Synaptic mechanisms of interference in working memory

Zachary P Kilpatrick[1,2]

1 Department of Applied Mathematics, University of Colorado, Boulder CO, USA
2 Department of Physiology and Biophysics, University of Colorado School of Medicine, Aurora CO, USA
* zpkilpat@colorado.edu

Abstract

In serial cognitive tasks, information from preceding trials can bias performance in the current trial, a phenomenon referred to as interference. Recent experiments by Papadimitriou et al. (2015) demonstrated such biases in spatial working memory tasks, wherein subjects recalled the location of a target presented in continuous space. Analyzing response correlations in serial trials, they found the recalled location in the current trial is biased in the direction of the target presented on the previous trial. We build on their heuristic computational model to: (a) provide a Bayesian interpretation of the history-dependent bias; and (b) derive a mechanistic model demonstrating short-term facilitation accounts for the dynamics of the observed bias. Our computational model is a bump attractor network whose architecture is reshaped dynamically during each trial, linking changes to network connectivity with a predictive distribution based on observations of prior trials. Applying timescale separation methods, we can obtain a low-dimensional description of the trial-to-trial bias based on the history of target locations. The model still has response statistics whose mean is centered at the true target location in the limit of a large number of trials. Furthermore, we demonstrate task protocols for which the plastic model performs better than a model with static connectivity. Thus, our work presents a testable hypothesis for the persistence of interference in uncorrelated spatial working memory trials.

Introduction

Parametric working memory experiments serve as a testbed for behavioral biases and errors, and help identify the neural mechanisms that underlies them (Funahashi et al., 1989; Romo et al., 1999; Pesaran et al., 2002). For instance, in spatial working memory, subjects identify, store, and recall target locations in continuous space. Responses tend to exhibit error that is normally distributed (White et al., 1994; Ploner et al., 1998; Wimmer et al., 2014), and most of this error is accumulated during the *delay period*, while subjects retain the target location in memory (Funahashi et al., 1989; Constantinidis et al., 2001; Wimmer et al., 2014). Networks that model the behavioral error and neural activity recorded from these tasks typically utilize stimulus-tuned neurons with slow excitation and broad inhibition (Goldman-Rakic, 1995; Camperi and Wang, 1998; Compte et al., 2000). Persistent activity emerges as a tuned pattern of activity called a "bump" state, whose peak or center-of-mass is thought to encode the remembered target position (Renart et al., 2003; Wimmer et al., 2014).

Most computational models of spatial working memory account for ensemble statistics of behavior and neural recordings, ignoring serial correlations that may exist across trials (Constantinidis and Klingberg, 2016). Recently, Papadimitriou et al. (2015) demonstrated a consistent serial bias by analyzing the impact that a target within one trial had on a subject's response in the subsequent trial (Fig. 1A). On average, a subject's response on a trial was biased toward the target presented on the previous trial (Fig. 1B). This effect was reduced by increasing the time interval between trials (Fig. 1C), while the effect increased for longer delay times within the current trial (Fig. 1D). To account for these data, Papadimitriou et al.

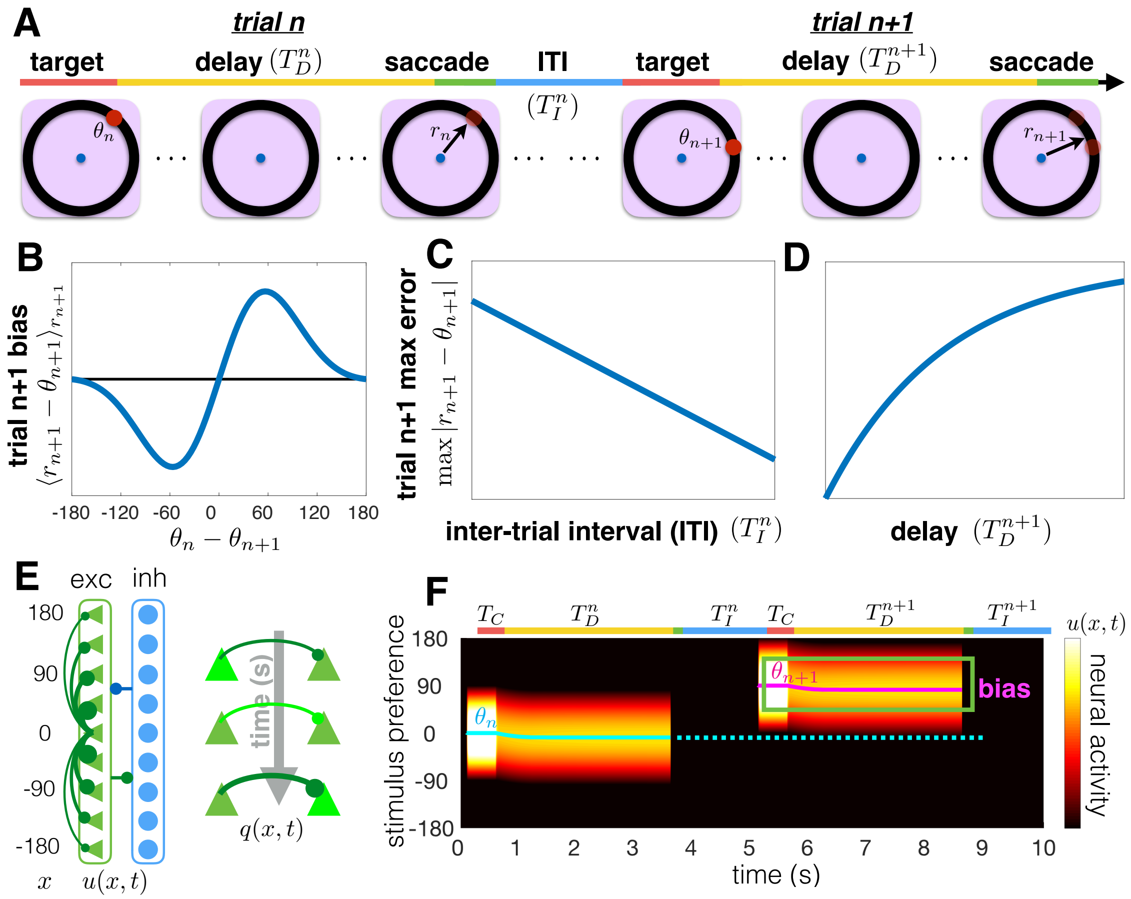

Fig 1. Interference in spatial working memory observed by Papadimitriou et al. (2015), and our corresponding recurrent network model with STF. (A) A spatial working memory task was administered in consecutive trials. The subject fixates on the central (blue) dot and a target (red dot) appears (θ_n and θ_{n+1}, 0-360°). The target then disappears, and the subject retains a memory of the target location during the delay period (T_D^n and T_D^{n+1}, 0-6000ms). Lastly, the subject makes a saccade (r_n and r_{n+1}) to the remembered target location. Papadimitriou et al. (2015) 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 errors in trial $n+1$ ($\langle r_{n+1} - \theta_{n+1} \rangle_{\theta_{n+1}}$) 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 error 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.

(2015) proposed a bump attractor model whose connectivity was impacted by two heuristic "memory stores," one which persists between trials and the other persists only within trials.

This previous study of interference leaves several open questions. First, what evidence accumulation strategy can account for the current trial response being biased by the previous trial's target? We will show that such biases emerge naturally in observers employing sequential Bayesian updating to predict the most likely next location of the target (Fig. 2). Bayesian inference has been used to account for both behavior and recorded neural activity in decision making experiments (Gold and Shadlen, 2002; Bogacz et al., 2006; Beck et al., 2008). Such models are obtained by iteratively applying Bayes' rule to a stream of noisy measurements to update an observer's belief of the most likely choice. In static environments, each measurement is given equal weight. In changing environments, older measurements are discounted at a rate that increases with the change rate of the environment (Glaze et al., 2015; Veliz-Cuba et al., 2016). Thus, if an observer believes the environment is changing rapidly, they may only use their most recent observation to determine the present state of the environment.

These principles can be applied to a sequence of spatial working memory trials, where the target in each trial is treated as a noisy observation. Depending on the observer's belief about the rate at which the environment is changing, the most recent trial may be weighted much more than the ensemble of previous trials. Our Bayesian model is based on the assumption that subjects erroneously integrate evidence as if only the most recent trial is relevant. Such suboptimal inference has been observed in experiments for which subjects have been trained extensively (Navarro and Newell, 2014; Summerfield and Tsetsos, 2015), and may be inevitable due to the complexity and time requirements of optimal strategies (Beck et al., 2012; Acerbi et al., 2014).

What neurophysiological processes can account for the slow relaxation of saccade bias within a trial and the transfer of the bias between trials? Note, the timescale of the buildup of the bias within a trial is roughly 1-5s (See Fig. 1D and Papadimitriou et al. (2015)). The decay timescale of slow-inactivating NMDA receptors is too short (roughly 100ms (Lester and Jahr, 1992)) to account for such dynamics. However, short-term synaptic plasticity can act on longer timescales (roughly 1s (Markram and Tsodyks, 1996; Tsodyks and Markram, 1997)). Thus, we propose short-term facilitation (STF) can slow the drift of a persistent bump of activity representing the stored target angle during the delay period (Fig. 1E). Previous models have also identified STF as a mechanism for lengthening the timescale of working memory (Mongillo et al., 2008; Itskov et al., 2011; Mi et al., 2017). Furthermore, STF can account for the latent bias present from the previous trial. Since the previous trial would have facilitated synapses originating from neurons tuned to the previous target, STF will attract the activity bump in the subsequent trial (Fig. 1F). Thus, as opposed to Papadimitriou et al. (2015), who employ two separate memory stores, we find that the slow timescale of the stored bias and incorporation of the bias into the subsequent response can be described by the same process.

Our neurocomputational model can account for the experimental observations of Papadimitriou et al. (2015). Due to the separation in timescales between the neural activity dynamics and the plasticity variable, we can derive a low-dimensional model that accounts for the bump position's interaction with the network's evolving synaptic weights. Subsequently, an analysis of the ensemble statistics of our model allows us to determine protocol conditions for which the mean estimated position of a target will still be at the true target position. Conversely, we propose target protocol sequences that will lead to a biased distribution 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 will tend to support bump attractors whose diffusion time course possesses two distinct phases, an experimental prediction we propose to validate our model.

Results

Our study presents two distinct frameworks for generating interference in a sequence of spatial 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 derive a probabilistic model that infers a distribution of possible target angles on the current trial based on observations of past trials (Fig. 2). In the limit of a rapidly-changing environment, the previous trial's target is the only information that shapes the prediction of the current trial's distribution. 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 profile of the STF variable represent the observer's evolving predictive distribution for the subsequent target (Figs. 1E,F and 3B). We show that these models can be directly related, and account for the bias inherited from the previous trial.

