32}$$

where in Eqs. (31,32) we have used the fact that patterns $\{\xi_i^\mu\}$ are independent, and that x, y, and z are independent and uncorrelated. We have also defined $C(t,u) = \frac{1}{N}\sum_i r_i(u)r_i(t)$. Solving Eqs. (31,32), we find

$$a(t,u) = \sqrt{1-\rho(u,t)} \tag{33}$$

$$b(t,u) = \sqrt{\rho(u,t)} \tag{34}$$

where we have defined

$$\rho(u,t) = \frac{\sum_k^{P-1} q_k(t)q_k(u) + \alpha C(t,u)}{\sigma_h(u)\sigma_h(t)}.$$

Averaging over the statistics of x, y, and z we obtain:

$$\begin{aligned}\tau\frac{dM}{dt} =& -2M + 2e^{-t/\tau}\int D\xi Dz \phi(\xi)\phi(\sigma_h(t)z) \\ &+ 2\int_0^t \frac{du}{\tau}e^{(u-t)/\tau}\int DxDyDz\phi\big(\sigma_h(u)\sqrt{1-\rho(u,t)}x + \sigma_h(u)\sqrt{\rho(u,t)}z\big) \\ &\phi\big(\sigma_h(t)\sqrt{1-\rho(t,u)}y + \sigma_h(t)\sqrt{\rho(t,u)}z\big).\end{aligned} \tag{35}$$

The time evolution of $C(t,u)$ can be derived similarly as for Eq. (28),

$$\begin{aligned}\tau\frac{dC(t,u)}{dt} =& -C(t,u) + e^{-u/\tau}\int D\xi Dz\phi(\xi)\phi(\sigma_h(t)z) \\ &+ \int_0^u \frac{dv}{\tau}e^{(v-u)/\tau}\int DxDyDz\phi\big(\sigma_h(v)\sqrt{1-\rho(v,t)}x + \sigma_h(v)\sqrt{\rho(v,t)}z\big) \\ &\phi\big(\sigma_h(t)\sqrt{1-\rho(t,v)}y + \sigma_h(t)\sqrt{\rho(t,v)}z\big).\end{aligned} \tag{36}$$

The dynamics of the average firing rate $\bar{r}(t) = \frac{1}{N}\sum_i r(t)$ are given by:

$$\tau\frac{d\bar{r}}{dt} = \int Dz\phi(\sigma_h(t)z) + e^{(-t/\tau)}\Big(-\bar{r}(0) - \int_0^t \frac{du}{\tau}e^{(u/\tau)}\int Dz\phi(\sigma_h(u)z)\Big). \tag{37}$$

In Supplementary Figure 4, solutions to equations (35-36) are plotted along with numerically computed values from a full network simulation of size $N = 50{,}000$.

2.3 Retrieval properties

To study the time-dependent properties of retrieval, we can analyze Eq. (21) for the case in which the gain is constant: $G = 1 + \epsilon$. With this approximation, it is straightforward to analytically derive several properties of the recalled sequence, including retrieval speed and the scaling of overlap widths.

For $l = 2, \ldots, P$, we have:

$$\tau \frac{dq_l}{dt} = -q_l + (1+\epsilon)q_{l-1}$$

which leads by recursion to

$$q_l(t) = \frac{q_1(t=0)(1+\epsilon)^{l-1}t^{l-1}\exp(-t/\tau)}{\tau^{l-1}(l-1)!} \tag{38}$$

From this equation, we can easily see that for $\epsilon > 0$, sequences grow unbounded, while for $\epsilon < 0$ sequences decay. Furthermore, by computing the derivative of $q_l(t)$ with respect to time, we find that the overlaps q_l peak at $t = \tau(l-1)$, which shows that the sequence progresses at a speed proportional to τ.

To determine the widths of the overlaps vs time curves, we compute the standard deviation of the distribution given by $q_l(t)/\int q_l(u)du$. We find that the mean is equal to τl, while the standard deviation is $\tau\sqrt{l}$. Thus, the width of the overlap vs time curves is proportional to the square root of the peak time. This prediction agrees well with the empirically measured values for full network simulations (see Figure 2 of main article).

2.4 Sequence capacity

In the following sections, we will calculate the maximum number of sequences that a network can store and successfully retrieve as a function of network parameters.

2.4.1 Conditions on transfer function for successful retrieval

In the previous section we found that recalled sequences with a constant gain above one grow unbounded, while those with gain below one decay. To make the notion of retrieval more precise, we can define a sequence of asymptotically long size as being successfully retrieved if the gain converges to a value larger or equal to one during the sequence.

The shape of the gain function G in Eq. (23) dictates whether it can take values that are greater than one, and its form depends on the transfer function parameters θ and σ. By examining the dependence of G on these two parameters, we can bound the region for which successful retrieval is possible. Note that the shape of G only determines whether it is in principle possible to successfully retrieve a sequence. Whether or not the temporal trajectory of the gain rises above one during recall depends on the initial condition of the overlaps and mean squared firing rate, as well as the number of stored sequences. We find that G is either a monotonically decreasing function (if $G'(0) < 0$), or that it is a monotonically increasing function for small values of its argument, reaching a maximum $G_{\max}$, and then monotonically decaying towards zero (see Fig. 3c for examples). Thus, successful retrieval is possible if at least one of the following conditions is satisfied:

1. $G(0)$ is larger than one

2. $G'(0)$ is positive, and the maximum of G, $G_{\max}$, is larger than one

These criteria lead to the following conditions for θ and σ, using G defined in Eq. (23):

1. If $|\theta| < \sigma\sqrt{-\log(\sqrt{2\pi\sigma^2})}$, then $G(0) > 1$

2. If $|\theta| > \sigma$, then $G'(0) > 0$

3. If $|\theta| < 1/\sqrt{2e\pi}$ and $|\theta| < \sigma\sqrt{-\log(\sqrt{2\pi\sigma^2})}$ then $G_{\max} > 1$.

These conditions define 6 possible regions, that are plotted in Supplementary Figure 3a. In region D and E (see figure), condition 1 is satisfied ($G(0) > 1$), so retrieval is possible for vanishingly small initial overlaps, provided α is sufficiently small. In region F, conditions 2 and 3 are satisfied but not 1 ($G(0) < 1$, but G initially increases with its argument and reaches a peak $G_{\max} > 1$), so retrieval is possible for initial overlaps of small but finite size, again provided α is small enough. In regions A, B, and C retrieval is not possible, as both $G(0) < 0$ and $G_{\max} < 1$.

2.4.2 Maximum load

As already mentioned, a necessary condition for sequence retrieval is that the function $G(x)$ has a maximum that is larger than 1. If this condition is satisfied, then let x_c be the largest value of x such that $G(x) = 1$. For $\alpha \to 0$, the condition $G_{max} > 1$ guarantees that a sequence can be retrieved, provided the initial overlap with the first pattern is large enough. If $G\big(\sum_{\mu=1}^{P-1} q_\mu^2(0)\big) > 1$, the overlaps initially grow until the norm of the overlap vector stabilizes at a value where $G\big(\sum_{\mu=1}^{P-1} q_\mu^2(t)\big) \sim 1$.

When $\alpha > 0$, the gain now depends on the additional 'noise term' $\alpha M(t)$, since the argument of G is now $\sum_{\mu=1}^{P-1} q_\mu^2(t) + \alpha M(t)$. If this noise term is larger than x_c, then the gain will be smaller than one for any $\vec{q}(t)$, and therefore sequences will decay starting from any initial condition. The maximal value of α for which sequences can be retrieved is therefore given by

$$G(\alpha_c M) = 1, \tag{39}$$

where M is given by its steady state value when $\vec{q} = 0$:

$$M = \int \mathcal{D}v \phi^2 \left(v\sqrt{\alpha_c M}\right). \tag{40}$$

When $\alpha < \alpha_c$, there exists an overlap vector for which $\sum_{\mu=1}^{P-1} q_\mu^2(0) + \alpha M = x_c$. With suitable initial conditions, the dynamics of the network will converge to a vector with such a norm, and sequences will be retrieved. When $\alpha > \alpha_c$, $\sum_{\mu=1}^{P-1} q_\mu^2(0) + \alpha M > x_c$ and hence $G < 1$ for any $\vec{q}$, and therefore stored sequences cannot be retrieved.

Equations (39-40) were solved numerically. Solutions are plotted in Fig. 3 of the main text as a function of σ for a few values of θ, and in Supplementary Fig. 7 in the whole σ-θ plane.

2.5 Sequence robustness

We assessed sequence robustness by perturbing the initial condition of the rates by a Gaussian pattern: $r(t = 0) = \phi(\xi^{1,1} + \sigma_z z_0)$, where σ_z controls the standard deviation of the standard Gaussian perturbation $z_0 \sim \mathcal{N}(0, 1)$. We focused specifically on transfer function parameters in region F (see section 2.4.1), as sequences with parameters falling in regions D and E will still be retrieved for arbitrarily small initial overlaps. Surprisingly, we find in region F that increasing α results in higher robustness to perturbations (Supplementary Figure 10). Mean-field analysis of the perturbation provides an explanation for this effect. The perturbed initial conditions for q and M are given by:

$$q_1(t = 0) = \int DvDzv\phi(v + \sigma_z z)$$
$$M(t = 0) = \int Dv\phi(v\sqrt{1 + \sigma_z^2})^2$$

with all other overlaps equal to zero, $q_l(t = 0) = 0, l \neq 1$. Increasing perturbation strength increases slightly $M(0)$ while more dramatically decreasing $q_1(0)$. Looking at the mean-field dynamics of M in Eqs. (35) and (36), we can also see that the effect of the perturbation decreases exponentially with τ.

