Conditions for replay of neuronal assemblies

Gaspar Cano[1], Richard Kempter[1,2,3*]

1 Institute for Theoretical Biology, Department of Biology, Humboldt-Universität zu Berlin, Germany
2 Bernstein Center for Computational Neuroscience Berlin, Germany
3 Einstein Center for Neurosciences Berlin, Germany

* r.kempter@biologie.hu-berlin.de

Abstract

From cortical synfire chains to hippocampal replay, the idea that neural populations can be activated sequentially with precise spike timing is thought to be essential for several brain functions. It has been shown that neuronal sequences with weak feedforward connectivity can be replayed due to amplification via intra-assembly recurrent connections. However, this phenomenon was thought to depend on inhibitory feedback, but its mechanisms were still unclear. Here, we arrive at a minimal spiking model that shows that feedback inhibition is not needed for this amplification to occur. We then introduce a population model of membrane-potential distributions that explains the spiking network behavior, and we analytically describe how different connectivity structures determine replay speed, with weaker feedforward connectivity generating slower pulses that can be sustained by recurrent connections. These pulses can only propagate if neuronal leak currents are slow enough with respect to the pulse speed. Together, our simulations and analytical results predict the conditions for replay of neuronal assemblies.

Introduction

The notion that groups of neurons could form assemblies that fire together and that these assemblies could then fire in a specific order is an old one [1]. In fact, the idea that some neural populations are activated sequentially is ubiquitous in neuroscience, with precise neuronal spike timing thought to be essential for several brain functions.

In the cortex, neuronal recordings in behaving animals have repeatedly yielded evidence of such precise spike-timing sequences [2–5]. Thus, networks of layers of neurons connected in a feedforward manner — called synfire chains — have been extensively studied to try and understand how such sequences of neuronal activity could propagate in the cortex [6–14]. These modeling studies have found that, given enough feedforward connectivity across groups of neurons, neuronal networks are capable of indefinitely propagating an activity pulse with precise spike timing. Interestingly, it has also been shown that adding feedback connections from one layer to the preceding one stabilizes this propagation [15].

In the hippocampus, the sequential activation of learned neuronal sequences is thought to be essential for memory consolidation [16–22]. This phenomenon, known as hippocampal replay, has been observed in the brain of many animals, including humans [23–29]. Several computational studies have investigated what properties the hippocampal neuronal network might need in order to successfully replay a learned sequence [30–34].

Computational studies have focused on how the interplay between recurrent connections (within an assembly) and feedforward connections (across assemblies) might influence stability and sequential activation of activity patterns [35, 36]. Such recurrent connections should be expected to exist in the brain due to Hebbian plasticity mechanisms ("fire together, wire together") and would be required for pattern completion, a key function attributed to the hippocampus [16, 17, 20, 22, 37]. More recently, [34] showed that in a balanced excitatory-inhibitory (EI) network, recurrent connections within an assembly of neurons can facilitate the sequential activation of assemblies. If feedforward connections are not strong enough to propagate a signal across assemblies of neurons, then adding recurrent connections at each layer/assembly can amplify the signal and allow activity propagation to succeed. In such cases, neurons in

an assembly need not fire more than once, just like what typically happens in synfire chains, where there are no recurrent connections. The authors of [34] hypothesized that, for recurrent connections to aid in sequence retrieval, a 'balanced amplification' mechanism [38] would be necessary. Such a mechanism requires a significant amount of feedback inhibition to the excitatory assemblies, which their model possessed. However, it remains unclear whether feedback inhibition and balanced amplification are really necessary for recurrent connections to facilitate sequence retrieval.

The goal of this work is to understand the mechanisms by which recurrent connections enable or facilitate the feedforward activation of sequences of neuronal assemblies. To this end, we first study spiking networks of leaky integrate-and-fire (LIF) neurons capable of retrieving an embedded sequence. Our results suggest that inhibition and balanced amplification are not necessary for recurrent connections to cooperate with weak feedforward connectivity and enable sequence retrieval. Then, we develop a time-discrete population model of the membrane-potential distributions for each assembly. Our population model provides analytical estimations that describe our simulations well and can predict how replay speed depends on network parameters, as well as the necessary connectivity a sequence must have to be replayed.

Results

Sequence retrieval in spiking networks

To understand how recurrent connections can amplify propagating activity pulses and enable the sequential activation of neuronal assemblies, we first perform spiking simulations of LIF neuronal networks with an embedded sequence. To study the ability of a network to recall the embedded sequence, we first ensure that it is in a low-firing asynchronous-irregular (AI) state. Then, all the neurons of the first assembly in the sequence are made to fire simultaneously. If the connectivity (feedforward across assemblies and/or recurrent within assemblies) is strong enough, an activity pulse successfully propagates throughout the entire sequence. Otherwise, the pulse dies out and the sequence cannot be retrieved. We consider an assembly to be 'fully activated' if a propagating activity pulse makes at least 90% of its neurons fire once. Accordingly, if an activity pulse generated at the first assembly in a sequence can successfully propagate through all of its assemblies, such that each of them becomes 'fully activated' in sequence order, down to the very last one, we say that the sequence has been successfully 'recalled'/'retrieved'/'replayed' (these terms are used interchangeably). If the activity pulse dies out before reaching the last assembly in the sequence, we say replay has failed. Moreover, neurons should not fire more than once, on average. Such cases, with bursting or exploding activity that can extend to the whole network, are considered failed events. When retrieval is successful, we measure the asymptotic pulse width and propagation speed. The width is obtained by filtering an assembly's activity and fitting it to a Gaussian to calculate its full width at half maximum (FWHM), which is averaged over the last three assemblies. The speed is calculated as the inverse of the time between filtered activity peaks of consecutive assemblies, averaged over the last four assemblies (for details, see 'Materials and methods').

Fig 1 summarizes the results obtained for each of the three spiking networks studied. In each of them, we embed a sequence of 10 assemblies. This number is chosen because it is short enough to allow for quick simulation times, but long enough to approximate the asymptotic pulse behavior in very long sequences. We embedded only one sequence in a network and assume having several different sequences does not change the results on replay if the number of sequences is well below the capacity of the network; memory capacity or competition between different sequences [39] is not the focus of this manuscript. The first network (Fig 1–A1) is similar to that of [34], with each assembly made up of both excitatory and inhibitory cells. In the second network, we remove the inhibitory cells from the assemblies and use them only to provide global inhibition to the excitatory population (Fig 1–A2). Finally, in the third network, inhibition is completely absent (Fig 1–A3). Below, the three models and simulation results are described in more detail (see also 'Materials and methods' and Table 1 for a full list of network parameters).

Excitatory-Inhibitory assemblies

Our first spiking model (Fig 1–A1) consists of N_E excitatory and $N_I = N_E/4$ inhibitory cells, all forming background connections to each other with low probability p_{bg}. Then, a unidirectional sequence of $q = 10$ non-overlapping EI assemblies is embedded in the network. Each assembly consists of M excitatory and $M/4$ inhibitory cells that form additional recurrent connections to each other with some probability

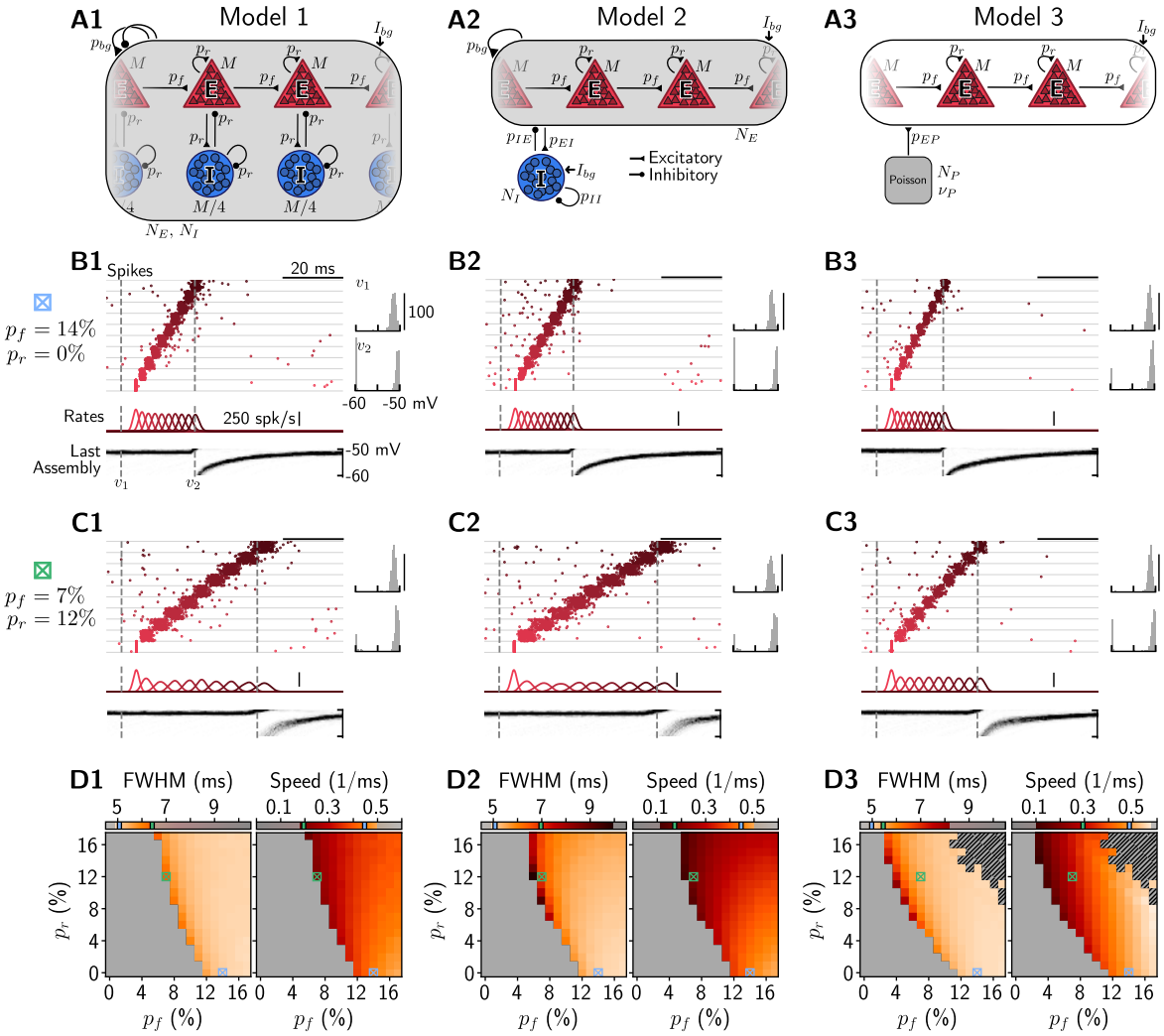

Fig 1. Sequential activation of neuronal assemblies in spiking networks. A1–A3, Network sketches of simulated spiking models. **A1**, Assemblies of M excitatory (E) and $M/4$ inhibitory (I) cells form a sequence embedded in a network with a total of N_E and N_I cells. Assembly cells form recurrent connections (probability p_r). The sequence of $q = 10$ assemblies is created by feedforward connections across the E cells of each assembly (probability p_f). All neurons can form background connections (low probability p_{bg}) and receive constant background current I_{bg}. The network is balanced by STDP in the $I \to E$ synapses until it reaches a low-firing AI state. **A2**, Similar to A1, but assemblies contain only E cells, with inhibition being global. The same STDP mechanism is used. **A3**, Similar to A2, but without background connections ($p_{bg} = 0$) nor I cells. E cells receive Poisson inputs, tuned to replicate the AI state of the first two models. **B1–B3**, Simulations with $p_f = 14\%$ and $p_r = 0\%$. Networks are in an AI state when all the neurons in the first assembly are made to fire instantly. An activity pulse propagates throughout the sequence. 'Spikes', raster plot of sampled E cells from each assembly (separated by gray lines), in sequence order from bottom to top. 'Rates', assembly population rates (Gaussian-filtered with width 1 ms). 'Last assembly', membrane potential distribution of the last assembly (snapshots on the top right: v_1 and v_2). **C1–C3**, Same as B1–B3, but for $p_f = 7\%$ and $p_r = 12\%$. **D1–D3**, Asymptotic pulse 'FWHM' and 'Speed' of successful retrievals (unsuccessful ones shown in gray). Color bars have the same scale but only the measured ranges are colored in each panel. Crossed blue and green squares correspond to the parameters used in panels B and C, respectively, with lines in color bars indicating the measured values for each square. The black-hatched squares in the upper-right corner of D3 represent failed retrievals due to activity explosions.

p_r. To make up the sequence, additional feedforward connections across the M excitatory cells in each assembly are created with some probability p_f. All synapses are modeled as an instantaneous jump in synaptic conductance that decays exponentially. Furthermore, the model uses a spike-timing-dependent plasticity (STDP) mechanism in the $I \to E$ synapses [40], which balances the network until it reaches an AI state. Once the network is balanced, plasticity is turned off. Then, to investigate replay, we make all the neurons in the first assembly fire at once. If the generated activity pulse successfully propagates throughout the sequence, such that all sequence neurons fire approximately once in the correct order, the trial is considered a success. To study the role of recurrent connections in facilitating recall, this is attempted for different combinations of p_f and p_r.

The example simulation in Fig 1–B1 indicates that recurrent connections are not always needed for replay. As shown in the summary figures in Fig 1–D1, p_f must be high enough to retrieve the sequence, for any given p_r. Moreover, recurrent connections enable the retrieval of sequences with lower p_f, as in the example simulation in Fig 1–C1. The lower the p_f, the higher the p_r required, as in [34]. Both simulations in Fig 1–B1,C1 show how the membrane potential distribution of the last assembly in the sequence behaves during the propagation of the activity pulse (Bottom and Right panels). We see that, in the AI state, the subthreshold membrane potential distribution of each assembly approximates a normal distribution, with the upper tail of the Gaussian clipped at the threshold and driving the fluctuation-driven low firing rate [as predicted by existing theoretical models, 41, 42]. When the propagating activity pulse reaches an assembly, this distribution is driven across the threshold, until all neurons have fired. After firing, the membrane potential of each neuron is reset and the distribution again slowly settles in the stationary AI state it had before the pulse disturbed it. In the simulations in Fig 1–B1,C1, all neurons in the assemblies fire approximately once during the replay of the sequence; thus, moving the membrane-potential distribution across the threshold, without explicitly considering the reset and the refractoriness of individual neurons, is an appropriate simplification to understand the fast dynamics during replay. The key difference in the behavior of the membrane potential distributions, as p_r or p_f are varied, is how long the distribution takes to cross the threshold. If p_f is high enough such that p_r is not needed (as in Fig 1–B1), the distribution is fast to cross the threshold, making all the neurons fire within a small time interval. As p_f decreases below a critical value, the membrane potential distribution will take longer to cross the threshold (as in Fig 1–C1), because the feedforward synapses alone are no longer sufficient to make all the neurons fire. Thus, recurrent excitatory synapses within an assembly need to cooperate with the feedforward synapses to ensure all neurons fire, which means more time is needed for those synaptic inputs to arrive. The smaller the p_f, the higher p_r must be for recall to succeed, and the slower those events will be.

Accordingly, Fig 1–D1 demonstrates that, while all activity pulses start out as a perfectly synchronous activation in the first assembly generated by external stimulation, the asymptotic width and speed of the pulse depend significantly on the chosen connectivity. In the figure, we see that a large width is highly correlated with a slow speed, while a narrow width correlates with a fast speed (see also Fig S2.1). The FWHM varies between about 5 and 7 ms. It decreases with increasing p_f and/or p_r, i.e., the higher the connectivity, the narrower the propagating pulse is (see 'FWHM' in Fig 1–D1). Panel 'Speed' in Fig 1–D1 shows a minimum pulse speed of $\sim$6 ms per assembly ($\sim$0.18 assemblies/ms) and a maximum pulse speed of 2 ms per assembly (0.50 assemblies/ms). Our maximum speed is roughly consistent with the used excitatory synaptic model, where synaptic conductances had a latency of 1 ms and exponential decay time constant of 2 ms. The pulse speed significantly increases with increasing feedforward connectivity p_f, on which it seems to depend the most. However, there is a more complex dependence on p_r. In the region where p_r is not needed for propagation ($p_f \geq 12\%$), increasing p_r significantly slows down the propagation speed. On the other hand, in the region where p_r is needed ($p_f < 12\%$), increasing p_r with respect to its minimal value slightly increases pulse speed, although this effect appears to fade away as one continues to increase p_r. To investigate whether feedback inhibition and balanced amplification are essential mechanisms at play here, we look at two simpler spiking network models in Fig 1.

Excitatory-assemblies balanced by global inhibition

The second spiking network model we simulate (Fig 1–A2) is as similar as possible to the first one, with the exception that assemblies now contain only excitatory cells. Unlike the first model, the second one does not have specific assemblies of interneurons targeting each excitatory assembly through a higher morphological connection probability. Instead, we have a homogeneous population of interneurons that has the same probability of receiving/sending synapses from/to any excitatory cell. A similar replay

model has been studied in [33]. The simulations in Fig 1–B2:D2 show that the obtained results are similar to those of the first model. Higher recurrent connectivity again allows for the retrieval of sequences with lower feedforward connectivity, and lower connectivity corresponds to slower and wider activity pulses: The observed pulse speed varies in a similar range as the first model, although it can reach lower values ($\sim$0.1 assemblies/ms). The minimal pulse width is also similar to the first model, but the maximum width can reach up to $\sim$10 ms. These slower and wider pulses likely occur due to each assembly receiving less feedback inhibition during replay when compared to the first model.

Assembly membrane potential distributions are similar to the first model both in the AI state and during retrieval. This is due to the use of the same spike-timing-dependent plasticity (STDP) mechanism in all $I \to E$ connections present in the first model [40]. This mechanism allows for the inhibitory neurons to change their synaptic weights and target each neuron differently, which could give rise to a selection of neurons based on the random $I \to E$ connectivity, thus somewhat mirroring the EI-assembly structure that was hard-wired in the first model. However, $E \to I$ and $I \to I$ connections, which do not possess any plasticity mechanism, are non-specific and constant in the second model, which could cause the differences observed in pulse width and disrupt the EI coupling needed for balanced amplification to occur. In contrast, in the first model, a change in p_r led to a change in all EI connections within an assembly.

Excitatory-assemblies only: a minimal spiking model

The third spiking model we simulate (Fig 1–A3) is the simplest that can exhibit replay of assembly sequences. It consists merely of the sequence of excitatory assemblies and has neither inhibitory neurons nor background connections ($p_{bg} = 0$). Thus, neurons can only connect to those in the same assembly with probability p_r, or to those in the next assembly with probability p_f. The existence of recurrent connections distinguishes this minimal model from the 'synfire chain' architecture [7], where only feedforward connections across assemblies are considered. To keep the network in a low-firing AI state similar to the first two models, the constant background current I_{bg} was decreased and noisy excitatory synaptic inputs generated by Poisson processes were added [8]. I_{bg} and the Poisson inputs were tuned such that the membrane potential distribution in the AI state was as similar as possible to the first two models.

The example simulations in Fig 1–B3,C3 show that successful replay events are similar to those of the first two models, and Fig 1–D3 shows that weak feedforward connections can again be compensated by high recurrent connectivity to allow for replay. The pulse width and speed of successful replay events are in a similar range to the first two models. However, a main difference is that retrieval fails in the regions where the total connectivity ($p_r + p_f$) is too high. The failed replay events in the upper-right corner (black and grey hatch in the two panels of Fig 1–D3) correspond to bursting/exploding activity, which occurs because there is no feedback inhibition to prevent this from happening.