Suboptimal inference model for updating target predictions

Interference can arise as a suboptimal probabilistic prediction of the target location in subsequent trials. There are two key features of our model. First, an observer computes a function describing the likelihood of having observed θ_j on the j^{th} trial, assuming the target is θ_{n+1} on the $(n+1)^{\text{th}}$ trial $(j < n+1)$ and the distribution of targets $s_n(\theta)$ is the same between those trials $(s_{n+1}(\theta) \equiv s_j(\theta))$ (Wilson et al., 2010). Thus, the observer has an internal model of the conditional probability $f_\theta(\theta') := \mathrm{P}(\theta_j = \theta' | \theta_{n+1} = \theta, s_{n+1}(\theta) \equiv s_j(\theta))$. Second, observers assume the distribution from which presented targets are drawn changes stochastically at a fixed rate $\epsilon := \mathrm{P}(s_{n+1}(\theta) \not\equiv s_n(\theta))$. Static environments have a change rate $\epsilon = 0$ while a constantly changing environment has $\epsilon = 1$. Most spatial working memory protocols fix the distribution of target angles throughout the task $(\epsilon = 0)$ (Funahashi et al., 1989; Pesaran et al., 2002; Wimmer et al., 2014; Papadimitriou et al., 2015), so note that we suppose the observer employs an incorrect model to estimate this distribution $(\epsilon > 0)$. As recently shown in (Glaze et al., 2015), subjects in psychophysical tasks can have a strong bias toward assuming environments change on a timescale of several seconds, and this bias is not easily trained away (Beck et al., 2012; Navarro and Newell, 2014).

Our model successively updates the distribution of possible target angles in trial $n+1$: θ_{n+1}. This algorithm is based on models recently developed to compute a predictive distribution for a stochastically moving target, given a sequence of noisy observations (Adams and MacKay, 2007; Wilson et al., 2010). The predictive distribution is computed using sequential analysis (Busemeyer and Townsend, 1993; Wald and Wolfowitz, 1948; Veliz-Cuba et al., 2016): The observer sees the target $\theta_j \in [-180, 180)^\circ$ at the beginning of the j^{th} trial $(j < n+1)$. We thus define the likelihood function $f_{\theta_{n+1}}(\theta_j)$ (Fig. 2B) to be the probability of having observed the target θ_j in the j^{th} trial assuming: (a) the target θ_{n+1} is observed in the $(n+1)^{\text{th}}$ trial and (b) the underlying probability distribution from which targets are sampled does not change from trial j to $n+1$ $(s_{n+1}(\theta) \equiv s_j(\theta))$. Further details on the assumptions and derivation of our probabilistic inference model are given in Methods.

We assume the observer utilizes a predictive distribution $L_{n+1,\theta} = \mathrm{P}(\theta_{n+1}|\theta_{1:n})$, which takes the previous targets $\theta_{1:n}$ (Fig. 2A) as observations to predict the subsequent target θ_{n+1}. If the distribution $s_{n+1}(\theta)$ from which targets are drawn in trial $n+1$ changes stochastically with a rate $\epsilon \in (0,1)$, then recent observations will be weighted more in determining $L_{n+1,\theta}$ (Wilson et al., 2010; Glaze et al., 2015; Veliz-Cuba et al., 2016). Each observation θ_j contributes to the current estimate of $L_{n+1,\theta}$ via the likelihood function $f_\theta(\theta_j)$ (Fig. 2B). The weighting of each observation is determined by assuming the observer has a fixed belief about the value ϵ, which specifies the average number of trials they expect the distribution $s_n(\theta)$ to remain the same. Leveraging techniques in probabilistic inference (See Methods), we

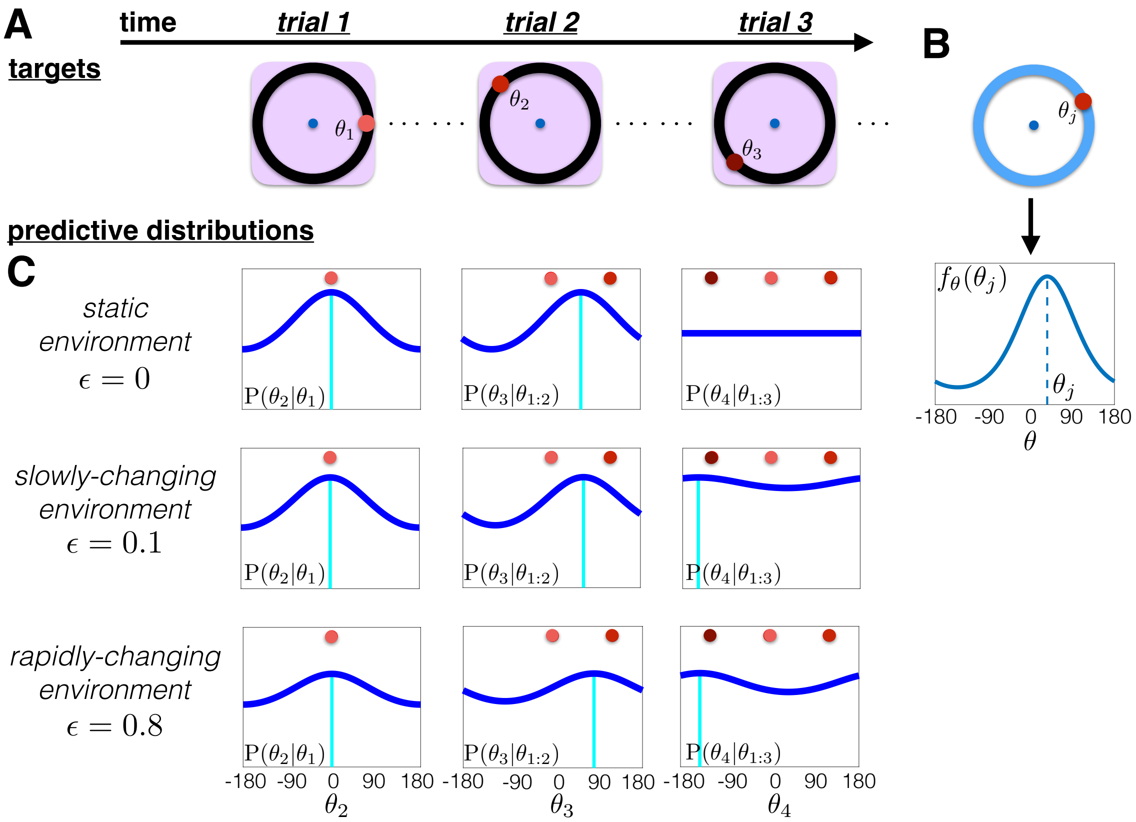

Fig 2. Updating the predictive distribution. The observer infers the predictive distribution for the subsequent target θ_{n+1} from prior observations $\theta_{1:n}$: $\mathrm{P}(\theta_{n+1}|\theta_{1:n})$. (A) A sequence of presented targets: $\theta_{1:3}$. (B) Self-conjugate likelihood function $f_\theta(\theta_j) \equiv f_{\theta_j}(\theta)$, peaked and centered at θ_j, showing the probability of observing $\theta_{n+1} = \theta$ 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 $\mathrm{P}(\theta_{n+1}|\theta_{1:n})$ for static ($\epsilon = 0$); slowly-changing ($\epsilon = 0.1$); and rapidly-changing ($\epsilon = 0.8$) environments. In static environments, all observations $\theta_{1:3}$ are weighted equally whereas in the rapidly-changing environment, the most recent observation dominates.

find that the predicted probability of seeing a target at location θ during trial $n+1$ is:

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

To understand Eq. (1), it is instructive to examine limits of the parameter ϵ that admit approximations or simple exact updates.

Static environments ($\epsilon \to 0$). In the limit $\epsilon \to 0$, the observer assumes the environment is static, so the predictive distribution is comprised of equal weightings of each observation (Fig. 2C and (Wald and Wolfowitz, 1948; Gold and Shadlen, 2002; Bogacz et al., 2006)):

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

As has been shown previously, Eq. (2) can be written iteratively (Beck et al., 2008):

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

suggesting such a computation could be implemented and represented by neural circuits. Most oculomotor delayed-response tasks use a distribution of targets $s(\theta)$ that is constant across trials (Funahashi et al., 1989; Pesaran et al., 2002; Wimmer et al., 2014; Papadimitriou et al., 2015). Therefore, Eq. (2) is the optimal strategy for obtaining an estimate of $s(\theta)$, assuming the observer has a correct representation of the likelihood $f_{\theta}(\theta_j)$.

Constantly changing ($\epsilon \to 1$). Sequential computations are trivial in the limit of a constantly-changing environment $\epsilon \to 1$, since the observer assumes the environment is reset after each trial. Prior observations provide no information about the present distribution, so the likelihood is always only comprised of the uniform prior $\bar{\mathrm{P}}_0$:

$$L_{n+1,\theta} = \bar{\mathrm{P}}_0. \tag{3}$$

Between these limits ($0 < \epsilon < 1$), the observer believes the environment probabilistically changes after each trial. Recently observed targets will be weighted more strongly than older targets. Veliz-Cuba et al. (2016) showed previously observed targets should be discounted at a rate that increases with ϵ.

Rapidly-changing environment ($\epsilon \approx 1$). Our work focuses on the limit where the environment changes rapidly, $0 < (1-\epsilon) \ll 1$ ($\epsilon \approx 1$), to account for biases that depend chiefly on the previous trial's target θ_n. Note that in the limit of slowly changing target distributions, $0 < \epsilon \ll 1$, the observer builds a predictive distribution that accounts for evidence in the multiple trial history $\theta_{1:n}$ (See Fig. 2C and Methods). On the other hand, when $\epsilon \approx 1$, the observer assumes the environment changes fast enough that each subsequent target is likely drawn from a new distribution ($s_{n+1}(\theta) \not\equiv s_n(\theta)$). Applying this assumption to Eq. (1), the formula for $L_{n+1,\theta}$ is dominated by terms of order $(1-\epsilon)$ and larger. Truncating to $\mathcal{O}(1-\epsilon)$ and normalizing the update equation (See Methods) then yields

$$\tilde{L}_{n+1,\theta} = \epsilon\bar{\mathrm{P}}_0 + (1-\epsilon) f_{\theta}(\theta_n). \tag{4}$$

Thus, the dominant contribution from $\theta_{1:n}$ to $L_{n+1,\theta}$ in the limit of rapidly-changing environments is the target θ_n observed during the previous trial n (Fig. 2C), similar to findings of Papadimitriou et al. (2015).

In summary, a suboptimal probabilistic inference model that assumes the distribution of targets is predictable over short timescales lead 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 commonly used to describe the degradation of target memory during the delay period of a spatial working memory task (Brody et al., 2003; Renart et al., 2003; Burak and Fiete, 2012; Kilpatrick et al., 2013).

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 $\theta \in [-180, 180)^{\circ}$. These dynamics can be implemented in recurrent neuronal networks with local excitation and effective inhibition that is broad and fast (Amari, 1977; Camperi and Wang, 1998; Compte et al., 2000). Before examining the effects of neural architecture, we discuss how to incorporate the predictive distribution update, Eq. (4), into the low-dimensional model. Our analysis draws a clear link between the update of the predictive distribution, and the spatial organization of attractors in a network. Importantly, working memory is degraded by dynamic input fluctuations that cause the recalled target angle to wander diffusively during the delay period (Compte et al., 2000; Burak and Fiete, 2012; Kilpatrick et al., 2013; Wimmer et al., 2014).

