Systems/Circuits

Balanced Cortical Microcircuitry for Spatial Working Memory Based on Corrective Feedback Control

Sukbin Lim[1] **and Mark S. Goldman**[1,2]

[1]Center for Neuroscience and [2]Departments of Neurobiology, Physiology, and Behavior, and Ophthalmology and Vision Science, University of California, Davis, Davis, California 95618

A hallmark of working memory is the ability to maintain graded representations of both the spatial location and amplitude of a memorized stimulus. Previous work has identified a neural correlate of spatial working memory in the persistent maintenance of spatially specific patterns of neural activity. How such activity is maintained by neocortical circuits remains unknown. Traditional models of working memory maintain analog representations of either the spatial location or the amplitude of a stimulus, but not both. Furthermore, although most previous models require local excitation and lateral inhibition to maintain spatially localized persistent activity stably, the substrate for lateral inhibitory feedback pathways is unclear. Here, we suggest an alternative model for spatial working memory that is capable of maintaining analog representations of both the spatial location and amplitude of a stimulus, and that does not rely on long-range feedback inhibition. The model consists of a functionally columnar network of recurrently connected excitatory and inhibitory neural populations. When excitation and inhibition are balanced in strength but offset in time, drifts in activity trigger spatially specific negative feedback that corrects memory decay. The resulting networks can temporally integrate inputs at any spatial location, are robust against many commonly considered perturbations in network parameters, and, when implemented in a spiking model, generate irregular neural firing characteristic of that observed experimentally during persistent activity. This work suggests balanced excitatory–inhibitory memory circuits implementing corrective negative feedback as a substrate for spatial working memory.

Key words: **balanced networks; computational model; decision making; derivative feedback; integration; working memory**

Introduction

Working memory refers to an ability to hold information "online" in the absence of sensory inputs. In spatial working memory, the item held in memory is the spatial location of an object that must be recalled after a delay period of up to several seconds. Electrophysiological recordings have revealed neurons in the parietal and frontal cortices that encode the remembered location of a cue through spatially tuned patterns of persistent neural firing (Funahashi et al., 1989; Constantinidis and Steinmetz, 1996; Chafee and Goldman-Rakic, 1998), but the circuit mechanisms maintaining this sustained neural activity remain poorly understood.

Computational modeling has been useful in suggesting possible mechanisms for the generation and storage of spatially specific patterns of persistent neural activity. The vast majority of models consist of networks of excitatory and inhibitory neuronal populations connected by short-range excitation and longer range inhibition (for review, see Ermentrout, 1998; Compte, 2006). Local recurrent excitation between neurons having similar preferred features provides positive feedback that supports long-lasting reverberation of activity, while long-range inhibition stabilizes and shapes the spatially localized patterns of activity. However, although long-range inhibition could be achieved through disynaptic pathways (Melchitzky et al., 2001) or large basket cells (Markram et al., 2004), the neural substrate for widespread inhibition in memory circuits remains unclear because inhibitory projections are typically shorter ranged than excitatory projections (Braitenberg and Schüz, 1998; Douglas and Martin, 2004).

Recent studies of frontal cortical microcircuitry suggest an alternative mechanism, based on negative-derivative feedback rather than positive feedback, may play a critical part in maintaining persistent neural activity. The key experimental observations motivating this hypothesis are that, first, inhibitory and excitatory inputs have been suggested to be balanced in strength in frontal cortical neurons (Shu et al., 2003; Haider et al., 2006) or more generally positively covary in other cortical neurons (Rudolph et al., 2007; Haider and McCormick, 2009), and, second, the kinetics of excitatory-to-excitatory synaptic connections are slower than those of excitatory-to-inhibitory connections (Wang et al., 2008; Wang and Gao, 2009; Rotaru et al., 2011). Recent modeling work (Lim and Goldman, 2013) has shown how these

Received Oct. 29, 2013; revised March 7, 2014; accepted April 4, 2014.

Author contributions: S.L. and M.S.G. designed research; S.L. and M.S.G. performed research; S.L. and M.S.G. analyzed data; S.L. and M.S.G. wrote the paper.

This research was supported by National Institutes of Health grants R01 MH069726 and R01 MH065034, National Science Foundation Grant IIS-1208218, and a UC Davis Ophthalmology Research to Prevent Blindness grant (M.S.G.). We thank N. Brunel, J. Ditterich, and T. Chartrand for valuable discussions and feedback on this manuscript. We thank D. Higgins for valuable discussions on simulations of spiking network models.

The authors declare no competing financial interests.

Correspondence should be addressed to either of the following: Sukbin Lim at her present address: Department of Neurobiology, University of Chicago, Chicago, IL 60637, E-mail: sukbin@uchicago.edu; or Mark S. Goldman, Department of Neuroscience, Department of Neurobiology, Physiology, and Behavior, and Department of Ophthalmology and Vision Science, University of California, Davis, Davis, California 95618, E-mail: msgoldman@ucdavis.edu.

DOI:10.1523/JNEUROSCI.4602-13.2014

 0270-6474/14/346790-17$15.00/0

two conditions provide a mechanism for maintaining persistent activity through negative-derivative feedback that opposes drifts in firing rate: changes in firing rate trigger fast negative feedback that opposes the drift, followed by slower excitatory feedback that rebalances the net synaptic input. Here, we show how such negative-derivative feedback can operate in a spatially specific manner to maintain spatial working memory. Unlike traditional spatial working memory networks that have stereotyped spatial profiles of activity, and thus lose information about stimulus amplitude, we show that negative-derivative feedback models can temporally integrate their inputs and store analog values of stimulus amplitudes as well as spatial locations. Furthermore, by examining the relationship between the structure of the synaptic connectivity and the spatial profiles of persistent activity, we show that derivative-feedback memory networks do not require widespread, lateral inhibition. Finally, we show that the balance of inhibition and excitation that underlies persistent activity is robustly maintained across a range of common perturbations and leads to irregular neuronal firing similar to that observed experimentally (Compte et al., 2003).

Materials and Methods

Here we describe how our firing rate and spiking networks are structured to maintain spatially tuned patterns of persistent firing through a negative-derivative feedback mechanism. Consistent with experimental observations in prefrontal cortex (Goldman-Rakic, 1995), the model networks are organized in a functionally columnar architecture of excitatory and inhibitory neurons (Fig. 1) with each column defined by having a similar preferred spatial feature of the stimulus. Following previous work (Ermentrout, 1998; Wang, 2001; Compte, 2006), we assume that these preferred spatial features are uniformly distributed along a ring and can be characterized by an angular variable θ. Below, we first describe the network structure and equations governing the dynamics of both the firing rate and spiking models. Then, we analytically derive conditions for producing spatially localized persistent activity in networks with either linear or nonlinear dynamics, and with or without translation-invariant symmetry.

Firing rate model of spatial memory network. In the firing rate models, the activities of, and synaptic interactions between, the neurons are parameterized by their preferred spatial feature θ, which ranges from $-\pi$ to π. The dynamics of the firing rates and synaptic state variables are governed by the equations:

$$\begin{aligned}
\tau_E \frac{dr_E(\theta,t)}{dt} &= -r_E(\theta,t) + f_E\left(\sum_{j=E,I}\int_{-\pi}^{\pi} J_{Ej}(\theta,\theta')\, s_{Ej}(\theta',t)d\theta' + i_E(\theta,t)\right)\\
\tau_I \frac{dr_I(\theta,t)}{dt} &= -r_I(\theta,t) + f_I\left(\sum_{j=E,I}\int_{-\pi}^{\pi} J_{Ij}(\theta,\theta')\, s_{Ij}(\theta',t)d\theta' + i_I(\theta,t)\right)\\
\tau_{ij} \frac{ds_{ij}(\theta',t)}{dt} &= -s_{ij}(\theta',t) + r_j(\theta',t) \qquad \text{for } i,j = E \text{ or } I
\end{aligned} \tag{1}$$

where $r_i(\theta,t)$ represents the mean firing rate of the excitatory (E) or inhibitory (I) population i with preferred feature θ. $s_{ij}(\theta',t)$ denotes the synaptic state variable for the connections from population j with preferred feature θ' onto population i for $i,j = E$ or I, and approaches the presynaptic firing rate $r_j(\theta',t)$ with time constant τ_{ij}.

The mean firing rate $r_i(\theta,t)$ approaches $f_i(x_i(\theta,t))$ with intrinsic time constant τ_i, where $f_i(x)$ represents the steady-state neuronal response to input current x. We consider two types of neuronal response functions: linear $f(x) = x$ (Figs. 4, 5*A–D*, top, 6–10) and a nonlinear neuronal response function (Fig. 5*A–D*, bottom) having the Naka–Rushton (Wilson, 1999) form

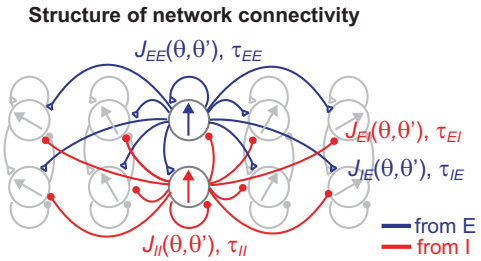

Figure 1. Structure of network model for spatial working memory. We consider a columnar architecture of excitatory and inhibitory neurons such that neurons in the same column have similar preferred feature θ. The connectivity strength J is dependent only on the preferred features θ and θ' of the presynaptic and postsynaptic neurons and τ represents the decay time of synaptic currents at the shown connections. Blue and red curves represent excitatory and inhibitory connections, respectively.

$$f(x) = M\frac{(x - x_\theta)^2}{x_0^2 + (x - x_\theta)^2}\, h(x - x_\theta), \tag{2}$$

where M represents the maximal neuronal response, x_θ represents the input threshold, x_0 defines the value of $(x - x_\theta)$ at which $f(x)$ reaches its half-maximal value, and $h(x)$ denotes the step function $h(x) = 1$ for $x \geq 0$ and $h(x) = 0$ for $x < 0$.

The input $x_i(\theta,t)$ to population i with the preferred feature θ is a sum of the recurrent synaptic currents $J_{ij}(\theta,\theta')\, s_{ij}(\theta',t)$ from population j with the preferred feature θ' and the external current $i_i(\theta,t)$ (not to be confused with the subscript i). $J_{ij}(\theta,\theta')$ represents the synaptic connectivity strength and, except for the nontranslationally invariant model described in the final section of the Materials and Methods, we assume that it depends only on the distance between θ and θ' and can be rewritten as $J_{ij}(\theta - \theta')$. In Figures 6–10, we consider networks with Gaussian-shaped profiles of synaptic connectivity $J_{ij}(\theta - \theta') = J_{ij}\exp[-(\theta - \theta')^2/\sigma_{ij}^2]$, where σ_{ij} here and below denotes $\sqrt{2}$ times the standard deviation of the Gaussian. In Figures 4 and 5, in which only the first cosine Fourier component is tuned to provide negative-derivative feedback, we consider networks that have a (different) Gaussian-shaped profile, plus additional constant and cosine components $J_{ij}(\theta - \theta') = J_{ij,\text{const}} + J_{ij,\cos}\cos(\theta - \theta') + J_{ij,\text{gaus}}\exp[-(\theta - \theta')^2/\sigma_{ij}^2]$.

We assume that the external input $i_i(\theta,t)$ is the sum of constant background input $i_{i,c}$ and time-varying input, where the time-varying component can be expressed separably as the product of a spatial component $i_{i,s}(\theta)$ and a temporal component $i_{i,t}(t)$, so that $i_i(\theta,t) = i_{i,c} + i_{i,s}(\theta)\, i_{i,t}(t)$. The temporal component $i_{i,t}(t)$ represents an external pulse of input that has undergone smoothing before its arrival at the memory network, and is modeled as a pulse of duration $t_{\text{window}} = 500$ ms that is exponentially filtered with time constant $\tau_{\text{ext}} = 100$ ms. The spatial component $i_{i,s}(\theta)$ is a Gaussian function centered at θ_0. For the unimodal activity described in most of the paper, $i_{i,s}(\theta) = i_{i,s,0} + i_{i,s,1}\exp[-(\theta - \theta_0)^2/\sigma_{iO}^2]$. For the multi-modal activity in Figure 6*D*, it is the sum of Gaussian functions $i_{i,s}(\theta) = i_{i,s,0} + \sum_{k=1}^{3} i_{i,s,1}^k \exp[-(\theta - \theta_0^k)^2/(\sigma_{iO}^k)^2]$, where the superscript k denotes the Gaussian component and is not an exponent. For the temporal integration of spatially localized input in Figure 5, $i_{i,s}(\theta)$ is given by $i_{i,s}(\theta) = i_{i,s,0} + i_{i,s,1}\cos(\theta - \theta_0)$.

Throughout the paper except in Figure 8, the intrinsic time constants of excitatory and inhibitory neurons, τ_E and τ_I, are 20 and 10 ms, respectively (McCormick et al., 1985). The time constants of $GABA_A$-type inhibitory synapses, τ_{EI} and τ_{IP}, are each 10 ms (Salin and Prince, 1996; Xiang et al., 1998). Based upon experimental measurements of excitatory synaptic currents in prefrontal cortex (Rotaru et al., 2011), the time constants of excitatory synaptic currents, τ_{EE} and τ_{IE}, were set to 100 and 25 ms, respectively. Note that these time constants reflect the kinetics of postsynaptic potentials triggered by activation of NMDA- and AMPA-type receptors, but likely include the effects of additional intrinsic ionic conductances since these experiments were performed without blocking intrinsic ionic currents (Rotaru et al., 2011).

For the nonlinear function of Naka–Rushton form in Equation 2, the maximal response $M = 100$, the half-activation parameter $x_0 = 40$, and the input threshold $x_\theta = 10$. The parameters for the spatial components of the synaptic connectivity and external input were assigned as follows: in Figures 4 and 5, the parameters for the Gaussian component of the connectivity are $J_{EE,gaus} = 50/\pi$, $J_{IE,gaus} = J_{EI,gaus} = J_{II,gaus} = 100/\pi$, $\sigma_{EE} = \sigma_{IE} = \sigma_{EI} = \sigma_{II} = 0.2\pi$. The parameters for the amplitudes of the constant and cosine terms of the connectivity were defined as $J_{EE,const} = 250/\pi - J_{EE,gaus}\, a_0$, $J_{IE,const} = J_{EI,const} = J_{II,const} = 300/\pi - J_{IE,gaus}\, a_0$, $J_{EE,cos} = 150/\pi - J_{EE,gaus}\, a_1$, $J_{IE,cos} = 300/\pi - J_{IE,gaus}\, a_1$, $J_{EI,cos} = 100/\pi - J_{EI,gaus}\, a_1$, $J_{II,cos} = 200/\pi - J_{II,gaus}\, a_1$, where a_0 and a_1 are multiplicative factors deriving from the constant and first cosine components of the Gaussian portion of the connectivity, and are defined as $a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} d\theta \exp[-(\theta/0.2\pi)^2] \approx 0.1\sqrt{\pi}$ and $a_1 = \frac{1}{\pi}\int_{-\pi}^{\pi} d\theta \cos(\theta) \exp[-(\theta/0.2\pi)^2] \approx 0.2e^{-0.01\pi^2}\sqrt{\pi}$. With these definitions, the overall first cosine component of the connectivity satisfied the balance condition of Equation 9 below. $\sigma_{EO} = 0.25\pi$, $i_{Ec} = 10{,}000$, $i_{Ic} = 9000$, $i_{E0} = 500$, $i_{I0} = 0$, and $i_{I1} = 0$ in Figures 4 and 5*A*,*B*. $i_{E1} = 300$ in Figures 4 and 5*A*, varies between 200 and 500 in Figure 5*B*, top, and varies between 200 and 800 in Figure 5*B*, bottom. $i_{Ec} = 5000$, $i_{Ic} = 0$, $i_{I0} = i_{I1} = 0$, $i_{E0} = i_{E1} = 80$ in Figure 5*C*,*D*. The parameters in Figures 6–9 are the following: $J_{EE,1} = 100$, $J_{IE,1} = 200$, $J_{EI,1} = 100$, $J_{II,1} = 200$, $\sigma_{EE} = \sigma_{IE} = 0.1\pi$, $\sigma_{EI} = \sigma_{II} = 0.2\pi$, $\sigma_{EO} = 0.4\pi$, $i_{Ec} = i_{Ic} = 0$, $i_{E0} = 100$, $i_{I0} = 0$, $i_{E1} = 135$, and $i_{I1} = 0$, except in Figure 6*D* where $i_{E0} = 100$, $i_{I0} = 0$, $i_{I1} = 0$, and $i^1_{E,1} = 150$, $i^2_{E,1} = i^3_{E,1} = 100$, $\theta^1_0 = 0$, $\theta^{2,3}_0 = \pm 2\pi/3$ and $\sigma^{1,2,3}_{EO} = \pi/6$.

