SCIENTIFIC REPORTS

OPEN

Synaptic mechanisms of interference in working memory

Zachary P. Kilpatrick[1,2]

Received: 17 November 2017
Accepted: 2 May 2018
Published online: 18 May 2018

Information from preceding trials of cognitive tasks can bias performance in the current trial, a phenomenon referred to as interference. Subjects performing visual working memory tasks exhibit interference in their responses: the recalled target location is biased in the direction of the target presented on the previous trial. We present modeling work that develops a probabilistic inference model of this history-dependent bias, and links our probabilistic model to computations of a recurrent network wherein short-term facilitation accounts for the observed bias. Network connectivity is reshaped dynamically during each trial, generating predictions from prior trial observations. Applying timescale separation methods, we obtain a low-dimensional description of the trial-to-trial bias based on the history of target locations. Furthermore, we demonstrate task protocols for which our model with facilitation performs better than a model with static connectivity: repetitively presented targets are better retained in working memory than targets drawn from uncorrelated sequences.

Parametric working memory experiments are a testbed for behavioral biases and errors, and help identify neural mechanisms that underlie them[1–3]. In visuospatial working memory, subjects identify, store, and recall target locations in trials lasting a few seconds. Response errors are normally distributed[4–6], and tend to accumulate during the delay-period, while subjects retain the target location in memory[1,6,7]. Complementary neural recordings suggest these working memories are implemented in circuits comprised of stimulus-tuned neurons with slow excitation and broad inhibition[8,9]. Persistent activity emerges as a tuned pattern of activity called a bump state, whose peak encodes the remembered target position[6,10].

Neuronal studies of visual working memory typically focus on population activity within a single trial, ignoring serial correlations across trials[11]. Several authors have identified behavioral biases that cause the previous trial's visual target to interfere with the subject's response on the subsequent trial[12,13]. For instance, in delayed match-to-sample tests, false alarms occur more often when comparison stimuli match samples from previous trials[14]. Interference was originally observed in verbal working memory tasks[15,16], and evidence suggests the effect impacts working memory capacity[17,18]. One consistent observation is that interference is reduced by increasing the time interval between trials[13,19,20], suggesting the effect persists for a few seconds. Investigations of interference in visuospatial working memory reveal other effects: Increasing the delay-period of working memory trials increases the bias strength, and responses are biased in the direction of the stimulus from the previous trial[13].

Our study focuses on why and how interference biases arise visuospatial working memory. First, what evidence accumulation strategy accounts for the bias introduced by the previous trial's target? We will show these biases emerge in observers using sequential Bayesian updating to predict the location of the next target. Such models are obtained by iteratively applying Bayes' rule to a stream of noisy measurements, updating an observer's belief of the most likely choice. In changing environments, older measurements are discounted at a rate that increases with the assumed change rate of the environment[21,22]. In our model, the sequence of targets observed on each trial is used to predict the next target. When subjects assume the environment changes rapidly, only the most recent target is used to make their prediction, leading to suboptimal inference of the subsequent target[23,24].

What neurophysiological processes could account for intertrial biases? Both the decay and activation timescales of the bias appear to be on the order of seconds. We propose short-term facilitation (STF), which acts on the timescale of seconds[25,26], can account for the dynamics of the bias. In a recurrent network that sustains persistent activity during a delay-period in the form of an activity bump, facilitated synapses from neurons tuned to the previous target attract the activity bump in the subsequent trial. Previous models identified STF as a possible mechanism for lengthening the timescale of working memory[27–29]. Our study proposes interference arises as a result of an irrelevant working memory remaining from the previous trial.

[1]Department of Applied Mathematics, University of Colorado, Boulder, Colorado, USA. [2]Department of Physiology & Biophysics, University of Colorado School of Medicine, Aurora, Colorado, USA. Correspondence and requests for materials should be addressed to Z.P.K. (email: zpkilpat@colorado.edu)

Our neurocomputational model accounts for recent observations of interference from visual working memory experiments[13], and makes novel predictions linking behavioral responses to corresponding neural and synaptic mechanisms. The separation in timescales between the neural activity dynamics and STF variable allows us to derive a low-dimensional model describing the bump's interaction with the network's evolving synaptic weights. We find that protocols with a uniform distribution of possible target angles lead to response error distributions that are normally distributed about zero as found previously[1,6,11]. Conversely, target protocol sequences with strong serial correlations can lead to a biased distributions in recalled target positions. Such biases may be advantageous in more complex tasks, where information from previous trials provides information about the target location in subsequent trials, as we show. Finally, we demonstrate that a recurrent network with STF supports bump attractors whose diffusion time course possesses two distinct phases, a prediction we propose to validate our model.

Results

Our study presents two frameworks for generating interference in a sequence of visual working memory trials. Both models use information about the target location on the previous trial to bias the response on the current trial. First, we develop a probabilistic inference model that predicts the distribution of possible target angles on the current trial based on observations of past trials. When the observer assumes the environment changes rapidly, the predictive distribution is primarily shaped by the previous trial's target. Second, we analyze a recurrent network model with STF wherein a localized bump of activity represents the observer's belief on the current trial and the spatial profile of STF represents the observer's evolving predictive distribution for the subsequent target. We show the attractor structure of the network model can be directly related to the predictive distribution of the inference model.

Interference in a visual working memory task. We focus specifically on an oculomotor delayed-response task with a single target presented in each trial[1,6]. On each trial, the subject views a target θ_n during a short cue period (Fig. 1A). They must remember the target location during a delay-period and saccade to the remembered location at the end. Response biases depend on the previous trial in three distinct ways[13]: (i) responses are attracted to the location of the previous target, graded with the difference between the current and previous target (Fig. 1B); (ii) the bias decreases as the interval between trials is increased (Fig. 1C); and (iii) the bias increases as the delay-period increases (Fig. 1D). As we will show, these biases are captured by a model of an observer that predicts the current target based on the previous target. These effects also emerge in a recurrent network model with slow excitation, subject to STF, and broad inhibition (Fig. 1E). This network represents the memory of the presented target as a bump of neural activity, which drifts in the direction of the target presented on the previous trial (Fig. 1F). Before analyzing the mechanics of this network model in more detail, we derive a probabilistic inference model that accounts for these tendencies.

Inference model for updating target predictions. Interference increases error in working memory tasks with independent trials, but may improve performance in tasks with probabilistically structured sequences of visual targets. We propose this as a biological origin of interference: subjects assume some predictable temporal structure in their environment. In fact, sequential Bayesian updating can account for interference observed in working memory, given specific constraints on a probabilistic updating algorithm. The observer attempts to predict the probability of observing target angle $\theta_{n+1} = \theta$ in trial $n+1$, given the targets $\theta_{1:n} = \{\theta_1, \theta_2, \ldots, \theta_n\}$ observed in the previous n trials (Fig. 2A). However, the target θ_j on the j^{th} trial will only help predict the target θ_{n+1} on the $n+1^{\text{th}}$ trial if the distribution $s_{n+1}(\theta)$ from which targets are drawn remains the same between trial j and trial $n+1$[30]. The observer assumes the distribution from which presented targets are drawn changes stochastically at a fixed rate $\varepsilon := \mathrm{P}(s_{n+1}(\theta) \not\equiv s_n(\theta))$. Most visual working memory protocols fix the distribution of target angles throughout the task $(\varepsilon = 0)$[1,3,6,13], as we do for most of our study, so the observer employs a potentially incorrect model to estimate this distribution $(\varepsilon > 0)$. Subjects in psychophysical tasks can have a strong bias toward assuming environments change on a timescale of several seconds[21], and this bias is not easily trained away[23,31]. Combining these features of the model, the observer updates their predictive distribution for the target during the $(n+1)^{\text{th}}$ trial.

Our algorithm is based on models that compute a predictive distribution for a stochastically moving target, given a sequence of noisy observations[30,32]. The predictive distribution is computed using sequential analysis[22,33]: Prior to trial $n+1$, the observer has seen n targets $\theta_{1:n} = \{\theta_1, \theta_2, \ldots, \theta_n\}$. The observer computes $f_{\theta'}(\theta) := \mathrm{P}(\theta_{n+1} = \theta | \theta_j = \theta', s_{n+1}(\theta) \equiv s_j(\theta))$ (Fig. 2B), the probability of observing the target θ_{n+1} in the $(n+1)^{\text{th}}$ trial assuming the underlying probability distribution from which targets are sampled does not change from trial j to $n+1$ $(s_{n+1}(\theta) \equiv s_j(\theta))$, for each trial $j = 1, \ldots, n$. The true distribution of target angles θ remains uniform throughout most our study, so the observer is applying suboptimal inference. Further details of our Bayesian nonparametric model are given in Methods.

The observer thus computes a predictive distribution $L_{n+1,\theta} = \mathrm{P}(\theta_{n+1} = \theta | \theta_{1:n}, \varepsilon)$, using the previous targets $\theta_{1:n}$ (Fig. 2A) to predict the subsequent target θ_{n+1}. If the observer assumes the distribution $s_{n+1}(\theta)$ from which targets are drawn in trial $n+1$ changes stochastically with a rate $\varepsilon \in (0, 1)$, recent observations will be weighted more in determining $L_{n+1,\theta}$[21,22,30]. Each observation θ_j contributes to the current estimate of $L_{n+1,\theta}$ via the probability $f_{\theta_j}(\theta)$ (Fig. 2B). Observations are weighted by assuming the observer has a fixed belief about the value ε, specifying the average number of trials they expect the distribution $s_n(\theta)$ to remain the same. Leveraging techniques in probabilistic inference (See Methods), we find

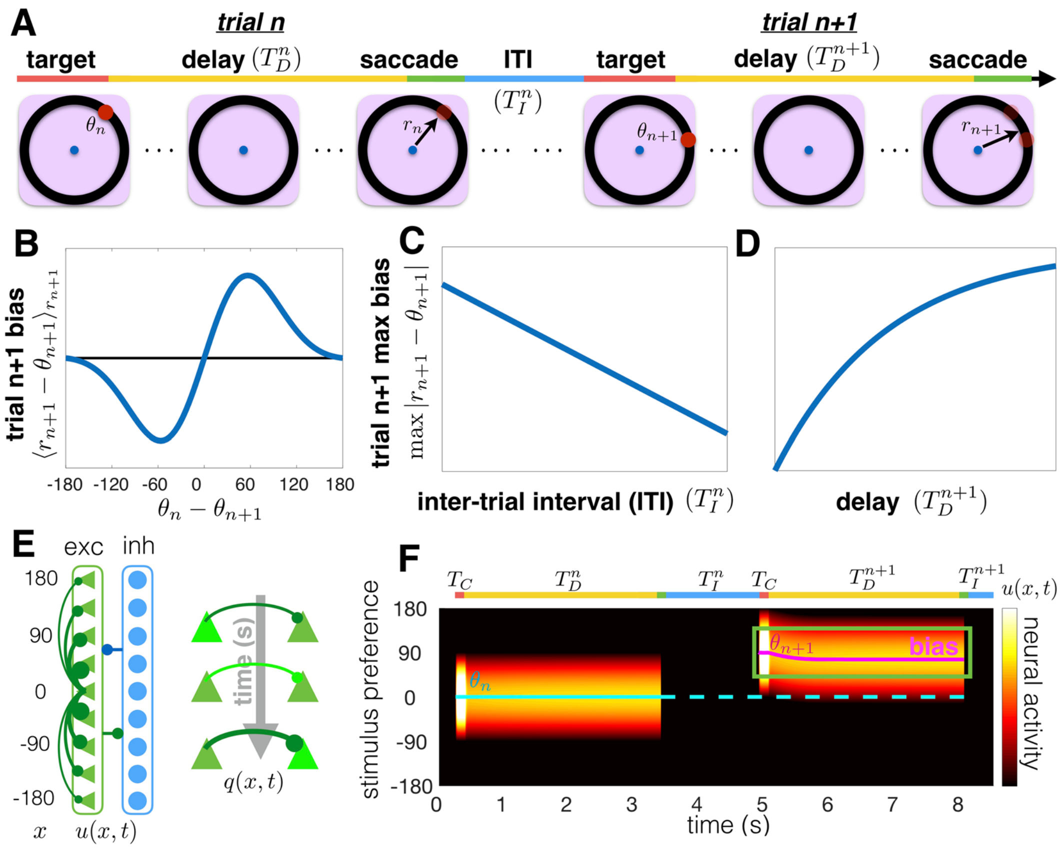

Figure 1. Interference in visuospatial working memory, and our corresponding recurrent network model with STF. (**A**) A visuospatial working memory task was administered in consecutive trials (schematics adapted from Papadimitriou *et al.*[13]). The subject fixates on the central (blue) dot and a target (red dot) appears at location θ_n on trial n. After the target disappears, the subject retains a memory of the target location during the delay-period (T_D^n and T_D^{n+1}, 0–6000 ms). Lastly, the subject makes a saccade (r_n and r_{n+1}) to the remembered target location. Papadimitriou *et al.*[13] found a systematic impact of the relative location ($\theta_n - \theta_{n+1}$) of the trial n target on the trial $n+1$ response r_{n+1}. (**B**) Response biases in trial $n+1$ $\left(\langle r_{n+1} - \theta_{n+1} \rangle_{\theta_{n+1}}\right)$ depend on the relative location of the target ($\theta_n - \theta_{n+1}$) in trial n. Responses err in the direction of the previous target θ_n, but this tendency is non-monotonic in $\theta_n - \theta_{n+1}$. (**C**,**D**) The maximum average bias in trial $n+1$ decreases with intertrial interval T_I^n (panel C) and increases with the trial $n+1$ delay-period T_D^{n+1} (panel D). (**E**) Schematic of our recurrent network model, showing excitatory (triangle) and inhibitory (circles) neurons. Connections between excitatory cells are distance-dependent. Effects of the inhibitory population are fast and spatially uniform, so excitatory and inhibitory populations are merged into single variable $u(x, t)$. STF increases the strength of recently used synapses, described by the variable $q(x, t)$. (**F**) A tuned input during the cue period (T_C) generates a bump of neural activity $u(x, t)$ centered at $x = \theta_n$ that persists during the delay-period of trial n (T_D^n) and ceases after the response. After the intertrial interval (T_I^n), the bump initially centered at $x = \theta_{n+1}$ drifts towards the position of the bump in the previous trial (dotted line) due to the attractive force of STF. Input fluctuations are ignored here to highlight the bias in a single trial.

$$L_{n+1,\theta} = \bar{\mathrm{P}}_0 \cdot \left[\frac{(1-\varepsilon)^n}{\mathrm{P}(\theta_{1:n})} \prod_{j=1}^{n} f_{\theta_j}(\theta) + \varepsilon \sum_{l=0}^{n-1} \frac{(1-\varepsilon)^l}{\mathrm{P}(\theta_{n-l+1:n})} \prod_{j=n-l+1}^{n} f_{\theta_j}(\theta) \right], \tag{1}$$

where we define $\mathrm{P}(\theta_{n+1:n}) = \prod_{j=n+1}^{n} f_{\theta_j}(\theta) = 1$ and $\bar{\mathrm{P}}_0 := 1/360$ is the uniform density for $-180° \le \theta < 180°$. To understand Eq. (1), it is instructive to examine limits of the parameter ε that admit approximations or exact updates.