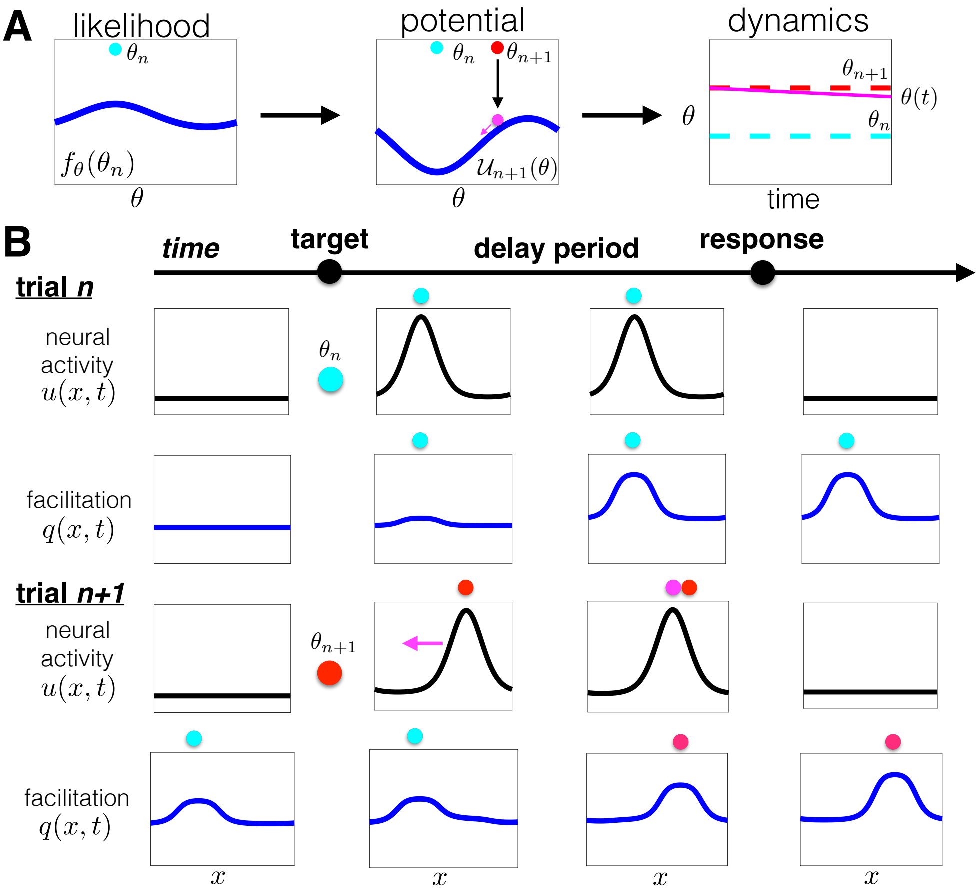

Fig 3. Encoding the predictive distribution in the potential function of an attractor network. (A) In a rapidly-changing environment, the predictive distribution is determined by the likelihood $f_\theta(\theta_n)$. In the low-dimensional system, with dynamics described by Eq. (5), this likelihood is represented by a potential function $\mathcal{U}_{n+1}(\theta)$ whose peak (valley) corresponds to the valley (peak) of $f_\theta(\theta_n)$, so the state $\theta(t)$ drifts towards the minimum of $\mathcal{U}_{n+1}(\theta)$ during the delay period. (B) A recurrent network with neurons distributed across $x \in [-180, 180)^\circ$ with STF (Fig. 1E) can implement these dynamics. The position of the trial n target is encoded by the peak location of the STF variable $q(x,t)$ during the early portion of trial $n+1$, attracting the neural activity $u(x,t)$ bump during the delay period.

Bump position $\theta(t)$ evolves according to a stochastic differential equation (Renart et al., 2003):

$$\mathrm{d}\theta(t) = -\frac{\mathrm{d}\mathcal{U}(\theta(t))}{\mathrm{d}\theta}\mathrm{d}t + \sigma_\theta \mathrm{d}\xi(t). \tag{5}$$

Here $\theta(t)$ is restricted to the periodic domain $\theta \in [-180, 180)^\circ$ and $\mathrm{d}\xi$ is a standard white noise process. The potential gradient $-\mathcal{U}'(\theta)$ in Eq. (5) models spatial heterogeneity in neural architecture that shapes attractor dynamics. During trial n, the potential $\mathcal{U}(\theta) \equiv \mathcal{U}_n(\theta)$. Classic models of bump attractors on a ring assume distance-dependent connectivity (Amari, 1977; Compte et al., 2000). The case $\mathcal{U}'_{n+1}(\theta) \not\equiv 0$ describes spatial heterogeneity in connectivity that may arise from a combination of training and synaptic plasticity (Renart et al., 2003; Klingberg, 2010), or random components of synaptic architecture (Wang et al., 2006). Our simplified model treats the working memory of the target angle $\theta(t)$ as a particle evolving on a potential landscape $\mathcal{U}_{n+1}(\theta)$ (Fig. 3A). We assume the potential landscape can be updated during each trial, so at the beginning of trial $n+1$ it has the form $\mathcal{U}_{n+1}(\theta)$. Thus, two qualitatively different scenarios are that the potential $\mathcal{U}_n(\theta)$ is: (a) always flat, so $\theta(t)$ evolves along a line attractor: $\mathcal{U}_n(\theta) \equiv 0$ for all $n = 1, 2, 3, ...$ (Burak and Fiete, 2012); or (b) heterogeneous, consisting of a combination of θ-dependent functions arising from some plasticity process acting during each trial: $\mathcal{U}_{n+1}(\theta) = \sum_{j=1}^{n} \mathcal{F}_j(\theta)$ (Renart et al., 2003; Kilpatrick et al., 2013). Prior information obtained from observing previous targets is incorporated by updating the potential $\mathcal{U}_n(\theta)$ after each trial.

To reflect the spatial working memory protocols, we assume that the observer receives $\widetilde{\theta}_n(0) = \theta_n$, the target, via sensory channels at the beginning of trial n. Thus, the observer initially estimates the angle perfectly. Eq. (5) evolves during the storage period of the working memory task, lasting for a delay time of T_D. After the delay period, $\widetilde{\theta}_n(T_D)$ will be the recalled angle of the memory layer. Typically, in non-human primate oculomotor tasks, the recalled angle is indicated via a saccade to a specific angle on a screen (Funahashi et al., 1989; Wimmer et al., 2014). Depending on the underlying potential $\mathcal{U}_n(\theta)$, there will be a strong bias to a subset of possible targets θ.

We derive a correspondence between the probabilistic inference model and attractor model by first ignoring kinetics of $\mathcal{U}_n(\theta)$ within a certain trial (See Methods). In the recurrent network model (Fig. 1E), we take these within-trial dynamics into account. Freezing $\mathcal{U}_n(\theta)$ 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. (5) to the desired likelihood function $L_{n+1,\theta}$ (See Methods). In this way, if information about the current trial's target θ_{n+1} is completely degraded, the probability of recalling the target angle θ is $L_{n+1,\theta}$. Focusing on the case suggested by experiments on interference, we aim to have the attractor structure of Eq. (5) represent the likelihood formula in Eq. (4). Our calculations yield the following relationship between the potential function prior to trial $n + 1$ and the likelihood function generated by the trial n target (Fig. 3A):

$$\mathcal{U}_{n+1}(\theta) \propto -f_\theta(\theta_n). \tag{6}$$

This suggests the potential update $\mathcal{U}_{n+1}(\theta)$ could be implemented by a decaying plasticity process that will potentiates portions of the network that represent the previous target. 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 comprised of neural activity $u(x, t)$ subject to STF $q(x, t)$ can incorporate predictive distribution updates we derived above. Predictions are stored in the dynamically changing synaptic weights of a recurrent neuronal network as the network is reshaped by STF. The recurrent network model spatially labels neurons and assigns each location in the network to a specific target preference, which determines 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 impacted dynamically by STF (Fig. 1E). See Methods for more details on the recurrent network model.

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 (n) has a cue period of time length T_C; a delay period of time length T_D^n; and a subsequent intertrial period of time length T_I^n before the next target is presented. The network receives a peaked current centered at the neurons preferring the presented target angle θ_n during the cue period of trial n; no external input during the delay period; and a strong inactivating current after the delay period (See Methods) (Camperi and Wang, 1998; Compte et al., 2000; Kilpatrick et al., 2013; Wimmer et al., 2014). The same protocol is applied during trial $n+1$ for a different target angle θ_{n+1}. 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 the attractor model, Eq (5), to account for STF. To reduce the recurrent network, we project the fast dynamics of bump solutions to an approximate equation for the bump's position (center-of-mass) $\theta(t)$ in trial n (Itskov et al., 2011; Kilpatrick and Ermentrout, 2013;

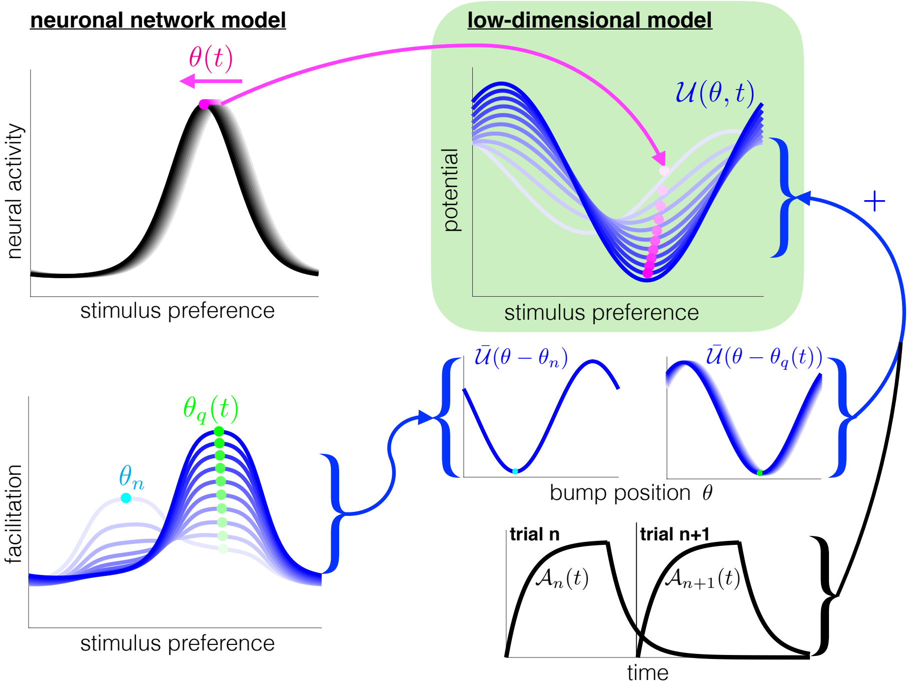

Fig 4. Low-dimensional system (green box) captures the motion of the bump ($\theta(t)$) and the evolving potential, $\mathcal{U}(\theta, t)$, shaped by the STF variable. The center-of-mass of the neural activity bump $\theta(t)$ is attracted by the most facilitated region of the network, $\text{argmin}_\theta\,[\mathcal{U}(\theta, t)]$. Both the previous trial's target θ_n and the current trial's bump location $\theta(t)$ attracts the center-of-mass of the STF variable's center-of-mass $\theta_q(t)$. 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 $\mathcal{U}(\theta - \theta_q(t))$. See Methods for a complete derivation.

Kilpatrick et al., 2013). This yields an evolving potential function $\mathcal{U}(\theta, t)$ of the network, determined by the STF variable $q(x, t)$ (Fig. 4). We use perturbation theory and timescale separation (See Methods) to derive a set of stochastic differential equations, which approximates the motion of the bump's position $\theta(t)$ along with the location of the STF variable, $\theta_q(t)$:

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

during trial $n+1$ ($t_n < t < t_{n+1}$). The slowly-evolving potential gradient $-\frac{\partial}{\partial\theta}\mathcal{U}(\theta, t)$ 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 bump position $\theta(t)$ moves towards the minimum of this dynamic potential during trial $n+1$, $\text{argmin}_\theta\,[\mathcal{U}(\theta, t)]$. The center-of-mass of the STF variable $\theta_q(t)$ in trial $n+1$ slowly moves toward the bump location $\theta(t)$.