Main features of spiking simulations

Our simulations show that recurrent connections enable the sequential activation of assemblies even after completely removing inhibitory feedback from these networks. These results suggest that inhibition and balanced amplification do not determine whether pulse propagation can be facilitated by intra-assembly recurrent connections. The asymptotic speed and width of the activity pulse depend significantly on p_r and p_f, with low values corresponding to slower and wider pulses. Large (narrow) width correlates with slow (fast) pulse speeds (see also Fig S2.1). The minimum values for pulse width are roughly the same across the three models, with the narrowest pulses having FWHM $\sim$5 ms. A similar maximum speed is also observed across all models ($\sim$0.5–0.6 assemblies/ms). These results are unsurprising since all three models used the same $E \to E$ synaptic dynamics. Interestingly, in all three models, the minimal speed is also roughly the same, with pulses slower than $\sim$0.1–0.2 assemblies/ms failing to propagate. These cases correspond to the widest pulses, although the maximum width does seem to be strongly influenced by inhibition, varying between $\sim$6–10 ms across models. Investigating how inhibition modulates these slower pulses is beyond the scope of our study. Without inhibition (third model), replay can only succeed if an assembly can go from the AI steady state to fully crossing the firing threshold in less than $\sim$10 ms, if we assume a normal distribution of firing times and the 'full width' as being approximately 4 SDs. A replay with wider pulses than this is probably prevented by the leak currents, which start having a substantial hyperpolarizing effect if the pulse width is too close to the membrane time constant, which in

all simulations was 20 ms.

Although inhibition can modulate the properties of propagating pulses, our simulations clearly show it is not needed for recurrently driven amplification to occur. Thus, the excitatory synapses within the sequence must be enough to achieve this. For a successful retrieval, connectivity must be strong enough to drive most sequence neurons across the firing threshold. This means that the subthreshold membrane potential distribution in the AI state (preserved in all three spiking models) must be essential in determining this. Then, how do excitatory recurrent connections cooperate with feedforward synapses to allow for the retrieval of sequences, and how does retrieval depend on the properties of the membrane potential distribution? Moreover, how does the interplay between recurrent and feedforward connections determine the pulse speed? In what follows, we develop a population model that utilizes membrane-potential distributions of assemblies to address these questions.

A time-discrete population model of sequence retrieval

In this section, we model the membrane potential distributions of neuronal assemblies during a replay event. To do this, we make several simplifying assumptions, which are detailed below.

Model assumptions

In our spiking networks, leak currents are essential in determining the subthreshold membrane potential distribution in the AI state. However, during successful retrieval events, these currents do not seem to have a substantial effect on the distribution. This can motivate a separation of time scales: the slow time scale of the leak currents, given by the membrane time constant (20 ms), and the fast time scale of sequence retrieval, governed by the time that excitatory synapses take to increase a neuron's membrane potential (1–5 ms). Thus, our first assumption is that we neglect the effect of leak currents during replay events. This assumption has two consequences for our population model. The first is that during retrieval the membrane potential distribution keeps its shape (the same shape as in the AI state), with only its mean changing as synaptic inputs arrive (this approximates the spiking network behavior seen in the histograms of Fig 1–B,C). The second consequence is that the mean of the membrane potential distribution can never decrease — it can only stay constant or increase.

In our spiking simulations, a condition for successful replay is that each neuron only fires once during an event, on average, with bursting and exploding activity being considered failures to retrieve the sequence. In a successful replay event, after a neuron has fired, its membrane potential is set to the reset potential, and the neuron does not make any additional contribution (as seen in the histograms in Fig 1–B,C). Thus, we neglect the activity of neurons in an assembly once their membrane potentials have crossed the firing threshold. This assumption means that for evaluating successful retrieval in our population model we only have to keep track of whether the distribution (which starts in the AI state) receives enough synaptic inputs to fully bring it across the threshold once.

In spiking networks, synaptic input during replay is governed by recurrent connections within assemblies and feedforward connections across assemblies, but there may also be background connections (as in the first two spiking networks). In this case, the activity of an assembly can excite many other neurons in the network. However, this effect is negligible in our spiking models because the number of neurons in an assembly is small compared to the network size and the background connectivity is low (background connections are even absent in the third spiking model). Thus, we neglect background connections in the population model.

Given the fast time scale at which excitatory synapses act on the sequence, the synaptic latency has a major impact on the speed and dynamics of replay. Our spiking simulations used a conductance-based synaptic model for all excitatory synapses. The conductance increases by a constant amount in the postsynaptic neuron after a delay of $\tau_l = 1$ ms and decays exponentially with a time constant of 2 ms. However, this specific choice is not necessary for sequence recall to occur. In 'S1 Appendix', we show that a minimal model with delta-current synapses that increase the membrane potential of the postsynaptic neuron by a constant amount w^{EE} after some delay can recall an embedded sequence showing the same features described above. This motivates our final assumption: the movement of the membrane potential distribution across the threshold during retrieval is approximated by a time-discrete model, where the time step corresponds to the delay between a presynaptic spike and a postsynaptic potential.

Model definition

Given these assumptions, we model the distribution of membrane potentials v of a given assembly $i \in \{1, 2, ..., q\}$ in a sequence of length q at some time step $t \in \{0, 1, 2, ...\}$ with a static-shaped probability density function $g_t^{(i)}(v)$. The static shape has finite width U and is described by a normalized function $h_U(v)$, where $h_U(v) \geq 0$ for $-U \leq v \leq 0$, and $h_U(v) = 0$ otherwise. If we denote $x_t^{(i)}$ as the position of the right edge of the distribution $g_t^{(i)}$, we can write

$$g_t^{(i)}(v) = h_U\left(v - x_t^{(i)}\right) . \tag{1}$$

We define the threshold to be located at the potential $v = 0$ and consider an assembly is active if the membrane potential distribution crosses the firing threshold. Since $x_t^{(i)}$ can never decrease with increasing t, the firing activity $a_t^{(i)}$ for some assembly i at some time t is the area under the distribution that crossed the firing threshold between the previous time $t-1$ and the current time t,

$$a_t^{(i)} \quad = \quad \int_0^{+\infty} \left[g_t^{(i)}(v) - g_{t-1}^{(i)}(v)\right] \mathrm{d}v \,, \quad \text{for} \quad t \geq 1 \,. \tag{2}$$

Because the function h_U is normalized, the total activity $\sum_k^t a_k^{(i)}$ is bounded between 0 and 1 and corresponds to the fraction of neurons in assembly i that crossed the threshold up to some time t.

In our spiking network models, the initial condition is that all sequence neurons are in a low-firing AI state, which is stable and does not cause sequential activation unless an external input is given to an assembly. In the population model, this corresponds to all assemblies starting at $t = 0$ with subthreshold membrane potential distributions at the same location and with zero activity, i.e.,

$$x_0^{(i)} = x_0 \leq 0 \,, \quad a_0^{(i)} = 0 \,, \quad \forall i \,. \tag{3}$$

For $t \geq 1$, in a given time step from $t-1$ to t, the right edge $x_t^{(i)}$ of the membrane potential distribution of assembly i increases with respect to its previous position $x_{t-1}^{(i)}$ by some amount proportional to the input to that assembly. Said input can either come from recurrent synapses from the assembly itself with a weight denoted by R, feedforward synapses from the previous assembly with weight F, or some external input $J_t^{(i)}$, which we assume is equally applied to the whole assembly i. These dynamics can be summarized as

$$x_t^{(i)} = x_{t-1}^{(i)} + R \cdot a_{t-1}^{(i)} + F \cdot a_{t-1}^{(i-1)} + J_t^{(i)} \ . \tag{4}$$

The first assembly in the sequence is $i = 1$, so, to keep Eq (4) consistent for the whole sequence, we define $a_t^{(0)} = 0$ for all t, and thus assembly 1 never receives feedforward input. Eq (4) can also be written in the explicit (non-recurrent) form

$$x_t^{(i)} = x_0 + R \cdot \sum_{j=0}^{t-1} a_j^{(i)} + F \cdot \sum_{k=0}^{t-1} a_k^{(i-1)} + \sum_{l=1}^{t} J_l^{(i)} \,, \tag{5}$$

where we have used our initial conditions (3).

In our spiking simulations, we retrieved sequences by stimulating the first assembly. To best match this protocol in the population model, we consider external inputs that act only on the first assembly at $t = 1$, i.e.

$$J_t^{(i)} = \begin{cases} J \geq 0 & , \quad \text{for } i = 1 \,, t = 1 \\ 0 & , \quad \text{otherwise} \,. \end{cases} \tag{6}$$

In general, we assume $0 \leq J \leq U - x_0$.

In our time-discrete population model, we will describe the membrane-potential distribution by two generic shapes: a clipped Gaussian and a rectangle. For the clipped-Gaussian model, we approximate the shape of the distribution of membrane potentials h_U by a Gaussian with standard deviation $\sigma = U/4$ and mean $\mu = -U/2$. The function h_U is then clipped at $\pm 2\sigma$ around the mean and normalized such that the integral is 1. A major aim of the population model is to obtain an analytical description of the

dynamics of the propagating activity pulse. Because the clipped-Gaussian distribution is too complex, we also approximate the shape of the distribution h_U by the simplest possible form it could take, i.e., a rectangle with finite width U and area 1, such that

$$h_U(v) = \begin{cases} 1/U & , \quad \text{for } -U \leq v \leq 0 \\ 0 & , \quad \text{otherwise.} \end{cases} \tag{7}$$

We shall refer to this particular case of our population model as the "rectangle model". In Fig 2, we sketch the behavior of both distributions with initial conditions (3) for $x_0 = 0$ and for $J > 0$.

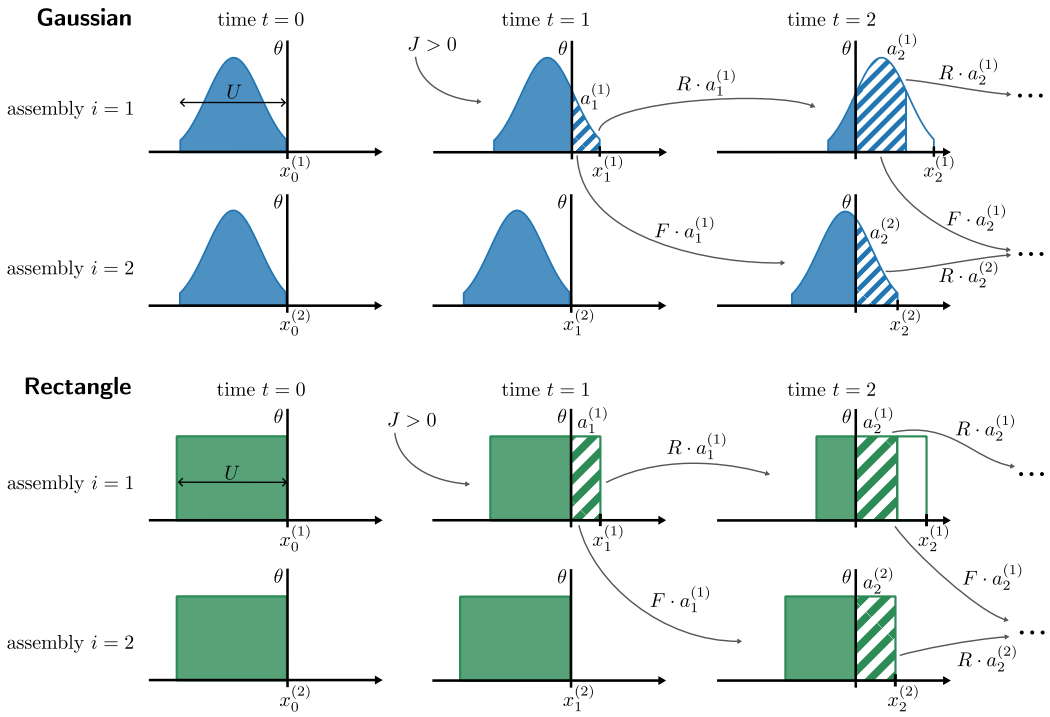

Fig 2. Behavior of the time-discrete population model. Sketch of the model's progression for the first two assemblies in a sequence at times $t = 0$, $t = 1$, and $t = 2$ for a clipped Gaussian (Top, blue) and rectangle (Bottom, green) distributions. Both assemblies start with initial condition (3) with $x_0 = 0$. At $t = 1$, the first assembly receives some positive external input $J > 0$, generating some activity $a_1^{(1)}$. That activity then propagates in the next time step $t = 2$ to that same assembly via recurrent connections (R) and to the next assembly via feedforward connections (F).

Quantification of model performance and numerical simulations

We consider that an assembly is successfully recalled if the distribution g fully crosses the threshold, and we define $\bar{t}_i$ as the minimal time at which this happens for assembly i, such that $x_t^{(i)} \geq U$ for any $t \geq \bar{t}_i$. For a sequence with q assemblies, we consider that the whole sequence has been successfully recalled if the membrane potential distribution of the q-th assembly fully crosses the threshold, i.e., if there exists some finite time $\bar{t}_q$ for which

$$x_{\bar{t}_q}^{(q)} \geq U \,. \tag{8}$$

We note that this definition of the successful recall of a whole sequence does not imply that all assemblies in the sequence are successfully recalled, which depends on initial conditions. However, we will show that this definition is useful, especially in the limit of large q.

For successful sequence retrieval, we calculate the pulse propagation speed and width. Analogously to the spiking models, in the population model we consider the pulse width to be the FWHM in time of the activities generated by a given assembly i:

$$2\sqrt{2\log 2} \cdot \sqrt{\sum_k (k - \langle t \rangle_i)^2 \cdot a_k^{(i)}} \tag{9}$$

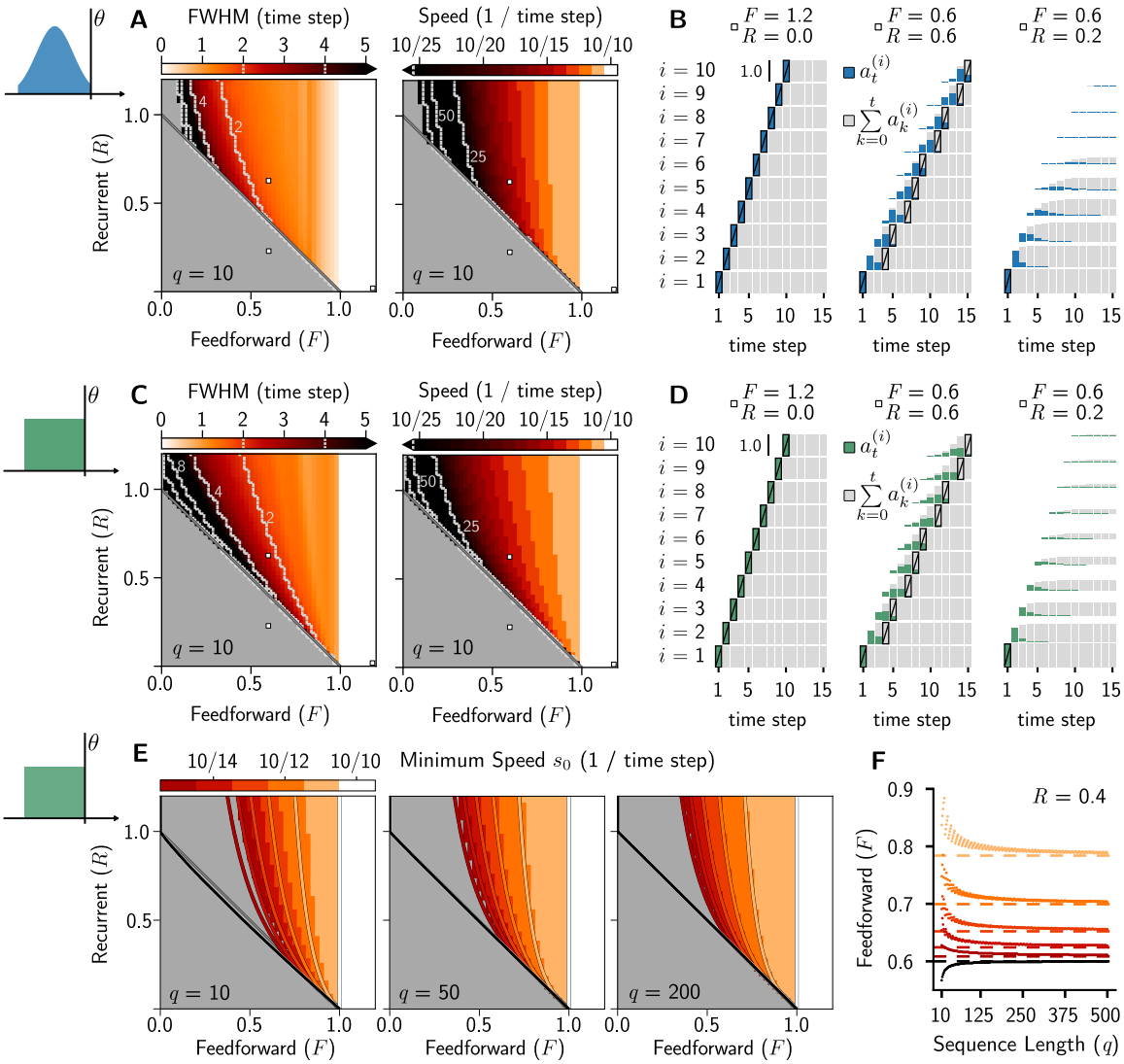

Fig 3. Population model simulations confirm analytical predictions and match spiking network behavior. **A**, Clipped-Gaussian model with $U = 4\sigma = 1$, and $x_0 = 0$. Simulated with external input $J = U$ in (6). Retrieval succeeds if the last assembly in a sequence of $q = 10$ assemblies fully crosses the threshold at some time $\bar{t}_q$. Color maps: Asymptotic pulse width (FWHM in time steps) and approximate pulse speed ($q/\bar{t}_q$, i.e. retrieved assemblies per time step) — failed replays in gray — for different values of R and F. Black regions show widths larger than 5 time steps and speeds slower than 10/25. The dashed white lines denote the upper borders for widths 16, 8, 4, and 2 (from left to right) and lower borders for speeds 10/100, 10/50, and 10/25. White squares correspond to the simulations on panel B. **B**, Example simulations; assemblies are stacked in sequence order from bottom to top. Colored bars represent activity $a_t^{(i)}$. Gray bars represent the sum of activities for that assembly. A black contour with a slash marks the time step at which an assembly becomes fully activated. **C–D**, Same as A–B, but for the rectangle model. **E**, Speed in rectangle model for different assembly sizes q. For $q = 10$, same panel as in B but the speed is capped at the minimum of 10/15 assemblies per time step. Overlain are colored lines, each corresponding to the R–F values that match the analytical Condition 1 for a given minimum pulse speed s_0. The black line corresponds to this analytical prediction when $s_0 \to 0$. The $R + F = 1$ line is drawn in gray. The same is then shown for $q = 50$ and $q = 200$, with the speeds rounded up to match the same color levels as in $q = 10$. **F**, F-values of the analytical curves drawn in panel E when $R = 0.4$ (dots). Each dot is the minimal F needed for replay to succeed, for the corresponding minimum speed, according to Condition 1. As q increases, the needed F converges to a fixed value, well approximated by Eq (17) (dashed lines).

where $\langle t \rangle_i = \sum_k k \cdot a_k^{(i)}$ and we assume the relation between FWHM to be $2\sqrt{2\log 2}$ times the SD, as in normal distributions. To approximate an asymptotic pulse width, we average (9) over the last three assemblies. The average pulse speed is approximated by $q/\bar{t}_q$, i.e., how many assemblies per time step were retrieved. For short sequences, the initial condition (magnitude of input J) might create a distortion between $q/\bar{t}_q$ and the asymptotic pulse speed. However, as the sequence length q increases, the asymptotic pulse speed gets closer and closer to $q/\bar{t}_q$.