Static environments ($\varepsilon \to 0$). In the limit $\varepsilon \to 0$, the observer assumes the environment is static, so the predictive distribution is comprised of equal weightings of each observation (See Fig. 2C and[34,35]):

$$L_{n+1,\theta} = \frac{\bar{\mathrm{P}}_0}{\mathrm{P}(\theta_{1:n})} \prod_{j=1}^{n} f_{\theta_j}(\theta). \tag{2}$$

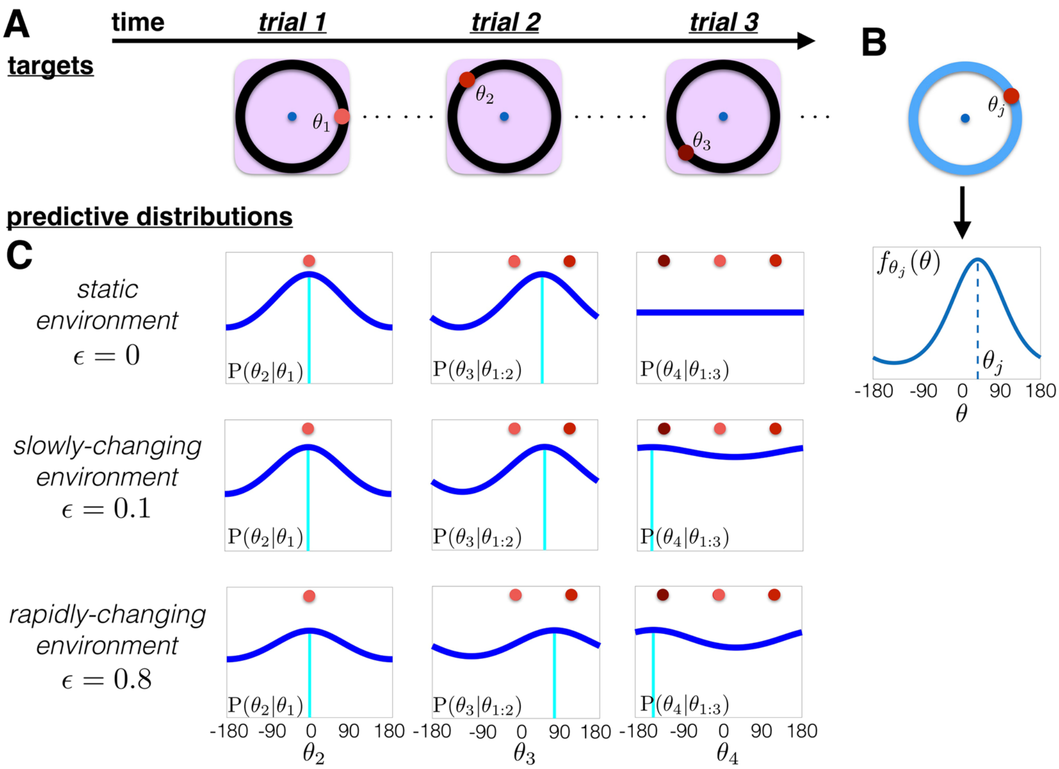

Figure 2. Updating the predictive distribution $L_{n+1,\theta}$. The observer infers the predictive distribution for the subsequent target θ_{n+1} from prior observations $\theta_{1:n}$, assuming a specific change rate ε of the environment: $L_{n+1,\theta} := \mathrm{P}(\theta_{n+1} = \theta|\theta_{1:n}, \varepsilon)$. (**A**) A sequence of presented targets: $\theta_{1:3}$. Note the environment is typically static, so $\varepsilon_{\text{true}} = 0$. (**B**) Probability $f_{\theta_j}(\theta)$, peaked and centered at θ_j, showing the observer's assumed probability that $\theta_{n+1} = \theta$, if θ_j is observed on trial j and the distribution remains the same in between ($s_{n+1}(\theta) \equiv s_j(\theta)$). (**C**) Evolution of the predictive distribution $L_{n+1,\theta} := \mathrm{P}(\theta_{n+1} = \theta|\theta_{1:n}, \varepsilon)$ for static ($\varepsilon = 0$); slowly-changing ($\varepsilon = 0.1$); and rapidly-changing ($\varepsilon = 0.8$) environments. In static environments, all observations $\theta_{1:3}$ are weighted equally whereas in the rapidly-changing environment, the most recent observation dominates.

As has been shown previously, Eq. (2) can be written iteratively[36]:

$$L_{n+1,\theta} = \frac{\mathrm{P}(\theta_{1:n-1})}{\mathrm{P}(\theta_{1:n})} f_{\theta_n}(\theta) L_{n,\theta},$$

suggesting such a computation could be implemented and represented by neural circuits. Temporal integration of tuned inputs has been demonstrated in both neural recordings[37–39] and circuit models[35,36,40] of decision-making tasks. Most oculomotor delayed-response tasks use a distribution of targets $s(\theta)$ that is constant across trials[1,3,6,13]. Therefore, Eq. (2) is the optimal strategy for obtaining an estimate of $s(\theta)$, assuming the observer has a correct representation of the probability $f_{\theta_j}(\theta)$. For instance, if the distribution $s(\theta)$ were peaked, repeated observations $\theta_{1:n}$ would gradually improve the observer's estimate of that peak in Eq. (2). In changing environments ($\varepsilon > 0$), recently observed targets are weighted more strongly than older targets, and the predictive distribution should down-weight the influence of past targets at a rate that increases with ε[22].

Rapidly-changing environment ($\varepsilon \approx 1$). Our work focuses on the limit where the environment changes rapidly, $\varepsilon \approx 1$ ($0 < (1 - \varepsilon) \ll 1$), to account for biases that depend on the previous trial's target θ_n (See Methods for other cases). In this case, the predictive distribution for trial $n + 1$ is a single peaked function centered at θ_n (Fig. 2C). The observer assumes the environment changes fast enough that each subsequent target is likely drawn from a new distribution ($s_{n+1}(\theta) \not\equiv s_n(\theta)$). This is a suboptimal strategy, but matches the typical trends of interference in working memory. Applying this assumption to Eq. (1), the formula for $L_{n+1,\theta}$ is dominated by terms of order $(1 - \varepsilon)$ and larger. Truncating to $\mathcal{O}(1 - \varepsilon)$ and normalizing the update equation (See Methods) then yields

$$\widetilde{L}_{n+1,\theta} = \varepsilon \bar{\mathrm{P}}_0 + (1 - \varepsilon) f_{\theta_n}(\theta). \tag{3}$$

Thus, the dominant contribution from $\theta_{1:n}$ to $\widetilde{L}_{n+1,\theta}$ is the target θ_n observed during the previous trial n (Fig. 2C), similar to the behavioral findings in Papadimitriou *et al.*[13].

Note, sequential computations are trivial in the limit of a constantly-changing environment $\varepsilon \to 1$, since the observer assumes the environment is reset after each trial. Prior observations provide no information about the present distribution, so the predictive distribution is always uniform: $L_{n+1,\theta} \equiv \bar{P}_0$.

In summary, a probabilistic inference model that assumes the distribution of targets is predictable over short timescales leads to response biases that depend mostly on the previous trial. We now demonstrate that this predictive distribution can be incorporated into a low-dimensional attractor model which describes the degradation of target memory during the delay-period of visual working memory tasks[10,41,42].

Incorporating suboptimal predictions into working memory. We model the loading, storage, and recall of a target angle θ using a low-dimensional attractor model spanning the space of possible target angles $\theta \in [-180, 180)°$. These dynamics can be implemented in recurrent neuronal networks with slow excitation and broad inhibition[6,9,43]. Before examining the effects of neural architecture, we discuss how to incorporate the predictive distribution update, Eq. (3), into an associated low-dimensional model. Our analysis links the update of the predictive distribution to the spatial organization of attractors in a network. Importantly, working memory is degraded by dynamic fluctuations, so the stored target angle wanders diffusively during the delay-period[6,9,42].

During the delay-period of a single trial, the stored target angle $\theta(t)$ evolves according to a stochastic differential equation[10]:

$$\mathrm{d}\theta(t) = -\frac{\mathrm{d}\mathcal{U}(\theta(t))}{\mathrm{d}\theta}\mathrm{d}t + \sigma_\theta \mathrm{d}\xi(t). \tag{4}$$

Here $\theta(t)$ is restricted to the periodic domain $\theta \in [-180, 180)°$ and $\mathrm{d}\xi$ is a standard white noise process. Equation (4) can be derived as the low-dimensional projection for the location of a bump attractor in a recurrent network. The potential gradient $-\mathcal{U}'(\theta)$ models spatial heterogeneity in neural architecture that shapes attractor dynamics (Fig. 3A). During trial $n+1$, we label the potential $\mathcal{U}(\theta) := \mathcal{U}_{n+1}(\theta)$. Classic models of bump attractors on a ring assume distance-dependent connectivity[9,43]. The case $\mathcal{U}'_{n+1}(\theta) \not\equiv 0$ accounts for spatial heterogeneity in connectivity that may arise from a combination of training and synaptic plasticity[10,44], or random components of synaptic architecture[45]. Potential models of bump attractors have been derived in detail for recurrent networks[46,47], and agree well with the qualitative dynamics of spiking network models[10,42]. The potential landscape of Eq. (4) is assumed to be updated during each trial, so at the beginning of trial $n+1$ it has the form $\mathcal{U}_{n+1}(\theta)$. When $\mathcal{U}_{n+1}(\theta) \equiv 0$, the potential is flat, so $\theta(t)$ evolves along a line attractor[46]. On the other hand, when the potential is heterogeneous, $\mathcal{U}_{n+1}(\theta) \not\equiv 0$, $\theta(t)$ tends to drift toward one of a finite number of discrete attractors[10,42]. We will incorporate a process whereby previous targets are used to update the potential, so $\mathcal{U}_{n+1}(\theta)$ is typically heterogeneous. The observer sees the target at the beginning of trial $n+1$, $\theta(0) = \theta_{n+1}$ (Fig. 3A), and the angle $\theta(t)$ evolves according to Eq. (4) during the delay-period, lasting T_D time units. After the delay-period, $\theta(T_D)$ is the recalled angle. Depending on the underlying potential $\mathcal{U}_n(\theta)$, there will be a strong bias to a subset of possible targets.

We derive a correspondence between the probabilistic inference model and attractor model by assuming stationarity of $\mathcal{U}_{n+1}(\theta)$ within each trial (See Methods). In the recurrent network model (Fig. 1E), we take these within-trial dynamics into account. Freezing $\mathcal{U}_{n+1}(\theta)$ during a trial allows us to relate the statistics of the position $\theta(t)$ to the shape of the potential. Specifically, we relate the stationary density of Eq. (4) to the desired predictive distribution $L_{n+1,\theta}$ (See Methods). Thus, if information about the current trial's target θ_{n+1} is degraded, the probability distribution associated with the recalled target angle θ is $L_{n+1,\theta}$. Focusing on interference trends in Fig. 1, we aim to have the attractor structure of Eq. (4) represent the predictive distribution in Eq. (3). Our calculations relate the potential function in trial $n+1$ to the probability generated by the trial n target (Fig. 3A) as

$$\mathcal{U}_{n+1}(\theta) \propto -f_{\theta_n}(\theta). \tag{5}$$

The potential $\mathcal{U}_{n+1}(\theta)$ can be implemented by a decaying plasticity process that facilitates synapses from neurons tuned to the previous target θ_n. The predictive distribution $L_{n+1,\theta}$ is therefore encoded by a potential $\mathcal{U}_{n+1}(\theta)$ that shapes the dynamics of the attractor model. As we will show, this can be accomplished via STF (Fig. 3B).

Short-term facilitation generates interference in working memory. We now show a neuronal network model describing neural activity $u(x, t)$ subject to STF $q(x, t)$ can incorporate predictive distribution updates derived above. Predictions are stored in the dynamically changing synaptic weights of a recurrent neuronal network. The recurrent network model spatially labels neurons according to their target orientation preference, determining the distance-dependent structure of inputs to the network. This is captured by a network with local excitation and effective inhibition that is fast and broad. Connectivity is shaped dynamically by STF (Fig. 1E). See Methods for more details.

A sequence of delayed-response protocols is implemented in the recurrent network by specifying a spatiotemporal input $I(x, t)$ across trials (top of Fig. 1F). Each trial has a cue period of time length T_C; a delay-period of time length T_D^{n+1}; and an intertrial period of time length T_I^{n+1} before the next target is presented. The network receives a peaked current centered at the neurons preferring the presented target angle θ_{n+1} during the cue period of trial $n+1$; no external input during the delay-period; and a strong inactivating current after the delay-period[6,9]. The resulting bump attractor drifts in the direction of the bump from trial n, due to the STF at the location of the trial n bump (Figs 1F and 3B).