In the spatial working memory networks without negative-derivative feedback (Fig. 8), τ_{EE} and τ_{IE} equal 100 ms, and the remaining time constants are the same as the corresponding ones for the negative-derivative feedback networks. The neuronal response (input current–output firing rate) functions in this figure were chosen to be linear for the inhibitory neurons and, for the excitatory neurons, a piecewise linear function given by $f(x) = 1.4(x-1) + 3.5$ for $x < 1$, $f(x) = 14(x-1) + 3.5$ for $1 \le x < 2$, and $f(x) = 7(x-2) + 17.5$ for $2 \le x$. The spatial component of the synaptic connectivity is a Gaussian function $J_{ij}(\theta - \theta') = J_{ij} \exp[-(\theta - \theta')^2/\sigma_{ij}^2]$ with no *I*-to-*I* connection (and, for Fig. 8 *D*,*H*,*L*,*P* only, with the addition of a constant function). The corresponding parameters are as follows: $J_{EE,gaus} = J_{IE,gaus} = 0.5/\pi$, $J_{EI,gaus} = 2.5/\pi$, $\sigma_{EE} = \sigma_{IE} = 0.2\pi$, $\sigma_{EI} = 0.1\pi$ for Figure 8*A*, *E*, *I*, and *M*; $J_{EE,gaus} = 0.5/\pi$, $J_{IE,gaus} = 1/\pi$, $J_{EI,gaus} = 0.5/\pi$, $\sigma_{EE} = 0.2\pi$, $\sigma_{IE} = \pi$, $\sigma_{EI} = 0.1\pi$ for Figure 8*B*, *F*, *J*, and N; $J_{EE,gaus} = J_{IE,gaus} = J_{EI,gaus} = 0.5/\pi$, $\sigma_{EE} = \sigma_{IE} = 0.2\ \pi$, $\sigma_{EI} = \pi$ for Figure 8*C*, *G*, *K*, and *O*; and $J_{EE,gaus} = J_{EI,gaus} = 0.5/\pi$, $J_{IE,gaus} = 1/\pi$, $\sigma_{EE} = \sigma_{IE} = 0.2\pi$, $\sigma_{EI} = 0.1\pi$, and with the addition of a constant value $0.1/\pi$ to the *I*-to-*E* connection for Figure 8 *D*, *H*, *L*, and *P*. The spatial profile of the transient external input is the same for all networks and is given by $i_{i,s}(\theta) = 0.5 + 0.5\cos(\theta)$.

All the simulations of the firing rate models were run with a fourth-order explicit Runge–Kutta method in MATLAB.

Spiking network of leaky integrate-and-fire neurons. In Figure 11, we constructed a recurrent network of excitatory and inhibitory populations of spiking neurons with balanced excitation and inhibition. The activities of, and synaptic interactions between, the neurons are parameterized by their preferred spatial feature θ, which ranges from $-\pi$ to π, as in the firing rate models. Here, we describe the intrinsic dynamics of the individual neurons and the synaptic currents connecting the neurons.

The spiking network consists of N_E excitatory and N_I inhibitory current-based leaky integrate-and-fire neurons that emit a spike when a threshold is reached and then return to a reset potential after a brief refractory period. The neurons are recurrently connected to each other and receive transient stimuli from an external population of N_O neurons. The connectivity between neurons is sparse and random with constant connection probability ρ_i so that, on average, each neuron receives $N_E\rho_E$, $N_I\rho_I$, and $N_O\rho_O$ synaptic inputs from the excitatory, inhibitory, and external populations, respectively. The strengths of the recurrent connections and connections from the external population are dependent on the difference between the preferred feature θ of the postsynaptic neuron and the preferred feature θ' of the presynaptic neuron.

The dynamics of the subthreshold membrane potential V_i^l of the lth neuron in population i and the dynamics of the synaptic input variables s_{ij}^{lm} onto this neuron from the mth neuron in population j are given as follows:

$$\tau_i \frac{dV_i^l}{dt} = -(V_i^l - V_L) + \sum_m \tilde{J}_{iE}^{lm} p_{iE}^{lm} (q_{iE}^N s_{iE}^{lm,N}(t) + q_{iE}^A s_{iE}^{lm,A}(t)) - \sum_m \tilde{J}_{iI}^{lm} p_{iI}^{lm} s_{iI}^{lm}(t) + \sum_m \tilde{J}_{iO}^{lm} p_{iO}^{lm} s_{iO}^{lm}(t) \quad (3)$$

$$\tau_{ij}^k \frac{ds_{ij}^{lm,k}}{dt} = -s_{ij}^{lm,k} + \sum_{t_j^m} \delta(t - t_j^m), \text{ for } j = E,I, \text{ or } O, \text{ and } k = N \text{ or } A. \quad (4)$$

The first term on the right-hand side of Equation 3 corresponds to a neuronal intrinsic leak process such that, without the input, the voltage decays to the resting potential V_L with time constant τ_i. The second term is the sum of the recurrent NMDA- and AMPA-mediated excitatory synaptic currents. The dynamic variables $s_{iE}^{lm,N}$ and $s_{iE}^{lm,A}$ represent NMDA- and AMPA-mediated synaptic currents from cell m of the excitatory population. The fractions of NMDA- and AMPA-mediated currents are assumed to be uniform across the population and are denoted by q_{iE}^N and $q_{iE}^A = 1 - q_{iE}^N$, respectively. p_{iE}^{lm} is a binary random variable with probability ρ_E and represents the random connectivity between neurons. The sum of the strengths of the NMDA- and AMPA-mediated synaptic currents is a Gaussian function given by the following:

$$\tilde{J}_{iE}^{lm} = \tilde{J}_{iE} \exp[-(\theta_i^l - \theta_E^m)^2/{\sigma_{iE}}^2] \text{ where } \theta_i^l = 2\pi l/N_i - \pi, \quad \theta_E^m = 2\pi m/N_E - \pi. \quad (5)$$

Similarly, the third and fourth terms represent the total synaptic inputs from the inhibitory population and the external population. The dynamic variables s_{iI}^{lm} and s_{iO}^{lm} denote inhibitory and external synaptic currents of strengths $\tilde{J}_{iI}^{lm}$ and $\tilde{J}_{iO}^{lm}$, respectively, and p_{iI}^{lm} and p_{iO}^{lm} are binary random variables with probability ρ_I and ρ_O, respectively.

In the dynamics of $s_{ij}^{lm,k}$ in Equation 4, a presynaptic spike at time t_j^m from neuron m in population j causes a discrete jump in synaptic current followed by an exponential decay with time constant τ_{ij}^k. The spikes arriving from the external population represent stimulus-driven inputs to be remembered and are generated by a Poisson process with rate r_O during a time window t_{window} ($r_O = 0$ during the memory period). Note that the strength of $s_{ij}^{lm,j}$, denoted by $\tilde{J}_{ij}^{lm}$ in Equation 3, corresponds to the integrated area under a single postsynaptic potential, not the height of a single postsynaptic potential. Furthermore, the connectivity strengths $\tilde{J}_{ij}^{lm}$ were scaled as follows:

$$\tilde{J}_{ij}^{lm} = \hat{J}_{ij}^{lm}/\sqrt{N_j p} \text{ for fixed } \hat{J}_{ij}^{lm}. \quad (6)$$

This scaling enabled the fluctuations in the input to remain of the same order of magnitude as the mean input as the network size varied (van Vreeswijk and Sompolinsky, 1996, 1998).

In Figure 11, *E* and *H*, the coefficients of variation of the interspike intervals were computed for 3 s from time 300 to 3300 ms using all excitatory neurons that exhibited >5 spikes during this period. CV_2 measures the variability of the interspike intervals locally when the activity is not stationary, and is defined as $\langle CV_2 \rangle = \frac{1}{N-1}\sum_n CV_2(n)$, $CV_2(n) = \frac{2|ISI_{n+1} - ISI_n|}{ISI_{n+1} + ISI_n}$ where ISI_n denotes the nth interspike interval (Holt et al., 1996).

In all spiking simulations, $N_E = 16000$, $N_I = 4000$, $N_O = 20000$, $\rho_E = \rho_O = 0.2$, and $\rho_I = 0.4$. The time constants and the fractions of NMDA-mediated currents were $\tau_E = 20$ms, $\tau_I = 10$ms, $\tau_{EI} = \tau_{II} = 10$ms, $\tau_{EE}^N = 150$ms, $\tau_{EE}^A = 50$ms, $\tau_{IE}^N = 45$ms, $\tau_{IE}^A = 20$ms, $q_{EE}^N = 0.5$, and $q_{IE}^N = 0.2$ (Rotaru et al., 2011). Note that, as in the rate models, these time constants reflect the kinetics of postsynaptic potentials triggered by activation of NMDA- and AMPA-type receptors, but likely include the effects of additional intrinsic ionic conductances since these experiments

were performed without blocking intrinsic ionic currents (Rotaru et al., 2011). The remaining parameters of the integrate-and-fire neuron, which were the same for both excitatory and inhibitory neurons, were $V_L = -60$ mV, $V_\theta = -40$ mV, and $V_{\text{reset}} = -52$ mV, with a refractory period $\tau_{\text{ref}} = 2$ ms. The parameters for the synaptic strengths were tuned to achieve a balance, on average, between the excitatory and inhibitory inputs arriving onto each population during sustained activity (Eq. 9), and were set as follows: $J_{EE} = J_{IE} = 29.70$, $J_{IE} = J_{II} = 42.43$, $J_{\text{EO},0} = 2.1$, $J_{\text{IO},0} = 0$, $J_{\text{EO},1} = 2.1$, $J_{\text{IO},1} = 0$, $\sigma_{EE} = \sigma_{IE} = 0.25\pi$, and $\sigma_{EI} = \sigma_{II} = 0.2\pi$. $r_O = 40$ Hz for excitatory external input neurons with indices from $0.45N_E$ (7200) to $0.55N_E$ (8800) and was zero otherwise.

The numerical integration of the network simulations was performed using the second-order Runge–Kutta algorithm. Spike times were approximated by linear interpolation, which maintains the second-order nature of the algorithm (Hansel et al., 1998).

Derivation of conditions for negative-derivative feedback using Fourier analysis: linear dynamics. Here, we analytically derive conditions for maintaining persistent spatial patterns of activity in firing rate models based on negative-derivative feedback control. First, to illustrate the conditions for negative-derivative feedback control in a simple manner, we assume that the network dynamics are linear and the connectivity pattern is translation invariant. In such a case, Fourier analysis can be used to obtain the conditions for negative-derivative feedback in terms of the Fourier coefficients of the synaptic strengths.

Under the assumption that the connectivity is translationally invariant, that is, the connectivity strength depends only on the difference $\theta - \theta'$ between the preferred features of the presynaptic and postsynaptic neurons so that $J_{ij}(\theta,\theta') = J_{ij}(\theta - \theta')$, all variables and functions of θ in Equation 1 can be rewritten in terms of their Fourier series so that

$$\begin{aligned}
\tau_E \sum_{n=-\infty}^{n=\infty} \frac{d\hat{r}_E(n,t)}{dt} e^{in\theta} &= -\sum_{n=-\infty}^{n=\infty} \hat{r}_E(n,t) e^{in\theta} + \\
&\quad f_E\left(\sum_{n=-\infty}^{n=\infty} e^{in\theta} \left[2\pi \sum_{j=E,I} \hat{J}_{Ej}(n)\hat{s}_{Ej}(n,t) + \hat{\imath}_E(n,t) \right] \right) \\
\tau_I \sum_{n=-\infty}^{n=\infty} \frac{d\hat{r}_I(n,t)}{dt} e^{in\theta} &= -\sum_{n=-\infty}^{n=\infty} \hat{r}_I(n,t) e^{in\theta} + \\
&\quad f_I\left(\sum_{n=-\infty}^{n=\infty} e^{in\theta} \left[2\pi \sum_{j=E,I} \hat{J}_{Ij}(n)\hat{s}_{Ij}(n,t) + \hat{\imath}_I(n,t) \right] \right) \\
\tau_{ij} \sum_{n=-\infty}^{n=\infty} \frac{d\hat{s}_{ij}(n,t)}{dt} e^{in\theta} &= [-\hat{s}_{ij}(n,t) + \hat{r}_j(n,t)] e^{in\theta} \qquad \text{for } i,j = E \text{ or } I,
\end{aligned} \tag{7}$$

where the $\hat{x}(n)$ are the Fourier coefficients of the function $x(\theta)$, and are defined by $\hat{x}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} x(\theta) e^{-in\theta}\, d\theta$ (Folland, 2009). The expression for the recurrent input is obtained by using the convolution theorem, which states that the Fourier coefficient of $\int_{-\pi}^{\pi} J_{ij}(\theta - \theta') s_{ij}(\theta',t) d\theta'$ is the product of the Fourier coefficients of $J_{ij}(\theta)$ and $s_j(\theta)$. Furthermore, if we assume linear dynamics with $f_{E,I}(x) = x$, the Fourier components of the different spatial frequencies do not interact with each other and the equation governing the dynamics of each Fourier coefficient is given by the following:

$$\begin{aligned}
\tau_E \frac{d\hat{r}_E(n,t)}{dt} &= -\hat{r}_E(n,t) + 2\pi \sum_{j=E,I} \hat{J}_{Ej}(n)\hat{s}_{Ej}(n,t) + \hat{\imath}_E(n,t) \\
\tau_I \frac{d\hat{r}_I(n,t)}{dt} &= -\hat{r}_I(n,t) + 2\pi \sum_{j=E,I} \hat{J}_{Ij}(n)\hat{s}_{Ij}(n,t) + \hat{\imath}_I(n,t) \\
\tau_{ij} \frac{d\hat{s}_{ij}(n,t)}{dt} &= -\hat{s}_{ij}(n,t) + \hat{r}_j(n,t) \qquad \text{for } i,j = E \text{ or } I
\end{aligned} \tag{8}$$

Thus, we obtain a 6D linear system for each Fourier component, obeying $d\vec{y}/dt = \overleftrightarrow{A}\vec{y}$ where $\vec{y} = (\hat{r}_E(n),\hat{r}_I(n),\hat{s}_{EE}(n),\hat{s}_{IE}(n),\hat{s}_{EI}(n),\hat{s}_{II}(n))$, and $\overleftrightarrow{A}$ is defined in terms of the time constants τ_E,τ_I, and τ_{ij} and the Fourier components $\hat{J}_{ij}(n)$.

The conditions for negative-derivative feedback control within each Fourier mode of this spatially structured network are analogous to those found previously for spatially uniform networks (Lim and Goldman, 2013). Here, we summarize the approach taken in the previous work, and refer the reader to that work for more extensive analysis. To analyze the linear networks, we used the eigenvector decomposition to decompose the coupled 6D system into noninteracting eigenvectors. For a linear system obeying $d\vec{y}/dt = \overleftrightarrow{A}\vec{y}$, the right eigenvectors $\vec{q}_i^{\,r}$ and corresponding eigenvalues λ_i satisfy the equation $\overleftrightarrow{A}\vec{q}_i^{\,r} = \lambda_i \vec{q}_i^{\,r}$ and the decay of each mode is exponential with time constant $\tau_{i,eff} = -1/\text{Re}(\lambda_i)$, where Re denotes the real part. To obtain persistent firing (large $\tau_{i,eff}$), the system should have at least one eigenvector with its corresponding eigenvalue equal to or close to zero. Also, to maintain persistent activity without unbounded growth of activity in the nonpersistent modes requires that all eigenvalues except those close to 0 have a negative real part (Lim and Goldman, 2013, their Supplementary information 1.2 and 1.3).