Fig 3 shows simulations of the time-discrete population model, both for the clipped Gaussian distribution (Fig 3–A,B) and for the rectangle distribution (Fig 3–C:E). For simplicity, we chose $U = 1$. To match the behavior of the AI state in the spiking model, where any input to an assembly leads to an increase in activity, we consider $x_0 = 0$. In the spiking network, we made all the neurons in the first assembly fire at once, so in the population model simulations we choose $J = U$ as the only input to the first assembly, such that it becomes instantly fully activated. Fig 3–A:D indicates that clipped Gaussian and rectangle distributions produce similar results, both with the same qualitative behavior as the spiking network. The asymptotic pulse width and average speed, as well as the ability to successfully replay the sequence, show a similar dependence on the strength of recurrent and feedforward connectivities; there are some minor differences between the two distributions for low feedforward connectivities. As in the spiking simulations, Fig 3–A,C shows that so too in the population model does a large width correlate with a slow pulse speed (see also Fig S2.1). To better understand these results, below we derive some analytical insights from the population model.

Partial activation may lead to full activation via recurrent and feedforward connections

In the simulations of Fig 3, we fully activated the first assembly, for a direct comparison of the pulse properties with the spiking network simulations. However, our population model allows us to make predictions on the conditions for which a partial activation of the first assembly would lead to the full activation of some sequence assembly i. To derive this, we will use the rectangle approximation:

Consider that $x_0 = 0$ such that any input to an assembly will generate some non-zero activity, as is the case in the AI state of our spiking networks. The activity of an assembly i at some time t in the rectangle model will be given by

$$a_t^{(i)} = \frac{\min\left(x_t^{(i)}, U\right) - \min\left(x_{t-1}^{(i)}, U\right)}{U} \,. \tag{10}$$

If an assembly has not yet fully crossed the threshold, the activity (10) becomes the simple linear relation

$$\hat{a}_t^{(i)} = \frac{x_t^{(i)} - x_{t-1}^{(i)}}{U} \,, \tag{11}$$

where $\hat{a}_t^{(i)} = a_t^{(i)}$ if both $x_{t-1}^{(i)}$ and $x_t^{(i)}$ are in the interval $[0, U]$. If $x_{t-1}^{(i)}$ or $x_t^{(i)}$ are outside this interval, we generally have $\hat{a}_t^{(i)} \geq a_t^{(i)}$, since the estimate given by (11) is no longer bounded between 0 and 1. Assume now that some assembly m is the first to become fully activated at some time $\bar{t}_m$. If $x_0 = 0$, then $\hat{a}_t^{(i)} = a_t^{(i)}$ is true for any $i \leq m$ and $t \leq \bar{t}_m$. Then, we can extend the linear relation (11) to all activities in (4) and obtain an equation for an estimate of the right edge of the distribution,

$$\hat{x}_t^{(i)} = \hat{x}_{t-1}^{(i)} + R \cdot \frac{\hat{x}_{t-1}^{(i)} - \hat{x}_{t-2}^{(i)}}{U} + F \cdot \frac{\hat{x}_{t-1}^{(i-1)} - \hat{x}_{t-2}^{(i-1)}}{U} \,. \tag{12}$$

Due to the linear nature of this equation, we can derive (see 'S3 Appendix' for details):

$$\hat{x}_t^{(i)} = \begin{cases} 0 & , \quad t < i \\ J \cdot (F/U)^{i-1} \sum_{k=i}^{t} \binom{k-1}{i-1} (R/U)^{k-i} & , \quad t \geq i \,. \end{cases} \tag{13}$$

Thus, m and $\bar{t}_m$ will correspond to the lowest i and t for which $\hat{x}_t^{(i)} \geq U$, i.e., for which the following condition is met:

$$(J/U) \cdot (F/U)^{i-1} \sum_{k=i}^{t} \binom{k-1}{i-1} (R/U)^{k-i} \geq 1 \qquad \text{(Condition 1)}$$

Note that if $i \leq m$ and $t \leq \bar{t}_m$, then $\hat{x}_t^{(i)} = x_t^{(i)}$ for the rectangle model, and Condition 1 is a necessary and sufficient condition for replay. Below, we will see that this linear estimate provides a good approximation of the model dynamics and can help explain the border between successful and failed replays seen in the spiking network simulations of Fig 1–D.

The estimate given by (13) and Condition 1 provides insight into the dynamics of the population model. In (13), for $t < i$, we find $\hat{x}_t^{(i)} = 0$ because the pulse propagation can never be faster than one assembly per time step. So, the right edge of the distribution of some assembly i can only start moving to the right for $t \geq i$.

At $t = i$, assembly i receives its first input, and $\hat{x}_i^{(i)} = J \cdot (F/U)^{i-1}$, which depends on the relative strength of only the feedforward connections F with respect to the distribution width U. At each assembly $i > 1$, the initial input J gets scaled by a factor of F/U, such that at assembly i this scaling is $(F/U)^{i-1}$. Because the recurrent inputs need additional time steps to arrive, they only contribute to an assembly's position for $t > i$. Thus, the dynamics of a sequence without recurrent connections (like a synfire chain) are fully described by what happens at $t = i$: If $F < U$, the initial input J will be scaled down with increasing i as the pulse propagates throughout the sequence, making replay impossible, regardless of the value of J. On the other hand, if $F > U$, the initial input is amplified as the pulse propagates, and, if the sequence is long enough, some assembly m will eventually become fully activated once $J \cdot (F/U)^{m-1} \geq U$ is true.

If there are recurrent connections, for $t > i$ the initial feedforward input gets scaled by an additional factor, which sums all the recurrent pathways by which the initial input J was scaled throughout the sequence up to that time t, which depends on the relative strength of the recurrent connections R with respect to the width U.

Recurrent connections are not needed if feedforward connections are strong enough

The numerical simulations in Fig 3–A:D (where $x_0 = 0$ and $U = 1$) confirm that recall always succeeds if $F \geq 1$, with those pulses propagating with minimal width at the maximum speed of one assembly per time step, and recurrent connections not being needed. This result can be shown analytically for any distribution, regardless of its shape h_U:

If a given assembly $m \geq 1$ becomes fully activated at some time $\bar{t}_m$ then we have $\sum_{k=0}^{\bar{t}_m} a_k^{(m)} = 1$. According to (5), the right edge of the next assembly $m+1$ at any time $t > \bar{t}_m$ will be given by

$$x_t^{(m+1)} = x_0 + R \cdot \sum_{j=0}^{t-1} a_j^{(m+1)} + F\,, \quad \text{for} \quad t > \bar{t}_m\,. \tag{14}$$

If $F \geq U - x_0$, we have $x_t^{(m+1)} \geq U$ already at $t = \bar{t}_m + 1$, i.e., assembly $m+1$ will instantly become fully activated within one time step, regardless of the value of R — recurrent connections are not needed. Consequently, the next assembly $m+2$ will become fully activated at time $\bar{t}_m + 2$. Thus, we can infer that, if $F \geq U - x_0$ ($F \geq 1$ in our numerical simulations, where $x_0 = 0$ and $U = 1$), the full activation of an assembly will lead to the full activation of all assemblies downstream from it, with minimal width and a propagation speed of one time-step per assembly, regardless of the distribution shape h_U.

Necessary and sufficient conditions for sequence retrieval

The simulations in Fig 3–A:D further show that the propagation of the pulse slows down as F decreases. But recall can only succeed if $R + F > 1$. The closer $R + F$ is to 1, the slower the replay is. These results can be shown analytically.

Without assuming any particular distribution shape h_U, we know that the sum of activities for any assembly is always smaller than or equal to 1 (since it can never be larger than the total area under the normalized distribution). Thus, $\sum_{k=0}^{\infty} a_k^{(i)} \leq 1$ for any i. We can then derive from Equations 5 and 8 that

$$R + F \geq U - x_0 \tag{Condition 2}$$

is a necessary condition for any assembly that does not receive external inputs (i.e., $i > 1$ in our model) to be retrieved. For $x_0 = 0$ and $U = 1$, as in our numerical simulations, the condition becomes $R + F \geq 1$.

Although this condition is necessary for both the clipped Gaussian and rectangle distributions, it is a sufficient condition only for the rectangle (see failed regions of replay in the upper-left corner of

the 'FWHM' and 'Speed' plots for the clipped Gaussian in Fig 3–A). The fact that this condition is both necessary and sufficient for the rectangle model, given enough simulation time steps, can also be shown analytically. Assuming a rectangle distribution with $x_0 = 0$, and some assembly m becoming fully activated at some time $\bar{t}_m$ such that (14) holds, we can derive the dynamics of assembly $m+1$ for some time before it is fully activated (see 'S3 Appendix' for details), i.e. for some time $\bar{t}_m < t \leq \bar{t}_{m+1}$:

$$x_t^{(m+1)} = F + (R/U) \cdot x_{t-1}^{(m+1)}\,, \quad \text{while} \quad x_t^{(m+1)} \leq U\,. \tag{15}$$

If $R \geq U$, the distribution of assembly $m+1$ will move across the threshold at a constant or increasing rate. If, however, $R < U$, the right edge of assembly $m+1$ will asymptotically tend to $F/(1-R/U)$ (see 'S3 Appendix' for details). Thus, if $F/(1-R/U) < U$, i.e. $R+F < U$, the assembly never becomes fully activated (as we already knew, since it violates the necessary Condition 2). If $R+F = U$, it would only become fully activated given an infinite number of time steps, which is not physically possible. Finally, if $R+F > U$, we know the assembly will eventually fully cross in a finite number of time steps (note that once $x_t^{(m+1)} > U$ the dynamics of Eq (15) no longer apply, but we can nevertheless know that at some point this will happen). Thus, the full activation of an assembly will lead to the sequential activation of the entire sequence if $R+F > U$, given enough time steps. The closer $R+F$ is to U, the more time steps are needed.

Relation between pulse speed and connectivity. The pulse speed s is approximately $q/\bar{t}_q$, where q is the sequence length and $\bar{t}_q$ is the time at which the last assembly becomes fully activated. To find how connectivity determines the pulse speed, we will use Eq (13), which is the linear estimate obtained from the rectangle model. This estimate tells us that replay can succeed with some speed s if Condition 1 is true when $i = q$ and $t = q/s$, i.e.,

$$(J/U) \cdot (F/U)^{q-1} \sum_{k=q}^{q/s} \binom{k-1}{q-1} (R/U)^{k-q} \geq 1\,. \tag{16}$$

Note that $s \in (0,1]$, with $s = 1$ when $\bar{t}_q = q$ and $s \to 0$ as $t_q \to \infty$. Importantly, the left-hand side of (16) is monotonously increasing as s decreases, such that if the condition is met for a certain speed s', it will also be met for all speeds $s < s'$. Thus, a pulse speed estimate $\hat{s}$ can be obtained by fixing R and F and finding the largest s for which (16) is true. If this condition is not met even for $s \to 0$, then replay is impossible.

In the limit when s is fixed and $q \to \infty$, it can be shown (see 'S3 Appendix') that Eq (16) becomes

$$F/U \geq \begin{cases} s \cdot \left(\frac{1-s}{R/U}\right)^{1/s-1} & , \quad \text{for} \quad R/U \geq 1-s \\ 1 - R/U & , \quad \text{for} \quad R/U < 1-s \end{cases}. \tag{17}$$

Remarkably, this condition is independent of the initial input $J > 0$ — the initial condition is irrelevant for $q \to \infty$. Although it is not trivial to see from (17), this condition has the same monotonous dependence on s as (16). Thus, Eq (17) can allow us to estimate the propagation speed $\hat{s}$ in dependence on connectivities R and F by finding the largest s for which the condition is true. Again, if this condition is not met even for $s \to 0$, then replay is impossible.

Besides allowing us to estimate the pulse speed for a given network connectivity, Equations 16–17 can be used to determine the necessary connectivity for replay to succeed, given a minimum desired propagation speed. This is explained in what follows.

Necessary connectivity given a minimum speed. Our spiking simulations in Fig 1 showed that too slow and too wide activity pulses do not propagate, presumably due to the effect of membrane leakiness. On the other hand, the dynamics of the population model was derived under the assumption that replay is fast enough so that leakiness can be neglected; thus, there are cases where extremely slow and wide pulses can successfully propagate (black regions in Fig 3–A,C). Therefore, the pulse propagation in the population model can match the spiking network only above some minimum pulse speed s_0. Note that speed and width are highly correlated (see Fig 1–D and Fig 3; see also Fig S2.1).

In Fig 3–E, we draw semi-numerically obtained curves that mark the border between success and failure of Condition 1. For a given sequence length q, we pick $t \geq q$ to set a minimum speed $s_0 = q/t$.

Then, for each F, we pick the lowest R that meets Condition 1 — this can be obtained by setting $s = s_0$ in (16). The resulting curve marks the needed connectivity for replay success for that minimum speed.

We first draw curves for $q = 10$ (Left panel) and vary the minimal speed from its fastest value $s_0 = 10/10$ (white vertical line) to its slowest $s_0 \sim 10/\infty$ (black curve), and 5 finite values in between (colored curves). For comparison with the numerical simulations, we chose $J = U = 1$. In the next two panels in Fig 3–E, we see what happens to this profile in longer sequences with $q = 50$ (Middle) and $q = 200$ (Right). Although the curves given by the border of (16) depend on q, we see that, as q increases and s_0 is kept fixed, the lines converge to a fixed location in the R–F plane, which is approximately given by (17). To confirm this, in Fig 3–F, we fix $R = 0.4$ and plot the needed feedforward connectivity F that meets (16), while increasing the sequence length q. We see that, for the minimum speeds considered in panel E, the necessary connectivity (dots) converges to a fixed value, which is well approximated by (17) (dashed line). Notice that in the limit when $s_0 \to 0$ (black line), both Eq (16) and Eq (17) become $R + F \geq U$, which is the bound given by Condition 2.

Comparing the linear estimate to rectangle-model simulations. In Fig 3–E, we also compare how well the rectangle model simulations match the analytical predictions given by Condition 1 for a minimum pulse speed. Our analytical curves slightly underestimate the needed connectivity for retrieval. An underestimation of the needed connectivity was expected, since our analytics were obtained from the linear estimate $\hat{x}_t^{(i)}$ of the location of the assembly, which always exceeds the true location $x_t^{(i)}$: Since $\hat{x}_t^{(i)} \geq x_t^{(i)}$, the connectivity needed to fully activate assembly i, i.e. $\hat{x}_t^{(i)} \geq U$, will be equal to or less than the connectivity needed for $x_t^{(i)} \geq U$.

We further see that the analytical condition becomes a better and better approximation of the model simulations as the sequence gets longer. We saw above that the linear estimate $\hat{x}_t^{(i)}$ is exact for $i \leq m$ and $t \leq \bar{t}_m$ if m is the first assembly to become fully activated, and this happens at time $\bar{t}_m$. This is never the case in our simulations of Fig 3, where we start the pulse by fully activating the first assembly and determine successful replay if there is a time for which $x_t^{(q)} \geq U$. For successful replays, all assemblies in the sequence get fully activated, so all assemblies $i < q$ can contribute to the error of the linear estimate $\hat{x}_t^{(q)}$. However, for failed replays, none of the other assemblies ($i > 1$) get fully activated. In those cases, the only error in $\hat{x}_t^{(q)}$ is caused by the first assembly. Thus, as the sequence length q increases, this error gets smaller and smaller. Since the analytical curves in Fig 3–E estimate the border between failed and successful replays, they become an exact match with the rectangle model simulations in the limit when $q \to \infty$.

A minimum feedforward connectivity is needed if a population is fully subthreshold

In the population model simulations of Fig 3 we chose $x_0 = 0$, such that any feedforward input to an assembly would lead to the generation of some non-zero activity, as happens in the AI state of our spiking simulations of Fig 1. However, we could assume without loss of generality that assemblies start with some positive distance below the threshold, such that $x_0^{(i)} < 0$. In a spiking network, that would correspond to having a distribution of membrane potentials fully subthreshold, with no spiking activity. Although this does not capture the behavior of the AI state, it might be useful to describe the activation of a fully subthreshold population via some feedforward input. Depending on how the excitatory-inhibitory connections are set in a network, it might be that an increase in activity in one assembly or sequence causes hyperpolarization of other sequences or assemblies via lateral inhibition. Moreover, it has been shown that many hippocampal pyramidal cells become hyperpolarized during sharp wave-ripples [43–45], a phenomenon known to co-occur with hippocampal replay.

For any activity to be generated at some assembly $i > 1$, the sum of feedforward inputs from the previous assembly $i - 1$ must be large enough to bring any portion of assembly i over the threshold, i.e.,

$$x_0 + F \cdot \sum_{k=0}^{t-1} a_j^{(i-1)} > 0 \ . \tag{18}$$

Since $\sum_{k=0}^{\infty} a_k^{(i-1)} \leq 1$, Eq (18) implies that

$$F > -x_0 \qquad \text{(Condition 3)}$$

is a necessary condition for any assembly that does not receive external inputs to be retrieved. If $x_0 = 0$ (as in our simulations of Fig 3), the condition becomes $F > 0$ and is trivial.

Spiking networks validate population model predictions

The time-discrete population model introduced in this paper allowed us to make several predictions about the ability of spiking networks to replay an embedded sequence of assemblies. These predictions were summarized in the analytical expressions given by Conditions 1, 2, and 3. In this section, we will compare these to spiking network simulations.

In the simulations of Fig 1, we changed the strength of recurrent and feedforward connections by varying the connection probabilities p_r and p_f, respectively. For comparison with our predictions derived from the population model, we must consider how these probabilities relate to the connectivity weights R and F. These correspond to the average connection strength between assemblies in the spiking network, which are given by the average number of inputs (the product of the number of neurons in an assembly, M, with the connections probabilities p_r or p_f) and the $E \to E$ synaptic weight w^{EE} (i.e. how much one synaptic input increases the membrane potential of the postsynaptic neuron), i.e.,

$$R = M \cdot p_r \cdot w^{EE}\,, \quad F = M \cdot p_f \cdot w^{EE}\,. \tag{19}$$

Moreover, we will need to determine the width of our membrane potential distributions U, which is a key parameter in our population model.

Slower synapses require higher connectivity

In our spiking network simulations of Fig 1, we measured the speed of replay as the number of assemblies retrieved per millisecond. To relate the spiking network speeds to our population model speeds, which are measured in assemblies per 'synaptic time step', we can express the spiking network speeds as $s/\Delta t$, where s is the speed measured in assemblies per synaptic time step and Δt is the duration of a time step, which depends on the synaptic dynamics of the spiking network. For simplicity, we will assume this to be the time distance between the firing of a presynaptic neuron and the centroid of the increase in membrane potential in the postsynaptic neuron, which in our spiking model is given by

$$\Delta t = \tau_l + \frac{1}{1/\tau_m + 1/\tau_d^E} \tag{20}$$

where τ_l is the delay between the presynaptic spike and the initial increase in synaptic conductance, τ_m is the membrane time constant, and τ_d^E the time constant at which synaptic conductance decays.

Our spiking simulations showed that, for each network, there is a minimal propagation speed for which replay can succeed, probably since too-slow activity pulses are prevented from propagating by the leakiness of the membrane. If that is the case, then the minimum allowed replay speed in a spiking network is a constant that we can write as $s_0/\Delta t$. If replay fails due to membrane leakiness, then this constant should not depend on the synaptic dynamics. Moreover, if $s_0/\Delta t$ is constant, an increase in Δt leads to a proportional increase in s_0, i.e., the longer the synaptic time step is, the larger the speed s_0 (in assemblies per time step) is. The population model showed that it is the minimum propagation speed s_0, measured in assemblies retrieved per synaptic time step, that determines the necessary recurrent and feedforward connectivity for replay to succeed. We further saw that, the higher the value of s_0 is, the higher the needed connectivity for replay to succeed is. Consequently, a prediction of our model is that increasing the duration of the synaptic time step Δt should increase the connectivity required for replay to succeed.