The mechanism underlying intertrial bias is determined by projecting our recurrent network model to a low-dimensional system that extends Eq. (4) to account for STF. To reduce the recurrent network, we project the fast dynamics of bump solutions to an equation for the bump's position $\theta(t)$ in trial $n+1$[28,42,47]. The STF variable

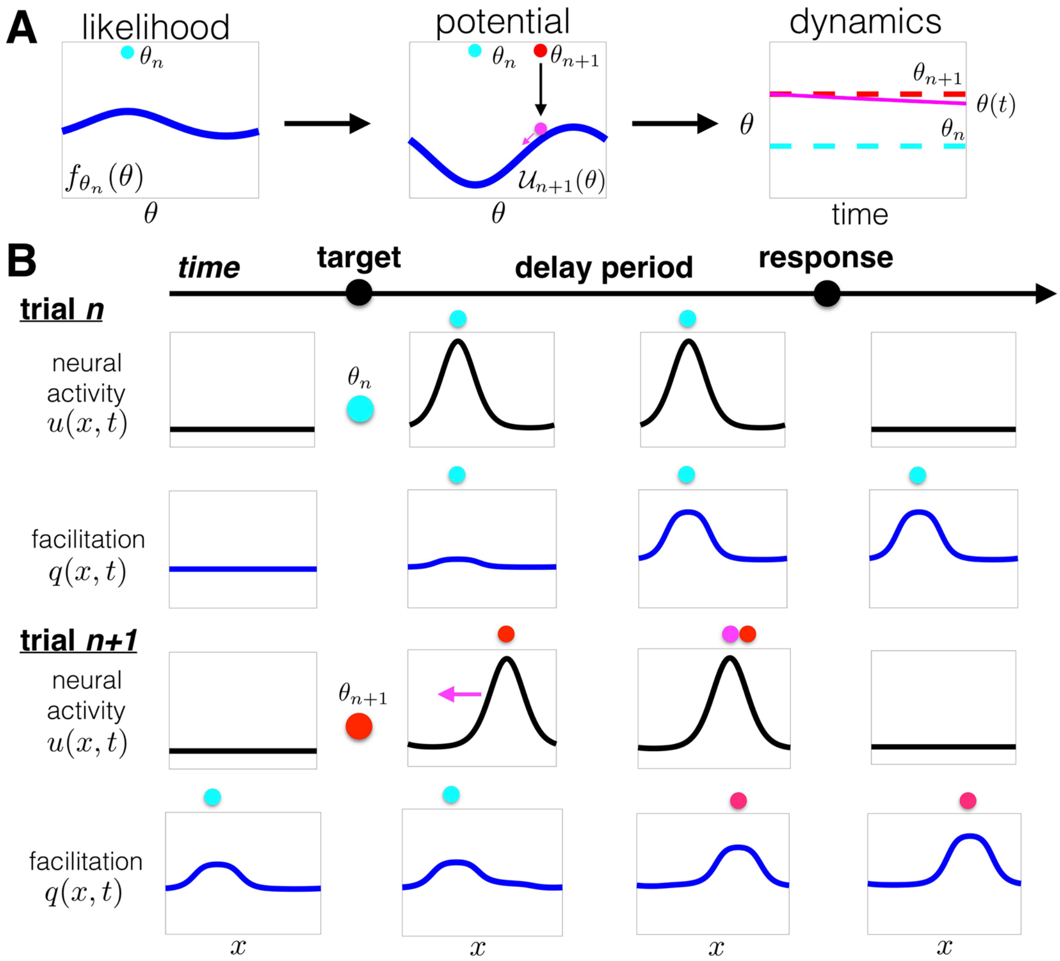

Figure 3. Encoding the predictive distribution in the potential function of an attractor network. (**A**) In a rapidly-changing environment, the predictive distribution is determined by the probability $f_{\theta_n}(\theta)$ (See Fig. 2C). In the low-dimensional system, with dynamics described by Eq. (4), this probability is represented by a potential function $\mathcal{U}_{n+1}(\theta)$ whose peak (valley) corresponds to the valley (peak) of $f_{\theta_n}(\theta)$, so the stored angle $\theta(t)$ drifts towards the minimum of $\mathcal{U}_{n+1}(\theta)$ during the delay-period. (**B**) A recurrent network with neurons distributed across $x \in [-180, 180)°$ with STF (Fig. 1E) can implement these dynamics. The position of the trial n target is encoded by the spatial profile of STF $q(x, t)$ during the early portion of trial $n+1$, attracting the neural activity $u(x, t)$ bump during the delay-period.

$q(x, t)$ determines an evolving potential function $\mathcal{U}(\theta,\ t)$ that shapes the bump's position (Fig. 4). We use timescale separation methods (See Methods) to derive a set of stochastic differential equations, which approximates the motion of the bump's position $\theta(t)$ and the location of STF $\theta_q(t)$:

$$d\theta(t) = -\mathcal{A}_n(t)\frac{d\overline{\mathcal{U}}(\theta(t) - \theta_n)}{d\theta}dt - \mathcal{A}_{n+1}(t)\frac{d\overline{\mathcal{U}}(\theta(t) - \theta_q(t))}{d\theta}dt + \sigma d\xi(t), \tag{6a}$$

$$\tau\dot{\theta}_q(t) = -d(\theta_q(t) - \theta(t)), \tag{6b}$$

during trial $n+1$ ($t_n < t < t_{n+1}$). The slowly-evolving potential gradient $-\frac{\partial}{\partial\theta}\mathcal{U}(\theta,\ t)$ shaping the dynamics of $\theta(t)$ is a mixture of STF contributions from trial n (decaying $\mathcal{A}_n(t)$) and trial $n+1$ (increasing $\mathcal{A}_{n+1}(t)$). The functions $\mathcal{A}_n(t)$ obey linear dynamics as shown in Fig. 4 (and see Methods). The bump position $\theta(t)$ moves towards the minimum of this dynamic potential, $\text{argmin}_\theta[\mathcal{U}(\theta,\ t)]$ (Fig. 4). The variable $\theta_q(t)$ is the location of STF originating in trial $n+1$, and its position slowly moves toward the bump location $\theta(t)$ via the scaled circular difference $d(\theta)$. The parametrized timescale τ of STF is inversely related to the observer's perceived environmental change rate ε in Eq. (3), since increasing ε corresponds to decreasing τ. While our derivation (See Methods) is performed assuming STF is slow and weak, we find the approximation agrees well with the full system for stronger STF.

The presence of STF provides two contributions to the slow dynamics of the bump position $\theta(t)$. The memory of the previous trial's target θ_n is reflected by the potential $\overline{\mathcal{U}}(\theta - \theta_n)$, whose effect decays slowly during trial $n+1$. This attracts $\theta(t)$, but the movement of $\theta(t)$ towards θ_n is slowed by the onset of the STF variable initially centered

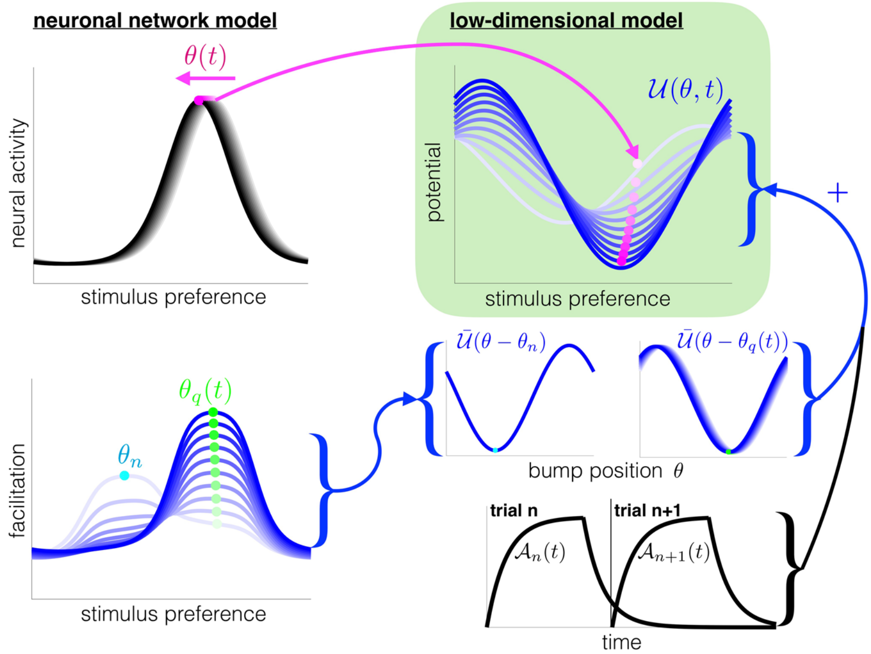

Figure 4. Low-dimensional system (green box) captures the motion of the bump ($\theta(t)$) and the evolving potential, $\mathcal{U}(\theta, t)$, shaped by STF. The center-of-mass of the neural activity bump $\theta(t)$ is attracted by the most facilitated region of the network, $\text{argmin}_\theta[\mathcal{U}(\theta, t)]$. The current trial's bump location $\theta(t)$ attracts the variable $\theta_q(t)$, indicating the location of STF in the current trial. The evolving potential $\mathcal{U}(\theta, t)$ is then comprised of the weighted sum of the potential arising from the previous target $\mathcal{U}(\theta - \theta_n)$ and the current bump $\mathcal{U}(\theta - \theta_q(t))$. Dynamic fluctuations also perturb the position $\theta(t)$, so memory would degrade diffusively in the case of a flat potential $\mathcal{U}(\theta) \equiv 0$. See Methods for a complete derivation.

at θ_{n+1}. The STF variable's center-of-mass $\theta_q(t)$ slowly drifts towards θ_n, which allows $\theta(t)$ to drift there as well, $\bar{\mathcal{U}}(\theta - \theta_q(t))$. This accounts for the slow build-up of the bias that increases with the length of the delay-period[13].

Target- and time-dependent trends match experimental observations. We now demonstrate that the bias observed in the visual working memory experiments of Papadimitriou *et al.*[13] can be accounted for by our recurrent network model (Fig. 1E) and our low-dimensional description of bump motion dynamics (Fig. 4). To represent a sequence of working memory trials, targets ($\theta_1, \theta_2, \theta_3, \ldots$) were presented to the recurrent network, and the center-of-mass of the bump was recorded at the end of each delay-period, representing the response ($r_1, r_2, r_3, \ldots$) (See Methods). The bias of responses was determined by computing the difference between the response and the presented target, $r_n - \theta_n$. Means and variances of the bias were determined under each condition.

Our results are summarized in Fig. 5, focusing on three conditions considered in Papadimitriou *et al.*[13]. First, we calculated the bias when conditioning on the angle between the trial n and trial $n+1$ targets, $\theta_n - \theta_{n+1}$ (Fig. 5A). Positive (negative) angles lead to positive (negative) bias; i.e. the bump drifts in the direction of the previous target θ_n. The bias is graded with the difference, $\theta_n - \theta_{n+1}$. To expose this effect, we averaged across trials, since the recurrent network incorporates dynamic input fluctuations, as in bump attractor models of visuospatial working memory[6,9]. We also calculated the peak bias as a function of the intertrial interval (ITI), the time between the trial n response (r_n) and the trial $n+1$ target presentation θ_{n+1}. Consistent with Papadimitriou *et al.*[13], the peak bias decreased with the ITI (Fig. 5B). The mechanism for this decrease is the slow decay in the STF of synapses utilized in the previous trial. Finally, the peak bias increased with the delay within a trial, since persistent activity was slowly attracted to the location of the previous target (Fig. 5C). This slow saturation arises due to the slow kinetics of STF within a trial. The bias produced is self-reinforcing, as the synapses originating from the newly-activated neurons undergo STF.

Not only did our recurrent network model recapitulate the main findings of Papadimitriou *et al.*[13], we also found our low-dimensional description of the bump and STF variable dynamics had these properties (blue curves in Fig. 5). The mechanics underlying the bias can be described with a model of a particle evolving in a slowly changing potential (Fig. 4), shaped by the dynamics of STF. Having established a mechanism for the bias, we consider how different protocols determine the unconditional statistics of responses.

Task protocol shapes ensemble statistics. Visual working memory tasks are often designed such that sequential target locations are independent[6,9]. In such protocols, there is no advantage in using previous trial information to predict targets within the current trial. Nonetheless, these biases persist in the intertrial response correlations discussed in Papadimitriou *et al.*[13] and Fig. 5. On the other hand, interference might be advantageous for tasks in which successive target angles θ_{n+1} depend on the previous θ_n. Consider object motion tracking tasks with transiently occluded objects[48,49]. The object's location prior to occlusion predicts its subsequent location

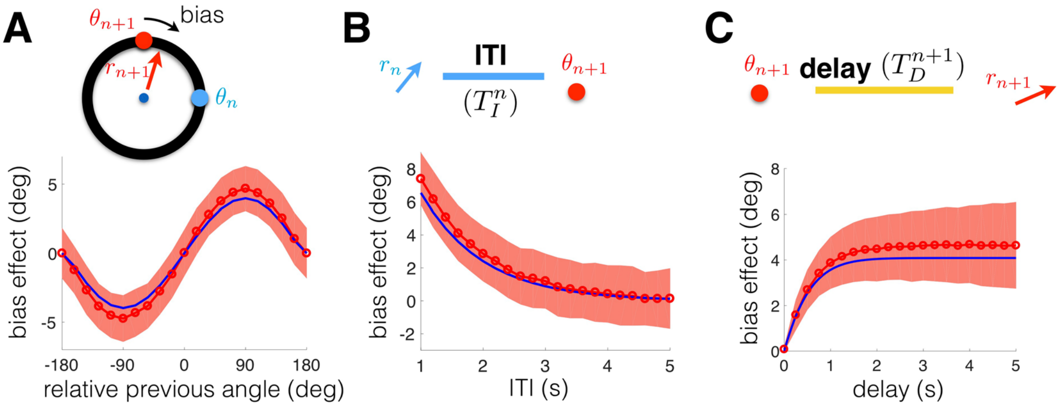

Figure 5. Intertrial bias is shaped by (**A**) the angle between targets θ_{n+1} and θ_n; (**B**) the interval between trials n and $n+1$ (ITI); and (**C**) the delay-period during trial $n+1$. (**A**) Responses in trial $n+1$ are biased in the direction of the previous trial target (θ_n). Simulations of the recurrent network (red circles) are compared with the low-dimensional model (blue line). Shaded region indicates one standard deviation (See Methods for details). (**B**) The peak bias decreases with the intertrial interval (ITI), due to the temporal decay of STF. (**C**) The peak bias increases with the delay since the bump drifts towards the equilibrium position determined by the STF profile.

when it reappears, so object memory that persists beyond a single trial can be useful for more naturally-inspired tasks.

We demonstrate this idea by comparing the network's performance in working memory tasks whose targets are drawn from distributions with different intertrial dependencies (Fig. 6). As a control, we consider the case of independent targets θ_n and θ_{n+1} (Fig. 6A). Responses are normally distributed about the true target angle. The dynamics of the bump encoding the target are shaped by both input fluctuations and a bias in the direction of the previous target on individual trials. However, the directional bias is not apparent in the entire distribution of response angles, since it samples from all possible pairs (θ_{n+1}, θ_n). An ensemble-wide measure of performance is given by the standard deviation of the response distribution ($\sigma \approx 4.42$). When the current target angle depends on the previous angle, the relative response distribution narrows (Fig. 6B). Memory of the previous trial's target θ_n stabilizes the memory of the current trial's target θ_{n+1}, decreasing the standard deviation of responses ($\sigma \approx 3.20$). There is a high probability the current target θ_{n+1} will be close to the previous target θ_n, so the timescale of the network's underlying inference process is reasonably well matched to the environment. However, such effects can be deleterious when the previous angle θ_n is skewed in comparison to the current angle θ_{n+1}. Protocols with this angle dependence lead to a systematic bias in the relative response distribution, so its peak is shifted away from zero (Fig. 6C).