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 $\bar{\mathcal{U}}(\theta - \theta_n)$, 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 at θ_{n+1}. The STF variable's center-of-mass $\theta_q(t)$ must slowly drift towards θ_n to allow $\theta(t)$ 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 (Papadimitriou et al., 2015).

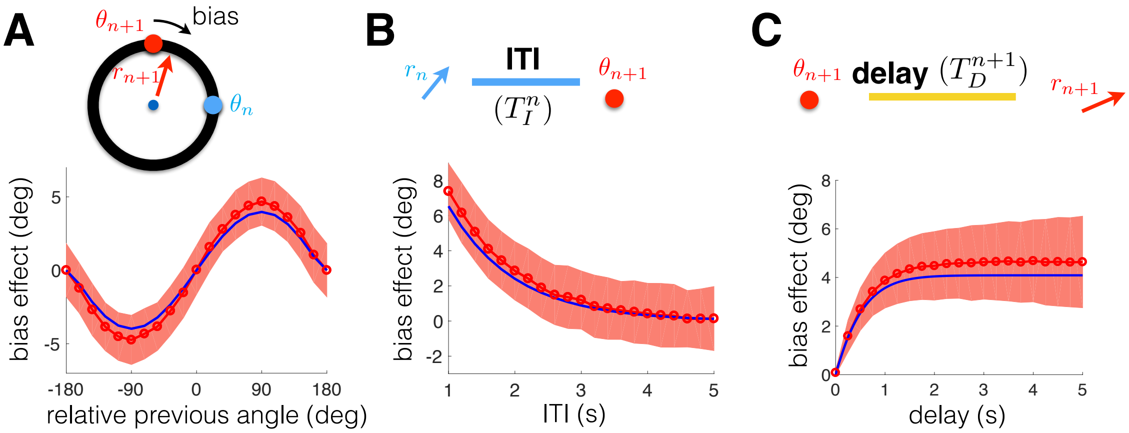

Fig 5. Intertrial bias is shaped by (A) the angle between targets θ_{n+1} and θ_n; (B) the interval between trials n and $n+1$ (ITI); and (C) the delay period during trial $n+1$. (A) Responses in trial $n+1$ are biased in the direction of the previous trial target (θ_n), with a peak bias occurring when $|\theta_{n+1} - \theta_n| \approx 90°$. 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.

Target- and time-dependent trends match experimental observations

We now demonstrate that the effects observed in the behavioral experiments of Papadimitriou et al. (2015) can be accounted for by our recurrent network model (Fig. 1E) and our low-dimensional description of bump motion dynamics (Fig. 4). A sequence of targets $(\theta_1, \theta_2, \theta_3, ...)$ was presented to the recurrent network at the beginning of each trial, and the remembered target location (response r_n) was determined by calculating the center-of-mass of the bump at the end of each delay period (See Methods). We computed the means and variances of the bias effect under each condition. Responses $(r_1, r_2, r_3, ...)$ were biased if the mean response $\langle r_n \rangle$ for a condition was different than the mean target angle $\langle \theta_n \rangle$.

Our results are summarized in Fig. 5, focusing on three conditions considered by Papadimitriou et al. (2015). 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. To expose this effect, we averaged across trials, since the recurrent network incorporates dynamic input fluctuations, consistent with typical bump attractor models of spatial working memory (Compte et al., 2000; Kilpatrick et al., 2013; Wimmer et al., 2014). 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. Consistent with Papadimitriou et al. (2015), the peak bias decreased with the ITI (Fig. 5B). The mechanism for this decrease is the decay in the STF of synapses utilized by the previous trial's persistent activity. 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.

Not only did our recurrent network model recapitulate the main findings of Papadimitriou et al. (2015), we also found our low-dimensional description of the bump and STF variable dynamics also had these properties (blue curves in Fig. 5). The mechanics underlying the bias can be described using a simple model of a particle evolving in a slowly changing potential (Fig. 4), due to the dynamics of STF. Having established a mechanism for the bias, we consider how different protocols shape the statistics of responses, not conditioned with sequential trial information.

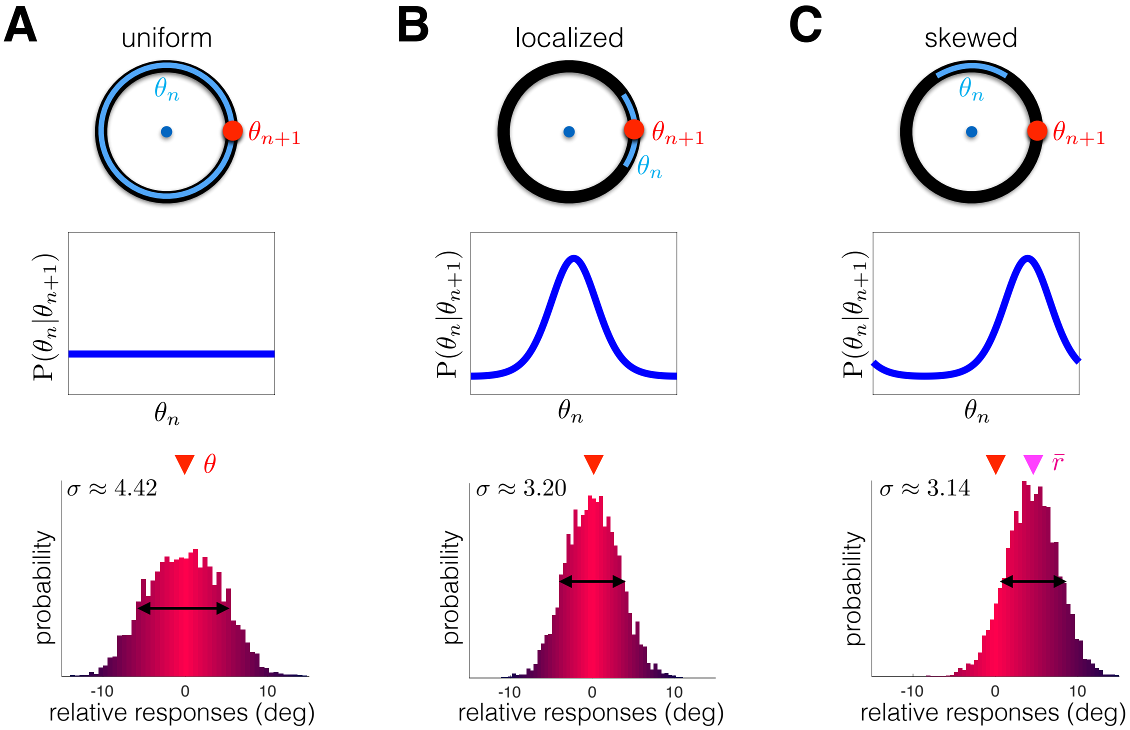

Fig 6. Response distribution is shaped by correlations between target angles in adjacent trials $P(\theta_n|\theta_{n+1})$. (A) Spatial working memory protocols typically use a sequence of target angles with no trial-to-trial correlations (uniform $P(\theta_n|\theta_{n+1})$) (Compte et al., 2000; Wimmer et al., 2014). Relative response angles ($r_n - \theta_n$) are normally distributed about the true target angle. (B) Prior target angle θ_n is correlated with the subsequent target angle θ_{n+1} according to a locally peaked distribution ($P(\theta_n|\theta_{n+1})$ shown for $\theta_{n+1} = 0°$). 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) Prior target θ_n is skewed counter-clockwise from current angle θ_{n+1}. The resulting response distribution is similarly skewed (note average response $\bar{r}$ is shifted). Numerical methods are described in Methods.

Task protocol shapes ensemble statistics

Spatial working memory tasks are often designed such that sequential target locations are uncorrelated (Compte et al., 2000; Wimmer et al., 2014). In such protocols, there is no advantage in using previous trial information to predict targets within the current trial. Nonetheless, such biases seem to persist in the extant intertrial response correlations discussed in Papadimitriou et al. (2015) and Fig. 5. On the other hand, such biases might be advantageous for tasks protocols with correlations between successive target angles, θ_n and θ_{n+1}. Consider object motion tracking tasks, where an object is transiently occluded (Scholl and Pylyshyn, 1999; Bennett and Barnes, 2006), so the object's location prior to occlusion predicts its subsequent location following occlusion. Memory of object location that persists beyond a single trial can therefore be useful for naturally-inspired tasks.

We demonstrate this idea by comparing the recurrent network's performance in working memory tasks whose targets are drawn from distributions with different intertrial correlation profiles (Fig. 6). As a control, we consider the case with no correlation between target θ_n and target θ_{n+1} (Fig. 6A). The distribution of responses is 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

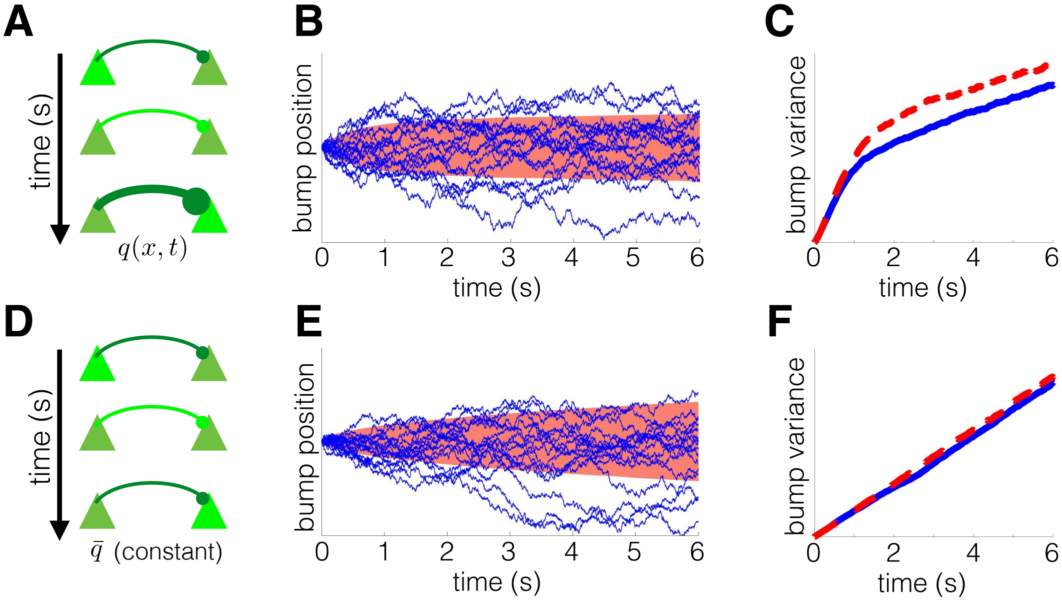

Fig 7. Recurrent networks with STF (panels A-C) exhibit two timescales of delay period dynamics in contrast to the single timescale of networks with static synapses (panels D-F). (A) STF strengthens synapses that were recently utilized. (B) In a facilitating network, bump trajectories (lines) stray less from their initial position due to the locally 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.

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, θ_{n+1}). An ensemble-wide measure of performance is given by the standard deviation of the response distribution ($\sigma \approx 4.42$). When target angles are correlated between trials, 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 ($\sigma \approx 3.20$). 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 correlation structure lead to a systematic bias in the relative response distribution, so its peak is shifted away from zero (Fig. 6C).

