Supplemental Materials for "Balanced amplification: a new mechanism of selective amplification of neural activity patterns"

Brendan K. Murphy and Kenneth D. Miller

November 26, 2015

Note: This is a corrected version of the original Supplement, which was dated Feb. 18, 2009. The changes made are indicated in the pages at the end of this document (following p. 45), where original versions are shown as red (and crossed out) and new versions underlined with squiggly line.

Contents

List of Figures

S1 Mathematical Solutions and Analysis of the Models Studied in the Main Text

S1.1 Two-Population Model

S1.1.1 Solutions

We begin with Eqs. 2 and 3 from the main text, with the addition of time-independent inputs $I_\pm = I_E \pm I_I$:

$$\tau \frac{dr_+}{dt} = -(1+w_+)r_+ + w_{FF}r_- + I_+ \tag{S1}$$

$$\tau \frac{dr_-}{dt} = -r_- + I_- \tag{S2}$$

Recall that $w_+ = w(k_I - 1)$ and $w_{FF} = w(k_I + 1)$.

To express the solutions, it is helpful to define the following pulse function:

$$g_{w_+}(t) = \frac{e^{-\frac{t}{\tau}} - e^{-(1+w_+)\frac{t}{\tau}}}{w_+} \tag{S3}$$

We will see that $w_{FF}g_{w_+}(t)$ represents the characteristic response in r_+ to input to r_-: an initial condition of $r_-(0)$ produces a response in r_+ of $r_-(0)w_{FF}g_{w_+}(t)$, while input to r_- is filtered by (convolved with) $\frac{1}{\tau}w_{FF}g_{w_+}(t)$ to produce response in r_+. $g_{w_+}(t)$, as a difference of exponentials, is a pulse response: it is 0 at $t = 0$, goes to 0 as $t \to \infty$, and peaks in between. It peaks at $t_{\text{peak}} = \tau\frac{\log(1+w_+)}{w_+}$, which decreases monotonically with increasing w_+ from τ for $w_+ \to 0$ (which represents perfectly balanced excitation and inhibition, $k_I \to 1$) to 0 for $w_+ \to \infty$. The value at the peak is $\left(\frac{1}{1+w_+}\right)^{\frac{w_++1}{w_+}}$ which decreases monotonically from 1 for $w_+ = 0$ to $1/w_+$ for $w_+ \to \infty$. Thus, the peak of $g_{w_+}(t)$ becomes smaller and occurs earlier as the eigenvalue associated with the sum mode becomes increasingly negative. After the peak, the decay to zero occurs essentially with timecourse $e^{-\frac{t}{\tau}}$. For $w_+ \to 0$, $g_{w_+}(t) \to \frac{t}{\tau}e^{-\frac{t}{\tau}}$, while as $w_+ \to \infty$, $g_{w_+}(t) \to e^{-\frac{t}{\tau}}/w_+$. We can think of $g_{w_+}(t)$ as interpolating between $\frac{t}{\tau}e^{-\frac{t}{\tau}}$ and $e^{-\frac{t}{\tau}}/w_+$ with increasing w_+.

The amplification in r_+ of the response to a steady-state input to r_- is equal to the integral of $\frac{1}{\tau}w_{FF}g_{w_+}(t)$ (section S2). This amplification factor is $\frac{w_{FF}}{1+w_+}$. For fluctuating inputs with no temporal correlations (white noise), the amplification in r_+ of input to r_- can be thought of as proportional to the square root of the integral of $\frac{1}{\tau}(w_{FF}g_{w_+}(t))^2$, while amplification for input with finite temporal correlations is likely to lie in between the amplification for white noise and the amplification for steady state inputs (section S2). This amplification factor for white noise inputs is $\frac{w_{FF}}{\sqrt{(1+w_+)(2+w_+)}}$.

More generally, $w_{FF}g_{w_+}(t)$ represents the characteristic response in the postsynaptic pattern with eigenvalue $-w_+$ to a unit initial condition in a presynaptic pattern with eigenvalue 0 that projects a feedforward connection of strength w_{FF}. Later, in Eq. S29 and section S3.4, we will see that the generalization of $g_{w_+}(t)$ to the case when the presynaptic pattern has eigenvalue $-w_-$ is $g_{w_+;w_-}(t) = \frac{e^{-(1+w_-)\frac{t}{\tau}} - e^{-(1+w_+)\frac{t}{\tau}}}{w_+ - w_-}$. The amplification of $w_{FF}g_{w_+;w_-}(t)$ for steady-state input is $\frac{w_{FF}}{(1+w_-)(1+w_+)}$, while its amplification for white noise input is $\frac{w_{FF}}{\sqrt{(1+w_-)(1+w_+)(2+w_-+w_+)}}$.

We define $\beta_+ = 1 + w_+$. The solutions to Eqs. S1-S2 are

$$\begin{aligned} r_+(t) &= r_+(0)e^{-\beta_+\frac{t}{\tau}} + \frac{I_+ + w_{FF}I_-}{\beta_+}\left(1 - e^{-\beta_+\frac{t}{\tau}}\right) \\ &\quad + w_{FF}\left(r_-(0) - I_-\right)g_{w_+}(t) && \text{(S4)} \\ r_-(t) &= (r_-(0) - I_-)e^{-\frac{t}{\tau}} + I_- && \text{(S5)} \end{aligned}$$

The terms multiplied by w_{FF} represent the effects of the feedforward connection from $\mathbf{r}_-$ to $\mathbf{r}_+$. w_{FF} scales the size of the amplification without affecting its time course or stability.

The E and I firing rates are $r_E = \frac{1}{2}(r_+ + r_-)$ and $r_I = \frac{1}{2}(r_+ - r_-)$. Then from Eqs. S4-S5 we can solve for $r_E(t)$ and $r_I(t)$, each as a sum of four terms representing the influence of each initial condition and each input:

$$\begin{aligned} r_E(t) &= r_E(0)e_{r_E}(t) - r_I(0)e_{r_I}(t) + I_E e_{I_E}(t) - I_I e_{I_I}(t) \qquad \text{with} \\ e_{r_E}(t) &= \frac{k_I e^{-\frac{t}{\tau}} - e^{-\beta_+\frac{t}{\tau}}}{k_I - 1} = e^{-\frac{t}{\tau}} + w g_{w_+}(t) \\ e_{r_I}(t) &= k_I\frac{e^{-\frac{t}{\tau}} - e^{-\beta_+\frac{t}{\tau}}}{k_I - 1} = k_I w g_{w_+}(t) \\ e_{I_E}(t) &= \frac{1}{\beta_+}\left((wk_I + 1) - \frac{k_I\beta_+ e^{-\frac{t}{\tau}} - e^{-\beta_+\frac{t}{\tau}}}{k_I - 1}\right) \\ &= \frac{1}{\beta_+}\left((wk_I + 1) - (w+1)e^{-\frac{t}{\tau}} - w g_{w_+}(t)\right) \\ e_{I_I}(t) &= \frac{k_I}{\beta_+}\left(w - \frac{\beta_+ e^{-\frac{t}{\tau}} - e^{-\beta_+\frac{t}{\tau}}}{k_I - 1}\right) = \frac{wk_I}{\beta_+}\left(1 - e^{-\frac{t}{\tau}} - g_{w_+}(t)\right) && \text{(S6)} \end{aligned}$$

$$\begin{aligned}
r_I(t) &= r_E(0)i_{r_E}(t) - r_I(0)i_{r_I}(t) + I_E i_{I_E}(t) - I_I i_{I_I}(t) \qquad \text{with} \\
i_{r_E}(t) &= \frac{e^{-\frac{t}{\tau}} - e^{-\beta_+\frac{t}{\tau}}}{k_I - 1} = w g_{w_+}(t) \\
i_{r_I}(t) &= \frac{e^{-\frac{t}{\tau}} - k_I e^{-\beta_+\frac{t}{\tau}}}{k_I - 1} = -e^{-\beta_+\frac{t}{\tau}} + w g_{w_+}(t) \\
i_{I_E}(t) &= \frac{1}{\beta_+}\left(w - \frac{\beta_+ e^{-\frac{t}{\tau}} - e^{-\beta_+\frac{t}{\tau}}}{k_I - 1}\right) = \frac{w}{\beta_+}\left(1 - e^{-\frac{t}{\tau}} - g_{w_+}(t)\right) \\
i_{I_I}(t) &= \frac{1}{\beta_+}\left(w - 1 - \frac{\beta_+ e^{-\frac{t}{\tau}} - k_I e^{-\beta_+\frac{t}{\tau}}}{k_I - 1}\right) \\
&= \frac{1}{\beta_+}\left((w-1)(1 - e^{-\frac{t}{\tau}}) - w k_I g_{w_+}(t)\right)
\end{aligned} \tag{S7}$$

The steady-state responses ($t \to \infty$) are $r_E^{\text{ss}} = [I_E + w k_I (I_E - I_I)]/\beta_+$, $r_I^{\text{ss}} = [I_I + w(I_E - I_I)]/\beta_+$ (or $r_+^{\text{ss}} = (I_+ + w_{FF} I_-)/\beta_+$, $r_-^{\text{ss}} = I_-$).

S1.1.2 Applications to Results in the Main Text

In Fig. 2 of the main text, we examined the steady-state response of r_E to a sustained input $I_E = 1$ ($I_I = 0$) starting from $r_E(0) = r_I(0) = 0$. This steady state response is $r_E^{\text{ss}} = (1 + w k_I)/(1 + w(k_I - 1))$. It is obtained for a given k_I by setting $w = (r_E^{\text{ss}} - 1)/(r_E^{\text{ss}} - k_I(r_E^{\text{ss}} - 1))$, which gives $w = 4\frac{2}{7}$ for $r_E^{\text{ss}} = 4$, $k_I = 1.1$. Note that this amplification factor, $(1 + w k_I)/(1 + w(k_I - 1))$, is also $\frac{1}{\tau}$ times the time integral of $e_{r_E}(t)$, as explained in section S2.1.

In Fig. 2F we saw that the time course of response of r_E to a steady input I_E grew faster with increasing amplification (increasing w). This time course is $e_{I_E}(t)$, whose time-dependent part is (negatively) proportional to $(w+1)e^{-\frac{t}{\tau}} + w g_{w_+}(t)$. As we saw above, as w is increased from zero to large values, the time course of $g_{w_+}(t)$ speeds up monotonically; in addition, the amplitude of $w g_{w_+}(t)$ increases from zero to $\frac{1}{k_I - 1}$. For $w \to 0$, $w g_{w_+}(t) \to w\frac{t}{\tau}e^{-\frac{t}{\tau}}$, while for $w \to \infty$, $w g_{w_+}(t) \to e^{-\frac{t}{\tau}}/(k_I - 1)$. In either limit, the time course of e_{I_E} becomes proportional simply to $e^{-\frac{t}{\tau}}$, *i.e.* the dynamics are just determined by the membrane time constant. For intermediate w, the term $w g_{w_+}(t)$ has a finite amplitude and a slightly slower time course than $e^{-\frac{t}{\tau}}$. The result is that, beginning at $w = 0$, the time course first slows slightly from $e^{-\frac{t}{\tau}}$ and then speeds up again back to $e^{-\frac{t}{\tau}}$ with increasing w. Note that this effect does not occur for perfectly balanced excitation and inhibition ($k_I = 1$). In this case, $w g_{w_+}(t) = w\frac{t}{\tau}e^{-\frac{t}{\tau}}$ for all w, so the time course is simply $(w+1)e^{-\frac{t}{\tau}} + w\frac{t}{\tau}e^{-\frac{t}{\tau}}$, which starts at $e^{-\frac{t}{\tau}}$ at $w = 0$ and slows toward an asymptotic (slowest possible, for large w) time course of $\left(1 + \frac{t}{\tau}\right)e^{-\frac{t}{\tau}}$, with amplitude w.

In Fig. 3C of the main text, we examined the time course of the response vector length $|\mathbf{r}(t)|$ to an initial condition in which one difference mode was set to $r_-(0) = 1$ and all other modes and inputs were set to zero. The result is that the difference mode decays, while serving as a source for its sum mode r_+. We noted that the first mode follows a time course $w_{FF}\frac{t}{\tau}e^{-\frac{t}{\tau}}$ once $w_{FF}\frac{t}{\tau} \gg 1$, corresponding to a zero eigenvalue, while subsequent modes had earlier and smaller peaks, reflecting the influence of increasingly negative eigenvalues. No modes other than the difference mode and its paired sum mode are activated, so we can write $|\mathbf{r}(t)| = \sqrt{r_-(t)^2 + r_+(t)^2}$. From Eqs. S4-S5, we find that $|\mathbf{r}(t)| = \sqrt{(w_{FF}g_{w_+}(t))^2 + (e^{-\frac{t}{\tau}})^2}$. For $w_{FF} \gg 1$, which was true for all of the pictured pairs of modes (all had $w_{FF} > 20$), this is well approximated by $|\mathbf{r}(t)| = w_{FF}g_{w_+}(t)$ except at very early times, $\frac{t}{\tau} \ll 1$, when $g_{w_+}(t)$ is very small. It is easy to show that at these early times $g_{w_+}(t) = \frac{t}{\tau} + O(\frac{t}{\tau}^2)$ (where $O(\frac{t}{\tau}^2)$ means terms involving $\frac{t}{\tau}$ raised to a power of 2 or greater). So once $w_{FF}\frac{t}{\tau} \gg 1$, *i.e.* $\frac{t}{\tau} \gg 1/w_{FF}$, then $|\mathbf{r}(t)| = w_{FF}g_{w_+}(t)$ and the described behavior follows immediately from this.

There is one slight wrinkle in this account. Except for the first sum mode which is an eigenvector of $\mathbf{W}$, each sum mode defined as the output of the corresponding difference mode actually is a linear combination of different sum eigenvectors with different negative eigenvalues, see section S1.2 below. We can regard this as the difference mode actually making feedforward connections to each of the underlying sum eigenvectors, with each eigenvector's dynamics described by Eq. S4 but with its own feedforward weight and eigenvalue. The sum of the squares of their feedforward weights will be equal to the square of the single feedforward weight shown in Fig. 3B. Since each eigenvector mode behaves as just described, a linear combination of them also behaves as just described.