As explained in the previous section, successful recall depends on the argument of G: $\sum_{\mu=1}^{P-1} q_\mu^2(t) + \alpha M$, specifically on whether or not this argument stabilizes to a value around x_c. For small α, this argument is dominated by the overlap norm, and as the perturbation strength increases, the initial value of this norm decreases. When the initial norm becomes too small, the dynamics of q and M decay to a region where $G\big(\sum_{\mu=1}^{P-1} q_\mu^2(0) + \alpha M\big) < 1$, and the sequence is not retrieved. For example, for $S = 1$ and $P = 16$ corresponding to an $\alpha = 0.08$, when $\sigma = 2.5$ the initial argument of G is too small to converge to x_c (Supplementary Figure 10, left).

For large α, the initial argument is dominated by $\alpha M(0)$. Increasing perturbation strength decreases the initial norm of the overlaps, as in the case for small α, but $\alpha M(0)$ can remain large enough such that the argument converges in time to x_c. This phenomenon ensures that the critical initial overlap below which a sequence is not retrieved (due to the dynamics of the argument not converging to x_c) is smaller, and thus confers a higher tolerance for perturbation strength.

3 Excitatory-inhibitory rate network

So far, we have analyzed a simplified rate network that ignores the separation between excitation and inhibition. To assess whether our results hold also in networks that obey Dale's law, we developed a procedure to build an E/I network that can store and retrieve sequences. Our goal is to transform a rate network of the form

$$\tau \frac{dh_i}{dt} = -h_i + \sum_{j=1}^{N} J_{ij}\phi(h_j) \tag{41}$$

where J_{ij} are unconstrained, into the following two-population network, composed of N_E E neurons and N_I I neurons:

$$\tau_E \frac{dh_i^E}{dt} = -h_i^E + \sum_{j=1}^{N_E} J_{ij}^{EE}\phi^E(h_j^E) - \sum_{j=1}^{N_I} J_{ij}^{EI}\phi^I(h_j^I) \tag{42}$$

$$\tau_I \frac{dh_i^I}{dt} = -h_i^I + \sum_{j=1}^{N_E} J_{ij}^{IE}\phi^E(h_j^E) \tag{43}$$

such that the excitatory population in Eq. (42) shares the same pattern overlap dynamics as those in Eq. (41), and excitatory (inhibitory) units send only positive (negative) projections. Note that for simplicity we ignore inhibitory to inhibitory connections,

The four connectivity matrices are given by

$$J_{ij}^{EE} = \frac{c_{ij}^{EE}}{\sqrt{K_{EE}}}\omega\left(\frac{A_{EE}}{\sqrt{K_{EE}}}\sum_{s=1}^{S}\sum_{\mu=1}^{P-1} \mathrm{f}(\xi_i^{s,\mu+1})\mathrm{g}(\xi_j^{s,\mu})\right) \tag{44}$$

$$J_{ij}^{EI} = \frac{c_{ij}^{EI} J_{EI}}{\sqrt{K_{EI}}} \tag{45}$$

$$J_{ij}^{IE} = \frac{c_{ij}^{IE} J_{IE}}{K_{IE}} \tag{46}$$

where the synaptic transfer function ω has zero support at negative values, and all types of connections have sparse connectivity (i.e. $\mathbb{E}(c_{ij}^{EE}N_E) = K_{EE} \ll N_E$, $\mathbb{E}(c_{ij}^{IE}N_E) = K_{IE} \ll N_E$, $\mathbb{E}(c_{ij}^{EI}N_I) = K_{EI} \ll N_I$). Note that in this model, both excitatory and inhibitory synaptic efficacies onto excitatory neurons scale as $1/\sqrt{K}$, as in the balanced network model [2, 3] and other associative memory models with separate E and I populations [4, 5], while excitatory synapses onto inhibitory neurons scale as $1/K$.

Note also that we have assumed that TAH learning takes place only in excitatory recurrent weights J_{ij}^{EE}, and that all other connection strengths are fixed. For the sake of simplicity, we make the following additional assumptions:

1. Inhibition is fast ($\tau_I \ll \tau_E$)

2. Inhibitory firing rates depend linearly on their input (i.e. $\phi^I(h^I) \approx g^I_\phi h^I$)

Using these assumptions, we can reduce Eqs. (42) and (43) to the following equation, that now only describes inputs to E neurons:

$$\tau_E \frac{dh_i^E}{dt} = -h_i^E + \sum_{j=1}^{N_E} J_{ij}^{EE}\phi(h_j^E) - \sum_{j=1}^{N_E} J_{ij}^{I}\phi(h_j^E) \tag{47}$$

where J^I_{ij} represents an effective inhibitory connectivity matrix, given by:

$$J^I_{ij} = g^I_\phi \sum_{k=1}^{N^I} J^{EI}_{ik} J^{IE}_{kj} = \frac{g^I_\phi J_{IE} J_{EI}}{K_{IE}\sqrt{K_{EI}}} \sum_{k=1}^{N^I} c^{EI}_{ik} c^{IE}_{kj}, \tag{48}$$

Taylor expanding the excitatory connectivity in Eq. (44) around all but the first stored sequence, we find that only the first and second order terms will contribute to the total synaptic inputs in the $K \to \infty$ limit:

$$\begin{aligned} J^{EE}_{ij} \approx \frac{c^{EE}_{ij}}{\sqrt{K_{EE}}} \Bigg(& \omega\Bigg(\frac{A_{EE}}{\sqrt{K_{EE}}} \sum_{s>1}^{S} \sum_{\mu=1}^{P-1} \mathrm{f}(\xi_i^{s,\mu+1}) \mathrm{g}(\xi_j^{s,\mu}) \Bigg) \\ & + \frac{A_{EE}}{\sqrt{K_{EE}}} \omega'\Bigg(\frac{A_{EE}}{\sqrt{K_{EE}}} \sum_{s>1}^{S} \sum_{\mu=1}^{P-1} \mathrm{f}(\xi_i^{s,\mu+1}) \mathrm{g}(\xi_j^{s,\mu}) \Bigg) \sum_{\mu=1}^{P-1} \mathrm{f}(\xi_i^{1,\mu+1}) \mathrm{g}(\xi_j^{1,\mu}) \Bigg). \end{aligned} \tag{49}$$

Plugging the above into Eq. (47) and averaging the field over sequences $s \neq 1$, we get:

$$\mathbb{E}\Bigg(\sum_{j=1}^{N_E} J^{EE}_{ij}\phi(h^E_j) - \sum_{j=1}^{N_E} J^I_{ij}\phi(h^E_j) \Bigg) = \big(\sqrt{K_{EE}}\bar{\omega}_{EE} - \sqrt{K_{EI}} g^I_\phi J_{IE} J_{EI}\big) R_E + A_{EE}\bar{\omega}'_{EE} \sum_{\mu=1}^{P-1} \mathrm{f}(\xi^{1,\mu+1}) m^\mu \tag{50}$$

where $\mathbb{E}$ denotes an average over patterns in all sequences except the retrieved one, and the random structural connectivity matrix c_{ij}, and in addition we have defined

$$\begin{aligned} \bar{\omega}_{EE} &= \int_{-\infty}^{\infty} \mathcal{D}x \omega(A_{EE}\sqrt{\alpha\gamma}x) \\ \bar{\omega}'_{EE} &= \int_{-\infty}^{\infty} \mathcal{D}x \omega'(A_{EE}\sqrt{\alpha\gamma}x) \\ \gamma &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{D}x\mathcal{D}y f(x)^2 g(y)^2 \end{aligned}$$

and introduced the following order parameters:

$$\begin{aligned} R_E &= \mathbb{E}\Big(\phi(h^E) \Big) \\ q^\mu &= \mathbb{E}\Big(g(\xi^\mu)\phi(h^E) \Big). \end{aligned}$$

R_E represents the mean firing rate of the excitatory population, while q^μs are the overlaps with the patterns of the retrieved sequence.

The mean field, Eq. (50), is composed of two terms: The first term in the r.h.s. is proportional to the mean firing rate of the excitatory population. Such a term does not appear in the one population network. It scales as $\sqrt{K}$, and therefore diverges in the large K limit, unless there is a balance between excitation and inhibition, $\sqrt{K_{EE}}\bar{\omega}_{EE} = \sqrt{K_{EI}} g^I_\phi J_{IE} J_{EI}$. The second term contains the sum over overlaps with the patterns in the sequence. This term is the same as in the one population network, except for the additional factor $\bar{\omega}'_{EE}$. Thus, for the E-I network to be described by the same equations as the one population network, a balance between excitation and inhibition is required [2, 3, 4, 5].

Thus, in the E-I network, we impose the condition

$$\sqrt{K_{EE}}\bar{\omega}_{EE} = \sqrt{K_{EI}} g_\phi^I J_{IE} J_{EI}. \tag{51}$$

In numerical simulations we used a rectified linear transformation $\omega(x) = [g_\omega \cdot x + o_\omega]_+$. Other parameters are specified in Table 7b.