Our neuronal network model therefore predicts that, if an intertrial bias is present, it should be detectable by varying the intertrial correlation structure of target angles θ_n. Furthermore, when there are strong local correlations between adjacent trials ($\mathrm{P}(\theta_n|\theta_{n+1})$ is large for $|\theta_n - \theta_{n+1}|$ small), the network will tend to perform better than for protocols with uncorrelated adjacent trial angles.

Two timescales of memory degradation

Bump attractor models are useful for linking observations from neurophysiology to behavioral psychophysics in oculomotor delayed-response tasks (Compte et al., 2000; Wimmer et al., 2014). Wimmer et al. (2014) showed that the normal distribution of saccade endpoints along with observed changes in neural firing rates during the delay period can be accounted for by a diffusing bump attractor model (Wimmer et al., 2014). Their analysis ruled out two other candidate models: a bump attractor whose amplitude decays during the delay period and a bump attractor in a network with strongly discretized attractor structure. However, they did not consider mixed models, to see if bump attractor networks with mild spatial heterogeneities might provide a better fit to data (Kilpatrick et al., 2013).

We have shown that the recurrent network with STF (Fig. 1E) still leads to a normal distribution of predicted response angles (Fig. 6A). Furthermore, this model provides novel predictions for the internal dynamics of memory degradation, which we compare with the standard diffusing bump attractor model (Compte et al., 2000; Burak and Fiete, 2012; Kilpatrick and Ermentrout, 2013; Wimmer et al., 2014) (Fig. 7). In the network with STF (Fig. 7A), bump trajectories evolve in a history-dependent fashion (Fig. 7B). Initially, bumps diffuse freely, and are then attracted toward their starting location by the resulting facilitated synapses (See also Fig. 4). This results in 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 the 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 (Itskov et al., 2011), but our identification of these multiple timescale dynamics in memory degradation is novel. This feature of the bump dynamics is not present in networks with static synapses (Fig. 7D). Bumps evolve as a noise-driven particle over a flat potential landscape (Fig. 7E), described by Brownian motion, a memoryless stochastic process (Brody et al., 2003; Burak and Fiete, 2012). Variance in the bump position scales purely linearly with time (Fig. 7F), and the diffusion coefficient can be computed using a low-dimensional approximation (Kilpatrick and Ermentrout, 2013).

The qualitative differences we have identified between the bump attractor with and without dynamic synapses should be detectable in both behavioral and neurophysiological recordings (Wimmer et al., 2014). Moreover, the observed intertrial bias identified in recent analyses of behavioral data requires some mechanism for transferring information between trials that is distinct from neural activity (Papadimitriou et al., 2015), as dynamic synapses are in our model. In total, our model provides both an intuition for the behavioral reason and potential neural and synaptic mechanisms behind such interference.

Discussion

Typical neural circuit models of spatial working memory tend to only consider neural activity variables as the encoders of target locations. We presented a computational model for interference in spatial working memory that arises through both suboptimal Bayesian inference and can be accounted for by STF acting on a recurrent network model of delay-period activity. The timescale and prior target dependence of attractive biases in our computational model correspond to psychophysical observations of behavioral experiments in monkeys (Papadimitriou et al., 2015). STF evolves dynamically over seconds (Hempel et al., 2000; Wang et al., 2006), apparently matching the kinetics of recently observed interference. The link we have drawn between our two models suggests neural circuits can implement probabilistic inference using short term plasticity.

Experimental predictions

A more complete description of the neural mechanics of spatial working memory can be captured by modulating and analyzing the effects of correlations in sequential target presentations. Since responses in subsequent trials are shaped by the previous trial's target (Papadimitriou et al., 2015), computational models can be validated by determining how well their response distributions reflect trial-to-trial target correlations (Fig. 6). Furthermore, our model predicts multiple timescales emerge in the statistics of memory degradation during the delay period of 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 (White et al., 1994; Qi et al., 2015). The dynamics of our model are thus inconsistent with the purely linear diffusion of recall error common in bump attractor models with static synapses (Compte et al., 2000; Wimmer et al., 2014).

The idea that STF may play a role in working memory is not new (Barak and Tsodyks, 2007; Mongillo et al., 2008), and there is evidence that prefrontal cortex neurons exhibit dynamic patterns of activity during the delay period, suggestive of an underlying modulatory process (Stokes et al., 2013). However, it remains unclear how the presence of STF may shape the encoding of working memories. Our model suggests STF as a plausible mechanism for transferring 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 (Jackman et al., 2016; Jackman and Regehr, 2017). We predict that reducing the effects of the STF should 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 (Itskov et al., 2011).

Alternative physiological mechanisms for intertrial bias

Our study was motivated by a specific behavioral data set (Papadimitriou et al., 2015), which identified an attractive bias 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 (Rose et al., 2016), suggesting working memory is maintained by mechanisms other than target-tuned persistent neural activity (Mongillo et al., 2008; Stokes et al., 2013). The principle of using short term plasticity to store memories of working memory targets could be extended to account for longer timescales and more intricate statistical structures. For instance, 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 (York and Van Rossum, 2009). 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 serve to 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 the entire experiment, but this distribution is biased towards a discrete set of possible angles (Kilpatrick et al., 2013). For a recurrent network to represent this distribution, it is necessary to retain information about the series of target presentations over a long timescale. Numerous biophysical processes underlying plasticity have slow enough timescales to encode information from such lengthy sequences (Bhalla, 2014; Benna and Fusi, 2016). Furthermore, the distributed nature of working memory suggests that there may indeed be brain regions whose task-relevant neural activity partially persists from one trial to the next (Christophel et al., 2017). Such activity could then shape low-level sensory interpretations of targets in subsequent trials via mechanisms like feature-based attention that would tend to bias working memory.

Memory and training across timescales

Modeling and analysis of working memory storage often focuses on statistics that ignore between-trial correlations in both behavioral responses and neural circuit activity (Wimmer et al., 2014; Constantinidis and Klingberg, 2016). Our work, along with previous experimental findings (Papadimitriou et al., 2015), suggests models of working memory should account for interference arising from the previous trial history. More generally, the multiple timescales of memory storage and degradation appear to not be separable into distinct modules. Neural and synaptic mechanisms for memory storage overlap in an interconnected network (Hasson et al., 2015; Benna and Fusi, 2016). Prior information that is less relevant to the present environment can still corrupt the storage and recall of current information, similar to the impact of distractors on task-relevant information (Vogel et al., 2005). An important next step in developing theories for working memory lies in linking its storage and recall across timescales to incorporate the effects of long-term memory (Cowan, 2008). Such work would benefit from analysis of behavioral data and physiological recordings during the training phase of working memory experiments, when long term memory consolidation occurs. Learning during working memory can generate small populations of highly selective neurons as task performance improves (Meyer et al., 2011; Meyers et al., 2012), and extensive training can lead to significant changes in working memory capacity that persists for months or even years (Klingberg, 2010). Synaptic plasticity on multiple timescales likely plays a major role in the neural underpinnings of these changes (Bhalla, 2014; Benna and Fusi, 2016).

Synaptic plasticity can stabilize working memory

The idea of incorporating synaptic dynamics into computational theories of working memory is not new (Barak and Tsodyks, 2014). Previous computational models proposed that short-term plasticity can help stabilize or encode working memory (Mongillo et al., 2008; Itskov et al., 2011). 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 (Itskov et al., 2011; Rolls et al., 2013). This is related to our finding that STF reduces the diffusion of bumps in response to dynamic fluctuations (Fig. 7B), generating two distinct 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 has also been shown to account for experimentally observed increases in spike train irregularity during the working memory retention period in neural circuit models that support tuned persistent activity (Hansel and Mato, 2013). Alternatively, homeostatic synaptic scaling has been suggested to compensate for spatial heterogeneity, which would otherwise cause persistent states to drift (Renart et al., 2003). However, the short homeostatic timescales often suggested in models do not often match experimental observations (Zenke and Gerstner, 2017).

Models of working memory have also replaced persistent neural firing with stimulus-selective STF, so that neuronal spiking is only required for recall at the end of the delay period (Mongillo et al., 2008). 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 (Mi et al., 2017). 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 (Wimmer et al., 2014; Markowitz et al., 2015).

Robust working memory via excitatory/inhibitory balance

Fast synaptic feedback is another recently proposed mechanism for balancing cortical circuitry, so networks can encode continuous parameters. Computational models have demonstrated that a balance of fast inhibition and slow excitation can stabilize networks, so they accurately integrate inputs. 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 incorporating strong friction into the kinetics governing a particle motion on a sloped landscape (Lim and Goldman, 2013). 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 (Boerlin et al., 2013). Mutually inhibiting balanced networks are similarly capable of representing working memory of continuous parameters (Shaham and Burak, 2017), and extending our framework by incorporating STF into such paradigm would be a fruitful direction of future study.

Multi-item working memory

Working memory can store multiple items at once, and the neural mechanisms responsible for interference between simultaneously stored items are the focus of ongoing work (Ma et al., 2014; Nassar et al., 2017). 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) (Zhang and Luck, 2008; Luck and Vogel, 2013) or continuous (e.g., fluid) (Bays and Husain, 2008; Ma et al., 2014). Spatially-organized neural circuit models have been successful in recapitulating inter-item biases observed in multi-item working memory experiments, and provide a theory for how network interactions produce such errors (Wei et al., 2012; Almeida et al., 2015). 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 (Bays, 2015). The model provides a theory for attractive bias and forgetting of items in that nearby activity bumps tend to merge with one another. This is related to the mechanism of attractive bias in our model, but a significant difference is that ours is generated by STF whereas previous models only required localized excitation. It would be interesting identify the temporal dynamics of biases in multi-item working memory, to see if they require 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 dynamically tuned through learning about 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 can use this information to weight their observations toward a decision (Glaze et al., 2015; Kim et al., 2017). Our model suggests that the trial-history inference the subjects utilize in Papadimitriou et al. (2015) is significantly suboptimal, so it may be difficult to infer the timescale of relevant past-trial information. There is also evidence that humans tend to be biased towards employing suboptimal and heuristic methods for accumulating evidence when they are much simpler than the optimal strategy (Gigerenzer and Gaissmaier, 2011; Beck et al., 2012; Glaze et al., 2015). 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 (Abraham, 2008; Hulme et al., 2014). Recently, a model of reward-based metaplasticity was proposed to account for adaptive learning observed in a probabilistic reversal learning task (Farashahi et al., 2017). 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

In total, our results suggest that interference observed in spatial working memory tasks can be accounted for by a persistently active neural circuit with STF. This is in contrast to the model of Papadimitriou et al. (2015), which required the use of two memory stores. 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 understanding how such biases arise will improve our understanding of how working memory fits into the brain's information-processing hierarchy.

Methods

Assumptions of the inference model

Before trial n, the observer assumes the next target θ will be drawn from a specific distribution $s_n(\theta|\zeta)$, parameterized by an unknown parameter set $\zeta \in \Omega$ that is distributed according to $Z(\zeta)$. We assume that marginalizing over all such distributions yields the uniform density $\bar{\mathrm{P}}_0 = \int_\Omega s_n(\theta|\zeta)Z(\zeta)\mathrm{d}\zeta = 1/360$. One possibility is that the distribution $s_n(\theta|\eta)$ is constructed by drawing N-tuples $\mathbf{a}$ and $\boldsymbol{\psi}$ (so $\zeta = (\mathbf{a}, \boldsymbol{\psi})$) from a uniform distribution over the hypercubes $[0, a_{max}]^N$ and $[-180°, 180°)^N$ and using the entries to construct an exponential distribution of a sum of cosines:

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

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

$$s_n(\theta|\zeta) = \mathcal{N}_s \exp\left[a_1 \cos(\omega_1 \cdot (\theta - \psi_1))\right],$$

peaked at ψ_1. For the main instantiation and reduction of our model, knowing the specific family of distributions is unnecessary. We simply assume the average over all possible distributions $s_n(\theta|\zeta)$ is $\bar{\mathrm{P}}_0 = 1/360$.