Applying this analysis to the system in Equation 8, we found conditions for the maintenance of persistent activity in each Fourier component by negative-derivative feedback. The conditions for each Fourier mode n are given by the following:

$$2\pi\hat{J}_{EE}(n) - (2\pi)^2 \hat{J}_{EI}(n)\hat{J}_{IE}(n)/(2\pi\hat{J}_{II}(n) + 1) \ll O(J),$$

$$\text{i.e., } \hat{J}_{EI}(n)\hat{J}_{IE}(n)/(\hat{J}_{EE}(n)\hat{J}_{II}(n)) \sim 1 \tag{9}$$

$$(\tau_{IE} + \tau_{EI})\hat{J}_{EE}(n) - (\tau_{EE} + \tau_{II})\hat{J}_{EI}(n)\hat{J}_{IE}(n)/\hat{J}_{II}(n) \sim O(J),$$

$$\text{i.e., } \tau_{IE} + \tau_{EI} \neq \tau_{EE} + \tau_{II}, \tag{10}$$

where here we have assumed that the magnitudes of the $\hat{J}_{ij}(n)$ are large so that lower order terms in $\hat{J}_{ij}(n)$ can be neglected. Equation 9 represents the balance between the strengths of positive feedback $\hat{J}_{EE}(n)$ and negative feedback $\hat{J}_{EI}(n)\hat{J}_{IE}(n)/\hat{J}_{II}(n)$ in each mode, and we thus refer to it as the balance condition (Fig. 3*B*). Equation 10 constrains the time constants of the positive and negative feedback. The time constants multiplying the feedback strengths correspond to the timescales for the positive and negative feedback, that is, $\tau_+ = \tau_{EE} + \tau_{II}$ and $\tau_- = \tau_{IE} + \tau_{EI}$, where we note that τ_{II} acts as a time constant for positive feedback since the *I*-to-*I* connection inhibits the negative feedback pathway. From Equation 10, these time constants must be unequal, $\tau_+ \neq \tau_-$. Under these conditions, the recurrent input approximates derivative feedback and thus defines the derivative-feedback models.

Additionally, we found the stability conditions on the network parameters for a system in which all eigenvalues except those close to 0 have a negative real part. Using the Routh–Hurwitz criterion (Nise, 2004), we found necessary conditions for stability given by the following:

$$\begin{aligned}
\frac{\hat{J}_{II}(n)}{\tau_I\tau_{II}} &> \frac{\hat{J}_{EE}(n)}{\tau_E\tau_{EE}} \\
\frac{\hat{J}_{II}(n)}{\tau_I\tau_{II}}\left(\frac{1}{\tau_E} + \frac{1}{\tau_{EI}} + \frac{1}{\tau_{IE}} + \frac{1}{\tau_{EE}}\right) &> \frac{\hat{J}_{EE}(n)}{\tau_E\tau_{EE}}\left(\frac{1}{\tau_I} + \frac{1}{\tau_{EI}} + \frac{1}{\tau_{IE}} + \frac{1}{\tau_{II}}\right) \\
\tau_{EE}\tau_{II} &> \tau_{IE}\tau_{EI} \\
\tau_{EE} + \tau_{II} &> \tau_{IE} + \tau_{IE}
\end{aligned} \tag{11}$$

The last condition is similar to Equation 10, which showed that the timescales for the positive and negative feedback must be different to have stable persistent firing. The stability condition above additionally specifies that the positive feedback should be slower than the negative feedback. The third condition is similar to the last condition, except that it constrains the product of the time constants, and the first two conditions require that the excitatory time constants be slower than the inhibitory ones.

Derivation of conditions for negative-derivative feedback using Fourier analysis: nonlinear dynamics. In this section, we consider a network

model in which the individual neurons have a nonlinear firing rate versus input current relationship. In the presence of such nonlinearity, the Fourier components of the firing rates and synaptic variables are no longer independent for the different Fourier modes. However, as shown below, the core principles for the conditions on the network parameters are similar to those for the linear networks, that is, negative-derivative feedback requires, first, a balance between the strengths of positive and negative feedback and, second, that positive feedback is slower than negative feedback.

To find analytically the conditions on negative-derivative feedback in nonlinear networks, we consider a simple model in which the connection strengths and the external input are described by their first two Fourier components, a constant mode and a cosine mode (Ben-Yishai et al., 1995). The first two Fourier coefficients of the quantity x, denoted $a_{0,x}$ and $a_{1,x}$, are given by $a_{l,x} = \frac{1}{\pi}\int_{-\pi}^{\pi}\cos(l\theta)x(\theta)d\theta$ for $l = 0$ or 1 (note that $a_{0,x} = 2\hat{x}(0)$ and $a_{1,x} = \hat{x}(1) + \hat{x}(-1) = \mathrm{Re}\{\hat{x}(1)\}$ in Eq. 8). Then, by projecting the system of Equation 1 onto the first two Fourier components, we obtain the following equations governing the dynamics of $a_{0,x}$ and $a_{1,x}$, for $x = r_E$, r_I, s_{EE}, s_{IE}, s_{EI}, or s_{II}, during the memory period (when the external input is zero):

$$\tau_i\frac{d}{dt}a_{0,r_i} = -a_{0,r_i} + \frac{1}{\pi}\int_{-\pi}^{\pi}d\theta f_i\left(2\pi\left[\sum_{j=E,I}\frac{a_{0,J_{ij}}a_{0,s_{ij}}}{2} + \cos(\theta)\sum_{j=E,I}a_{1,J_{ij}}a_{1,s_{ij}}\right]\right)$$

$$\tau_i\frac{d}{dt}a_{1,r_i} = -a_{1,r_i} + \frac{1}{\pi}\int_{-\pi}^{\pi}d\theta\cos(\theta) f_i\left(2\pi\left[\sum_{j=E,I}\frac{a_{0,J_{ij}}a_{0,s_{ij}}}{2} + \cos(\theta)\sum_{j=E,I}a_{1,J_{ij}}a_{1,s_{ij}}\right]\right)$$

$$\tau_{ij}\frac{d}{dt}a_{0,S_{ij}} = -a_{0,S_{ij}} + a_{0,r_j} \quad \text{for } i,j = E, \text{ or } I$$

$$\tau_{ij}\frac{d}{dt}a_{1,S_{ij}} = -a_{1,S_{ij}} + a_{1,r_j} \quad \text{for } i,j = E, \text{ or } I. \tag{12}$$

In the presence of nonlinearity, global analysis of the network dynamics through the eigenvector decomposition is not possible. Instead, we find the conditions by locally linearizing the system around possible steady states and note that the conditions obtained must hold for all steady states that can be maintained persistently. For the steady state to belong to a continuous attractor, there should be at least one eigenvector equal to or close to 0 in the local linearization. If we assume that there exists a steady state and denote it by the superscript SS as $a_{0,x}^{SS}$ and $a_{1,x}^{SS}$ for $x = r_E$, r_I, s_{EE}, s_{IE}, s_{EI}, or s_{II}, Equation 12 becomes

$$\tau_i\frac{d}{dt}\delta a_{0,r_i} = -\delta a_{0,r_i} + 2\pi c_{0,i}\sum_{j=E,I}\frac{a_{0,J_{ij}}}{2}\delta a_{0,s_{ij}} + 2\pi c_{1,i}\sum_{j=E,I}a_{1,J_{ij}}\delta a_{1,s_{ij}}$$

$$\tau_i\frac{d}{dt}\delta a_{1,r_i} = -\delta a_{1,r_i} + 2\pi c_{1,i}\sum_{j=E,I}\frac{a_{0,J_{ij}}}{2}\delta a_{0,s_{ij}} + 2\pi c_{2,i}\sum_{j=E,I}a_{1,J_{ij}}\delta a_{1,s_{ij}}$$

$$\tau_{ij}\frac{d}{dt}\delta a_{0,S_{ij}} = -\delta a_{0,S_{ij}} + \delta a_{0,r_j} \quad \text{for } i,j = E, \text{ or } I$$

$$\tau_{ij}\frac{d}{dt}\delta a_{1,S_{ij}} = -\delta a_{1,S_{ij}} + \delta a_{1,r_j} \quad \text{for } i,j = E, \text{ or } I, \tag{13}$$

where $c_{0,i} = \frac{1}{\pi}\int_{-\pi}^{\pi}d\theta f_i'\left(2\pi\left[\sum_{j=E,I}\frac{a_{0,J_{ij}}a_{0,s_{ij}}^{SS}}{2} + \cos(\theta)\sum_{j=E,I}a_{1,J_{ij}}a_{1,s_{ij}}^{SS}\right]\right)$

$$c_{1,i} = \frac{1}{\pi}\int_{-\pi}^{\pi}d\theta\cos(\theta) f_i'\left(2\pi\left[\sum_{j=E,I}\frac{a_{0,J_{ij}}a_{0,s_{ij}}^{SS}}{2} + \cos(\theta)\sum_{j=E,I}a_{1,J_{ij}}a_{1,s_{ij}}^{SS}\right]\right)$$

$$c_{2,i} = \frac{1}{\pi}\int_{-\pi}^{\pi}d\theta\cos^2(\theta) f_i'\left(2\pi\left[\sum_{j=E,I}\frac{a_{0,J_{ij}}a_{0,s_{ij}}^{SS}}{2} + \cos(\theta)\sum_{j=E,I}a_{1,J_{ij}}a_{1,s_{ij}}^{SS}\right]\right).$$

In the above, $f_i'(x_i)$ denotes the derivative of $f_i(x)$ evaluated at x_i, $\delta a_{0,x} = a_{0,x} - a_{0,x}^{SS}$, and $\delta a_{1,x} = a_{1,x} - a_{1,x}^{SS}$ for $x = r_E$, r_I, s_{EE}, s_{IE}, s_{EI}, or s_{II}. Thus, these equations describe a 12D linear system (two coupled 6D systems, one for the constant mode and the other for the cosine mode). As shown in the previous section, we obtain the conditions for negative-derivative feedback by examining the conditions for the system given by Equation 13 to have an eigenvalue close to 0. These conditions are given by the following:

$$(a_{0,J_{EE}}a_{0,J_{II}} - a_{0,J_{EI}}a_{0,J_{IE}})(a_{1,J_{EE}}a_{1,J_{II}} - a_{1,J_{EI}}a_{1,J_{IE}}) << O(J^4), \tag{14}$$

$$\tau_+ = (\tau_{EE} + \tau_{II}) > (\tau_{IE} + \tau_{EI}) = \tau_-. \tag{15}$$

Equation 15 is the condition for slower positive feedback, which is the same as Equation 11 for the linear networks. Equation 14 can be achieved either when $a_{0,J_{EE}}a_{0,J_{II}} - a_{0,J_{EI}}a_{0,J_{IE}} \ll O(J^2)$ or $a_{1,J_{EE}}a_{1,J_{II}} - a_{1,J_{EI}}a_{1,J_{IE}} \ll O(J^2)$, that is, when either the constant mode or the first cosine mode satisfies a balance condition identical to Equation 9 for the linear networks. Additional inequality conditions for the stability of the system can likewise be obtained by analogy to the analysis underlying Equation 11 for the linear networks.

We note that, for both the linear and nonlinear networks, the condition that positive and negative feedback are balanced leads to a corresponding requirement that the excitatory and inhibitory inputs onto at least the excitatory cells (and, unless J_{EI} and J_{IE} are very different, also the inhibitory cells) are closely balanced as well. The reason for this is that achieving large negative-derivative feedback requires correspondingly large excitatory and inhibitory recurrent inputs. If these inputs were unbalanced, then the total current driving the neural response functions f_E and f_I would be very large. This would cause very large synaptic input to the neurons that would drive strong changes in firing rates rather than maintaining persistent activity. Thus, even in the presence of higher Fourier components of the connection strengths or nonlinear response functions, the balance condition remains the same (derivation not shown) and the core principles for negative-derivative feedback remain the same as in the linear networks.

Derivation of conditions for negative-derivative feedback to maintain arbitrary patterns of activity in nontranslationally invariant networks. In the previous sections, we found the conditions necessary for negative-derivative feedback when the connection strengths are translationally invariant. In this section, we extend our analysis to networks without translation invariance and generalize the conditions for negative-derivative feedback control to such networks.

For simplicity, we assume the network obeys linear dynamics and assume that the neuronal index θ is discrete and uniformly spaced along the ring, with the total number of neurons in either the excitatory or inhibitory population equal to N_θ. Then, in Equation 1, the firing activities and synaptic variables are vectors of length N_θ, the connection strengths are $N_\theta \times N_\theta$ matrices that we denote as $\overleftrightarrow{M}_{ij}$ for $i,j = E$ or I, and Equation 1 can be rewritten as follows:

$$\tau_E \frac{d\vec{r}_E(t)}{dt} = -\vec{r}_E(t) + \overleftrightarrow{M}_{EE}\vec{s}_{EE}(t) + \overleftrightarrow{M}_{EI}\vec{s}_{EI}(t) + \vec{i}_E(t)$$

$$\tau_I \frac{d\vec{r}_I(t)}{dt} = -\vec{r}_I(t) + \overleftrightarrow{M}_{IE}\vec{s}_{IE}(t) + \overleftrightarrow{M}_{II}\vec{s}_{EI}(t) + \vec{i}_I(t)$$

$$\tau_{ij} \frac{d\vec{s}_{ij}(t)}{dt} = -\vec{s}_{ij}(t) + \vec{r}_j(t) \qquad \text{for } i,j = E, \text{ or } I. \tag{16}$$

In this case, the slower positive than negative feedback can be achieved under the same conditions (bottom two equations of Equation 11) found for the translationally invariant networks. On the other hand, the balance condition now is expressed as a relation between the connectivity matrices

$$[(\overleftrightarrow{M}_{EE} - \overleftrightarrow{I}) - \overleftrightarrow{M}_{EI}(\overleftrightarrow{M}_{II} + \overleftrightarrow{I})^{-1}\overleftrightarrow{M}_{IE}]\vec{v} \sim 0 \text{ for some } \vec{v} \neq 0, \tag{17}$$

and the persistent pattern of activity under this condition is

$$\vec{r}_E \sim \vec{v} \text{ and } \vec{r}_I \sim \overleftrightarrow{M}_{EI}^{-1}(\overleftrightarrow{M}_{EE} - \overleftrightarrow{I})\vec{r}_E \sim (\overleftrightarrow{M}_{II} + \overleftrightarrow{I})^{-1}\overleftrightarrow{M}_{IE}\vec{r}_E. \tag{18}$$

For example, if the $\overleftrightarrow{M}_{ij}$'s commute with each other and have a common eigenvector $\vec{v}$ such that $\overleftrightarrow{M}_{ij}\vec{v} = \lambda_{ij}\vec{v}$, then the balance condition becomes $\lambda_{EE} \sim \lambda_{EI}\lambda_{IE}/\lambda_{II}$ for large λ, and $\vec{r}_E \sim \vec{v}$ and $\vec{r}_I \sim \lambda_{EE}/\lambda_{EI}\vec{r}_E \sim \lambda_{IE}/\lambda_{II}\vec{r}_E$. Note that if $\overleftrightarrow{M}_{ij}$ is translationally invariant, the common eigenvectors of $\overleftrightarrow{M}_{ij}$ are the Fourier components discussed previously.

Results

Principle of negative-derivative feedback control for spatial working memory

We consider a spatial working memory model that maintains persistent activity through a negative-derivative feedback mechanism that counteracts drift in memory representations. In this section, we review recent work (Lim and Goldman, 2013) showing how a negative-derivative feedback mechanism can maintain spatially uniform patterns of persistent activity in networks with no spatial structure. In the following sections, we show how this framework can be extended to networks whose spatial structure allows them to maintain stimulus-dependent spatial patterns of activity, and we describe salient properties of these networks.