To test this prediction, we will simulate the spiking network with different values of τ_l, thus changing our time step Δt. In Fig 4–A, we perform simulations of our minimal spiking model (the third model in Fig 1), and vary the synaptic delay τ_l in the range 0.1–4.0 ms (in the simulations of Fig 1 we used $\tau_l = 1$ ms). In each panel, we show the regions where replay succeeds (black and colored squares) and fails (gray), for different values of connection probabilities p_r and p_f. As predicted by the population model, a shorter synaptic time step corresponds to a lower needed value of p_r and p_f for replay to succeed. Conversely, the longer the time step is, the higher the required values of p_r and p_f are.

As predicted, we also see that the slowest speed for which replay can succeed is approximately the same in all cases (∼0.06–0.11 assemblies/ms), regardless of the chosen synaptic delay τ_l. This justifies

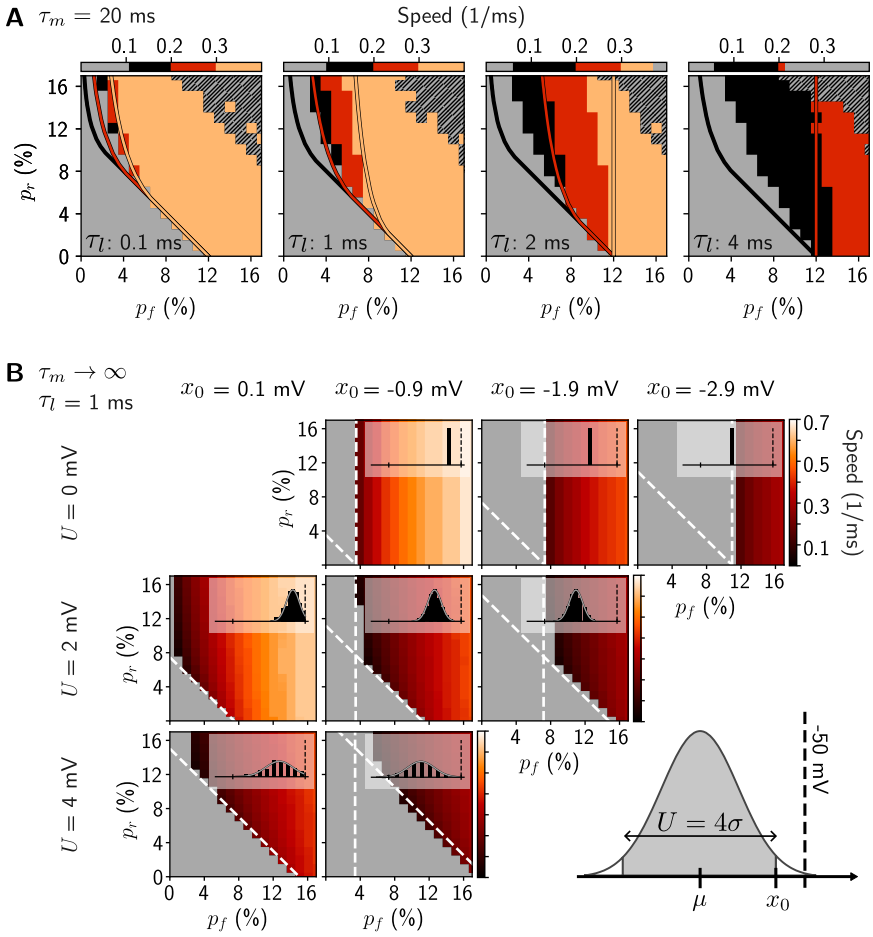

Fig 4. Spiking networks match population model predictions. A, Spiking network simulations of Fig 1–D3, but for different values of synaptic delay τ_l (indicated on the bottom of each panel). The simulations with $\tau_l = 1$ ms are the same shown in Fig 1–D3. The black and colored squares represent successful replay; failed replays are shown in gray. The yellow squares correspond to replays with propagation speed ≥ 0.3 assemblies/ms; the red squares to speeds in [0.2, 0.3) assemblies/ms; and the black squares to speeds < 0.2 assemblies/ms. In the scale bars, the left edge of the black region marks the minimum speed for each panel, which varied between ∼0.06 assemblies/ms (for $\tau_l = 4$ ms) and ∼0.11 assemblies/ms (for $\tau_l = 0.1$ ms). The maximum speed varied between ∼1.42 assemblies/ms (for $\tau_l = 0.1$ ms) and ∼0.21 assemblies/ms (for $\tau_l = 4$ ms) — marked by the left edge of the grey region on the right of the scale bar (visible on the third and fourth panels). The black-hatched squares in the upper-right corners (when $p_r + p_f$ is high) represent failed retrievals due to activity explosions. We see that the slower the synapses (i.e. the higher τ_l is), the higher the connectivity required for replay. Overlain, are curves corresponding to the analytical predictions given by (21) for three different values of speed: minimal speed measured in simulations (black curve); 0.2 assemblies/ms (red curve); and 0.3 assemblies/ms (yellow curves). **B**, Spiking simulations without leak currents ($\tau_m \to \infty$), $\tau_l = 1$ ms, and fixed number of synaptic inputs per neuron $M \cdot (p_r + p_f)$. The firing threshold is located at -50 mV and the membrane potentials for each assembly before retrieval are drawn from a normal distribution with mean μ and SD σ, resulting in distributions (see insets) with different values of U and x_0 (indicated to the left of and above each panel) — see sketch on the bottom-right to see how μ and σ relate to U and x_0. We chose $U = 4\sigma$ and obtained a good match with our analytical Condition 2 (23) (tilted dashed lines) and Condition 3 (25) (vertical dashed lines).

our assumption that the minimum speed $s_0/\Delta t$ is a constant that mostly depends on the membrane time constant and not on the duration of synapses. On the other hand, and also in accordance with our population model, changes in synaptic delay drastically change the maximum values of replay speed observed in the simulations of Fig 4–A. This maximum speed changes from ∼1.42 assemblies/ms when $\tau_l = 0.1$ ms (not visible in the scale bar) to ∼0.21 assemblies/ms when $\tau_l = 4$ ms.

Recurrent and feedforward connectivity determine pulse speed

The population model also allowed us to estimate the relation between recurrent and feedforward connectivity and the pulse speed s, obtained from our analytical Condition 1. To see how this prediction is matched by the spiking network, we will evaluate the condition

$$p_f/u \geq \begin{cases} s \cdot \left(\frac{1-s}{p_r/u}\right)^{1/s-1} & , \quad \text{for} \quad p_r/u \geq 1-s \\ 1 - p_r/u & , \quad \text{for} \quad p_r/u < 1-s \end{cases} \tag{21}$$

obtained from (17) and the definition (19). We further introduced the normalized width of the membrane-potential distribution,

$$u := \frac{U}{M \cdot w^{EE}} \,. \tag{22}$$

Eq (21) tells us that for recurrent connectivity $p_r = 0$ the minimum feedforward connectivity for replay to succeed is $p_f = u$. Since in all our simulations of Fig 1 and Fig 4–A the border between success and failure for $p_r = 0\%$ is between 11% and 12%, we set $u = 12\%$ in (21). Note that (21) relates p_f and p_r with the population model speed s, measured in assemblies per synaptic time step. Thus, to find the value of s, we multiply the spiking network speeds (given in assemblies/ms) by the synaptic time step Δt, given by (20).

In each panel of Fig 4–A, we plot a curve corresponding to the success border of Condition 1 in its asymptotic limit when $q \to \infty$, as expressed by (21), for three different values of replay speed: the minimum speed measured for each case (black); 0.2 assemblies/ms (red); and 0.3 assemblies/ms (yellow). The resulting curves allow us to compare our analytical prediction to the spiking simulations. We see that the analytical curves provide very good matches for speeds 0.2–0.3 assemblies/ms, and a worse match for the minimal speed. This discrepancy is not surprising, since our population model was derived under the assumption that leak currents can be neglected. Thus, if these leaks are causing replay to fail in the spiking network, that border will not be well captured by our theory. On the other hand, if replay is fast enough such that the leakiness can be neglected, our model seems to capture very well the relation between connectivity and pulse speed. These results validate our assumption that the spiking network behavior can be described by the time-discrete population model and that the linear estimate used to derive Condition 1 provides a good prediction of necessary conditions for assembly sequences to be replayed with a given speed.

Membrane potential distributions determine necessary connectivity for replay

Our analysis of the population model showed that retrieval should only be possible if $R + F > U - x_0$ (Condition 2). We further saw that this condition is a limit of Condition 1 when $s \to 0$. Thus, if a spiking network did not have a minimum propagation speed determined by leak currents, the border between successful and failed replays would be approximately given by

$$p_r + p_f = u - \delta \tag{23}$$

where we used Condition 2, (22) and defined

$$\delta := \frac{x_0}{M \cdot w^{EE}} \tag{24}$$

Moreover, Condition 3 tells us that there is another border between successful and failed replays, located at

$$p_f = -\delta \,. \tag{25}$$

To compare these analytical predictions to spiking networks, in Fig 4–B we simulate a simpler version of the minimal spiking model. To be able to set our minimum speed to $s = 0$, we remove the leak currents

from the network. Then, the parameter region of successful replays should be approximately delineated by (23) and (25).

We initialize the membrane potentials to follow a given normal distribution with mean μ and SD σ. We consider $U = 4\sigma$, as in the simulations of the population model with a clipped-Gaussian (Fig 3–A,B), and thus obtain $x_0 = \mu + 2\sigma - \theta$, where θ is the firing threshold, located at -50 mV (see Fig 4–B, Bottom–Right for a sketch). We use the mean μ to estimate the synaptic strength of our conductance-based synapses, such that

$$w^{EE} \approx g^{EE} \tau_d^E (V^E - \mu)/C \,, \tag{26}$$

where g^{EE} is the postsynaptic increase in conductance, V^E is the reversal potential for excitatory synapses, and C is the membrane capacitance (for model parameters, see 'Materials and methods'). Moreover, to avoid noise in the number of inputs per neuron, we fix the in-degree of the connectivity matrix such that every neuron receives exactly $M \cdot p_r$ recurrent and $M \cdot p_f$ feedforward synapses. We can then calculate how μ and σ lead to different values of u (22) and δ (24).

Our membrane potential distributions in the AI state of the spiking network simulations of Fig 1 and Fig 4–A fit well to a Gaussian with mean ~ -51 mV and SD $\sim$0.5 mV. In Fig 4–B, we vary μ in the range $[-52.9, -50.9]$ mV, and σ in the range $[0, 2]$ mV. In each panel, we choose a different combination of these parameters and show the regions where replay succeeds (colored squares) and fails (gray). We then calculate the corresponding u and δ and overlay the lines given by (23) and (25) (dashed oblique and vertical lines, respectively). We find that our choice of $U = 4\sigma$ provides a very good match between the regions of successful replay in the spiking simulations and our analytical predictions.

Although our membrane potential distributions in our spiking network simulations were approximately Gaussian (see histograms in Fig 1–B,C), our simulations and analysis of the population model suggested that the lines given by (23) and (25) should border successful replays in parameter space, regardless of the exact shape of the distribution (compare panels A and C in Fig 3). Here, this is confirmed by the results of Fig 4–B, where different Gaussian distributions were used. This underscores the validity of our population model in determining conditions for sequence replay. Overall, our models predict that the required connectivity for sequential activation is strongly determined by the width and position of the membrane potential distribution with respect to the firing threshold.

Discussion

In this work, we considered the dynamics of sequential activation of neuronal assemblies by focusing on the cooperation between recurrent synapses within an assembly and feedforward synapses across assemblies. We simulated different spiking network models capable of recalling embedded sequences and developed a time-discrete population model to gain analytical insight into how a given connectivity scheme influences sequential activation. We found that weak feedforward pulses can be amplified by intra-assembly recurrent connections to retrieve neuronal sequences without the need for feedback inhibition. Moreover, we introduced a population model that provides analytical conditions that must be met for a sequence to be replayed and allows for an estimation of replay speed from network parameters. Our findings contribute to understanding the mechanisms of hippocampal replay, as well as the propagation of activity across feedforward neuronal circuits in general.

The role of Hebbian assemblies in sequence replay

Hopfield [46] introduced a recurrent network capable of storing multiple patterns of activity by strengthening auto-associative synapses. However, the same study showed that this network was not robustly able to store and retrieve long sequences of more than a few patterns. Kleinfeld [35] and Sompolinsky and Kanter [36] later extended this to show that they could store discrete patterns of activity using Hopfield dynamics and embed them in a hetero-associative network with slower feedforward connections across patterns, thus being capable of reliably storing and retrieving sequences. Other authors have hypothesized that auto- and hetero-association might be implemented by complementary but distinct hippocampal subnetworks [47].

A challenge arises if one assumes that assemblies representing related concepts share neurons. In that case, the distinction between feedforward (across-assembly) and recurrent (within-assembly) connections becomes blurred, and a single isolated assembly is difficult to distinguish within a sequence. Many models of sequential activity, particularly those based on hippocampal place cells, focus on continuous

sequences of neurons, where distinct assemblies cannot be isolated [30, 32, 48]. In contrast, classical models of sequential activity such as the synfire chain [7] consist of purely feedforward networks, in which connections across layers (or assemblies) of neurons are considered, but intra-assembly recurrent connections are typically overlooked.

Nevertheless, the abundant evidence on the existence of engram cell assemblies that encode external representations [see 49, for a recent review] suggests that it is worth considering the importance of both recurrent and feedforward connections within a sequence. From a theoretical point of view, the most straightforward way of investigating this is to consider discrete non-overlapping assemblies [34, 39], which we have done here.

In our spiking network models, we embedded sequences by artificially creating synapses within and across assemblies, with some probability larger than the overall background connectivity. In a biological system, this learning process is thought to emerge via plastic changes in synaptic strength. Different types of plasticity rules have been found in the brain, from the classical Hebbian asymmetrical temporal window of STDP in Bi and Poo [50], to symmetrical [51] or anti-Hebbian rules [52]. More recently, [53] used temporally asymmetric Hebbian learning rules to study the encoding and retrieval of random activity sequences in recurrently connected networks. This model was extended in [54, 55] to show that heterogeneity of plasticity rules allows for the control of the retrieval speed, with symmetric rules functioning as brakes, and asymmetric rules as accelerators [55]. Other models of hippocampal replay have used symmetric [32] or asymmetric [48] STDP learning rules to encode neuronal sequences of place cells representing continuous space. In our work, we remained agnostic about the type of plasticity needed to embed our sequences and focused only on the network properties that are needed to retrieve them.

Hebbian assemblies can amplify activity pulses without inhibitory feedback

Our spiking simulations of a network with balanced EI assemblies confirm that recurrent connections play a crucial role in maintaining the propagation of activity pulses when feedforward connectivity is low, as discovered by [34]. The stronger the recurrent connectivity, the better the system is at amplifying weak signals, ensuring that activity pulses do not die out as they propagate across assemblies with weak feedforward connections. The authors of [34] hypothesized that in balanced EI networks, increasing recurrent connectivity would lead to faster and stronger pulses through a 'balanced amplification' mechanism, requiring strong feedback inhibition [38]. However, our simulations show that propagation speed has a weak dependence on recurrent connections. We further simulated a spiking network where global inhibition balances excitatory assemblies, and a minimal spiking model, made up exclusively of excitatory assemblies without any inhibition. These two networks produce a similar amplification effect of weak feedforward signals by recurrent synapses. This suggests that balanced amplification is not the key mechanism at play when recurrent connections amplify weak activity pulses. Hebbian amplification alone, governed by intra-assembly recurrent excitatory connections, is sufficient and reliable in allowing for the retrieval of sequences when feedforward connectivity is low.

In this work, we studied the replay of only unidirectional sequences. However, hippocampal replay is known to occur bidirectionally [23, 28]. [34] already showed that a configuration such as our first model with EI assemblies is capable of replaying events bidirectionally if one embeds connections across assemblies in both directions. Because in our simulations the activation of one assembly ends due to the refractoriness following the activation of most of its neurons, we can infer that even in our minimal model without inhibition, such a bidirectional replay of sequences would also be possible.

Recurrent and feedforward connections determine properties of activity pulse

Both our spiking simulations and population model show that, when feedforward connectivity is strong, activity pulses can propagate without needing recurrent amplification, leading to fast and narrow activity pulses, with all neurons within an assembly firing very synchronously. This is the case typically observed in studies of cortical synfire chains [7–9]. Such studies have found that, for a given feedforward connectivity, different pulses may fail or succeed in propagating depending on the activity and width with which they are initiated [8, 9]. In [34], it was shown that having EI assemblies allowing for recurrent amplification increases the range of pulses that can be sustained during feedforward propagation. Nevertheless, these studies found that if pulses do succeed in propagating, their activity and width converge to fixed values after the first few layers. Although we initiate all of our retrieval events with perfectly synchronized pulses in the first assembly, our work describes how the asymptotic features of the activity pulse depend

on the network properties. Both our spiking simulations and population model show that the relation between feedforward and recurrent connectivity shapes the propagation width and speed of sequential activity pulses. For networks without recurrent connections, our model predicts that successful pulses must eventually converge to a fast and narrow-width propagation, as happens in synfire chains. As feedforward connectivity decreases below a critical value, we predict that successful pulses would require amplification through recurrent synapses, needing more synaptic pathways to fully activate an assembly, and thus propagating at a slower speed and with a larger width, with neurons within an assembly firing less synchronously. Moreover, our population model allowed us to obtain an analytical estimate of how the asymptotic replay speed is determined by recurrent and feedforward connections. Our spiking simulations show this estimation can be very good if replay is fast enough such that the membrane leakiness of a neuron can be neglected.

Membrane potential distribution and leak currents determine retrieval success

Our population model allowed us to make analytical predictions of necessary conditions that must be met for a sequence to be successfully retrieved. We showed that, even if infinitely slow pulses were possible in theory, these can only emerge if the sum of the recurrent and feedforward synaptic weights is larger than or equal to the width of the distribution of the neuronal membrane potentials. This suggests that the width and position of the membrane potential distribution of an assembly with respect to the firing threshold is essential in determining whether a sequence can be retrieved or not. This is in line with modeling literature underlying the impact of noisy and balanced AI states in determining successful propagation in synfire chains [11, 13, 56]. Our model shows that, if all assembly neurons have a membrane potential distribution close to the threshold when the propagating activity pulse arrives, then Hebbian amplification by the recurrent excitatory synapses can be fast and reliable enough to allow for the activation of sequences where feedforward connectivity alone would not suffice. Balanced EI networks of the type studied by [34] do possess such membrane potential distributions when they are in a low-firing AI state. However, such a state could also be achieved without inhibition by having noisy and uncorrelated external inputs to the sequence [8], as our minimal spiking model shows.

Another key insight from our study is that the membrane time constant imposes a limit on how slow activity pulses can be to successfully propagate through a sequence of assemblies. Our findings suggest that, for a given membrane time constant, activity pulses can only propagate down to a minimum speed. If the propagation speed of the pulse (determined by the synaptic delays and interplay between recurrent and feedforward synapses) is too slow, then the neuronal leak currents governed by the membrane time constant will dissipate the generated activity before it can propagate to the next assembly. This suggests that neurons with faster membrane time constants may be more susceptible to disruptions in sequential activation, requiring stronger or faster feedforward synapses to propagate a highly synchronous pulse of activity. On the other hand, having slower time constants could be an advantage to recall sequences whose feedforward connectivity is weak, but could nevertheless be retrieved by strong recurrent connections as a less synchronous pulse of activity. These predictions could be particularly interesting for understanding replay in the different regions of the hippocampus. It has been shown that pyramidal neurons in region CA3 have a significantly higher membrane time constant ($\sim$50 ms) than in CA1 ($\sim$10 ms) [57]. These long time constants could be useful in retrieving assemblies and sequences in CA3, a region known to have strong auto-associative connections [33, 58–63]. Moreover, CA3 has been found to exhibit more robust replay than CA1 [64]. This supports the hypothesis that hippocampal sequences may first be learned and replayed in CA3 and then relayed to CA1 [65].