The likelihood function $f_\theta(\theta') := \mathrm{P}(\theta_n = \theta'|\theta_{n+1} = \theta, s_{n+1}(\theta|\zeta) \equiv s_n(\theta|\zeta))$ is defined under static conditions ($s_{n+1}(\theta|\zeta) \equiv s_n(\theta|\zeta)$) to separate out the dynamic effects of sampling distribution $s_n(\theta)$ changes. Conjugate distributions $f_\theta(\theta')$ are not a necessary assumption of our model, but aid in conceptualizing the link between $\mathrm{P}(\theta_{n+1}|\theta_n)$ and $\mathrm{P}(\theta_n|\theta_{n+1})$. Several univariate priors on periodic domains are self-conjugate (Diaconis et al., 1979). To illustrate, we consider a family of distributions given by an exponential of cosines:

$$f_\theta(\theta') = \mathcal{N}_\theta \exp\left[\sum_{j=1}^{N} a_j \cos(\omega_j \cdot (\theta' - \theta))\right], \tag{8}$$

which is self-conjugate: $f_\theta(\theta') \equiv f_{\theta'}(\theta)$ (Diaconis et al., 1979). The example $f_\theta(\theta')$ we use for comparison with our recurrent network with STF is close to the case of Eq. (8) with $N = 1$.

Derivation of the probabilistic inference model

The observer's predictive distribution $L_{n+1,\theta} = \mathrm{P}(\theta_{n+1}|\theta_{1:n})$ is derived by marginalizing over possible run lengths $r_n = r$, corresponding to the number of trials the assumed underlying distribution $s_n(\theta|\zeta)$ has remained the same (Adams and MacKay, 2007; Wilson et al., 2010). Thus, $r_n = n$ indicates the environment has remained the same since the first trial, and $r_n = 0$ indicates the environment changes between trial n and $n + 1$. Summing over all possible run lengths, the marginal predictive distribution is

$$L_{n+1,\theta} = \sum_{r=0}^{n} \mathrm{P}(\theta_{n+1}|r_n = r, \theta_{1:n}^r)\mathrm{P}(r_n = r|\theta_{1:n}), \tag{9}$$

where $\mathrm{P}(\theta_{n+1}|r_n = r, \theta_{1:n}^r)$ is the conditional predictive distribution assuming run length $r_n = r$ and $\mathrm{P}(r_n = r|\theta_{1:n})$ is the conditional probability of the run length $r_n = r$ given the series of target angles $\theta_{1:n}$. We further simplify Eq. (9) as follows: First, utilizing sequential analysis, we find that if the present run length is $r_n = r$, the conditional predictive distribution is given by the product of likelihood functions corresponding to the last r observations (Veliz-Cuba et al., 2016):

$$\mathrm{P}(\theta_{n+1}|r_n = r, \theta^r_{1:n}) = \frac{\bar{\mathrm{P}}_0}{\mathrm{P}(\theta_{n-r+1:n})} \prod_{j=n-r+1}^{n} f_\theta(\theta_j). \tag{10}$$

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 primarily uses their knowledge of the change rate of the environment ϵ to determine the probability of a given run length $r_n = r$, and the conditional probability can be computed

$$\mathrm{P}(r_n = r|\theta_{1:n}) = \mathrm{P}(r_n = r) = \begin{cases} \epsilon(1-\epsilon)^r, & r < n, \\ (1-\epsilon)^n, & r = n. \end{cases} \tag{11}$$

Plugging Eqs. (10–11) into the update Eq. (9), we find the likelihood of the next target being at angle $\theta_{n+1} = \theta$, given that the previous n targets were $\theta_{1:n}$, is:

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

Limit of slowly-changing environment (small ϵ)

Here, we examine the case $0 < \epsilon \ll 1$, where the environment changes very slowly. Assuming independence of the target angles selected on each trial $\theta_{1:n}$ (Bogacz et al., 2006), $\mathrm{P}(\theta_{n-r:n}) = \mathrm{P}(\theta_{n-r:n-1})\mathrm{P}(\theta_n)$, we can split the probabilities over the target sequences $\theta_{n-r:n}$ into products: $\mathrm{P}(\theta_{n-r:n}) = \prod_{j=n-r}^{n} \mathrm{P}(\theta_j) = \bar{\mathrm{P}}_0^{r+1}$. The last equality holds since the family of possible distributions $s_n(\theta|\zeta)$ averages to a constant $\bar{\mathrm{P}}_0$, the uniform density. Applying this assumption to Eq. (1) and truncating to $\mathcal{O}(\epsilon)$, we have

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

noting 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 ($\epsilon \approx 1$)

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

$$L_{3,\theta} = \epsilon\bar{\mathrm{P}}_0 + \frac{1-\epsilon}{\mathrm{P}(\theta_2)} f_\theta(\theta_2) \left[\epsilon\bar{\mathrm{P}}_0 + \frac{1-\epsilon}{\mathrm{P}(\theta_1)} f_\theta(\theta_1)\bar{\mathrm{P}}_0 \right] \approx \epsilon\bar{\mathrm{P}}_0 \cdot \left[1 + \frac{(1-\epsilon)}{\mathrm{P}(\theta_2)} f_\theta(\theta_2) \right],$$

dropping the term of $\mathcal{O}((1-\epsilon)^2)$. Extending to arbitrary n, this reduces Eq. (1) to

$$L_{n+1,\theta} \approx \epsilon\bar{\mathrm{P}}_0 \left[1 + \frac{1-\epsilon}{\mathrm{P}(\theta_n)} f_\theta(\theta_n) \right]. \tag{12}$$

Furthermore, if we apply a uniform assumption to the unconditional probability of each observed target, $\mathrm{P}(\theta_n) = \bar{\mathrm{P}}_0$, we ensure the expression in Eq. (12) is normalized by writing

$$\tilde{L}_{n+1,\theta} = \frac{\bar{\mathrm{P}}_0 + (1-\epsilon) f_\theta(\theta_n)}{2-\epsilon},$$

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

$$\tilde{L}_{n+1,\theta} = \epsilon \bar{\mathrm{P}}_0 + (1-\epsilon) f_\theta(\theta_n),$$

which is the key update equation we focus upon in our Results (Figs. 2 and 3A). Higher order approximations are obtained by keeping more terms from Eq. (1). For instance, a second order approximation yields

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

successively downweighting likelihood products from previous observations (θ_{n-1}).

Relating likelihood to potential function of attractor model

We can understanding how a likelihood might be represented by an attractor model by determining the formula of the stationary distribution of Eq. (5), given an arbitrary potential function $\mathcal{U}_n(\theta)$. Eq. (5) can be reformulated as an equivalent Fokker-Planck equation for the represented angle θ_n during trial n assuming the present potential function is $\mathcal{U}_n(\theta)$ (Risken, 1996),

$$\frac{\partial p_n(\theta,t)}{\partial t} = \frac{\partial}{\partial \theta} \left[\frac{\mathrm{d}\mathcal{U}_n(\theta)}{\mathrm{d}\theta} p_n(\theta,t) \right] + \frac{\sigma_\theta^2}{2} \frac{\partial^2 p_n(\theta,t)}{\partial \theta^2}, \tag{13}$$

where $p_n(\theta,t)$ is the probability density corresponding to the target angle estimate $\widetilde{\theta}_n = \theta$ at time t.

We now derive the specific form of $\mathcal{U}_{n+1}(\theta)$ that would lead to a stationary density corresponding the predictive distribution $L_{n+1,\theta}$ in the limit $t \to \infty$ in Eq. (13). The stationary density $\bar{p}_{n+1}(\theta)$ is analogous to a likelihood function represented by Eq. (5) because it is the probability distribution represented by the system when no information about the current trial's target remains. Thus, we build a rule to update $\mathcal{U}_{n+1}(\theta)$ to mirror the update of $L_{n+1,\theta}$. To obtain this result, we seek to match the stationary density for Eq. (13) to the updated likelihood:

$$\lim_{t \to \infty} p_{n+1}(\theta,t) = \bar{p}_{n+1}(\theta) = L_{n+1,\theta}. \tag{14}$$

Solving Eq. (13) for its stationary solution, we find that during trial $n+1$:

$$\bar{p}_{n+1}(\theta) = \chi_{n+1} \exp\left[-\frac{2\mathcal{U}_{n+1}(\theta)}{\sigma_\theta^2} \right], \tag{15}$$

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. (15) into Eq. (14) and solving for $\mathcal{U}_{n+1}(\theta)$, we obtain

$$\mathcal{U}_{n+1}(\theta) = \frac{\sigma_\theta^2}{2} \ln \frac{\chi_p}{L_{n+1,\theta}}. \tag{16}$$

In the case of a rapidly changing environment $0 < (1-\epsilon) \ll 1$, we employ the approximation given by Eq. (4) so that

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

where we have linearized in $(1-\epsilon)$. However, for Eq. (5), only the derivative of $\mathcal{U}_{n+1}(\theta)$ will impact the dynamics, so we drop the additive constants and examine proportionality

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

so in the limit of weak interactions between trials, the potential $\mathcal{U}_{n+1}(\theta)$ should be shaped like the negative of the likelihood $f_\theta(\theta_n)$ 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 $t > 0$

$$\tau_u \mathrm{d}u(x,t) = \left[-u(x,t) + \int_{-180}^{180} w(x-y)(1+q(y,t))F(u(y,t))\mathrm{d}y\right]\mathrm{d}t + \mathrm{d}J(x,t), \quad (17a)$$

$$\tau \dot{q}(x,t) = -q(x,t) + \beta F(u(x,t))(q_+ - q(x,t)), \quad (17b)$$

where $u(x,t)$ describes the evolution of the normalized synaptic input at location x. The model Eq. (17) can be derived as the large system size limit of a population of synaptically coupled spiking neurons (Bressloff, 2012), and similar dynamics have been validated in spiking networks with lateral inhibitory connectivity (Compte et al., 2000; Wimmer et al., 2014). 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 (Häusser and Roth, 1997). 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.

Each location x in the network receives recurrent coupling described by the weight function $w(x-y)$. We expect this 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 (Goldman-Rakic, 1995). We do not separately model excitatory and inhibitory populations, but Eq. (17) can be derived from a model with distinct excitatory and inhibitory populations in the limit of fast inhibitory synapses (Amari, 1977; Carroll et al., 2014). 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 $F(u) = 1/\left[1 + \mathrm{e}^{-\gamma(u-\kappa)}\right]$ (Wilson and Cowan, 1973), 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 (Amari, 1977; Bressloff, 2012).

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 increase in synaptic utilization and the rate at which facilitation occurs (Tsodyks and Markram, 1997; Tsodyks et al., 1998). For our numerical simulations, we consider the parameters values $q_+ = 2$ and $\beta = 0.01$, consistent with previous models employing facilitation in working memory circuits (Itskov et al., 2011; Mongillo et al., 2008; Mi et al., 2017) and experimental findings for facilitation responses in prefrontal cortex (Hempel et al., 2000; Wang et al., 2006). The timescale of plasticity is slow, $\tau = 1000\text{ms} \gg 10\text{ms}$, consistent with experimental measurements (Tsodyks and Markram, 1997). 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 (Mongillo et al., 2008; Mi et al., 2017).

