Recall tempo of Hebbian sequences depends on the interplay of Hebbian kernel with tutor signal timing

Matthew Farrell\textsuperscript{a,b} and Cengiz Pehlevan\textsuperscript{a,b,c}

\textsuperscript{a}John A. Paulson School of Engineering and Applied Sciences, Harvard University; \textsuperscript{b}Center for Brain Science, Harvard University; \textsuperscript{c}Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University

Abstract
Understanding how neural circuits generate sequential activity is a longstanding challenge. While foundational theoretical models have shown how sequences can be stored as memories with Hebbian plasticity rules, these models considered only a narrow range of Hebbian rules. Here we introduce a model for arbitrary Hebbian plasticity rules, capturing the diversity of spike-timing-dependent synaptic plasticity seen in experiments, and show how the choice of these rules and of neural activity patterns influences sequence memory formation and retrieval. In particular, we derive a general theory that predicts the speed of sequence replay. This theory lays a foundation for explaining how cortical tutor signals might give rise to motor actions that eventually become “automatic”. Our theory also captures the impact of changing the speed of the tutor signal. Beyond shedding light on biological circuits, this theory has relevance in artificial intelligence by laying a foundation for frameworks whereby slow and computationally expensive deliberation can be stored as memories and eventually replaced by inexpensive recall.

Introduction
An important class of animal behaviors are those that are consolidated into “automatic”, well-practiced sequential routines (1–9). An analogy is the process of learning a tennis serve, which starts as a slow and deliberate process, but eventually becomes nearly muscle-memory. Importantly, the behavior can be thought of as a sequence that, once initialized in a starting state, progresses in a highly stereotyped, automatic fashion. From this perspective, the behavior can be thought of as a stored sequence memory that is recalled by an initial prompt.

For such sequence memories, the timing of the sub-elements of these sequences and overall tempo are essential components (10–16). It is therefore important to understand how temporally specific sequences can be learned, memorized, and recalled.

To answer this question, we need to look at the mechanisms of learning and generation of sequential behavior. One common perspective holds that these behaviors are generated by stereotyped sequential neural activity (17–21). Further, several experiments showed a tight correlation between neural activity and timing of behavior (22–30). This suggests that automatic sequential behavior may at least in part be “stored” in the synapses of a neural circuit, such that the behavior can be generated by setting the network to a state that corresponds to the first point in the sequence.

Despite the importance of these topics to understanding brain function, it is unclear what actually determines the tempo of sequence generation in neural circuits, and how this is connected to the learning process. Here we hypothesize that the tempo arises from an interaction between the temporal dependence of the synaptic learning rules and the temporal structure of the network activity when learning the sequential behavior. We refer to the mechanism that sets network activity during sequence learning as the “tutor signal”; this tutor signal may come from a higher-order brain area such as motor cortex, or from minimally processed sensory inputs.

This perspective is embodied by models featuring temporally asymmetric Hebbian (TAH) learning rules. Despite the long and illustrious legacy of this class of models (31–41), previous studies were seriously limited by the very narrow class of learning rules that these preexisting models can describe, and the fact that these learning rules do not correspond with those observed in the brain. In this work we extend the basic TAH models previously considered to a much richer class of models. Not only do we derive a theory that describes the essential elements of sequence dynamics, including tempo and noise robustness, for learning rules currently seen in the brain; our theory describes a wide class of TAH learning rules that gives rise to sequential dynamics, covering the case of new learning rules that may be discovered in biology in the future.

TAH network models are built upon Hebbian learning rules, wherein the network state is set by the tutor to a sequence of activity patterns one after another. These activity patterns correspond to moments in behavioral sequences, and the Hebbian learning rules store the structure of these sequences in the network as a memory. The sequences can then be recalled by setting the network to the first state in a desired sequence, resulting in the execution of the now “automatic”, well-practiced sequential behavior. These models can be seen as a generalization of Hopfield networks (42) to storing sequences rather than static patterns; therefore, many of the motivations and rationale behind the Hopfield model is applicable here, such as the ability to “complete” a sequence based on partial inputs (see Discussion for more in-depth interpretations of this model).

While our theory is very general, we demonstrate its applicability and utility by focusing on two fundamental and biologically-relevant cases: (1) TAH learning rules with fast timescales relative to the tutor signal and (2) Double-sided exponential TAH learning rules as are commonly seen in biology (43). In both cases we find useful principles that are interpretable in the context of biology. For example, weight structures (and TAH learning rules) can be adjusted such that the tempo remains the same, but robustness to noise increases. We also show how networks with the same qualitative dynamical properties can arise from very different weight structure. For double-sided exponential TAH kernels, we show that the tempo of sequence recall depends linearly on the speed of the tutor signal, with a slope that depends in a simple way on...
Fig. 1. Illustration of network model and stereotypical behavior. a) The interaction of a tutor signal (the $\xi^\mu$) with Hebbian temporal kernel $w(\delta t) = w(s - t)$ gives rise to weight matrix $W$. b) Illustrations of different possible kernels. Left: kernel that would give rise to two terms for each $\mu$, $a_0^\mu \xi^\mu(\xi^\mu)^T$ and $a_1^\mu \xi^{\mu+1}(\xi^\mu)^T$. Second from left: kernel that would give rise to three terms for each $\mu$, $a_0^\mu \xi^\mu(\xi^\mu)^T$, $a_1^\mu \xi^{\mu+1}(\xi^\mu)^T$ and $a_2^\mu \xi^{\mu-1}(\xi^\mu)^T$. Third from left: kernel that would give rise to five terms for each $\mu$, corresponding to $\mu - 2$, $\mu - 1$, $\mu$, $\mu + 1$, and $\mu + 2$. Right: double-sided exponential kernel. c) Activity $r_\mu$ of ten randomly selected units in the network for coefficient values $a_0 = 0$ and $a_1 = 1$. d) Overlaps $q_\mu$ of the network activity with patterns $\xi^\mu$ (here $a_0 = 0$ and $a_1 = 1$). Color corresponds to pattern index $\mu$. The vertical lines indicate the locations $t_\mu$ of the peaks. The first overlap $q_1$ is not included in the plot for ease of visualization.

The sequence is stored and produced by a recurrently connected network of $N$ neurons. The connectivity is given by a recurrent weight matrix $W$, where $W_{ij}$ is the connectivity strength of neuron $j$ to neuron $i$. The dynamics of the network are then given by

$$\dot{r} = -r + \phi(Wr + \eta(t))$$  \hspace{1cm} [1]$$

where $\eta(t)$ is white noise ($\langle \eta(t) \rangle = 0$ and $\langle \eta(t)\eta(t') \rangle = \rho^2 \delta(t-t')$) and $\phi$ is a sigmoidal nonlinearity (see Appendix I for more details). Note that it is common to add a timescale factor $\tau$ which multiplies the term $\dot{r}$ and $\sqrt{\tau}$ that multiplies $\eta(t)$, which have the effect of rescaling the timescale of the dynamics. Since this effect is relatively simple to understand, we simply take $\tau = 1$ and focus our attention on other attributes.

The sequence is stored in this network by a tutor signal (Fig. 1a). Given a sequence of length $P$, this signal sets the state of the network sequentially to an ordered set of $P$ patterns $\langle \xi^\mu \rangle_{\mu=1}^P$. These $P$ patterns represent $P$ neural states that correspond to the states making up the sequential behavior. The recurrent network memorizes these patterns by synaptic plasticity at the recurrent synapses.

To store a memory of this sequence, this synaptic plasticity can take the form of a temporally asymmetric Hebbian (TAH) learning rule active during application of the tutor signal (31, 32, 37, 44). An example of a TAH learning rule commonly seen in the literature is

$$W_{ij} \leftarrow \frac{1}{N} \int_0^{T-\Delta t} dt \int_0^{T-\Delta t} ds w(s-t)r_i(t)r_j(t),$$ \hspace{1cm} [2]$$

where $\Delta t$ is an interval of time that spans the amount of time the network is placed in each pattern state and $T$ is the duration for the tutor signal associated with the sequence. With this learning rule, connections are strengthened according to the coincident firing of neurons for the current pattern with those of the next pattern; if weights are initially zero, this learning rule results in final weights

$$W = \frac{1}{N} \sum_{\mu=1}^{P-1} \xi^{\mu+1}(\xi^\mu)^T.$$ \hspace{1cm} [3]$$

We refer to this weight structure as sequi-associative. Intuitively, this set of synaptic weights calculates the overlap between the current network state with the stored patterns, and if the overlap with one of the patterns dominates over others (say pattern $\nu$), it steers the network activity to the pattern that is after the dominant pattern (pattern $\nu + 1$).

While sequi-associative weights capture some temporal aspects of synaptic plasticity, they fall short in their ability to express the full complexity of plasticity dynamics observed in experiments (43, 45, 46) which can significantly affect synaptic structure and network dynamics (47–49). Mathematically, a more general Hebbian learning rule could be written as

$$W_{ij} \leftarrow \frac{1}{N} \int_0^{T-\Delta t} dt \int_0^{T-\Delta t} ds w(s-t)r_i(s)r_j(t),$$ \hspace{1cm} [4]$$

where $w$ is a general kernel/filter that weights the Hebbian learning rule as a function of time offset. One sees that Eq. (2) corresponds to $w(s-t) = \delta(s-t-\Delta t)$. However, biologically observed kernels take much richer and different

the parameters of the kernel. While we focus here on applications to neuroscience, considering its generality our analysis may also be of interest to theorists in other domains studying dynamical systems with asymmetric weight structures.

Results

Sequence model definition

The sequence is stored and produced by a recurrently connected network of $N$ neurons. The connectivity is given by a recurrent weight matrix $W$, where $W_{ij}$ is the connectivity strength of neuron $j$ to neuron $i$. The dynamics of the network are then given by

$$\dot{r} = -r + \phi(Wr + \eta(t))$$  \hspace{1cm} [1]$$

where $\eta(t)$ is white noise ($\langle \eta(t) \rangle = 0$ and $\langle \eta(t)\eta(t') \rangle = \rho^2 \delta(t-t')$) and $\phi$ is a sigmoidal nonlinearity (see Appendix I for more details). Note that it is common to add a timescale factor $\tau$ which multiplies the term $\dot{r}$ and $\sqrt{\tau}$ that multiplies $\eta(t)$, which have the effect of rescaling the timescale of the dynamics. Since this effect is relatively simple to understand, we simply take $\tau = 1$ and focus our attention on other attributes.

The sequence is stored in this network by a tutor signal (Fig. 1a). Given a sequence of length $P$, this signal sets the state of the network sequentially to an ordered set of $P$ patterns $\langle \xi^\mu \rangle_{\mu=1}^P$. These $P$ patterns represent $P$ neural states that correspond to the states making up the sequential behavior. The recurrent network memorizes these patterns by synaptic plasticity at the recurrent synapses.