S1.1.3 Paradoxical Effects of Input to Inhibitory Cells

In the discussion, we suggest that one test for the dynamics underlying balanced amplification is the "paradoxical" effect caused by adding input to inhibitory cells when the excitatory recurrence is strong enough that the excitatory subnetwork would be unstable by itself (*i.e.*, $w > 1$ in the two-population model) [Tsodyks et al. 1997, Ozeki et al. 2009]. The effect is that adding excitatory input to the inhibitory cells causes them to decrease their firing rate in the new steady state, and conversely adding inhibitory input or withdrawing excitatory input causes them to raise their firing rate. This is true for an arbitrary two-population model, but it is easy to see using the two-population model of Fig. 1 of the main text. We think of steady input I_I to the inhibitory cells as a set of delta-pulse inputs (inputs confined to a single instant dt) of size $I_I dt$, or equivalently as a set of initial conditions $r_I^0 = I_I dt/\tau$ induced by such an input; the steady-state response is then the superposition of the responses to all such past initial conditions, which is just the integral of the response to a single such initial condition (section S2.1). Each delta-pulse of excitation to the inhibitory population represents an equal-sized positive increase r_I^0 of r_+ and negative increase $-r_I^0$ of

r_-. Thus, we can think of the dynamics, up to a multiplicative scaling by r_I^0, as being as in Fig. 1 except with initial condition $r_-(0) = -1$ and $r_+(0) = 1$, rather than both having both initial conditions $= +1$ as in that figure. r_- will exponentially decay back to zero, and will induce a *negative* pulse response $-w_{FF}g_{w_+}(t)$ in r_+. This will add to the exponential decay of the r_+ initial condition to give the total response in r_+. The response of the inhibitory cell is $r_I = \frac{1}{2}(r_+ - r_-)$, which is an average of the two exponential decays, one with time constant τ and one with time constant $\frac{\tau}{1+w_+}$, plus the negative pulse response $-\frac{w_{FF}}{2}g_{w_+}(t)$. The overall amplification of the steady-state input I_I is just $1/\tau$ times the integral of this response. This yields $\frac{1+w_+/2}{(1+w_+)}$ for the average of the two exponentials and $-\frac{w_{FF}/2}{1+w_+}$ for the negative pulse, so the total integrated response is negative when $1 + w_+/2 < w_{FF}/2$ or $1 + w(k-1)/2 < w(k+1)/2$, which is precisely the condition $w > 1$.

This effect can also be seen from Eq. S7, where we see that the response of an inhibitory cell to input I_I is $I_I(-i_{I_I}(t))$ where $-i_{I_I}(t) = \frac{1}{\beta_+}\Big((1-w)(1-e^{-\frac{t}{\tau}}) + wk_I g_{w_+}(t)\Big)$. Thus we see immediately that the steady-state response, $1 - w$, is negative for $w > 1$. The response rises briefly due to the $g_{w_+}(t)$ term, representing the immediate response to the input, before falling and becoming negative as excitatory firing rates fall and feedback excitation is reduced.

S1.2 Multi-Neuron Model

We consider the weight matrix $\mathbf{W} = \begin{pmatrix} \mathbf{W}_E & -\mathbf{W}_I \\ \mathbf{W}_E & -\mathbf{W}_I \end{pmatrix}$, an example of which was studied in Fig. 3.

We first characterize the eigenvectors and eigenvalues of $\mathbf{W}$. Let $\mathbf{W}_E$ and $\mathbf{W}_I$ be $N \times N$, and let the normalized eigenvectors of $\mathbf{W}_E - \mathbf{W}_I$ be $\mathbf{e}_i^D$ with eigenvalues $-\lambda_i^D$, $(\mathbf{W}_E - \mathbf{W}_I)\mathbf{e}_i^D = -\lambda_i^D \mathbf{e}_i^D$, $i = 1, \ldots, N$.[1] We will imagine that inhibition balances or dominates excitation in such a manner that no pattern can excite itself – all the eigenvalues of $(\mathbf{W}_E - \mathbf{W}_I)$ have real part ≤ 0 – so we have taken the eigenvalue to be $-\lambda_i^D$ so that λ_i^D will have positive real part. Then $\mathbf{W}$ has N eigenvalues equal to the $-\lambda_i^D$, with corresponding normalized eigenvectors $\mathbf{p}_i^{D+} = \frac{1}{\sqrt{2}}\begin{pmatrix} \mathbf{e}_i^D \\ \mathbf{e}_i^D \end{pmatrix}$ (the + is used to indicate that these are sum modes), as can be seen directly by applying $\mathbf{W}$ to $\mathbf{p}_i^{D+}$. An additional N eigenvalues of $\mathbf{W}$ are equal to zero, because the top N rows are identical to the bottom N rows. If either $\mathbf{W}_E$ or $\mathbf{W}_I$ are invertible, the corresponding eigenvectors can be written as proportional

[1] In the main text we used the convention for basis vectors of denoting both which basis vector (i) and which type of basis vector (+) as superscripts, $\mathbf{p}^{i+}$, so that subscripts could be used to designate elements of the vector. In the supplement we will revert to the more usual convention $\mathbf{p}_i^+$; should we need to refer to the j^{th} element, we would write $(\mathbf{p}_i^+)_j$.

to $\begin{pmatrix} {\mathbf{W}_E}^{-1}\mathbf{W}_I\mathbf{v} \\ \mathbf{v} \end{pmatrix}$ or $\begin{pmatrix} \mathbf{v} \\ {\mathbf{W}_I}^{-1}\mathbf{W}_E\mathbf{v} \end{pmatrix}$ for any N-dimensional basis $\mathbf{v}$. Note that, with the assumption that inhibition appropriately balances or dominates excitation, $\mathbf{W}$ has no eigenvalues with positive real part.

We now consider the feedforward connectivity. We let $\mathbf{e}_i^S$ be the normalized eigenvectors of $\mathbf{W}_E + \mathbf{W}_I$ with eigenvalues λ_i^S, and note that $\mathbf{W}_E + \mathbf{W}_I$ is a nonnegative matrix with large entries (if excitation and inhibition are large) so that some of these eigenvalues will be large and positive. We define the difference modes $\mathbf{p}_i^{S-} = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_i^S \\ -\mathbf{e}_i^S \end{pmatrix}$ and the sum modes $\mathbf{p}_i^{S+} = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_i^S \\ \mathbf{e}_i^S \end{pmatrix}$ and find that $\mathbf{W}\mathbf{p}_i^{S-} = \lambda_i^S \mathbf{p}_i^{S+}$. Thus, each pair $\mathbf{p}_i^{S-}$, $\mathbf{p}_i^{S+}$ behaves much like the difference and sum modes, $\mathbf{p}^-$ and $\mathbf{p}^+$, in the simpler, two-neuron model we studied previously, with feedforward weight $w_i^{FF} = \lambda_i^S$.

There is one difference, however. Each $\mathbf{p}_i^{S+}$ is a linear combination[2] of the $\mathbf{p}_i^{D+}$, each of which in turn decays at its own rate (determined by its λ_j^D). So the decay of $\mathbf{p}_i^{S+}$ is actually a mix of decays at different rates, rather than a decay at a single rate as before. Instead of thinking in terms of $\mathbf{p}_i^{S-}$ making a single feedforward connection to $\mathbf{p}_i^{S+}$, which then decays as a mixture of modes, one can alternatively think of $\mathbf{p}_i^{S-}$ making a set of feedforward connections to the different $\mathbf{p}_i^{D+}$'s, each of which decays at its own rate. If $\mathbf{p}_i^{S+} = \sum_j c_{ij}\mathbf{p}_j^{D+}$, then the feedforward connection from $\mathbf{p}_i^{S-}$ to $\mathbf{p}_j^{D+}$ is equal to $\lambda_i^S c_{ij}$. If the $\mathbf{e}_j^D$ and thus the $\mathbf{p}_j^{D+}$ are mutually orthogonal (see below), then $c_{ij} = \mathbf{p}_j^{D+} \cdot \mathbf{p}_i^{S+} = \mathbf{e}_j^D \cdot \mathbf{e}_i^S$.

There is one other slight wrinkle. If the matrix $\mathbf{W}_E + \mathbf{W}_I$ is not normal, then the $\mathbf{p}_i^{S-}$ will not be mutually orthogonal, nor will the $\mathbf{p}_i^{S+}$, though each $\mathbf{p}_i^{S-}$ will be orthogonal to each $\mathbf{p}_j^{S+}$. Similarly, if $\mathbf{W}_E - \mathbf{W}_I$ is not normal, the $\mathbf{p}_j^{D+}$ will not be mutually orthogonal. If this is true, this description, while correct, could be misleading in the same way that the solution in the eigenvector basis is misleading when the eigenvectors are not orthogonal, namely the size or dynamics of the basis pattern amplitudes may not directly reflect the size or dynamics of the rates. The $\mathbf{W}_E$ and $\mathbf{W}_I$ matrices we used in Fig. 3 are slightly nonnormal, because the normalization of total excitatory and inhibitory weights onto each neuron (see Methods) results in small asymmetries. However, this non-normality is very small, as assessed by measures such as $f^{\mathbf{M}}$ (see section S3.2), so the vast majority of the non-normality of the overall matrix $\mathbf{W}$ is the result of the arrangement of the submatrices, not the non-normality of the submatrices themselves. In other words, these basis patterns should be close to orthogonal to one another, if not orthogonal, so distortions, if any, should be small. Our guess is that this will be typical of biological connection matrices.

[2] This is true because the $\mathbf{p}_i^{S+}$ and the $\mathbf{p}_i^{D+}$ each span the N-dimensional space of vectors that have identical patterns of activity in the excitatory and the inhibitory neurons

We can write down the solution in a basis of the $\mathbf{p}_i^{S-}$ and either of the group of sum modes; we choose to use $\mathbf{p}_j^{D+}$. Each $\mathbf{p}_i^{S-}$ is orthogonal to each $\mathbf{p}_j^{D+}$, and if $\mathbf{W}_E + \mathbf{W}_I$ and $\mathbf{W}_E - \mathbf{W}_I$ are normal (or close to normal), this is an orthonormal (or close to orthonormal) basis. We let $\mathbf{C}$ be the matrix with elements $C_{ij} = c_{ji}\lambda_j^S$, and let $\mathbf{L}^D$ be the diagonal matrix of the the $-\lambda_i^D$. Then in the basis $\{\mathbf{p}_1^{D+}, \ldots, \mathbf{p}_N^{D+}, \mathbf{p}_1^{S-}, \ldots, \mathbf{p}_N^{S-}\}$, the matrix $\mathbf{W}$ becomes $\begin{pmatrix} \mathbf{L}^D & \mathbf{C} \\ 0 & 0 \end{pmatrix}$.

The solution to $\tau \frac{d}{dt}\mathbf{r} = -\mathbf{r} + \mathbf{W}\mathbf{r} + \mathbf{I}$ for time-independent $\mathbf{I}$ can be formally written $\mathbf{r}(t) = e^{-\frac{t}{\tau}(\mathbf{1}-\mathbf{W})}\mathbf{r}(0) + (\mathbf{1} - e^{-(\mathbf{1}-\mathbf{W})\frac{t}{\tau}})(\mathbf{1}-\mathbf{W})^{-1}\mathbf{I}$, where, for a matrix $\mathbf{M}$, the matrix $e^{\mathbf{M}}$ is defined by the same power series as for the ordinary exponential, $e^{\mathbf{M}} = \mathbf{1} + \mathbf{M} + \mathbf{M}^2/2! + \mathbf{M}^3/3! + \ldots$. Thus, calculating $e^{-\frac{t}{\tau}(\mathbf{1}-\mathbf{W})} = e^{-\frac{t}{\tau}} e^{\frac{t}{\tau}\mathbf{W}}$ amounts to solving the equation. This turns out to be easy to do, and we can write the solution as follows. Let $\mathcal{L}^D$ be the diagonal matrix of $e^{-\lambda_i^D \frac{t}{\tau}}$, and define $\mathbf{K}$ as the matrix with entries $K_{ij} = c_{ji}\lambda_j^S(1 - e^{-\lambda_i^D \frac{t}{\tau}})/\lambda_i^D$. Then $e^{-\frac{t}{\tau}(\mathbf{1}-\mathbf{W})} = e^{-\frac{t}{\tau}} \begin{pmatrix} \mathcal{L}^D & \mathbf{K} \\ 0 & \mathbf{1} \end{pmatrix}$.

This solution tells us that an initial condition of size 1 of $\mathbf{p}_j^{S-}$ causes a response in the sum pattern $\mathbf{p}_i^{D+}$ equal to $e^{\frac{t}{\tau}}\mathbf{K}_{ij} = \lambda_j^S c_{ji} \left(e^{-\frac{t}{\tau}} - e^{-(1+\lambda_i^D)\frac{t}{\tau}} \right) / \lambda_i^D = w_{FF} g_{\lambda_i^D(t)}$ with $w_{FF} = \lambda_j^S c_{ji}$ and $g_{\lambda_j^D}(t) = g_{w_+}(t)$ (defined in Section S1.1.1) for $w_+ = \lambda_j^D$. This is precisely the response we derived for the sum mode amplitude $r_+(t)$ in the two-population model in response to an initial difference mode amplitude $r_-(0) = 1$, Eq. S4. More generally, if, in Eqs. S4-S5, I_- and I_+ are understood to be the inputs to mode $\mathbf{p}_j^{S-}$ and $\mathbf{p}_i^{D+}$, respectively, and r_- and r_+ their respective amplitudes, then, with the substitutions $w_{FF} \to \lambda_j^S c_{ji}$ and $w_+ \to \lambda_i^D$ (and thus $\beta_+ = 1 + \lambda_i^D$), Eqs. S4-S5 describe the solution for the amplitudes of $\mathbf{p}_j^{S-}$ and $\mathbf{p}_i^{D+}$ arising from initial conditions and inputs of these two modes. Other difference modes $\mathbf{p}_k^{S-}$ might also project to $\mathbf{p}_i^{D+}$. In this case, the terms that the various difference modes generate for $\mathbf{p}_i^{D+}$ under equation S4 (those involving $r_-(0)$ and I_-) must be added together, along with a single instance of the terms involving $r_+(0)$ and I_+, to yield the solution for the amplitude of $\mathbf{p}_i^{D+}$.