Our neuronal network model predicts that, if an intertrial bias is present in the computations of a neural circuit, it should be detectable by varying the intertrial dependence of target angles θ_n. Furthermore, when there are strong local correlations between adjacent trials (P(θ_{n+1}, θ_n) is large for $|\theta_{n+1} - \theta_n|$ small), responses are more accurate than for protocols with independent adjacent trial angles. Since the strength of the bias increases as the intertrial interval is decreased, due to the decay of STF, we expect the effect to be more pronounced for trials taken closer together.

Two timescales of memory degradation. Wimmer *et al.*[6] have shown that both the normal distribution of saccade endpoints and observed changes in neural firing rates during the delay-period can be replicated by a diffusing bump attractor model[6]. We have shown that a recurrent network with STF (Fig. 1E) still leads to a normal distribution of predicted response angles (Fig. 6A). Our model also provides new predictions for the dynamics of memory degradation, which we now compare with the standard diffusing bump attractor model[9,47] (Fig. 7). In a network with STF (Fig. 7A), bump trajectories evolve in a history-dependent fashion (Fig. 7B). Initially, bumps diffuse freely, but are eventually drawn to their starting location by facilitated synapses (See also Fig. 4). This leads to two distinct phases of diffusion, as shown in plots of the bump variance (Fig. 7C). Rapid diffusion occurs initially, as the bump equilibrates to the quasistationary density determined by the slowly evolving potential (Fig. 4). Slower diffusion occurs subsequently, as spatial heterogeneity in synaptic architecture gradually responds to changes in bump position via STF. Stabilizing effects of STF on bump attractors have been analyzed previously[28], but our identification of these multiple timescale dynamics is novel. This feature of the bump dynamics is not present in networks with static synapses (Fig. 7D). Here, bumps evolve as a noise-driven particle over a flat potential landscape (Fig. 7E), described by Brownian motion: a memoryless stochastic process[41,46]. Variance in bump position scales purely linearly with time (Fig. 7F), and the diffusion coefficient can be computed using a low-dimensional approximation[47].

The qualitative differences between the bump attractor with and without dynamic synapses should be detectable in both behavioral and neurophysiological recordings[6]. Moreover, the observed intertrial bias identified in recent analyses of behavioral data requires some mechanism for transferring information between trials that is

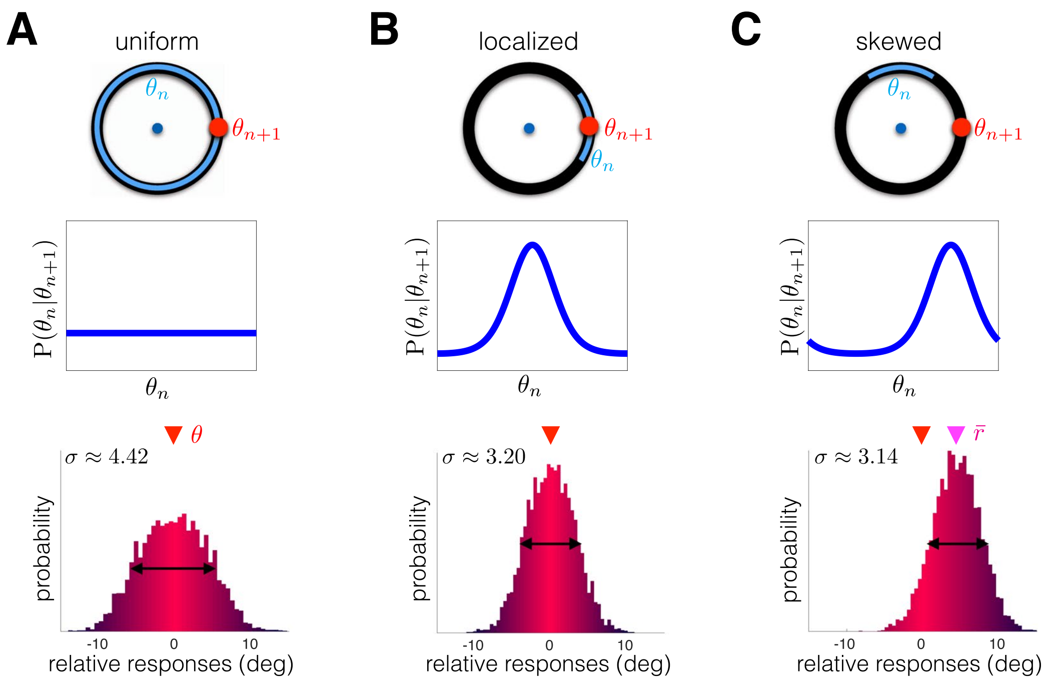

Figure 6. Response distribution is shaped by dependence between target angles in adjacent trials $P(\theta_{n+1}|\theta_n)$. (**A**) Visual working memory protocols typically use a sequence of targets with no trial-to-trial dependence (uniform $P(\theta_{n+1}, \theta_n)$, shown for $\theta_n \equiv 0°$)[6,9]. Relative responses $(r_n - \theta_n)$ are normally distributed about the true target. (**B**) Current target θ_{n+1} depends on the previous target θ_n according to a locally peaked distribution. The response distribution narrows (note decreased standard deviation σ), since the target θ_{n+1} is often close to the previous target θ_n. (**C**) Current target θ_{n+1} is skewed clockwise from previous angle θ_n. Responses are thus skewed counter-clockwise towards the previous target (note average response $\bar{r}$ is shifted). Numerical methods are described in Methods.

distinct from neural activity[13], as dynamic synapses are in our model. In total, our model provides both an intuition for the behavioral motivation as well as neurophysiological mechanisms that produce such interference.

Discussion

Neural circuit models of visual working memory tend to use neural activity variables as the encoders of target locations. Our computational models account for interference in visual working memory using both suboptimal Bayesian inference and STF acting on a recurrent network model of delay-period activity. The timescale and prior target dependence of attractive biases we observe correspond to psychophysical observations of behavioral experiments in monkeys[13]. STF evolves dynamically over seconds[45,50], matching the kinetics of interference in visual working memory. The link we have drawn between our two models suggests neural circuits can implement probabilistic inference using short-term plasticity.

Experimental predictions. More complete descriptions of the neural mechanics of visual working memory could be accomplished by analyzing the effects of correlations in sequential target presentations. Since responses in subsequent trials are shaped by the previous trial's target[13], computational models can be validated by determining how well their response distributions reflect trial-to-trial target correlations (Fig. 6). It is also possible that the introduction of target sequences whose distributions change in time could impact quantitative features of interference. For instance, implementing tasks with target sequences that have multiple trial correlations may extend the timescale of interference beyond a single trial. Furthermore, our model predicts that multiple timescales emerge in the statistics of delay-period activity during a working memory task (Fig. 7). Variance of recall error increases sublinearly in our model, consistent with a recent reanalysis of psychophysical data of saccades to remembered visual targets[4,51]. The dynamics of our model are thus inconsistent with the purely linear diffusion of recall error common in bump attractor models with static synapses[6,9].

The idea that STF may play a role in working memory is not new[27,52], and there is evidence that prefrontal cortex neurons exhibit dynamic patterns of activity during the delay-period, suggestive of an underlying modulatory process[53]. However, it remains unclear how the presence of STF may shape the encoding of working memories. Our model suggests STF can transfer attractive biases between trials. Recent findings on the biophysics of STF could be harnessed to examine how blocking STF shapes behavioral biases in monkey experiments[54]. We predict that reducing the effects of the STF would both decrease the systematic bias in responses and increase the amplitude of errors, since the stabilizing effect of STF on the persistent activity will be diminished[28].

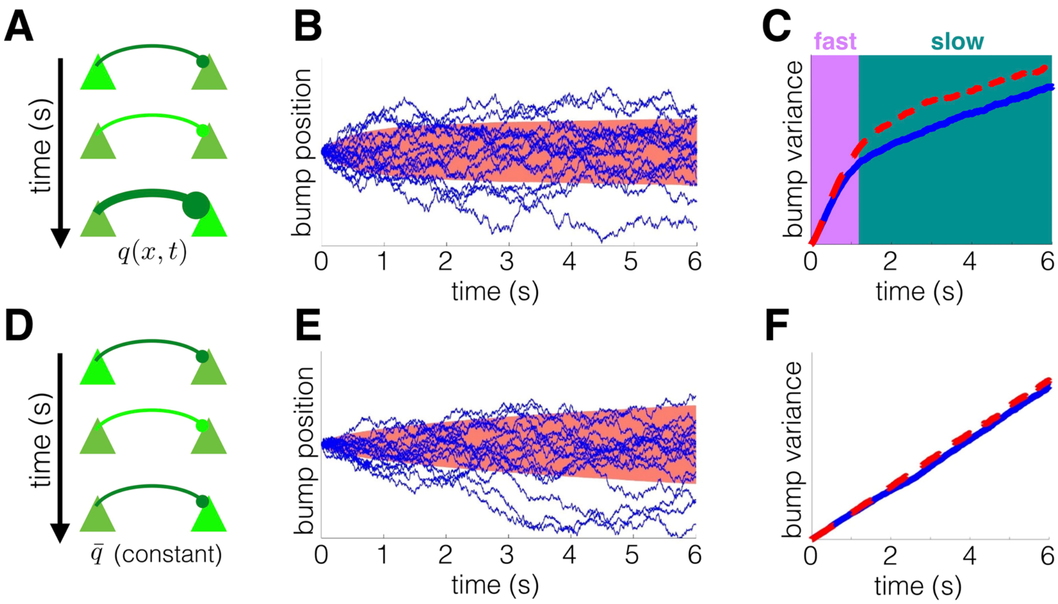

Figure 7. Recurrent networks with STF (panels A–C) exhibit two timescales of delay-period dynamics, in contrast to the single timescale dynamics of networks with static synapses (panels D–F). (**A**) Release of neurotransmitter leads to the strengthening of the synapse via STF. (**B**) In a facilitating network, bump trajectories (lines) stray less from their initial position due to the attractive effect of STF. Large ensemble standard deviation shown in red. (**C**) STF generates two phases of variance scaling. An initial fast phase is followed by a slower phase due to the dampening effect of STF in both neuronal network (red dashed) and low-dimensional (blue solid) simulations. (**D**) Network with static synapses. (**E**) Bump trajectories obey linear diffusion, due to the spatial homogeneity of the network. (**F**) Variance grows linearly with time, a hallmark of pure diffusion. See Methods for further details.

Comparison with previous work. The work of Papadimitriou *et al.*[13,55] also contains modeling studies, accounting for some aspects of their experimental observations. Our computational model differs from and extends their findings in several important ways. We propose that interference can arise as a suboptimal inference process, which can be implemented by concrete synaptic mechanisms. This conclusion can only be drawn from a tractable model, allowing us to reduce our recurrent network's dynamics to the low-dimensional system, Eq. (6). Furthermore, Papadimitriou *et al.*[13] employ a two-store model of memory that is not linked to a specific physiological mechanism, whereas we propose STF and use a well tested model of its kinetics[56]. Lastly, Papadimitriou *et al.*[55] present recorded data from the frontal eye fields showing no firing rate tuning to the previous target during the current target onset. While this observation contradicts their purely activity-based description of the bias proposed in Papadimitriou *et al.*[13], this is not an issue for the STF-based bias we propose here. The mechanism we propose is gradual and initially silent within the current trial, revealing its effects toward the end of the delay period, so it is consistent with the findings of Papadimitriou *et al.*[55].

Alternative neurophysiological mechanisms for intertrial bias. Our study accounts for biases observed by Papadimitriou *et al.*[13], who identified an attraction between the previous target and current response. Strengthening synapses that originate from recently active neurons can attract neural activity states in subsequent trials. This is consistent with recent experiments showing latent and "activity-silent" working memories can be reactivated in humans using transcranial magnetic stimulation[57], suggesting working memory is maintained by mechanisms other than target-tuned persistent neural activity[27,53]. The principle of using short-term plasticity to store memories of visual working memory targets could be extended to account for longer timescales and more intricate statistical structures. Short-term depression (STD) could effect a repulsive bias on subsequent responses, since neural activity would be less likely to persist in recently-activated depressed regions of the network[58]. In this way, STD could encode a predictive distribution for targets that are anti-correlated to the previously present target.

Other physiological mechanisms could also shape network responses to encode past observations in a predictive distribution. Long-term plasticity is a more viable candidate for encoding predictive distributions that accumulate observations over long timescales. Consider a protocol that uses the same distribution of target angles throughout an entire experiment, but this distribution is biased towards a discrete set of possible angles[42]. For a recurrent network to represent this distribution, it must retain information about past target presentations over a long timescale. Many biophysical processes underlying plasticity are slow enough to encode information from such lengthy sequences[59,60]. Furthermore, the distributed nature of working memory suggests that there may be brain regions whose task-relevant neural activity partially persists from one trial to the next[61]. Such activity could shape low-level sensory interpretations of targets in subsequent trials.

Synaptic plasticity can stabilize working memory. Several modeling studies of working memory have focused on the computational capability of synaptic dynamics[62]. For instance, STF can prolong the lifetime of working memories in spatially heterogeneous networks, since facilitated synapses slow the systematic drift of bump attractor states[28,63]. This is related to our finding that STF reduces the diffusion of bumps in response to dynamic fluctuations (Fig. 7B), generating two timescales of memory degradation, corresponding to the bump variance (Fig. 7C). This scaling may be detectable in neural recordings or behavioral data, since recall errors may saturate if stabilized by STF. Facilitation can also account for experimentally observed increases in spike train irregularity during the delay-period in circuit models that support tuned persistent activity[64]. Alternatively, homeostatic synaptic scaling can compensate for spatial heterogeneity, which would otherwise cause persistent states to drift[10]. However, the short homeostatic timescales often suggested in models do not often match experimental observations[65].

Models of working memory have also replaced persistent neural firing with stimulus-selective STF, so neuronal spiking is only required for recall at the end of the delay-period[27]. One advantage of this model is that multiple items can be stored in the dynamic efficacy of synapses, and the item capacity can be regulated by external excitation for different task load demands[29]. Our model proposes that STF plays a supporting rather than a primary role, and there is extensive neurophysiological evidence corroborating persistent neural activity as a primary working memory encoder[6,66].

Robust working memory via excitatory/inhibitory balance. Computational modeling studies have demonstrated that a balance of fast inhibition and slow excitation can stabilize networks, so they accurately integrate inputs[40,46,67]. Drift in the representation of a continuous parameter can be reduced by incorporating negative-derivative feedback into the firing rate dynamics of a network, similar to introducing strong friction into the mechanics of particle motion on a sloped landscape[68]. Fast inhibition balanced by slower excitation produces negative feedback that is proportional to the time-derivative of population activity. A related mechanism can be implemented in spiking networks wherein fast inhibition rapidly prevents runaway excitation, and the resulting network still elicits highly irregular activity characteristic of cortical population discharges[69]. Mutually inhibiting balanced networks are similarly capable of representing working memory of continuous parameters[70], and extending our framework by incorporating STF into this paradigm would be a fruitful direction of future study.