To store a memory of this sequence, this synaptic plasticity can take the form of a temporally asymmetric Hebbian (TAH) learning rule active during application of the tutor signal (31, 32, 37, 44). An example of a TAH learning rule commonly seen in the literature is

$$W_{ij} \leftarrow \frac{1}{N} \int_0^{T-\Delta t} dt \int_0^{T-\Delta t} ds w(s-t)r_i(t)r_j(t),$$ \hspace{1cm} [2]$$

where $\Delta t$ is an interval of time that spans the amount of time the network is placed in each pattern state and $T$ is the duration for the tutor signal associated with the sequence. With this learning rule, connections are strengthened according to the coincident firing of neurons for the current pattern with those of the next pattern; if weights are initially zero, this learning rule results in final weights

$$W = \frac{1}{N} \sum_{\mu=1}^{P-1} \xi^{\mu+1}(\xi^\mu)^T.$$ \hspace{1cm} [3]$$

We refer to this weight structure as sequi-associative. Intuitively, this set of synaptic weights calculates the overlap between the current network state with the stored patterns, and if the overlap with one of the patterns dominates over others (say pattern $\nu$), it steers the network activity to the pattern that is after the dominant pattern (pattern $\nu + 1$).

While sequi-associative weights capture some temporal aspects of synaptic plasticity, they fall short in their ability to express the full complexity of plasticity dynamics observed in experiments (43, 45, 46) which can significantly affect synaptic structure and network dynamics (47–49). Mathematically, a more general Hebbian learning rule could be written as

$$W_{ij} \leftarrow \frac{1}{N} \int_0^{T-\Delta t} dt \int_0^{T-\Delta t} ds w(s-t)r_i(s)r_j(t),$$ \hspace{1cm} [4]$$

where $w$ is a general kernel/filter that weights the Hebbian learning rule as a function of time offset. One sees that Eq. (2) corresponds to $w(s-t) = \delta(s-t-\Delta t)$. However, biologically observed kernels take much richer and different
forms, including the famous double-sided exponential STDP curve of Bi and Poo (43) (Fig. 1b, right). Note that the neuroscience literature often plots the kernel as a function of $-\delta t = -(s - t)$ instead of $\delta t = s - t$ (i.e. “pre-post” instead of the “post-pre” convention used here).

In this paper, we consider the learning of sequences under a tutor signal with the general TAH rule (Eq. (4)). This rule can give rise to a much richer set of weights defined by

$$W = \frac{1}{N} \sum_{\mu=1}^{P} \sum_{\nu=1}^{P} a_{\nu}^{\mu} (\xi^{\mu})^{\top}$$

where $a_{\nu}^{\mu}$ are coefficients that satisfy

$$a_{\nu}^{\mu} = \int_{t_{\nu}^{\mu}}^{T_{\nu}} dt \int_{t_{\nu}^{\mu} + t_{\nu}^{\mu}}^{t_{\nu}^{\mu} + t_{\nu}^{\mu}} ds w(s - t).$$

Here $t_{\nu}^{\mu}$ is the time at which the tutor signal for pattern $\nu$ begins and $T_{\nu}^{\mu} = t_{\nu}^{\mu} + t_{\nu}^{\mu}$ is the duration for which pattern $\mu$ is shown to the network. Our goal is to understand the sequential dynamics and its tempo arising from such learning rules.

We should note that other works have also considered extensions to Eq. (3). For example, some works (32, 34, 50, 51) include an autoassociative component $\xi^{\mu} (\xi^{\mu})^{\top}$ to the weights; (32, 34) use discrete dynamics that selectively low-pass filter the sequi-associative component, and (50) multiplies the sequi-associative term with random noise. The autoassociative term modulates the pattern transition speed. The authors of (40) conduct simulation studies of a model with an additional pre-associate term $\xi^{\mu-1} (\xi^{\mu})^{\top}$ that has the same strength and opposite sign as the sequi-associative component. (35) derive a discrete dynamics in terms of the parameters of an exponentially decaying kernel, giving rise to terms of the form $\xi^{\mu+k} (\xi^{\mu})^{\top}$ for $k \geq 0$; however, mathematical analysis of this model is limited to deriving the mean-field dynamics and computing some aspects of stability (there are no calculations related to tempo). (36) similarly consider terms of the form $\xi^{\mu+k} (\xi^{\mu})^{\top}$, but again focus on capacity and stability properties.

Going back to our model, to quantify the sequential behavior of the network dynamics, we define the overlaps

$$q_{\mu}(t) := \frac{1}{N} (\xi^{\mu})^{\top} r(t),$$

which measure the similarity of the recurrent network activity $r$ at time $t$ to stored patterns $\xi^{\mu}$. For our initial condition, we will always initialize the network state to the first pattern $\xi^{1}$. Note that the dynamics of the network in the basis of the neurons may not reveal the sequential structure (Fig. 1c), while the overlaps $q_{\mu}$ clearly display this structure (Fig. 1d).

To measure sequence progression, we track the location in time $t_{\nu}$ of the maxima of the overlaps $q_{\mu}(t)$; at time $t_{\nu}$ we say the sequence is in state $\mu$ (see vertical lines in Fig. 1d).* We also define $p_{\nu} := q_{\mu}(t_{\nu})$ to be the height of the peaks. We are particularly interested in the speed of the sequence progression, which we measure by the peak difference $d_{\nu} := t_{\nu} - t_{\nu-1}$. Note that the coefficients $a_{\nu}^{\mu}$ can inhomogeneously depend on $\mu$, resulting in a tempo of sequence recall that can be controlled based on the pattern state (i.e. the same sequence can have both fast and slow transitions between patterns, resulting in a $d_{\mu}$ that depends strongly on $\mu$). If the tutor signal presents patterns at regular intervals, then the coefficients take the form $a_{\nu}^{\mu} = a_{\nu-\mu}$. In this case the coefficient for the $\xi^{\mu} (\xi^{\mu})^{\top}$ term only depends on the difference between $\nu$ and $\mu$ rather than depending on $\mu$ independent of $\nu$:

$$W = \frac{1}{N} \sum_{\mu=1}^{P} \sum_{\nu=1}^{P} a_{\nu-\mu} (\xi^{\mu})^{\top} (\xi^{\nu})^{\top}$$

This means that the tempo can no longer be controlled in a sequence-state-dependent (i.e. $\mu$-dependent) manner. Even stronger, often the tempo is approximately independent of $\mu$ (it is a global property of the sequence), so that $d_{\mu}$ is approximately constant across $\mu$.

Sequence stability

Beside tempo, we are also interested in the stability properties of sequential activity. Note that since we are using a saturating nonlinearity $\phi$, each $r_{i}$ is bounded which implies that the overlap $q_{\mu}$ is also bounded. Hence we are primarily interested in whether or not the solutions $q_{\mu}(t)$ decay to 0 with increasing $\mu$, and we say that the sequence of solutions $(q_{\mu}(t))_{\mu}$ is stable if the solutions $q_{\mu}(t)$ do not decay to the zero solution $q(t) = 0$ with increasing $\mu$.$^\dagger$ Formally, in the $T \to \infty$ limit, we say that the sequence is stable if there exists a positive constant $\epsilon$ such that for all positive integers $\mu$, there exists a $s_{\mu}$ such that $q_{\mu}(s_{\mu}) > \epsilon$. Note that this definition requires that we take $\mu$ (and hence $T$) arbitrarily large. If $\mu$ were bounded in the definition of stability then almost every solution would be stable trivially.

Mean-field theory

To gain analytical insight about the behavior of our model, we extend the mean-field analysis of (37) to the general weights described by Eq. (5). We assume that the patterns are drawn identically and independently from a Gaussian distribution, $\xi^{\mu} \sim N(0,1)$, and take the network population size to infinity, $N \to \infty$.

In this mean-field limit the overlaps evolve with dynamics given by (see Appendix 1)

$$\dot{q}_{\mu} = -q_{\mu} + g(t) \sum_{\nu=1}^{P} a_{\nu}^{\mu} q_{\nu}$$

where

$$g(t) = G \left( \frac{||W r(t)||^2}{\mu^2} + \rho^2 \right)^{\dagger}$$

for some function $G$. We give the full functional form of $G(x)$ in Appendix 1 (Eq. (20)), which depends on the shape of $\phi$. For the analysis here, the important aspect of $G$ is the value of $G(\rho^2)$, which serves as an upper bound on $g(t)$. In particular, for $\rho = 0$, $g(t)$ is bounded from above by $G(0)$, and $G(0) = 20/\sqrt{2\pi} \approx 8$ for our choice of parameters for $\phi$.

The convergence with increasing $N$ of the mean-field solutions to the full network solutions is illustrated in Fig. 2. This

---

*We can also define the sequence state represented at a time $t$ by $\arg\max_{\mu} q_{\mu}(t)$, but the $t_{\nu}$ are more mathematically convenient objects to consider.

$^\dagger$Note the contrast to the dynamical systems literature, where stability typically means that the solutions remain finite.
Fig. 2. Comparison of full network simulations Eq. (1) (solid lines) with the mean-field equations Eq. (9) (dashed lines). Throughout, every fifth overlap \( q_n(t) \) is shown, starting with \( \mu = 2 \). a) Overlaps \( q_n(t) \) for network with \( N = 5000 \) neurons. The only nonzero coefficient is \( a_1 = 1.5 \). b) As in a, but for \( N = 20000 \) neurons. c) As in a, but the coefficients are \( a_\mu^\nu = 0 \) and \( a_\mu^\nu = 1.3 \) for \( \mu < 11 \); and \( a_0^\nu = -0.1 \) and \( a_1^\nu = 1.5 \) for \( \mu \geq 11 \). d) As in c but for \( N = 20000 \).

We validate this assumption in simulations below. Assuming that way. To see this, we first use Eq. (6):

For the following, we assume that tutor signal overlaps are uniform in temporal duration, so \( a_\mu^\nu = a_{\mu\nu} \). We will revisit the case of non-uniform intervals in the final subsection of Results.

We next analyze the mean-field theory defined by Eq. (9) to gain insight into the stability properties of sequence generation. To analyze Eq. (9), we use the assumption that \( g(t) \) asymptotes to a constant value over time, and define \( g_\infty = \lim_{t \to \infty} g(t) \). We validate this assumption in simulations below.

A fundamental quantity in determining sequence dynamics is \( \tilde{g} := \left( \sum_{k=0}^{\nu-1} a_k \right)^{-1} \), where the sum is over all coefficients. For the following we consider the case of large \( N \), and reiterate that these analyses treat the case of uniform tutor signal intervals. In Appendix 2 we show that in this case, and under other mild assumptions, the solutions to the network dynamics Eq. (1) are stable if and only if \( g_\infty = \tilde{g} \). Further, we reason and observe that the network will endeavor to satisfy this stability condition, so that \( \tilde{g} < G(\rho^2) \) is necessary and sufficient for \( g_\infty = \tilde{g} \). Motivated by this, we define \( \tilde{a}_k := \tilde{g} a_k \) and consider the linear equations

\[
\dot{q}_\mu = -q_\mu + \sum_{\nu=1}^{\nu} \tilde{a}_{\mu\nu} q_\nu, \quad [10]
\]

Eq. (10) can be approximated by Eq. (9) when the solutions to Eq. (9) are stable (equivalently, when \( \tilde{g} < G(\rho^2) \)). When the solutions to Eq. (10) are unstable (when \( \tilde{g} > G(\rho^2) \)), these solutions will vanish with increasing \( \mu \) while the solutions to Eq. (10) will remain stable. We will test the goodness of the linear approximation Eq. (10) to Eq. (9) in simulations.