3.1 Simulation procedure

To construct the excitatory-inhibitory rate network, we take the following steps:

1. We begin by simulating the single population rate network of Eq. (41) with connectivity specified by Eq. (3). We use a rectified linear rate transfer function, $\phi(h) = [g_\phi \cdot h]_+$, and the threshold plasticity rule of Eqs. (6-7) in the main text. We fix x_f and q_g to a desired coding level, and find values for A and g_ϕ that lead to sequence recall for a given sequence length.

2. We next construct a two-population rate network (Eqs. 42-43). We start by defining N_E excitatory neurons, each with the same rectified linear rate transfer function as in the previous step, $\phi^E(h) = [g_\phi \cdot h]_+$. We then construct the recurrent excitatory connectivity as specified in Eq. (44), with all plasticity-related free parameters fixed as before. To impose non-negative weights, we choose a rectified linear transformation for the synaptic transfer function $\omega(x) = [g_\omega \cdot x + o_\omega]_+$.

3. We next add N_I inhibitory neurons, each with a rectified linear rate transfer function, $\phi^I(h) = [g_\phi^I \cdot h]_+$. For sparse E-I and I-E connectivity, the weights J_{ij}^{IE} and J_{ij}^{EI} are constrained by our choice of ω, and Eq. 51. The initial condition for the inhibitory neurons is fixed to the initial excitatory population average firing rate (i.e. $\mathbb{E}(\phi^E(\mathrm{f}(\xi_i^{1,1}))))$.

4 Spiking network

To transform the rate network of Eqs. (42,43) into a spiking network, we mapped dynamics to a current-based leaky integrate-and-fire network. Single unit dynamics were governed by the following current-based equations. For $\alpha, \beta \in \{E, I\}$:

$$\tau_m^\alpha \frac{dV_\alpha^i}{dt} = \Theta\left(V_\alpha^i - V_\alpha^{\text{floor}}\right)\left(-V_\alpha^i + V_\alpha^{\text{rest}} + \sum_\beta \sum_{j\neq i}^{K_{\alpha\beta}} S_{\alpha\beta}^{ij} + I_\alpha + \sigma_\alpha \sqrt{\tau_m^\alpha} W_\alpha(t)\right) \tag{52}$$

$$\tau_s^\alpha \frac{dS_{\alpha\beta}^{ij}}{dt} = -S_{\alpha\beta}^{ij} + J_{\alpha\beta}^{ij} \tau_s^\alpha \sum_{t_\beta^k} \delta\big(t - t_\beta^k - D\big) \tag{53}$$

where D controls the synaptic delay, I_α controls the external input drive, and τ_{rp} controls the refractory period. σ_α controls the strength of the stochastic fluctuations induced by a white noise input $W_\alpha(t)$ with unit variance density. The Heaviside function Θ sets a lower bound on the attainable voltage, so that the membrane potential cannot be more hyperpolarized than a 'floor' V_α^{floor}. This lower bound captures in a simplified fashion the inhibitory reversal potential that prevents the neuronal membrane potential from going to arbitrarily hyperpolarized values. Without this lower bound, many neurons have voltages with unrealistically large hyperpolarizing deflections, as large as 100 mV below the resting potential. Note that retrieval still occurs without the implementation of this lower bound.

We simulated a spiking network with N_E = 20,000 excitatory units, and N_I = 5,000 inhibitory units. V_α^{floor} was set to -80 mV for excitatory units and $-\infty$ mV for inhibitory units, with $V_\alpha^{\text{rest}} = -70$ mV. We set $D = 1$ ms, $\tau_{rp} = 1$ ms, the reset potential $V_\alpha^{\text{reset}} = V_\alpha^{\text{rest}}$, and the input drive $I_\alpha = 0$ mV. We used the Euler-Maruyama method with a time step of 1 ms. A full list of parameters can be found in Table 7c.

4.1 Simulation procedure

To construct the full spiking network, we take the following steps:

1. We start by simulating an excitatory-inhibitory rate network, following the steps as outlined in section 4.2.

2. We then match both rectified linear rate transfer functions to those derived from leaky-integrate and fire (LIF) units operating in the presence of noise. We use the following equation for the transfer function of unit activity in population α [6]:

$$\nu_\alpha(x) = \tau_{rp} + \tau_m^\alpha \sqrt{\pi} \int_{\frac{V_\alpha^{\text{reset}} - x}{\sigma_\alpha}}^{\frac{V_\alpha^{\text{thresh}} - x}{\sigma_\alpha}} du e^{u^2} (1 + \text{erf}(u)).$$

   To fit the LIF transfer function, we fix the reset potential to zero, and the refractory period and membrane time constant to desired values. The desired membrane time constant should be less than the synaptic time constant (see final step). We leave as free parameters the threshold and strength of external white noise (σ_α). We minimize the Euclidean distance between the two transfer functions over the interval $x \in \{x_{\text{lower}}, x_{\text{upper}}\}$ using the L-BFGS-B optimization algorithm, bounding both the threshold and noise strength from below at zero. Starting from many random initial conditions we find that a global minimum is reached by the procedure.

3. To produce threshold, reset, and resting membrane potentials within a physiological range, we apply a linear transformation $\phi_V(x) = \lambda_V^\alpha x + \Delta_V^\alpha$ that shifts and scales these parameters. All connectivity weights are also scaled by the same factor. To impose a nonzero resting membrane potential, we fix $V_\alpha^{\text{rest}} = V_\alpha^{\text{reset}} = \Delta_V^\alpha$. Note that all weights $J_{ij}^{\alpha\beta}$ taken from the rate network are also divided by the synaptic time constant τ_α. The new parameters and connectivities are therefore:

$$\begin{aligned} V_\alpha^{\text{thresh}^*} &= \lambda_V^\alpha V_\alpha^{\text{thresh}} + \Delta_V^\alpha \\ V_\alpha^{\text{reset}^*} &= \Delta_V^\alpha \\ J_{\alpha\beta}^{ij} &= \lambda_V^\alpha J_{ij}^{\alpha\beta} / \tau_\alpha \\ \sigma_\alpha &= \lambda_V^\alpha \sigma_\alpha. \end{aligned}$$

   The initial conditions are $S_{EE}(0) = \phi^E(\lambda_V^E f(\xi^{1,1}))$, $S_{IE}(0) = \mathbb{E}[S_{EE}(0)]$, $S_{EI}(0) = 0$, $V_E(0) = \Delta_V^E$, and $V_I(0) = \Delta_V^I$, where z is a standard Gaussian variable.

4. We aim to keep the range of neural firing rates below saturation, and adjust several parameters to this effect:

   (a) The dynamic range of excitatory firing can be adjusted by scaling the gain of the excitatory linear transfer function, g_ϕ^E. The dynamic range of inhibitory firing can be adjusted by rescaling J_{EI} and J_{IE} while keeping their product constant.

   (b) The gain of the excitatory LIF transfer function is fixed by the gain of the original linear transfer function, g_ϕ. The gain of the inhibitory LIF transfer function can be controlled by adjusting the gain of the corresponding linear transfer function, while scaling J_{EI} inversely.

5. Finally, we build the LIF spiking network of Eqs. (52,53). We now have two timescales, one for the synapses (τ_s) and one for the single units (τ_m). The time constants of the rate network are mapped to the synaptic time constants. All other parameters, including weight matrices, are taken from the previous two steps. Note that if the fitted σ_α^V is unrealistically large and disrupts retrieval, we reduce it to a more realistic value. This is the case for the parameters of Figure 7 in the main article, and so we lower it to a value equal to half the spiking threshold (see Table 7c). An alternative approach would be to fix σ_α to a desired value, and fit the LIF transfer function using only the threshold as a free parameter.

5 Supplemental Procedures

All rate network simulations were performed using an adapative Runge-Kutta method of order 2(3) with relative and absolute error sizes of 1e-3 and 1e-6, respectively.

Measuring robustness

To measure robustness in the top row of Supplementary Figure 10, we computed the difference in the time-averaged norm of the overlaps between perturbed and unperturbed trials:

$$\kappa(t_0, t_1) = \mathbb{E}_{t_0<t<t_1}(\|\vec{m}_{\text{unperturbed}}(t)\|_2) - \mathbb{E}_{t_0<t<t_1}(\|\vec{m}_{\text{perturbed}}(t)\|_2)$$

where the expectation is over the time interval starting at t_0 and ending at t_1. Unperturbed overlaps $\vec{m}_{\text{unperturbed}}(t)$ in Supplementary Figure 10 correspond to those in Figure 1 of the main article. To generate the perturbed overlaps $\vec{m}_{\text{perturbed}}(t)$, we fixed the initial condition of rate units to $r(t=0) = \phi(\xi^{1,1} + \sigma_z z_0)$, where σ_z controls the standard deviation of the the standard Gaussian perturbation $z_0 \sim \mathcal{N}(0, 1)$.

We measured κ in two time intervals: 1) in the time interval leading up to the observed retrieval time, where $t_0 = 0$ and $t_1 = \text{argmax}_t\, m_P(t)$, and 2) in the latter half of this interval, where $t_0 = \text{argmax}_t\, m_P(t)/2$ and $t_1 = \text{argmax}_t\, m_P(t)$.