Role of inhibition

Our minimal spiking model showed that feedback inhibition is not needed for the amplification of weak feedforward pulses to occur. Nevertheless, our spiking simulations do show that inhibition modulates this amplification and affects the properties of the propagating activity pulse. Thus, we should expect that inhibition plays an essential role in retrieving sequences in biological networks. Although it is important to study these inhibitory effects, they are beyond the scope of the work here presented.

In the hippocampus, inhibition increases during replay events [66], with several studies showing different interneuron types are involved in the mechanisms of replay-associated sharp wave-ripples [67–72]. [73] reported that inhibitory activity determines sequential spike times of pyramidal cells during sharp wave-ripples. Moreover, the balance between excitation and inhibition determines the membrane potential

distribution of neurons in an AI state, which we have shown is a key component in the dynamics of replay in our model. This was achieved by STDP of the inhibitory-to-excitatory connections [40] in our first two models and had to be manually tuned by Poisson inputs for the third model. In [13], it was shown that embedding inhibitory assemblies (called 'shadow pools') in a network with an excitatory sequence of assemblies allowed for the propagation of an activity pulse on top of AI background activity even in small networks.

Additionally, feedback inhibition can increase the range for which retrieval is possible if connectivity is high, by keeping the transient increase in activity observed during replay from leading to bursting and activity explosions. For the range of connectivity studied in Fig 1, these explosions are only visible in the third model, where feedback inhibition is completely absent. However, if we were to increase p_r or p_f beyond what is shown, we would also see the first and second models creating these bursts [as was observed by 34, for high feedforward connectivity]. Conversely, if we were to keep the same connection probabilities but increase the synaptic weights, the same thing would occur [as seen in 33]. Moreover, if the synaptic weights were strong enough, a low-firing AI state would be impossible to reach without inhibition in our minimal spiking model. As we aimed to understand when connectivity is strong enough for replay to succeed, we did not concern ourselves with the regions where it fails because connectivity is too high. It should be noted that such bursting or exploding activity could also be prevented in a network without inhibition by adding synaptic depression or spike-frequency adaptation mechanisms, which we have also not considered.

To conclude, our combination of spiking simulations with an analytically tractable model allowed us to gain new insights into the mechanisms that regulate sequence retrieval and formulate explicit predictions about the necessary relative strength of different parameters. Our results contribute to the understanding of how neuronal networks may use recurrent connections within Hebbian assemblies to facilitate the activation of neuronal sequences.

Materials and methods

Using spiking network models, we simulated the activity of three different networks and their ability to retrieve an embedded sequence of assemblies. Assemblies are subgroups of cells whose recurrent connections are distinctly stronger than the average network connectivity. We also strengthened the feedforward connections across the embedded assemblies such that they form a sequence. This setup allowed us to study the conditions under which that sequence can be retrieved. Below, we provide a detailed description of the structure of each of the three networks and the tests we have performed on them.

Network simulations and analyses of the spiking network data were performed in Python (`www.python.org`), with the neural network being implemented with the package Brian [74], using a time step of 0.1 ms. Numerical values for all the parameters mentioned below are listed in Table 1. Sample code to reproduce all simulations is available in the GitHub repository `github.com/gaspar-c/replay-conditions`.

Neuron and synapse models

To keep our network models as simple as possible, neurons were described as LIF units. The subthreshold membrane potential $V_j(t)$ of a cell j obeyed

$$C\frac{\mathrm{d}V_j}{\mathrm{d}t} = g^{\mathrm{leak}}(V^{\mathrm{rest}} - V_j) + I_j^{\mathrm{syn}}(t) + I_{bg} \tag{27}$$

where C is the membrane capacitance, g^{leak} is the leak conductance, V^{rest} is the resting membrane potential, $I_j^{\mathrm{syn}}(t)$ is the total synaptic current received by neuron j at time t, and I_{bg} is a constant background current injected to all neurons. Every time a neuron's membrane potential reaches the threshold V^{thr}, a spike is emitted and V_j is reset to the reset potential (for simplicity, it equals V^{rest}), where it is clamped for a refractory period τ_{refr}.

In the first two models (Fig 1, Left and Center panels) there were both excitatory E and inhibitory I cells, and a conductance-based model for synapses was used, such that

$$I_j^{\mathrm{syn}}(t) = g_j^E(t)(V^E - V_j) + g_j^I(t)(V^I - V_j) \tag{28}$$

where V^E and V^I are the reversal potentials of excitation and inhibition, respectively. The time-dependent variables $g_j^E(t)$ and $g_j^I(t)$ describe the total synaptic conductances resulting from incoming inputs to neuron j. The conductance dynamics were described by

$$\begin{aligned} \frac{\mathrm{d}g_j^E}{\mathrm{d}t} &= -\frac{g_j^E}{\tau_d^E} + \sum_{k,f} g_{jk}^{JE}\delta(t - t_k^{(f)} - \tau_l) \\ \frac{\mathrm{d}g_j^I}{\mathrm{d}t} &= -\frac{g_j^I}{\tau_d^I} + \sum_{k,f} g_{jk}^{JI}\delta(t - t_k^{(f)} - \tau_l), \qquad J \in \{E, I\}, \end{aligned} \tag{29}$$

where $\delta(t - t_k^{(f)} - \tau_l)$ is the contribution of the f-th incoming spike from neuron k at time $t_k^{(f)}$, with δ being the Dirac delta function and τ_l the latency between a presynaptic spike and the postsynaptic response onset. The quantities g_{jk}^{JK} denote the unitary conductance increase resulting from a single spike in presynaptic neuron k of population K connected to postsynaptic neuron j of population J, where $J, K \in \{E, I\}$, with E corresponding to the excitatory synapses and I to the inhibitory ones. The conductances decay exponentially with time constants τ_d^E and τ_d^I.

In the third spiking network model (Fig 1, Right panels) there were only excitatory cells ($N_I = 0$), and neurons could receive additional synaptic inputs from an external population P, with each unit generating spikes according to a Poisson process with mean rate ν_P. The synapses coming from the Poisson processes were modeled as an instantaneous jump in the membrane potential of the postsynaptic cell, such that

$$I_j^{\mathrm{syn}}(t) = g_j^E(t)(V^E - V_j) + C\sum_{p,f} w^{EP}\delta(t - t_p^{(f)} - \tau_l) \tag{30}$$

with δ being the Dirac delta function, $t_p^{(f)}$ the f-th incoming spike from the Poisson unit p, and τ_l the latency between a presynaptic spike and the postsynaptic response. The time-dependent variable $g_j^E(t)$ was described by the first equation in (29), while the quantity w^{EP} denotes the unitary membrane potential increase resulting from a single Poisson-generated spike.

Network and connectivity

The first spiking model had a total of N_E excitatory and N_I inhibitory cells. All cells could form connections to each other with a low background probability of p_{bg}. Then, 10 assemblies were formed by selecting groups of non-overlapping M excitatory and $M/4$ inhibitory cells and creating connections among them with some recurrent probability p_r. After the assemblies were formed, a unidirectional sequence was created by selecting the excitatory neurons in each assembly and forming connections to the excitatory neurons in the next assembly with some feedforward connection probability p_f. All $I \to E$ synapses were plastic (see 'Balancing the network' for details).

In the second model, none of the N_I inhibitory cells were part of the sequence, whose assemblies were now made up exclusively of M excitatory cells each. The background connections with low probability p_{bg} were also formed exclusively by the excitatory cells. Instead, the inhibitory cells had fixed connection probabilities to themselves and to/from the excitatory cells, not varying with the values of p_f or p_r. All inhibitory cells formed connections to each other with probability p_{II}, to the excitatory cells with probability p_{EI}, and received connections from all N_E excitatory cells with probability p_{IE}. As in the first model, all $I \to E$ synapses were plastic (see 'Balancing the network' for details).

In the third model, there were no inhibitory cells ($N_I = 0$) nor background connections ($p_{bg} = 0$). To keep the sequence in a low-firing AI state, all neurons received synapses from an external population P containing N_P units, each generating spikes from a Poisson process with rate ν_P. Each of the N_P units could form connections to each of the N_E neurons with probability p_{EP}. There was no plasticity in the third model.

Balancing the network

In the first two models, the inhibitory-to-excitatory synaptic weights were subject to a plasticity mechanism that allows for excitatory cells to receive different amounts of inhibition in order to balance the network, allowing it to reach an AI state [40]. Near coincident pre- and postsynaptic firing potentiates the synapse and lone presynaptic spikes depress them. To implement this plasticity rule, we assigned a synaptic trace

variable x_j to every neuron j, such that x_j is incremented with each spike of that neuron and decays with a time constant τ_{STDP}:

$$\begin{aligned} x_j &\to x_j + 1, \quad \text{if neuron } j \text{ fires,} \\ \tau_{\mathrm{STDP}} \frac{\mathrm{d}x_j}{\mathrm{d}t} &= -x_j, \quad \text{otherwise.} \end{aligned} \tag{31}$$

The synaptic weight g_{jk}^{EI} from inhibitory neuron k to excitatory neuron j was initialized with the same value g_0^{EI} for all $I \to E$ synapses. The conductances were then updated at the times of pre/postsynaptic events in the following manner:

$$\begin{aligned} g_{jk}^{EI} &\to g_{jk}^{EI} + \eta g_0^{EI}(x_j - \alpha), \quad \text{for a presynaptic spike} \\ & \qquad\qquad\qquad\qquad\qquad\qquad \text{in neuron } k, \\ g_{jk}^{EI} &\to g_{jk}^{EI} + \eta g_0^{EI} x_k, \quad \text{for a postsyaptic spike} \\ & \qquad\qquad\qquad\qquad\qquad\qquad \text{in neuron } j, \end{aligned} \tag{32}$$

where η is the learning-rate parameter, and the bias $\alpha = 2\nu_0\tau_{\mathrm{STDP}}$ is determined by the desired firing rate ν_0 of the excitatory postsynaptic neurons.

In the third model, we tuned the external Poisson processes and constant background current to mimic the membrane potential distribution achieved by this plasticity mechanism in the first two models, which fits well to a Gaussian with mean $\sim$ -51 mV and SD $\sim$0.5 mV. We thus chose parameters N_P, ν_P, p^{EP}, w^{EP}, and I_{bg} ad-hoc, such that the third model showed a similar membrane-potential distribution and low-firing AI state in the middle of the explored parameter range i.e. for $p_f = p_f = 8\%$.

Quantifying the retrieval of the embedded sequence

To test if a sequence can be retrieved, we first initialized the network with all its synaptic connections, including the embedded sequence we wished to retrieve. In the case of the first two models, we let the network run with STDP of the $I \to E$ synapses, in order to balance the activity of the neuronal populations. Once the network reached an AI state (which typically occurs after 5 s of simulation), the plasticity was turned off. In the third model, we waited 1 second from the simulation start to ensure the initial conditions did not affect the stimulation. Then, once the neurons in the sequence were in an AI state, all the neurons in the first assembly were made to fire at the same time, by manually setting their membrane potential to the firing threshold. The population activity of each assembly was then filtered with a Gaussian kernel with SD 2 ms, and its peaks were detected using a script by [75].

For a retrieval event to be successful, there were several requirements: First, the peaks must be larger than 30 spikes/s for each assembly, and the peaks between consecutive assemblies must happen in order with some positive delay of at least τ_l. Then, the filtered activity of each assembly is fitted to a Gaussian to find its standard deviation (SD). We then count the spikes within $\pm$ 3 SDs of the filtered peak and check that all assembly neurons fire approximately once without bursting (number of assembly spikes within 90–110% of the number of neurons). If there are background neurons (as in the first two models), their filtered population activity must stay below 30 spikes/s). If these requirements are met, we calculate the asymptotic properties of the activity pulse. The asymptotic speed of propagation was calculated by the inverse of the time difference in the peaks of consecutive assemblies, averaged over the last four assemblies. To obtain the asymptotic pulse width, we further calculate the FWHM for each assembly by multiplying the fitted SD by $2\sqrt{2\log 2}$, and average this across the last three assemblies in the sequence. For each simulated parameter configuration, we performed 5 attempts (with one-second intervals between attempts) on 5 different pseudo-random instantiations of the network. The fraction of successful retrievals (out of the 25) gave us a percentage of success for each point in parameter space. We considered retrieval successful for a given parameter configuration if it succeeded in at least 80% of the attempted times. For all the successful attempts, we average our calculated values of the asymptotic pulse speed and width.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project-ID 327654276 — SFB 1315. We thank Stefano Masserini and Natalie Schieferstein for very helpful discussions and feedback on the manuscript.

Table 1. Numerical parameters for the spiking network models. Used in the simulations of Fig 1 and Fig 4. In Fig 4–B, we set $g^{\mathrm{leak}} = 0$ nS.

Parameter		Model 1	Model 2	Model 3	Definition
N_E		20,000	20,000	5,000	Number of excitatory (E) cells
N_I		5,000	5,000	0	Number of inhibitory (I) cells
M		500	500	500	Number of E cells in one assembly
q		10	10	10	Sequence length (# of assemblies)
C	(pF)	200	200	200	Membrane capacitance
g^{leak}	(nS)	10	10	10	Leak conductance
V^{rest}	(mV)	-60	-60	-60	Resting potential
V^{thr}	(mV)	-50	-50	-50	Voltage threshold
τ_{refr}	(ms)	1	1	1	Refractory period
τ_l	(ms)	1	1	1	Synaptic latency
I_{bg}	(pA)	200	200	58	Constant background current
p^{bg}	(%)	1	1	0	Background connectivity
p^{IE}	(%)	-	1	-	(Global) E-to-I connectivity
p^{II}	(%)	-	1	-	(Global) I-to-I connectivity
p^{EI}	(%)	-	1	-	(Global) I-to-E connectivity
V^E	(mV)	0	0	0	Excitatory reversal potential
g^{EE}	(nS)	0.1	0.1	0.1	E-to-E synaptic weight
g^{IE}	(nS)	0.1	0.1	-	E-to-I synaptic weight
τ_d^E	(ms)	2	2	2	Excitatory synaptic decay time constant
V^I	(mV)	-80	-80	-	Inhibitory reversal potential
g^{II}	(nS)	0.4	0.4	-	I-to-I synaptic weight
g_0^{EI}	(nS)	0.4	0.4	-	I-to-E initial synaptic weight
τ_d^I	(ms)	4	4	-	Inhibitory synaptic decay time constant
η		0.01	0.01	-	STDP learning rate
τ_{STDP}	(ms)	20	20	-	STDP time constant
ν_0	(spk/s)	1	1	-	STDP target firing rate
N_P		0	0	5,000	Number of Poisson units
ν_P	(Hz)	-	-	50	Firing rate of Poisson units
p^{EP}	(%)	-	-	1	P-to-E connection probability
w^{EP}	(μV)	-	-	60	P-to-E synaptic weight

References

[1] D. O. Hebb. *The Organization of Behavior; a Neuropsychological Theory*. Oxford, England: Wiley, 1949, pp. xix, 335. xix, 335.

[2] M. Abeles, H. Bergman, E. Margalit, and E. Vaadia. "Spatiotemporal Firing Patterns in the Frontal Cortex of Behaving Monkeys". In: *Journal of Neurophysiology* 70.4 (1993), pp. 1629–1638. DOI: 10.1152/jn.1993.70.4.1629.

[3] Y. Prut, E. Vaadia, H. Bergman, I. Haalman, H. Slovin, and M. Abeles. "Spatiotemporal Structure of Cortical Activity: Properties and Behavioral Relevance". In: *Journal of Neurophysiology* 79.6 (1998), pp. 2857–2874. DOI: 10.1152/jn.1998.79.6.2857.

[4] T. Shmiel, R. Drori, O. Shmiel, Y. Ben-Shaul, Z. Nadasdy, M. Shemesh, M. Teicher, and M. Abeles. "Neurons of the Cerebral Cortex Exhibit Precise Interspike Timing in Correspondence to Behavior". In: *Proceedings of the National Academy of Sciences* 102.51 (2005), pp. 18655–18657. DOI: 10.1073/pnas.0509346102.

[5] M. A. Long, D. Z. Jin, and M. S. Fee. "Support for a Synaptic Chain Model of Neuronal Sequence Generation". In: *Nature* 468.7322 (2010), pp. 394–399. DOI: 10.1038/nature09514.

[6] M. Abeles. *Local Cortical Circuits*. Berlin, Heidelberg: Springer, 1982. DOI: 10.1007/978-3-642-81708-3.

[7] M. Abeles. *Corticonics: Neural Circuits of the Cerebral Cortex*. Cambridge: Cambridge University Press, 1991. DOI: 10.1017/CBO9780511574566.

[8] M. Diesmann, M.-O. Gewaltig, and A. Aertsen. "Stable Propagation of Synchronous Spiking in Cortical Neural Networks". In: *Nature* 402.6761 (6761 1999), pp. 529–533. DOI: 10.1038/990101.

[9] S. Goedeke and M. Diesmann. "The Mechanism of Synchronization in Feed-Forward Neuronal Networks". In: *New Journal of Physics* 10.1 (2008), p. 015007. DOI: 10.1088/1367-2630/10/1/015007.

[10] J. L. Riquelme, M. Hemberger, G. Laurent, and J. Gjorgjieva. "Single Spikes Drive Sequential Propagation and Routing of Activity in a Cortical Network". In: *eLife* 12 (2023), e79928. DOI: 10.7554/eLife.79928.

[11] Y. Aviel, C. Mehring, M. Abeles, and D. Horn. "On Embedding Synfire Chains in a Balanced Network". In: *Neural Computation* 15.6 (2003), pp. 1321–1340. DOI: 10.1162/089976603321780290.

[12] Y. Aviel, E. Pavlov, M. Abeles, and D. Horn. "Synfire Chain in a Balanced Network". In: *Neurocomputing* 44–46 (2002), pp. 285–292. DOI: 10.1016/S0925-2312(02)00352-1.

[13] Y. Aviel, D. Horn, and M. Abeles. "Synfire Waves in Small Balanced Networks". In: *Neurocomputing* 58–60 (2004), pp. 123–127. DOI: 10.1016/j.neucom.2004.01.032.

[14] H. Câteau and T. Fukai. "Fokker–Planck Approach to the Pulse Packet Propagation in Synfire Chain". In: *Neural Networks* 14.6 (2001), pp. 675–685. DOI: 10.1016/S0893-6080(01)00065-X.

[15] S. Moldakarimov, M. Bazhenov, and T. J. Sejnowski. "Feedback Stabilizes Propagation of Synchronous Spiking in Cortical Neural Networks". In: *Proceedings of the National Academy of Sciences* 112.8 (2015), pp. 2545–2550. DOI: 10.1073/pnas.1500643112.

[16] B. L. McNaughton and R. G. M. Morris. "Hippocampal Synaptic Enhancement and Information Storage within a Distributed Memory System". In: *Trends in Neurosciences* 10.10 (1987), pp. 408–415. DOI: 10.1016/0166-2236(87)90011-7.

[17] E. T. Rolls. "Functions of Neuronal Networks in the Hippocampus and Cerebral Cortex in Memory". In: *Models of Brain Function*. Ed. by R. M. J. Cotterill. Cambridge University Press, 1989, pp. 15–33.

[18] A. Treves and E. T. Rolls. "Computational Analysis of the Role of the Hippocampus in Memory". In: *Hippocampus* 4.3 (1994), pp. 374–391. DOI: 10.1002/hipo.450040319.

[19] E. T. Rolls. "An Attractor Network in the Hippocampus: Theory and Neurophysiology". In: *Learning & Memory* 14.11 (2007), pp. 714–731. DOI: 10.1101/lm.631207.