Note that \( \tilde{g} \) can be related to the TAH kernel \( w \) in a simple way. To see this, we first use Eq. (6):

\[
\tilde{g}^{-1} = \sum_k a_k = \int_0^{T_\xi} dt \int_{-\infty}^{T_\xi/2} ds \ w(s-t), \quad P \text{ even}
\]

\[
\tilde{g}^{-1} = \int_0^{T_\xi} dt \int_{-T_\xi/2}^{T_\xi/2} ds \ w(s-t), \quad P \text{ odd}
\]

Assuming that \( w(\delta t) \) becomes vanishingly small for large \( |\delta t| \) and that \( T \gg T_\xi \), the inner integral can be approximately replaced with one over the real line:

\[
\tilde{g}^{-1} \approx T_\xi \int_{-\infty}^{\infty} ds \ w(s-t).
\]

Note that this equation means the network can be stabilized by extending the duration of the tutor signal \( T_\xi \). Therefore, a biologically-plausible solution to embedding stable sequences is by extending the tutor signal duration. As we will show below, increasing \( \tilde{g}^{-1} \) also increases the robustness of the sequence dynamics to noise.

One forward term

We next turn to analyzing sequence progression in the recurrent neural network. We start by considering the case with one forward term of Eq. (10), where \( a_\mu = 0 \) for \( k \notin [0,1] \). This case is relevant in the regime where the duration of sequence elements, \( T_\xi \), is significantly longer than the timescale of \( w(t) \) (i.e., \( w(\delta t) \approx 0 \) for \( \delta t > T_\xi \) and \( w(\delta t) = 0 \) for \( \delta t < 0 \)). Figure 3a illustrates two possible kernels \( w \) corresponding to this case.

In this case Eq. (10) becomes

\[
\dot{q}_\mu = \begin{cases} -q_\mu + \tilde{a}_0 q_\mu + \tilde{a}_1 q_{\mu-1}, & \mu > 1 \\ -q_\mu + \tilde{a}_0 q_\mu, & \mu = 1 \end{cases}
\]

The solutions to these equations are (recall the initial conditions \( q_\mu(0) = \delta(\mu - 1) \))

\[
q_\mu(t) = \frac{(\tilde{a}_1)^{\mu-1} \tilde{a}_0^{-1}}{(\mu - 1)!} e^{(\tilde{a}_0-1)t}
\]

which have maxima at

\[
t_\mu = \frac{\mu - 1}{1 - \tilde{a}_0}. \quad [11]
\]

Note that \( \tilde{a}_0 + \tilde{a}_1 = 1 \) by definition, which places some restrictions on the values that \( \tilde{a}_0 \) and \( \tilde{a}_1 \) can take. For instance, \( \tilde{a}_0 > 1 \) implies \( \tilde{a}_1 < 1 \). This ensures that \( 0 < q_\mu(t) \leq 1 \) for all choices of \( \tilde{a}_0 \) and \( \tilde{a}_1 \). The (inverse) speed as measured by peak difference is

\[
d_\mu = \frac{1}{1 - \tilde{a}_0} + 1 = \frac{\tilde{a}_0}{\tilde{a}_1}.
\]

Figure 3 shows the simulation results for various values of \( \tilde{a}_0, \tilde{a}_1, \) and \( \rho \). Figure 3b plots peak times \( t_\mu \) for \( \rho = 0 \). The full network simulations with \( N = 30000 \) match the linear approximations Eq. (11) closely, except for parameter values where the solutions for the full network decay to zero (denoted by C). This happens because \( \tilde{g} \) is larger than the critical value of \( G(0) \approx 8 \). Note the significant role of \( \tilde{a}_0 \) in determining the sequence speed. In particular, for \( \tilde{a}_0 = 0 \) changing \( \tilde{a}_1 \) has no impact on the speed (\( \tilde{a}_1 \) symbol in Fig. 3b). Note also that, unlike in standard Hopfield models, the autoassociative term \( \tilde{a}_0 \) can be negative. The bottom panel of Fig. 3a shows an
example of a kernel that gives rise to negative $a_0$ and positive $a_1$.

Figure 3c plots the function $g(t)$. This plot supports our original assumption that $g(t)$ asymptotes to a constant value, and shows the close match between $g_{\text{asc}}$ and $\bar{g}$, except in the case where $\bar{g} > G(0) \approx 8$.

To look more closely at the relative impacts of $a_0$ and $a_1$ on sequence speed, we plot sequence speed for different combinations of $a_0$ and $a_1$ in Fig. 3d. Figure 3d shows that the dependence of $d_\mu$ on $a_0$ is roughly linear in the range plotted, with a slope that decreases with $a_1$. Parameter ranges were chosen in an attempt to capture the full behavior of the dynamics within the region of stable sequence generation; in particular, this requires that $a_1 > 0$ and $a_0 + a_1 > 1/G(0)$. Note that for $a_0 < 0$ sequence speed decreases, counter-intuitively, with increasing $a_1$, while the opposite relationship holds for $a_0 > 0$. This can be seen from the expression for $d_\mu$ above.

Figure 3e quantifies the stability properties of sequences by plotting the peak magnitude $p_\mu$ for a pattern of large index (here $\mu = 70$) as a function of $\bar{g}^{-1} = a_0 + a_1$. Color denotes the strength of noise injected into the network dynamics ($\rho$ in Eq. (1)). Note that such noise could come from interference, such as that which arises if there are multiple stored sequences (37). The dashed vertical lines mark the critical values of stability where $\bar{g} = G(\rho^2)$. This plot illustrates how $\bar{g}$ determines the stability of sequences, and that smaller $\bar{g}$ (equiv. larger $a_0 + a_1$) is required for stability in the presence of noise. The plot shows that, as predicted, peak magnitudes $p_\mu$ have a sharp inflection point at $\bar{g} = G(\rho^2)$, where $p_\mu$ goes from being approximately zero to nonzero.

Note that $a_0$ and $a_1$ can be adjusted so that sequence progression speed $d_\mu$ remains constant, but stability is improved by decreasing $\bar{g}$. This can most easily be seen by considering lines of constant $d_\mu$ in Fig. 3d. The most straightforward example is the line $d_\mu = 1$ in the right panel of Figure 3d, which clearly corresponds to many values of $a_0 + a_1$. This is particularly important in the presence of noise, as shown by Fig. 3e. This degree of freedom of the system makes some kernels $w(t)$ strictly better than others, even if they result in the same speed.

**General approximate solution**

While we were able to explicitly solve Eq. (10) in the case of two nonzero terms in the preceding section, in general an explicit solution is not available and we need to resort to approximations. To approximately solve Eq. (10), we replace the equation with one that is periodic in $\mu$. A general approximate solution to the equations with these periodic boundary conditions can then be found (see Appendix 2):

$$t_\mu \approx \frac{\mu - 1}{\alpha} - \frac{\beta}{2\alpha^2}, \quad \mu > 1$$  [12]

where $\alpha = \sum_{k=1}^{\lfloor p/2 \rfloor} (\bar{a} + \bar{\alpha})k$ and $\beta = \sum_{k=1}^{\lfloor p/2 \rfloor} (\bar{a} + \bar{\alpha})k^2$.

Sequence speed as measured by peak difference is then

$$d_\mu \approx \frac{1}{\alpha}, \quad \mu > 2$$  [13]

When $\alpha > 0$, these equations to describe forward propagation of the sequence.

In the following sections we will look at special cases of interest that illustrate the theory.
Three terms with bidirectional connectivity

Next we consider the case with three terms and bidirectional connectivity in Eq. (10), where $a_k = 0$ for $k \notin \{-1, 0, 1\}$. Bidirectional connectivity occurs with TAH kernels commonly found in biology, such as the double-sided exponential decay kernels (52) (see Fig. 4a for an illustration). Three terms in particular are relevant when the TAH kernel’s timescale is faster than the tutor signal’s, for instance when the support of the kernel vanishes outside of time windows spanning more than two pattern presentations. Figure 4a illustrates a temporally asymmetric kernel that would give rise to such terms.

To understand sequence progression speed, we use Eq. (12):

$$t_\mu \approx \frac{\mu - 1}{a_1 - a_{-1}} = \frac{\bar{a}_1 + a_1}{2(a_1 - a_{-1})^2}.$$ [14]

Sequence speed as measured by peak difference is then

$$d_\mu \approx \frac{1}{a_1 - a_{-1}} = \frac{a_0 + a_1}{a_1 - a_{-1}}.$$ [15]

The effect of changing $a_{-1}$ on sequence speed is revealed by taking the derivative of $d_\mu$ with respect to $a_{-1}$:

$$\frac{d}{da_{-1}} d_\mu \approx -\frac{a_0 + 2a_1}{(a_1 - a_{-1})^2}.$$ [16]

This shows that increasing $a_{-1}$ decreases $d_\mu$ provided $a_0 + 2a_1 > 0$. To compare the relative impact of changing $a_{-1}$ and $a_1$ on sequence speed, we can take the ratio of derivatives:

$$\frac{dd_\mu}{da_1} / \frac{dd_\mu}{da_{-1}} \approx -\frac{a_0 + 2a_1}{a_0 + 2a_1}.$$ [17]

This equation reveals that changes in $a_{-1}$ will impact sequence speed more than changes in $a_1$ when $a_{-1}$ is smaller in magnitude than $a_1$. Hence $a_{-1}$ is a natural parameter to use to control speed.

Figure 4 compares the full network simulations, the simulations of the linear system Eq. (10), and the approximation Eq. (14). The peak times $t_\mu$ are plotted in Fig. 4b, showing that Eq. (14) is a good estimate of both the linear and nonlinear systems. Figure 4c shows the relationship between $d_\mu$, $a_0$, and $a_1$ for fixed $a_{-1} = -0.2$. The left panel reveals that $d_\mu$ increases approximately linearly with $a_0$, with a slope that decreases with $a_1$. Comparison with Fig. 3d shows the impact of introducing $a_{-1}$, which can be mixed but is generally to decrease $d_\mu$ over the parameter regime considered.

Figure 4d shows the relationship between $d_\mu$, $a_0$, and $a_1$ for fixed positive $a_{-1} = 0.2$. Again, the left panel reveals that $d_\mu$ increases approximately linearly with $a_0$, with a slope that decreases with $a_1$. Comparing with Fig. 3d shows how $a_{-1} = 0.2$ typically increases $d_\mu$ relative to $a_{-1} = 0$.