Retrieval time ratio

To compute the retrieval time ratio (RTR) in Supplementary Figure 2, we divided the observed retrieval time by the predicted retrieval time (see Section 2.3.1): $\text{RTR} = \text{argmax}_t\, m_P(t)/(\tau(P-1))$.

Peak distributions

Neurons

To find firing rate peaks for a given unit during a single trial of recall, we computed the average firing rate $\bar{r}_i$ for unit i and selected all continuous intervals of firing occurring one standard deviation above this threshold. The nth peak midpoint time for unit i was computed using a weighted average,

$$t_{i,n}^{\text{midpoint}} = \frac{\sum_{t_0^n<t<t_1^n} r(t) \cdot t}{\sum_{t_0^n<t<t_1^n} r(t)}.$$

where t_0^n and t_1^n mark the beginning and end of the nth continuous interval. The width $\sigma_{i,n}^{\text{width}}$ of a peak was defined as the length of the continuous interval: $t_1^n - t_0^n$. Single unit peaks that had maximal firing rate at t=0 were excluded from the analysis, but do not qualitatively alter the distribution if included.

Overlaps

To measure the width of each overlap, we used the following equation:

$$\sigma_\mu^{\text{width}} = \sqrt{\frac{\sum_t m_\mu(t) \cdot (t - t_\mu^{\text{mean}})^2}{\sum_t m_\mu(t)}}$$

where t_μ^{mean} is defined as:

$$t_\mu^{\text{mean}} = \frac{\sum_t t \cdot m(t)}{\sum_t m(t)}.$$

To compute the cumulative density of peak times for single neurons and overlaps, we measured the time interval starting at $t = 0$ and lasting up to the retrieval time (defined in the previous sections).

Measuring capacity in a finite network

To compute the mean-field correlation in Fig 3b, we use both M and $\bar{r}$ to normalize the overlaps by the standard deviation of the firing rates (see section 2.2).

To compute the mean-field capacity curve in Fig. 3c (dashed curves), we use a bisection method to converge on the smallest α for which the maximal overlap values in the sequence $\{\text{argmax}_t\, q_1(t), \text{argmax}_t\, q_2(t), ...,$ $\text{argmax}_t\, q_S(t)\}$ are a monotonically decaying function.

Retrieval with nonlinear rule

We initialize the network to $r(t = 0) = \phi(f(\xi^1))$. To measure the pattern correlations in Figure 4d, we compute the Pearson correlation coefficient between $r(t)$ and $g(\xi^l)$ for each pattern ξ^l.

Selectivity criterion

To generate the sorted raster plots of Fig. 5, we first divided units into active (silent) pools according to their maximum firing rate: $\text{argmax}_t\, r(t) > \theta$ ($\text{argmax}_t\, r(t) < \theta$), where θ defines a minimal activity threshold. We set $\theta = 0.05 \cdot r_{\text{max}}$. Selective units were defined by the union or intersection of these pools in different stimulus contexts.

Turn selective units

Right (left) selective units were those that had membership in the active pool during recall of the right (left) sequence, and were in the silent pool during recall of the left (right) sequence.

Non-specific units

To determine non-specific units, we first sought neurons that were active in both left and right stimulus contexts. We then computed the mean absolute difference (MD) in activity between the left and right trial for each of these units: $\text{MD} = \frac{1}{NT}\sum_i^N \sum_t^T \left[r_i^{\text{left}}(t) - r_i^{\text{right}}(t)\right]$. Units with $\text{MD} < 0.075$ were identified as non-specific.

Decaying weight perturbations

To generate the sorted raster plots of Fig. 6a, we included units that had a maximal activity of at least $0.15 \cdot r_{\text{max}}$ on day 1 (for day 1 sorted) or day 30 (for day 30 sorted). These accounted for roughly half of the total population of neurons.

Activity profile correlation

To measure the similarity of activity across days in Fig. 6c, we computed the Pearson correlation coefficient between rate trajectories $\vec{r}_n^{\,i}$ and $\vec{r}_m^{\,i}$, where n and m are simulation days, and i is the index of the neuron. We averaged this quantity across all neurons.

Spiking network

Quantifying correlations

To compute pattern correlations in Figure 7c, we transformed spiking activity into rates by convolving spikes with a Gaussian kernel (20 ms standard deviation). We then computed the Pearson correlation between rates and transformed patterns as in the rate network case with a nonlinear rule.

Sequential capacity of the nonlinear rule

To compute the storage capacity in Supplementary Figure 9, we numerically simulated retrieval of a sequence of length $P = 16$ in the presence of an unretrieved stored sequence of length P'. We used a bisection method to find the largest P' for which the final retrieved pattern correlation exceeded a threshold of $\kappa = 0.025$. If the final pattern correlation was less than κ for $P' = 0$, then the sequence capacity was $\alpha_c = 0$, otherwise it was reported as $\alpha_c = \frac{(P-1)+(P'-1)}{K}$.

We choose for simplicity $\sigma \to 0$ and rescaled the transfer function threshold θ by defining $\theta_0 = \theta/(p(1-p)(p(1-q)^2 + (1-p)q^2))$, where $p = 1 - q_g$ and $q = 1 - q_f$. We have assumed that the plasticity rule thresholds are the same (i.e. $x_f = x_g$), and as in Section 1.2, that $q_g = F_z(x_f)$. This rescaling of the threshold θ is necessary as the variance of the field scales with p, and so the threshold must be adjusted to maximize capacity [7, 1].

6 Supplemental Figures

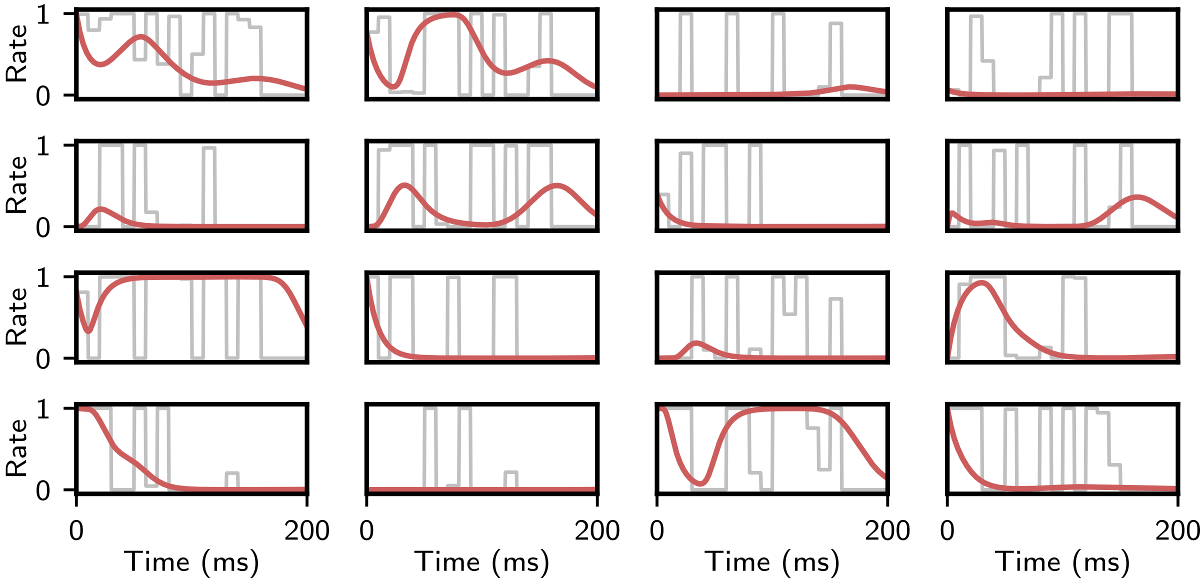

Figure 1: Single unit retrieval examples for the network in Figure 1 in the main text. Units are randomly selected from the whole population.

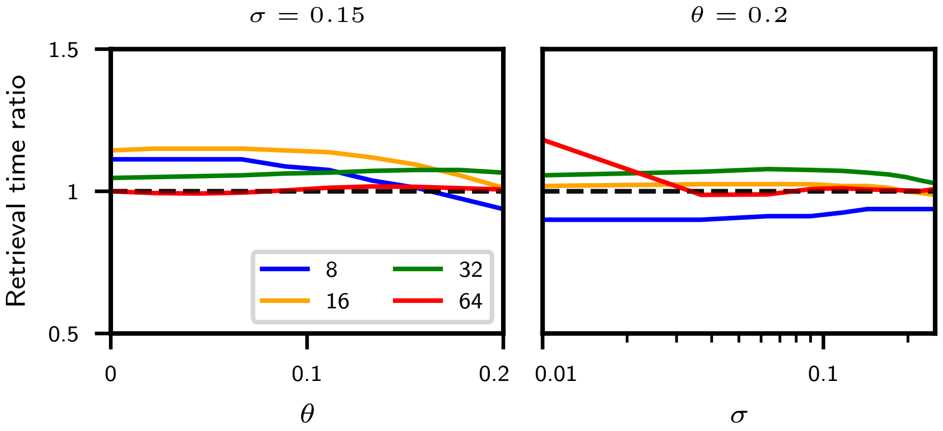

Figure 2: The ratio of the observed and predicted retrieval times (see Supplemental Procedures), for varying P, as a function of rate transfer function parameters θ and σ. All other parameters are as in Figure 1 of the main text.