In summary, the differences mode $\mathbf{p}_j^{S-}$, in which excitation and inhibition have spatial patterns of activity $\mathbf{e}_j^S$ with opposite signs, is amplified into the sum pattern $\mathbf{p}_j^{S+}$, in which excitation and inhibition have the same spatial pattern $\mathbf{e}_j^S$ but now of the same sign, with feedforward weight λ_j^S. The spatial pattern $\mathbf{e}_j^S$ in turn is a mixture of the patterns $\mathbf{e}_i^D$ with weights c_{ji}, so that we can instead take the $\mathbf{p}_j^{S-}$ to send feedforward weights $\lambda_j^S c_{ji}$ to the various sum eigenvector patterns $\mathbf{p}_i^{D+}$, which have eigenvalues λ_i^D. The amplitudes of $\mathbf{p}_j^{S-}$ and $\mathbf{p}_i^{D+}$ receiving inputs I_- and I_+, respectively are then described precisely by the solutions for the amplitudes r^- and r^+, respectively, of the two-population system (Eqs. S4-S5), with λ_i^D replacing w_+. If $\mathbf{p}_i^{D+}$ receives inputs from multiple difference modes $\mathbf{p}_k^{S-}$, each of their contributions to $\mathbf{p}_i^{D+}$ under Eqs. S4 simply add together to yield the amplitude of $\mathbf{p}_i^{D+}$.

S2 Amplification in the Data and the Models

S2.1 Relationship Between Transient Response to an Initial Condition and Sustained Response to Onset of a Steady-State Stimulus

Here we remind the reader of a simple result for a linear model that we refer to in several places: the response at time t to the onset of a sustained stimulus is just proportional to the integral from 0 to t of the transient response to an initial condition created by a delta-pulse of that input (by delta-pulse, we mean input restricted to a single instant of time, represented by the infinitesimal width dt). As specific examples, in Eqs. S4-S7, each term multiplied by I_X ($X =$E,I,+,$-$), which represents the time course of response to the onset of I_X, is $\frac{1}{\tau}$ times the time integral of the corresponding term that is multiplied by $r_X(0)$, which represents the transient response to the initial condition $r_X(0)$.

We consider the response $r_j(t)$ of pattern or neuron j to the onset of steady-state input I_k to pattern k with onset time $t = 0$. We suppose the response in j to an initial condition $r_k(0)$ is $r_k(0)K_{jk}(t)$ for some temporal response function $K_{jk}(t)$ with $K_{jk}(t) = 0$ for $t < 0$. Given the differential equation that says $\tau\frac{d}{dt}r_k = \ldots + I_k$, we see that a delta-pulse of input I_k to k – an input confined to a the time dt – evokes an immediate change in r_k of $dr_k = I_k(dt/\tau)$. Thus, the instantaneous delta-pulse of input at time $t' > 0$ evokes an "initial condition" $r_k(t') = I_k(dt/\tau)$, and at time $t > t'$ the response of r_j to this initial condition has become $\Delta r_j(t) = R_{jk}(t-t')I_k dt/\tau$. Since it is a linear model, the responses to the input delta-pulses at different times superpose, so the full response $r(t)$ to $I(k)$ is obtained by integrating $\Delta r(t)$ over the t''s for which the stimulus has been on:

$$r_j(t) = \int_0^t dt' \frac{1}{\tau} K_{jk}(t-t') I_k = I_k \int_0^t dt' \frac{1}{\tau} K_{jk}(t') \tag{S8}$$

(where we have changed variables $t-t' \to t'$ in the last step). We can think of $\frac{1}{\tau}K_{jk}(t)$ either as $\frac{1}{\tau}$ times the temporal kernel describing the response of r_j to a unit initial condition of r_k, or as the kernel describing the response of r_j to a delta-pulse of input I_k.

We can describe this result in different words: the rate of rise of an onset response to a steady stimulus is just determined by the rate of accumulation of the area under the curve of the transient response to a delta-pulse of the stimulus. This is the reason why the slow response to a delta-pulse input in Fig. 2A yields the slow onset response in Fig. 2C, while the fast response to a delta-pulse input in Fig. 2B yields the fast onset response in Fig. 2D. This result also tells us that the steady-state response to I_k is just the full time integral

$I_k \int_0^\infty dt' \frac{1}{\tau} K_{jk}(t')$, so the steady-state amplification of a constant stimulus is determined by the area under the curve of the transient pulse response.

S2.2 Amplification in the Data

We are motivated by the optical recording data of Kenet et al. [2003]. There, amplification was measured in the fluctuating spontaneous activity, presumably driven by fluctuating inputs. The amplification of a given pattern (in their case, the average evoked response to an oriented grating) was measured as the increase in the width of the distribution of correlation coefficients between spontaneous activity and that pattern, relative to the width of the similar distribution for a control pattern (in their case, a mirror-reflection of the evoked pattern). A reason for using the correlation coefficient, rather than simply the amplitude, is that biologically there are factors that nonspecifically elevate or suppress the size of all patterns in the data (changes in overall excitability including those due to changes in anesthesia level; changes in overall signal due to fluctuations in the illuminant). These would increase the standard deviation of the amplitude but are factored out of the correlation coefficient. Our analysis of their published data (their Fig. 2) indicates the amplification by this measure was by a factor of 2. Our spiking model simulations (Fig. 5 of main text) used precisely this measure to assay the amplification. The simulations also show amplification of 2, and values ranging from 1-3 (or more) are easily achieved by strengthening or weakening all recurrent weights (Fig. S3A, blue line).

There are several uncertainties in the estimate that the patterns in Kenet et al. [2003] showed amplification by a factor of 2. First, the evoked map patterns probably do not perfectly correspond to an eigenvector (Hebbian amplification) or a Schur basis vector (balanced amplification), in which case the amplification of the most similar eigenvector(s) or Schur basis vector(s) would need to be by a factor greater than 2 in order for the evoked maps to be amplified by a factor of 2. Second, the control pattern might itself be amplified, in which case the evoked patterns must be amplified by the circuit by a factor greater than 2. Alternatively, the control pattern might be a mixture of eigenvectors that have negative eigenvalues and might not receive significant effective feedforward input, and thus be diminished rather than amplified by the network. In this case, to obtain a relative amplification of 2 for the evoked map relative to the control pattern, the absolute amplification of the evoked map would be less than 2. Finally, the measured degree of amplification is dependent to some extent on the filtering of the image, since filtering reduces the number of degrees of freedom or dimensions of the data and thus changes the denominator of a correlation coefficient (Goldberg et al. 2004 and Fig. S4), and the data analyzed by Kenet et al. [2003] was filtered both by the optics and the brain tissue [Polimeni et al. 2005] and by subsequent processing.

S2.3 Amplification in Linear Models

We define the amplification for a fluctuating input and the amplification to a steady-state input produced in the linear models. For fluctuating input, we take the amplification A_j of a pattern j to be the measure used by Kenet et al. [2003]: the standard deviation of the correlation coefficient of pattern j with the fluctuating response, relative to the same measure for an unamplified pattern. By an unamplified pattern, we mean the response of any pattern in the network with all recurrent weights set to zero (we assume that all patterns have statistically identical input). We show that this A_j, if factors that nonspecifically change overall activity or signal levels are eliminated, is equivalent to the standard deviation of the response $r_j(t)$, suitably normalized to give 1 for the case of no recurrent connections. For a steady state input, we define the amplification A_{jk} of pattern or neuron j in response to a nonzero input I_k to pattern or neuron k to be r_j/I_k.

We first present the answers; we will then present the details of their derivation. We assume that the response of r_j to a unit initial condition for k, $r_k(0) = 1$, is described by the function $K_{jk}(t)$. Then for a steady-state input to pattern or neuron k, the amplification of pattern or neuron j is

$$A_{jk} = \int_0^\infty dt \frac{1}{\tau} K_{jk}(t) \text{ (Steady State)} \tag{S9}$$

as shown in section S2.1. For fluctuating input, the amplification of pattern j depends on the statistics of the input. We can work out two limits, giving:

$$\begin{aligned} A_j &= \left(\sum_k A_{jk}^2 \right)^{1/2} \text{ (Fluctuating Input); where} \\ A_{jk} &= \left(2 \int_0^\infty dr \frac{1}{\tau} K_{jk}(r)^2 \right)^{1/2} \text{ (White Noise Input);} && \text{(S10)} \\ A_{jk} &= \int_0^\infty dt \frac{1}{\tau} K_{jk}(t) \text{ (Input with long correlation times)} && \text{(S11)} \end{aligned}$$

In the limit of long correlation times, A_{jk} for fluctuating input is the same as for steady state input. We also suggest that amplification to input with finite correlation times might be well thought of as bounded between these two limits.

The input to the patterns we study are typically dominated by the "feedforward" input from sum mode to difference mode. If there is a single dominant input k, then for the fluctuating input case $A_j \approx A_{jk}$. Based on this, elsewhere in this supplement we simply use A_{jk} from equation S10 as the amplification expected for white noise input from a mode supplying such a feedforward link.

The details for the case of fluctuating input follow, but can be safely skipped.

The Details: Fluctuating Input

We generalize the approach of section S2.1 to the case of multiple inputs. We assume that the response $r_j(t)$ of pattern j arises from a sum over patterns k of a filtering of their inputs $I_k(t)$. The filters or kernels for different inputs can be different: for example, in our two-population model, the sum response pattern has one kernel of response (exponential decay) to a sum initial condition and another (a pulse response) to a difference initial condition (Fig. 1; Eq. S4). Following the reasoning presented in section S2.1, but with multiple time-dependent inputs, we arrive at

$$r_j(t) = \sum_k \int_{-\infty}^{t} dt' \frac{1}{\tau} K_{jk}(t - t') I_k(t') \tag{S12}$$

For fluctuating inputs, we assume the input patterns are all independent and have identical statistical properties, so that any differences in output (*i.e.*, amplification) of different patterns result from differences in their kernels. We also assume that the inputs have zero means, that is, either polarity of a given pattern is equally likely in the input noise.

We begin with the measure of Kenet et al. [2003], in which amplification is measured as the standard deviation of the set of correlation coefficients of the pattern j. Suppose we have a set of orthonormal basis patterns $\mathbf{p}_i$. The spontaneous activity is $\mathbf{r}(t) = \sum_i r_i(t)\mathbf{p}_i$. The correlation coefficient with pattern $\mathbf{p}_j$ is then $cc_j(t) = \frac{\mathbf{r}(t)\cdot\mathbf{p}_j}{|\mathbf{r}(t)||\mathbf{p}_j|} = \frac{r_j(t)}{|\sum_i r_i(t)^2|}$. The width of the distribution of cc_j's, measured as the standard deviation of the distribution, is $\langle cc_j(t)^2\rangle_t^{1/2} = \left\langle \frac{r_j(t)^2}{\sum_i r_i(t)^2} \right\rangle_t^{1/2}$ where $\langle x(t)\rangle_t$ is the time average of x. We argue that the numerator and denominator can be taken to be independent. In the biological data from optical imaging there will be factors that scale all patterns up or down together, as discussed above, which will scale numerator and denominator together, but the correlation coefficient factors these out, and they are not present in our models of amplification, so we will ignore such effects. Then the numerator is still correlated with the denominator, because when r_j is large, it will contribute a correspondingly larger amount to the denominator. But if the system has many independent basis patterns that contribute significantly to the denominator, this will be a very small effect, so that it should be a good approximation to treat the numerator and denominator as independent.

This approximation means that, for purposes of computing the correlation coefficient cc_j, we can ignore any dependence of the denominator on K_{jk}. So we arrive at the conclusion that, for any patterns j, p, the ratio of their correlation coefficient standard deviations is just the ratio of their response standard deviations:

$$\frac{\langle cc_j(t)^2\rangle_t^{1/2}}{\langle cc_p(t)^2\rangle_t^{1/2}} = \frac{\langle r_j(t)^2\rangle_t^{1/2}}{\langle r_p(t)^2\rangle_t^{1/2}} \tag{S13}$$

Thus, if we take our amplification measure to be the response standard deviation, normalized to be 1 for the network with no recurrent connections, then this amplification measure will correctly assay the increase in correlation coefficient width of a pattern relative to an unamplified pattern.