As in the case with two terms, the speed of the sequence can be held constant while the stability term $\tilde{g}$ is decreased. Indeed, the introduction of $a_{-1}$ introduces an additional degree of freedom. Examples of keeping $d_\mu$ fixed while varying $\tilde{g}$ can be seen by following horizontal lines of fixed $d_\mu$ in Figs. 4c and 4d.

Three terms with forward connectivity

Now we consider the case with two forward terms, where $a_k = 0$ for $k \notin \{0, 1, 2\}$. This illustrates the effects of having slower timescale TAH kernels, roughly equal to or a bit slower than the timescale of the tutor signal. Figure 5a illustrates a TAH kernel that would give rise to two forward terms. Using Eq. (12):

$$t_\mu \approx \frac{\mu - 1}{a_1 + 2a_2} = \frac{\bar{a}_1 + 4a_2}{2(a_1 + 2a_2)^2}. $$ [18]

Sequence speed as measured by peak difference is then

$$d_\mu \approx \frac{1}{a_1 + 2a_2} = \frac{a_0 + a_1 + a_2}{a_1 + 2a_2}. $$ [19]

Figure 5 compares the full network simulations, the simulations of the linear system Eq. (10), and the asymptotic approxi-
This is most clearly illustrated by the horizontal line $t < \mu$ in biology (43, 52). These have the form in the right panel of Fig. 5c. Fig. 5b); that is, $d_{\mu}$, depending on the values of $a_0$ and $a_1$. Here $a_2 = 0.2$. Shaded regions are 95% confidence intervals for $\mu \in \{2, ..., 10\}$, and lines are the means. Left: $a_0$ is plotted on the horizontal axis and $a_1$ is denoted by color. Right: $a_1$ is plotted on the horizontal axis and $a_0$ is denoted by color. A) Overlaps $q_{\mu}(t)$ corresponding to parameter value $(0.4, 0.6, -0.1)$. B) Overlaps $q_{\mu}(t)$ corresponding to parameter value $(-0.4, 0.6, 0.4)$.

Fig. 5. Sequence progression for two forward terms. a): illustration of a TAH kernel $w$ that would give rise to two forward terms. Plots show the full network simulations (solid lines), linear approximations Eq. (15) (dashed lines), and approximation Eq. (12) (dotted lines). b) Peak times $t_\mu$ for a variety of coefficient combinations $(a_0, a_1, a_2)$, denoted by color. Dotted lines are the linear approximation given by Eq. (15). The $\dagger$ symbol denotes parameter values that give rise to the same peak times. c) Plots of peak differences $d_{\mu}$ as a function of $a_0$ and $a_1$. The $a_2 = 0.2$. The relationship between $d_{\mu}$ and different coefficient combinations is further elucidated in Fig. 5c, which plots $d_{\mu}$ as a function of $a_0$ and $a_1$ where $a_2 = 0.2$. As in previous sections, the dependence of $d_{\mu}$ on $a_0$ is approximately linear over the coefficient values considered, with a slope that decreases with $a_1$. Comparison with Fig. 3d shows the (multiplexed) effect of positive $a_2$.

Note that, as in the previous sections, the stability properties of the sequence (determined by the sum of coefficients) can be improved even as the timing is held constant (\dagger symbol in Fig. 5b); that is, $1/(a_1 + 2a_2)$ can be fixed while $\bar{g}$ is decreased. This is most clearly illustrated by the horizontal line $d_{\mu} = 1$ in the right panel of Fig. 5c.

Here we consider the special case of (double-sided) exponential kernels as commonly used to model TDP kernels seen in biology (43, 52). These have the form $w(t) = -m_1e^{-\alpha t}$ for $t < 0$ and $w(t) = m_2e^{-\beta t}$ for $t \geq 0$ (see Fig. 6a for an illustration).

Exponential kernels

We compute in Appendix 2 that the approximate peak times are given by Eq. (12) where

$$\alpha = \bar{g}(m_1\tau_1^2 + m_2\tau_2^2)$$

$$\beta = \bar{g} \left(m_1\tau_1^2 \left(1 - e^{-\tau_1/\tau_1}\right) + m_2\tau_2^2 \left(1 - e^{-\tau_2/\tau_2}\right)\right)$$

and

$$\bar{g} = (m_1\tau_1 \xi - m_2\tau_2 \xi)^{-1}.\tag{17}$$

Here $T_\xi$ is the length of the tutor signal interval. Note in particular that the (inverse) sequence speed as measured by

$$d_{\mu} \approx \frac{1}{\alpha} = \frac{T_\xi \tau_1 m_1 - m_2 \tau_2}{m_1 \tau_1^2 + m_2 \tau_2^2}$$

scales linearly with $T_\xi$. This dependence provides a biologically-plausible mechanism to control the speed of recalled sequence by the tutor.

The quantity $d_{\mu}$ is plotted in Figs. 6b and 6c for a variety of different kernels as a function of tutor interval $T_\xi$. These plots compare the full network simulations with the linear approximation Eq. (10) and the approximation Eq. (12), showing a close match, especially for small $T_\xi$.

In Fig. 6d we plot the dependence of inverse sequence speed $d_{\mu}$ as a function of $\tau_1$, $\tau_2$, and $m_2$. The left panel shows that $d_{\mu}$ typically decreases with increasing $\tau_2$, but this can be reversed for large enough $\tau_1$. The center panel shows that, for the parameter values chosen, $d_{\mu}$ decreases with $\tau_1$. The right panel shows the influence of changing $m_2$ on $d_{\mu}$; while increasing $m_2$ increases $d_{\mu}$, the effect is small. Parameter values were chosen to demonstrate the full spectrum of behaviors within the regime of stable, forward propagating dynamics.

To look more closely at the interplay between the timescale of the TAH kernel and the timescale of the tutor signal, we look at the case where $\tau_1$ is small relative to $\tau_2$. Then $\alpha \approx \tau_2/T_\xi$, indicating that in this case the ratio of the timescales sets the sequence speed.
Sequences with fast and slow parts

In naturalistic settings, different states of a sequence may have different duration. For instance, in playing a piece of music different notes are held with different durations. Our analysis indicates that changing these durations causes a change in the sequence progression speed, meaning that variable duration tutor signals $T_k^2$ will result in sequences with faster and slower parts (Fig. 7). Consider a sequence with sections that pass from one set of uniform durations $\{T_k\}_{\mu \in S_1}$ to another $\{T_k\}_{\mu \in S_2}$. The overlaps $q_\mu$ for $\mu \in S_1$ will be governed by the first interval length, with some disturbance from $q_\nu$ for $\nu \in S_2$ mediated by the backward coefficients $a_{k\nu}$ for $k > 0$. For $\mu$ far from the boundary this disturbance should be small. On the other hand, $q_\mu$ far from the boundary will be governed by the second interval length with small disturbance from $q_\nu$, with the caveat that the evolution up to time $t$ has been influenced significantly by the dynamics of $q_\nu$. With these caveats in mind, we should expect the speed in each of these regions far from the boundary to progress at speeds determined by the corresponding tutor signal interval length, with more complex behavior occurring near the boundary of the transition. In Fig. 7 we consider a case with two transitions, where $T_1 = 1$ for $\mu \in [1, 20]$, $T_2 = .5$ for $\mu \in [21, 30]$, and $T_3 = 1$ for $\mu \in [31, 80]$ (slow to fast back to slow). Here we use a double-sided exponential kernel with $\tau_1 = 0.5, m_1 = 5, \tau_2 = 1$, and $m_2 = 2$. This figure illustrates that after transitioning, the sequence speeds up or slows down to match the new duration.

However, these transitions have undesirable characteristics. For one, in the first transition the sequence slows down just before entering the faster sequence section. A second issue is that the transition from $T_1^\mu$ to $T_2^\mu$ is very gradual and takes a significant amount of time before recovering the speed of the $T_1^\mu$ section. These issues, and particularly the third, are probably related to how the overlaps $q_\mu$ “spread out” for increasing index $\mu$ (see Fig. 1 for an example); that is, the width of the bumps traced out by $q_\mu(t)$ increases. This suggests a modified sequence model.
where the behavior of the model is agnostic to the index of the
pattern being represented, perhaps with the use of nonlinear
learning rules as in (37, 51) or “interaction modulation” as
suggested by (53).

Discussion
The ability to store experiences as memories is fundamental
to intelligent behavior. While memories are typically modeled
as fixed points in recurrent networks, sequences constitute
information that is also of primary importance. This is clear
when considering the ubiquity of well-practiced, automatic
motor behaviors.

Our model stores memories by assuming that the network is
placed in states by a “tutor signal” during the learning process.
There are multiple possible interpretations for this. The state
of the network could be set by sensory inputs, so that TAH
learning rules constitute an unsupervised method for learning
sequential structure in the world. Another interpretation is
that higher-level brain areas provide this tutor signal. For
instance, motor cortex could be directing motor outputs (a
process involving slow, deliberate thought) while sending tutor
signals to a subcortical network which makes a copy of the
behavioral control that allows sequential behavior to be later
recalled, thereby skipping the slow deliberation process and
responding to automatic behavior (54, 55).

This work seeks to fill a significant gap in theory describing
the ability to form sequence memories. Namely, it is not
clear how general and commonly observed Hebbian learning
rules could give rise to such sequence memories. Hebbian
learning rules have the advantage of being simple, biologically
plausible, implementable in neural circuits, and continuously
differentiable (as opposed to, say, storing data in a database).
Our theory may also be useful for analysis of dynamical systems
with sequence dynamics more generally, since the mean-field
equations Eq. (9) are quite general.

Our analysis reveals many interesting phenomena. For
one, Hebbian kernels can be chosen that produce the same
sequence (in terms of timing), but that are more or less ro-
 bust to noise. Our theory lays out a quantitative framework
that allows such kernels to be delineated. We derive simple
relationships between the parameters of the (double-sided)
exponential kernels commonly seen in biology, and show that
sequence timing scales approximately linearly with the timing
of the tutor signal, and the slope of this linear relationship
depends on the parameters of the kernel. Finally, we show the
result of changing the timing of the tutor signal throughout
the sequence, so that the sequence has faster and slower parts.
While the sequence timing roughly follows these changes, we
find several characteristics of these changes that leave some-
thing to be desired and suggest possible goals for improving
the model.

A shortcoming of our model is that the possibility of cor-
related patterns is not addressed. It is reasonable to think
that sequence state patterns could be correlated in biological
circuits, and addressing this may provide a more useful and
biologically faithful model. An additional limitation is the
interpretation of discrete sequence states, and on understand-
ing what the biologically relevant timescales for tutor signals
are. It is possible that neural control signals in higher order
brain areas are “chunked” – that is, a relatively constant con-
trol signal initiates another motor area that enacts a motor
motif (1). In addition, we consider only linear TAH learning
rules, although we expect our analysis would extend in spirit
to nonlinear learning rules such as those considered in (37).