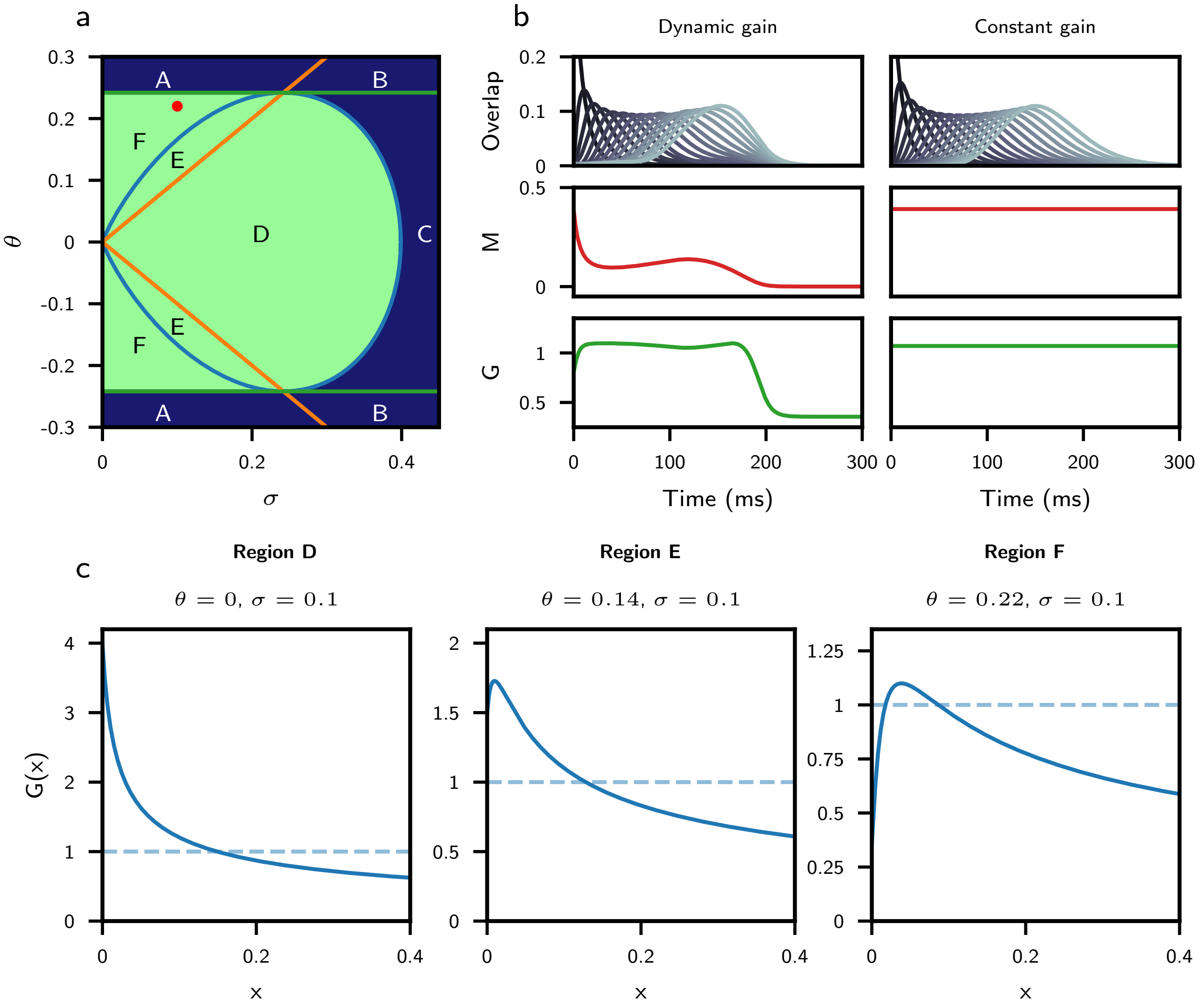

Figure 3: Gain function behavior **a**. Conditions for successful retrieval. Sequences can be retrieved for appropriate α and $q_1(0)$ in the green regions, and cannot be retrieved in the blue regions. Within D and E, retrieval is possible for vanishingly small initial overlaps. In region F, retrieval is possible for initial overlaps of small but finite size. In regions A, B, and C retrieval is not possible, as both $G(0) < 1$ and $G_{\max} < 1$. The blue line corresponds to condition 1, the orange line to condition 2, and the green line to condition 3 (see section 2.4.1). The maximal capacity within the green region is shown in Supplementary Figure 5. The red dot corresponds to the parameters in panel b. **b**. Overlaps as a function of time (top), average squared rate (middle), and gain function (bottom) for full network (left), and overlap dynamics with a constant M and G approximation (right). In the constant gain case, G has been fixed to the average value of G during retrieval in the dynamic gain case: 1.0718, and M is shown purely for illustration. All parameters in the "dynamic gain" case are as in Figure 1 of the main text. **c**. Solid lines are profiles of gain function G(x) as a function of x, for three sets of parameters corresponding to the three possible regions of successful retrieval. Dashed lines indicate threshold at one. The parameters chosen for region F correspond to the the red dot in panel a.

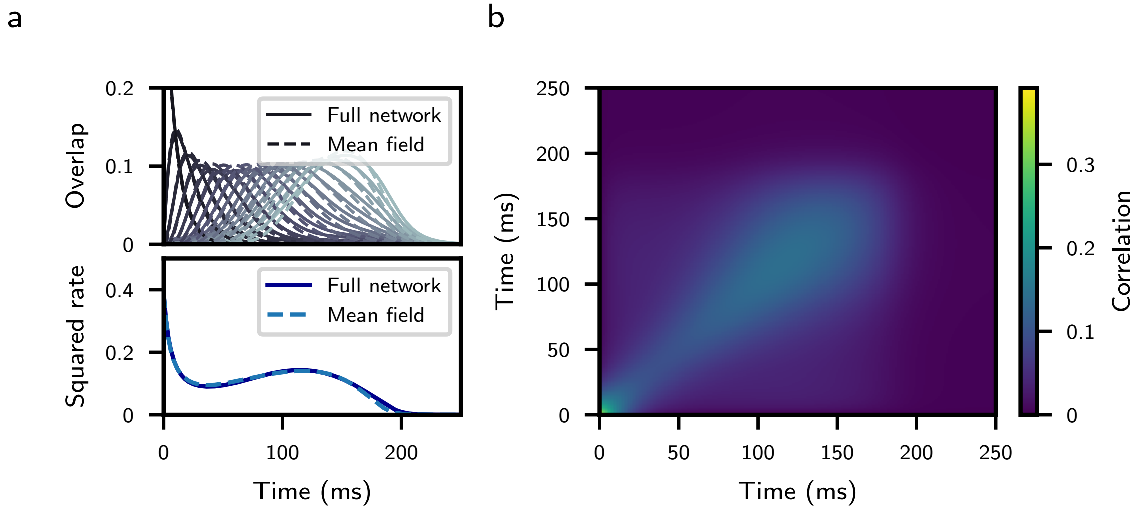

Figure 4: Comparison of mean-field with full network simulations **a**. Solid lines show overlaps computed using a simulation of the full network. Dashed lines are solutions to the mean-field equations. Top: Overlaps $\{q_\mu\}$ of network activity with stored patterns. Bottom: Average squared firing rates, M. **b**. Mean-field two time autocorrelation function C, as defined in section 2.2. All parameters are as in Figure 1 of the main text, except $N = 50,000$. Discrepancies in the mean-field are finite-size effects, and decrease with larger N.

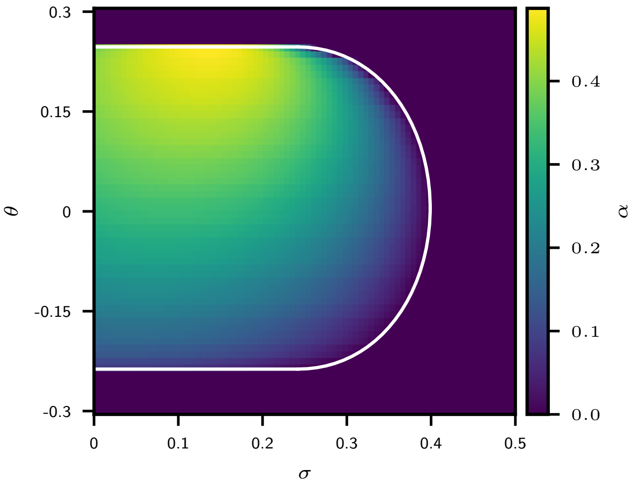

Figure 5: Maximal capacity, computed analytically using Eqs. (39,40) as a function of θ and σ, for the bilinear plasticity rule and the error-function transfer function. The white boundary corresponds to the successful retrieval region in Figure 3. Note that the full curves in Fig. 3c of the main text correspond to horizontal cuts in this plane.