[20] G. Papp, M. P. Witter, and A. Treves. "The CA3 Network as a Memory Store for Spatial Representations". In: *Learning & Memory* 14.11 (2007), pp. 732–744. DOI: 10.1101/lm.687407.

[21] D. Marr. "Simple Memory: A Theory for Archicortex". In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 262.841 (1971), pp. 23–81. DOI: 10.1098/rstb.1971.0078.

[22] E. Rolls. "The Mechanisms for Pattern Completion and Pattern Separation in the Hippocampus". In: *Frontiers in Systems Neuroscience* 7 (2013). DOI: 10.3389/fnsys.2013.00074.

[23] K. Diba and G. Buzsáki. "Forward and Reverse Hippocampal Place-Cell Sequences during Ripples". In: *Nature Neuroscience* 10.10 (2007), pp. 1241–1242. DOI: 10.1038/nn1961.

[24] W. E. Skaggs and B. L. McNaughton. "Replay of Neuronal Firing Sequences in Rat Hippocampus During Sleep Following Spatial Experience". In: *Science* 271.5257 (1996), pp. 1870–1873. DOI: 10.1126/science.271.5257.1870.

[25] Y.-L. Qin, B. L. Mcnaughton, W. E. Skaggs, and C. A. Barnes. "Memory Reprocessing in Corticocortical and Hippocampocortical Neuronal Ensembles". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 352.1360 (1997), pp. 1525–1533. DOI: 10.1098/rstb.1997.0139.

[26] Z. Nádasdy, H. Hirase, A. Czurkó, J. Csicsvari, and G. Buzsáki. "Replay and Time Compression of Recurring Spike Sequences in the Hippocampus". In: *Journal of Neuroscience* 19.21 (1999), pp. 9497–9507. DOI: 10.1523/JNEUROSCI.19-21-09497.1999.

[27] A. K. Lee and M. A. Wilson. "Memory of Sequential Experience in the Hippocampus during Slow Wave Sleep". In: *Neuron* 36.6 (2002), pp. 1183–1194. DOI: 10.1016/S0896-6273(02)01096-6.

[28] D. J. Foster and M. A. Wilson. "Reverse Replay of Behavioural Sequences in Hippocampal Place Cells during the Awake State". In: *Nature* 440.7084 (2006), pp. 680–683. DOI: 10.1038/nature04587.

[29] A. C. Schapiro, E. A. McDevitt, T. T. Rogers, S. C. Mednick, and K. A. Norman. "Human Hippocampal Replay during Rest Prioritizes Weakly Learned Information and Predicts Memory Performance". In: *Nature Communications* 9.1 (2018), p. 3920. DOI: 10.1038/s41467-018-06213-1.

[30] S. Romani and M. Tsodyks. "Short-Term Plasticity Based Network Model of Place Cells Dynamics". In: *Hippocampus* 25.1 (2015), pp. 94–105. DOI: 10.1002/hipo.22355.

[31] A. H. Azizi, L. Wiskott, and S. Cheng. "A Computational Model for Preplay in the Hippocampus". In: *Frontiers in Computational Neuroscience* 7 (2013). DOI: 10.3389/fncom.2013.00161.

[32] A. Ecker, B. Bagi, E. Vértes, O. Steinbach-Németh, M. R. Karlócai, O. I. Papp, I. Miklós, N. Hájos, T. F. Freund, A. I. Gulyás, and S. Káli. "Hippocampal Sharp Wave-Ripples and the Associated Sequence Replay Emerge from Structured Synaptic Interactions in a Network Model of Area CA3". In: *eLife* 11 (2022), e71850. DOI: 10.7554/eLife.71850.

[33] R. P. Sammons, M. Vezir, L. Moreno-Velasquez, G. Cano, M. Orlando, M. Sievers, E. Grasso, V. D. Metodieva, R. Kempter, H. Schmidt, and D. Schmitz. "Structure and Function of the Hippocampal CA3 Module". In: *Proceedings of the National Academy of Sciences* 121.6 (2024), e2312281120. DOI: 10.1073/pnas.2312281120.

[34] N. Chenkov, H. Sprekeler, and R. Kempter. "Memory Replay in Balanced Recurrent Networks". In: *PLOS Computational Biology* 13.1 (2017), e1005359. DOI: 10.1371/journal.pcbi.1005359.

[35] D. Kleinfeld. "Sequential State Generation by Model Neural Networks." In: *Proceedings of the National Academy of Sciences* 83.24 (1986), pp. 9469–9473. DOI: 10.1073/pnas.83.24.9469.

[36] H. Sompolinsky and I. Kanter. "Temporal Association in Asymmetric Neural Networks". In: *Physical Review Letters* 57.22 (1986), pp. 2861–2864. DOI: 10.1103/PhysRevLett.57.2861.

[37] A. Treves and E. T. Rolls. "What Determines the Capacity of Autoassociative Memories in the Brain?" In: *Network: Computation in Neural Systems* 2.4 (1991), pp. 371–397. DOI: 10.1088/0954-898X_2_4_004.

[38] B. K. Murphy and K. D. Miller. "Balanced Amplification: A New Mechanism of Selective Amplification of Neural Activity Patterns". In: *Neuron* 61.4 (2009), pp. 635–648. DOI: 10.1016/j.neuron.2009.02.005.

[39] T. M. Stöber, A. B. Lehr, M. Fyhn, and A. Kumar. *Competition and Cooperation of Assembly Sequences in Recurrent Neural Networks*. 2023. DOI: 10.1101/2023.11.03.565506. Pre-published.

[40] T. P. Vogels, H. Sprekeler, F. Zenke, C. Clopath, and W. Gerstner. "Inhibitory Plasticity Balances Excitation and Inhibition in Sensory Pathways and Memory Networks". In: *Science* 334.6062 (2011), pp. 1569–1573. DOI: 10.1126/science.1211095.

[41] W. Gerstner and W. M. Kistler. *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge: Cambridge University Press, 2002. ISBN: 978-0-521-89079-3. DOI: 10.1017/CBO9780511815706.

[42] N. Brunel. "Dynamics of Sparsely Connected Networks of Excitatory and Inhibitory Spiking Neurons". In: *Journal of Computational Neuroscience* 8.3 (2000), pp. 183–208. DOI: 10.1023/A:1008925309027.

[43] N. Maier, V. Nimmrich, and A. Draguhn. "Cellular and Network Mechanisms Underlying Spontaneous Sharp Wave–Ripple Complexes in Mouse Hippocampal Slices". In: *The Journal of Physiology* 550.3 (2003), pp. 873–887. DOI: 10.1113/jphysiol.2003.044602.

[44] K. T. Hofer, Á. Kandrács, I. Ulbert, I. Pál, C. Szabó, L. Héja, and L. Wittner. "The Hippocampal CA3 Region Can Generate Two Distinct Types of Sharp Wave-Ripple Complexes, in Vitro". In: *Hippocampus* 25.2 (2015), pp. 169–186. DOI: 10.1002/hipo.22361.

[45] K. Kajikawa, B. K. Hulse, A. G. Siapas, and E. V. Lubenov. "UP-DOWN States and Ripples Differentially Modulate Membrane Potential Dynamics across DG, CA3, and CA1 in Awake Mice". In: *eLife* 11 (2022), e69596. DOI: 10.7554/eLife.69596.

[46] J. J. Hopfield. "Neural Networks and Physical Systems with Emergent Collective Computational Abilities." In: *Proceedings of the National Academy of Sciences* 79.8 (1982), pp. 2554–2558. DOI: 10.1073/pnas.79.8.2554.

[47] J. E. Lisman. "Relating Hippocampal Circuitry to Function: Recall of Memory Sequences by Reciprocal Dentate–CA3 Interactions". In: *Neuron* 22.2 (1999), pp. 233–242. DOI: 10.1016/S0896-6273(00)81085-5.

[48] S. Jahnke, M. Timme, and R.-M. Memmesheimer. "A Unified Dynamic Model for Learning, Replay, and Sharp-Wave/Ripples". In: *Journal of Neuroscience* 35.49 (2015), pp. 16236–16258. DOI: 10.1523/JNEUROSCI.3977-14.2015.

[49] S. A. Josselyn and S. Tonegawa. "Memory Engrams: Recalling the Past and Imagining the Future". In: *Science* 367.6473 (2020), eaaw4325. DOI: 10.1126/science.aaw4325.

[50] G.-q. Bi and M.-m. Poo. "Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type". In: *Journal of Neuroscience* 18.24 (1998), pp. 10464–10472. DOI: 10.1523/JNEUROSCI.18-24-10464.1998.

[51] R. K. Mishra, S. Kim, S. J. Guzman, and P. Jonas. "Symmetric Spike Timing-Dependent Plasticity at CA3–CA3 Synapses Optimizes Storage and Recall in Autoassociative Networks". In: *Nature Communications* 7.1 (2016), p. 11552. DOI: 10.1038/ncomms11552.

[52] E. Fino, J. Glowinski, and L. Venance. "Bidirectional Activity-Dependent Plasticity at Corticostriatal Synapses". In: *Journal of Neuroscience* 25.49 (2005), pp. 11279–11287. DOI: 10.1523/JNEUROSCI.4476-05.2005.

[53] M. Gillett, U. Pereira, and N. Brunel. "Characteristics of Sequential Activity in Networks with Temporally Asymmetric Hebbian Learning". In: *Proceedings of the National Academy of Sciences* 117.47 (2020), pp. 29948–29958. DOI: 10.1073/pnas.1918674117.

[54] M. Farrell and C. Pehlevan. "Recall Tempo of Hebbian Sequences Depends on the Interplay of Hebbian Kernel with Tutor Signal Timing". In: *Proceedings of the National Academy of Sciences* 121.32 (2024), e2309876121. DOI: 10.1073/pnas.2309876121.

[55] M. Gillett and N. Brunel. "Dynamic Control of Sequential Retrieval Speed in Networks with Heterogeneous Learning Rules". In: *eLife* 12 (2024). DOI: 10.7554/eLife.88805.2.

[56] A. Kumar, S. Rotter, and A. Aertsen. "Conditions for Propagating Synchronous Spiking and Asynchronous Firing Rates in a Cortical Network Model". In: *Journal of Neuroscience* 28.20 (2008), pp. 5268–5280. DOI: 10.1523/JNEUROSCI.2542-07.2008.

[57] J. Kowalski, J. Gan, P. Jonas, and A. J. Pernía-Andrade. "Intrinsic Membrane Properties Determine Hippocampal Differential Firing Pattern in Vivo in Anesthetized Rats". In: *Hippocampus* 26.5 (2016), pp. 668–682. DOI: 10.1002/hipo.22550.

[58] L. W. Swanson, J. M. Wyss, and W. M. Cowan. "An Autoradiographic Study of the Organization of Intrahippocampal Association Pathways in the Rat". In: *Journal of Comparative Neurology* 181.4 (1978), pp. 681–715. DOI: 10.1002/cne.901810402.

[59] S. Laurberg and K. E. Sørensen. "Associational and Commissural Collaterals of Neurons in the Hippocampal Formation (Hilus Fasciae Dentatae and Subfield CA3)". In: *Brain Research* 212.2 (1981), pp. 287–300. DOI: 10.1016/0006-8993(81)90463-7.

[60] N. Ishizuka, J. Weber, and D. G. Amaral. "Organization of Intrahippocampal Projections Originating from CA3 Pyramidal Cells in the Rat". In: *Journal of Comparative Neurology* 295.4 (1990), pp. 580–623. DOI: 10.1002/cne.902950407.

[61] L. Wittner, D. A. Henze, L. Záborszky, and G. Buzsáki. "Three-Dimensional Reconstruction of the Axon Arbor of a CA3 Pyramidal Cell Recorded and Filled in Vivo". In: *Brain Structure and Function* 212.1 (2007), pp. 75–83. DOI: 10.1007/s00429-007-0148-y.

[62] R. P. Sammons, S. Masserini, L. Moreno-Velasquez, V. D. Metodieva, G. Cano, A. Sannio, M. Orlando, N. Maier, R. Kempter, and D. Schmitz. "Sub-Type Specific Connectivity between CA3 Pyramidal Neurons May Underlie Their Sequential Activation during Sharp Waves". In: *eLife* 13 (2024). DOI: 10.7554/eLife.98653.1.

[63] R. Layous, B. Tamás, A. Mike, E. Sipos, A. Arszovszki, J. Brunner, Á. Szatai, F. Yaseen, T. Andrási, and J. Szabadics. *Optical Recordings of Unitary Synaptic Connections Reveal High and Random Local Connectivity between CA3 Pyramidal Cells*. 2025. DOI: 10.1101/2025.01.15.633173. Pre-published.

[64] M. P. Karlsson and L. M. Frank. "Awake Replay of Remote Experiences in the Hippocampus". In: *Nature Neuroscience* 12.7 (2009), pp. 913–918. DOI: 10.1038/nn.2344.

[65] M. F. Carr, S. P. Jadhav, and L. M. Frank. "Hippocampal Replay in the Awake State: A Potential Substrate for Memory Consolidation and Retrieval". In: *Nature Neuroscience* 14.2 (2 2011), pp. 147–153. DOI: 10.1038/nn.2732.

[66] T. Klausberger and P. Somogyi. "Neuronal Diversity and Temporal Dynamics: The Unity of Hippocampal Circuit Operations". In: *Science* 321.5885 (2008), pp. 53–57. DOI: 10.1126/science.1149381.

[67] R. Evangelista, G. Cano, C. Cooper, D. Schmitz, N. Maier, and R. Kempter. "Generation of Sharp Wave-Ripple Events by Disinhibition". In: *The Journal of Neuroscience* 40.41 (2020), pp. 7811–7836. DOI: 10.1523/JNEUROSCI.2174-19.2020.

[68] D. Schlingloff, S. Káli, T. F. Freund, N. Hájos, and A. I. Gulyás. "Mechanisms of Sharp Wave Initiation and Ripple Generation". In: *Journal of Neuroscience* 34.34 (2014), pp. 11385–11398. DOI: 10.1523/JNEUROSCI.0867-14.2014.

[69] Z. Kohus, S. Káli, L. Rovira-Esteban, D. Schlingloff, O. Papp, T. F. Freund, N. Hájos, and A. I. Gulyás. "Properties and Dynamics of Inhibitory Synaptic Communication within the CA3 Microcircuits of Pyramidal Cells and Interneurons Expressing Parvalbumin or Cholecystokinin". In: *The Journal of Physiology* 594.13 (2016), pp. 3745–3774. DOI: 10.1113/JP272231.

[70] J. R. Donoso, D. Schmitz, N. Maier, and R. Kempter. "Hippocampal Ripple Oscillations and Inhibition-First Network Models: Frequency Dynamics and Response to GABA Modulators". In: *Journal of Neuroscience* 38.12 (2018), pp. 3124–3146. DOI: 10.1523/JNEUROSCI.0188-17.2018.

[71] N. Schieferstein, T. Schwalger, B. Lindner, and R. Kempter. "Intra-Ripple Frequency Accommodation in an Inhibitory Network Model for Hippocampal Ripple Oscillations". In: *PLOS Computational Biology* 20.2 (2024), e1011886. DOI: 10.1371/journal.pcbi.1011886.

[72] N. Brunel and X.-J. Wang. "What Determines the Frequency of Fast Network Oscillations with Irregular Neural Discharges? I. Synaptic Dynamics and Excitation-Inhibition Balance". In: *Journal of Neurophysiology* 90.1 (2003), pp. 415–430. DOI: 10.1152/jn.01095.2002.

[73] A. Noguchi, R. Huszár, S. Morikawa, G. Buzsáki, and Y. Ikegaya. "Inhibition Allocates Spikes during Hippocampal Ripples". In: *Nature Communications* 13.1 (2022), p. 1280. DOI: 10.1038/s41467-022-28890-9.

[74] D. F. Goodman and R. Brette. "The Brian Simulator". In: *Frontiers in Neuroscience* 3 (2009). DOI: 10.3389/neuro.01.026.2009.

[75] M. Duarte and R. N. Watanabe. *Notes on Scientific Computing for Biomechanics and Motor Control*. Version v0.0.2. 2021. DOI: 10.5281/zenodo.4599319.

Supporting information

S1 Appendix Minimal spiking model with instant synapses

In all spiking models simulated in the main text of this chapter, $E \to E$ synapses were modeled as jumps in conductance that decay exponentially — see (28) and (29). Here, we simulate our third spiking model (see Fig 1–A3) and replace the conductance-based synapses with synapses that act as delta-currents, causing an instantaneous jump w^{EE} in the membrane potential of the postsynaptic neuron after a delay τ_l^E. We perform the same simulation protocol used in Fig 1. The results are shown below:

These results show that the choice of synaptic model used in the main text is not essential to reproduce the dynamics being investigated. Moreover, we show that strengthening or weakening the weight of excitatory synapses w^{EE} moves the regions where replay can succeed, as expected (see Eq (23) in the main text).

The synaptic currents a LIF neuron j receives are now given by

$$I_j^{\text{syn}}(t) = C \sum_{e,f} w^{EE} \delta(t - t_e^{(f)} - \tau_l^E) + C \sum_{p,f} w^{EP} \delta(t - t_p^{(f)} - \tau_l) \tag{S1.1}$$

with δ being the Dirac delta function, $t_e^{(f)}$ the f-th incoming spike from excitatory neuron e, $t_p^{(f)}$ the f-th incoming spike from the Poisson unit p, and τ_l the latency between a presynaptic spike in a Poisson unit and the postsynaptic response. The quantity w^{EE} denotes the unitary membrane potential increase resulting from a single spike in the excitatory population E, whereas w^{EP} is the unitary increase resulting from a single spike in the Poisson population P. In our simulations in the main text, we used conductance-based synapses with $\tau_l = 1$ ms and a conductance exponential decay of 2 ms. As our neurons have time constant $\tau_m = 20$ ms, the average point in time at which the postsynaptic potential increases was ∼3 ms after the presynaptic neuron spikes. To approximate this, here we chose $\tau_l^E = 3$ ms.

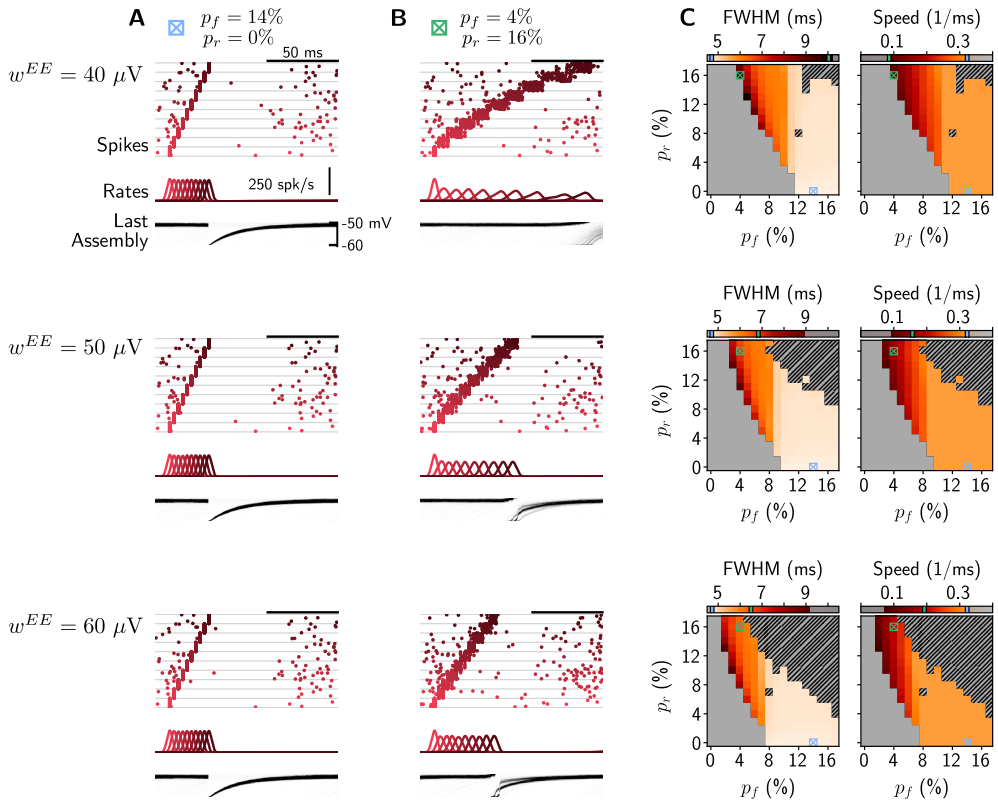

Fig S1.1. Same as Fig 1, but we simulate Model 3 with delta-current synapses with synaptic delay $\tau_l^E = 3$ ms and three different synaptic weights w^{EE}.