In turn,

$$\begin{aligned}
\left\langle r_j(t)^2 \right\rangle_t^{1/2} &= \left\langle \left(\sum_k \frac{1}{\tau} K_{jk} * I_k(t) \right)^2 \right\rangle_t^{1/2} \\
&= \left\langle \sum_{kl} \int_{-\infty}^{t} dp \int_{-\infty}^{t} dq \frac{1}{\tau^2} K_{jk}(t-p) K_{jl}(t-q) I_k(p) I_l(q) \right\rangle_t^{1/2} \\
&= \sum_{kl} \int_0^{\infty} dr \int_0^{\infty} ds \frac{1}{\tau^2} K_{jk}(r) K_{jl}(s) \left\langle I_k(t-r) I_l(t-s) \right\rangle_t^{1/2} \\
&= \left(\sum_{kl} \int_0^{\infty} dr \int_0^{\infty} ds \frac{1}{\tau} K_{jk}(r) C_{kl}^{input}(r-s) \frac{1}{\tau} K_{jl}(s) \right)^{1/2} \qquad \text{(S14)}
\end{aligned}$$

where $C_{kl}^{input}(x) = \langle I_k(t) I_l(t+x) \rangle_t$ is the input correlation function. For $k \neq l$, C_{kl}^{input} is just the square of the mean of any input pattern, which is 0. For $k = l$, $C_{kl}^{input}(x)$ is the correlation function of any individual input pattern, which we will call $C^{input}(x)$. We have assumed that the input statistics are stationary in time, so that C^{input} only depends on the difference in time between two samples of the input pattern. Thus,

$$\left\langle r_j(t)^2 \right\rangle_t^{1/2} = \left(\sum_k \int_0^{\infty} dr \int_0^{\infty} ds \frac{1}{\tau} K_{jk}(r) C^{input}(r-s) \frac{1}{\tau} K_{jk}(s) \right)^{1/2} \qquad \text{(S15)}$$

In general, this depends on the structure of both K_{jk} and C^{input}. For the special case that the input is temporally white, $C^{input}(r-s) = C^2 \tau \delta(r-s)$ (where the τ is included so that both C^2 and C^{input} have the dimension of i_k^2), it becomes

$$\begin{aligned}
\left\langle r_j(t)^2 \right\rangle_t^{1/2} &= C \left(\sum_k A_{jk}^2 / 2 \right)^{1/2} \quad \text{where} \\
A_{jk} &= \left(2 \int_0^{\infty} dr \frac{1}{\tau} K_{jk}(r)^2 \right)^{1/2} \qquad \text{(S16)}
\end{aligned}$$

A_{jk} represents the contribution to the amplification of pattern j of input to pattern k when the input is white noise. The factor of 2 is included so that $A_{jk} = \delta_{jk}$ and the amplification is 1 for the network without recurrent connections, for which $K_{jk} = \delta_{jk} e^{-t/\tau}$.

On the other hand, as the input temporal correlations become long, A_{jk} goes to the amplification seen for a steady-state input. Intuitively, the temporal kernel extends only some finite extent T in time, *i.e.* there is a limit to how long the response to a delta-pulse input or an initial condition will endure. As input temporal correlations become comparable to and then longer than this time, the input within the window seen by the kernel will become more and more constant, and so the amplification will become the same as that for a steady-state input. Mathematically, in Eq. S14, if $K_{jk}(r) \approx 0$ for $r \geq T$, then when $C^{input}(x)$ becomes roughly constant (with value $C^2 = \langle i_k(t)^2 \rangle_t$) over $-T \leq x \leq T$, the expression becomes

$$\begin{aligned} \left\langle r_j(t)^2 \right\rangle_t^{1/2} &= C \left(\sum_k A_{jk}^2 \right)^{1/2} \quad \text{where} \\ A_{jk} &= \int_0^\infty dr \frac{1}{\tau} K_{jk}(r) \end{aligned} \tag{S17}$$

Thus, the factor by which the input is amplified is just the integral of the kernel, as in the steady-state case. It seems reasonable to guess that the amplification to input with finite temporal correlations will lie somewhere between the bounds of the amplification to white noise input and the amplification to steady-state input, though there is no guarantee of this.

S3 Non-normal matrices, Neurobiological Connection Matrices, and the Schur Decomposition

Normal matrices are matrices $\mathbf{M}$ that satisfy $\mathbf{M}^\dagger\mathbf{M} = \mathbf{M}\mathbf{M}^\dagger$ where $\mathbf{M}^\dagger$ is the complex conjugate of the transpose of $\mathbf{M}$, or equivalently, matrices that have a complete orthonormal basis of eigenvectors.[3] For real matrices, $\mathbf{M}^\dagger = \mathbf{M}^T$, the transpose of $\mathbf{M}$.

S3.1 Neurobiological connection matrices are non-normal

Neurobiological connection matrices are of the form $\mathbf{W} = \begin{pmatrix} \mathbf{W}_{EE} & -\mathbf{W}_{EI} \\ \mathbf{W}_{IE} & -\mathbf{W}_{II} \end{pmatrix}$, with all entries of the $\mathbf{W}_{XY}$ being non-negative. The simplest way to see that these are non-normal is just to consider the arrangement of the signs of the nonzero entries: $\begin{pmatrix} + & - \\ + & - \end{pmatrix}$. For such

[3] The overall idea underlying this equivalence is: the right eigenvectors of $\mathbf{M}^\dagger$ are the conjugate transpose of the left eigenvectors of $\mathbf{M}$. Two matrices share a common basis of eigenvectors if and only if they commute. Thus, iff $\mathbf{M}^\dagger$ and $\mathbf{M}$ commute, the right and left eigenvectors of $\mathbf{M}$ are identical (meaning that one set is the conjugate transpose of the other). These are mutually orthonormal, so iff they are identical, they constitute an orthonormal basis.

a matrix, $\mathbf{W}\mathbf{W}^T$ has signs $\begin{pmatrix} + & + \\ + & + \end{pmatrix}$, while $\mathbf{W}^T\mathbf{W}$ has signs $\begin{pmatrix} + & - \\ - & + \end{pmatrix}$. So, assuming the off-diagonal blocks are not all zero, $\mathbf{W}$ is non-normal.

More generally, $\mathbf{W}^T = \begin{pmatrix} \mathbf{W}_{EE}^T & \mathbf{W}_{IE}^T \\ -\mathbf{W}_{EI}^T & -\mathbf{W}_{II}^T \end{pmatrix}$. So

$$\mathbf{W}\mathbf{W}^T = \begin{pmatrix} \mathbf{W}_{EE}\mathbf{W}_{EE}^T + \mathbf{W}_{EI}\mathbf{W}_{EI}^T & \mathbf{W}_{EE}\mathbf{W}_{IE}^T + \mathbf{W}_{EI}\mathbf{W}_{II}^T \\ \mathbf{W}_{IE}\mathbf{W}_{EE}^T + \mathbf{W}_{II}\mathbf{W}_{EI}^T & \mathbf{W}_{IE}\mathbf{W}_{IE}^T + \mathbf{W}_{II}\mathbf{W}_{II}^T \end{pmatrix} \tag{S18}$$

while

$$\mathbf{W}^T\mathbf{W} = \begin{pmatrix} \mathbf{W}_{EE}^T\mathbf{W}_{EE} + \mathbf{W}_{IE}^T\mathbf{W}_{IE} & -\mathbf{W}_{EE}^T\mathbf{W}_{EI} - \mathbf{W}_{IE}^T\mathbf{W}_{II} \\ -\mathbf{W}_{EI}^T\mathbf{W}_{EE} - \mathbf{W}_{II}^T\mathbf{W}_{IE} & \mathbf{W}_{EI}^T\mathbf{W}_{EI} + \mathbf{W}_{II}^T\mathbf{W}_{II} \end{pmatrix}. \tag{S19}$$

$\mathbf{W}$ cannot be normal unless all the submatrices are symmetric and $\mathbf{W}_{EI} = \mathbf{W}_{IE}$. In this case, the requirements for $\mathbf{W}$ to be normal reduce to $\mathbf{W}_{EE}\mathbf{W}_{EI} = 0$ and $\mathbf{W}_{EI}\mathbf{W}_{II} = 0$. The first is equivalent to saying that no excitatory cell that receives a connection from an inhibitory cell makes a projection to another excitatory cell, while the second is equivalent to saying that no inhibitory cell that receives a connection from another inhibitory cell makes a projection to an excitatory cell. Clearly, no plausible connectivity pattern will be normal.

We are of course ignoring many elements of biological complexity, starting with the fact that a connection matrix is used to describe connections onto cells that linearly sums their inputs, and may not be an adequate description to the extent that summation on dendritic trees is nonlinear [*e.g.* Spruston 2008]. Even within the connection matrix formalism, we are ignoring the fact that there are gap junctions among inhibitory neurons of a given subtype [Beierlein et al. 2003], which represent an excitatory influence of one inhibitory neuron on another. Thus, some elements of $\mathbf{W}_{II}$ conceivably could be negative. We imagine that these effects are not critical to the rate dynamics we are studying (though they may be very important to spike synchronization and rhythms, [*e.g.* Pfeuty et al. 2005]), but of course we cannot be certain.

S3.2 The Schur Decomposition

The Schur decomposition gives a “simplest” orthonormal basis for a non-normal matrix. Before we describe the Schur decomposition itself, we explain the motivation for using an orthonormal basis rather than the non-orthogonal eigenvector basis.

In the text, we stated that when eigenvectors that are far from orthogonal are used as basis vectors, the size and time course of their amplitudes can give a misleading picture of the dynamics. To see why the decomposition into the eigenvector basis is deceiving, we examine the dynamics of the simple two-population model of Fig. 1 in the r_E/r_I plane, and its decomposition into the basis of the non-orthogonal eigenvectors of $\mathbf{W}$ (Fig. S1A) or of the orthogonal sum and difference modes (Fig. S1B). Recall that the dynamics are given by

$$\frac{d}{dt}\mathbf{r} = -\mathbf{r} + \mathbf{W}\mathbf{r} + \mathbf{I} \tag{S20}$$

with $\mathbf{r} = \begin{pmatrix} r_e \\ r_i \end{pmatrix}$ and $\mathbf{W} = \begin{pmatrix} w & -k_I w \\ w & -k_I w \end{pmatrix}$ in the $\mathbf{r}_e$, $\mathbf{r}_i$ basis. We start the dynamics from the initial condition $\mathbf{r}(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, that is, with excitation but not inhibition active.

The eigenvectors of $\mathbf{W}$ are the sum mode, proportional to $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, with eigenvalue $-w_+ = -w(k_I-1)$, and another very similar pattern, proportional to $\begin{pmatrix} k_I \\ 1 \end{pmatrix}$, with eigenvalue 0. The amplitudes of the two eigenvectors are initially large, and both monotonically (exponentially) decay to zero with time constants $\tau_+ = \tau/(1 + w_+)$ and τ respectively (Fig. S1A).[4] From the eigenvalues and the corresponding monotonic amplitude decays, there is no hint that the neural activities, given by the sum of the eigenvectors weighted by their amplitudes, are actually growing. Rather, this fact is hidden in the non-orthogonal geometry of the eigenvectors and the complicated ways in which they can cancel one another. Because the two eigenvectors are not orthogonal, a small initial condition at a large angle from both must be represented as a sum of one eigenvector with a large positive amplitude and the other with a large negative amplitude – two large contributions must cancel to produce the small initial condition. Then each component independently decays, but at different rates. As a result, one large component is increasingly revealed as the other decays away; the overall network activity grows, moving from the small initial condition toward the large remaining component.

If orthonormal basis patterns (meaning mutually orthogonal and normalized to length 1) are used, then the sum of the squares of the amplitudes of the basis patterns is equal to the sum of the squares of the neuronal firing rates, so the amplitudes accurately reflect the

[4] To express the initial condition as the sum of the two unnormalized eigenvectors, we write $\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{k_I-1}\left(\begin{pmatrix} k_I \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right)$. That is, the initial amplitudes of the two eigenvectors are large ($\frac{1}{k_I-1} \gg 1$) but of opposite sign, largely cancelling one another to create the much smaller initial condition. These amplitudes then each exponentially decay to zero, giving $\mathbf{r}(t) = \frac{1}{k_I-1}\left(e^{t/\tau}\begin{pmatrix} k_I \\ 1 \end{pmatrix} - e^{t/\tau_+}\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right)$. Because $\tau_+ < \tau$, the $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ term decays away more quickly, leaving $\mathbf{r}(t)$ dominated by the more slowly decaying $\begin{pmatrix} k_I \\ 1 \end{pmatrix}$ vector.

size and the growth or decay of overall activity. Using the difference and sum modes (p^- and p^+, normalized to length 1) as a basis (Fig. S1B), the amplitude of the difference mode monotonically decays, while that of the sum mode first grows, because of its feedforward connection from the difference, and then decays (these amplitudes are plotted in Fig. 1), directly revealing the non-monotonic dynamics of the firing rates. Orthogonal components cannot cancel one another, so one cannot have the situation in which large components cancel to create a small resultant, and thus in which the decay of some components reveals hidden large components in other directions.

These effects are quite general: in higher dimensions, if the eigenvectors are not orthogonal, but are all normalized to length 1, we can say that some directions (unit vectors) are poorly represented if they are close to orthogonal with (have small dot products with) all of the eigenvectors. An example of such a direction for $\mathbf{W}$ would be the difference direction, $\frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$; the initial condition $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ has a significant component in (significant dot product with) that direction. Then to represent a small vector with a significant component in a poorly represented direction, there must be a weighted sum of large but cancelling amplitudes of various eigenvectors. That is, in the eigenvector basis, the amplitudes decay independently – the change in one does not depend on the values of the others – but the eigenvectors are dependent in a hidden way, namely in the way they combine and cancel to represent a given vector (*e.g.* the initial condition): the weight assigned to one eigenvector depends on the other eigenvectors.[5] In an orthonormal basis, each basis pattern's contribution to the representation is independent (it is given by the dot product of the basis pattern with the represented vector). The dependence between basis patterns becomes explicit in the dependencies between their amplitudes – the evolution of one amplitude depends on the values of others – as represented for our 2-D example in the feedforward connection from p^- to p^+.

The problem with non-orthogonal bases can be stated more generally as follows: a transformation to a non-orthogonal basis is not *unitary*, and unitary transformations are the only ones that preserve vector length and the angles between vectors. Unitary transformations are precisely the set of transformations to an orthonormal set of basis vectors. When we transform to a non-orthonormal basis, the trajectory is sheared and stretched. Thus, in the eigenvector basis (meaning that we plot the eigenvector amplitudes on orthogonal axes), the

[5] Mathematically, to represent a vector $\mathbf{v}$ as a weighted sum of non-orthonormal basis vectors, the weight of one basis vector $\mathbf{e}_i$ depends on all the other basis vectors: one finds the direction that is orthogonal to all other basis vectors $\mathbf{e}_j$ for $j \neq i$, lets $\mathbf{l}_i$ (the left eigenvector corresponding to $\mathbf{e}_i$) be a vector in that direction with length such that $\mathbf{l}_i \cdot \mathbf{e}_i = 1$, and then takes the dot product of $\mathbf{v}$ with $\mathbf{l}_i$ to obtain the weight of $\mathbf{e}_i$. If $\mathbf{E}$ is the matrix whose columns are the eigenvectors, then the left eigenvector corresponding to the j^{th} column is found as the j^{th} row of $\mathbf{E}^{-1}$.

trajectory of Fig. S1 would become a trajectory that monotonically decays from an initially very large vector length, yet in the original basis (the firing rates) the trajectory is that of Fig. S1B and Fig. 1, a transient increase in firing rates followed by their decay, starting from initially relatively small firing rates. When we transform to an orthonormal basis, we do not stretch or shear the trajectory, it keeps exactly the same geometric structure, we only do a rigid rotation of the coordinates in which we view the trajectory (*e.g.* the sum and difference Schur basis vectors in Fig. S1B are a rigid rotation of the unit-length vectors along the r_e and r_i axes, which were the original basis vectors; the trajectory has the same size and shape relative to either basis, it is only rigidly rotated when the basis is changed).