In the longterm, memory models should be combined with
other models, such as models of motor cortex supporting
flexible motor control. Tutor signals would then be generated
by these other models, and gradually the memory model
would take over control of motor actions as they become more
practiced (56). In addition, mechanisms for refinement based
on reward signals could be added to the memory modules.
For instance, by gating TAH learning rules based on reward
signals, this learning process can be turned into a reinforcement
learning process.

In summary, our analysis shows that sequence memories can
be formed by simple biologically plausible learning rules with
general TAH kernels and sequence speed is set by the interplay
between the TAH kernels and the duration that the network
is tutored. Our work indicates areas of potential improvement
and future development that we believe are important for
advancing the field.

Acknowledgements
MF thanks the Swartz Foundation and HDSI for its support.
CP is supported by NSF grant DMS-2134157.

References
1. AM Graybiel, The basal ganglia and chunking of action reper-
2. O Hikosaka, et al., Parallel neural networks for learning se-
3. P Redgrave, et al., Goal-directed and habitual control in the
4. N Kadmon Harpaz, K Hardcastle, BP Ölveczky, Learning-
induced changes in the neural circuits underlying motor se-
5. A Moors, J De Houwer, Automaticity: A theoretical and
6. KGC Mizes, J Lindsey, GS Escola, BP Ölveczky, Dissociating
the contributions of sensorimotor striatum to automatic and
7. FG Ashby, BO Turner, JC Horvitz, Cortical and basal gan-
glia contributions to habit learning and automaticity. Trends
8. AM Haith, JW Krakauer, The multiple effects of practice: Skill,
habit and reduced cognitive load. Curr. opinion behavioral
9. S Hélie, D Cousineau, The cognitive neuroscience of automatic-
ity: Behavioral and brain signatures, in Advances In Cognitive
and Behavioral Sciences., Psychology Research Progress. (Nova
10. M Carter, D Shapiro, Control of sequential movements: Evi-
dence for generalized motor programs. J. neurophysiology 52,
11. DR Gentner, Timing of skilled motor performance: Tests of
12. S Sternberg, RL Knoll, S Monsell, CE Wright, Motor programs
and hierarchical organization in the control of rapid speech.
13. MA Long, MS Fee, Using temperature to analyse temporal
dynamics in the songbird motor pathway. Nature 456, 189–194
(2008).
14. MA Long, DZ Jin, MS Fee, Support for a synaptic chain model

Code availability

Code for reproducing the plots can be found at https://github.com/msf235/Recall-tempo-of-Hebbian-sequences. For details of the parameters used in the simulations for producing the plots, see Appendix 3.
Appendix 1

Derivation of mean-field equations

Here we derive the mean-field equations Eq. (9). This is a fairly straightforward generalization of the derivation in (37). The full network equations are \( \dot{r} = -r + \phi(Wr + \eta) \), where \( \eta(t) \) is a white noise vector \((\langle \eta(t) \rangle = 0 \text{ and } \langle \eta(t)\eta(t') \rangle = \rho^2 \delta(t - t') \)\) and \( \phi \) is a sigmoidal nonlinearity:

\[
\phi(x) = \frac{r_{\text{span}}}{2} \left( \text{erfc} \left( \frac{x - \theta}{\sqrt{2\sigma}} \right) \right).
\]

The weights \( W \) are

\[
W = \frac{1}{N} \sum_{\mu=1}^{P} \sum_{\nu=1}^{P} a_{\mu}^{\nu} (\xi^\mu) \top.
\]

Here each pattern \( \xi^\nu \) is standard normal: \( \xi^\nu \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \). The mean-field equations are written in terms of the overlaps \( q_{\mu}(t) = \langle \xi^\mu \rangle \top r(t)/N \). If we let \( h = Wr + \eta \) we can then write

\[
h = \sum_{\mu=1}^{P} \sum_{\nu=1}^{P} a_{\mu}^{\nu} q_{\mu} + \eta.
\]

We now investigate the evolution of the overlaps \( q_{\mu} \):

\[
\dot{q}_{\mu} = \langle \xi^\mu \rangle \top \dot{r}/N
\]

\[
= -q_{\mu} + \langle \xi^\mu \rangle \top \phi(h)/N.
\]

Let’s look more closely at the second term on the right hand side. As \( N \to \infty \) this term approaches an average by the law of large numbers, yielding \( \langle \xi^\nu \rangle \top \phi(h)/N \to \langle \xi^\nu \phi(h) \rangle \) where \( \xi^\nu \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \) and

\[
h = \sum_{\mu=1}^{P} \sum_{\nu=1}^{P} a_{\mu}^{\nu} q_{\mu} + \eta.
\]

Here \( \eta \) is a scalar white noise term: \( \langle \eta(t) \rangle = 0 \) and \( \langle \eta(t)\eta(t') \rangle = \rho^2 \delta(t - t') \). Hence,

\[
\dot{q}_{\mu} = -q_{\mu} + \left( \sum_{\nu=1}^{P} a_{\mu}^{\nu} (\xi^\nu q_{\nu} + \eta) \right)
\]

\[
= -q_{\mu} + \left( \sum_{\nu=1}^{P} a_{\mu}^{\nu} q_{\nu} + \sum_{\nu \neq \mu}^{P} a_{\mu}^{\nu} q_{\nu} + \eta \right)
\]

We next note that each term of

\[
\sum_{\nu \neq \mu}^{P} a_{\mu}^{\nu} q_{\nu} + \eta
\]

is an independent normally distributed random variable with mean zero and variance \((\sum_{\nu=1}^{P} a_{\mu}^{\nu} q_{\nu})^2 \). Hence the sum \( \sum_{\nu \neq \mu}^{P} a_{\mu}^{\nu} q_{\nu} + \eta \) is distributed like \( xR_{\mu} \) where \( R_{\mu}^2 = \sum_{\nu \neq \mu}^{P} (\sum_{\ell=1}^{P} a_{\mu}^{\nu} q_{\nu})^2 + \rho^2 \) and \( x \) is standard normal. We similarly define \( S_{\mu} = \sum_{\ell=1}^{P} a_{\mu}^{\nu} q_{\nu} \) and write the average as a Gaussian integral:

\[
\dot{q}_{\mu} = -q_{\mu} + \int D\xi^\mu D\xi^\nu \xi^\mu (S_{\mu} + xR_{\mu})
\]

where \( D\xi = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \).

The remainder of the calculation follows closely (37). We next define the change of coordinates

\[
u = \frac{\xi^\mu S_{\mu} + xR_{\mu}}{\sqrt{S_{\mu}^2 + R_{\mu}^2}} = \frac{\xi^\mu S_{\mu} + xR_{\mu}}{c}
\]

\[
u = \frac{\xi^\mu R_{\mu} - xS_{\mu}}{\sqrt{S_{\mu}^2 + R_{\mu}^2}} = \frac{\xi^\mu R_{\mu} - xS_{\mu}}{c}
\]
where \( v \) and \( u \) are uncorrelated standard normal random variables and where we have defined \( c = \sqrt{S\mu + R\mu} \). To perform this change of variables we first compute the determinant of the Jacobian:

\[
\left| \frac{\partial (u,v)}{\partial (\mu,\xi)} \right| = \frac{1}{c^2} \left| \begin{array}{cc} S\mu & R\mu \\ R\mu & -S\mu \end{array} \right| = \frac{1}{c^2} (S^2\mu + R^2\mu) = 1.
\]

Hence we can use the substitution \( dudv = d\xi dx \). Furthermore,

\[
2\pi Duv = e^{-u^2/2}e^{-v^2/2}dudv \\
= \exp \left( -\frac{(\xi^\mu R\mu - xS\mu)^2}{2c^2} \right) \exp \left( -\frac{(\xi^\mu S\mu + xR\mu)^2}{2c^2} \right) dudv \\
= \exp \left( -\frac{(\xi^\mu R\mu)^2 + (\xi^\mu S\mu)^2 + (xS\mu)^2 + (xR\mu)^2}{2c^2} \right) dudv \\
= \exp \left[ -\frac{(\xi^\mu)^2 + x^2}{2} \right] d\xi dx \\
= 2\pi D\xi^\mu D\mu.
\]

Inverting the change of coordinate equations above, we find that \( \xi^\mu = (S\mu v + R\mu u)/c \). Using these substitutions, we can rewrite the integral

\[
\int D\xi^\mu D\mu \xi^\mu \phi (\xi^\mu S\mu + xR\mu) = \frac{1}{c} \int Duv (vS\mu + uR\mu) \phi (cv) \\
= \frac{S\mu}{c} \int Dv v\phi (cv).
\]

We now define

\[
G(x) = \frac{1}{\sqrt{2\pi}} \int Dv v\phi(v\sqrt{x}).
\]

Using the definition of \( \phi \), this integral can be evaluated with integration by parts and is

\[
G(x) = \frac{\text{rspan}}{\sqrt{2\pi(\sigma^2 + x)}} \exp \left( -\frac{\theta^2}{2(\sigma^2 + x)} \right).
\]

Our equation becomes

\[
q_\mu = -q_\mu + G(R^2\mu + S^2\mu)S\mu \\
= -q_\mu + G \left( \sum_\nu \left( \sum_\ell a_\nu^\ell q_\ell \right)^2 + \rho^2 \right) \sum_\ell a_\nu^\ell q_\ell \\
= -q_\mu + G \left( \|h(t)^2 + \rho^2 \right) \sum_\ell a_\nu^\ell q_\ell.
\]

Finally, we define

\[
g(t) = G \left( \|h(t)^2 + \rho^2 \right),
\]

which results in Eq. (9). Plots showing the convergence of the mean-field approximation to the network simulations for increasing \( N \) are shown in Fig. 2. Note that in subsequent analysis, the number of patterns \( P \) will also be taken to infinity. In order for these mean-field dynamics to hold exactly, \( P/N \) must vanish as \( N \) and \( P \) are both taken large.

Note that stability is determined by the shape of the function \( G \). For the values of \( \phi \) considered in the main text (\( \text{rspan} = 2, \text{rspan} = 0, \theta = 0, \) and \( \sigma = 0.1 \)), \( G \) has the shape plotted in Fig. S1. In particular, \( G \) is monotonically decreasing from its maximum at \( x = 0 \). Since \( \|h(t)^2 \geq 0 \), the maximal value that \( g(t) \) can take during the ongoing dynamics is bounded from above by \( G(\rho^2) \). Note that the function \( G \) is the same as in (37), and the supplementary material of (37) contains plots of \( G \) for various values of \( \text{rspan}, \theta \), and \( \sigma \).