To illustrate how negative-derivative feedback networks slow memory decay and maintain a graded range of spatially uniform persistent activity, we consider a simple mathematical model of a memory cell with mean firing rate $r(t)$, which receives transient input $I(t)$ to be integrated and maintained during a delay period (Fig. 2A):

$$\tau \frac{dr}{dt} = -r - W_{der}\frac{dr}{dt} + I(t)$$

$$\Rightarrow (\tau + W_{der})\frac{dr}{dt} = -r + I(t). \tag{19}$$

The first term on the right side of the top equation, $-r$, represents intrinsic leak processes that lead to activity decay with time constant τ in the absence of feedback. The second term, $-W_{der}\frac{dr}{dt}$, represents negative-derivative feedback that resists changes in activity such that increases (decreases) in firing rates result in a feedback signal of negative (positive) sign (Fig. 2B). For strong derivative feedback, $W_{der} \gg \tau$, the effective time constant of activity decay $\tau_{eff} = \tau + W_{der}$ is dominated by this derivative feedback, so that the system becomes proportionately more resistant to memory decay as the strength of derivative feedback increases.

Mechanistically, this negative-derivative feedback can arise from recurrent network interactions in memory-storing circuits that contain positive and negative feedback pathways (Fig. 2C). When positive feedback mediated by recurrent excitation and negative feedback mediated by recurrent inhibition have equal strength, but positive feedback has slower kinetics, a neuron receives derivative-like recurrent input: the equal-strength positive and negative feedback lead to nearly zero net input during persistent activity, but the faster negative feedback leads to large input that opposes changes in activity whenever activity fluctuates. In spatially uniform networks, the strength of negative-derivative feedback has been shown (Lim and Goldman, 2013) to be proportional to the strength of the balanced positive and negative feedback and the difference in their timescales, so that

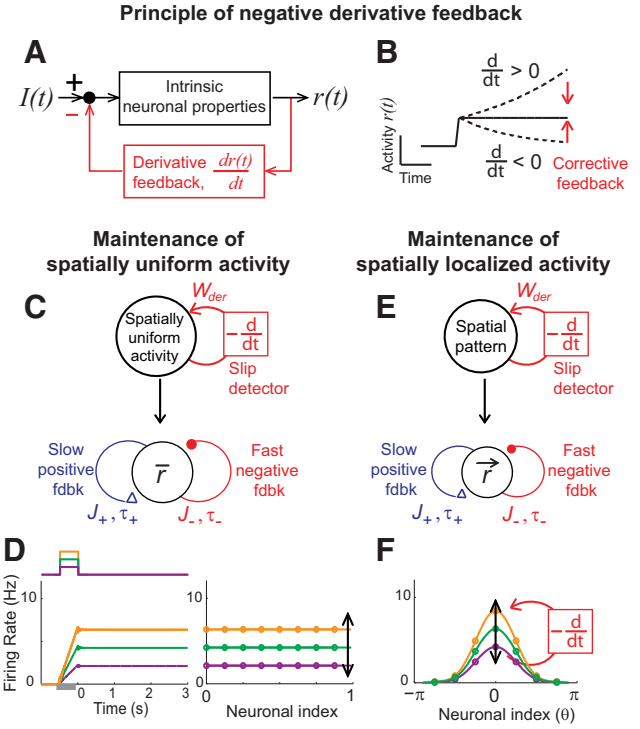

Figure 2. Memory networks with negative-derivative feedback. ***A***, Block diagram illustrating the principle of negative-derivative feedback control for a system with transient external input $I(t)$ and output firing rate $r(t)$. ***B***, A negative-derivative feedback mechanism maintains persistent activity by providing corrective feedback that opposes upward or downward changes in activity. ***C***, Simple black-box model of a neural population with negative-derivative feedback that maintains spatially uniform patterns of persistent activity (top) and simplified network illustrating key components underlying the conditions on the positive and negative feedback pathways for negative-derivative feedback (bottom). ***D***, Time course of average firing rates in spatially uniform negative-derivative feedback networks in response to three example amplitudes of transient stimuli (left) and corresponding maintenance of spatially uniform patterns of activity at different amplitudes (right) during the delay period. ***E***, ***F***, Extension of the mechanism of negative-derivative feedback control to maintaining spatially localized patterns of activity. The basic principles are the same as for the negative-derivative feedback networks for spatially uniform patterns of activity (***E***), but what negative-derivative feedback detects and corrects is the amplitude of particular spatial patterns of activity (***F***). In ***D***, the neurons were rank ordered and the neuronal index is the neuron's order number divided by the network size. In ***F***, the neuronal index is the neuron's preferred spatial feature θ.

$$W_{der} \sim J(\tau_+ - \tau_-), \tag{20}$$

where J denotes the strength of the balanced positive and negative feedback pathways, and τ_+ and τ_- denote the timescales of positive and negative feedback, respectively (Fig. 2*C*). Thus, when the recurrent synaptic interactions contain strong positive and negative feedback that are balanced in strength (large J) but with slower positive feedback ($\tau_+ > \tau_-$), the network temporally integrates its input with long integration time constant $\tau_{eff} \approx W_{der}$, showing step-like activity in response to spatially uniform transient input (Fig. 2*D*). We note that, although the derivative-feedback mechanism maintains persistent activity by resisting changes in firing rate, this does not keep the system from responding to external inputs as long as these inputs are of the same scale as the recurrent synaptic inputs, which would be expected if the strengths of recurrent and external inputs both scale with population size. Furthermore, external input can transiently imbalance the recurrent excitatory and inhibitory feedback, allowing for more rapid response to external inputs (Lim and Goldman, 2013).

Requirements for negative-derivative feedback in circuits with functionally columnar architecture

Here we describe how the mechanism of negative-derivative feedback described above can be extended to networks that maintain spatially localized patterns of persistent neural activity characteristic of those observed during spatial working memory tasks (Fig. 2*E*). The basic concept is the same as above, but for spatial working memory, the feature that negative-derivative feedback detects and corrects is a deviation in the amplitude of a particular spatial pattern of activity $\vec{r} = (r_1, r_2, \ldots, r_n)$, where r_i is the firing rate of the ith neuron in the network (Fig. 2*F*). That is, for any maintained spatial pattern $\vec{r}$, we require that this activity drives recurrent synaptic interactions containing positive and negative feedback signals of equal strength but with slower kinetics for the positive feedback (Fig. 2*E*). Below, we show mathematically how these conditions can be met in a spatially structured network and find the conditions on the spatial profile and kinetics of the synaptic connectivity for negative-derivative feedback control.

We consider networks of excitatory and inhibitory populations that store the angular location of a transiently presented spatial cue that must be remembered during a subsequent delay period. Recordings of the persistent activity of spatially selective memory cells identified in such tasks suggest a functionally columnar architecture in which neurons in the same column have similar preferred features of the stimulus (Goldman-Rakic, 1995; Wimmer et al., 2014). To capture this functional organization, we parameterize the activities of the excitatory and inhibitory neurons by their preferred feature θ, which we assume to be uniformly distributed along a ring (Fig. 1). The connection strength between a presynaptic neuron from the jth population with preferred feature θ' and a postsynaptic neuron from the ith population with preferred feature θ is denoted by $J_{ij}(\theta,\theta')$, where $i = E$ or I denotes whether the presynaptic and postsynaptic neurons are part of the excitatory (E) or inhibitory (I) populations. Time constants for these connections similarly are denoted as τ_{ij}, which is assumed to be independent of θ and θ' for given population types i and j (Fig. 1).

The core requirements for negative-derivative feedback, a balance between the strengths of the positive and negative feedback pathways and slower positive than negative feedback, impose a tuning condition on the connection strengths $J_{ij}(\theta,\theta')$ and a constraint on the time constants of the connections, τ_{ij}. To derive the tuning condition on $J_{ij}(\theta,\theta')$, we assume as in most previous models of orientation-selective spatial working memory that the connectivity $J_{ij}(\theta,\theta')$ is translationally invariant, that is, independent of the absolute values of θ and θ' but dependent on the difference between θ and θ' as $J_{ij}(\theta - \theta')$ (Ermentrout, 1998; Wang, 2001; Compte, 2006). Furthermore, if the dynamics of the system is linear, Fourier analysis can be used to decompose the spatial activity and recurrent interactions into cosine and sine functions of θ that do not interact with each other (Fig. 3*A*). In this case, the strengths of the recurrent connections within each Fourier component are denoted by $\hat{J}_{ij}(n)$ and their timescales are given by τ_{ij} (Fig. 3*B*, top; see Materials and Methods). However, we note that, although translation invariance and linear dynamics are helpful in building intuition and providing a simple illustration of conditions for negative-derivative feedback, neither of these features are necessary requirements for negative-derivative feedback (see Materials and Methods and Fig. 5 for networks with nonlinear dynamics and Materials and Methods for linear networks without translationally invariant connectivity).

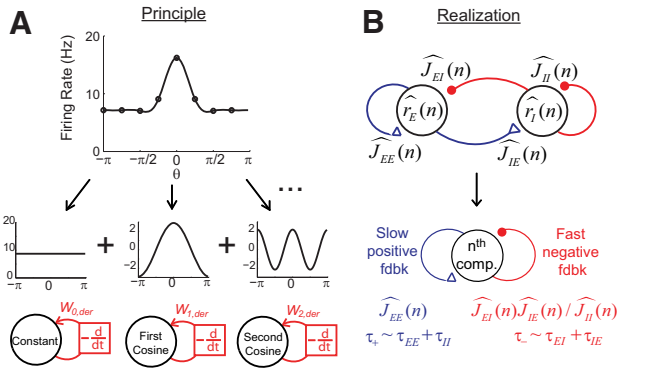

Figure 3. Negative-derivative feedback in linear networks can be analyzed through the Fourier decomposition. ***A***, Example spatial pattern of persistent activity (top) and its Fourier decomposition (middle). Negative-derivative feedback can occur within any or all of the Fourier components (bottom). ***B***, Conditions for negative-derivative feedback in each Fourier component. Top, Illustration of the projection onto the nth Fourier component of the network's activity $\hat{r}_i$ and connectivity $\hat{J}_{ij}$. To have negative-derivative feedback in each Fourier component, the positive feedback within this component must have equal strength, but slower kinetics, than the negative feedback within the component.

Since the dynamics of each Fourier component are independent, negative-derivative feedback can be achieved independently for each component (Fig. 3*A*). Specifically, for the nth Fourier component to be governed by negative-derivative feedback, the positive feedback and negative feedback pathways onto this Fourier component should have equal strength, and the positive feedback pathway should have slower kinetics than the negative feedback pathway. This can be accomplished when two conditions are met:

$$\hat{J}_{EE}(n) \sim \hat{J}_{EI}(n)\hat{J}_{IE}(n)/\hat{J}_{II}(n), \tag{21}$$

$$\tau_+ = \tau_{EE} + \tau_{II} > \tau_{EI} + \tau_{IE} = \tau_-. \tag{22}$$

Equation 21 is the condition for balancing positive feedback and negative feedback for the nth Fourier component. The left side of this condition represents the strength of positive feedback in this Fourier component, which is mediated by the E-to-E connection. The right side represents the strength of negative feedback and is mediated by the E-to-I-to-E feedback loop, with normalization of the strength of this loop provided by the I-to-I connection (Fig. 3*B*, bottom). Equation 22 is the condition for slower positive than

negative feedback. The sum $\tau_+ = \tau_{EE} + \tau_{II}$ represents the sum of the positive feedback contributions, where τ_{II} plays the role of a positive feedback time constant because the *I*-to-*I* connection inhibits the negative feedback pathway, and this feedback must be slower than the time constant associated with the traversal time around the negative feedback loop $\tau_- = \tau_{EI} + \tau_{IE}$ (Fig. 3*B*, bottom; see Materials and Methods).

Throughout this paper, we assume that τ_{EE} is longer than the time constants of the other connections. This assumption is based upon recent experimental observations in prefrontal cortex that found that *E*-to-*E* connections are much slower than *E*-to-*I* connections, due to a relative prominence of slow NMDA-type synapses (Wang et al., 2008; Wang and Gao, 2009; Rotaru et al., 2011). Thus, because the time constants are independent of the particular Fourier component, Equation 22 is satisfied for all Fourier components.

In contrast, the balance condition, given by Equation 21, can be satisfied independently for each Fourier component. To maintain spatially nonuniform persistent activity across the population, this condition must be satisfied by at least one of the nonconstant Fourier components, and the specific spatial profile of persistent activity observed during the delay period reflects the relative balance of the different components satisfying the balance condition.

Maintenance of spatially modulated activity based on a balance between excitation and inhibition

To illustrate the dynamics of the negative-derivative feedback networks and how they maintain spatially localized patterns of persistent activity, we first consider a simple network that has been structured to receive negative-derivative feedback only in its first cosine component (Fig. 4*A*). This network's synaptic connectivity profile contained three components, an untuned uniform component of the connectivity, an untuned component with Gaussian connectivity profile, and a tuned component with cosine profile (see Materials and Methods). The network received a spatially localized input of narrow Gaussian profile centered at 0 degrees during a brief cue period, plus a constant background input that was present during both the cue and delay periods (Fig. 4*B*,*C*). During the cue presentation and shortly after the offset of the cue, the spatial profile of the network activity had a narrow width that directly followed the spatial profile of the transient input (Fig. 4*B*, bright horizontal band centered at 0 degrees during the cue period; *C*, left). However, during the delay period, the activity profile quickly broadened so that only the activity pattern of the first cosine component was maintained (Fig. 4*B*, *C*, middle and right). This is because all Fourier components except the first cosine component decayed quickly back to their baseline activity, which was zero for the higher Fourier components and a constant level driven by the tonic background input for the constant component. In contrast, the first cosine component was maintained throughout the delay period by the negative-derivative feedback (Fig. 4*B*, broad brighter region during delay period; *C*, middle and right). More generally, this example illustrates that the profile of activity maintained by the network reflects only those components that receive negative-derivative feedback.

A feature of the derivative-feedback networks is that, during the delay period, neurons in the network receive strong excitatory and inhibitory inputs that are closely balanced with each other (Fig. 4*D*; see Materials and Methods). The cosine component receives a balance of recurrent excitatory and inhibitory synaptic inputs, as required by the balance condition (Fig. 4*F*). The constant component likewise receives a balance of excitation and inhibition (Fig. 4*E*). However, this balance is achieved through inclusion of the external background input; the recurrent inputs, in contrast, are dominated by inhibition so that the network does not contain negative-derivative feedback in this component and cannot maintain spatially uniform activity in the absence of background input (data not shown). This reflects that both the excitatory and inhibitory inputs to each neuronal population (but not necessarily the excitatory and inhibitory tuning curves or connectivity, as shown in Fig. 7) are spatially localized and have the same spatial tuning widths.

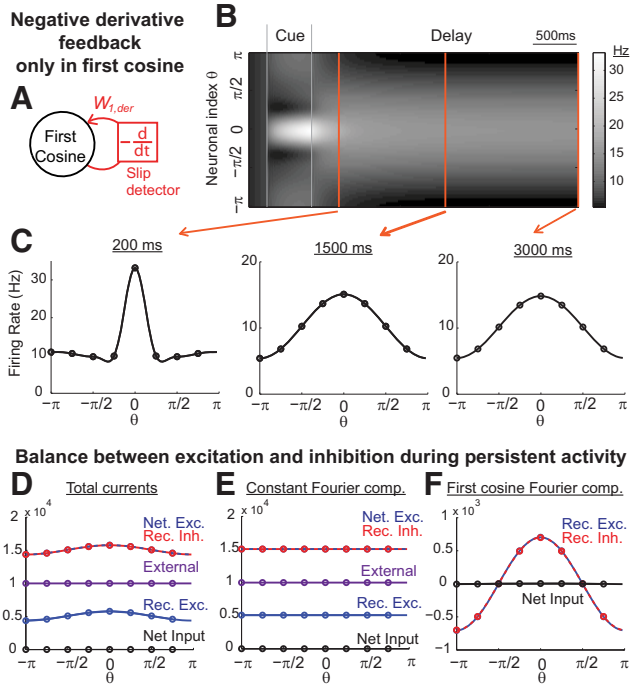

Figure 4. Temporal evolution of activity and balance between excitation and inhibition during memory performance. ***A***, Example network with negative-derivative feedback only in the first cosine component. ***B***, Grayscale-coded spatiotemporal activity pattern. The *x*-axis represents time and the *y*-axis represents the neuronal index parameterized by its preferred cue direction θ. ***C***, Spatial profiles of firing activity at different times corresponding to the orange vertical slices in ***B***. ***D–F***, Close balance between excitation and inhibition during the delay period. The shown neuron receives strong net excitation and inhibition that are balanced in strength. The first cosine Fourier component, but not the constant Fourier component, is maintained by negative-derivative feedback, as indicated in ***F*** by the balance of recurrent excitation with recurrent inhibition within this first cosine component. Values in ***D–F*** are illustrated for the inputs onto an excitatory neuron at 3 s into the delay period.