Thus, if we wish to understand the changes in firing rate (the original basis), we do well to restrict ourselves to basis sets that preserve the size and shape of the trajectory, that is, to orthonormal basis sets or to unitary transformation. A matrix $\mathbf{M}$ is particularly simple in a basis in which it is diagonal, because that means each basis vector behaves independently of all others. But if $\mathbf{M}$ is non-normal, it cannot be diagonalized by a unitary transformation – it is diagonalized by the basis of eigenvectors, with the eigenvalues on the diagonal, but the eigenvectors are not orthogonal. How close to diagonal can we make the matrix by transformation to an orthogonal basis? The answer, given by the *Schur decomposition*, is that we can make the matrix upper triangular, with the eigenvalues on the diagonal and all other nonzero entries above the diagonal; this matrix will be diagonal (no nonzero entries above the diagonal) if and only if the matrix is normal [Horn and Johnson 1985].[6]

We interpret the Schur decomposition as follows. The strictly upper triangular part of the matrix (excluding the diagonal) corresponds to a strictly feedforward hierarchy of connections: connectivity flows from node j to node i only for $j > i$. The diagonal entries correspond to recurrent connectivity: node i connects to itself with a strength corresponding to an eigenvalue. In the transformed orthonormal basis in which $\mathbf{M}$ is upper triangular, each node corresponds to an activity pattern. Thus, non-normal matrices, in addition to the recurrent connectivity represented by the eigenvalues, have a hidden feedforward connectivity pattern between activity patterns, which results in amplification not predicted by the eigenvalues. In essence, the hidden dependency between the eigenvectors represented by their overlaps (nonzero dot products) is transformed into an explicit dependency (feedforward connections between orthogonal basis patterns). The purely feedforward nature of the connectivity also makes computation of the dynamics tractable (section S3.4).

For the generic case in which a matrix has a complete basis of eigenvectors, a Schur

[6] The Schur Decomposition should not be confused with the Jordan normal form of a matrix. The Jordan normal form involves non-unitary transformations, and is diagonal for any matrix, non-normal or normal, with a complete basis of eigenvectors. It has nonzero entries above the diagonal only for matrices that are missing one or more eigenvectors. The Schur Decomposition involves only unitary transformations, and is diagonal only for normal matrices; it has nonzero entries above the diagonal for all non-normal matrices.

decomposition is found by transforming to an orthogonal basis obtained by Gram-Schmidt orthonormalization of the eigenvector basis. A problem with the Schur decomposition is that it is not unique. For a non-normal matrix, each ordering of the non-orthogonal eigenvectors may lead, under the Gram-Schmidt orthonormalization process, to a distinct orthogonal basis. Since there are $N!$ possible orderings of the eigenvectors, a non-normal matrix may have $N!$ distinct Schur decompositions (not counting decompositions that differ only by a reordering of the orthonormal basis vectors). Thus, we cannot describe a unique feedforward structure between activity patterns that characterizes a given matrix.

In reality the set of Schur bases may be more restricted. For example, for the $2N \times 2N$ matrix $\mathbf{W} = \begin{pmatrix} \mathbf{W}_E & -\mathbf{W}_I \\ \mathbf{W}_E & -\mathbf{W}_I \end{pmatrix}$ studied in section S1.2, if $\mathbf{W}_E - \mathbf{W}_I$ is normal, the N eigenvectors $\mathbf{p}_i^{D+}$ with eigenvalues λ_i^D (which are eigenvalues of $\mathbf{W}_E - \mathbf{W}_I$) are mutually orthogonal. The other N eigenvectors correspond to eigenvalues of 0. The $\mathbf{p}_i^{D+}$ eigenvectors, which are of the form $\begin{pmatrix} \mathbf{q} \\ \mathbf{q} \end{pmatrix}$ for some vectors q, are analogous to the $\mathbf{p}^+$ eigenvector, $\propto \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, with eigenvalue $w_+ = w_E - w_I$ in the 2×2 case of Fig. 1 of the main text. The eigenvectors corresponding to the 0 eigenvalues are analogous to the very similar eigenvector proportional to $\begin{pmatrix} k \\ 1 \end{pmatrix}$ with eigenvalue 0 in the 2×2 case. They also can be mutually orthogonal, but they are not very different from, and not orthogonal to, the $\mathbf{p}_i^{D+}$ eigenvectors. If in constructing the Schur basis we start with the $\mathbf{p}_i^{D+}$ vectors, they will become part of the Schur basis in a manner that does not depend on their ordering. The remaining Schur basis vectors will all be perfect difference vectors, *i.e.* of the form $\begin{pmatrix} \mathbf{v} \\ -\mathbf{v} \end{pmatrix}$ for some orthonormal set of N-dimensional vectors $\mathbf{v}$, because these are the only vectors that are orthonormal to a complete basis of perfect sum vectors (the $\mathbf{p}_i^{D+}$), i.e. vectors of the form $\begin{pmatrix} \mathbf{q} \\ \mathbf{q} \end{pmatrix}$. Thus, if we start with the sum vectors $\mathbf{p}_i^{D+}$, the only ambiguity in the Schur basis will be the choice of orthonormal basis for the space of difference vectors.

Although the Schur decomposition is not uniquely specified, we can uniquely characterize the overall strength of the feedforward connectivity of a matrix. All the different Schur decompositions of a matrix are related to one another by unitary transformations. The sum of the absolute squares of all of the elements of $\mathbf{M}$ is a unitary invariant (unchanged by unitary transformations of $\mathbf{M}$, and thus identical for all Schur decompositions of $\mathbf{M}$), and is equal to $\mathrm{Tr}\,\mathbf{M}\mathbf{M}^\dagger$, where Tr is the trace, which in turn is equal to the sum of the squares of the singular values $\sigma_a^{\mathbf{M}}$ of $\mathbf{M}$. The eigenvalues of $\mathbf{M}$ are also unitary invariants, and so in particular the sum of the absolute squares of the eigenvalues of $\mathbf{M}$, $\sum_a |\beta_a^M|^2$, is a unitary invariant. But since all Schur decompositions have the eigenvalues on the diagonal, this is the sum of the absolute squares of the diagonal elements of any Schur decomposition of $\mathbf{M}$. Thus, the sum of the absolute squares of the off-diagonal or feedforward elements of any Schur decomposition of $\mathbf{M}$, as a proportion of the sum of the absolute squares of all of the elements, is $f^{\mathbf{M}} = \left(\mathrm{Tr}\,\left(\mathbf{M}\mathbf{M}^\dagger\right) - \sum_a |\beta_a^M|^2\right)/\mathrm{Tr}\,\left(\mathbf{M}\mathbf{M}^\dagger\right) = 1 - \frac{\sum_a |\beta_a^M|^2}{\sum_a (\sigma_a^{\mathbf{M}})^2}$. The size of $f^{\mathbf{M}}$ is a measure of the strength of hidden feedforward connectivity and thus of the strength of transient response and of the non-normality of the matrix. Note that, in the special case that all of the eigenvalues of $\mathbf{M}$ are real, $\sum_a |\beta_a^M|^2 = \mathrm{Tr}\,\mathbf{M}^2$ and $f^{\mathbf{M}} = \mathrm{Tr}\,\left(\mathbf{M}\mathbf{M}^\dagger - \mathbf{M}^2\right)/\mathrm{Tr}\,\left(\mathbf{M}\mathbf{M}^\dagger\right)$. For $\mathbf{W}$ given by the orientation-specific connectivity matrix used in Fig. 3 in the main text (based on 32 × 32 grids of E cells and of I cells), $f^{\mathbf{M}} = 0.55$, that is, 55% of the total power in the matrix driving the dynamics is in the feedforward links.

S3.3 The Schur Decomposition for the General 2 × 2 case

We consider the general 2 × 2 connection matrix $\mathbf{W} = \begin{pmatrix} w_{EE} & -w_{EI} \\ w_{IE} & -w_{II} \end{pmatrix}$. We assume w_{EI} is nonzero. Our results will otherwise be valid for any 2 × 2 matrix. However, we will regard w_{EE}, w_{EI}, w_{IE}, w_{II} as all positive, and refer to modes as sum or difference modes based on this assumption.

We define

$$\begin{aligned} X &= \frac{w_{EE} + w_{II}}{2w_{EI}} \\ Y &= \sqrt{X^2 - w_{IE}/w_{EI}} \end{aligned} \tag{S21}$$

and also $Z = \frac{w_{EE} - w_{II}}{2w_{EI}}$. We note X and Z are real, $X > 0$, and Y is either real and positive or else is pure imaginary. We also note that $|\Re(Y)| < X$ where $\Re(Y)$ is the real part of Y. The eigenvectors of $\mathbf{W}$ are $\mathbf{e}_\pm = \frac{1}{\sqrt{1+|x_\pm|^2}}\begin{pmatrix} 1 \\ x_\pm \end{pmatrix}$ where $x_\pm = X \pm Y$, and the corresponding eigenvalues are $\lambda_\pm = w_{EI}(Z \mp Y)$.

Since $X > 0$ and $|\Re(Y)| < X$, the real parts of x_+ and x_- are both positive. This means that both eigenvectors are sum modes, in the generalized sense that both entries have real parts of the same sign.

To make a Schur basis, we start with $\mathbf{e}_+$, and construct a second vector orthonormal to it, which we'll call $\mathbf{q}$. We can immediately see that to have $\mathbf{q} \cdot \mathbf{e}_+ = 0$, we must have $\mathbf{q} = \pm\frac{1}{\sqrt{1+|x_+|^2}}\begin{pmatrix} x_+^* \\ -1 \end{pmatrix}$ (the * indicates complex conjugate; recall that, for possibly complex vectors, $\mathbf{q}\cdot\mathbf{e}_+ = \mathbf{q}^\dagger\mathbf{e}_+$ where † means conjugate transpose). We choose the + of the ± choice. Note that $\mathbf{q}$ is a difference vector in the generalized sense that its two entries have real parts of opposite signs. We can also write this as $\mathbf{q} = \frac{\mathbf{e}_- - (\mathbf{e}_- \cdot \mathbf{e}_+)\mathbf{e}_+}{\sqrt{1 - |\mathbf{e}_- \cdot \mathbf{e}_+|^2}}$, since this is the Gram-Schmidt formula for finding a unit-length vector in the $\mathbf{e}_+/\mathbf{e}_-$ plane that is orthogonal to $\mathbf{e}_+$, and the sign turns out to agree with the + choice for $\mathbf{q}$ above.

Then we can compute $\mathbf{Wq} = \frac{\lambda_- \mathbf{e}_- - \lambda_+ (\mathbf{e}_- \cdot \mathbf{e}_+)\mathbf{e}_+}{\sqrt{1 - |\mathbf{e}_- \cdot \mathbf{e}_+|^2}} = \lambda_- \mathbf{q} + \frac{(\lambda_- - \lambda_+)(\mathbf{e}_- \cdot \mathbf{e}_+)}{\sqrt{1 - |\mathbf{e}_- \cdot \mathbf{e}_+|^2}} \mathbf{e}_+$. Letting $\beta = \frac{(\lambda_- - \lambda_+)(\mathbf{e}_- \cdot \mathbf{e}_+)}{\sqrt{1 - |\mathbf{e}_- \cdot \mathbf{e}_+|^2}}$, we have $\mathbf{We}_+ = \lambda_+ \mathbf{e}_+$, $\mathbf{Wq} = \lambda_- \mathbf{q} + \beta \mathbf{e}_+$. In other words, in the Schur basis $(\mathbf{e}_+, \mathbf{q})$, $\mathbf{W}$ takes the upper triangular form

$$\begin{aligned} \mathbf{W} &= \begin{pmatrix} \lambda_+ & \beta \\ 0 & \lambda^- \end{pmatrix} \text{ with} \\ \beta &= \frac{(\lambda_- - \lambda_+)(\mathbf{e}_- \cdot \mathbf{e}_+)}{\sqrt{1 - |\mathbf{e}_- \cdot \mathbf{e}_+|^2}} \end{aligned} \tag{S22}$$

β is the effective feedforward weight from the difference mode $\mathbf{q}$ to the sum mode $\mathbf{e}_+$.

Note that this definition of β defines a Schur decomposition for *any* 2×2 matrix with distinct eigenvectors $\mathbf{e}_i$ and corresponding eigenvalues λ_i, $i = 1, 2$. We didn't use any specific information about the structure of the eigenvectors and eigenvalues to compute this. Thus, for any 2×2 matrix, the feedforward weight can become large when $|\mathbf{e}_1 \cdot \mathbf{e}_2|$ is close to one, i.e. when the angle between the eigenvectors is small. On the other hand, it becomes zero when the matrix is normal, so that $|\mathbf{e}_1 \cdot \mathbf{e}_2| = 0$. (It also becomes zero if $\lambda_1 = \lambda_2$, but this also means that the matrix is normal, because, assuming that there are two distinct eigenvectors, then when the two eigenvalues are equal, any linear combination of the two eigenvectors is also an eigenvector so we can always choose the eigenvectors to be orthonormal.)