**Appendix 2**

In this appendix we show the full details of the mathematical derivations of the peak times in the main text.
Derivation of $g_\infty$

Here we derive how the value of $g_\infty = \lim_{t \to \infty} g(t)$ depends on the coefficients $a_k$ and the nonlinearity $\phi$. We start with the mean-field equations Eq. (9) and assume that the coefficients are shift-invariant, so that $a_\nu^t = a_{\mu-\nu}$:

$$\dot{q}_\mu = -q_\mu + g(t) \sum_{\nu=1}^{P} a_{\mu-\nu} q_\nu.$$

We show that

$$\lim_{t \to \infty} g(t) := g_\infty = \bar{g} := \left( \sum_{k=-P/2}^{P/2} \alpha_k \right)^{-1}$$

provided the solution to Eq. (21) is stable, and under a few other mild assumptions. First, we assume that $g_\infty$ exists. Using this assumption, we note that for large enough $t$ the equations Eq. (21) become approximately linear (since $g(t)$ becomes approximately constant). Hence we can approximate the limiting behavior of Eq. (21) by the linear system

$$\dot{q}_\mu = -q_\mu + g_\infty \sum_{\nu=1}^{P} a_{\mu-\nu} q_\nu$$

provided we use as initial condition for Eq. (23) the solution of Eq. (21) evaluated at some large enough $t$. Since the limiting behavior (in terms of stability) of the linear system Eq. (23) does not depend on initial condition, this choice of initial condition is not of substance. Hence to analyze the asymptotic stability properties of Eq. (21), we can instead analyze the asymptotic stability properties of Eq. (23).

To analyze the asymptotic stability properties of Eq. (23), we instead consider solutions to this equation with periodic boundary conditions (in $\mu$). While this change in boundary conditions may influence the final solution, it should not affect the asymptotic stability behavior as we take $P \to \infty$. This change in boundary conditions requires approximating our kernel function $w(t)$ with a kernel function that is $T$-periodic; this approximation will be good if $w(t)$ decays to zero away from $t=0$ sufficiently fast so that $\sum_{k=-[P/2]+1}^{[P/2]} |\alpha_k| \approx 0$ and $\sum_{k> [P/2]} |\alpha_k| \approx 0$. This change of boundary conditions gives rise to a linear system of ordinary differential equations with a circulant coefficient matrix that we call $A$:

$$\ddot{q} = -q + g_\infty Aq.$$

The following analysis simplifies if we reindex the coefficients as $a_k$ for $k \in \{-[P/2],-[P/2]+1,\ldots,[P/2]\}$. For even $P$ this will leave us with $P+1$ coefficients (one more than is contained in $A$), so in this case we require $a_{-[P/2]} = 0$. The periodic boundary conditions (in $\mu$) imply that $A$ is circulant, which means that $A$ can be diagonalized by Fourier modes, with eigenvalues that take a simple form. Specifically, $A$ being circulant means that $-I + g_\infty A$ has eigenvalues $\lambda_k = -1 + g_\infty \sum_{k=-[P/2]}^{[P/2]} a_k e^{2\pi ik\mu/P}$.

Since the network dynamics are bounded by the saturating nonlinearity $\phi$, overlaps $q_\mu(t)$ will remain finite over time, which implies that $\max_\mu \Re \lambda_k \leq 0$. To be stable, the solutions must also not decay to zero, which implies that $\max_\mu \Re \lambda_k \geq 0$. Hence stable solutions of the system with periodic boundary conditions satisfy $\max_\mu \Re \lambda_k = 0$. We next relate the condition $\max_\mu \Re \lambda_k = 0$ to the coefficients $a_k$. We first note that

$$\Re \lambda_k = -1 + g_\infty \sum_{k=-[P/2]}^{[P/2]} a_k \cos(2\pi k\mu/P).$$

![Fig. S1. The function $G(x)$ for $r_{\text{span}} = 2$, $\theta = 0$, and $\sigma = 0.1$.](image-url)
Each of the cosine factors is maximized at \( \mu = 0 \); however, some of the \( a_k \) may be negative, which makes the location of the maximum for the sum less clear. Hence, we pair the terms in Eq. (25) as follows:

\[
\Re \lambda_\mu = -1 + g_\infty a_0 + g_\infty \sum_{k=1}^{[P/2]} (a_k \cos(2\pi k\mu/P) + a_{-k} \cos(-2\pi k\mu/P))
\]

\[
= -1 + g_\infty a_0 + g_\infty \sum_{k=1}^{[P/2]} (a_k + a_{-k}) \cos(2\pi k\mu/P).
\]

(For \( P \to \infty \), this rearrangement is valid provided the sum is absolutely convergent). We next assume that each of the forward terms \( a_k \) is positive and larger in magnitude than the corresponding backward term \( a_{-k} \). Under this assumption, each of the terms in the above sum is positive, so that the maximum occurs at \( \mu = 0 \). Note that this assumption is stronger than necessary; the maximum will occur at \( \mu = 0 \) provided the positive \( a_k + a_{-k} \) terms significantly outweigh the negative. This is a reasonable assumption for getting forward-propagating sequential behavior. Hence we can write

\[
\max \Re \lambda_\mu = -1 + g_\infty \sum_{k=-[P/2]}^{[P/2]} a_k.
\]

Hence \( \Re \lambda_\mu = 0 \) when Eq. (22) holds (using again the assumption that the terms \( a_k \) for \( k > [P/2] \) and \( k < [P/2] \) sum to zero).

In summary, we found that stable dynamics implies Eq. (22) under the following assumptions: (1) the limit \( g_\infty = \lim_{t \to \infty} g(t) \) exists, (2) \( \sum_{k=-[P/2]}^{[P/2]} |a_k| \approx 0 \) and \( \sum_{k=[P/2]}^{[P/2]} |a_k| \approx 0 \) (so that a periodic version of the nonperiodic system can be defined), and (3) the forward terms \( a_k \) for positive \( k \) sufficiently dominate the backward terms \( a_{-k} \), so that the maximum of \( \sum_{k=[P/2]}^{[P/2]} (a_k + a_{-k}) \cos(2\pi k\mu/P) \) occurs at \( \mu = 0 \). If taking \( P \to \infty \), we additionally require the series \( \sum_{k=-[P/2]}^{[P/2]} a_k \cos(2\pi k\mu/P) \) to be absolutely convergent so that the terms can be rearranged.

When will the network dynamics satisfy the stability condition of Eq. (22)? Since \( g(t) \) is bounded from above by \( G(\rho^2) \) (see Appendix 1), clearly the stability condition cannot be satisfied if \( g > G(\rho^2) \); in this case the network activity will decay to zero. What if \( g < G(\rho^2) \)? We can reason that the network will indeed satisfy the stability conditions. This is because, if the network activity is vanishing, then the argument \( ||h(t)||^2 + \rho^2 \) to \( G \) will approach \( \rho^2 \). This causes \( g(t) \approx G(||h(t)||^2 + \rho^2) \) to grow to a value where it causes the network activity to grow (since \( G(\rho^2) \) is larger than the critical boundary of stability of \( g \)). A similar line of reasoning can be found in (37).

Bidirectional connectivity — explicit solution

Here we derive an expression for a solution to the mean-field system with bidirectional connectivity. We include this derivation for completeness and to illustrate the difficulties that arise in deriving a simple exact expression for the solution. The reader may prefer to skip to the next sections showing our methods for finding approximate solutions. We start with the system of equations

\[
\dot{q}_\mu = \begin{cases} 
-q_\mu + \bar{a}_0 q_\mu + \bar{a}_{-1} q_{\mu+1} + \bar{a}_1 q_{\mu-1}, & \mu > 1 \\
-q_\mu + \bar{a}_0 q_\mu + \bar{a}_{-1} q_{\mu+1}, & \mu = 1 \\
-q_\mu + \bar{a}_0 q_\mu + \bar{a}_1 q_{\mu-1}, & \mu = P. 
\end{cases} \tag{26}
\]

In this case, the coefficient matrix \( A \) for the system written as \( \dot{q} = -q + Aq \) has eigenvalues

\[
\lambda_\mu = \bar{a}_0 + 2\sqrt{\bar{a}_{-1}\bar{a}_1} \cos \frac{\mu \pi}{P+1}, \quad \mu \in \{1, 2, \ldots, P\}
\]

and eigenvectors \( v^\mu \):

\[
v^\mu_k = \sqrt{\frac{2}{P+1}} \frac{\bar{a}_{-1}}{\bar{a}_1} \left(\bar{a}_1/\bar{a}_{-1}\right)^{k/2} \sin \frac{\mu k \pi}{P+1}, \quad \mu, k \in \{1, 2, \ldots, P\} \tag{27}
\]

The solution to Eq. (26) is then

\[
q(t) = VDV^{-1}q(0). \tag{28}
\]

where the columns of \( V \) are the eigenvectors \( v^\mu \) and where \( D \) is a diagonal matrix with \( D_{\mu\mu} = e^{(-1+\lambda_\mu)t} \). Letting \( V^* \) denote the conjugate transpose of \( V \), we see that \( VV^* \) is a diagonal matrix with the first entry being equal to 1. To see this, an easy calculation shows that

\[
\sum_{\mu=1}^{P} v_{\mu k}^* (v_{\mu k}^*)^* = 1, \quad \sum_{\mu=1}^{P} v_{\mu k}^* (v_{\mu k'}^*)^* = 0, \quad k \neq k'
\]
where the asterisk denotes the complex conjugate. Hence, if we use the initial conditions $q_{\mu}(0) = \delta(\mu - 1)$, we have that $(V^{-1}q(0))_k = (v_k^1)^*$. Hence we can write Eq. (28) as

$$q_\mu(t) = \sum_j v_j^\mu \exp((-1 + \lambda_j)t)(v_j^1)^*$$

$$= \frac{2}{P + 1} \left| \bar{a}_{-1}/a_1 \right|^\mu \sum_j \sin \frac{j\mu\pi}{P + 1} \exp \left[ -1 + a_0 + 2\frac{\sqrt{a_1/a_1}}{\sqrt{a_1/a_1}} \cos \frac{j\pi}{P + 1} \right] \sin \frac{j\pi}{P + 1}$$

Taking $P \to \infty$ this becomes an integral

$$q_\mu(t) = 2\left| \bar{a}_{-1}/a_1 \right|^\mu \sum_j \sin \left\{ \frac{j\mu\pi}{P + 1} \right\} \exp \left[ -1 + a_0 + 2\frac{\sqrt{a_1/a_1}}{\sqrt{a_1/a_1}} \cos \frac{j\pi}{P + 1} \right] \sin \frac{j\pi}{P + 1}$$

$$= 2\left| \bar{a}_{-1}/a_1 \right|^\mu \sum_j K_\mu(t)$$

where

$$K_\mu(t) = \int_0^1 ds e^{\mu \cos \pi s} \sin \mu \pi \sin \pi s.$$  

Noting that

$$I_\mu(t) = K_\mu(t)/\mu$$

is the modified Bessel function of the first kind, this equation can be written

$$q_\mu(t) = 2\left| \bar{a}_{-1}/a_1 \right|^\mu |e^{(-1 + a_0)t}K_\mu(2\sqrt{a_1/a_1})|$$

Instead of analyzing these equations further, we proceed with approximate methods that will generalize to more cases (i.e. more nonzero coefficients $a_k$).