A close balance between excitatory and inhibitory inputs in memory cells is a distinct feature of negative-derivative feedback. In most previous studies, it has been suggested that spatially localized activity patterns result from excess excitation in high-firing rate neurons and widespread lateral inhibition that stabilizes the bump of activity during the delay period (Ermentrout, 1998; Wang, 2001; Compte, 2006). This leads to inhibitory synaptic inputs onto a postsynaptic cell being more broadly tuned than excitatory inputs in such networks, whereas the spatial tuning of excitatory and inhibitory inputs are similar in negative-derivative feedback networks. Thus, a balance between excitation and inhibition is one prediction of the negative-derivative feedback mechanism that can be tested experimentally (see Discussion).

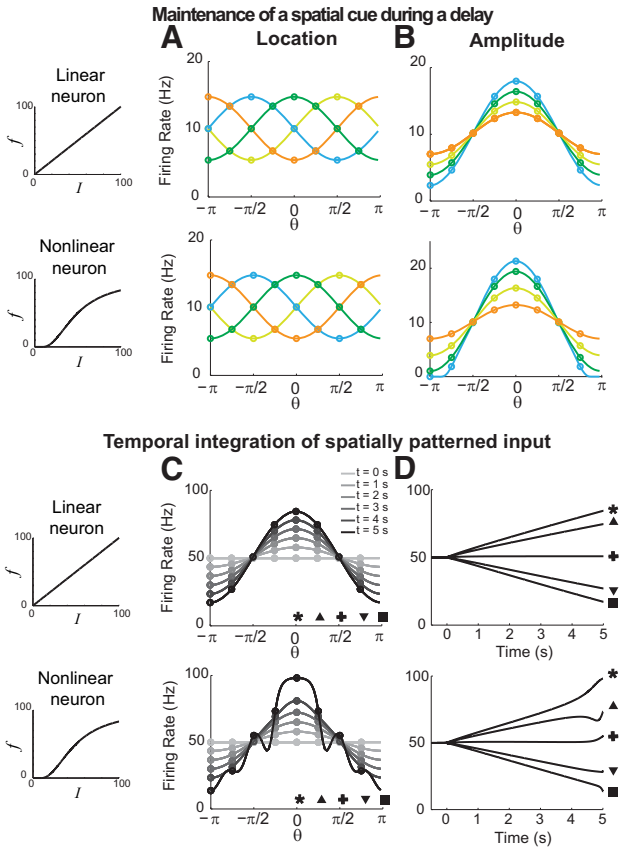

Figure 5. Location codes, amplitude codes, and temporal integration of negative-derivative feedback networks with linear (top) and nonlinear (bottom) firing rate (f) versus input current (I) relationships. ***A***, Maintenance during the delay period of spatial patterns of activity centered at different locations in the network, corresponding to different spatial cues (shown as different colors). ***B***, Maintenance of spatial patterns of activity of different amplitudes during a delay period for inputs at the same spatial location but with different strengths. Input amplitudes for the linear case were smaller than those for the nonlinear case to avoid the negative firing rates that are permitted in linear networks. ***C***, ***D***, Temporal integration of spatially structured input. After time 0, a spatially structured input is continuously present and the amplitude of the spatial pattern of activity linearly increases in time (***C***), resulting in ramp-like changes (***D***). Patterns in ***A*** and ***B*** are shown at 3 s into the delay period. In ***C*** and ***D*** (bottom), the ripply activity pattern at $t = 5$ s reflects the effect of reaching the extreme, saturating limit of the neuronal response function, when the neuron with preferred location $\theta = 0$ has approached its limiting firing rate of 100 Hz and becomes insensitive to additional input.

Location codes, amplitude codes, and neural integration in negative-derivative feedback networks

Traditional spatial working memory models maintain the analog spatial location of a stimulus through stereotyped patterns of network activity centered on the maintained stimulus location, as observed experimentally (Goldman-Rakic, 1995; Wang, 2001). However, a fundamental feature of these models is that the amplitude of the pattern of neuronal activity during the delay period is bistable, either exhibiting untuned background activity or participating in a fixed-amplitude pattern of activity corresponding to the location of the maintained stimulus (Ermentrout, 1998; Wang, 2001; Compte, 2006). Because of this bistability, only the location of the cue can be stored in such networks and, for example, the amplitude or value of the cue cannot be distinguished beyond a binary discrimination.

Negative-derivative feedback networks likewise can maintain an analog spatial location in memory (Fig. 5*A*) and, as in traditional memory models, this can be achieved by having a translation-invariant network connectivity profile that permits the network to maintain a given spatial pattern of activity centered at any location in the network. However, because the negative-derivative feedback models operate by resisting changes in activity, without regard for the absolute level of activity, they can also maintain analog amplitudes of activity at a given location (Fig. 5*B*). Thus, these networks can convey information simultaneously about the amplitude and location of a spatial cue (for related examples in the context of optimal Bayesian cue combination and storage and efficient spike-based coding, see Boerlin and Denéve, 2011; Boerlin et al., 2013).

A related feature of the negative-derivative feedback networks is that they can temporally integrate their inputs. Temporal integration is the defining property of neural accumulators that integrate evidence over time (in the sense of calculus) during decision-making processes (Gold and Shadlen, 2007). However, most previous work modeling evidence accumulation has focused primarily upon temporal aspects of this facility, without considering that the accumulated evidence could occur across an analog range of spatial locations. A hallmark of feedback control theory is that the input–output transformation performed by systems with strong negative feedback is approximately equal to the inverse of the function that was fed back. In the case of the negative-derivative feedback networks, the signal that is negatively fed back is the derivative of the activity pattern. Thus, since the functional inverse of a temporal derivative is a temporal integral, these networks output a temporal integral of their inputs. For example, if the inputs are spatially structured, but constant in time, the negative-derivative feedback networks accumulate these signals into a uniformly increasing spatial pattern of activity (Fig. 5*C*,*D*). Thus, negative-derivative feedback networks can maintain in memory both the spatial identity of accumulated evidence as well as its running total.

Notably, even in the presence of nonlinearities in intrinsic neuronal dynamics such as thresholds and saturation, negative-derivative feedback networks accumulate and maintain spatially localized activity under the same conditions as in linear networks: a balance between positive and negative feedback, with slower positive feedback than negative feedback, leads to negative-derivative feedback. This occurs even though the Fourier components in a nonlinear network are no longer decoupled and cannot easily be decomposed into independent components (see Materials and Methods). Furthermore, the features of negative-derivative feedback discussed for linear dynamics are maintained under nonlinear dynamics, that is, the networks receive balanced excitation and inhibition during persistent activity (data not shown), and can accumulate and maintain spatially localized patterns of activity at different locations (Fig. 5*A*, bottom) or at different amplitudes (Fig. 5*B–D*, bottom; note that at $t = 5$ s, the neuron with preferred location $\theta = 0$ has approached its absolute maximum firing rate of 100 Hz, demonstrating that in this extreme case the profile does become significantly affected by the nonlinearity).

Maintaining multiple bumps of activity in negative-derivative feedback networks

In the previous sections, we considered networks receiving negative-derivative feedback only in the first cosine component and used this example to illustrate important features of the negative-derivative feedback mechanism–a close balance between excitation and inhibition during persistent activity (Fig. 4*D–F*) and the ability to encode information both in the location and in the amplitude of spatial patterns of activity (Fig. 5). While these features are hallmarks of negative-derivative feedback net-

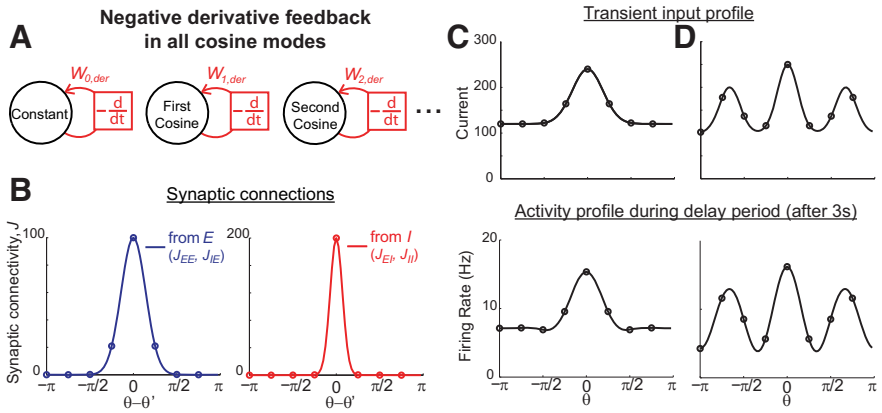

Figure 6. Networks receiving negative-derivative feedback in multiple Fourier components and maintenance of multiple bumps of activity. ***A***, ***B***, Example networks with negative-derivative feedback in multiple Fourier components (***A***). The synaptic connectivity profiles for this example were Gaussian shaped (***B***). ***C***, ***D***, The presence of negative-derivative feedback in higher order Fourier components permits the maintenance of narrowly tuned patterns of activity (***C***) and multiple bumps of activity (***D***). Data in ***C*** and ***D*** are shown at 3 s into the delay period.

works, the specific activity profile that is maintained during persistent activity is not constrained to simple sinusoids and ultimately is determined by which Fourier components receive negative-derivative feedback. Here, we consider more general networks that receive negative-derivative feedback in all Fourier components and show that such networks can be obtained by a condition analogous to the tuning condition used for the simple cosine example discussed above.

To construct more general networks receiving negative-derivative feedback (Fig. 6*A*), we consider networks with the same spatial profiles of the excitatory *E*-to-*E* and *E*-to-*I* connections (Fig. 6*B*, left) and the same spatial profiles of the inhibitory *I*-to-*E* and *I*-to-*I* connections (Fig. 6*B*, right), so that $J_{ij}(\theta) = \tilde{J}_{ij} w_j(\theta)$ for $i, j = E$ or I (note that this assumption leads to a simple form of the balance condition, but is not essential to tuning negative-derivative networks more generally). In this case, the Fourier components of the synaptic connectivity profiles are given by $\hat{J}_{ij}(n) = \tilde{J}_{ij}\hat{w}_j(n)$, and the condition for having a balance in strength of positive and negative feedback in a given Fourier mode is given by $\hat{J}_{EE}(n)\hat{J}_{II}(n) = \tilde{J}_{EE}\tilde{J}_{II}\hat{w}_E(n)\hat{w}_I(n) \sim \tilde{J}_{EI}\tilde{J}_{IE}\hat{w}_E(n)\hat{w}_I(n) = \hat{J}_{EI}(n)\hat{J}_{IE}(n)$, so that $\tilde{J}_{EE}\tilde{J}_{II} \sim \tilde{J}_{EI}\tilde{J}_{IE}$ for large values of $\tilde{J}_{ij}$ for all n. When, in addition, positive feedback is slower than negative feedback (due to a relatively slow combination of self-excitatory and self-inhibitory time constants $\tau_{EE} + \tau_{II} > \tau_{EI} + \tau_{EI}$), the network interactions provide negative-derivative feedback to all Fourier components.

Unlike the network of Figure 4, which only could maintain broad patterns of activity corresponding to its tuned, first cosine component (Fig. 6*C*), networks that receive negative-derivative feedback in multiple Fourier components can maintain spatially localized activity with narrower tuning widths that reflect higher order Fourier components. Furthermore, these networks can maintain more general spatial patterns of activity comprised of these different Fourier components, such as activity profiles with multiple bumps (Fig. 6*D*), which have been suggested as a neural correlate of the storage of multiple items (Laing et al., 2002; Edin et al., 2009; Wei et al., 2012). Thus, networks receiving negative-derivative feedback in multiple Fourier components have a higher memory capacity than those that receive negative-derivative feedback only in a single cosine component. Note, however, that the strength of negative-derivative feedback in each Fourier component, and thus the integration time constant associated with this component, in general will not be the same for all Fourier components, because this strength depends linearly upon the amount of the frequency component that is present within the synaptic connectivity profile. For this reason, the network capacity over a given timescale will in general depend both upon the specific form of the connectivity and the shape of the profile to be maintained so that, for example, networks with broad synaptic connectivity profiles would not be expected to maintain very long-lasting activity for high-frequency components that are minimally represented in their synaptic connectivity. This feature may explain why the long-lasting profiles observed experimentally during spatial working memory tend to be of relatively broad width that likely reflects features of the underlying connectivity profile.

Relation between the profile of synaptic connectivity and tuning widths of activity

Traditional spatial working memory networks require long-range inhibition to maintain the stability of localized patterns of activity in memory (Ermentrout, 1998; Wang, 2001; Compte, 2006). Such long-range inhibition is not prevalent anatomically in cortical networks, although it might be achieved functionally through disynaptic connections (Melchitzky et al., 2001) or through the broadly projecting basket cell subclass of inhibitory interneurons (Markram et al., 2004). In any case, an interesting question is whether long-range inhibition is critical for storing spatial working memory, and what constraints experimental observations may place upon the form of synaptic connectivity.