Now, for our particular matrix, we wish to compute β. To begin, we compute $|\beta|^2$ by using the fact, discussed in the last paragraph of the previous section, that the sum of the absolute squares of the matrix elements is a unitary invariant, and hence is the same in the original basis as in the Schur basis. Therefore,

$$|\beta|^2 = w_{EE}^2 + w_{EI}^2 + w_{IE}^2 + w_{II}^2 - |\lambda_+|^2 - |\lambda_-|^2 \tag{S23}$$

When the eigenvalues are real ($|Y|^2 = X^2 - w_{IE}/w_{EI}$), the sum of their absolute squares is $2w_{EI}^2 \left(Z^2 + Y^2\right) = w_{EE}^2 + w_{II}^2 - 2w_{IE}w_{EI}$, so

$$|\beta|^2 = \left(w_{EI} + w_{IE}\right)^2 \qquad \text{(eigenvalues real)} \tag{S24}$$

When the eigenvalues are complex ($|Y|^2 = w_{IE}/w_{EI} - X^2$), the sum of their absolute squares is $2w_{EI}^2 \left(Z^2 + |Y|^2\right) = -2w_{EE}w_{II} + 2w_{IE}w_{EI}$ so

$$|\beta|^2 = \left(w_{EI} - w_{IE}\right)^2 + \left(w_{EE} + w_{II}\right)^2 \qquad \text{(eigenvalues complex)} \tag{S25}$$

Note that, when eigenvalues are real, β is a measure of the deviation of $\mathbf{W}$ from symmetry (a symmetric matrix would have $w_{IE} = -w_{EI}$), while when eigenvalues are complex, β is a measure of the deviation of $\mathbf{W}$ from antisymmetry (an antisymmetric matrix would have $w_{EE} = w_{II} = 0$ and $w_{IE} = w_{EI}$). Symmetric and antisymmetric real matrices are both normal matrices with real or imaginary eigenvalues, respectively. Thus, β could be thought of as a measure of distance from these "canonical" normal matrix classes whose eigenvalues are real or imaginary, respectively.[7]

To determine the phase of β, we must explicitly compute its value. We find[8] that, in the orthonormal Schur basis $\{\mathbf{e}_+, \mathbf{q}\}$:

$$\begin{aligned} \beta &= w_{EI} + w_{IE}, \text{ Y real;} && \text{(S26)} \\ &= -\frac{1}{2w_{EI}}\Big((w_{EE}+w_{II})\sqrt{4w_{EI}w_{IE} - (w_{EE}+w_{II})^2} \\ &\quad + i\left(2w_{EI}(w_{EI} - w_{IE}) + (w_{EE}+w_{II})^2\right)\Big), \text{ Y imaginary} && \text{(S27)} \end{aligned}$$

Direct computation confirms that $|\beta|^2$ is then as given by Eqs. S24-S25.

We have noted (section S1.2) that the solution to the dynamical equation for $\mathbf{r}$ with time-independent input $\mathbf{I}$ can be written in terms of the matrix $e^{-(\mathbf{1}-\mathbf{W})t/\tau} = e^{-t/\tau}e^{\mathbf{W}t/\tau}$ as $\mathbf{r}(t) = e^{-(\mathbf{1}-\mathbf{W})t/\tau}\mathbf{r}(0) + (\mathbf{1} - e^{-(\mathbf{1}-\mathbf{W})t/\tau})(\mathbf{1}-\mathbf{W})^{-1}\mathbf{I}$ (or with time-dependent input $\mathbf{I}(t)$, $\mathbf{r}(t) = e^{-(\mathbf{1}-\mathbf{W})t/\tau}\mathbf{r}(0) + \frac{1}{\tau}\int_0^t dt' e^{-(\mathbf{1}-\mathbf{W})(t-t')/\tau}\mathbf{I}(t')$). We can compute this matrix:

$$e^{-(\mathbf{1}-\mathbf{W})t/\tau} = e^{-t/\tau}\begin{pmatrix} e^{\lambda_+ t/\tau} & \beta\frac{e^{\lambda_- t/\tau} - e^{\lambda_+ t/\tau}}{\lambda_- - \lambda_+} \\ 0 & e^{\lambda_- t/\tau} \end{pmatrix} \qquad \text{(S28)}$$

[7] We thank Yashar Ahmadian for pointing out to us this simple derivation and interpretation.

[8] First we note that $(\lambda_- - \lambda_+) = 2w_{EI}Y$. Next, $\mathbf{e}_- \cdot \mathbf{e}_+ = \frac{1+x_-^* x_+}{\sqrt{(1+|x_+|^2)(1+|x_-|^2)}}$. Write this as A/B where A, the numerator, may be complex, and B is real. Then the term $\frac{(\mathbf{e}_- \cdot \mathbf{e}_+)}{\sqrt{1-|\mathbf{e}_- \cdot \mathbf{e}_+|^2}} = \frac{A}{B\sqrt{1-|A|^2/B^2}} = \frac{A}{\sqrt{B^2-|A|^2}}$. Now $A = 1 + (X - Y^*)(X+Y) = 1 + X^2 - |Y|^2 + X(Y - Y*)$, with $X(Y - Y*) = 0$ if Y is real and $= 2Y$ if Y is imaginary. After some manipulation, we find that $\sqrt{B^2 - |A|^2} = 2|Y|$.

Putting this all together, we find $\beta = w_{EI}(Y/|Y|)(1 + X^2 - |Y|^2 + X(Y - Y*))$. $Y/|Y| = 1$, Y real; $= i$, Y imaginary (where $i = \sqrt{-1}$). So we arrive at

$$\begin{aligned} \beta &= w_{EI}(1 + X^2 - Y^2), \qquad Y \text{ real} \\ &= iw_{EI}(1 + X^2 - |Y|^2 + 2XY), \qquad Y \text{ imaginary} \end{aligned}$$

For the case Y real, which means $\left(\frac{w_{EE}+w_{II}}{2}\right)^2 \geq w_{EI}w_{IE}$, we have $X^2 - Y^2 = X^2 - (X^2 - w_{IE}/w_{EI}) = w_{IE}/w_{EI}$. Thus $\beta = w_{EI}(1 + w_{IE}/w_{EI}) = w_{EI} + w_{IE}$. In other words, the feedforward weight is just given by the sum of the two feedback inhibition terms, so if feedback inhibition is strong, there is strong balanced amplification.

For the case Y imaginary, which means $w_{EI}w_{IE} > \left(\frac{w_{EE}+w_{II}}{2}\right)^2$, $X^2 - |Y|^2 = X^2 - (w_{IE}/w_{EI} - X^2) = 2X^2 - w_{IE}/w_{EI}$. Thus $\beta = iw_{EI}(1 - w_{IE}/w_{EI} + 2X(X+Y)) = i(w_{EI} - w_{IE} + 2w_{EI}X(X+Y))$. The expression simplifies somewhat if we define $\xi = \sqrt{X^2/(w_{IE}/w_{EI})} = \frac{w_{EE}+w_{II}}{2\sqrt{w_{EI}w_{IE}}}$, and note Y is imaginary if and only if $\xi < 1$. Then $X = \sqrt{w_{IE}/w_{EI}}\xi$, $Y = i\sqrt{(w_{IE}/w_{EI})(1-\xi^2)}$, and $X(X+Y) = (w_{IE}/w_{EI})(\xi^2 + i\xi\sqrt{1-\xi^2})$. Thus we can write $\beta = i(w_{EI} - w_{IE} + 2w_{IE}\xi(\xi + i\sqrt{1-\xi^2})$. Substituting back for ξ and simplifying yields Eq. S27.

The term multiplying β,

$$g_{\lambda_+;\lambda_-}(t) = \frac{e^{(\lambda_- -1)t/\tau} - e^{(\lambda_+ -1)t/\tau}}{\lambda_- - \lambda_+} \tag{S29}$$

is the generalization of the pulse function $g_{w_+}(t)$ that we saw in section S1.2. $g_{w_+}(t)$ is this term for the case $\lambda_- = 0$, $\lambda_+ = -w_+$, which was true for that model. $g_{\lambda_+;\lambda_-}(t)$ just arises from concatenating the filter $\frac{1}{\tau}e^{(\lambda_- -1)t/\tau}$ that the difference mode $\mathbf{q}$ applies to its input, with the filter $\frac{1}{\tau}e^{(\lambda_+ -1)t/\tau}$ that the sum mode $\mathbf{e}_+$ applies to its input, as explained in section S3.4. In the limit $\lambda_+ \to \lambda_-$, $g_{\lambda_+;\lambda_-}(t) \to (t/\tau)e^{\lambda_- t/\tau}$.

S3.4 Solution of the Dynamics in a Schur Basis, and Coexistence of Hebbian and Balanced amplification

The dynamics can, at least in principle, be simply solved in a Schur Basis. Let the eigenvalues of $\mathbf{W}$ be λ_i. Let the Schur basis patterns be $\mathbf{p}_i$, with associated eigenvalues λ_i and amplitudes $r_i(t)$, and feedforward weights w_{ij}^{FF} between the patterns with $i < j$. Then the i^{th} pattern simply filters (convolves) its input with its filter, $f_i(t) = \frac{1}{\tau}e^{-(1-\lambda_i)t/\tau}$ (where convolution of $I(t)$ with f_i is $f_i \star I(t) = \int_0^t dt'\, f_i(t-t')I(t')$), and any initial condition $r_i(0)$ is multiplied by $\tau f_i(t)$. The sum of these gives the activity $r_i(t)$ of the i^{th} pattern at time t. This is the same prescription used to solve the dynamics in the eigenvector basis (or in any basis, if the eigenvalue is replaced by the self-connection in that basis).

There are two differences from the eigenvector basis. First, in the Schur basis, the inputs to patterns include inputs via feedforward links from other patterns, while in the eigenvector basis there are no connections between patterns. If there were loops in the connectivity, this prescription would not suffice to write down a solution. To compute a neuron's response would require concatenating infinite loops of exponential filters. But because the connectivity is purely feedforward, this prescription yields a finite solution.

Second, in the eigenvector basis, the components r_i of the patterns and I_i of the inputs must be found by dot product of the corresponding *left* eigenvectors with the rate vector $\mathbf{r}$ or input vector $\mathbf{I}$, and the left eigenvectors are not equal to the right eigenvectors when the eigenvectors are not mutually orthogonal (the left eigenvectors are found as the rows of the inverse of the matrix whose columns are the eigenvectors). This is why the components of $\mathbf{r}$ in the eigenvector basis are so nonintuitively related to $\mathbf{r}$ for a non-normal matrix. For an orthonormal basis set, the components are found simply as dot products with the basis patterns.

How is a solution written down in the Schur basis? One begins with patterns receiving no feedforward input from other patterns, and then propagates the activity forward through the feedforward tree of patterns. This continues until reaching the end of all chains and branches of the tree. The total input to pattern i at time t is $I_i^{total}(t) = I_i(t) + \sum_{j>i} w_{ij} r_j(t)$

where $r_j(t)$ is the activity of node j. Then node i's activity $r_i(t) = f_i \star I_i^{total}(t) + r_i(0)\tau f_i(t)$ where $\star$ indicates convolution.

Alternatively, one can compute the matrix $e^{-(\mathbf{1}-\mathbf{W})t/\tau}$, from which the solution can be computed as $\mathbf{r}(t) = e^{-(\mathbf{1}-\mathbf{W})t/\tau}\mathbf{r}(0) + \frac{1}{\tau}\int_0^t dt' e^{-(\mathbf{1}-\mathbf{W})(t-t')/\tau}\mathbf{I}(t')$. The element $\left(e^{-(\mathbf{1}-\mathbf{W})t/\tau}\right)_{ij}$ is computed as follows. It is τ times the sum, over all feedforward paths from j to i, of the following for each path: the concatenation (convolutions) of the filters for each pattern in the path (including j and i), multiplied by the product of all the feedforward weights along the path. If $i = j$ (diagonal elements), this is just the filter for the node j; if there are no feedforward paths, the element is 0.

As a simple example: for the case in which a difference mode $\mathbf{p}_-$ corresponding to eigenvalue λ_- sends a feedforward connection w_{FF} to a sum mode $\mathbf{p}_+$ corresponding to eigenvalue λ_+, then the result of concatenating the two exponential filters of $\mathbf{p}_-$ and $\mathbf{p}_+$, multiplied by w_{FF}, is $\frac{1}{\tau}w_{FF}g_{\lambda_+;\lambda_-}(t)$ (Eq. S29), the pulse function that describes the response of the sum mode to input to the difference mode.

We have focused on the case in which there are no positive eigenvalues, so that there is no Hebbian slowing and the only mechanism of amplification is balanced amplification. However, it is important to point out that balanced amplification and amplification by slowing down will coexist if there are eigenvalues of $\mathbf{W}$, λ_i, with positive real part but in the stable regime, $0 < \Re(\lambda_i) < 1$. The basic mechanism is as just described and as illustrated in Fig. 4 in the main text: the eigenvalues control the dynamics of each pattern, while the feedforward connections transmit between them. If a pattern's dynamics are slowed by its eigenvalues, this will affect both its integration of any input it receives, including feedforward input, and the time course of any feedforward input it sends to other patterns.

S3.5 The general case of distinct $\mathbf{W}_{EE}$, $\mathbf{W}_{EI}$, $\mathbf{W}_{IE}$, and $\mathbf{W}_{II}$

In the general case in which $\mathbf{W} = \begin{pmatrix} \mathbf{W}_{EE} & -\mathbf{W}_{EI} \\ \mathbf{W}_{IE} & -\mathbf{W}_{II} \end{pmatrix}$, with each submatrix $\mathbf{W}_{XY}$ having non-negative entries, we cannot form a general solution or make a general argument as to the size or structure of the balanced amplification that will arise. However, we can make a number of more limited arguments to suggest that, when recurrent excitation is large but is balanced by large feedback inhibition, we should expect large balanced amplification, with the dominant feedforward links being from difference modes to sum modes. In addition, for many simple connectivities (translation-invariant connectivity), feedforward chains are only a single link long.