**Bidirectional connectivity – periodic boundary conditions**

To further analyze the behavior, we convert the boundary conditions of Eq. (26) to periodic. This yields the following system of equations

$$q_\mu = \begin{cases} 
-q_\mu + \bar{a}_0 q_\mu + \bar{a}_{-1} q_{\mu+1} + \bar{a}_1 q_{\mu-1}, & \mu > 1 \\
-q_\mu + \bar{a}_0 q_\mu + \bar{a}_{-1} q_\mu + \bar{a}_1 q_\mu, & \mu = 1 \\
-q_\mu + \bar{a}_0 q_\mu + \bar{a}_{-1} q_\mu + \bar{a}_1 q_\mu - \bar{a}_0, & \mu = P.
\end{cases}$$

Note that the periodic equations may not match the solutions to the nonperiodic equations exactly, even as $P \to \infty$. This makes our simulations important for verifying that the approximations are useful.

The periodic equations can be written

$$\dot{q} = -Aq$$

where

$$A = \begin{pmatrix}
\bar{a}_0 & \bar{a}_{-1} & 0 & 0 & \cdots & \bar{a}_1 \\
\bar{a}_1 & \bar{a}_0 & \bar{a}_{-1} & 0 & \cdots & 0 \\
0 & \bar{a}_1 & \bar{a}_0 & \bar{a}_{-1} & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \bar{a}_1 & \bar{a}_0 & \bar{a}_{-1} \\
\bar{a}_{-1} & 0 & 0 & 0 & \bar{a}_0 & \bar{a}_1
\end{pmatrix}$$

Since $A$ is circulant, it can be diagonalized by Fourier modes. Specifically, the $\mu$th eigenvalue and eigenvector of $-I + A$ can be written $\lambda_\mu = -1 + \bar{a}_0 + \bar{a}_{-1} e^{2\mu \pi i/P} + \bar{a}_1 e^{2\mu \pi i/P}$ and $v_\mu = e^{-2\mu \pi k i/P}/\sqrt{P}$ for $\mu, k \in \{0, ..., P - 1\}$. Using the initial condition $q_\mu(0) = \delta(\mu - 1)$, the solution can then be written

$$q_{\mu+1}(t) = \frac{1}{\sqrt{P}} \sum_{k=0}^{P-1} v_k^\mu \exp \lambda_k t$$

which becomes for $P \to \infty$

$$q_{\mu+1}(t) = \frac{e^{(-1 + \bar{a}_0)t}}{2\pi} \int_0^{2\pi} \exp -\mu s \exp (\bar{a}_1 e^{2\pi i s} + \bar{a}_0 e^{-2\pi i s}) ds.$$  

Note that in taking $P \to \infty$, we still require that $N$ is much larger than $P$ so that the mean-field equations are valid. Simplifying the integrand results in

$$q_{\mu+1}(t) = \frac{e^{(-1 + \bar{a}_0)t}}{2\pi} \int_0^{2\pi} \exp (t(\bar{a}_1 e^{-2\pi i s} + \bar{a}_0 e^{2\pi i s})) ds.$$
and using the $2\pi$-periodicity of the integrand

$$q_{\mu+1}(t) = \frac{e^{(-1+\bar{a}_0)t}}{2\pi} \int_{-\pi}^{\pi} e^{t(\bar{a}_1+\bar{a}_{-1})} \cos s e^{t(\bar{a}_1-\bar{a}_{-1}) \sin s - \mu s} ds. \quad [31]$$

We next have some choice of method for approximating this integral. A simple approach would involve taking a truncated Taylor expansion of the trigonometric functions in the integrand, with the hope that this resolves into an integral that can be evaluated. A more sophisticated and principled approach would be to use a saddle point approximation, where $t$ is treated as a variable that is becoming large. We will start with saddle point approximations and show the limitations of this approach in this instance. We will then compare the result with the naive Taylor expansion approach. To get the essential information, the reader may wish to skip to this Taylor expansion approach.

### Bidirectional connectivity – saddle point approximation

We proceed by performing a saddle point approximation of the integral in Eq. (31). To do so, we write the integral as

$$I(t) = \int_{-\pi}^{\pi} e^{t\phi(s)} ds$$

where $\phi(s) = \beta \cos s + i\alpha \sin s - \gamma s$ and $\beta = \bar{a}_1 + \bar{a}_{-1}$ and $\alpha = \bar{a}_1 - \bar{a}_{-1}$ and $\gamma = \mu/t$. Note that we are treating $\mu/t$ as a finite scalar, so $\mu$ and $t$ are growing large together.

For the saddle point approximation, we first extend $\phi$ to a function over the complex numbers and find the critical points of $\phi$ (where $\phi$ are saddle points at these critical points in the complex plane). When $\gamma$ is large enough, there are two saddle points that both lie on the imaginary axis. To find these, we can take $s = iy$ and find the roots of $\phi'(iy)$. Letting $u = e^y$, $\phi'(iy)$ can be written $\phi'(iy) = -i\beta(u - 1/u)/2 + i\alpha(u + 1/u)/2 - iy$. The roots of the resulting quadratic are

$$u_{\pm} = \left(\gamma \pm \sqrt{\beta^2 - \alpha^2 + \gamma^2}\right)/(\alpha - \beta).$$

Here we see that the condition that these roots exist is that $\gamma^2 \geq \beta^2 - \alpha^2$ since $u$ must be real. We next need to calculate the angle of approach for these saddle points. This is given by $\arg \phi''$ evaluated at the saddle points. This shows that the angles are $-\pi/2$ and $0$ for the saddle points in the $y > 0$ positive half-space and $y < 0$ negative half-space, respectively. The contribution of the saddle point in the positive half-space is pure imaginary, so we can focus on the saddle point in the negative half space, which we denote $s^\ast$. Given the negative root $u_{-}$ above, it is straightforward to compute that

$$\phi(s^\ast) = \sqrt{\beta^2 - \alpha^2 + \gamma^2} + \log \left(\gamma - \sqrt{\beta^2 - \alpha^2 + \gamma^2}\right)/\alpha - \beta, \quad \phi''(s^\ast) = -\sqrt{\beta^2 - \alpha^2 + \gamma^2}.$$

The resulting saddle point approximation yields

$$I(t) \approx I_{\phi}(t) := \sqrt{\frac{2\pi}{|\phi''(s^\ast)|}} e^{t\phi(s^\ast)}$$

so that

$$q_{\mu+1}(t) \approx e^{(-1+\bar{a}_0)t} \frac{1}{2\pi|\phi''(s^\ast)|} e^{t\phi(s^\ast)}. \quad [32]$$

To find the peak times, we seek the roots of $q'_{\mu}(t)$. Given the saddle point approximation above, it is straightforward to compute $q'_{\mu}(t)$. This expression, even in the asymptotic limit $t \to \infty$ (with $\mu/t$ constant), does not appear to have roots that can be expressed in an explicit equation for $t$.

Given that the roots of the saddle point approximation are unhelpfully complex expressions, we next try taking a saddle point approximation of $q_{\mu+1}(t)$; i.e., instead of finding a saddle point approximation and then taking a derivative, we first take a derivative and then find the saddle point approximation. The derivative $q'_{\mu+1}(t)$ is

$$q'_{\mu+1}(t) = \frac{e^{(-1+\bar{a}_0)t}}{2\pi} \cdot ((-1+\bar{a}_0)I(t) + I'(t))$$

where

$$I'(t) = \int_{-\pi}^{\pi} \psi(s) e^{t\phi(s)} ds, \quad \psi(s) = \beta \cos s + i\alpha \sin s.$$

Taking a saddle point approximation of this involves taking a saddle point approximation of $I(t)$ and $I'(t)$; we have already done the former, and the latter is simply

$$I'(t) \approx \sqrt{\frac{2\pi}{|\phi''(s^\ast)|}} \psi(s^\ast) e^{t\phi(s^\ast)}$$
Fig. S2. Comparing approximations of $t_\mu$ to true peak times. In these plots, the difference between the true peak times $t_\mu$ and approximations $\hat{t}_\mu$ is plotted across $\mu$. Blue line, orange line, and green line correspond to the peaks of the saddle point approximation of $q_\mu(t)$, the roots of the saddle point approximation of $q'_\mu(t)$, and the peaks of the Taylor expansion approximation of $q_\mu(t)$, respectively. The python scipy.signal utility find_peaks is used to find the peak times numerically for the blue line. True peak times (dashed lines) are found by numerical quadrature of the integral expression in Eq. (19), followed by using the find_peaks. a) Coefficient values $a_{-1} = -3$, $a_0 = .1$, and $a_1 = 1.8$. b) Coefficient values $a_{-1} = 2$, $a_0 = 2$, and $a_1 = .8$. c-d) Difference of peak differences $d_\mu = t_\mu - \hat{t}_\mu$ with approximations $\hat{d}_\mu$. Note that the orange and green lines coincide. Dotted lines denote the precision ceiling due to the step size $dt = 0.001$.

c) Coefficient values $a_{-1} = -3$, $a_0 = .1$, and $a_1 = 1.8$. d) Coefficient values $a_{-1} = 2$, $a_0 = .2$, and $a_1 = .8$.

where $s^*$ is the critical point of $\phi$ as above. Hence

$$q_{\mu+1}(t) \approx \frac{1}{2\pi t |\phi''(s^*)|} e^{(-1+a_0)t} ((-1+a_0) + \psi(s^*)).$$

We again look for the peak times by finding the roots of this expression, which occur at $((-1+a_0) + \psi(s^*)) = 0$. Evaluating $\psi(s^*)$ is similar to evaluating $\phi''(s^*)$ and yields $\psi(s^*) = \sqrt{\beta^2 - \alpha^2 + \gamma^2}$. Substituting $\beta = \bar{a}_1 + \bar{a}_{-1}$, $\alpha = \bar{a}_1 - \bar{a}_{-1}$, $a_0 = 1 - \bar{a}_1 - \bar{a}_{-1}$, and $\gamma = \mu/t$ into this expression and solving $(-1+a_0) + \sqrt{\beta^2 - \alpha^2 + \gamma^2} = 0$ for $t$ yields

$$t_\mu \approx (\mu - 1)/(\bar{a}_1 - \bar{a}_{-1}).$$

This is the same as Eq. (12) up to a constant offset.

While this is a fair approximation, the differences between this estimate and the true peak times typically remain non-vanishing even as $t$ and $\mu$ approach infinity; see the blue and orange lines of Figs. S2a and S2b. However, the differences between the true peak differences $d_\mu = t_\mu - \hat{t}_\mu$ and this estimate do vanish with increasing $t$ and $\mu$; see the blue and orange lines of Figs. S2c and S2d.

Considering that the saddle point method yields either (1) an estimate for $t_\mu$ that doesn’t appear to have a simple closed-form expression, or (2) an estimate that is not asymptotically precise, we seek a simpler approach. This simpler approach derived in the next section has the additional advantage of generalizing to the case of arbitrarily many nonzero coefficients $a_k$.