Unlike traditional models, negative-derivative feedback networks are capable of maintaining spatially localized patterns of activity regardless of the relative widths of excitatory and inhibitory connections (Fig. 7*A*, *C–E*). In fact, narrower inhibitory connections are required for our models to generate the experimental observation (Rao et al., 1999, 2000; Constantinidis and Goldman-Rakic, 2002) that inhibitory neurons have broader tuning of activity (after subtracting off any constant baseline) than excitatory neurons (Fig. 7*B*). When we define "widths" of the activity or connectivity as the spatial spread of the tuned portion after subtracting off any constant, untuned baseline (Constantinidis and Goldman-Rakic, 2002), short-range excitation and long-range inhibition lead to a spatially localized activity profile with the excitatory neurons having broader tuning of activity than the inhibitory neurons (Fig. 7*C*,*F*). On the other hand, the reverse relationship of the excitatory and inhibitory synaptic projections, that is, long-range excitation and short-range inhibition (Fig. 7*D*, or with the addition of nonselective inhibitory projections, Fig. 7*E*) lead to stable persistent activity with broader tuning of the inhibitory neurons than that of the excitatory neurons (Fig. 7*G*,*H*), as seen experimentally (Rao et al., 1999, 2000; Constantinidis and Goldman-Rakic, 2002). In all cases, neurons receive closely balanced excitation and inhibition and thus, the excitatory and inhibitory inputs show the same tuning widths (Fig. 7*I–K*). This balance of excitation and inhibition with the same spatial tuning is a general feature of negative-derivative feedback networks, since the large amount of excitation and in-

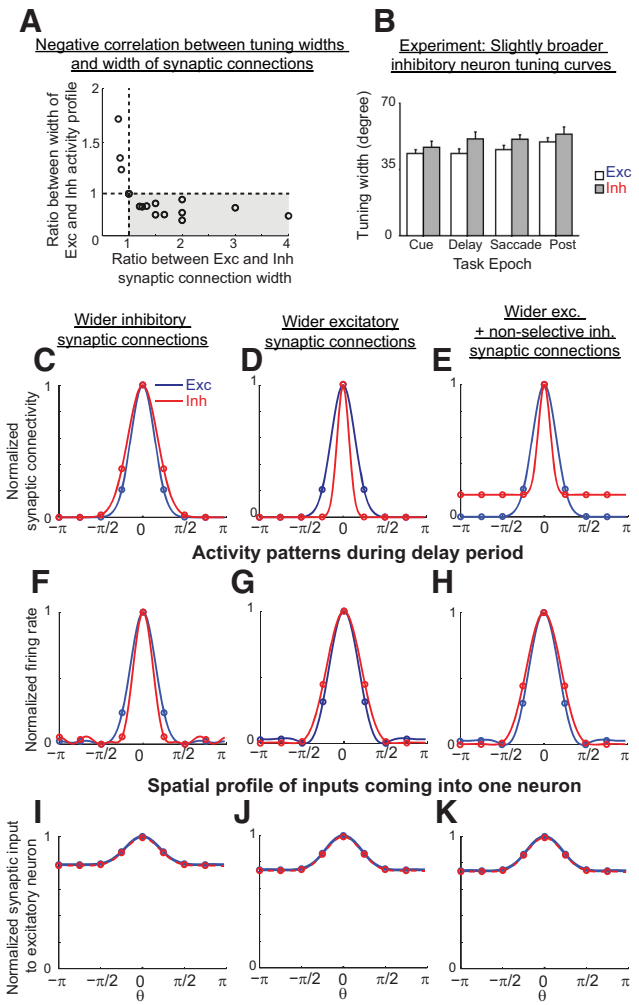

Figure 7. Maintenance of spatially localized, persistent activity with short-range inhibition and long-range excitation and relation between the profile of synaptic connectivity and the tuning widths of activity. ***A***, Negative correlation between tuning widths and width of synaptic connections in negative-derivative feedback networks. The spatial profiles of the excitatory and inhibitory connections in this example were Gaussian, and the widths of both the excitatory and inhibitory connections were varied. To obtain the tuning widths of the activity of the excitatory and inhibitory populations, the activity was fit by the sum of constant and Gaussian functions and the tuning width was defined as the width of the Gaussian function during the delay period as in Constantinidis and Goldman-Rakic, (2002). ***B***, Experimentally observed broader tuning of inhibitory neurons than of excitatory neurons during the different task epochs of a delayed saccade task. (Adapted from Constantinidis and Goldman-Rakic, 2002). ***C–E***, Spatial profile of excitatory (blue; same for the *E*-to-*E* and *E*-to-*I* connections) and inhibitory (red; same for the *I*-to-*E* and *I*-to-*I* connections) connectivity with broader inhibitory connections (***C***), broader excitatory connections (***D***), or broader tuned component of excitatory than inhibitory connections but with inhibitory connections additionally having a nonselective (constant) component (***E***). Spatial profiles of synaptic connectivity are normalized by their maximum height. ***F–H***, Activity patterns of excitatory (blue) and inhibitory (red) populations in response to a stimulus centered at $\theta = 0$, illustrated 3 s into the delay period. Firing rates were normalized for comparison by subtracting off the minimum activity and scaling the resulting firing rates to unit amplitude. A broader tuned component of the inhibitory connections (***C***) results in a narrower profile of the inhibitory population activity (***F***). A broader tuned component of excitatory connections, with (***E***) or without (***D***) an additional nonselective component of inhibitory projections, results in a narrower profile of sustained excitatory activity (***G***, ***H***). ***I–K***, Spatial profiles of balanced excitatory (blue) and inhibitory (red) synaptic inputs coming into an excitatory neuron. Inputs are normalized by the maximum of the excitatory inputs.

hibition required for strong derivative feedback must cancel to avoid saturation or total silencing of firing rates.

Thus, in the negative-derivative feedback networks, the relative tuning widths of the excitatory and inhibitory neurons are inversely correlated with the widths of the excitatory and inhibitory synaptic connections (Fig. 7*A*). This reciprocal relationship between the tuning widths of the neurons and the widths of synaptic projections is a consequence of the balance of excitatory and inhibitory inputs (Fig. 7*I–K*): because the tuning width of the total excitatory or inhibitory synaptic input onto a neuron is given by a convolution of the synaptic connectivity onto this neuron and the width of the presynaptic neurons' tuning curves, achieving balanced inhibitory and excitatory inputs requires that the experimentally observed broader inhibitory (compared with excitatory) tuning curves be offset by relatively narrower inhibitory synaptic connectivity profiles. This is different from most previous models for spatial working memory, which require broader negative feedback and show no reciprocal relationship between tuning widths of synaptic connectivity and activity profiles. Without different timescales for positive and negative feedback pathways (and, thus, without negative-derivative feedback), narrower negative feedback cannot sustain spatially localized activity (Fig. 8*A*,*E*,*I*,*M*; see Ermentrout and Cowan, 1980 for a mathematical proof). To stabilize spatially localized activity in such traditional lateral inhibitory models, broader negative feedback than positive feedback is required. This can be achieved either by long-range *E*-to-*I* connections (Fig. 8*B*,*F*) or long-range *I*-to-*E* synaptic connections (Fig. 8*C*,*D*,*G*,*H*). With broader negative feedback, the excitatory neurons receive broader inhibitory inputs than excitatory inputs (Fig. 8*N–P*). With no requirement of a close balance between excitation and inhibition, the reciprocal relationship between tuning widths and widths of synaptic projections is not observed in these previous models (Fig. 8*J–L*). Thus, this reciprocal relationship is a distinct feature of the negative-derivative feedback networks that highlights the mechanism underlying spatial working memory based on balanced excitatory and inhibitory inputs.

Robust memory performance against common perturbations to synaptic weights

A major challenge in short-term memory networks is stably maintaining analog memory representations in the face of perturbations. Although many types of memory networks, including the negative-derivative feedback networks, are quite robust against random noise in synaptic weights that largely can be averaged out across the network or random noise inputs that are filtered out by the slow network dynamics underlying persistent activity, resisting systematic perturbations in weights or intrinsic neuronal response properties has proven to be more challenging. An advantage of negative-derivative networks is that the balance condition that defines these networks is robust against many types of such naturally occurring perturbations. For example, global increases in the intrinsic gains of all neurons, which is equivalent to multiplicatively scaling the strengths of all synaptic connections, does not affect the balance of excitation and inhibition upon which negative-derivative feedback depends. As a result, such perturbations have minimal effect upon the ability of the network to maintain spatially localized persistent activity (Fig. 9*A*). Conceptually, this is because each neuronal population participates in both positive (through the *E*-to-*E* and, effectively, the *I*-to-*I* connections) and negative (through *E*-to-*I* and *I*-to-*E* connections) feedback loops so that such perturbations produce offsetting changes in positive and negative

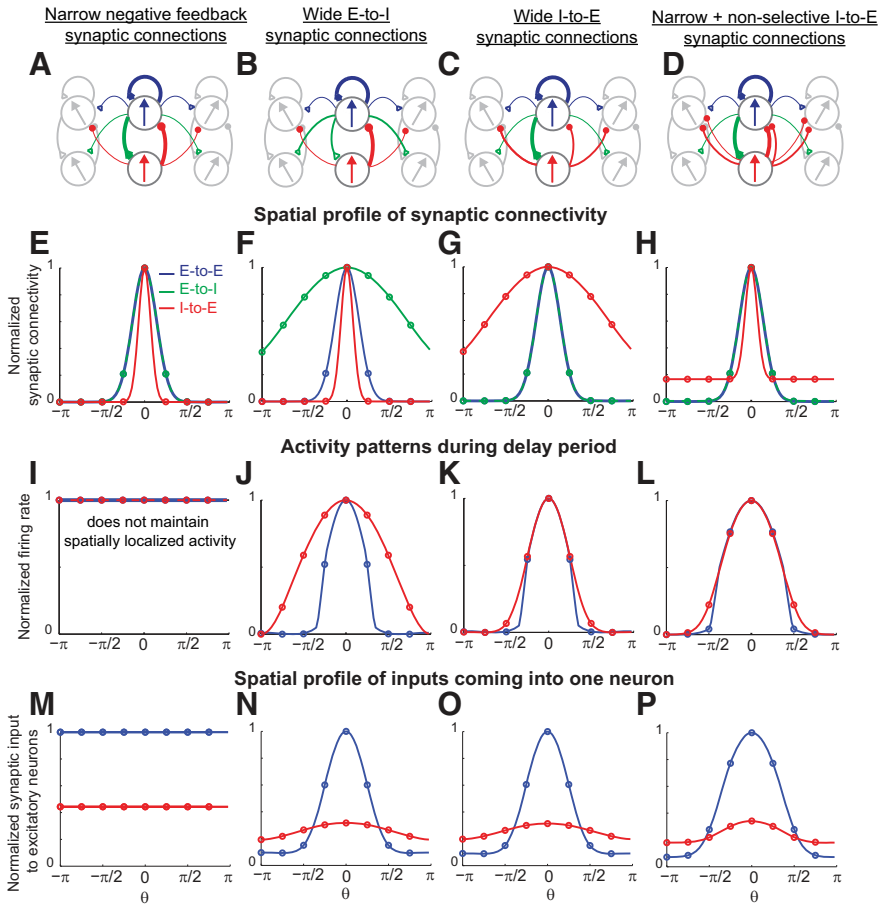

Figure 8. Comparison of activity patterns and synaptic inputs with spatial memory networks without negative-derivative feedback networks that have equal timescales for positive and negative feedback pathways. ***A–H***, Network structures and spatial profile of synaptic connectivity. For simplicity, the *E*-to-*E* (blue), *E*-to-*I* (green), and *I*-to-*E* (red) connections without the *I*-to-*I* connection are included and the spatial profile of synaptic connectivity is normalized by its maximum height (***E–H***). ***I–L***, Activity patterns of excitatory (blue) and inhibitory (red) populations in response to a stimulus centered at $\theta = 0$ (3 s into the delay period). When both the *E*-to-*I* and *I*-to-*E* connections are narrow, the network cannot sustain spatially localized activity, resulting in spatially uniform activity. Firing rates were normalized for comparison by subtracting off the minimum activity and scaling the resulting firing rates to unit amplitude except for the case showing spatially uniform patterns of activity (***I***). With broader negative feedback (***B–D***, ***F–H***), spatially localized activity can be stabilized, with broader tuning of the inhibitory population. ***M–P***, Spatial profiles of balanced excitatory (blue) and inhibitory (red) synaptic inputs coming into an excitatory neuron. Inputs are normalized by the maximum of the excitatory inputs. The networks receive broader inhibitory than excitatory inputs for spatially localized activity, resulting in an overall Mexican-hat (i.e., center-surround) type input (***N–P***).

feedback. Quantitatively, this result reflects that the balance condition for derivative-feedback networks is ratiometric, depending only upon the ratio of the synaptic strengths $\hat{J}_{EE}(n)\hat{J}_{II}(n)/\hat{J}_{EI}(n)\hat{J}_{IE}(n) \sim 1$ (see Eq. 21). Similarly, examination of this ratiometric condition shows that maintenance of persistent activity with negative-derivative feedback is also robust against global changes in the intrinsic gain of excitatory neurons alone (changes in $\hat{J}_{EE}(n)$ and $\hat{J}_{EI}(n)$; Fig. 9*B*) or inhibitory neurons alone (changes in $\hat{J}_{IE}(n)$ and $\hat{J}_{II}(n)$; Fig. 9*C*). Likewise, global changes in excitatory synaptic inputs (Figs. 9*E*, Fig. 10*A*,*B*; changes in $\hat{J}_{EE}(n)$ and $\hat{J}_{IE}(n)$), inhibitory synaptic inputs (Fig. 9*F*; changes in $\hat{J}_{EI}(n)$ and $\hat{J}_{II}(n)$) or all synaptic inputs (Fig. 9*D*) have minimal effect upon the maintenance of persistent activity, as does loss of a fraction of a subpopulation of neurons, which is equivalent to loss of a fraction of the corresponding excitatory or inhibitory synaptic inputs as in Figure 9*D–F*.

Furthermore, persistent neural activity in negative-derivative feedback networks is quite robust even against perturbations that occur locally in clusters of neurons with similar preferred spatial locations. To test how well the networks responded to local perturbations, we presented a transient input centered at a location $\theta = 0$ (Fig. 9*G–L*) and asked how well this item could be maintained in memory following a local perturbation that affected 1/8 of the network. When the perturbation was centered on the preferred location (possibly modeling, for example, effects of attention that changed the gains of neurons triggered by the stimulus), the amplitude of activity increased or decreased mildly for neuronal gain or synaptic weight increases or decreases, respectively, but the time course of persistent activity was only mildly affected (Fig. 10*C*,*D*), with the change in time constant approximately linearly related to the perturbation size (data not shown). When the perturbation was located on the flanks of the presented stimulus location (Fig. 9*G–L*; black bar along *x*-axis), activity was again maintained persistently in time (data not shown), although there was a small warping of the Gaussian-shaped bump that reflected that the perturbation disrupted the translation-invariant form of the network's structure. Thus, in this case, the perturbation would slightly bias the observation of the cue location if the readout of the network activity remained the same as before the perturbation. However, because the local perturbation does not affect the balance of positive and negative feedback that maintains persistent activity, the cue would remain in memory and, if the perturbation were continually present, a change in network readout could in principle learn to compensate for the changes in shape of the maintained activity profile.

The negative-derivative feedback networks are not robust against all forms of perturbations, in particular those that break the balance between excitation and inhibition that underlies the balance in strength of the positive and negative feedback components of negative-derivative feedback. For example, global or local perturbations in specific excitatory pathways, such as the *E*-to-*E* pathways that are dominated by NMDA-type synapses, do disrupt persistent activity (Fig. 10*E–H*). This is because NMDA-mediated currents are stronger at *E*-to-*E* than *E*-to-*I* connections (Wang et al., 2008; Wang and Gao, 2009; Rotaru et al., 2011); therefore their disruption imbalances the positive and negative feedback pathways, consistent with recent experimental observations of lack of robustness of working memory to pharmacological blockade of NMDA receptors (Wang et al., 2013). The disruption of persistent activity under such perturbations can be quantified by changes of the time constant of decay of activity at the perturbed location, τ_{eff}. As the ratio between the strengths of the positive and negative feedback J_{pos}/J_{neg} deviates from 1, τ_{eff} decreases inversely proportional to $1 - J_{pos}/J_{neg}$ (Fig. 10*I*,*J*). Thus, while

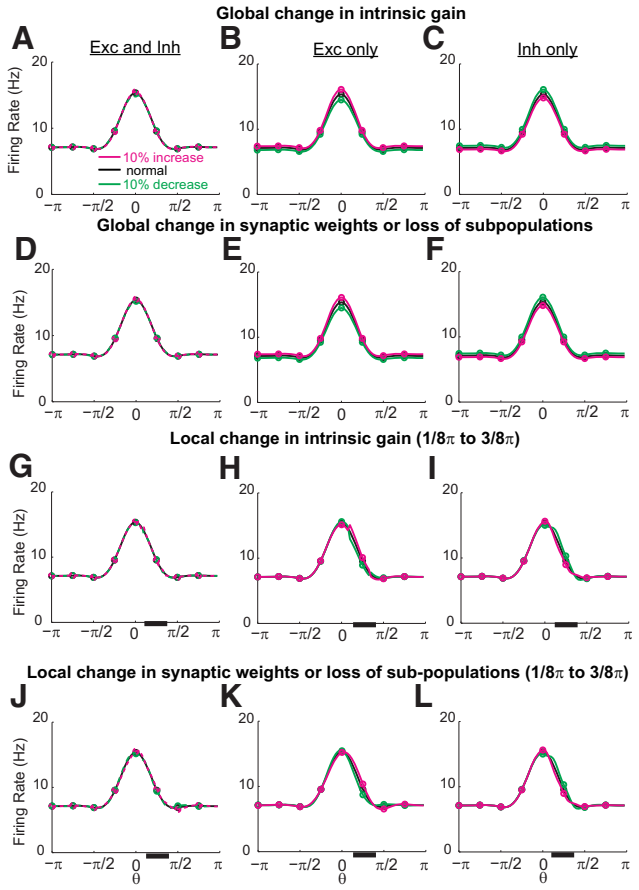

Figure 9. Robustness to common perturbations in memory networks with negative-derivative feedback. The networks can maintain spatially localized, persistent activity robustly against either global (***A–F***) or local (***G–L***) perturbations that preserve the balance between excitation and inhibition such as changes in intrinsic gain (***A–C***, ***G–I***) or changes in synaptic weights or loss of a fraction of the neurons in a given subpopulation (***D–F***, ***J–L***). The local perturbations were in neurons whose preferred directions lie between $1/8\pi$ and $3/8\ \pi$, represented by the black bar along the *x*-axis. Activity profiles are shown for 3 s into the delay period.

many common perturbations such as loss of neurons, changes in intrinsic neuronal gains, or uniform changes in synaptic strengths maintain J_{pos}/J_{neg} close to 1 (Fig. 10*I*, dashed diagonal line), the negative-derivative feedback networks are susceptible to perturbations that break the tuning of $J_{pos}/J_{neg} \sim 1$ (Fig. 10*I*, off-diagonal portions).