Extensions to multi-item working memory. Working memory can store multiple items at once, and the neural mechanisms of interference between simultaneously stored items are the focus of ongoing work[71,72]. While there is consensus that working memory is a limited resource allocated across stored items, controversy remains over whether resource allocation is quantized (e.g., slots)[73,74] or continuous (e.g., fluid)[71,75]. Spatially-organized neural circuit models can recapitulate inter-item biases observed in multi-item working memory experiments, and provide a theory for how network interactions produce such errors[76,77]. In these models, each remembered item corresponds to an activity bump, and the spatial scale of lateral inhibition determines the relationship between recall error and item number[78]. The model provides a theory for attractive bias and forgetting of items since nearby activity bumps merge with one another. This is related to the mechanism of attractive bias in our model, but a significant difference is that previous models only required localized excitation whereas we use STF. It would be interesting to identify the temporal dynamics of biases in multi-item working memory to see if they suggest slower timescale processes like short-term plasticity.

Tuning short-term plasticity to the environmental timescale. We have not identified a mechanism whereby our network model's timescale of inference could be tuned to learn the inherent timescale of the environment. There is recent evidence from decision-making experiments that humans can learn the timescale on which their environment changes and use this information to weight their observations toward a decision[21,79]. Our model suggests that the trial-history inference utilized by subjects in Papadimitriou *et al.*[13] is significantly suboptimal, perhaps because it is difficult to infer the timescale of relevant past-trial information. The complexity, sensitivity, and resource expense of optimal inference in most contexts likely makes it impossible to implement exactly in neural circuits[80,81]. This may explain why humans often use suboptimal methods for accumulating evidence[21,23,82]. Plasticity processes that determine the timescale of evidence accumulation may be shaped across generations by evolution, or across a lifetime of development. Nonetheless, metaplasticity processes can internally tune the dynamics of plasticity responses in networks without changing synaptic efficacy itself, and these changes could occur in a reward-dependent way[83,84]. Recently, a model of reward-based metaplasticity was proposed to account for adaptive learning observed in a probabilistic reversal learning task[85]. Such a process could modify the timescale and other features of short-term plasticity in ways that improve task performance in working memory as well.

Conclusions

Our results suggest that interference observed in visual working memory tasks can be accounted for by a persistently active neural circuit with STF. Importantly, interference is graded by the time between trials and during a trial. The interplay of synaptic and neural processes involved in interference may have arisen as a robust system for processing visual information that changes on the timescale of seconds. More work is need to determine how information about the environment stretches across multiple timescales to shape responses in cognitive tasks. We expect that identifying the neural origin of such biases will improve our understanding of how working memory fits into the brain's information-processing hierarchy.

Symbol	Description
θ_n	target observed on trial n
$s_{n+1}(\theta)$	unknown target distribution observer attempts to infer before trial $n+1$
$f_{\theta_n}(\theta)$	probability of observing target θ in the next trial given θ_n was observed previously, and the environment has not changed since
ε	assumed environmental change rate used to discount evidence
$L_{n+1,\theta}$	*predictive distribution*: inferred probability of seeing target θ in trial $n+1$, given past observations $\theta_{1:n}$ and assumed change rate ε
l_n	*run-length*: number of trials prior to trial $n+1$ with same distribution
$\bar{\mathrm{P}}_0$	*uniform prior*: distribution defining $\theta \in [-180°, 180°)$ with same probability

Table 1. Variables and parameters of the probabilistic inference model.

Methods

Assumptions of the inference model. Our model performs nonparametric density estimation to approximate the distribution $s_{n+1}(\theta)$ from which a target θ will be drawn in trial $n+1$. The observer assumes the possible distributions $s(\theta)$ are drawn from a function space $s \in S$ according to the prior $p(s)$. We assume that marginalizing over this space of distributions yields the uniform density $\bar{\mathrm{P}}_0 = \int_S s(\theta)p(s)\mathrm{d}s = 1/360$. One possibility is that the distribution $s_{n+1}(\theta)$ is constructed by drawing N-tuples $\mathbf{a}$ and ψ from a uniform distribution over the hypercubes $[0, a_{max}]^N$ and $[-180°, 180°)^N$ and using the entries in an exponential distribution of a sum of cosines:

$$s_{n+1}(\theta) = \mathcal{N}_s \exp\left[\sum_{j=1}^{N} a_j \cos(\omega_j \cdot (\theta - \psi_j))\right],$$

where $\omega_j = j\pi/180$ and $\mathcal{N}_s$ is a normalization constant. For instance, when $N=1$,

$$s_{n+1}(\theta) = \mathcal{N}_s \exp\left[a_1 \cos(\omega_1 \cdot (\theta - \psi_1))\right],$$

peaked at ψ_1. For the main instantiation and reduction of our model, knowing the specific family of distributions is unnecessary.

The probability $f_{\theta'}(\theta) := \mathrm{P}(\theta_{n+1} = \theta | \theta_n = \theta', s_{n+1}(\theta) \equiv s_n(\theta))$ is defined under static conditions ($s_{n+1}(\theta) \equiv s_n(\theta)$) to separate the dynamic effects of sampling distribution $s_n(\theta)$ changes. We are performing nonparametric Bayesian estimation of the distribution, and the probability $f_{\theta'}(\theta)$ is already marginalized over the space of distributions $s(\theta)$. Thus, we do not model the intermediate step of inferring the probability of each distribution $s(\theta)$ and marginalizing, but it could be computed by integrating over the prior on the function space, $f_{\theta'}(\theta) = \int_S s(\theta) f(s|\theta')\mathrm{d}s$. Each observation θ' would give the probability $f(s|\theta')$ that the current distribution is $s(\theta)$. Integrating over the space of all distributions $s \in S$ provides the probability the next target will be θ, based on the previous observation θ' alone and the assumption that the distribution remains the same from trial n to $n+1$. Further details on the difference between parametric and nonparametric Bayesian estimation of densities can be found in Orbanz and Teh[86]. Note, we assume self-conjugacy of $f_{\theta'}(\theta) = f_{\theta}(\theta')$, which follows since the order of observations does not matter while the environment remains fixed. This relationship will also make the predictiveness of our model more apparent. It is important to note that the observer assumes the form of $f_{\theta'}(\theta)$, but this is not necessarily the distribution an ideal observer should use. For illustration, we consider a family of distributions given by an exponential of cosines:

$$f_{\theta'}(\theta) = \mathcal{N}_\theta \exp\left[\sum_{j=1}^{N} a_j \cos(\omega_j \cdot (\theta - \theta'))\right], \tag{7}$$

for $\omega_j = j\pi/180$, which is self-conjugate: $f_{\theta'}(\theta) \equiv f_{\theta}(\theta')$[87]. A distribution like Eq. (7) would emerge from a generative model with distance-dependent spatial correlations in the ensemble of produced targets. The example $f_{\theta'}(\theta)$ we use for comparison with our recurrent network with STF is close to the case of Eq. (7) with $N=1$. A description of the parameters and variables in our model is provided in Table 1.

Derivation of the probabilistic inference model. The observer's predictive distribution $L_{n+1,\theta} = \mathrm{P}(\theta_{n+1}|\theta_{1:n}, \varepsilon)$ is derived by computing the probability of observing θ_{n+1} given each prior observation θ_j, $j = 1, \ldots, n$. Importantly, we must compute the probability of each run length $l_n = l$, $l = 0, \ldots, n$, corresponding to the number of trials the assumed underlying distribution $s_n(\theta)$ has remained the same[30,32]. Knowing the probability of each run length will inform us of how much to weight each observation θ_j, $j = 1, \ldots, n$. In particular, $l_n = n$ indicates the environment has remained the same since the first trial, and $l_n = 0$ indicates the environment changes between trial n and $n+1$. Summing over all possible run lengths, the marginal predictive distribution is

$$L_{n+1,\theta} = \sum_{l=0}^{n} \mathrm{P}(\theta_{n+1}|l_n = l, \theta_{1:n}^{l})\mathrm{P}(l_n = l|\theta_{1:n}, \varepsilon), \tag{8}$$

where $\mathrm{P}(\theta_{n+1}|l_n = l, \theta_{1:n}^{l})$ is the conditional predictive distribution assuming run length $l_n = l$ with the special case $\mathrm{P}(\theta_{n+1}|l_n = 0, \theta_{1:n}^{0}) = \bar{\mathrm{P}}_0$ (the uniform distribution), and $\mathrm{P}(l_n = l|\theta_{1:n})$ is the conditional probability of the run length $l_n = l$ given the series of target angles $\theta_{1:n}$. We further simplify Eq. (8) as follows: First, utilizing sequential analysis, we find that if the present run length is $l_n = l$, the conditional predictive distribution is given by the product of probabilties from the last l observations[22]:

$$\mathrm{P}(\theta_{n+1}|l_n = l,\, \theta_{1:n}^l) = \frac{\overline{\mathrm{P}}_0}{\mathrm{P}(\theta_{n-l+1:n})} \prod_{j=n-l+1}^{n} f_{\theta_j}(\theta), \tag{9}$$

where $\overline{\mathrm{P}}_0$ is the uniform distribution, $\mathrm{P}(\theta_{n+1:n}) = \prod_{j=n+1}^{n} f_{\theta_j}(\theta) = 1$, and we have utilized our self-conjugacy assumption for $f_{\theta'}(\theta) \equiv f_\theta(\theta')$. Next, we assume that observations provide no information about the present run length r_n, which would be a consequence of the observer making no a priori assumptions on the overall distribution from which targets $\theta_{1:n}$ are drawn. Thus, the observer only uses their knowledge of the change rate of the environment ε to determine the probability of a given run length $l_n = l$, and the conditional probability can be computed

$$\mathrm{P}(l_n = l|\theta_{1:n}, \varepsilon) = \mathrm{P}(l_n = l|\varepsilon) = \begin{cases} \varepsilon(1-\varepsilon)^l, & l < n, \\ (1-\varepsilon)^n, & l = n. \end{cases} \tag{10}$$

Plugging Eqs (9–10) into the update Eq. (8), we find the probability of the next target being at angle $\theta_{n+1} = \theta$, given that the previous n targets were $\theta_{1:n}$, is:

$$L_{n+1,\theta} = \overline{\mathrm{P}}_0 \cdot \left[\frac{(1-\varepsilon)^n}{\mathrm{P}(\theta_{1:n})} \prod_{j=1}^{n} f_{\theta_j}(\theta) + \varepsilon \sum_{l=0}^{n-1} \frac{(1-\varepsilon)^l}{\mathrm{P}(\theta_{n-l+1:n})} \prod_{j=n-l+1}^{n} f_{\theta_j}(\theta) \right].$$

Limit of slowly-changing environment (small ε). Here, we examine the case $0 < \varepsilon \ll 1$, where the environment changes very slowly. Assuming independence of the target angles selected on each trial $\theta_{1:n}$[35], $\mathrm{P}(\theta_{n-l:n}) = \mathrm{P}(\theta_{n-l:n-1})\mathrm{P}(\theta_n)$, we can split the probabilities over the target sequences $\theta_{n-l:n}$ into products: $\mathrm{P}(\theta_{n-l:n}) = \prod_{j=n-l}^{n} \mathrm{P}(\theta_j) = \overline{\mathrm{P}}_0^{l+1}$. The last equality holds because the unconditioned probability of a particular target location is uniform $\overline{\mathrm{P}}_0$. Applying this assumption to Eq. (1) and truncating to $\mathcal{O}(\varepsilon)$, we have

$$\widetilde{L}_{n+1,\theta} = \mathcal{N}_s \cdot \left[(1-n\varepsilon) \prod_{j=1}^{n} \frac{f_{\theta_j}(\theta)}{\overline{\mathrm{P}}_0} + \varepsilon \sum_{l=0}^{n-1} \prod_{j=n-l+1}^{n} \frac{f_{\theta_j}(\theta)}{\overline{\mathrm{P}}_0} \right],$$

noting $\prod_{j=n+1}^{n} \frac{f_{\theta_j}(\theta)}{\overline{\mathrm{P}}_0} = 1$, and we must choose $\mathcal{N}_s$ so $\int_{-180}^{180} L_{n+1,\theta} \mathrm{d}\theta = 1$, normalized at each step.

Limit of rapidly-changing environment ($\varepsilon \approx 1$). Here, we examine the case $\varepsilon \approx 1$ $(0 < (1-\varepsilon) \ll 1)$, a rapidly-changing environment. Applying this assumption to Eq. (1), we find $L_{n+1,\theta}$ is dominated by terms of order $(1-\varepsilon)$ and larger. Terms of order $(1-\varepsilon)^2$ are much smaller. For instance, we can approximate to linear order, dropping terms of $\mathcal{O}((1-\varepsilon)^2)$, to reduce Eq. (1) to

$$L_{n+1,\theta} \approx \varepsilon \overline{\mathrm{P}}_0 \left[1 + \frac{1-\varepsilon}{\mathrm{P}(\theta_n)} f_{\theta_n}(\theta) \right]. \tag{11}$$

Furthermore, we ensure the expression in Eq. (11) is normalized by writing

$$\widetilde{L}_{n+1,\theta} = \frac{\overline{\mathrm{P}}_0 + (1-\varepsilon) f_{\theta_n}(\theta)}{2-\varepsilon},$$

since $\int_{-180}^{180} [\overline{\mathrm{P}}_0 + (1-\varepsilon) f_{\theta_n}(\theta)]\, \mathrm{d}\theta = 2-\varepsilon$. Alternatively, we can truncate by multiplying through by $[1-(1-\varepsilon)]/[1-(1-\varepsilon)]$, truncating to $\mathcal{O}(1-\varepsilon)$ and normalizing to yield

$$\widetilde{L}_{n+1,\theta} = \varepsilon \overline{\mathrm{P}}_0 + (1-\varepsilon) f_{\theta_n}(\theta),$$

the key update equation in our Results (Figs 2 and 3A). Higher order approximations are obtained by keeping more terms from Eq. (1); e.g., a second order approximation yields

$$L_{n+1,\theta} \approx \varepsilon \overline{\mathrm{P}}_0 + \varepsilon(1-\varepsilon) f_{\theta_n}(\theta) + \frac{\varepsilon(1-\varepsilon)^2 \overline{\mathrm{P}}_0}{\mathrm{P}(\theta_{n-1:n})} f_{\theta_n}(\theta) f_{\theta_{n-1}}(\theta),$$

successively downweighting the influence of previous observations (θ_{n-1}).