External inputs and dynamic input fluctuations are described by $\mathrm{d}J(x,t) = I(x,t)\mathrm{d}t + \mathrm{d}W(x,t)$, a spatially-extended noisy process. The effects of the target and the response are described by the deterministic spatiotemporal process $I(x,t)$, which we discuss more in detail below. The noise process $W(x,t)$ is white in time and has an increment with mean $\langle \mathrm{d}W(x,t)\rangle \equiv 0$ and spatial correlation function $\langle \mathrm{d}W(x,t)\mathrm{d}W(y,s)\rangle = C(x-y)\delta(t-s)\mathrm{d}t\mathrm{d}s$. 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 1-10s (Funahashi et al., 1989; White et al., 1994; Wimmer et al., 2014).

Implementing sequential delayed-response task protocol

A series of oculomotor delayed-response tasks is executed by the network Eq. (17) 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

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

for all $n = 1, 2, 3, ...$, 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 Papadimitriou et al. (2015). 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 $T_C = 500$ms; $T_A = 500$ms; $I_0 = 1$; $I_1 = 1$; and $I_R = 2$. Target locations θ_n are drawn from a uniform probability mass function (pmf) for the discrete set of angles $\theta_n \in \{-180°, -162°, ..., 162°$ to generate statistics in Figs. 5A, which adequately resolves the bias effect curves for comparison with the results in Papadimitriou et al. (2015). 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 $T_I^n \in \{1000, 1200, ..., 5000\}$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 $T_I^n \in \{0, 200, ..., 5000\}$ms and following a similar procedure to Fig. 5B. Draws from a uniform density function $\mathrm{P}(\theta_n)$, 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 a von Mises distribution $\mathrm{P}(\theta_n|\theta_{n+1}) = \mathcal{N}_v \mathrm{e}^{25\cos(\theta_n - \theta_{n+1} - \mu)}$ with $\mu = 0$ for local correlations (Fig. 6B) and $\mu = 90$ for skewed correlations (Fig. 6C).

The recurrent network, Eq. (17), is assumed to encode the initial target θ_n during trial n via the center-of-mass $\widetilde{\theta}_n(t)$ 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 $\widetilde{\theta}_n(t) = \mathrm{argmax}_x u(x,t)$ (when $t \in [t_n, t_n + T_C + T_D^n)$) is an accurate and efficient approximation (Kilpatrick et al., 2013; Wimmer et al., 2014). Bias (relative saccade endpoint) on each trial n is then determined by computing the difference $\widetilde{\theta}_n(t) - \theta_n$ (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. (17) has stationary bump solutions that are well studied and defined by the implicit equation (Amari, 1977; Camperi and Wang, 1998; Bressloff, 2012; Kilpatrick and Ermentrout, 2013):

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

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. (17) and truncating to linear order yields

$$\mathrm{d}\Phi(x,t) - \mathcal{L}\Phi(x,t)\mathrm{d}t = U'(x)\mathrm{d}\theta + \int_{-180}^{180} w(x-y) q(y+\theta, t_s) F(U(y)) \mathrm{d}y\mathrm{d}t + \mathrm{d}W, \tag{18}$$

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 $u(x,t)$. We ensure a bounded solution by requiring the right hand side of Eq. (18) 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. (18) is zero. It can be shown $V(x) = F'(U(x))U'(x)$ spans the nullspace of $\mathcal{L}^*$ (Kilpatrick and Ermentrout, 2013). This yields the following equation for the evolution of the bump position:

$$\mathrm{d}\theta(t) = K(\theta(t), t_s)\mathrm{d}t + \mathrm{d}\mathcal{W}(t), \tag{19}$$

where the slowly evolving nonlinearity

$$K(\theta, t_s) = \frac{\int_{-180}^{180} \int_{-180}^{180} w(x-y) q(y+\theta, t_s) F(U(y)) \mathrm{d}y F'(U(x)) U'(x) \mathrm{d}x}{\int_{-180}^{180} U'(x)^2 F'(U(x)) \mathrm{d}x} \tag{20}$$

is shaped by the form of $q(x,t_s)$ and the noise

$$\mathrm{d}\mathcal{W}(t) = \frac{\int_{-180}^{180} V(x) \mathrm{d}W(x,t) \mathrm{d}x}{\int_{-180}^{180} U'(x) V(x) \mathrm{d}x}$$

is a standard Wiener process that comes from filtering the full spatiotemporal noise process $\mathrm{d}W(x,t)$.

Eq. (19) has the same form as Eq. (5), up to the scaling of the noise $\mathrm{d}\mathcal{W}$. Thus, if the facilitation variable $q(x,t_s)$ evolves trial-to-trial such that $K(\theta, t_s)$ has similar shape to $-\frac{\mathrm{d}\mathcal{U}_n}{\mathrm{d}\theta}(\theta)$ at the beginning of the n^{th} trial ($t = t_n$), the dynamics of the network Eq. (17) can reflect a prior distribution based on the previous target(s). Given the approximation we derived in Eq. (6), we enforce proportionality $K(\theta, t_{n+1}) \propto -\frac{\mathrm{d}\mathcal{U}_{n+1}}{\mathrm{d}\theta}(\theta)$:

$$K(\theta, t_{n+1}) = \alpha \frac{\mathrm{d}f_\theta(\theta_n)}{\mathrm{d}\theta}, \tag{21}$$

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 likelihood f_θ 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 likelihood 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 $\omega_1 = 180/\pi$. 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 $x_0 \equiv 0$) (Kilpatrick and Ermentrout, 2013). Furthermore, we can determine the form of the evolution of $q(x,t)$ by studying the stationary solutions to Eq. (17) 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 $x \in (-a,a)$ (modulo the 360° period) 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. (20) combined with the relation Eq. (21) yields:

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

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

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

for $|\theta-\theta_n| < 2a$, and $f_\theta(\theta_n)$ constant otherwise. Thus, with the STF dynamics we have incorporated into our network, the network architecture will represent a prior that is peaked at the previous target location and decays for $|\theta-\theta_n|$ increasing (Fig. 3). The amplitude of the θ-dependent portion of the likelihood 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 continue to focus on the limit in which $F(u) \equiv H(u-\kappa)$. Our approximation for $q(x,t)$ is constructed by summing the contributions from each of the $n+1$ trials up to trial $n+1$. This yields

$$q(x,t) \approx \sum_{j=1}^{n} \mathcal{A}_j(t) Q(x - \theta_q(t_j + T_C + T_D^n)) + \mathcal{A}_{n+1}(t) Q(x - \theta_q(t)) \tag{22}$$

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 $\mathcal{A}_1(t) \ll \mathcal{A}_2(t) \ll \cdots \ll \mathcal{A}_n(t)$, we only keep the terms $\mathcal{A}_n(t)$ and $\mathcal{A}_{n+1}(t)$ in Eq. (22). 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 previously presented target rather than the response in trial n as the variable influencing the bias in Papadimitriou et al. (2015). Therefore, we start with the following ansatz for the evolution of the facilitation variable during trial $n+1$:

$$q(x,t) = \mathcal{A}_n(t)Q(x-\theta_n) + \mathcal{A}_{n+1}(t)Q(x-\theta_q(t)). \tag{23}$$

A bump centered at $\theta(t)$, $U(x-\theta(t))$, will attract a facilitation 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. (17b) about $(u,q) = (U(x-\theta(t)), Q(x-\theta(t))) + (0, \phi(x,t))$ (with $|\phi| \ll 1$) to find

$$\tau\dot{\phi}(x,t) = -\phi(x,t) - \beta F(U(x-\theta(t)))\phi(x,t). \tag{24}$$

The perturbation $\phi(x,t)$ describes the displacement of the variable q away from its equilibrium position. Following (Kilpatrick and Bressloff, 2010), we introduce the field $\Phi(x,t) = \int_{-180}^{180} w(x-y)\phi(y,t)F(U(y-\theta(t)))\mathrm{d}y$, which reduces Eq. (24) to

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

so separating variables $\Phi(x,t) = \bar{\Phi}(x)\mathrm{e}^{\lambda t}$ 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. (19) can be described by integrating Eq. (20) using the ansatz Eq. (23) 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:

$$\mathrm{d}\theta(t) = -\mathcal{A}_n(t)\frac{\mathrm{d}\bar{\mathcal{U}}(\theta-\theta_n)}{\mathrm{d}\theta}\mathrm{d}t - \mathcal{A}_{n+1}(t)\frac{\mathrm{d}\bar{\mathcal{U}}(\theta-\theta_q(t))}{\mathrm{d}\theta}\mathrm{d}t + \mathrm{d}\mathcal{W}(t), \tag{25a}$$

$$\tau\dot{\theta}_q(t) = -d(\theta_q(t) - \theta(t)), \tag{25b}$$

where we have defined a parametrized time-invariant potential gradient $\frac{\mathrm{d}\bar{\mathcal{U}}(\theta-\theta')}{\mathrm{d}\theta}$ 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:

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

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 the neuronal network model, we use the parameters $\kappa = 0.1$; $q_+ = 2$; $\beta = 0.01$; and $\tau/\tau_u = 100$ for comparisons with the full network simulations in Fig. 5.

Numerical simulations of the neuronal network model

Numerical simulations of the neuronal network model Eq. (17) were done in MATLAB using an Euler-Maruyama method with timestep $\mathrm{d}t = 0.1$ms and spatial step $\mathrm{d}x = 0.18°$ with initial conditions generated randomly by starting $u(x,0) \equiv q(x,0) \equiv 0$ and then allowing the system to evolve in response to the noise input 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. (25) were also performed using an Euler-Maruyama method with timestep d$t = 0.1$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).

Acknowledgments

This work was funded by National Science Foundation grants NSF-DMS-1615737 and NSF-DMS-1517629. We thank Krešimir Josić and Brent Doiron for comments on the manuscript.

References

Abraham W.C. (2008). Metaplasticity: tuning synapses and networks for plasticity. Nature Reviews Neuroscience *9*, 387–387.

Acerbi L., Vijayakumar S., and Wolpert D.M. (2014). On the origins of suboptimality in human probabilistic inference. PLoS Comput Biol *10*, e1003661.

Adams R.P., and MacKay D.J. (2007). Bayesian online changepoint detection. arXiv preprint arXiv:0710.3742 .

Almeida R., Barbosa J., and Compte A. (2015). Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. Journal of neurophysiology *114*, 1806–1818.

Amari S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. Biological cybernetics *27*, 77–87.

Barak O., and Tsodyks M. (2007). Persistent activity in neural networks with dynamic synapses. PLoS Comput Biol *3*, e35.

Barak O., and Tsodyks M. (2014). Working models of working memory. Current opinion in neurobiology *25*, 20–24.

Bays P.M. (2015). Spikes not slots: noise in neural populations limits working memory. Trends in cognitive sciences *19*, 431–438.

Bays P.M., and Husain M. (2008). Dynamic shifts of limited working memory resources in human vision. Science *321*, 851–854.

Beck J.M., Ma W.J., Kiani R., Hanks T., Churchland A.K., Roitman J., Shadlen M.N., Latham P.E., and Pouget A. (2008). Probabilistic population codes for bayesian decision making. Neuron *60*, 1142–1152.

Beck J.M., Ma W.J., Pitkow X., Latham P.E., and Pouget A. (2012). Not noisy, just wrong: the role of suboptimal inference in behavioral variability. Neuron *74*, 30–39.