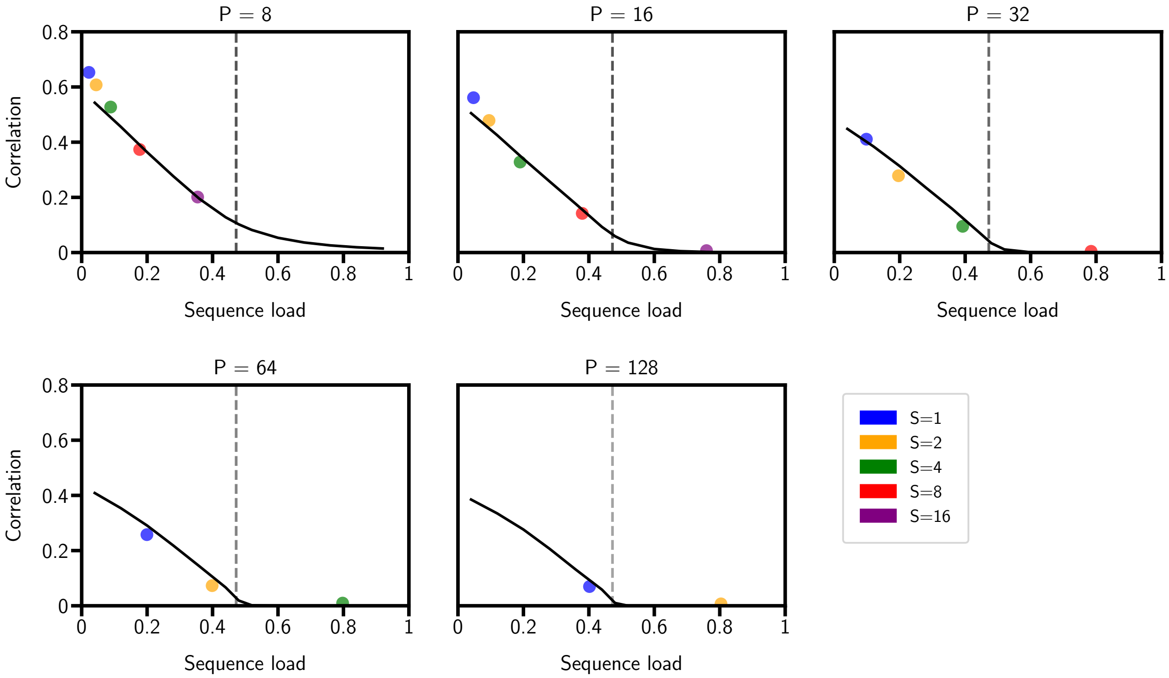

Figure 6: Maximal correlation with the final pattern in the stored sequence obtained from numerical simulations of the full network for various S and P. $N = 100{,}000$, $K = \sqrt{N}$ and all other parameters are as in Figure 3b of the main text. The dashed vertical lines correspond to the capacity computed in Figure 3b of the main text, and the solid curves to the correlations obtained from solving numerically the mean-field equations.

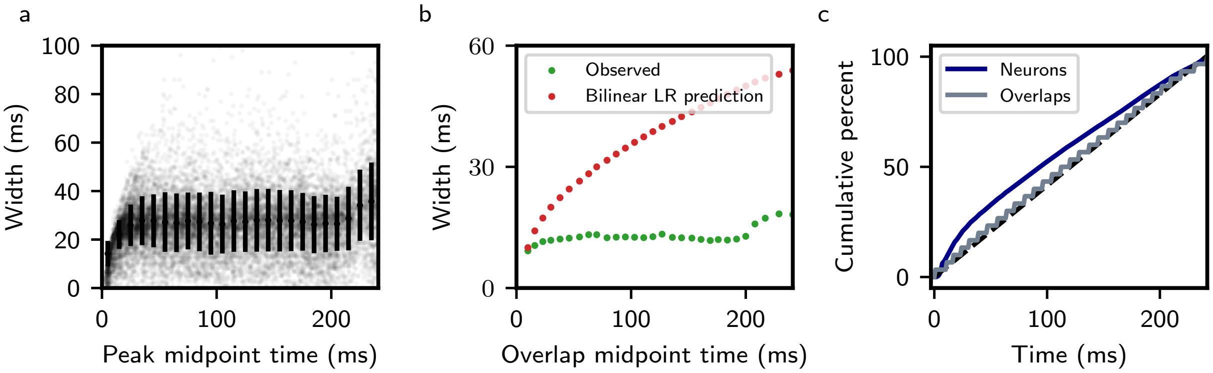

Figure 7: Temporal characteristics of a retrieved sequence stored with a nonlinear learning rule. **a**. Distribution of single neuron peak widths, defined as continuous firing intervals occurring one standard deviation above the time-averaged firing rate. Black error bars denote mean and standard deviation of widths within each 10 ms interval. **b**. Green dots indicate observed overlap widths (see Supp. Procedures). Analytic predictions for Gaussian patterns stored with bilinear learning rule (see Figure 2 of main text) are shown in red. **c**. Cumulative percentage of peak times for single neurons (blue) and overlaps (grey). The dashed black line represents a uniform distribution. All parameters are as in Figure 4 of the main text.

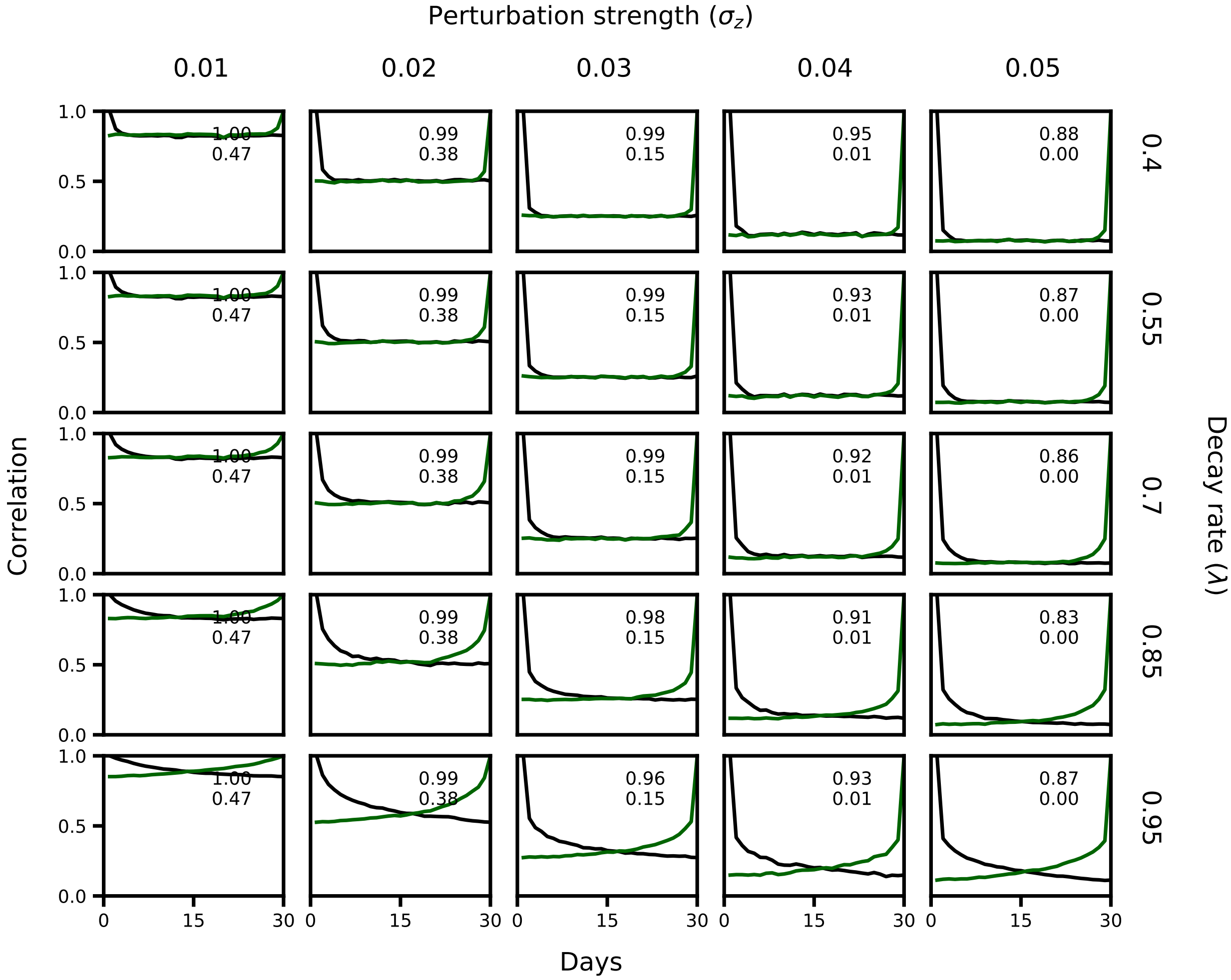

Figure 8: Average correlation of activity profiles between day n and either day 1 (black) or day 30 (green) for various σ_z and λ. The time-averaged correlation value between sequential overlaps on day 1 and 30 is given by the upper number in each panel. The lower number indicates the maximal correlation value with the final pattern on day 1. All other parameters are as in Figure 6 of the main text.