Relating the predictive distribution to the potential of an attractor model. A predictive distribution can be represented by an attractor model by first determining the formula of the stationary distribution of Eq. (4), given an arbitrary potential function $\mathcal{U}_{n+1}(\theta)$. Equation (4) can be reformulated as an equivalent Fokker-Planck equation for the represented angle θ during trial $n+1$ assuming the present potential function is $\mathcal{U}_{n+1}(\theta)$[88],

Symbol	Description
$\theta(t)$	observer's estimate of the remembered target during the delay-period
$\mathcal{U}_{n+1}(\theta)$	potential determining the dynamics of θ during the delay-period
σ_θ	standard deviation of the dynamic fluctuations
$p_{n+1}(\theta, t)$	probability density function (pdf) of observer's estimate θ at time t
$\bar{p}_{n+1}(\theta)$	*stationary density*: pdf of estimate θ over long time, $t \to \infty$

Table 2. Variables and parameters of the particle evolving on a potential model.

$$\frac{\partial p_{n+1}(\theta, t)}{\partial t} = \frac{\partial}{\partial \theta}\left[\frac{\mathrm{d}\mathcal{U}_{n+1}(\theta)}{\mathrm{d}\theta} p_{n+1}(\theta, t)\right] + \frac{\sigma_\theta^2}{2}\frac{\partial^2 p_{n+1}(\theta, t)}{\partial \theta^2}, \tag{12}$$

where $p_{n+1}(\theta, t)$ is the probability density corresponding to the target angle estimate θ at time t. The initial estimate of the target is exact, $\theta(0) = \theta_{n+1}$, so $p_{n+1}(\theta, 0) = \delta(\theta - \theta_{n+1})$ is the initial condition. We summarize the constituent variables and model parameters in Table 2.

We now derive the form of $\mathcal{U}_{n+1}(\theta)$ that leads to a stationary density corresponding the predictive distribution $L_{n+1,\theta}$ in the limit $t \to \infty$ in Eq. (12). The stationary density $\bar{p}_{n+1}(\theta)$ is analogous to a predictive distribution represented by Eq. (4) since it is the probability the system represents when no information about the current trial's target θ_{n+1} remains. Thus, we build a rule to update $\mathcal{U}_{n+1}(\theta)$ to mirror the update of $L_{n+1,\theta}$ in Eq. (3). To obtain this result, we match the stationary density for Eq. (12) to the updated predictive distribution:

$$\lim_{t\to\infty} p_{n+1}(\theta, t) = \bar{p}_{n+1}(\theta) = L_{n+1,\theta}. \tag{13}$$

Solving Eq. (12) for its stationary solution, we find that during trial $n+1$:

$$\bar{p}_{n+1}(\theta) = \chi_{n+1}\exp\left[-\frac{2\mathcal{U}_{n+1}(\theta)}{\sigma_\theta^2}\right], \tag{14}$$

where χ_{n+1} is a normalization factor chosen so that $\int_{-180}^{180}\bar{p}_{n+1}(\theta)\,\mathrm{d}\theta = 1$. Plugging Eq. (14) into Eq. (13) and solving for $\mathcal{U}_{n+1}(\theta)$, we obtain

$$\mathcal{U}_{n+1}(\theta) = \frac{\sigma_\theta^2}{2}\ln\frac{\chi_{n+1}}{L_{n+1,\theta}}.$$

For a rapidly changing environment $0 < (1-\varepsilon) \ll 1$, we approximate $L_{n+1,\theta}$ using Eq. (3) so that

$$\begin{aligned}\mathcal{U}_{n+1}(\theta) &= \frac{\sigma_\theta^2}{2}[\ln\chi_{n+1} - \ln(\varepsilon\bar{\mathrm{P}}_0 + (1-\varepsilon)f_{\theta_n}(\theta))]\\ &\approx \frac{\sigma_\theta^2}{2}\left[\ln\frac{\chi_{n+1}}{\bar{\mathrm{P}}_0} - (1-\varepsilon)\frac{f_{\theta_n}(\theta) - \bar{\mathrm{P}}_0}{\bar{\mathrm{P}}_0}\right],\end{aligned}$$

where we have linearized in $(1-\varepsilon)$. However, for Eq. (4), only the derivative of $\mathcal{U}_{n+1}(\theta)$ impacts the dynamics, so we drop the additive constants and examine the proportionality

$$\mathcal{U}_{n+1}(\theta) \propto -f_{\theta_n}(\theta).$$

In the limit of weak interactions between trials, the potential $\mathcal{U}_{n+1}(\theta)$ should be shaped like the negative of the probability $f_{\theta_n}(\theta)$ based on the previous trial's target θ_n.

Bump attractor model with short-term facilitation. Our neuronal network model is comprised of two variables evolving in space $x \in [-180, 180)°$, corresponding to the stimulus preference of neurons at that location, and time $t > 0$. Variables and parameters are summarized in Table 3, and the evolution equations consist of one stochastic integrodifferential equation and one auxiliary differential equation:

$$\tau_u du(x, t) = [-u(x, t) + w(x) * [(1 + q(x, t))F(u(x, t))] + I(x, t)]\,\mathrm{d}t + \mathrm{d}W(x, t), \tag{15a}$$

$$\tau\dot{q}(x, t) = -q(x, t) + \beta F(u(x, t))\,(q_+ - q(x, t)), \tag{15b}$$

where $u(x, t)$ describes the evolution of the normalized synaptic input at location x. The model Eq. (15) can be derived as the large system size limit of a population of synaptically coupled spiking neurons[89], and similar dynamics have been validated in spiking networks with lateral inhibitory connectivity[6,9]. We fix the timescale of dynamics by setting $\tau_u = 10$ ms, so time evolves according to units of a typical excitatory synaptic time constant[90].

Symbol	Description
$u(x, t)$	synaptic input to location x at time t
$q(x, t)$	short term facilitation (STF) increases the strength of connectivity from neurons at location x in response to their activation
τ_u	timescale of synaptic excitation, set to 10 ms in simulations[90]
$w(x-y)$	baseline strength and polarity of synaptic connectivity from y to x
$F(u)$	nonlinearity that converts synaptic input to output population rate
$I(x, t)$	external input due to visual stimuli
$W(x, t)$	spatiotemporal noise due to dynamic fluctuations
τ	timescale of STF, set to 1 s in simulations[26]
β	onset rate of STF, set to 0.01 in simulations[27]
q_+	maximal increase in synaptic utilization, set to 2 in simulations[27]

Table 3. Variables and parameters of the recurrent network model.

This population rate model can be explicitly analyzed to link the architecture of the network to a low-dimensional description of the dynamics of a bump attractor as described by Eq. (4).

Each location x in the network receives recurrent coupling defined by the weight function $w(x-y)$ via a convolution $w(x) * g(x) = \int_{-180}^{180} w(x-y)g(y)\,dy$. We take this function to be peaked when $x=y$ and decreasing as the distance $|x-y|$ grows, in line with anatomical studies of delay-period neurons in prefrontal cortex[8]. We do not separately model excitatory and inhibitory populations, but Eq. (15) can be derived from a model with distinct excitatory and inhibitory populations in the limit of fast inhibitory synapses[43,67]. Thus, we have combined excitatory and inhibitory populations, so $w(x-y)$ takes on both positive and negative values. Our analysis can be applied to a general class of distance-dependent connectivity functions, given by an arbitrary sum of cosines $w(x-y) = \sum_{n=0}^{\infty} \alpha_n \cos(\omega_n(x-y))$ where $\omega_n = n\pi/180$, and we will use a single cosine to illustrate in examples: $w(x-y) = \cos(\omega_1(x-y))$. The nonlinearity $F(u)$ converts the normalized synaptic input $u(x, t)$ into a normalized firing rate, $F(u) \in [0, 1]$. We take this to be sigmoidal $F(u) = 1/[1 + e^{-\gamma(u-\kappa)}]$[91], with a gain of $\gamma = 20$ and a threshold of $\kappa = 0.1$ in numerical simulations. In the high-gain limit ($\gamma \to \infty$), a Heaviside step function $F(u) = H(u-\kappa)$ allows for explicit calculations[43,89].

Recurrent coupling is shaped by STF in active regions of the network ($F(u) > 0$), as described by the variable $q(x, t) \in [0, q_+]$; $q_+ > 0$ and β determine the maximal increase in synaptic utilization and the rate at which facilitation occurs[26,56]. For our numerical simulations, we consider the parameter values $q_+ = 2$ and $\beta = 0.01$, consistent with previous models employing facilitation in working memory circuits[27–29] and experimental findings for facilitation responses in prefrontal cortex[45,50]. The timescale of plasticity is slow, $\tau = 1000\,\text{ms} \gg 10\,\text{ms}$, consistent with experimental measurements[26]. Our qualitative results are robust to parameter changes. Information from the previous trial is maintained by the slow-decaying kinetics of the facilitation variable $q(x, t)$, even in the absence of neural activity[27,29].

Effects of the target and the response are described by the deterministic spatiotemporal input $I(x, t)$, which we discuss more in detail below. The noise process $W(x, t)$ is white in time and has an increment with mean $\langle dW(x, t)\rangle \equiv 0$ and spatial correlation function $\langle dW(x, t)dW(y, s)\rangle = C(x-y)\delta(t-s)dtds$. In numerical simulations, we take our correlation function to be $C(x-y) = \sigma_W^2 \cos(x-y)$ with $\sigma_W = 0.005$, so the model recapitulates the typical 1–5% standard deviation in saccade endpoints observed in oculomotor delayed-response tasks with delay-periods from 1–10 s[1,4,6].

Implementing sequential delayed-response task protocol. A series of oculomotor delayed-response tasks is executed by the network Eq. (15) by specifying a schedule of peaked inputs occurring during the cue periods of length T_C, no input during trial n's delay-period of length T_D^n, and brief and strong inhibitory input of length T_A after the response has been recorded, and then no input until the next trial. This is described by the spatiotemporal function

$$I(x, t) = \begin{cases} I_0 \exp[I_1(\cos(x-\theta_n)-1)], & t \in [t_n, t_n + T_C), \\ 0, & t \in [t_n + T_C, t_n + T_C + T_D^n), \\ -I_R, & t \in [t_n + T_C + T_D^n, t_n + T_C + T_D^n + T_A), \\ 0, & t \in [t_n + T_C + T_D^n + T_A, t_{n+1}), \end{cases}$$

for all $n = 1, 2, 3, \ldots$, where t_n is the starting time of the n^{th} trial which has cue period T_C, delay-period T_D^n, inactivation period T_A, and subsequent intertrial interval T_I^n. Note that the delay and intertrial interval times may vary trial-to-trial, but the cue is always presented for the same period of time as in[13]. The amplitude of the cue-related stimulus is controlled by I_0, and I_1 controls is sharpness. Activity from trial n is ceased by the global inactivating stimulus of amplitude I_R.

In numerical simulations, we fix the parameters $T_C = 150$ ms; $T_A = 500$ ms; $I_0 = I_1 = 1$; and $I_R = 2$. Target locations θ_n are drawn from a uniform probability mass function (pmf) for the discrete set of angles $\theta_n \in \{-180°, -162°, \ldots, 162°\}$ to generate statistics in Fig. 5A, which adequately resolves the bias effect curves for comparison with the results in[13]. Intertrial intervals are varied to produce Fig. 5B by drawing $T_I^n := t_{n+1} - (T_C + T_D^n + T_A)$ randomly from a uniform pmf for the discrete set of times $T_I^n \in \{1000, 1200, \ldots, 5000\}$ ms and θ_n randomly as in Fig. 5A and identifying the θ_n that produces the maximal bias for each value of T_I^n. Delay-periods are varied to produce Fig. 5C by drawing T_D^n randomly from a uniform pmf for the discrete set of times $T_I^n \in \{0, 200, \ldots, 5000\}$ ms and following a similar procedure to Fig. 5B. Draws from a uniform density function $\mathrm{P}(\theta_n) \equiv \overline{\mathrm{P}}_0$, defined on $\theta_n \in [-180, 180)°$ are used to generate the distribution in Fig. 6A and plots in Fig. 7. Nontrivial correlation structure in target selection is defined by the sum of a von Mises distribution and uniform distribution $\mathrm{corr}\,(\theta_{n+1}, \theta_n) = (1 - \varepsilon)\mathcal{N}_v e^{25\cos(\theta_n - \theta_{n+1} - \mu)} + \varepsilon\overline{\mathrm{P}}_0$ for fixed θ_n with $\varepsilon = 0.5$; $\mu = 0$ for local correlations (Fig. 6B) and $\mu = 90$ for skewed correlations (Fig. 6C).

The recurrent network, Eq. (15), is assumed to encode the initial target θ_n during trial n via the center-of-mass $\theta(t)$ of the corresponding bump attractor. Representation of the cue at the end of the trial is determined by performing a readout on the neural activity $u(x, t)$ at the end of the delay time for trial n: $t = t_n + T_C + T_D^n$. One way of doing this would be to compute a circular mean over x weighted by $u(x, t)$, but since $u(x, t)$ is a roughly symmetric and peaked function in x, computing $\theta(t) := \mathrm{argmax}_x u(x, t)$ (when $t \in [t_n, t_n + T_C + T_D^n)$) is an accurate and efficient approximation[6,42]. The bias and relative saccade endpoint on each trial n are then determined by computing the difference $\theta(t) - \theta_n$ (Figs 5, 6 and 7).

