Synaptic mechanisms of interference in working memory

Zachary P. Kilpatrick[1,2]

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

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

Neuronal studies of visual working memory typically focus on population activity within a single trial, ignoring serial correlations across trials[11]. Several authors have identi ed behavioral biases that cause the previous trial's visual target to interfere with the subject's response on the subsequent trial[12,13]. For instance, in delayed match-to-sample tests, false alarms occur more o en when comparison stimuli match samples from previous trials[14]

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

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

suggesting such a computation could be implemented and represented by neural circuits. Temporal integration of tuned inputs has been demonstrated in both neural recordings[37–39] and circuit models

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

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

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

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

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

Here $\theta(t$

$q(x, t)$ determines an evolving potential function

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

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

Discussion

Comparison with previous work. e work of Papadimitriou *et al.*[13,55] also contains modeling studies, accounting for some aspects of their experimental observations. Our computational model di ers from and extends their ndings in several important ways. We propose that interference can arise as a suboptimal inference

Methods

Assumptions of the inference model.

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

Each location x in the network receives recurrent coupling de ned by the weight function $w(x - y)$ via a convolution $w(x) \ g(x) = \int_{-180}^{180} w(x - y)g(y)\,dy$. We take this function to be peaked when $x = y$ and decreasing as the distance $|x - y$

in[13]. Intertrial intervals are varied to produce Fig. 5B by drawing $T_I^n := t_{n+1} - (T_C + T_D^n + T_A)$ randomly from a uniform pmf for the discrete set of times $T_I^n \in \{1000, 1200, \ldots, 5000\}$ ms and θ_n randomly as in Fig. 5A and identifying the θ_n that produces the maximal bias for each value of T_I^n. Delay-periods are varied to produce Fig. 5C by drawing T_D^n randomly from a uniform pmf for the discrete set of times $T_I^n \in \{0, 200, \ldots, 5000\}$ ms and following a similar procedure to Fig. 5B. Draws from a uniform density function $P(\theta_n) \equiv \bar{P}_0$, de ned on $\theta_n \in [-180, 180)°$ are used to generate the distribution in Fig. 6A and plots in Fig. 7. Nontrivial correlation structure in target selection is defined by the sum of a von Mises distribution and uniform distribution $\mathrm{corr}(\theta_{n+1}, \theta_n) = (1 - \varepsilon)\ {}_{\nu}e^{25\cos(\theta_n - \theta_{n+1} - \mu)} + \varepsilon\bar{P}_0$ for xed θ_n with $\varepsilon = 0.5$; $\mu = 0$ for local correlations (Fig. 6B) and $\mu = 90$ for skewed correlations (Fig. 6C).

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

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

$$U(x) = \int_{-} \qquad - \qquad .$$

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

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

Explicit solutions for high-gain f ring rate nonlinearities. To explicitly calculate solutions, we take the limit of high-gain, so that $F(u) \to H(u - \kappa)$ and $w(x) = \cos(\omega_1 x)$, note $\omega_1 = 180/\pi$. Note, we have compared our predictions here with the results of numerical simulations for sigmoidal ring rates $F(u) = 1/[1 + \mathrm{e}^{-\gamma(u-\kappa)}]$ with gain $\gamma = 20$, and the results are in good agreement. In this case, the bump solution $U(x - x_0) = (2\sin(a)/\omega_1)\cos(\omega_1(x - x_0))$ for $U(\pm a) = \kappa$ and null vector $V(x - x_0) = \delta(x - x_0 - a) - \delta(x - x_0 + a)$ (without loss of generality we take $x_0 \equiv 0$)[47]. Furthermore, we can determine the form of the evolution of $q(x, t)$ by studying the stationary solutions to Eq. (15) in the absence of noise ($W \equiv 0$). For a bump $U(x)$ centered at $x_0 = 0$, the associated stationary form for $Q(x)$ assuming $H(U(x) - \kappa) = 1$ for $x \in (-a, a)$ and zero otherwise is $Q(x) = \beta q_+/(1 + \beta)$ for $x \in (-a, a)$ and zero otherwise. us, if the previous target was at θ_n, we expect $q(x, t)$ to have a shape resembling $Q(x - \theta_n)$ a er 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 $_n(T^n) = (1 - \mathrm{e}^{-(T_C + T_D^n)/\tau})\mathrm{e}^{-T_I^n/\tau}$ ($T^n = T_C + T_D^n + T_I^n$ is the total time g]TJ EMC /T6 tt

68. Lim, S. & Goldman, M. S. Balanced cortical microcircuitry for maintaining information in working memory. *Nat. Neurosci.* **16**, 1306–1314 (2013).
69. Boerlin, M., Machens, C. K. & Denève, S. Predictive coding of dynamical variables in balanced spiking networks. *PLoS Comput. Biol.* **9**, e1003258 (2013).
70. Shaham, N. & Burak, Y. Slow di usive dynamics in a chaotic balanced neural network. *PLoS Comput. Biol.* **13**, e1005505 (2017).
71. Ma, W. J., Husain, M. & Bays, P. M. Changing concepts of working memory. *Nat. Neurosci.* **17**, 347–356 (2014).
72. Nassar, M. R., Helmers, J. C. & Frank, M. J. Chunking as a rational strategy for lossy data compression in visual working memory tasks. *bioRxiv* 098939 (2017).
73. Zhang, W. & Luck, S. J. Discrete xed-resolution representations in visual working memory. *Nature* **453**, 233–235 (2008).
74. Luck, S. J. & Vogel, E. K. Visual working memory capacity: from psychophysics and neurobiology to individual di erences. *Trends. Cogn. Sci.* **17**, 391–400 (2013).
75. Bays, P. M. & Husain, M. Dynamic shi s of limited working memory resources in human vision. *Science* **321**, 851–854 (2008).
76. Wei, Z., Wang, X.-J. & Wang, D.-H. From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. *J. Neurosci.* **32**, 11228–11240 (2012).
77. Almeida, R., Barbosa, J. & Compte, A. Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. *J. Neurophysiol.* **114**, 1806–1818 (2015).
78. Bays, P. M. Spikes not slots: noise in neural populations limits working memory. *Trends. Cogn. Sci.* **19**, 431–438 (2015).
79.