Bidirectional connectivity - Taylor expansion

In our next approach to approximating the integral Eq. (31), we expand the trigonometric functions around $s = 0$ to get

$$q_{\mu+1}(t) \approx e^{(-1+a_0)t} \int_{-\pi}^{\pi} e^{(t(\bar{a}_1+\bar{a}_{-1})/(1-s^2/2))} e^{(t(\bar{a}_1-\bar{a}_{-1}) s - \mu)} ds$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-s^2 t (\bar{a}_1+\bar{a}_{-1})} e^{i\pi t (\bar{a}_1-\bar{a}_{-1}) - \mu} ds$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-s^2 t (\bar{a}_1+\bar{a}_{-1})} e^{i\pi t (\bar{a}_1-\bar{a}_{-1}) - \mu} ds$$

where in the last step we have used that $\sum_k \bar{a}_k = 1$. Letting $\alpha := \bar{a}_1 - \bar{a}_{-1}$ and $\beta := \bar{a}_1 + \bar{a}_{-1}$, we rewrite the integral as

$$q_{\mu+1}(t) \approx \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-\beta t/2} e^{i(\alpha t - \mu)} ds.$$
While we can write this integral in terms of error functions, a simpler expression results if we are able to integrate over the whole real line. We note that across a range of reasonable choices of the $a_k$, the integrand is vanishingly small outside of the interval $[-\pi, \pi]$. Hence we can indeed approximate the integral with one over the real line. This results in

$$q_{\mu+1}(t) \approx \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(\beta t/2)k^2 + i(\alpha t - \mu)s} ds$$

The locations $t_\mu$ of the peaks are given by the positive root of $q_\mu(t) = 0$, which when using the approximation above yields

$$t_\mu \approx -\beta + \sqrt{\beta^2 + 4\alpha^2(\mu - 1)^2}$$

Taking the asymptotic limit of large $\mu$ and taking the positive root yields Eq. (14):

$$t_\mu \approx \frac{\mu - 1}{\alpha} - \frac{\beta}{2\alpha^2} = \frac{\mu - 1}{\bar{a}_1 - \bar{a}_1 - \frac{\bar{a}_1 + \bar{a}_1}{2(\bar{a}_1 - \bar{a}_1)^2}}$$

Similar to the estimate yielded by taking a saddle point approximation of $q_\mu'(t)$, this estimate is not asymptotically exact; see the green lines in Figs. S2a and S2b. Depending on the choice of $a_k$, the approximation can be better or worse than the saddle point approximation Eq. (33); compare Fig. S2a with Fig. S2b. However, as before the differences $d_\mu := t_\mu - t_{\mu-1}$ are asymptotically correct; see Figs. S2c and S2d. Hence we can consider this approximation to be practically at least as useful as that yielded by taking saddle point approximations. An extra benefit of this approach is that it generalizes; see the next section.

**Generalizing the computation to shift-invariant connectivity**

The above technique of using Taylor expansions to find an approximate solution for $q_\mu$ extends to the general case of Eq. (10) with periodic boundary conditions in $\mu$. As in the derivation of $\tilde{g}$ above, we index the coefficients as $a_k$ for $k \in \{[P/2], [P/2] + 1, \ldots, [P/2]\}$. Then the coefficient matrix $A$ is circulant, and the eigenvalues of $I - A$ are $\lambda_k = -1 + \sum_{k=[P/2]}^{[P/2]} a_k e^{2\pi ik/\pi}$. Then grouping terms and taking the limit as $P \to \infty$ results in

$$e^{\lambda(s) t} = e^{(-1+\bar{a}_0)t} e^{\sum_{k=1}^{[P/2]} e^{i(\pi k/\pi)} \cos \frac{2\pi k}{P} + i e^{i(\pi k/\pi)} \sin \frac{2\pi k}{P} s}$$

where $c_k^+ = \bar{a}_k + \bar{a}_{-k}$, and $s \in [0, 1]$. Hence, using the eigenvectors Eq. (27) of circulant matrices in the expression Eq. (30) (which is also valid here),

$$q_\mu(t) \approx \frac{1}{P} \sum_{k=1}^{P} e^{i\lambda_k t} \exp \lambda_k t \to \frac{1}{2\pi} \int_{0}^{\infty} e^{\left(\sum_{k=1}^{[P/2]} e^{i(\pi k/\pi)} \cos \frac{2\pi k}{P} + i e^{i(\pi k/\pi)} \sin \frac{2\pi k}{P} s\right) ds}$$

as $P \to \infty$. Redefining the bounds of integration to $[-\pi, \pi]$ and expanding the trigonometric functions around $s = 0$ as above yields

$$q_\mu(t) \approx \frac{e^{(-1+\bar{a}_0)t}}{2\pi} \int_{-\pi}^{\pi} e^{i\sum_{k=1}^{[P/2]} e^{i(\pi k/\pi)} \cos \frac{2\pi k}{P} + i e^{i(\pi k/\pi)} \sin \frac{2\pi k}{P} s} ds = \frac{e^{(-1+\sum_{k=[P/2]}^{[P/2]} \bar{a}_k)}}{2\pi} \int_{-\pi}^{\pi} e^{-(\beta t/2)k^2 + i(\alpha t - \mu)s} ds$$

where $\beta = \sum_{k=1}^{[P/2]} e^{i \bar{a}_k}$ and $\alpha = \sum_{k=-[P/2]}^{[P/2]} e^{i \bar{a}_k}$. Again extending this to an integral over the real line as before, as well as noting that $\sum_{k=[P/2]}^{[P/2]} a_k = 1$, yields the solution

$$q_{\mu+1}(t) \approx \frac{1}{\sqrt{2\pi\beta t}} e^{\frac{(\alpha - \mu)^2}{2\alpha^2}}.$$

As before, the positive root of $q_\mu(t)$ is asymptotically

$$t_\mu \approx \frac{\mu - 1}{\alpha} - \frac{\beta}{2\alpha^2},$$

which is Eq. (12). This approximation will be used for the remainder of the Appendices.

**Two forward terms**

In the case of two forward terms, the system of equations is

$$\dot{q}_\mu = \begin{cases} -q_\mu + \bar{a}_0 q_{\mu-1} + \bar{a}_1 q_{\mu-2} + \bar{a}_2 q_{\mu-3}, & \mu > 2 \\ -q_\mu + \bar{a}_0 q_{\mu-1}, & \mu = 2 \\ -q_\mu + \bar{a}_0 q_{\mu-1}, & \mu = 1. \end{cases}$$

Using Eq. (12) yields the approximate peak times of Eq. (15):

$$t_\mu \approx \frac{\mu - 1}{\bar{a}_1 + 2\bar{a}_2} - \frac{\bar{a}_1 + 4\bar{a}_2}{2(\bar{a}_1 + 2\bar{a}_2)^2}$$
Exponential kernels

Here we derive the peak times $t_\mu$ for exponential kernels of the form $w(t) = -m_1 e^{t/\tau_1}$ for $t < 0$ and $w(t) = m_2 e^{-t/\tau_2}$ otherwise. As usual, we suppose that the coefficients $a_k$ are indexed as $k \in \{-P', -P' + 1, \ldots, P'\}$ for some integer $P'$ and define $a_k = 0$ for $k > P/2$ if necessary. Given the approximation Eq. (12), we need only compute $\alpha = \sum_{k=1}^{P'} (\bar{a}_k - \bar{a}_{-k})k$ and $\beta = \sum_{k=1}^{P'} (\bar{a}_k + \bar{a}_{-k})k^2$. To start we note that the relationship between the coefficients $a_k$ and the kernel $w$ takes the form

$$a_k = \frac{1}{N} \int_0^{T_\xi} dt \int_{kT_\xi}^{(k+1)T_\xi} ds \ w(s-t).$$

Then, for $k > 0$,

$$a_k = m_2 \int_0^{T_\xi} \int_{kT_\xi}^{(k+1)T_\xi} e^{-(s-t)/\tau} \ ds \ dt = m_2 \tau^2 e^{-T/\tau} \left( e^{T/\tau} - 1 \right)^2 \ e^{-kT/\tau_2}$$

$$a_{-k} = -m_1 \int_0^{T_\xi} \int_{kT_\xi}^{(k+1)T_\xi} e^{s/\tau_1 - t} \ ds \ dt = -m_1 \tau_1^2 e^{-T/\tau_1} \left( e^{T/\tau_1} - 1 \right)^2 \ e^{-kT/\tau_1}$$

$$a_0 = m_2 \tau_2^2 \left( T/\tau_1 + e^{-T/\tau_1} - 1 \right) - m_1 \tau_1^2 \left( T/\tau_1 + e^{-T/\tau_1} - 1 \right).$$

From this we compute

$$\lim_{P' \to \infty} \sum_{k=1}^{P'} a_k = m_2 \left( 1 - e^{-T/\tau_2} \right) \tau_2^2$$

$$\lim_{P' \to \infty} \sum_{k=1}^{P'} a_{-k} = m_1 \left( 1 - e^{-T/\tau_1} \right) \tau_1^2$$

Hence, for large $P'$,

$$\bar{g} \approx T^{-1} (m_2 \tau_2 - m_1 \tau_1)^{-1}$$

Similarly

$$\lim_{P' \to \infty} \sum_{k=1}^{P'} k a_k = \frac{e^{T/\tau_2}}{(e^{T/\tau_2} - 1)^2}$$

$$\lim_{P' \to \infty} \sum_{k=1}^{P'} -k a_{-k} = \frac{e^{T/\tau_1}}{(e^{T/\tau_1} - 1)^2}$$

Hence, for large $P'$,

$$\alpha \approx \bar{g} \left( m_1 \tau_1^2 + m_2 \tau_2^2 \right).$$

A similar calculation yields

$$\beta \approx \bar{g} \left( \frac{\tau_2^2 m_2 \left( e^{T/\tau_2} + 1 \right)}{(e^{T/\tau_2} - 1)^2} - \frac{\tau_1^2 m_1 \left( e^{T/\tau_1} + 1 \right)}{(e^{T/\tau_1} - 1)^2} \right).$$

Appendix 3

Here we give the details of the network simulations. Differential equations are solved with forward Euler integration with a timestep of $dt = 0.075$, except for Fig. 1 which uses $dt = 0.0375$ and Fig. 3e, which uses $dt = 1/15$. Recall that the length of the input patterns is denoted by $N$ and the number of patterns in the sequence is $P$. In Figs. 1 and 3 to 5 and Fig. S2, $N = 35000$. In Figs. 6 and 7, $N = 40000$. In Fig. 1, $P = 40$. In Figs. 3 to 5, 6b, 6c and 7 and Fig. S2, $P = 100$. In Fig. 6d, $P = 60$. In Figs. 2 to 6 and Fig. S2, $P = 100$.

Throughout, we use the nonlinearity Eq. (18) with $r_{\text{span}} = 2$, $r_{\text{center}} = 0$, $\theta = 0$, and $\sigma = 0.1$.

Error bars (shaded regions) are computed automatically by the Seaborn plotting library, which uses bootstrapping to compute 95% confidence intervals.