Benna M.K., and Fusi S. (2016). Computational principles of synaptic memory consolidation. Nature neuroscience .

Bennett S.J., and Barnes G.R. (2006). Combined smooth and saccadic ocular pursuit during the transient occlusion of a moving visual object. Experimental Brain Research *168*, 313–321.

Bhalla U.S. (2014). Molecular computation in neurons: a modeling perspective. Current opinion in neurobiology *25*, 31–37.

Boerlin M., Machens C.K., and Denève S. (2013). Predictive coding of dynamical variables in balanced spiking networks. PLoS Comput Biol *9*, e1003258.

Bogacz R., Brown E., Moehlis J., Holmes P., and Cohen J.D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychological review *113*, 700.

Bressloff P.C. (2012). Spatiotemporal dynamics of continuum neural fields. Journal of Physics A: Mathematical and Theoretical *45*, 033001.

Brody C.D., Romo R., and Kepecs A. (2003). Basic mechanisms for graded persistent activity: discrete attractors, continuous attractors, and dynamic representations. Current opinion in neurobiology *13*, 204–211.

Burak Y., and Fiete I.R. (2012). Fundamental limits on persistent activity in networks of noisy neurons. Proceedings of the National Academy of Sciences *109*, 17645–17650.

Busemeyer J.R., and Townsend J.T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. Psychological review *100*, 432–59.

Camperi M., and Wang X.J. (1998). A model of visuospatial working memory in prefrontal cortex: recurrent network and cellular bistability. Journal of computational neuroscience *5*, 383–405.

Carroll S., Josić K., and Kilpatrick Z.P. (2014). Encoding certainty in bump attractors. Journal of computational neuroscience *37*, 29–48.

Christophel T.B., Klink P.C., Spitzer B., Roelfsema P.R., and Haynes J.D. (2017). The distributed nature of working memory. Trends in Cognitive Sciences .

Compte A., Brunel N., Goldman-Rakic P.S., and Wang X.J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. Cerebral Cortex *10*, 910–923.

Constantinidis C., Franowicz M.N., and Goldman-Rakic P.S. (2001). The sensory nature of mnemonic representation in the primate prefrontal cortex. Nature neuroscience *4*, 311–316.

Constantinidis C., and Klingberg T. (2016). The neuroscience of working memory capacity and training. Nature Reviews Neuroscience .

Cowan N. (2008). What are the differences between long-term, short-term, and working memory? Progress in brain research *169*, 323–338.

Diaconis P., Ylvisaker D., et al. (1979). Conjugate priors for exponential families. The Annals of statistics *7*, 269–281.

Farashahi S., Donahue C.H., Khorsand P., Seo H., Lee D., and Soltani A. (2017). Metaplasticity as a neural substrate for adaptive learning and choice under uncertainty. Neuron *94*, 401–414.

Funahashi S., Bruce C.J., and Goldman-Rakic P.S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. Journal of neurophysiology *61*, 331–349.

Gigerenzer G., and Gaissmaier W. (2011). Heuristic decision making. Annual review of psychology *62*, 451–482.

Glaze C.M., Kable J.W., and Gold J.I. (2015). Normative evidence accumulation in unpredictable environments. Elife *4*, e08825.

Gold J.I., and Shadlen M.N. (2002). Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. Neuron *36*, 299–308.

Goldman-Rakic P.S. (1995). Cellular basis of working memory. Neuron *14*, 477–485.

Hansel D., and Mato G. (2013). Short-term plasticity explains irregular persistent activity in working memory tasks. Journal of Neuroscience *33*, 133–149.

Hasson U., Chen J., and Honey C.J. (2015). Hierarchical process memory: memory as an integral component of information processing. Trends in cognitive sciences *19*, 304–313.

Häusser M., and Roth A. (1997). Estimating the time course of the excitatory synaptic conductance in neocortical pyramidal cells using a novel voltage jump method. Journal of Neuroscience *17*, 7606–7625.

Hempel C.M., Hartman K.H., Wang X.J., Turrigiano G.G., and Nelson S.B. (2000). Multiple forms of short-term plasticity at excitatory synapses in rat medial prefrontal cortex. Journal of neurophysiology *83*, 3031–3041.

Hulme S.R., Jones O.D., Raymond C.R., Sah P., and Abraham W.C. (2014). Mechanisms of heterosynaptic metaplasticity. Phil. Trans. R. Soc. B *369*, 20130148.

Itskov V., Hansel D., and Tsodyks M. (2011). Short-term facilitation may stabilize parametric working memory trace. Frontiers in computational neuroscience *5*, 40.

Jackman S.L., and Regehr W.G. (2017). The mechanisms and functions of synaptic facilitation. Neuron *94*, 447–464.

Jackman S.L., Turecek J., Belinsky J.E., and Regehr W.G. (2016). The calcium sensor synaptotagmin 7 is required for synaptic facilitation. Nature *529*, 88–91.

Kilpatrick Z.P., and Bressloff P.C. (2010). Stability of bumps in piecewise smooth neural fields with nonlinear adaptation. Physica D: Nonlinear Phenomena *239*, 1048–1060.

Kilpatrick Z.P., and Ermentrout B. (2013). Wandering bumps in stochastic neural fields. SIAM Journal on Applied Dynamical Systems *12*, 61–94.

Kilpatrick Z.P., Ermentrout B., and Doiron B. (2013). Optimizing working memory with heterogeneity of recurrent cortical excitation. The Journal of Neuroscience *33*, 18999–19011.

Kim T.D., Kabir M., and Gold J.I. (2017). Coupled decision processes update and maintain saccadic priors in a dynamic environment. Journal of Neuroscience *37*, 3632–3645.

Klingberg T. (2010). Training and plasticity of working memory. Trends in cognitive sciences *14*, 317–324.

Lester R., and Jahr C.E. (1992). Nmda channel behavior depends on agonist affinity. Journal of Neuroscience *12*, 635–643.

Lim S., and Goldman M.S. (2013). Balanced cortical microcircuitry for maintaining information in working memory. Nature neuroscience *16*, 1306–1314.

Luck S.J., and Vogel E.K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. Trends in cognitive sciences *17*, 391–400.

Ma W.J., Husain M., and Bays P.M. (2014). Changing concepts of working memory. Nature neuroscience *17*, 347–356.

Markowitz D.A., Curtis C.E., and Pesaran B. (2015). Multiple component networks support working memory in prefrontal cortex. Proceedings of the National Academy of Sciences *112*, 11084–11089.

Markram H., and Tsodyks M. (1996). Redistribution of synaptic efficacy between neocortical pyramidal neurons. Nature *382*, 807.

Meyer T., Qi X.L., Stanford T.R., and Constantinidis C. (2011). Stimulus selectivity in dorsal and ventral prefrontal cortex after training in working memory tasks. Journal of Neuroscience *31*, 6266–6276.

Meyers E.M., Qi X.L., and Constantinidis C. (2012). Incorporation of new information into prefrontal cortical activity after learning working memory tasks. Proceedings of the National Academy of Sciences *109*, 4651–4656.

Mi Y., Katkov M., and Tsodyks M. (2017). Synaptic correlates of working memory capacity. Neuron *93*, 323–330.

Mongillo G., Barak O., and Tsodyks M. (2008). Synaptic theory of working memory. Science *319*, 1543–1546.

Nassar M.R., Helmers J.C., and Frank M.J. (2017). Chunking as a rational strategy for lossy data compression in visual working memory tasks. bioRxiv p. 098939.

Navarro D.J., and Newell B. (2014). Information versus reward in a changing world. 36th Annual Meeting of the Cognitive Science Society .

Papadimitriou C., Ferdoash A., and Snyder L.H. (2015). Ghosts in the machine: memory interference from the previous trial. Journal of neurophysiology *113*, 567–577.

Pesaran B., Pezaris J.S., Sahani M., Mitra P.P., and Andersen R.A. (2002). Temporal structure in neuronal activity during working memory in macaque parietal cortex. Nature neuroscience *5*, 805–811.

Ploner C.J., Gaymard B., Rivaud S., Agid Y., and Pierrot-Deseilligny C. (1998). Temporal limits of spatial working memory in humans. European Journal of Neuroscience *10*, 794–797.

Qi Y., Breakspear M., and Gong P. (2015). Subdiffusive dynamics of bump attractors: mechanisms and functional roles. Neural computation .

Renart A., Song P., and Wang X.J. (2003). Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. Neuron *38*, 473–485.

Risken H. (1996). The Fokker-Planck equation (Springer).

Rolls E.T., Dempere-Marco L., and Deco G. (2013). Holding multiple items in short term memory: a neural mechanism. PloS one *8*, e61078.

Romo R., Brody C.D., Hernández A., and Lemus L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. Nature *399*, 470–473.

Rose N.S., LaRocque J.J., Riggall A.C., Gosseries O., Starrett M.J., Meyering E.E., and Postle B.R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. Science *354*, 1136–1139.

Scholl B.J., and Pylyshyn Z.W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. Cognitive psychology *38*, 259–290.

Shaham N., and Burak Y. (2017). Slow diffusive dynamics in a chaotic balanced neural network. PLoS computational biology *13*, e1005505.

Stokes M.G., Kusunoki M., Sigala N., Nili H., Gaffan D., and Duncan J. (2013). Dynamic coding for cognitive control in prefrontal cortex. Neuron *78*, 364–375.

Summerfield C., and Tsetsos K. (2015). Do humans make good decisions? Trends in cognitive sciences *19*, 27–34.

Tsodyks M., Pawelzik K., and Markram H. (1998). Neural networks with dynamic synapses. Neural computation *10*, 821–835.

Tsodyks M.V., and Markram H. (1997). The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. Proceedings of the National Academy of Sciences *94*, 719–723.

Veliz-Cuba A., Kilpatrick Z.P., and Josic K. (2016). Stochastic models of evidence accumulation in changing environments. SIAM Review *58*, 264–289.

Vogel E.K., McCollough A.W., and Machizawa M.G. (2005). Neural measures reveal individual differences in controlling access to working memory. Nature *438*, 500–503.

Wald A., and Wolfowitz J. (1948). Optimum character of the sequential probability ratio test. The Annals of Mathematical Statistics *19*, 326–339.

Wang Y., Markram H., Goodman P.H., Berger T.K., Ma J., and Goldman-Rakic P.S. (2006). Heterogeneity in the pyramidal network of the medial prefrontal cortex. Nature neuroscience *9*, 534–542.

Wei Z., Wang X.J., and Wang D.H. (2012). From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. Journal of Neuroscience *32*, 11228–11240.

White J.M., Sparks D.L., and Stanford T.R. (1994). Saccades to remembered target locations: an analysis of systematic and variable errors. Vision research *34*, 79–92.

Wilson H.R., and Cowan J.D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. Biological Cybernetics *13*, 55–80.

Wilson R.C., Nassar M.R., and Gold J.I. (2010). Bayesian online learning of the hazard rate in change-point problems. Neural computation *22*, 2452–2476.

Wimmer K., Nykamp D.Q., Constantinidis C., and Compte A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. Nature neuroscience *17*, 431–439.

York L.C., and Van Rossum M.C. (2009). Recurrent networks with short term synaptic depression. Journal of computational neuroscience *27*, 607–620.

Zenke F., and Gerstner W. (2017). Hebbian plasticity requires compensatory processes on multiple timescales. Phil. Trans. R. Soc. B *372*, 20160259.

Zhang W., and Luck S.J. (2008). Discrete fixed-resolution representations in visual working memory. Nature *453*, 233–235.