The dynamics is driven by $\mathbf{W} - \mathbf{1}$, but the identity matrix remains the identity in any basis, subtracting 1 from each diagonal element and making no contribution to feedforward weights. So, we focus on $\mathbf{W}$, knowing that we must simply subtract 1 from each diagonal element. We think of $\mathbf{W}$ as the mean connectivity matrix in the linear model, which defines the probabilities from which the sparse random connectivity of the spiking model was drawn. However, some of our arguments would also apply to a sparse random connectivity matrix.

S3.5.1 Balanced Networks Should Have Large Feedforward Weights

Here we make two arguments. The first is that presented in the main text, which we slightly amplify here. The sum of the absolute squares of the matrix entries of any matrix is a unitary invariant (invariant under unitary transformations). Since both excitation and inhibition are strong, this sum is large for $\mathbf{W}$. In the basis of a Schur decomposition, this is equal to the sum of the absolute squares of the eigenvalues plus the sum of the absolute squares of the effective feedforward connections. If we define balanced inhibition to mean that all of the eigenvalues are small, then it follows that there will be large effective feedforward connections and therefore large balanced amplification. However, some connectivities that might otherwise be interpreted as "balanced inhibition" might produce eigenvalues with large negative real parts and/or large imaginary parts, and conceivably these eigenvalues could be large and/or numerous enough to account for most of the sum, leaving only relatively small feedforward connections; we cannot rule this out or state conditions under which it will or will not happen.

Second, we compute the invariant $f^{\mathbf{W}}$ defined above in section S3.2, which measures the relative strength of the effective feedforward connectivity and thus of the balanced amplification, in a special case: we assume that all of the eigenvalues of $\mathbf{W}$ are real. In this case, $f^{\mathbf{W}} = \mathrm{Tr}\left(\mathbf{W}\mathbf{W}^\dagger - \mathbf{W}^2\right)/\mathrm{Tr}\left(\mathbf{W}\mathbf{W}^\dagger\right)$. Let $\mathbf{W}^A_{EE} = (\mathbf{W}^\dagger_{EE} - \mathbf{W}_{EE})/2$ and $\mathbf{W}^A_{II} = (\mathbf{W}^\dagger_{II} - \mathbf{W}_{II})/2$ be the antisymmetric parts of $\mathbf{W}_{EE}$ and $\mathbf{W}_{II}$ respectively. Then we can compute

$$f^{\mathbf{W}} = \frac{\mathrm{Tr}\left(\mathbf{W}_{EI}\mathbf{W}^\dagger_{EI} + \mathbf{W}_{IE}\mathbf{W}^\dagger_{IE} + \mathbf{W}_{EI}\mathbf{W}_{IE} + \mathbf{W}_{IE}\mathbf{W}_{EI} + 2\mathbf{W}_{EE}\mathbf{W}^A_{EE} + 2\mathbf{W}_{II}\mathbf{W}^A_{II}\right)}{\mathrm{Tr}\left(\mathbf{W}_{EI}\mathbf{W}^\dagger_{EI} + \mathbf{W}_{IE}\mathbf{W}^\dagger_{IE} + \mathbf{W}_{EE}\mathbf{W}^\dagger_{EE} + \mathbf{W}_{II}\mathbf{W}^\dagger_{II}\right)} \tag{S30}$$

In particular, if all of the submatrices $\mathbf{W}_{XY}$ are symmetric, this becomes

$$f^{\mathbf{W}} = \frac{\mathrm{Tr}\left((\mathbf{W}_{EI} + \mathbf{W}_{IE})^2\right)}{\mathrm{Tr}\left(\mathbf{W}^2_{EI} + \mathbf{W}^2_{IE} + \mathbf{W}^2_{EE} + \mathbf{W}^2_{II}\right)} \tag{S31}$$

Thus, if feedback inhibitory terms $\mathbf{W}_{IE}$ and $\mathbf{W}_{EI}$ are at least comparable in size to the recurrent terms $\mathbf{W}_{EE}$ and $\mathbf{W}_{II}$, as they must be for inhibition to balance excitation, then the numerator of $f^{\mathbf{W}}$ should be comparable to the denominator and $f^{\mathbf{W}}$ should be significantly nonzero. This becomes particularly clear in the symmetric case, in which $f^{\mathbf{W}}$ becomes essentially a measure of the size of the feedback inhibition relative to the overall connectivity, similar to our finding in the case in which the different submatrices can be simultaneously diagonalized.

S3.5.2 The Dominant Feedforward Links in Biological Connection Matrices Should Be From Difference Modes to Sum Modes

Here we argue that the overall structure of $\mathbf{W}$, namely its two nonnegative submatrices on the left and two nonpositive submatrices on the right, causes the dominant feedforward links to be from difference modes to sum modes. We take $\mathbf{W}$ to be $2N$-dimensional, so that each submatrix is N-dimensional.

Our argument is based on the following fact: in any $2N$-dimensional orthonormal basis $\{\mathbf{f}_i\}$, $i = 1, \ldots, 2N$, $\mathbf{W}$ can be written $\mathbf{W} = \sum_{ij} W^f_{ij} \mathbf{f}_i \mathbf{f}_j^\dagger$ where $W^f_{ij} = \mathbf{f}_i^\dagger \mathbf{W} \mathbf{f}_j$. W^f_{ij} is the i-j^{th} element of $\mathbf{W}$ when it is expressed in the $\{\mathbf{f}_i\}$ basis. Each term of the form $W^f_{ij} \mathbf{f}_i \mathbf{f}_j^\dagger$ takes input in the $\mathbf{f}_j$ direction, multiplies it by W_{ij}, and converts it to output in the $\mathbf{f}_i$ direction. That is, it can be thought of as a link from pattern $\mathbf{f}_j$ to pattern $\mathbf{f}_i$ with weight W_{ij}.

Let $\{\mathbf{e}_i\}$ be any set of N orthonormal N-dimensional basis vectors. Form the $2N$-dimensional orthonormal basis consisting of the sum vectors $\mathbf{e}_i^+ = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_i \\ \mathbf{e}_i \end{pmatrix}$ and the difference vectors $\mathbf{e}_i^- = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_i \\ -\mathbf{e}_i \end{pmatrix}$. We then can write

$$\mathbf{W} = \sum_{ij} W^{++}_{ij} \mathbf{e}_i^+ \mathbf{e}_j^{+\dagger} + \sum_{ij} W^{--}_{ij} \mathbf{e}_i^- \mathbf{e}_j^{-\dagger} + \sum_{ij} W^{+-}_{ij} \mathbf{e}_i^+ \mathbf{e}_j^{-\dagger} + \sum_{ij} W^{-+}_{ij} \mathbf{e}_i^- \mathbf{e}_j^{+\dagger} \tag{S32}$$

We define $\mathbf{W}^{++} = \sum_{ij} W^{++}_{ij} \mathbf{e}_i^+ \mathbf{e}_j^{+\dagger}$ and similarly for the other three terms. $\mathbf{W}^{++}$ represents links from sum patterns to sum patterns; $\mathbf{W}^{--}$ represents links from difference patterns to difference patterns; $\mathbf{W}^{+-}$ represents links from difference patterns to sum patterns; and $\mathbf{W}^{-+}$ represents links from sum patterns to difference patterns.

By considering the structure of individual terms in the sums, it is easy to see that these matrices have the form

$$\begin{aligned} \mathbf{W}^{++} &= \begin{pmatrix} \mathbf{A} & \mathbf{A} \\ \mathbf{A} & \mathbf{A} \end{pmatrix} \\ \mathbf{W}^{--} &= \begin{pmatrix} \mathbf{B} & -\mathbf{B} \\ -\mathbf{B} & \mathbf{B} \end{pmatrix} \\ \mathbf{W}^{+-} &= \begin{pmatrix} \mathbf{C} & -\mathbf{C} \\ \mathbf{C} & -\mathbf{C} \end{pmatrix} \\ \mathbf{W}^{-+} &= \begin{pmatrix} \mathbf{D} & \mathbf{D} \\ -\mathbf{D} & -\mathbf{D} \end{pmatrix} \end{aligned} \tag{S33}$$

for some submatrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$. From the fact that $\mathbf{W} = \mathbf{W}^{++} + \mathbf{W}^{--} + \mathbf{W}^{+-} + \mathbf{W}^{-+}$, we can find

$$\begin{aligned}
\mathbf{A} &= \frac{1}{4}\left(\mathbf{W}_{EE} - \mathbf{W}_{EI} + \mathbf{W}_{IE} - \mathbf{W}_{II}\right) && \text{(S34)}\\
\mathbf{B} &= \frac{1}{4}\left(\mathbf{W}_{EE} + \mathbf{W}_{EI} - \mathbf{W}_{IE} - \mathbf{W}_{II}\right) && \text{(S35)}\\
\mathbf{C} &= \frac{1}{4}\left(\mathbf{W}_{EE} + \mathbf{W}_{EI} + \mathbf{W}_{IE} + \mathbf{W}_{II}\right) && \text{(S36)}\\
\mathbf{D} &= \frac{1}{4}\left(\mathbf{W}_{EE} - \mathbf{W}_{EI} - \mathbf{W}_{IE} + \mathbf{W}_{II}\right) && \text{(S37)}
\end{aligned}$$

$\mathbf{C}$ is the average of the four nonnegative submatrices of $\mathbf{W}$, so it is nonnegative and it will have large entries if $\mathbf{W}$ does. $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{D}$ all are averages of two of these submatrices plus the negatives of two others, meaning that $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{D}$ should be relatively small by some measure (for example, in the case of sparse random submatrices, then $\mathbf{C}$ would have leading eigenvalue of order N while $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{D}$ would have leading eigenvalue of order $\sqrt{N}$ [Rajan and Abbott 2006]).

Thus, the dominant contribution to $\mathbf{W}$ should be from $\mathbf{W}^{+-}$, which has the same structure of signs as $\mathbf{W}$, and which involves links from difference patterns to sum patterns. The contributions from $\mathbf{W}^{++}$, $\mathbf{W}^{--}$, and $\mathbf{W}^{-+}$ should be relatively small: these account for the differences between $\mathbf{W}_{EE}, \mathbf{W}_{EI}, \mathbf{W}_{IE}$ and $\mathbf{W}_{II}$, that is, for their deviations from their average, while $\mathbf{W}^{+-}$ accounts for their average. Since $\mathbf{W}^{+-}$ makes the dominant contribution to $\mathbf{W}$ and $\mathbf{W}$ overall is large (involving large recurrent excitation balanced by large feedback inhibition), we expect $\mathbf{W}$ to involve large balanced amplification dominantly involving feedforward links from difference-like patterns (in which inhibition and excitation largely or entirely have opposite signs) to sum-like patterns (in which they largely or entirely have the same sign).

We can see more generally the sources of other kinds of links from Eq. S33. $\mathbf{W}^{+-}$, and thus difference-to-sum links, account for equal strengths of E→E, E→I, I→E, and I→I connections. Sum-to-sum links ($\mathbf{W}^{++}$) add something of the same sign to all four kinds of connections; if positive, this makes excitatory connections stronger and inhibitory weaker, if negative it does the reverse. So sum-to-sum links create overall imbalances between excitatory and inhibitory connections. Difference-to-difference links ($\mathbf{W}^{--}$) make connections to E cells stronger and connections to I cells weaker, or vice versa, so these create overall imbalances between connections to E cells and connections to I cells. Finally, sum-to-difference links $\mathbf{W}^{-+}$ make same-type coupling (E→E and I→I) stronger and opposite-type (I→E and E→I) weaker, or vice versa, so these create overall imbalances between same-type connections and opposite-type connections. Thus, by looking at the strengths of the average and of each type of imbalance, one can obtain some sense of the strength of each type of link. This analysis is limited, because the sum and difference basis with which we're working is not likely to render a general $\mathbf{W}$ matrix upper triangular, *i.e.* the links will involve loops and not be strictly feedforward. The Schur basis, which is strictly feedforward except for the eigenvalues, will be a different basis, and we cannot predict exactly how this picture will translate into the Schur picture. We do suggest that, if certain kinds of links are especially strong or especially weak in the sum and difference basis, this is likely to also be evident in the feedforward links in the Schur picture. Thus, given that difference-to-sum links dominate the matrix, we expect the dominant feedforward links in the Schur basis to be from difference-like modes to sum-like modes.

S3.5.3 Translation-Invariant Connectivity Leads to Independent Two-by-Two Connection Matrices for Each Spatial Frequency

Third, consider the case in which the N-dimensional submatrices $\mathbf{W}_{EE}$, $\mathbf{W}_{EI}$, $\mathbf{W}_{IE}$, and $\mathbf{W}_{II}$ can all be simultaneously diagonalized in an orthonormal basis. In this case, we can show quite generally that the matrix breaks down into a set of N independent 2×2 submatrices, one for each eigenvector, so that feedforward chains can only have length one, between the two Schur vectors of a single 2×2 submatrix. We argue further that when recurrent excitation is large but is balanced by large feedback inhibition, there will be large amplification from difference modes to sum modes.

This assumption is not unreasonable for the mean connectivity. In particular, it will be true if the mean connectivity submatrix is translation-invariant, meaning that it looks the same at any spatial position or at any position in feature space on which the connectivity depends, so that translating a spatial or feature-space position by the same amount for all neurons will not change the connectivity between them. In this case, the submatrices can be simultaneously diagonalized by the Fourier transform, which represents a transformation to an orthonormal basis. If connections depend on spatial position and preferred orientation, they will be translation invariant if changing every neuron's position by 100 μm or changing every neuron's preferred orientation by 30^o will leave the connection matrix unchanged. This will be true if the connection strength between two neurons depend only on differences between their positions or features, if boundary effects are negligible, and if the different spatial or feature dimensions on which the connections depend are not coupled, so that changing one does not also cause changes in others. An example in which this last condition is violated, *i.e.* features are coupled, is Fig. 3: the orientation map is spread over space, so the preferred orientation of a neuron depends on its position. Thus, if all neurons are translated in space, this will also change their preferred orientations, and for most orientation maps the preferred orientations of different neurons will not change by the same amount, so the connectivity is not translation invariant. If instead we assume that each position contains cells of all preferred orientations, then connectivity that only depends on differences of position and of preferred orientation would be translation invariant. To the extent that this latter form of model is adequate to understand many features of V1, the conclusions drawn from considering the behavior of translation-invariant matrices will apply.