Deriving the low-dimensional description of bump motion. We analyze the mechanisms by which STF shapes the bias on subsequent trials by deriving a low-dimensional description for the motion of the bump position $\theta(t)$. To begin, note that in the absence of facilitation ($\beta \equiv 0$), the variable $q(x, t) \equiv 0$. In the absence of noise ($W(x, t) \equiv 0$), the resulting deterministic Eq. (15) has stationary bump solutions that are well studied and defined by the implicit equation[43,47,89]:

$$U(x) = \int_{-180}^{180} w(x - y)F(U(y))\,\mathrm{d}y.$$

Assuming the stimulus $I(x, t)$ presented during the cue period of trial n ($t \in [t_n, t_n + T_C)$) is strong enough to form a stationary bump solution, the impact of the facilitation variable $q(x, t)$ and noise $W(x, t)$ on $u(x, t)$ during the delay-period ($t \in [t_n + T_C, t_n + T_C + T_D^n)$) can be determined perturbatively, assuming $|q| \ll 1$ and $|\mathrm{d}W| \ll 1$. Since $\tau \gg \tau_u$, $u(x, t)$ will rapidly equilibrate to a quasi-steady-state determined by the profile of $q(x, t)$. We thus approximate the neural activity dynamics as $u(x, t) \approx U(x - \theta(t)) + \Phi(x, t)$, where $\theta(t)$ describes the dynamics of the bump center-of-mass during the delay-period ($|\theta| \ll 1$ and $|\mathrm{d}\theta| \ll 1$), and $\Phi(x, t)$ describes perturbations to the bump's shape ($|\Phi| \ll 1$). Plugging this approximation into Eq. (15) and truncating to linear order yields

$$d\Phi(x, t) - \mathcal{L}\Phi(x, t)\,\mathrm{d}t = U'(x)\,\mathrm{d}\theta + \int_{-180}^{180} w(x - y)q(y + \theta, t_s)F(U(y))\,\mathrm{d}y\mathrm{d}t + \mathrm{d}W, \tag{16}$$

where $\mathcal{L}u = -u + \int_{-180}^{180} w(x - y)F'(U(y))u(y)\,\mathrm{d}y$ is a linear operator and $q(x, t_s)$ is the facilitation variable evolving on the slow timescale $t_s = \tau_u t/\tau \ll t$, quasi-stationary on the fast timescale of $u(x, t)$. From numerical simulations, we know that the synaptic input variable remains finite, so any terms in the approximation $u \approx U + \Phi$ should also be bounded, including $\Phi(x, t)$. Therefore, we require a bounded solution to Eq. (16) by requiring the right hand side is orthogonal to the nullspace $V(x)$ of the adjoint linear operator $\mathcal{L}^* v = -v + F'(U)\int_{-180}^{180} w(x - y)v(y)\,\mathrm{d}y$. Orthogonality is enforced by requiring the inner product $\langle u, v\rangle = \int_{-180}^{180} u(x)v(x)\,\mathrm{d}x$ of the nullspace $V(x)$ with the inhomogeneous portion of Eq. (16) is zero. It can be shown $V(x) = F'(U(x))U'(x)$ spans the nullspace of $\mathcal{L}^*$[47]. This yields the following equation for the evolution of the bump position:

$$\mathrm{d}\theta(t) = K(\theta(t), t_s)\mathrm{d}t + \sigma\mathrm{d}\xi(t), \tag{17}$$

where the slowly evolving nonlinearity

$$K(\theta, t_s) = \frac{\int_{-180}^{180}\int_{-180}^{180} w(x - y)q(y + \theta, t_s)F(U(y))\mathrm{d}yF'(U(x))U'(x)\mathrm{d}x}{\int_{-180}^{180} U'(x)^2 F'(U(x))\mathrm{d}x} \tag{18}$$

is shaped by the form of $q(x, t_s)$ and the noise $\xi(t)$ is a standard Wiener process that comes from filtering the full spatiotemporal noise process $\mathrm{d}W(x, t)$, so the diffusion coefficient

$$D := \frac{\sigma^2}{2} = \frac{\int_{-180}^{180}\int_{-180}^{180} V(x)C(x - y)\mathrm{d}y\mathrm{d}x}{\left[\int_{-180}^{180} U'(x)V(x)\mathrm{d}x\right]^2}.$$

Equation (17) has the same form as Eq. (4). Thus, if the facilitation variable $q(x, t_s)$ evolves trial-to-trial such that $K(\theta, t_s)$ has similar shape to $-\frac{\mathrm{d}\mathcal{U}_{n+1}}{\mathrm{d}\theta}(\theta)$ at the beginning of the $(n+1)^{\mathrm{th}}$ trial ($t = t_{n+1}$), the dynamics of the network Eq. (15) can reflect a prior distribution based on the previous target(s). Given the approximation we derived in Eq. (5), we enforce proportionality $K(\theta, t_{n+1}) \propto -\frac{\mathrm{d}\mathcal{U}_{n+1}}{\mathrm{d}\theta}(\theta)$:

$$K(\theta,\, t_{n+1}) = \alpha \frac{\mathrm{d}f_{\theta_n}(\theta)}{\mathrm{d}\theta}, \tag{19}$$

where α is a scaling constant and t_{n+1} is the starting time of trial $n+1$ in the original time units $t = \tau t_s/\tau_u$. The form of the probability $f_{\theta'}(\theta)$ that can be represented is therefore restricted by the dynamics of the facilitation variable $q(x, t)$. We can perform a direct calculation to identify how $q(x, t)$ relates to the predictive distribution it represents in the following special case.

Explicit solutions for high-gain firing rate nonlinearities. To explicitly calculate solutions, we take the limit of high-gain, so that $F(u) \to H(u-\kappa)$ and $w(x) = \cos(\omega_1 x)$, note $\omega_1 = 180/\pi$. Note, we have compared our predictions here with the results of numerical simulations for sigmoidal firing rates $F(u) = 1/[1 + \mathrm{e}^{-\gamma(u-\kappa)}]$ with gain $\gamma = 20$, and the results are in good agreement. In this case, the bump solution $U(x-x_0) = (2\sin(a)/\omega_1) \cos(\omega_1(x-x_0))$ for $U(\pm a) = \kappa$ and null vector $V(x-x_0) = \delta(x-x_0-a) - \delta(x-x_0+a)$ (without loss of generality we take $x_0 \equiv 0$)[47]. Furthermore, we can determine the form of the evolution of $q(x, t)$ by studying the stationary solutions to Eq. (15) in the absence of noise ($W \equiv 0$). For a bump $U(x)$ centered at $x_0 = 0$, the associated stationary form for $Q(x)$ assuming $H(U(x)-\kappa) = 1$ for $x \in (-a, a)$ and zero otherwise is $Q(x) = \beta q_+/(1+\beta)$ for $x \in (-a, a)$ and zero otherwise. Thus, if the previous target was at θ_n, we expect $q(x, t)$ to have a shape resembling $Q(x-\theta_n)$ after trial n. Assuming the cue plus delay time during trial n was $T_C + T_D^n$ and the intertrial interval is T_I^n, slow dynamics will reshape the amplitude of $q(x, t)$ so $\mathcal{A}_n(T^n) = (1 - \mathrm{e}^{-(T_C+T_D^n)/\tau})\mathrm{e}^{-T_I^n/\tau}$ ($T^n = T_C + T_D^n + T_I^n$ is the total time block of each trial) and so $q(x,\, t) \approx \mathcal{A}_n(T^n) \cdot Q(x - \theta_n)$ at the beginning of trial $n+1$. A lengthy calculation of Eq. (18) combined with the relation Eq. (19) yields:

$$\alpha \frac{\mathrm{d}f_{\theta_n}(\theta)}{\mathrm{d}\theta} = \frac{\beta q_+ \mathcal{A}_n(T^n)}{2(1+\beta)\tan(a)} [\mathrm{sign}(\theta - \theta_n)\,(1 - \cos(\omega_1(\theta - \theta_n))) - \tan(a)\sin(\omega_1(\theta - \theta_n))],$$

for $|\theta - \theta_n| < 2a$, and $\frac{\mathrm{d}f_{\theta_n}(\theta)}{\mathrm{d}\theta} \equiv 0$ otherwise. Integrating, we find this implies

$$f_{\theta_n}(\theta) \propto |\theta - \theta_n| - \sin|\theta - \theta_n| + \tan(a)\cos(\theta - \theta_n),$$

for $|\theta - \theta_n| < 2a$, and $f_{\theta_n}(\theta)$ constant otherwise. Thus, the STF dynamics allows the network architecture to represent a predictive distribution that is peaked at the previous target location (Fig. 3). The amplitude of the θ-dependent portion of the predictive distribution during trial $n+1$ is then controlled by cue, delay, and intertrial times (T_C, T_D^{n+1}, T_I^{n+1}) and the facilitation parameters (β, q_+, τ).

To derive a coupled pair of equations (Fig. 4) describing the dynamics of the bump location $\theta(t)$ and the slow evolution of the nonlinearity $K(\theta, t)$, we focus on the limit $F(u) \to H(u-\kappa)$. We approximate $q(x, t)$ by summing the contributions from each of the $n+1$ trials. This yields

$$q(x,\, t) \approx \sum_{j=1}^{n} \mathcal{A}_j(t) Q(x - \theta_q(t_j + T_C + T_D^n)) + \mathcal{A}_{n+1}(t) Q(x - \theta_q(t)) \tag{20}$$

where the slowly evolving function $\mathcal{A}_n(t)$ defines the rising and falling kinetics of the facilitation variable originating in trial n:

$$\tau \dot{\mathcal{A}}_n(t) = \begin{cases} 1 - \mathcal{A}_n(t) & t_n < t < t_n + T_C + T_D^n, \\ -\mathcal{A}(t) & t > t_n + T_C + T_D^n, \end{cases}$$

increasing towards saturation ($\mathcal{A}_n \to 1$) during the cue and delay-period $[t_n,\, t_n + T_C + T_D^n)$ and decaying afterward ($\mathcal{A}_n \to 0$). The variable $\theta_q(t)$ describes the slow movement of the center-of-mass of the saturating portion of the facilitation variable $q(x, t)$ due to the drift of the neural activity $u(x, t)$ described by $\theta(t)$. However, since typically $\mathcal{A}_1(t) \ll \mathcal{A}_2(t) \ll \cdots \ll \mathcal{A}_n(t)$, we only keep the terms $\mathcal{A}_n(t)$ and $\mathcal{A}_{n+1}(t)$ in Eq. (20). It is possible that the memory of previous cues could persist for multiple trials, but the probability of this is exponentially small, since it would require many stochastic perturbations of the bump in the tail of the noise amplitude distribution. Therefore, we exclude a consideration of these cases from our linear perturbation approximation. Furthermore, since $\mathcal{A}_n(t)$ becomes much smaller than $\mathcal{A}_{n+1}(t)$ for most times $t > t_{n+1}$ in trial $n+1$, we approximate $\theta_q(t_n + T_C + T_D^n) \approx \theta_n$. This provides intuition as to why it is sufficient to only consider the previous target rather than the response in trial n as the variable influencing the bias in Papadimitriou *et al.*[13]. Therefore, we start with the following ansatz for the evolution of the facilitation variable during trial $n+1$:

$$q(x,\, t) = \mathcal{A}_n(t) Q(x - \theta_n) + \mathcal{A}_{n+1}(t) Q(x - \theta_q(t)). \tag{21}$$

A bump centered at $\theta(t)$, $U(x - \theta(t))$, attracts the STF variable to the same location $q \to Q(x - \theta(t))$, but the dynamics of q are much slower ($\tau \gg 1$). Thus, we model the evolution of $\theta_q(t)$ by linearizing the slow dynamics of Eq. (15b) about $(u, q) = (U(x - \theta(t)), Q(x - \theta(t))) + (0, \phi(x, t))$ (with $|\phi| \ll 1$) to find

$$\tau \dot{\phi}(x,\, t) = -\phi(x,\, t) - \beta F(U(x - \theta(t)))\phi(x,\, t). \tag{22}$$

The perturbation $\phi(x, t)$ describes the displacement of the variable q away from its equilibrium position. We now introduce the field $\Phi(x, t) = \int_{-180}^{180} w(x - y)\phi(y, t)F(U(y - \theta(t)))\,dy$[92], which reduces Eq. (22) to

$$\tau\dot{\Phi}(x, t) = -(1 + \beta)\Phi(x, t),$$

so separating variables $\Phi(x, t) = \overline{\Phi}(x)e^{\lambda t}$ we see that perturbations of the facilitation variable's center-of-mass $\theta_q(t)$ away from $\theta(t)$ should relax at rate $\lambda_\tau = -(1+\beta)/\tau$.

Therefore, the slow evolution of the potential gradient function $K(\theta, t_s)$ in Eq. (17) can be described by integrating Eq. (18) using the ansatz Eq. (21) for $q(x, t)$. Our low-dimensional system for the dynamics of the bump location $\theta(t)$ and leading order facilitation bump $\theta_q(t)$ during the delay-period of trial $n + 1$ ($t \in [t_{n+1} + T_C, t_{n+1} + T_C + T_D^{n+1})$) is given by the set of non-autonomous stochastic differential equations:

$$\begin{aligned} d\theta(t) &= -\mathcal{A}_n(t)\frac{d\overline{\mathcal{U}}(\theta - \theta_n)}{d\theta}dt - \mathcal{A}_{n+1}(t)\frac{d\overline{\mathcal{U}}(\theta - \theta_q(t))}{d\theta}dt + \sigma d\xi(t), \\ \tau\dot{\theta}_q(t) &= -d(\theta_q(t) - \theta(t)), \end{aligned}$$

where we have defined a parametrized time-invariant potential gradient $\frac{d\overline{\mathcal{U}}(\theta - \theta')}{d\theta}$ corresponding to the stationary profile of the facilitation variable centered at θ': $Q(x - \theta_n)$. For our specific choices of weight function and firing rate nonlinearity, we find the potential gradient is:

$$-\frac{d\overline{\mathcal{U}}(\theta - \theta')}{d\theta} = \frac{\beta q_+}{2(1 + \beta)\tan(a)}[\mathrm{sign}(\theta - \theta')\,(1 - \cos(\theta - \theta')) - \tan(a)\sin(\theta - \theta')],$$

and

$$d(\theta_q - \theta) = (1 + \beta)\begin{cases} \theta_q - \theta, & |\theta_q - \theta| \le \pi \\ \mathrm{sign}(\theta_q)\,(2\pi - |\theta_q - \theta|), & |\theta_q - \theta| > \pi \end{cases}$$

calculates the shorter difference on the periodic domain. As in our recurrent network, we use the parameters $\kappa = 0.1$; $q_+ = 2$; $\beta = 0.01$; and $\tau/\tau_u = 100$ to compare with network simulations in Fig. 5.

Numerical simulations of the neuronal network model. Numerical simulations of the recurrent network Eq. (15) were done in MATLAB using an Euler-Maruyama method with timestep $dt = 0.1$ ms and spatial step $dx = 0.18°$ with initial conditions generated randomly by starting $u(x, 0) \equiv q(x, 0) \equiv 0$ and allowing the system to evolve in response to the dynamic fluctuations for $t = 2$ s prior to applying the sequence of stimuli $I(x, t)$ described for each numerical experiment in Figs 5, 6 and 7. Numerical simulations of Eq. (6) were also performed using an Euler-Maruyama method with timestep $dt = 0.1$ ms. The effects of the target θ_n on each trial n were incorporated by holding $\theta(t) = \theta_n$ during the cue period $t \in [t_n, t_n + T_C)$. Otherwise, the dynamics were allowed to evolve as described.

Data analysis. MATLAB was used for statistical analysis of all numerical simulations. The bias effects in Fig. 5 were determined by identifying the centroid of the bump at the end of the delay-period. Means were computed across 10^5 simulations each, and standard deviations were determined by taking the square root of the var command applied to the vector of endpoints. Histograms in Fig. 6 were computed for 10^5 simulations using the hist and bar commands applied to the vector of endpoints for each correlation condition. Bump positions were computed in Fig. 7 by determining the centroid of the bump at each timepoint, and 10^5 simulations were then used to determine the standard deviation and variance plots (using var again).