S2 Appendix Relation between pulse width and speed

For successful sequence retrieval, we measured the pulse speed and width, both in the spiking network and population model simulations. The pulse speed measures how fast the pulse travels from one assembly to the next, thus telling us how long the sequence will take to be replayed. On the other hand, the pulse width, given by the FWHM of an assembly's population rate, is proportional to how long it takes for an assembly to fully cross the threshold. As we saw in our spiking simulations, when feedforward connectivity is very strong, recurrent connections are not needed for retrieval, with propagating pulses being fast and narrow. Conversely, when feedforward connectivity is weak, recurrent connections are needed for retrieval, with pulses being slower to propagate and having a larger width. Thus, we expect pulse with and speed to be correlated, as suggested by Fig 1–D. Below, we quantitatively compare the width and speed of activity pulses in the simulations of Fig 1 and Fig 3–A,C. Our analysis shows that both pulse width and speed are clearly correlated across all models.

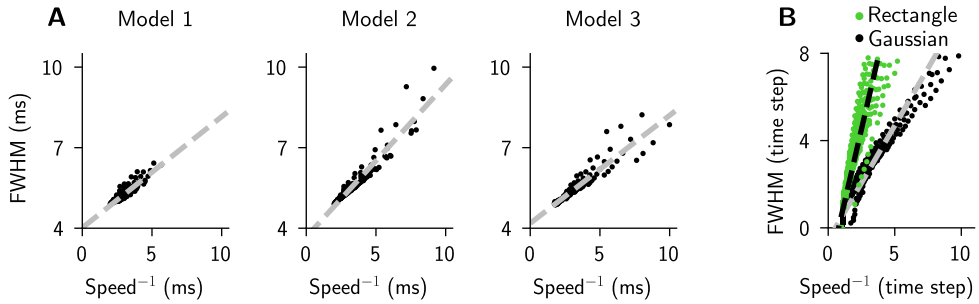

Fig S2.1. Correlation between pulse width and speed. A, Asymptotic pulse speed and width (FWHM) of the three spiking models in Fig 1–D1:D3. The Pearson correlation coefficients are: 0.916 for the first model, 0.958 for the second model, and 0.945 for the third model (p-value < 0.001 for all three). The dashed line corresponds to the best-fit linear regression. **B**, Same as panel A, but for the population model (values shown in Fig 3–A,C), with the clipped-Gaussian (black dots) and rectangle distributions (green dots). For better visibility, we limit the analysis to pulses with a FWHM no wider than 8 time steps. The Pearson correlation coefficients are: 0.938 for the clipped Gaussian, and 0.921 for the rectangle (p-values < 0.001 for both cases). The dashed line corresponds to the best-fit linear regression.

S3 Appendix Analytical derivations for the rectangle model

This appendix provides a detailed description of the analytical solutions that can be derived from the rectangle model and its linear approximation. To facilitate reading, the text has been written to be self-contained. Thus, there is some repetition of the contents of the main text.

Model definition

We model the distribution of membrane potentials v of a given assembly $i \in \{1, 2, ..., q\}$ in a sequence of length q at some time step $t \in \{0, 1, 2, ...\}$ with a uniform distribution with constant finite width U, such that probability density function $g_t^{(i)}(v)$ is

$$g_t^{(i)}(v) = \begin{cases} 1/U & , \quad x_t^{(i)} - U \leq v \leq x_t^{(i)} \\ 0 & , \quad \text{otherwise.} \end{cases} \tag{S3.1}$$

For $t \geq 1$, the corresponding firing activity $a_t^{(i)}$ for some assembly i at that time will be the area under the distribution that crossed the firing threshold between the previous time $t-1$ and the current time t

$$a_t^{(i)} \quad = \quad \int_0^{+\infty} \left[g_t^{(i)}(v) - g_{t-1}^{(i)}(v) \right] \, \mathrm{d}v \, , \tag{S3.2}$$

where we define the threshold to be located at $v = 0$. We further assume initial conditions

$$x_0^{(i)} := 0 \, , \quad a_0^{(i)} = 0 \, , \quad \forall i \, , \tag{S3.3}$$

such that all assemblies start out with zero activity and the right edge of their distribution located at the threshold.

In a given time t, the right edge $x_t^{(i)}$ of the membrane potential distribution of assembly i increases with respect to its previous position $x_{t-1}^{(i)}$ by some amount proportional to the synaptic input to that assembly. Said input can either come from recurrent connections from the assembly itself, feedforward connections from the previous assembly, or some external input $J_t^{(i)}$ which we assume is equally applied to each neuron in assembly i. These dynamics can be written as

$$x_t^{(i)} = x_{t-1}^{(i)} + R \cdot a_{t-1}^{(i)} + F \cdot a_{t-1}^{(i-1)} + J_t^{(i)} \,, \tag{S3.4}$$

where R and F are the weights of the recurrent and feedforward connections, respectively. Alternatively, these dynamics can be written in a non-recurrent form:

$$x_t^{(i)} = R \cdot \sum_{j=0}^{t-1} a_j^{(i)} + F \cdot \sum_{k=0}^{t-1} a_k^{(i-1)} + \sum_{l=1}^{t} J_l^{(i)} \,, \tag{S3.5}$$

where we used our initial condition $x_0^{(i)} = 0$. For Equations S3.4–S3.5 to apply to the first assembly, we define $a_t^{(0)} = 0$ for all t.

We consider the only external input received by the sequence is given to the first assembly as some constant $0 \leq J \leq U$ at $t = 1$, such that

$$J_t^{(i)} = \begin{cases} J \,, & i = 1,\, t = 1 \\ 0 \,, & \text{otherwise.} \end{cases} \tag{S3.6}$$

We then consider that an assembly is successfully recalled if its distribution fully crosses the threshold, and we define $\bar{t}_i$ as the minimal time at which this happens for assembly i, such that $x_t^{(i)} \geq U$ for any $t \geq \bar{t}_i$. For a sequence with q assemblies, we consider that the whole sequence has been successfully recalled if the membrane potential distribution of the last assembly fully crosses the threshold, i.e., if there exists some finite time $\bar{t}_q$ for which

$$x_{\bar{t}_q}^{(q)} \geq U \,, \tag{S3.7}$$

If we choose initial conditions (S3.3), the activity $a_t^{(i)}$ given by (S3.2) generated by the rectangle distribution (S3.1) with width U crossing the threshold (defined to be at $v = 0$) is given by

$$a_t^{(i)} = \frac{\min\left(x_t^{(i)},\, U\right) - \min\left(x_{t-1}^{(i)},\, U\right)}{U} \,, \quad \text{for} \quad x_{t-1}^{(i)}, x_t^{(i)} \geq 0 \,. \tag{S3.8}$$

If the assembly has not yet fully crossed the threshold, the activity (S3.8) is the simple linear relation

$$a_t^{(i)} = \frac{x_t^{(i)} - x_{t-1}^{(i)}}{U} \,, \quad \text{for} \quad 0 \leq x_{t-1}^{(i)}, x_t^{(i)} \leq U \,. \tag{S3.9}$$

Conditions for sequential activation

Activating an isolated assembly

To determine the conditions for activating an isolated assembly, let us look at the behavior of the first assembly, $i = 1$, which does not receive any feedforward inputs. Consider the dynamics of assembly 1, given by (S3.4), when it receives some constant input $0 < J \leq U$ at $t = 1$,

$$x_1^{(1)} = J \,, \quad a_1^{(1)} = J/U \,. \tag{S3.10}$$

If $J = U$, the first assembly becomes fully activated at $t = 1$, by definition. If, however, $J < U$, we must determine what happens in the subsequent time steps. Following (S3.5), the right edge of assembly 1 is given by

$$x_t^{(1)} = \begin{cases} 0 & , \quad \text{if} \quad t = 0 \\ R \sum_{k=1}^{t} a_{k-1}^{(1)} + J & , \quad \text{if} \quad t \geq 1 \end{cases} \tag{S3.11}$$

since the only external input is J at $t = 1$, and there are no feedforward inputs. Moreover, we also know that at any step t for which an assembly has not yet been fully activated, its activity is given by the linear relation (S3.9). Using (S3.9) and (S3.11), we see that for any $t > 1$ at which assembly 1 has not yet become fully activated, i.e., $x_t^{(1)} \leq U$:

$$a_t^{(1)} = \frac{x_t^{(1)} - x_{t-1}^{(1)}}{U} = (R/U) \cdot a_{t-1}^{(1)} \tag{S3.12}$$

We can solve this equation with initial condition $a_1^{(1)} = J/U$ (S3.10) to find

$$a_t^{(1)} = (J/U) \cdot (R/U)^{t-1}\,, \quad \text{for} \quad 0 \leq x_t^{(1)} \leq U\,. \tag{S3.13}$$

Inserting (S3.3) and (S3.13) into (S3.11), we find that, while $x_t^{(1)} \leq U$:

$$x_t^{(1)} = \begin{cases} 0 & , \quad \text{if} \quad t = 0 \\ J \cdot \sum_{k=1}^{t} (R/U)^{k-1} & , \quad \text{if} \quad t \geq 1 \end{cases} \tag{S3.14}$$

Our goal is to determine whether assembly 1 will eventually become fully activated if given enough time steps, i.e.: Is there some finite time $\bar{t}_1$ for which $x_{\bar{t}_1}^{(1)} \geq U$? To determine this, we start by noting that:

$$J \cdot \sum_{k=1}^{t} (R/U)^{k-1} = \begin{cases} \frac{J}{1-R/U} \cdot (1 - (R/U)^t) & , \quad R \neq U \\ J \cdot t & , \quad R = U \end{cases}$$

If $R \geq U$, we immediately see that $x_t^{(1)}$ will grow at a constant or increasing rate, and assembly 1 will fully cross the threshold if given enough time steps. If $R < U$, $x_t^{(1)}$ is a monotonically increasing function of t that asymptotically tends to the limit $J/(1 - R/U)$ as $t \to \infty$. Thus, if given enough time steps, assembly 1 will become fully activated if $J/(1 - R/U) > U \Leftrightarrow J + R > U$ — note that once $x_t^{(1)} > U$ the dynamics of (S3.14) no longer apply, but we can nevertheless know whether this will happen at some point. If $J + R = 1$, the assembly will never become fully activated within a finite number of time steps.

Sequence retrieval without recurrent connections

Let us now consider a sequence with multiple assemblies but without recurrent connections ($R = 0, F > 0$). Given the same input (S3.6), assemblies only move once to the right, which happens when the feedforward input arrives at time $t = i$. Assembly 1 will thus only generate positive activity at $t = 1 : a_1^{(1)} = J/U$, which is the only input assembly 2 will receive.

Let us first consider the case where $F < U$. The dynamics of assembly 2 will be

$$x_2^{(2)} = F \cdot J/U\,, \quad a_2^{(2)} = F \cdot J/U^2 \tag{S3.15}$$

Thus, assembly 3 will only move to the right at time $t = 3$, with

$$x_3^{(3)} = F^2 \cdot J/U^2\,, \quad a_3^{(3)} = F^2 \cdot J/U^3 \tag{S3.16}$$

If we continue iteratively, we find that, in general:

$$x_i^{(i)} = (F/U)^{i-1} \cdot J\,, \quad \text{if} \quad F < U\,. \tag{S3.17}$$

Thus, we see that for $F < U$, as the generated pulse moves through the sequence, it gets weaker and weaker, with assemblies becoming progressively less activated. Thus, retrieval is not possible for a sequence of any length $q \geq 2$, even if we fully activate the first assembly by setting $J = U$.

Let us now consider the case where $F \geq U$. We can derive for the second assembly that

$$x_2^{(2)} = F \cdot J/U\,, \quad a_2^{(2)} = \begin{cases} (F/U) \cdot J/U & , \quad \text{if} \quad F \cdot J < U^2 \\ 1 & , \quad \text{if} \quad F \cdot J \geq U^2 \end{cases} \tag{S3.18}$$

Subsequently, $a_2^{(2)}$ is the only input assembly 3 will receive, and we can derive that

$$x_3^{(3)} = \begin{cases} (F/U)^2 \cdot J & , \quad \text{if} \quad F \cdot J < U^2 \\ F & , \quad \text{if} \quad F \cdot J \geq U^2 \end{cases} \tag{S3.19}$$

$$a_3^{(3)} = \begin{cases} (F/U)^2 \cdot J/U & , \quad \text{if} \quad F \cdot J < U^2 \wedge F^2 \cdot J < U^3 \\ 1 & , \quad \text{otherwise.} \end{cases} \tag{S3.20}$$

Since $F \geq U$, the condition $F \cdot J < U^2 \wedge F^2 \cdot J < U^3$ is simply $F^2 \cdot J < U^3$. If we continue iteratively, we find that, in general, for $i \geq 2$:

$$x_i^{(i)} = \begin{cases} (F/U)^{i-1} \cdot J & , \quad \text{if} \quad F^{i-2} \cdot J < U^{i-1} \\ F & , \quad \text{if} \quad F^{i-2} \cdot J \geq U^{i-1} \end{cases} . \tag{S3.21}$$

Thus, the pulse will get stronger and stronger and assemblies become progressively more activated until one of them eventually fully crosses the threshold when the condition $F^{i-1} \cdot J \geq U^i$ is met. After that, all downstream assemblies will become fully activated as well, and the rest of the sequence is retrieved at the speed of one assembly per time step (since $F \geq U$). Thus, for $F \geq U$, recurrent connections are not needed for sequence retrieval.

Partial activation may lead to full activation via recurrent and feedforward connections

If an assembly i has not yet fully crossed the threshold at some time t, the activity (S3.2) is given by the simple linear relation (S3.9). Let us assume that $J < U$, such that assembly 1 does not become instantly fully activated. If the connectivity is strong enough, some assembly m will be the first assembly to become fully activated at some time $\bar{t}_m$. Then, we can extend the linear relation (S3.9) to all activities in (S3.4), such that

$$\hat{x}_t^{(i)} = \hat{x}_{t-1}^{(i)} + R \cdot \frac{\hat{x}_{t-1}^{(i)} - \hat{x}_{t-2}^{(i)}}{U} + F \cdot \frac{\hat{x}_{t-1}^{(i-1)} - \hat{x}_{t-2}^{(i-1)}}{U} , \tag{S3.22}$$

with $\hat{x}_t^{(i)} = x_t^{(i)}$ for $i \leq m$ and $t \leq \bar{t}_m$. Otherwise, we consider $\hat{x}_t^{(i)}$ to be an estimate of $x_t^{(i)}$, whith $\hat{x}_t^{(i)} \geq x_t^{(i)}$.

Solution for the first assembly. We have already found the general solution for the first assembly before it becomes fully activated (S3.14). From it, we know that:

$$\hat{x}_t^{(1)} = \begin{cases} 0 & , \quad t = 0 \\ J \cdot \sum_{k=1}^{t} (R/U)^{k-1} & , \quad t \geq 1 \end{cases} \tag{S3.23}$$

Solution for the second assembly. From (S3.23), we can see that

$$\hat{x}_t^{(1)} - \hat{x}_{t-1}^{(1)} = J \cdot (R/U)^{t-1} , \quad \forall \quad t \geq 1 \tag{S3.24}$$

With this, we can easily compute the solutions for the second assembly for $t \geq 2$:

$$\begin{aligned} \hat{x}_2^{(2)} &= J \cdot F/U \\ \hat{x}_3^{(2)} &= J \cdot (F/U) \cdot (1 + 2R/U) \\ \hat{x}_4^{(2)} &= J \cdot (F/U) \cdot (1 + 2R/U + 3(R/U)^2) \\ \hat{x}_5^{(2)} &= J \cdot (F/U) \cdot (1 + 2R/U + 3(R/U)^2 + 4(R/U)^3) \\ &\ldots \end{aligned}$$

which in general can be written as

$$\hat{x}_t^{(2)} = \begin{cases} 0 & , \quad t < 2 \\ J \cdot (F/U) \cdot \sum_{k=2}^{t} (k-1)(R/U)^{k-2} & , \quad t \geq 2 \end{cases} \tag{S3.25}$$

Solution for the third assembly. From (S3.25), we can see that

$$\hat{x}_t^{(2)} - \hat{x}_{t-1}^{(2)} = J(F/U)(t-1)(R/U)^{t-2}\,, \quad \forall \quad t \geq 2 \tag{S3.26}$$

Thus, the solutions for the third assembly for $t \geq 3$ are:

$$\begin{aligned}
\hat{x}_3^{(3)} &= J \cdot (F/U)^2 \\
\hat{x}_4^{(3)} &= J \cdot (F/U)^2 \cdot (1 + 3R/U) \\
\hat{x}_5^{(3)} &= J \cdot (F/U)^2 \cdot (1 + 3R/U + 6(R/U)^2) \\
\hat{x}_6^{(3)} &= J \cdot (F/U)^2 \cdot (1 + 3R/U + 6(R/U)^2 + 10(R/U)^3) \\
&\ldots
\end{aligned}$$

which in general can be written as

$$\hat{x}_t^{(3)} = \begin{cases} 0 & , \quad t < 3 \\ J \cdot (F/U)^2 \cdot \displaystyle\sum_{k=3}^{t} \binom{k-1}{2} (R/U)^{k-3} & , \quad t \geq 3 \end{cases} \tag{S3.27}$$

where $\binom{n}{p}$ is the corresponding binomial coefficient.

General solution. If we continue this process iteratively, we will find that

$$\hat{x}_t^{(i)} = \begin{cases} 0 & , \quad t < i \\ J \cdot (F/U)^{i-1} \cdot \displaystyle\sum_{k=i}^{t} \binom{k-1}{i-1} (R/U)^{k-i} & , \quad t \geq i \end{cases} \tag{S3.28}$$

Proof by induction. We have just inferred a general expression for $\hat{x}_t^{(i)}$ by iterating (S3.22) from our initial conditions. Here, we will show by induction that (S3.28) is indeed the solution to the difference equation (S3.22) for any $i \geq 1$ and $t \geq 1$, given initial conditions $\hat{x}_0^{(i)} = 0$.