We note that the lack of robustness to perturbations that disrupt the excitatory–inhibitory balance in our model is different from the behavior observed in previous lateral inhibition models that require rough but not exact balance between excitation and inhibition and therefore exhibit robust memory performance across a wider range of perturbations in connectivity. For example, mild perturbations of the strength of the *E*-to-*E* connection alone or the *E*-to-*I* or *I*-to-*E* connections alone do not affect the memory performance of lateral inhibition models (Camperi and Wang, 1998; Hansel and Sompolinsky, 1998). However, in these models, the spatial patterns of activity can be maintained only at a fixed amplitude, rather than the graded range of amplitudes that can be sustained in models based upon derivative feedback. Thus, the more stringent tuning conditions on synaptic connections in the negative-derivative feedback networks reflects a trade-off between robustness to excitatory–inhibitory imbalance and being able to encode the amplitude of spatial patterns of activity and temporally integrate the strength of inputs.

Irregular firing activity during persistent activity

A characteristic feature of persistent neural activity during spatial working memory tasks is the irregular, Poisson-like nature of the firing activity (Compte et al., 2003). This has been a challenge for most previous spatial working memory models because, in these models, elevated persistent activity is maintained by a constant, suprathreshold excitatory drive that causes relatively regular persistent firing unless large external sources of noise are included (Barbieri and Brunel, 2008; but see Barbieri and Brunel, 2007; Renart et al., 2007; Roudi and Latham, 2007; Lundqvist et al., 2010; Boerlin and Denéve, 2011; Mongillo et al., 2012; Boerlin et al., 2013; Hansel and Mato, 2013). In contrast, negative-derivative feedback networks operate in a regime of closely balanced excitation and inhibition, so that the mean synaptic input is subthreshold and firing is driven largely by fluctuations that lead to a high coefficient of variation of the interspike intervals (Shadlen and Newsome, 1994; van Vreeswijk and Sompolinsky, 1996; Amit and Brunel, 1997; Troyer and Miller, 1997; Renart et al., 2007; Roudi and Latham, 2007). This spike-train irregularity was shown previously in spatially uniform negative-derivative feedback models (Lim and Goldman, 2013). Here, we show that the same result occurs in negative-derivative feedback networks with spatial structure.

We constructed spiking network models with the same columnar structure as in the firing rate models (Fig. 1). Each column consisted of excitatory and inhibitory integrate-and-fire neurons with similar preferred spatial features, and the connectivity between neurons within and across the columns was random and sparse (van Vreeswijk and Sompolinsky, 1996). For connected neurons, the strength of synaptic connections was assumed to be a Gaussian function of the difference between the preferred features of the presynaptic and postsynaptic neurons (Fig. 6*B*), and the strengths of the excitatory and inhibitory connections on average were set to satisfy the balance condition of Equation 21. However, we note that the connection strengths onto individual neurons were not precisely balanced and were heterogeneous due to the sparse and random connectivity of the network. Inhibitory currents were mediated by $GABA_A$ receptors and recurrent excitatory currents were mediated by a mixture of AMPA and NMDA receptors, with a greater proportion of and slower kinetics of NMDA receptors in the excitatory feedback pathways (Wang et al., 2008; Wang and Gao, 2009; Rotaru et al., 2011). The networks receive spatially patterned input during the stimulus presentation, but no input during the delay period.

As in the firing rate models, these spiking networks implementing negative-derivative feedback showed spatially tuned persistent activity encoding the cue location of the transiently presented stimulus (Fig. 11*B*). Due to the balance between excitation and inhibition, the neuronal spike trains were highly irregular during the delay period (Fig. 11*C*,*D*,*F*,*G*). Quantitative analysis of the spike-train irregularity using the local coefficient of variation CV_2 (see Materials and Methods) found that the model distributions were similar to those observed experimentally in memory cells receiving preferred cue or nonpreferred cue stimuli (Fig. 11*A*, experiments; *E*,*H*, model). Thus, the principle of negative-derivative feedback also is applicable to spiking networks and networks incorporating this principle can reproduce salient properties of biological working memory networks such as spatially tuned delay period activity and irregular firing.

Discussion

Here we suggest a new model for spatial working memory based on negative-derivative feedback control. When recurrent inhibi-

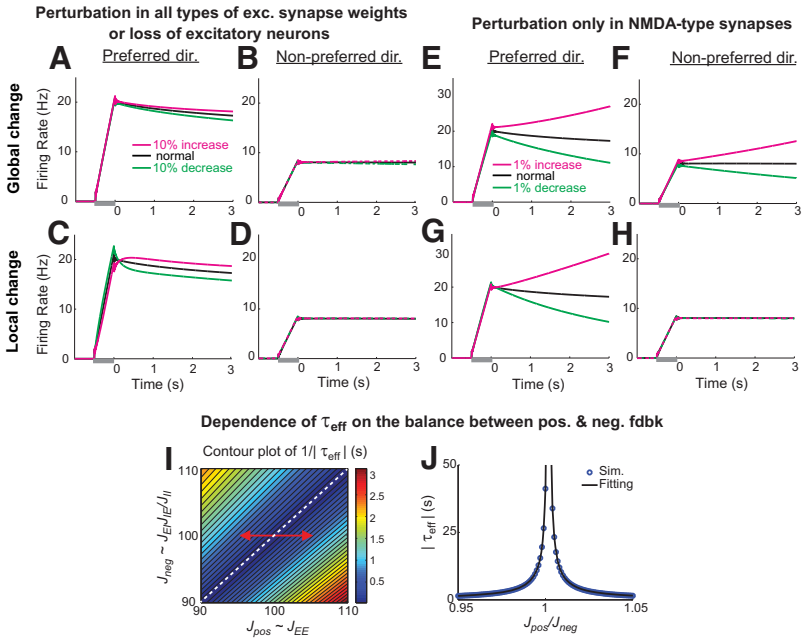

Figure 10. Memory performance following perturbations under which the *E-I* balance is maintained (***A–D***) or disrupted (***E–H***). ***A–D***, Time course of activity under global (***A***, ***B***) or local (***C***, ***D***) perturbations in all excitatory synaptic weights, or equivalently under loss of a corresponding fraction of the excitatory subpopulation. Since loss of a fraction of the cells or of all excitatory synapses affects both the positive and negative feedback pathways equally, the balance between positive and negative feedback is maintained and persistent firing is minimally affected. The time course of activity is shown for an example neuron receiving an external input stimulus centered on its preferred direction and for an example neuron receiving external input in its nonpreferred directions. ***E–H***, Disruption of persistent firing under global (***E***, ***F***) or local (***G***, ***H***) perturbations in the *E*-to-*E* connections, mimicking perturbations in slow NR2B subunit-containing NMDA receptors located predominantly in this connection (Wang et al., 2013). Global perturbations led to similar changes in activity throughout the entire network (***E***, ***F***), while local perturbations disrupted activity most severely in the perturbed neurons (***G***, ***H***). ***I***, ***J***, Time constant of activity decay τ_{eff} in a neuron receiving an external input stimulus centered on its preferred direction as a function of the strengths of positive and negative feedback pathways. τ_{eff} is estimated by fitting the time course of delay activity between 500 and 2000 ms with an exponential function. As the overall strengths J_{pos} and J_{neg} of the positive and negative feedback pathways change, τ_{eff} decays approximately in inverse proportion to $1 - J_{pos}/J_{neg}$ (linear change in $1/|\tau_{eff}|$ in ***I***). Local perturbations (data not shown) provided similar results. In ***J***, J_{neg} was fixed and only J_{pos} changed, corresponding to the red arrow in ***I***. The black fit curve is given by $0.075 \text{ sec}/|1 - J_{pos}/J_{neg}| \sim (\tau_{pos} - \tau_{neg})/(1 - J_{pos}/J_{neg})$.

tion and excitation are balanced, and the feedback pathways mediating positive feedback are slower than those mediating negative feedback, we have shown how a network with functionally columnar architecture can maintain analog amplitude signals corresponding to any spatial location. Furthermore, we have demonstrated that these negative-derivative feedback networks can temporally integrate their inputs, thus showing how accumulation of sensory input can be performed in a spatially specific manner. Given that recent experiments in frontal cortex suggest a balance of inhibition and excitation (Shu et al., 2003; Haider et al., 2006) as well as differential kinetics in the *E*-to-*E* versus *E*-to-*I* pathways mediating positive versus negative feedback (Wang et al., 2008; Wang and Gao, 2009; Rotaru et al., 2011), this suggests that negative-derivative feedback may serve as a fundamental principle underlying the accumulation and storage of signals in spatial working memory.

Comparison to network models with lateral inhibition and experimental predictions

A "Mexican-hat" network architecture with a broader range of inhibitory interactions than excitatory interactions between neurons is prevalent in cortical circuit models that generate spatial patterns of activity for working memory (Ermentrout, 1998; Wang, 2001; Compte, 2006). Compared with most previous network models with functionally long-range inhibition, negative-derivative feedback networks exhibit several distinct features. First, negative-derivative feedback networks do not require long-range inhibition. Second, they receive massive amounts of excitatory and inhibitory inputs that are closely balanced. Third, this close balance between excitation and inhibition leads to irregular firing activity during the delay period, consistent with experimental observations in cortical memory circuits (Compte et al., 2003). Fourth, negative-derivative feedback networks can encode information about a transient stimulus not only in the location but also in the amplitude of its spatial patterns of activity and, in principle, can maintain arbitrary spatial patterns of activity (see also Carroll et al., 2014 for a special network with fast inhibitory neurons and tuning of both the form of neuronal response nonlinearity and connectivity to allow graded-amplitude spatial patterns of activity to be maintained with long-range inhibition).

The negative-derivative feedback networks require a tight tuning condition on network connectivity to have positive and negative feedback of equal strengths. This tuning condition is more stringent than that of the previous lateral inhibition models, which require only rough balance between the strengths of positive and negative feedback (Camperi and Wang, 1998; Hansel and Sompolinsky, 1998). However, in such systems, spatial patterns of activity can be maintained only at a fixed amplitude and thus, the strictness of the tuning condition in derivative feedback models can be considered as a trade-off with being able to maintain activity across a graded range of levels and to temporally integrate inputs. Somewhat mitigating the strictness of this tuning condition is the fact that it applies only to the average connectivity of the different populations and does not require that the tuning be exact for each individual neuron. As shown in Figure 9, this tuning condition may be preserved following many natural perturbations, such as changes in intrinsic or synaptic gains or loss of subpopulations of neurons that occur globally or locally in the circuits. On a slower timescale, recent work suggests that the excitatory–inhibitory balance in cortical cells may be actively maintained by homeostatic mechanisms (Liu, 2004; Vogels et al., 2011), or may be achieved gradually through the developmental refinement of synaptic connections (Tao and Poo, 2005). Thus, the balance of inhibition and excitation required for derivative-feedback memory networks may be quite robust in normal situations.

The distinct features of our model provide testable predictions. First, negative-derivative feedback networks predict similar spatial tuning between excitatory and inhibitory inputs due to a close balance between them (Figs. 4*D*, 7*I–K*). Experimentally, a balance between strong excitatory and inhibitory synaptic inputs (Shu et al., 2003; Haider et al., 2006) and co-tuning between them (Destexhe et al., 2003; Wehr and Zador, 2003; Priebe and Ferster,

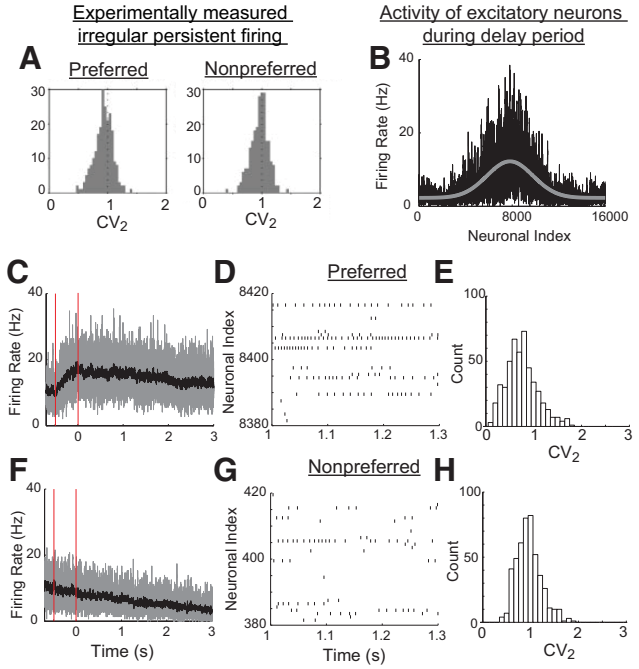

Figure 11. Irregular firing in spiking networks. ***A***, Experimentally measured irregular firing, as captured by the local coefficient of variation CV_2 during persistent activity in a delayed-saccade tasks. (Adapted from Compte et al., 2003). ***B–H***, Simulation of spiking network models with negative-derivative feedback. ***B***, Activity of excitatory neurons during the delay period in response to a transient, Gaussian-shaped stimulus. The firing rate distribution shown represents averages of neuronal activity between 2500 and 3000 ms after the stimulus offset (black curve) and is fit with a sum of constant and Gaussian functions (gray curve). ***C–H***, Activity of neurons receiving a stimulus in their preferred directions (***C–E***, computed for neurons 7601–8400 in ***B***) or neurons receiving stimuli in their nonpreferred directions (***F–H***, computed for neurons 1–400 and 15601–16000). ***C***, ***F***, Instantaneous, population-averaged activity, computed within time bins of 1 ms (gray) or 10 ms (black). Vertical red lines indicate the times of the start and end of the stimulus. ***D***, ***G***, Raster plots illustrating the irregular persistent firing of 40 example neurons. ***E***, ***H***, Histogram of CV_2 values of active neurons during the persistent firing. Note that, for the activity of neurons receiving preferred stimuli, a small set of neurons fired regularly at high firing rate and exhibited low CV_2 values. This is due to the heterogeneity of our completely randomly connected networks, resulting in excess positive feedback in some clusters of neurons.

2005; Rudolph et al., 2007) have been observed in cortical neurons, although the ultimate test, intracellular recordings of memory cells in a behaving animal, has yet to be performed and currently stands as a prediction of our model. Second, perturbations in specific synaptic pathways–such as blockade of excitatory or inhibitory transmission exclusively onto excitatory or inhibitory neurons that break the balance between excitation and inhibition–would cause more severe impairments of persistent activity than completely silencing a subset of excitatory or inhibitory neurons. Consistent with this prediction, a recent experiment blocking slow NMDA-mediated currents that are especially prominent in the pyramidal-to-pyramidal (*E*-to-*E*) connections of prefrontal cortex (Wang et al., 2013) did lead to strong impairments in working memory performance. Third, the balance condition between excitation and inhibition provides negative correlation between the relative tuning widths of the excitatory and inhibitory neurons and the widths of the excitatory and inhibitory synaptic connections, such that if the inhibitory synaptic projection is shorter range, the balance in inputs is preserved under broader tuning of the inhibitory neurons. Finally, we note that it is possible that different mechanisms are used in networks that maintain graded representations from those that maintain only spatial location information, so experiments designed to test these predictions ideally should be performed using paradigms that require both the spatial location and amplitude (or duration for integrators) of stimuli to be encoded.