References

1. Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* **61**, 331–349 (1989).
2. Romo, R., Brody, C. D., Hernández, A. & Lemus, L. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nat.* **399**, 470–473 (1999).
3. Pesaran, B., Pezaris, J. S., Sahani, M., Mitra, P. P. & Andersen, R. A. Temporal structure in neuronal activity during working memory in macaque parietal cortex. *Nat. Neurosci.* **5**, 805–811 (2002).
4. White, J. M., Sparks, D. L. & Stanford, T. R. Saccades to remembered target locations: an analysis of systematic and variable errors. *Vision Res.* **34**, 79–92 (1994).
5. Ploner, C. J., Gaymard, B., Rivaud, S., Agid, Y. & Pierrot-Deseilligny, C. Temporal limits of spatial working memory in humans. *Eur. J. Neurosci.* **10**, 794–797 (1998).
6. Wimmer, K., Nykamp, D. Q., Constantinidis, C. & Compte, A. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat. Neuroscience* **17**, 431–439 (2014).
7. Constantinidis, C., Franowicz, M. N. & Goldman-Rakic, P. S. The sensory nature of mnemonic representation in the primate prefrontal cortex. *Nat. Neurosci.* **4**, 311–316 (2001).
8. Goldman-Rakic, P. S. Cellular basis of working memory. *Neuron* **14**, 477–485 (1995).
9. Compte, A., Brunel, N., Goldman-Rakic, P. S. & Wang, X.-J. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb. Cortex* **10**, 910–923 (2000).
10. Renart, A., Song, P. & Wang, X.-J. Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. *Neuron* **38**, 473–485 (2003).
11. Constantinidis, C. & Klingberg, T. The neuroscience of working memory capacity and training. *Nature Reviews Neurosci* (2016).
12. Hartshorne, J. K. Visual working memory capacity and proactive interference. *PLoS one* **3**, e2716 (2008).

13. Papadimitriou, C., Ferdoash, A. & Snyder, L. H. Ghosts in the machine: memory interference from the previous trial. *J. Neurophysiol.* **113**, 567–577 (2015).
14. Makovski, T. & Jiang, Y. V. Proactive interference from items previously stored in visual working memory. *Mem. & cognition* **36**, 43–52 (2008).
15. Keppel, G. & Underwood, B. J. Proactive inhibition in short-term retention of single items. *J. verbal learning verbal behavior* **1**, 153–161 (1962).
16. Jonides, J. & Nee, D. E. Brain mechanisms of proactive interference in working memory. *Neurosci.* **139**, 181–193 (2006).
17. Kane, M. J. & Engle, R. W. Working-memory capacity, proactive interference, and divided attention: limits on long-term memory retrieval. *J. Exp. Psychol. Learn. Mem. Cogn.* **26**, 336 (2000).
18. Lustig, C., May, C. P. & Hasher, L. Working memory span and the role of proactive interference. *J. Exp. Psychol. Gen.* **130**, 199 (2001).
19. Baddeley, A. & Scott, D. Short term forgetting in the absence of proactive interference. *Q. J. Exp. Psychol.* **23**, 275–283 (1971).
20. Dunnett, S. B. & Martel, F. L. Proactive interference effects on short-term memory in rats: I. basic parameters and drug effects. *Behav. Neurosci.* **104**, 655 (1990).
21. Glaze, C. M., Kable, J. W. & Gold, J. I. Normative evidence accumulation in unpredictable environments. *eLife* **4**, e08825 (2015).
22. Veliz-Cuba, A., Kilpatrick, Z. P. & Josic, K. Stochastic models of evidence accumulation in changing environments. *SIAM Review* **58**, 264–289 (2016).
23. Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E. & Pouget, A. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron* **74**, 30–39 (2012).
24. Summerfield, C. & Tsetsos, K. Do humans make good decisions? *Trends. Cogn. Sci.* **19**, 27–34 (2015).
25. Markram, H. & Tsodyks, M. Redistribution of synaptic efficacy between neocortical pyramidal neurons. *Nature* **382**, 807 (1996).
26. Tsodyks, M. V. & Markram, H. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proceedings of the National Academy of Sciences* **94**, 719–723 (1997).
27. Mongillo, G., Barak, O. & Tsodyks, M. Synaptic theory of working memory. *Science* **319**, 1543–1546 (2008).
28. Itskov, V., Hansel, D. & Tsodyks, M. Short-term facilitation may stabilize parametric working memory trace. *Front. Comput. Neurosci.* **5**, 40 (2011).
29. Mi, Y., Katkov, M. & Tsodyks, M. Synaptic correlates of working memory capacity. *Neuron* **93**, 323–330 (2017).
30. Wilson, R. C., Nassar, M. R. & Gold, J. I. Bayesian online learning of the hazard rate in change-point problems. *Neural Comput.* **22**, 2452–2476 (2010).
31. Navarro, D. J. & Newell, B. Information versus reward in a changing world. *36th Annual Meeting of the Cognitive Science Society* (2014).
32. Adams, R. P. & MacKay, D. J. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742* (2007).
33. Wald, A. & Wolfowitz, J. Optimum character of the sequential probability ratio test. *The Annals Math. Statistics* **19**, 326–339 (1948).
34. Gold, J. I. & Shadlen, M. N. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron* **36**, 299–308 (2002).
35. Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700 (2006).
36. Beck, J. M. *et al.* Probabilistic population codes for bayesian decision making. *Neuron* **60**, 1142–1152 (2008).
37. Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).
38. Churchland, A. K., Kiani, R. & Shadlen, M. N. Decision-making with multiple alternatives. *Nat. Neurosci.* **11**, 693–702 (2008).
39. Kable, J. W. & Glimcher, P. W. The neurobiology of decision: consensus and controversy. *Neuron* **63**, 733–745 (2009).
40. Machens, C. K., Romo, R. & Brody, C. D. Flexible control of mutual inhibition: a neural model of two-interval discrimination. *Science* **307**, 1121–1124 (2005).
41. Brody, C. D., Romo, R. & Kepecs, A. Basic mechanisms for graded persistent activity: discrete attractors, continuous attractors, and dynamic representations. *Curr. Opin. Neurobiol.* **13**, 204–211 (2003).
42. Kilpatrick, Z. P., Ermentrout, B. & Doiron, B. Optimizing working memory with heterogeneity of recurrent cortical excitation. *J. Neurosci.* **33**, 18999–19011 (2013).
43. Amari, S. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol. Cybern.* **27**, 77–87 (1977).
44. Klingberg, T. Training and plasticity of working memory. *Trends. Cogn. Sci.* **14**, 317–324 (2010).
45. Wang, Y. *et al.* Heterogeneity in the pyramidal network of the medial prefrontal cortex. *Nat. Neurosci.* **9**, 534–542 (2006).
46. Burak, Y. & Fiete, I. R. Fundamental limits on persistent activity in networks of noisy neurons. *Proceedings of the National Academy of Sciences* **109**, 17645–17650 (2012).
47. Kilpatrick, Z. P. & Ermentrout, B. Wandering bumps in stochastic neural fields. *SIAM Journal on Applied Dynamical Systems* **12**, 61–94 (2013).
48. Scholl, B. J. & Pylyshyn, Z. W. Tracking multiple items through occlusion: Clues to visual objecthood. *Cognit. Psychol.* **38**, 259–290 (1999).
49. Bennett, S. J. & Barnes, G. R. Combined smooth and saccadic ocular pursuit during the transient occlusion of a moving visual object. *Exp. Brain Res.* **168**, 313–321 (2006).
50. Hempel, C. M., Hartman, K. H., Wang, X.-J., Turrigiano, G. G. & Nelson, S. B. Multiple forms of short-term plasticity at excitatory synapses in rat medial prefrontal cortex. *J. Neurophysiol.* **83**, 3031–3041 (2000).
51. Qi, Y., Breakspear, M. & Gong, P. Subdiffusive dynamics of bump attractors: mechanisms and functional roles. *Neural computation* (2015).
52. Barak, O. & Tsodyks, M. Persistent activity in neural networks with dynamic synapses. *PLoS Comput. Biol.* **3**, e35 (2007).
53. Stokes, M. G. *et al.* Dynamic coding for cognitive control in prefrontal cortex. *Neuron* **78**, 364–375 (2013).
54. Jackman, S. L. & Regehr, W. G. The mechanisms and functions of synaptic facilitation. *Neuron* **94**, 447–464 (2017).
55. Papadimitriou, C., White, R. L. & Snyder, L. H. Ghosts in the machine ii: Neural correlates of memory interference from the previous trial. *Cerebral Cortex* bhw106 (2016).
56. Tsodyks, M., Pawelzik, K. & Markram, H. Neural networks with dynamic synapses. *Neural Comput.* **10**, 821–835 (1998).
57. Rose, N. S. *et al.* Reactivation of latent working memories with transcranial magnetic stimulation. *Science* **354**, 1136–1139 (2016).
58. York, L. C. & Van Rossum, M. C. Recurrent networks with short term synaptic depression. *J. Comput. Neurosci.* **27**, 607–620 (2009).
59. Bhalla, U. S. Molecular computation in neurons: a modeling perspective. *Curr. Opin. Neurobiol.* **25**, 31–37 (2014).
60. Benna, M. K. & Fusi, S. Computational principles of synaptic memory consolidation. *Nature Neurosci* (2016).
61. Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R. & Haynes, J.-D. The distributed nature of working memory. *Trends in Cognitive Sciences* (2017).
62. Barak, O. & Tsodyks, M. Working models of working memory. *Curr. Opin. Neurobiol.* **25**, 20–24 (2014).
63. Rolls, E. T., Dempere-Marco, L. & Deco, G. Holding multiple items in short term memory: a neural mechanism. *PLoS One* **8**, e61078 (2013).
64. Hansel, D. & Mato, G. Short-term plasticity explains irregular persistent activity in working memory tasks. *J. Neurosci.* **33**, 133–149 (2013).
65. Zenke, F. & Gerstner, W. Hebbian plasticity requires compensatory processes on multiple timescales. *Phil. Trans. R. Soc. B* **372**, 20160259 (2017).
66. Markowitz, D. A., Curtis, C. E. & Pesaran, B. Multiple component networks support working memory in prefrontal cortex. *Proceedings of the National Academy of Sciences* **112**, 11084–11089 (2015).
67. Carroll, S., Josić, K. & Kilpatrick, Z. P. Encoding certainty in bump attractors. *J. Comput. Neurosci.* **37**, 29–48 (2014).

68. Lim, S. & Goldman, M. S. Balanced cortical microcircuitry for maintaining information in working memory. *Nat. Neurosci.* **16**, 1306–1314 (2013).
69. Boerlin, M., Machens, C. K. & Denève, S. Predictive coding of dynamical variables in balanced spiking networks. *PLoS Comput. Biol.* **9**, e1003258 (2013).
70. Shaham, N. & Burak, Y. Slow diffusive dynamics in a chaotic balanced neural network. *PLoS Comput. Biol.* **13**, e1005505 (2017).
71. Ma, W. J., Husain, M. & Bays, P. M. Changing concepts of working memory. *Nat. Neurosci.* **17**, 347–356 (2014).
72. Nassar, M. R., Helmers, J. C. & Frank, M. J. Chunking as a rational strategy for lossy data compression in visual working memory tasks. *bioRxiv* 098939 (2017).
73. Zhang, W. & Luck, S. J. Discrete fixed-resolution representations in visual working memory. *Nature* **453**, 233–235 (2008).
74. Luck, S. J. & Vogel, E. K. Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends. Cogn. Sci.* **17**, 391–400 (2013).
75. Bays, P. M. & Husain, M. Dynamic shifts of limited working memory resources in human vision. *Science* **321**, 851–854 (2008).
76. Wei, Z., Wang, X.-J. & Wang, D.-H. From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. *J. Neurosci.* **32**, 11228–11240 (2012).
77. Almeida, R., Barbosa, J. & Compte, A. Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. *J. Neurophysiol.* **114**, 1806–1818 (2015).
78. Bays, P. M. Spikes not slots: noise in neural populations limits working memory. *Trends. Cogn. Sci.* **19**, 431–438 (2015).
79. Kim, T. D., Kabir, M. & Gold, J. I. Coupled decision processes update and maintain saccadic priors in a dynamic environment. *J. Neurosci.* **37**, 3632–3645 (2017).
80. Brighton, H. & Gigerenzer, G. Bayesian brains and cognitive mechanisms: Harmony or dissonance. *The probabilistic mind*: *Prospects for Bayesian cognitive science*, ed. Chater, N. & Oaksford, M. 189–208 (2008).
81. Austerweil, J. L., Gershman, S. J., Tenenbaum, J. B. & Griffiths, T. L. Structure and flexibility in bayesian models of cognition. *Oxford handbook computational mathematical psychology* 187–208 (2015).
82. Gigerenzer, G. & Gaissmaier, W. Heuristic decision making. *Annu. Rev. Psychol.* **62**, 451–482 (2011).
83. Abraham, W. C. Metaplasticity: tuning synapses and networks for plasticity. *Nat. Rev. Neurosci.* **9**, 387–387 (2008).
84. Hulme, S. R., Jones, O. D., Raymond, C. R., Sah, P. & Abraham, W. C. Mechanisms of heterosynaptic metaplasticity. *Phil. Trans. R. Soc. B* **369**, 20130148 (2014).
85. Farashahi, S. *et al.* Metaplasticity as a neural substrate for adaptive learning and choice under uncertainty. *Neuron* **94**, 401–414 (2017).
86. Orbanz, P. & Teh, Y. W. Bayesian nonparametric models. In *Encyclopedia of Machine Learn.*, 81–89 (Springer, 2011).
87. Diaconis, P. & Ylvisaker, D. *et al.* Conjugate priors for exponential families. *The Annals statistics* **7**, 269–281 (1979).
88. Risken, H. *The Fokker-Planck equation* (Springer, 1996).
89. Bressloff, P. C. Spatiotemporal dynamics of continuum neural fields. *J. Phys. A: Math. Theor.* **45**, 033001 (2012).
90. Häusser, M. & Roth, A. Estimating the time course of the excitatory synaptic conductance in neocortical pyramidal cells using a novel voltage jump method. *J. Neurosci.* **17**, 7606–7625 (1997).
91. Wilson, H. R. & Cowan, J. D. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Biol. Cybern.* **13**, 55–80 (1973).
92. Kilpatrick, Z. P. & Bressloff, P. C. Stability of bumps in piecewise smooth neural fields with nonlinear adaptation. *Phys. D: Nonlinear Phenom.* **239**, 1048–1060 (2010).

Acknowledgements
We thank Krešimir Josić and Brent Doiron for helpful conversations and comments on the manuscript. This work was supported by grants from the National Science Foundation: NSF-DMS-1615737 and NSF-DMS-1517629.

Author Contributions
Z.P.K. Conception and design, Analysis and interpretation of data, Drafting and revising the article.

Additional Information
Competing Interests: The author declares no competing interests.