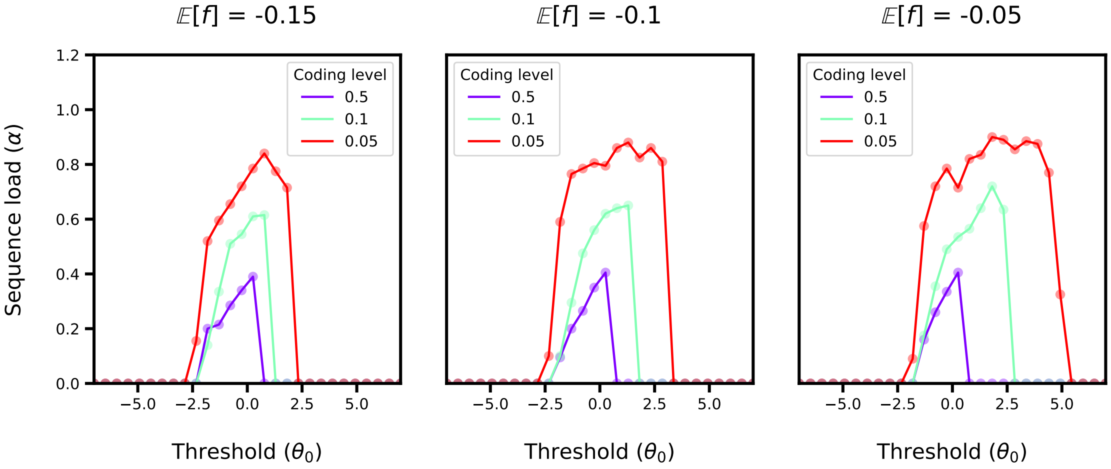

Figure 9: Storage capacity of the nonlinear learning rule, in the limit $\sigma \to 0$. Capacity is plotted as a function of the rescaled transfer function threshold for several coding levels and values of the average of the function f. $N = 50{,}000$, and all other parameters are as in Figure 4 of the main text. See Supplemental Procedures for details on how storage capacity was determined.

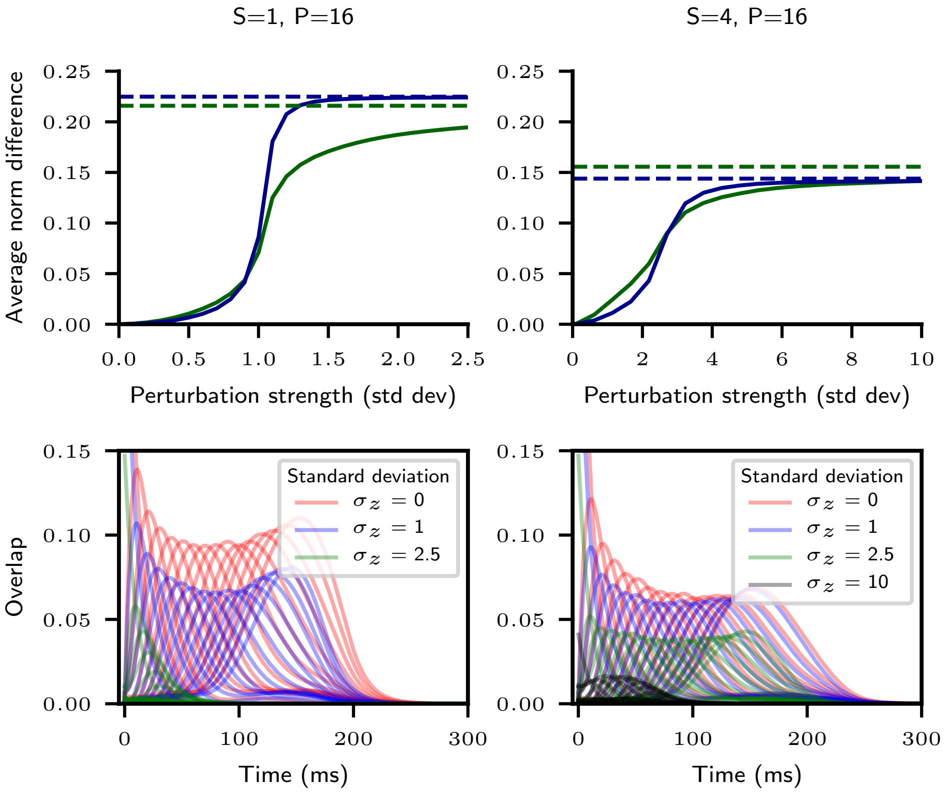

Figure 10: Robustness to initial perturbations of a network storing sequence length $P = 16$ for $S = 1$ (left column), and $S = 4$ (right column) sequence number. Top row: The difference in the time-averaged norm of the overlaps between perturbed and unperturbed trajectories, given as a function of the standard deviation of the initial Gaussian perturbation σ_z. This perturbation is added to the initial condition (ξ_1). In green, the difference is computed using the time interval leading up to retrieval time (see Supp. Procedures). In blue, the difference is computed using only the latter half of this interval. The green and blue dashed lines display the average overlap norms of the unperturbed trajectory for the full and latter half of retrieval, respectively. Bottom row: Overlaps during retrieval following an initial perturbation of strength σ_z. All other parameters are as in Figure 1 of the main text.

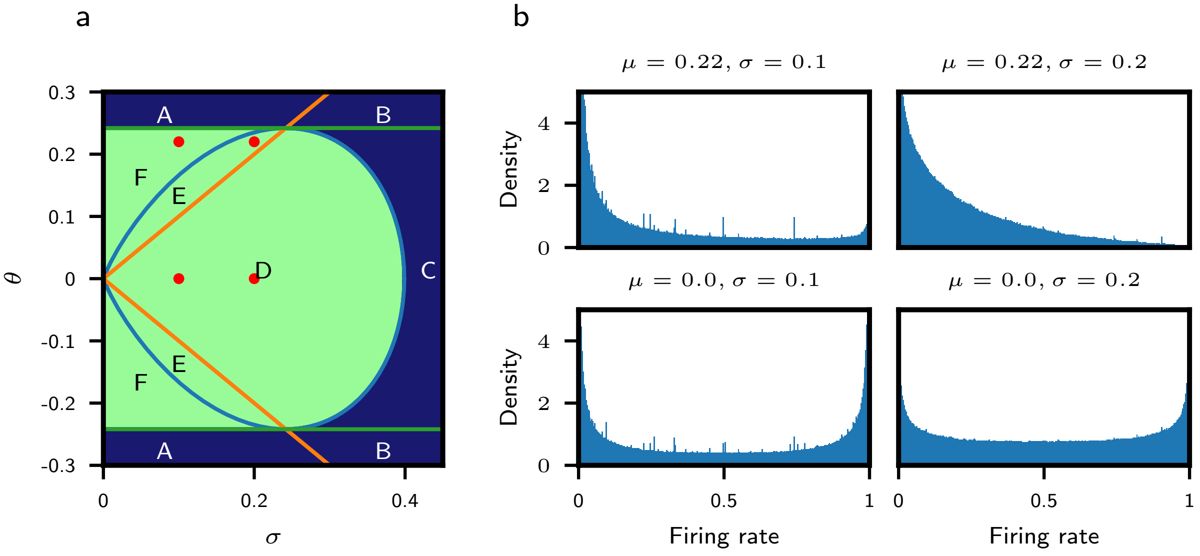

Figure 11: **a**. Conditions for successful retrieval (see Supp. Figure 3). Red dots correspond to panels in (b). **b**. Distribution of firing rates across neurons and time (the interval between time 0 and retrieval time). All other parameters are as in Figure 1 of the main text.

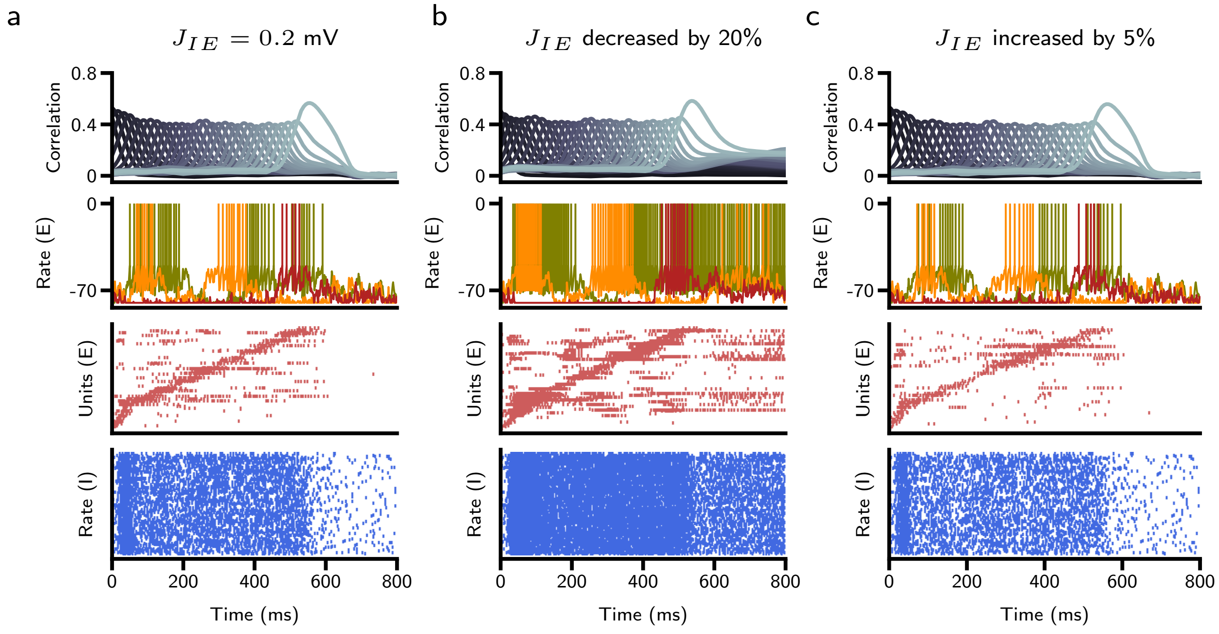

Figure 12: **a**. Spiking network from Figure 7 of the main text. **b**. Excitatory-to-inhibitory connection strength is decreased by 20%. **c**. Excitatory-to-inhibitory connection strength is increased by 5%. All other parameters are as in Figure 7 of the main text.