Irregular firing statistics based on balanced excitation and inhibition

The irregular firing activity observed during working memory performance (Compte et al., 2003) provides indirect support for a balance between excitation and inhibition in the neurons supporting this activity. This balance has been a challenge to achieve in most models of working memory, because these models depend upon stronger excitation than inhibition to maintain elevated firing rates, and such imbalanced excitation tends to lead to regular patterns of neuronal firing (Barbieri and Brunel, 2008). In contrast, negative-derivative feedback networks inherently depend upon a balance of inhibition and excitation throughout an analog range of firing rates, leading to irregular firing at all rates.

To account for irregular spiking activity during a delay period, recent works have suggested bistable memory circuits based on balanced excitation and inhibition. These balanced networks can maintain elevated (UP) states through neuronal nonlinearities (Barbieri and Brunel, 2007; Renart et al., 2007; Roudi and Latham, 2007; Lundqvist et al., 2010) or through synaptic nonlinearities associated with short-term synaptic plasticity (Mongillo et al., 2012; Hansel and Mato, 2013). However, with the exception of Hansel and Mato (2013), these networks used identical time constants for positive and negative feedback pathways so that they do not contain negative-derivative feedback, are not able to maintain a graded range of persistent activity and perform temporal integration, and can be distinguished from the negative-derivative feedback networks by their essential dependence upon lateral inhibition to stabilize spatially localized persistent activity (Fig. 8).

Like our networks, the spatial working memory networks of Hansel and Mato (2013) show similar tuning of excitatory and inhibitory inputs and contain an asymmetric ratio of NMDA/AMPA currents in the *E*-to-*E* versus *E*-to-*I* connections, resulting in slower positive than negative feedback. Thus, these networks also may contain a derivative-feedback signal that contributes to their robustness, and an interesting question is whether the principle of negative-derivative feedback could be useful in bistable, as well as analog, spatial memory networks. In separate work, networks built upon the principle of optimal inference of external inputs and efficient spike-based coding can maintain analog-valued amplitudes of irregular persistent activity (Boerlin and Denéve, 2011; Boerlin et al., 2013). These networks require balanced excitation and inhibition with slower excitation, and thus likely also depend upon a large negative-derivative feedback component, suggesting that the principle of negative-derivative feedback control may be derived independently from the theory of Bayesian inference and spike-based coding.

Memory capacity of negative-derivative networks

A distinctive feature of the negative-derivative feedback networks is that they can maintain spatially localized activity of different amplitudes as well as at different locations (Fig. 5). This could be useful in modulating network response as a function of attention (Reynolds and Chelazzi, 2004) or reward (Schultz et al., 1993; Watanabe, 1996; Leon and Shadlen, 1999; Amemori and Sawaguchi, 2006). Alternatively, simultaneously being able to vary amplitude and location could be useful in encoding quantities such as the color of a patch that can vary in an analog manner in

both spatial and nonspatial dimensions (Luck and Vogel, 1997; Zhang and Luck, 2008), or in accumulating the value of a single quantity over time (Gold and Shadlen, 2007) in a spatially specific manner. However, we note that the ability to integrate external inputs makes negative-derivative feedback networks relatively more sensitive to noisy or interfering input present during memory performance. To enhance the signal-to-noise ratio of negative-derivative feedback networks, additional mechanisms may be required to suppress external inputs during memory performance, for example, by dopamine regulation that is triggered with the onset of task-related input (Sawaguchi et al., 1988; Durstewitz et al., 1999).

Negative-derivative feedback networks also can maintain activities with multiple bumps when negative-derivative feedback is present in higher order Fourier components. Previous studies have suggested that the width of recurrent excitatory connections is a critical factor determining the maximal number of items that can be stored in the network (Edin et al., 2009; Wei et al., 2012). On the other hand, the memory capacity of negative-derivative feedback networks is determined by the amount of negative-derivative feedback in higher order Fourier components, and thus the width of both the excitatory and inhibitory connections affects memory capacity. Since storing more items requires the maintenance of narrower, higher frequency-containing patterns, this may provide a fundamental constraint on the forms of synaptic connectivity in memory networks. Further work is needed to explore the relationship between memory capacity and connectivity structure, and to compare the performance of negative-derivative feedback networks with that of previous network models.

References

Amemori K, Sawaguchi T (2006) Contrasting effects of reward expectation on sensory and motor memories in primate prefrontal neurons. Cereb Cortex 16:1002–1015. Medline

Amit DJ, Brunel N (1997) Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. Cereb Cortex 7:237–252. CrossRef Medline

Barbieri F, Brunel N (2007) Irregular persistent activity induced by synaptic excitatory feedback. Front Comput Neurosci 1:5. CrossRef Medline

Barbieri F, Brunel N (2008) Can attractor network models account for the statistics of firing during persistent activity in prefrontal cortex? Front Neurosci 2:114–122. CrossRef Medline

Ben-Yishai R, Bar-Or RL, Sompolinsky H (1995) Theory of orientation tuning in visual cortex. Proc Natl Acad Sci U S A 92:3844–3848. CrossRef Medline

Boerlin M, Denève S (2011) Spike-based population coding and working memory. PLoS Comput Biol 7:e1001080. CrossRef Medline

Boerlin M, Machens CK, Denève S (2013) Predictive coding of dynamical variables in balanced spiking networks. PLoS Comput Biol 9:e1003258. CrossRef Medline

Braitenberg V, Schüz A (1998) Cortex: statistics and geometry of neuronal connectivity, Ed 2. New York: Springer.

Camperi M, Wang XJ (1998) A model of visuospatial working memory in prefrontal cortex: recurrent network and cellular bistability. J Comput Neurosci 5:383–405. CrossRef Medline

Carroll S, Josic K, Kilpatrick ZP (2013) Encoding certainty in bump attractors. J Comput Neurosci. Advance online publication. doi:10.1007/s10827-013-0486-0. CrossRef Medline

Chafee MV, Goldman-Rakic PS (1998) Matching patterns of activity in primate prefrontal area 8a and parietal area 7ip neurons during a spatial working memory task. J Neurophysiol 79:2919–2940. Medline

Compte A (2006) Computational and in vitro studies of persistent activity: edging towards cellular and synaptic mechanisms of working memory. Neuroscience 139:135–151. CrossRef Medline

Compte A, Constantinidis C, Tegner J, Raghavachari S, Chafee MV, Goldman-Rakic PS, Wang XJ (2003) Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task. J Neurophysiol 90:3441–3454. CrossRef Medline

Constantinidis C, Goldman-Rakic PS (2002) Correlated discharges among putative pyramidal neurons and interneurons in the primate prefrontal cortex. J Neurophysiol 88:3487–3497. CrossRef Medline

Constantinidis C, Steinmetz MA (1996) Neuronal activity in posterior parietal area 7a during the delay periods of a spatial memory task. J Neurophysiol 76:1352–1355. Medline

Destexhe A, Rudolph M, Paré D (2003) The high-conductance state of neocortical neurons in vivo. Nat Rev Neurosci 4:739–751. CrossRef Medline

Douglas RJ, Martin KA (2004) Neuronal circuits of the neocortex. Annu Rev Neurosci 27:419–451. CrossRef Medline

Durstewitz D, Kelc M, Güntürkün O (1999) A neurocomputational theory of the dopaminergic modulation of working memory functions. J Neurosci 19:2807–2822. Medline

Edin F, Klingberg T, Johansson P, McNab F, Tegnér J, Compte A (2009) Mechanism for top-down control of working memory capacity. Proc Natl Acad Sci U S A 106:6802–6807. CrossRef Medline

Ermentrout B (1998) Neural networks as spatio-temporal pattern-forming systems. Rep Prog Phys 61:353–430. CrossRef

Ermentrout B, Cowan JD (1980) Large scale spatially organized activity in neural nets. SIAM J Appl Math 38:1–21. CrossRef

Folland GB (2009) Fourier analysis and its applications. Providence, RI: American Mathematical Society.

Funahashi S, Bruce CJ, Goldman-Rakic PS (1989) Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. J Neurophysiol 61:331–349. Medline

Gold JI, Shadlen MN (2007) The neural basis of decision making. Annu Rev Neurosci 30:535–574. CrossRef Medline

Goldman-Rakic PS (1995) Cellular basis of working memory. Neuron 14:477–485. CrossRef Medline

Haider B, McCormick DA (2009) Rapid neocortical dynamics: cellular and network mechanisms. Neuron 62:171–189. CrossRef Medline

Haider B, Duque A, Hasenstaub AR, McCormick DA (2006) Neocortical network activity in vivo is generated through a dynamic balance of excitation and inhibition. J Neurosci 26:4535–4545. CrossRef Medline

Hansel D, Mato G (2013) Short-term plasticity explains irregular persistent activity in working memory tasks. J Neurosci 33:133–149. CrossRef Medline

Hansel D, Sompolinsky H (1998) Modeling feature selectivity in local cortical circuits. In: Methods in neuronal modeling: from synapses to networks, Ed 2 (Koch C, Segev I, eds). Cambridge, MA: MIT.

Hansel D, Mato G, Meunier C, Neltner L (1998) On numerical simulations of integrate-and-fire neural networks. Neural Comput 10:467–483. CrossRef Medline

Holt GR, Softky WR, Koch C, Douglas RJ (1996) Comparison of discharge variability in vitro and in vivo in cat visual cortex neurons. J Neurophysiol 75:1806–1814. Medline

Laing CR, Troy WC, Gutkin B, Ermentrout GB (2002) Multiple bumps in a neuronal model of working memory. SIAM J Appl Math 63:62–97. CrossRef

Leon MI, Shadlen MN (1999) Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. Neuron 24:415–425. CrossRef Medline

Lim S, Goldman MS (2013) Balanced cortical microcircuitry for maintaining information in working memory. Nat Neurosci 16:1306–1314. CrossRef Medline

Liu G (2004) Local structural balance and functional interaction of excitatory and inhibitory synapses in hippocampal dendrites. Nat Neurosci 7:373–379. CrossRef Medline

Luck SJ, Vogel EK (1997) The capacity of visual working memory for features and conjunctions. Nature 390:279–281. CrossRef Medline

Lundqvist M, Compte A, Lansner A (2010) Bistable, irregular firing and population oscillations in a modular attractor memory network. PLoS Comput Biol 6:e1000803. CrossRef Medline

Markram H, Toledo-Rodriguez M, Wang Y, Gupta A, Silberberg G, Wu C (2004) Interneurons of the neocortical inhibitory system. Nat Rev Neurosci 5:793–807. CrossRef Medline

McCormick DA, Connors BW, Lighthall JW, Prince DA (1985) Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex. J Neurophysiol 54:782–806. Medline

Melchitzky DS, González-Burgos G, Barrionuevo G, Lewis DA (2001) Syn-

aptic targets of the intrinsic axon collaterals of supragranular pyramidal neurons in monkey prefrontal cortex. J Comp Neurol 430:209–221. CrossRef Medline

Mongillo G, Hansel D, van Vreeswijk C (2012) Bistability and spatiotemporal irregularity in neuronal networks with nonlinear synaptic transmission. Phys Rev Lett 108:158101. CrossRef Medline

Nise NS (2004) Control systems engineering, Ed 4. New York: Wiley.

Priebe NJ, Ferster D (2005) Direction selectivity of excitation and inhibition in simple cells of the cat primary visual cortex. Neuron 45:133–145. CrossRef Medline

Rao SG, Williams GV, Goldman-Rakic PS (1999) Isodirectional tuning of adjacent interneurons and pyramidal cells during working memory: evidence for microcolumnar organization in PFC. J Neurophysiol 81:1903–1916. Medline

Rao SG, Williams GV, Goldman-Rakic PS (2000) Destruction and creation of spatial tuning by disinhibition: GABA(A) blockade of prefrontal cortical neurons engaged by working memory. J Neurosci 20:485–494. Medline

Renart A, Moreno-Bote R, Wang XJ, Parga N (2007) Mean-driven and fluctuation-driven persistent activity in recurrent networks. Neural Comput 19:1–46. CrossRef Medline

Reynolds JH, Chelazzi L (2004) Attentional modulation of visual processing. Annu Rev Neurosci 27:611–647. CrossRef Medline

Rotaru DC, Yoshino H, Lewis DA, Ermentrout GB, Gonzalez-Burgos G (2011) Glutamate receptor subtypes mediating synaptic activation of prefrontal cortex neurons: relevance for schizophrenia. J Neurosci 31:142–156. CrossRef Medline

Roudi Y, Latham PE (2007) A balanced memory network. PLoS Comput Biol 3:1679–1700. Medline

Rudolph M, Pospischil M, Timofeev I, Destexhe A (2007) Inhibition determines membrane potential dynamics and controls action potential generation in awake and sleeping cat cortex. J Neurosci 27:5280–5290. CrossRef Medline

Salin PA, Prince DA (1996) Spontaneous GABAA receptor-mediated inhibitory currents in adult rat somatosensory cortex. J Neurophysiol 75:1573–1588. Medline

Sawaguchi T, Matsumura M, Kubota K (1988) Dopamine enhances the neuronal activity of spatial short-term memory task in the primate prefrontal cortex. Neurosci Res 5:465–473. CrossRef Medline

Schultz W, Apicella P, Ljungberg T (1993) Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. J Neurosci 13:900–913. Medline

Shadlen MN, Newsome WT (1994) Noise, neural codes and cortical organization. Curr Opin Neurobiol 4:569–579. CrossRef Medline

Shu Y, Hasenstaub A, McCormick DA (2003) Turning on and off recurrent balanced cortical activity. Nature 423:288–293. CrossRef Medline

Tao HW, Poo MM (2005) Activity-dependent matching of excitatory and inhibitory inputs during refinement of visual receptive fields. Neuron 45:829–836. CrossRef Medline

Troyer TW, Miller KD (1997) Physiological gain leads to high ISI variability in a simple model of a cortical regular spiking cell. Neural Comput 9:971–983. CrossRef Medline

van Vreeswijk C, Sompolinsky H (1996) Chaos in neuronal networks with balanced excitatory and inhibitory activity. Science 274:1724–1726. CrossRef Medline

van Vreeswijk C, Sompolinsky H (1998) Chaotic balanced state in a model of cortical circuits. Neural Comput 10:1321–1371. CrossRef Medline

Vogels TP, Sprekeler H, Zenke F, Clopath C, Gerstner W (2011) Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. Science 334:1569–1573. CrossRef Medline

Wang HX, Gao WJ (2009) Cell type-specific development of NMDA receptors in the interneurons of rat prefrontal cortex. Neuropsychopharmacology 34:2028–2040. CrossRef Medline

Wang H, Stradtman GG 3rd, Wang XJ, Gao WJ (2008) A specialized NMDA receptor function in layer 5 recurrent microcircuitry of the adult rat prefrontal cortex. Proc Natl Acad Sci U S A 105:16791–16796. CrossRef Medline

Wang M, Yang Y, Wang CJ, Gamo NJ, Jin LE, Mazer JA, Morrison JH, Wang XJ, Arnsten AF (2013) NMDA receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex. Neuron 77:736–749. CrossRef Medline

Wang XJ (2001) Synaptic reverberation underlying mnemonic persistent activity. Trends Neurosci 24:455–463. CrossRef Medline

Watanabe M (1996) Reward expectancy in primate prefrontal neurons. Nature 382:629–632. CrossRef Medline

Wehr M, Zador AM (2003) Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. Nature 426:442–446. CrossRef Medline

Wei Z, Wang XJ, Wang DH (2012) From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. J Neurosci 32:11228–11240. CrossRef Medline

Wilson HR (1999) Spikes, decisions, and actions. New York: Oxford UP.

Wimmer K, Nykamp DQ, Constantinidis C, Compte A (2014) Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. Nat Neurosci 17:431–439. CrossRef Medline

Xiang Z, Huguenard JR, Prince DA (1998) GABAA receptor-mediated currents in interneurons and pyramidal cells of rat visual cortex. J Physiol 506:715–730. CrossRef Medline

Zhang W, Luck SJ (2008) Discrete fixed-resolution representations in visual working memory. Nature 453:233–235. CrossRef Medline