Let us first consider the base case, $i = 1$. There, (S3.28) becomes

$$\hat{x}_t^{(1)} = \begin{cases} 0 & , \quad t < 1 \\ J \cdot \displaystyle\sum_{k=1}^{t} (R/U)^{k-1} & , \quad t \geq 1\,. \end{cases} \tag{S3.29}$$

Eq (S3.29) is equivalent to (S3.14), which we have already shown is the solution of $x_t^{(1)}$ while $0 \leq x_t^{(1)} \leq U$. Since $\hat{x}_t^{(1)} = x_t^{(1)}$ for $0 \leq x_t^{(1)} \leq U$, we know that (S3.29) is the solution to (S3.22) for the base case $i = 1$. □

We will now do the induction step, by assuming that the solution given by (S3.28) is true for some assembly $i = l - 1$, i.e.,

$$\hat{x}_t^{(l-1)} = \begin{cases} 0 & , \quad t < l-1 \\ J \cdot (F/U)^{l-2} \cdot \displaystyle\sum_{k=l-1}^{t} \binom{k-1}{l-2} (R/U)^{k-(l-1)} & , \quad t \geq l-1 \end{cases} \tag{S3.30}$$

Let us write the difference equation (S3.22) for $i = l$, i.e.,

$$\hat{x}_t^{(l)} = \hat{x}_{t-1}^{(l)} + R \cdot \frac{\hat{x}_{t-1}^{(l)} - \hat{x}_{t-2}^{(l)}}{U} + F \cdot \frac{\hat{x}_{t-1}^{(l-1)} - \hat{x}_{t-2}^{(l-1)}}{U}\,. \tag{S3.31}$$

We first consider $t < l$. In this case, we know that $t - 1 < l - 1$ and so our induction hypothesis (S3.30) tellls that $\hat{x}_{t-1}^{(l-1)} = \hat{x}_{t-2}^{(l-1)} = 0$. We can plug this into (S3.31) and find

$$\hat{x}_t^{(l)} = \hat{x}_{t-1}^{(l)} + R \cdot \frac{\hat{x}_{t-1}^{(l)} - \hat{x}_{t-2}^{(l)}}{U}, \quad \text{for} \quad t < l\,. \tag{S3.32}$$

Let us define the change in $\hat{x}_t^{(i)}$ within one time step:

$$\Delta\hat{x}_t^{(i)} := \begin{cases} 0 & , \quad \text{if} \quad t = 0 \\ \hat{x}_t^{(i)} - \hat{x}_{t-1}^{(i)} & , \quad \text{if} \quad t \geq 1\,. \end{cases} \tag{S3.33}$$

such that we can write

$$\hat{x}_t^{(i)} = \sum_{k=0}^{t} \Delta\hat{x}_k^{(i)}\,, \quad \text{for} \quad t \geq 0\,. \tag{S3.34}$$

We can now write (S3.32) as

$$\Delta\hat{x}_t^{(l)} = (R/U) \cdot \Delta\hat{x}_{t-1}^{(l)}, \quad \text{for} \quad t < l \tag{S3.35}$$

which, given our initial condition $\Delta\hat{x}_0^{(l)} = 0$, has solution $\Delta\hat{x}_t^{(l)} = 0$ for any $t < l$ and, consequently, $\hat{x}_t^{(l)} = 0$ for any $t < l$. Thus, we have shown by induction that $\hat{x}_t^{(i)} = 0$ is true for any $i \geq 1$ and any $0 \leq t < i$.

□

To complete our proof by induction, we will now consider the case when $t \geq l$. The induction hypothesis (S3.30) for time $t-1$ is

$$\hat{x}_{t-1}^{(l-1)} = \begin{cases} 0 & , \quad \text{if} \quad t-1 < l-1 \\ J \cdot (F/U)^{l-2} \cdot \displaystyle\sum_{k=l-1}^{t-1} \binom{k-1}{l-2} (R/U)^{k-(l-1)} & , \quad \text{if} \quad t-1 \geq l-1 \end{cases} \tag{S3.36}$$

and for time $t-2$ is

$$\hat{x}_{t-2}^{(l-1)} = \begin{cases} 0 & , \quad \text{if} \quad t-2 < l-1 \\ J \cdot (F/U)^{l-2} \cdot \displaystyle\sum_{k=l-1}^{t-2} \binom{k-1}{l-2} (R/U)^{k-(l-1)} & , \quad \text{if} \quad t-2 \geq l-1\,. \end{cases} \tag{S3.37}$$

Thus, whether $t = l$ or $t > l$, the hypothesis (S3.30) leads to

$$\hat{x}_{t-1}^{(l-1)} - \hat{x}_{t-2}^{(l-1)} = J \cdot (F/U)^{l-2} \cdot \binom{t-2}{l-2} (R/U)^{t-l}\,, \quad \text{for} \quad t \geq l\,. \tag{S3.38}$$

Plugging this into (S3.31), we obtain

$$\hat{x}_t^{(l)} = \hat{x}_{t-1}^{(l)} + R \cdot \frac{\hat{x}_{t-1}^{(l)} - \hat{x}_{t-2}^{(l)}}{U} + J \cdot (F/U)^{l-1} \cdot \binom{t-2}{l-2} (R/U)^{t-l}, \quad \text{for} \quad t \geq l \tag{S3.39}$$

which we can write as

$$\Delta\hat{x}_t^{(l)} = (R/U) \cdot \Delta\hat{x}_{t-1}^{(l)} + J \cdot (F/U)^{l-1} \cdot \binom{t-2}{l-2} (R/U)^{t-l} \tag{S3.40}$$

for $t \geq l$, where we used the definition (S3.33). This is a non-homogenous equation and we can solve it by using initial condition $\Delta\hat{x}_{l-1}^{(l)} = 0$, since we have shown above that $\Delta x_t^{(l)} = 0$ for $t < l$. Thus, we can find that the solution to (S3.40) is:

$$\Delta\hat{x}_t^{(l)} = J \cdot (F/U)^{l-1} \cdot \binom{t-1}{l-1} (R/U)^{t-l}, \quad \text{for} \quad t \geq l\,. \tag{S3.41}$$

To show that (S3.41) is indeed the solution to (S3.40), we can plug in $\Delta\hat{x}_{t-1}^{(l)}$ as given by (S3.41) into (S3.40) and obtain:

$$\begin{aligned} \Delta\hat{x}_t^{(l)} &= J \cdot (F/U)^{l-1} \cdot \left[\binom{t-2}{l-1} + \binom{t-2}{l-2} \right] (R/U)^{t-l} \\ &= J \cdot (F/U)^{l-1} \cdot \binom{t-1}{l-1} (R/U)^{t-l} \quad \square \end{aligned} \tag{S3.42}$$

To find $\hat{x}_t^{(l)}$ for $t \geq l$, we now only have to use (S3.34), which becomes

$$\hat{x}_t^{(l)} = \sum_{k=l}^{t} \Delta \hat{x}_k^{(l)} \,, \tag{S3.43}$$

since we already showed that $\Delta \hat{x}_t^{(l)} = 0$ for $t < l$. Thus, we obtain

$$\hat{x}_t^{(l)} = J \cdot (F/U)^{l-1} \cdot \sum_{k=l}^{t} \binom{k-1}{l-1} (R/U)^{k-l} \,, \quad \text{for} \quad t \geq l \tag{S3.44}$$

which completes our proof by induction.

□

Recovering solution for $R = 0$. Note that if $R = 0$, the solution given by (S3.28) becomes

$$\hat{x}_t^{(i)} = \begin{cases} 0 & , \quad t < i \\ J \cdot (F/U)^{i-1} & , \quad t \geq i \end{cases} \tag{S3.45}$$

which matches what we found for a sequence without recurrent connections.

Condition 1 emerges from the linear estimate. We have derived a general solution for $\hat{x}_t^{(i)}$ in (S3.28), with $\hat{x}_t^{(i)} = x_t^{(i)}$ for $i \leq m$ and $t \leq \bar{t}_m$. Thus, we know that m and $\bar{t}_m$ will correspond to the lowest i and t for which the following condition is met:

$$(J/U) \cdot (F/U)^{i-1} \sum_{k=i}^{t} \binom{k-1}{i-1} (R/U)^{k-i} \geq 1 \tag{Condition 1}$$

The limit of Condition 1 for infinitely long sequences

Let us start by writing out Condition 1 for $i = q$ and $t = q/s$:

$$(J/U) \cdot (F/U)^{q-1} \sum_{k=q}^{q/s} \binom{k-1}{q-1} (R/U)^{k-q} \geq 1 \tag{S3.46}$$

which can be rearranged as

$$(F/U) \left(\sum_{k=q}^{q/s} \binom{k-1}{q-1} (R/U)^{k-q} \right)^{1/q} \geq \sqrt[q]{F/J} \tag{S3.47}$$

Note that we have assumed q/s is an integer. To be rigorous, we could have explicitly written $\lfloor q/s \rfloor$, which denotes the integer value just under q/s. However, as we are interested in the limit $q \to \infty$, we assume $q/s \approx \lfloor q/s \rfloor$ for simplicity.

We first look at the right-hand side of (S3.47) and note that

$$\lim_{q \to \infty} \sqrt[q]{F/J} = 1 \,, \tag{S3.48}$$

since F and J are positive constants. So, the limit of Condition 1 will not depend on J. Although the left-hand side of (S3.47) has a similar power $1/q$, its limit is less trivial to determine because the term in parentheses depends on q. Let us first define the function

$$S(x) := \sum_{k=q}^{q/s} \binom{k-1}{q-1} \cdot x^{k-q} \tag{S3.49}$$

To find the limit of Condition 1 we need to find

$$\lim_{q \to \infty} \left[S(R/U) \right]^{1/q} \,. \tag{S3.50}$$

We have defined S as a sum of terms over the integer indices

$$\{k \in \mathbb{Z} \mid q \leq k \leq q/s\} \tag{S3.51}$$

For convenience, we will now define it in terms of a new non-integer index m and write

$$S(x) = \sum_{m \in M} \binom{m \cdot q - 1}{q - 1} \cdot x^{(m-1)q}\,, \tag{S3.52}$$

where

$$M = \left\{ \frac{k}{q} \mid k \in \mathbb{Z}\,,\, q \leq k \leq q/s \right\}\,. \tag{S3.53}$$

Let us consider each term in the sum

$$T_m^{(q)}(x) := \binom{m \cdot q - 1}{q - 1} \cdot x^{(m-1)q}\,. \tag{S3.54}$$

We can write this as

$$T_m^{(q)}(x) = \frac{(mq-1)!}{(q-1)! \cdot ((m-1)q)!} \cdot x^{(m-1)q}\,, \tag{S3.55}$$

where we used

$$\binom{n}{p} = \frac{n!}{p!\,(n-p)!}\,. \tag{S3.56}$$

The factorials in the above expression can be written using Stirling's formula, i.e.,

$$n! = n^n e^{-n} P(\sqrt{n})\,, \quad \text{with} \quad P\left(\sqrt{n}\right) = \sqrt{2\pi n}\left(1 + \frac{1}{12n} + \mathcal{O}\left(\frac{1}{n^2}\right)\right)\,. \tag{S3.57}$$

Note that $P\left(\sqrt{n}\right)$ is a polynomial that at most grows with $\sqrt{n}$. By doing this, we obtain

$$\begin{aligned} T_m^{(q)}(x) &= \frac{(mq-1)^{mq-1}}{(q-1)^{(q-1)}\left((m-1)q\right)^{(m-1)q}} \cdot \frac{P(\sqrt{mq-1})}{P(\sqrt{q-1}) \cdot P(\sqrt{(m-1)q})} \cdot x^{(m-1)q} \\ &= \left[\frac{(m-1/q)^m}{1-1/q} \cdot \left(\frac{x}{m-1}\right)^{(m-1)}\right]^q \cdot \tilde{P}\,, \end{aligned} \tag{S3.58}$$

where in the second line we rearranged the terms and represented the fraction of polynomials in the first equation by the symbol $\tilde{P}$. Let us analyze the term inside the brackets, which we will define as

$$T_m^{(q)}(x) = \left[A_m^{(q)}(x)\right]^q \cdot \tilde{P}\,, \quad A_m^{(q)}(x) := \frac{(m-1/q)^m}{1-1/q} \cdot \left(\frac{x}{m-1}\right)^{(m-1)}\,. \tag{S3.59}$$

Since $1 \leq m \leq 1/s$, we first calculate

$$\lim_{m \to 1} A_m^{(q)}(x) = 1\,. \tag{S3.60}$$

To determine if $A_m^{(q)}$ is growing or decreasing with m, we want to calculate the sign of its derivative with respect to m. Since $A_m^{(q)}$ is always non-negative, we can do this by looking at the simpler derivative of its logarithm

$$\frac{\partial}{\partial m} \log\left(A_m^{(q)}(x)\right) > 0 \tag{S3.61}$$

$$\Leftrightarrow \quad \log\left(\frac{m-1/q}{m-1} \cdot x\right) > 0$$

$$\Leftrightarrow \quad \left(m < \frac{1-x/q}{1-x} \wedge x < 1\right) \vee \left(m > -\frac{1-x/q}{x-1} \wedge x \geq 1\right) \tag{S3.62}$$

Here, we consider $1 - x/q > 0$ since we are interested in large values of q. Thus, we find that, for $x < 1$, $A_m^{(q)}$ is monotonously increasing until a maximum is reached at $(1-x/q)/(1-x)$. For $x > 1$, $A_m^{(q)}$ is

always increasing. Thus, we have found that there are always terms inside the brackets in (S3.58) which are larger than 1. This means that, as $q \to \infty$, $S(x)$ will be dominated by the $T_m^{(q)}(x)$ with the largest $A_m^{(q)}$, such that

$$\lim_{q\to\infty} [S(R/U)]^{1/q} \approx \max_{m\in M} \left\{ m^m \cdot \left(\frac{R/U}{m-1}\right)^{(m-1)} \right\} . \tag{S3.63}$$

In this limit, the terms $1/q$ in (S3.58) have vanished, as have the polynomial terms, since

$$\lim_{q\to\infty} \tilde{P}^{1/q} = 1 . \tag{S3.64}$$

Thus, to find the limit of Condition 1, we need to find the value of m which maximizes

$$m^m \cdot \left(\frac{R/U}{m-1}\right)^{(m-1)} . \tag{S3.65}$$

From our analysis of $A_m^{(q)}(x)$, we know that, if $R > U$, the maximum value will be located at $m = 1/s$, which is the highest m can be. For $R < U$, the maximum will be located at $m = 1/(1 - R/U)$. However, this value might be outside the domain $1 \leq m \leq 1/s$. If that is the case, i.e.

$$\frac{1}{1-R/U} > 1/s \Leftrightarrow R/U > 1-s \tag{S3.66}$$

then the maximizing m will again be $1/s$. However, if

$$\frac{1}{1-R/U} \leq 1/s \Leftrightarrow R/U \leq 1-s \tag{S3.67}$$

then the maximizing m will be $1/(1 - R/U)$, which is a local maximum. If we insert these two conditions, and respective maximizing values $m = 1/s$ or $m = 1/(1 - R/U)$, into (S3.63) we get

$$\lim_{q\to\infty} [S(R/U)]^{1/q} \approx \begin{cases} \frac{1}{1-R/U} & , \quad \text{for} \quad R/U \leq 1-s \\ (1/s)\left(\frac{R/U}{1-s}\right)^{(1/s-1)} & , \quad \text{for} \quad R/U > 1-s \end{cases} \tag{S3.68}$$

Plugging this into (S3.47), we finally get that the limit of Condition 1 when $q \to \infty$ is approximately

$$F/U \geq \begin{cases} 1-R/U & , \quad \text{for} \quad R/U \leq 1-s \\ s\left(\frac{1-s}{R/U}\right)^{(1/s-1)} & , \quad \text{for} \quad R/U > 1-s \end{cases} \tag{S3.69}$$

Note that in the limit of infinitely slow propagation, i.e., $\lim_{s\to 0}$, the condition becomes

$$F/U \geq \begin{cases} 1-R/U & , \quad \text{for} \quad R/U \leq 1 \\ 0 & , \quad \text{for} \quad R/U > 1 \end{cases} \tag{S3.70}$$

which can simply be written as

$$F + R \geq U \tag{S3.71}$$

since both R and F are non-negative parameters. This is what we saw in the main text as Condition 2, a necessary condition for replay for any distribution shape.

A general condition for sequence retrieval

Let us assume that some assembly m became fully activated at some time $\bar{t}_m$, such that $\sum_{j=0}^{\bar{t}_m} a_j^{(i-1)} = 1$. Following (S3.5), the position of the right edge of the subsequent assembly $m+1$ at any time step $t > \bar{t}_m$ is given by

$$x_t^{(m+1)} = R \cdot \sum_{k=0}^{t-1} a_k^{(m+1)} + F . \tag{S3.72}$$

If we assume assembly i has not yet been fully activated, we can use (S3.9), such that (S3.72) becomes

$$\begin{aligned} x_t^{(m+1)} &= R \sum_{k=1}^{t-1} \frac{x_k^{(m+1)} - x_{k-1}^{(m+1)}}{U} + F \\ &= (R/U) \cdot \left(x_{t-1}^{(m+1)} - x_0 \right) + F \\ &= F + (R/U) \cdot x_{t-1}^{(m+1)}, \quad \text{while} \quad x_t^{(m+1)} < U \end{aligned} \tag{S3.73}$$

which is a first-order inhomogeneous difference equation.

If $R = U$, (S3.73) becomes

$$x_t^{(m+1)} = F + x_{t-1}^{(m+1)}, \tag{S3.74}$$

where $x_t^{(m+1)}$ increases at a constant rate F and can thus certainly cross U within a finite number of time steps.

If $R \neq U$, we have to find the solution to the inhomogeneous difference equation. The general solution of the inhomogeneous difference equation (S3.73) is given by

$$x_t^{(m+1)} = x_t^{(m+1),h} + x_t^{(m+1),p} \tag{S3.75}$$

where $x_t^{(m+1),h}$ is the general solution to the corresponding homogeneous equation

$$x_t^{(m+1),h} = (R/U) \cdot x_{t-1}^{(m+1),h} \tag{S3.76}$$

and $x_t^{(m+1),p}$ is a particular solution to the inhomogeneous equation (S3.73). The general solution of the homogeneous equation is trivial and given by

$$x_t^{(m+1),h} = K \cdot (R/U)^t \tag{S3.77}$$

where $K \in \mathbb{R}$ is some constant value. To find the particular solution, we attempt a constant solution $x^{(m+1),p}$ that solves (S3.73), such that

$$\begin{aligned} x^{(m+1),p} &= F + (R/U) \cdot x^{(m+1),p} \\ \Leftrightarrow x^{(m+1),p} &= \frac{F}{1 - R/U} \end{aligned}$$

Thus, we have found the general solution for (S3.73) when $R \neq U$, which is given by

$$x_t^{(m+1)} = \frac{F}{1 - R/U} + K \cdot (R/U)^t \tag{S3.78}$$

To determine the value of constant K we would need to use initial conditions. However, our equation (S3.73) only applies for $t > \bar{t}_m$ and we cannot use the value of $x_{\bar{t}_m}^{(m+1)}$ as an initial condition, since this is also unknown. Thus, we cannot find the exact value of K. Nevertheless, we can determine whether it is a negative or positive constant, depending on the value of R. According to (S3.78),

$$x_t^{(m+1)} - x_{t-1}^{(m+1)} = K \cdot \left((R/U)^t - (R/U)^{t-1} \right) = K \cdot (R/U)^{t-1} \cdot (R/U - 1) \tag{S3.79}$$

We know that this quantity has to be non-negative since, by definition, $x_t^{(m+1)}$ can never decrease. Thus, if $R > U$, it follows that $K \geq 0$ must be true. We can write (S3.78) for $R > U$ as

$$x_t^{(m+1)} = \frac{F}{1 - R/U} + C_1 \cdot (R/U)^t, \quad C_1 \geq 0 \tag{S3.80}$$

It is then trivial to see that $x_t^{(m+1)}$ will grow at an increasing rate and can reach U within a finite number of time steps.

On the other hand, when $R < U$, equation (S3.79) tells us that for $x_t^{(m+1)}$ never to decrease, K must be a non-positive constant. In that case, we can write (S3.78) as

$$x_t^{(m+1)} = \frac{F}{1 - R/U} - C_2 \cdot R^t, \quad C_2 \geq 0 \tag{S3.81}$$

We thus see that, if $R < U$, $x_t^{(m+1)}$ is a monotonically increasing function that asymptotically tends to the limit $F/(1 - R/U)$ as $t \to \infty$. This result tells us that, if assembly m crosses the threshold at some time $\bar{t}_m$, assembly $m + 1$ will become fully activated if $F/(1 - R/U) > U \Leftrightarrow F + R > U$, if given enough time steps. If $F + R = U$, the assembly will never become fully activated within a finite number of time steps.