7 Parameter tables

Figure 1, 2: Retrieval of a stored sequence

Parameter	Value	Comment
N	40,000	Neuron count
c	0.005	Connection probability
θ	0.22	Rate transfer function offset
σ	0.1	Rate transfer function inverse gain
τ	10 ms	Time constant
S	1	Sequence number
P	16	Sequence length

Figure 3a: Sequence capacity

Parameter	Value	Comment
N	40,000	Neuron count
c	0.005	Connection probability
θ	0.22	Rate transfer function offset
σ	0.1	Rate transfer function inverse gain
τ	10 ms	Time constant
S	2	Sequence number
P	16	Sequence length

Figure 3b: Sequence capacity

Parameter	Value	Comment
θ	0.22	Rate transfer function offset
σ	0.1	Rate transfer function inverse gain
τ	10 ms	Time constant

Figure 3c: Sequence capacity

Parameter	Value	Comment
N	40,000	Neuron count
c	0.005	Connection probability
θ	0.2	Rate transfer function offset
σ	0.1	Rate transfer function inverse gain
τ	10 ms	Time constant

Figure 4: Retrieval with nonlinear learning rules

Parameter	Value	Comment
N	40,000	Neuron count
c	0.005	Connection probability
θ	0.005	Rate transfer function offset
σ	0.00357	Rate transfer function inverse gain
τ	10 ms	Time constant
S	1	Sequence number
P	30	Sequence length
x_{f}	1.645	Post-synaptic threshold of plasticity rule
x_{g}	1.645	Pre-synaptic threshold of plasticity rule
q_{f}	0.8	Plasticity rule parameter
q_{g}	0.95	Plasticity rule parameter

Figure 5: Selectivity emerges from random input patterns

Parameter	Value	Comment
N	40,000	Neuron count
c	0.005	Connection probability
θ	0.005	Rate transfer function offset
σ	0.00357	Rate transfer function inverse gain
τ	10 ms	Time constant
S	2	Sequence number
P	30	Sequence length
x_{f}	1.645	Post-synaptic threshold of plasticity rule
x_{g}	1.645	Pre-synaptic threshold of plasticity rule
q_{f}	0.8	Plasticity rule parameter
q_{g}	0.95	Plasticity rule parameter

Figure 6: Changes in synaptic connectivity preserve collective sequence retrieval

Parameter	Value	Comment
N	40,000	Neuron count
c	0.005	Connection probability
θ	0.005	Rate transfer function offset
σ	0.00357	Rate transfer function inverse gain
τ	10 ms	Time constant
S	1	Sequence number
P	30	Sequence length
x_{f}	1.645	Post-synaptic threshold of plasticity rule
x_{g}	1.645	Pre-synaptic threshold of plasticity rule
q_{f}	0.8	Plasticity rule parameter
q_{g}	0.95	Plasticity rule parameter
λ	0.85	Decay rate
σ_z	0.03	Perturbation strength

Figure 7a: One population rate network

Parameter	Value	Comment
N	20,000	Neuron count
c	0.04	Connection probability
g_ϕ	12	Rate transfer function gain
τ	20 ms	Time constant
S	1	Sequence number
P	32	Sequence length
A	6.3	Learning strength
x_{f}	1.5	Post-synaptic threshold of plasticity rule
x_{g}	1.5	Pre-synaptic threshold of plasticity rule
q_{f}	0.8	Plasticity rule parameter
q_{g}	0.933	Plasticity rule parameter

Figure 7b: Two population rate network

Parameter	Value	Comment
N_E	20,000	Excitatory neuron count
N_I	5,000	Inhibitory neuron count
c_{EE}	0.04	Connection probability
c_{IE}	0.04	Connection probability
c_{EI}	0.04	Connection probability
g_ϕ^{E}	12	Rate transfer function gain, excitatory
τ_E	20 ms	Time constant, excitatory
g_ϕ^{I}	20	Rate transfer function gain, inhibitory
τ_I	5 ms	Time constant, inhibitory
S	1	Sequence number
P	32	Sequence length
A_{EE}	6.3	Learning strength
x_{f}	1.5	Post-synaptic threshold of plasticity rule
x_{g}	1.5	Pre-synaptic threshold of plasticity rule
q_{f}	0.8	Plasticity rule parameter
q_{g}	0.933	Plasticity rule parameter
g_ω	1	Synaptic transfer function gain
o_ω	-0.0038	Synaptic transfer function offset
J^{IE}	0.0673	Synaptic weight
J^{EI}	0.0250	Synaptic weight

Figure 7c: Two population spiking network

Parameter	Value	Comment
N_E	20,000	Excitatory neuron count
N_I	5,000	Inhibitory neuron count
c_{EE}	0.04	Connection probability
c_{IE}	0.04	Connection probability
c_{EI}	0.04	Connection probability
τ_m^{E}	10 ms	Membrane time constant, excitatory
τ_s^{E}	20 ms	Synaptic time constant, excitatory
$V_{\mathrm{thresh}}^{\mathrm{E}}$	-50 mV	Spiking threshold, excitatory
$V_{\mathrm{reset}}^{\mathrm{E}}$	-70 mV	Voltage reset, excitatory
$V_{\mathrm{rest}}^{\mathrm{E}}$	-70 mV	Voltage resting potential, excitatory
$V_{\mathrm{reversal}}^{\mathrm{E}}$	-80 mV	Voltage floor potential, excitatory
$\tau_{\mathrm{rp}}^{\mathrm{E}}$	1 ms	Refractory period, excitatory
σ_{E}	10 mV	White noise strength (std dev), excitatory
τ_m^{I}	2 ms	Membrane time constant, inhibitory
τ_s^{I}	5 ms	Synaptic time constant, inhibitory
$V_{\mathrm{thresh}}^{\mathrm{I}}$	-50 mV	Spiking threshold, inhibitory
$V_{\mathrm{reset}}^{\mathrm{I}}$	-70 mV	Voltage reset, inhibitory
$V_{\mathrm{rest}}^{\mathrm{I}}$	-70 mV	Voltage floor potential, inhibitory
$V_{\mathrm{reversal}}^{\mathrm{I}}$	$-\infty$ mV	Voltage reversal potential, inhibitory
$\tau_{\mathrm{rp}}^{\mathrm{I}}$	1 ms	Refractory period, inhibitory
σ_{I}	10 mV	White noise strength (std dev), inhibitory
S	1	Sequence number
P	32	Sequence length
A	6.3	Learning strength
x_{f}	1.5	Post-synaptic threshold of plasticity rule
x_{g}	1.5	Pre-synaptic threshold of plasticity rule
q_{f}	0.8	Plasticity rule parameter
q_{g}	0.933	Plasticity rule parameter
g_ω	1	Synaptic transfer function gain
o_ω	-0.0038	Synaptic transfer function offset
J^{IE}/K_{IE}	0.204 mV	Synaptic weight
$J^{\mathrm{EI}}/\sqrt{K_{EI}}$	0.228 mV	Synaptic weight
λ_V^E	5.167	Excitatory rescaling factor
λ_V^I	3.089	Inhibitory rescaling factor

References

[1] U. Pereira and N. Brunel. Attractor Dynamics in Networks with Learning Rules Inferred from In Vivo Data. *Neuron*, 99(1):227–238, Jul 2018.

[2] C. van Vreeswijk. Partial synchronization in populations of pulse-coupled oscillators. *Phys. Rev. E*, 54:5522–5537, 1996.

[3] C. van Vreeswijk and H. Sompolinsky. Chaotic balanced state in a model of cortical circuits. *Neural Computation*, 10:1321–1371, 1998.

[4] R. Rubin, L. F. Abbott, and H. Sompolinsky. Balanced excitation and inhibition are required for high-capacity, noise-robust neuronal selectivity. *Proc. Natl. Acad. Sci. U.S.A.*, 114(44):E9366–E9375, 10 2017.

[5] G. Mongillo, S. Rumpel, and Y. Loewenstein. Inhibitory connectivity defines the realm of excitatory plasticity. *Nat. Neurosci.*, 21(10):1463–1470, 10 2018.

[6] L. M. Ricciardi. *Diffusion processes and Related topics on biology.* Springer-Verlag, Berlin, 1977.

[7] M. V Tsodyks. Associative Memory in Asymmetric Diluted Network with Low Level of Activity. *Europhysics Letters (EPL)*, 7(3):203–208, October 1988.