Let $\mathbf{e}_i$ be the orthonormal basis of the $N \times N$ subspace in which all of the submatrices are diagonal. Let $D_{EE}(i)$ be the eigenvalue of $\mathbf{W}_{EE}$ corresponding to $\mathbf{e}_i$, and similarly for the other submatrices. For a translation-invariant matrix in which connectivity depends only on space, i corresponds to a spatial frequency, and $D_{EE}(i)$ is the Fourier transform of the excitatory connectivity at frequency i. For a translation-invariant matrix that depends on multiple spatial or feature dimensions, i represents a particular set of frequencies, one for each dimension, and $D_{EE}(i)$ is the product of the Fourier transforms of the excitatory connectivity along each dimension at the corresponding frequency for that dimension.

Define orthonormal basis vectors of the full space by the excitatory cell vector $\mathbf{e}_i^E = \begin{pmatrix} \mathbf{e}_i \\ \mathbf{0} \end{pmatrix}$ and inhibitory cell vector $\mathbf{e}_i^I = \begin{pmatrix} \mathbf{0} \\ \mathbf{e}_i \end{pmatrix}$, where $\mathbf{0}$ is the N-dimensional vector of all 0's, and work in the basis $\{\mathbf{e}_1^E, \mathbf{e}_1^I, \mathbf{e}_2^E, \mathbf{e}_2^I, \ldots, \mathbf{e}_N^E, \mathbf{e}_N^I\}$. In this basis, the matrix $\mathbf{W}$ becomes a set of N 2×2 matrices arrayed along the diagonal, with the k^{th} such matrix corresponding to the basis vectors $\mathbf{e}_k^E, \mathbf{e}_k^I$ and being of the form $D(k) = \begin{pmatrix} D_{EE}(k) & -D_{EI}(k) \\ D_{IE}(k) & -D_{II}(k) \end{pmatrix}$. Thus, the dynamics break up into independent two-dimensional subspaces, one for each N-dimensional eigenvector. E and I amplitudes for a given eigenvector interact with one another by the corresponding 2×2 matrix, but do not interact with the amplitudes for any other eigenvector.

In section S3.3, we computed the Schur decomposition for this 2×2 matrix. We showed that, if all of the D_{XY}'s were positive, the Schur basis showed a feedforward connection of size β from a difference-like mode to a sum-like mode. Here, we cannot be certain that all the D_{XY}'s will be positive, but if the connection strengths decrease smoothly with distance (in all the dimensions on which they depend), then they are likely to be, particularly when they are large. We also showed (Eqs. S24-S25), on the assumption that the $D_{XY}(k)$ are real (as they will be *e.g.* if the submatrices $\mathbf{W}_{XY}$ are symmetric), that, when the eigenvalues of $D(k)$ are real, the feedforward connection strength is $\beta = D_{EI}(k) + D_{IE}(k)$; while when the eigenvalues are complex, $|\beta| = \left((D_{EE}(k) + D_{II}(k))^2 + (D_{EI}(k) - D_{IE}(k))^2\right)^{1/2}$. Assuming each submatrix individually has large elements, each of the D_{XY}'s must take large values for some k's (*e.g.*, the sum of the absolute squares of the $D_{EI}(k)$'s is equal to the sum of the absolute squares of the elements of $\mathbf{W}_{EI}$, etc.). If they are positive (for example, the Fourier transform of a Gaussian connectivity function is a Gaussian, and similar results are expected for any connectivity that falls off gradually with distance in the the relevant real or feature spaces that define connectivity), or more generally if there is no conspiracy by which D_{EI} and D_{IE} (or D_{EE} and D_{II}) tend to be of opposite sign and cancel, then there should be large feedforward weights and large balanced amplification.

None of these arguments are general or definitive, but all are consistent with the hypothesis that large balanced amplification should be expected when large recurrent excitation is balanced by large feedback inhibition. It obviously remains an important open question to define more precisely when this will or will not be true.

S4 Issues related to the Model and the Experimental Data

S4.1 Asynchronous, irregular activity in the spiking model, and the correspondence between spiking and rate models

The spiking model studied here operates in the "asynchronous irregular" regime [Brunel 2000] characterized by irregular spiking response and absence of global rate oscillations (Fig. S2), as in previous models of sparse balanced networks with unstructured random connectivity [Brunel 2000, van Vreeswijk and Sompolinsky 1996] or orientation-specific connectivity [Lerchner et al. 2006]. The coefficient of variation for inter-spike intervals (ISIs) is around 1 (Fig. S2A), and the ISI distribution is essentially exponential (Fig. S2C), indicating Poisson-like firing. The average firing rate in spontaneous activity fluctuates around 7 Hz without oscillations (Fig. S2B).

In the asynchronous irregular regime, mean field theory can be applied to derive expressions for firing rates from a spiking model [*e.g.* Brunel 2000, Lerchner et al. 2006, Shriki et al. 2003, Sompolinsky and White 2005]. Furthermore, for a statistically stationary input for which the system is fluctuating relatively weakly around the mean rates it would have in response to the mean input, as is the case for the spontaneous activity studied here, one can derive linear dynamical equations for the rate (although the best linear description has a band-pass temporal filter, rather than the low-pass filter used here for the rate model) [Shriki et al. 2003, Sompolinsky and White 2005].[9] One imagines that the mean connectivity matrix from which the sparse random connectivity is drawn should provide a reasonable description

[9] This correspondence was derived by [Shriki et al. 2003, Sompolinsky and White 2005] on the assumption that a neuron receives a large enough number of uncorrelated pre-synaptic spikes in one integration time that fluctuations in this number for a fixed network firing rate can be neglected. We speculate that, even for the sparsely connected network studied here, this approximation is sufficient to explain why a simple linear model captures key aspects of spiking model behavior, although this requires further study. Our neurons have about a 10 ms time constant, so at a 7 Hz average firing rate, with 100 excitatory and 25 inhibitory connections, they will receive a mean of about 14 excitatory and 3.5 inhibitory inputs in one integration time. Fluctuations in number, relative to the mean N, are expected to be of size $1/\sqrt{N}$, that is, about 25% for excitatory inputs and about 50% for inhibitory inputs.

of the connectivity in this linear model, again by mean field arguments (given enough inputs, the input a neuron receives from the sparse random sampling should show small deviations from the input it would receive under the mean connectivity matrix). Together these provide an intuitive but speculative reasoning as to why the linear rate model we studied should capture key aspects of the behavior of the spiking model we studied. Obviously, these ideas need more careful study.

S4.2 The relationship between the auto-correlation function (ACF) and the response rise time

We looked at two measures of network dynamics: the ACF of a pattern's amplitude relative to the ACF of its input, which is a natural measure of network time scales for fluctuating (spontaneous) activity and which we used to characterize the spiking model; and the onset time for response to stimuli that drive that pattern, which is a natural measure of dynamics for responses to stimuli and which we examined in some studies of the linear two-population model. We showed that neither is slowed by balanced amplification. What is the relationship between these two measures?

Intuitively, the relationship in a linear model is as follows. We are measuring the ACF of, essentially, the amplitude of the orientation-map-like pattern in the model of Fig. 3. This sum mode is driven both by input to the sum mode and by input to the difference mode, each with different temporal responses. A pulse of instantaneous input to the difference mode drives a pulse of activity in the sum mode that grows and decays, while a delta-pulse of input to the sum mode simply drives a decaying exponential of activity; the response of the E population in Fig. 2 is just the sum of these two.

We abstract from this to consider just a single input $I(t)$ that drives a response $r(t)$. In section S2.1, we show the following. Suppose an initial condition $r(0)$ evokes a response time course $R(t)$, $R(t) = 0$ for $t < 0$. Then, if a steady input I is turned on at time zero, the response at t is just the integral of R: $r(t) = I \int_0^t dt' \frac{1}{\tau} R(t')$. The response grows with time at a rate corresponding to the rate of accumulation of area under the curve $R(t)$, as can be seen by comparing Figs 3A-B (which represent $R(t)$ to their corresponding Figs 3C-D (which represent the response to onset of I.

In the noise response, each instant of noisy input evokes the same response time course $R(t)$, and these responses to the different noisy inputs at different times just superpose. So the noisy input is just being filtered by $R(t)$ to determine the response, the same filter function whose integral determines the rise time. There is a slight complication because the ACF involves a product of $r(t)$ and $r(t + \tau)$ and thus two factors of R. But, the increase in the time over which the noisy response is correlated, relative to the correlation time of the noisy input, is determined by the same time scales that determine $R(t)$ and thus determine the rise time. They are in a sense two measures of the same thing. In footnote 11 and Eq. S38 we show that this is mathematically true in the simple case that noise is derived by exponential filtering of white noise and the response function is also an exponential function.

In particular, we have seen that in the models of balanced amplification, amplification occurs with little or no widening of the ACF and little or no slowing of the rise time to stimulus onset.

S4.3 Further evidence of balanced amplification in the spiking network

In the main text (Fig. 6) we provide evidence of balanced amplification in the spiking model by showing that the time course of the amplified patterns is not slowed even as the strengths of the recurrent connections, and thus the strength of the amplification, is increased (by scaling all synapses, both excitatory and inhibitory, by a common factor). As a control, we now also examine an alternative spiking model that is identical except for a modification in connectivity that, in the linear model, yields a positive eigenvalue for patterns resembling evoked orientation maps. In this case, the time course of the amplified patterns should be increasingly slowed with increasing strength of recurrent connectivity and amplification, showing that the lack of slowing in the original model is not a general attribute of spiking models.

In the original spiking model of Fig. 5, excitation and inhibition have identical orientation tuning ($w^e_\theta = w^i_\theta = 20°$). In this case, all eigenvalues in the linear model are ≤ 0. In the modified spiking model, a “Mexican hat” connectivity is used, in which inhibitory connections have wider orientation tuning than excitatory inputs ($w^i_\theta = 50°$). In the linear rate model with this circuitry, orientation-map-like patterns have positive eigenvalues, so there is both Hebbian and balanced amplification. The overall excitatory and inhibitory synaptic strengths are equal in the two models: each neuron in the second model receives exactly the same summed excitatory and summed inhibitory input as in the first model (both in the linear versions of the models and in the spiking versions of the models). However, in the second model, a cell receives more excitation than inhibition from cells with nearby preferred orientations, and more inhibition than excitation from cells with more distant preferred orientations. Thus, orientation-map-like patterns, in which neurons with similar preferred orientation have similar activity and neurons with more distant preferred orientations have opposite activity, acquire positive eigenvalues.

In Fig. S3A,C we compare the effect of increasing recurrent strength in these two types of network, from 0% (no recurrent circuitry) to 200% (twice the strength used for the original model in Fig. 5.) Blue lines indicate the original model, using the same data as in Fig. 6, while green lines indicate the model with a positive eigenvalue. In both models, increasing recurrent strength increases the amplification of patterns resembling evoked orientation maps, as measured by the width of the distribution of correlation coefficients (Fig. S3A). The increase is less for the original model, for reasons that can be understood from the linear model. First the patterns in the $w_\theta^i = 50°$ network are amplified both by slowing associated with a positive eigenvalue and by balanced amplification, while the $w_\theta^i = 20°$ network has only the balanced amplification.

Second, one expects the correlation coefficient to grow to a plateau with increasing recurrent strength for the network that only has balanced amplification, but not for the network that also has Hebbian amplification, for the following reason. Recall from S1.2 that the response in a sum pattern, $\mathbf{p}_i^{D+}$, when its corresponding difference pattern, $\mathbf{p}_j^{S-}$, is activated is $w_{FF}g_{\lambda_i^D(t)}$ with $w_{FF} = \lambda_j^S c_{ji}$. For the network with only balanced amplification, there are no positive eigenvalues, and we assume also that there are no zero eigenvalues (real part of $-\lambda_i^D < 0$), as was true for the relevant patterns in our simulation.[10] Then, the degree of amplification produced for the sum pattern, as discussed in S2 and S1.1.1, is proportional to $\frac{w_{FF}}{1+\lambda_i^D}$ (recall, we defined λ_i^D as a positive quantity whose negative is the eigenvalue of the sum pattern) for a steady state input and $\frac{w_{FF}}{\sqrt{(1+\lambda_i^D)(2+\lambda_i^D)}}$, for white noise input. The amplification for temporally correlated input is likely to be between these two quantities. As the recurrent circuitry is scaled up, both w_{FF} and λ_i^D are scaled up by equal factors, with their ratio remaining constant, and the amplification in either case asymptotes to $\frac{w_{FF}}{\lambda_i^D}$.

In contrast, for the network with both balanced and Hebbian amplification, there is a positive eigenvalue (real part of $-\lambda_i^D > 0$), so one expects Hebbian amplification by a factor of between $\frac{1}{1+\lambda_i^D}$ (steady state) and $\frac{1}{\sqrt{(1+\lambda_i^D)}}$ (white noise). This grows without bound for real λ_i^D as $-\lambda_i^D$ approaches 1, so the amplification need not plateau.

The decay time of the amplified patterns actually decreases with increasing recurrent strength for the original model, whereas it increases with increasing recurrent strength for the modified model (Fig. S3C). The reason for the decrease is that, with increasing recurrent strength, the neurons receive more synaptic conductance and hence their average membrane time constant is reduced. To see this, we subtract the average membrane time constant from the decay time and plot the difference (dashed lines in Fig. S3C). The time added beyond the membrane time constant is roughly constant for the original model, regardless of recurrent strength. In contrast, a steep increase in decay time with increasing recurrent

[10] In our model circuit, all the $-\lambda_i^D$ have real part < 0 except one, that corresponding to the first pattern shown in Fig. 3B, which is a spatially uniform or "DC" pattern and has $\lambda^D = 0$. Following the methods used in experiments, we subtract the DC component from the frames before computing their correlation coefficient with the evoked map, so we do not consider the correlation coefficients for this pattern.

strength occurs in the modified model.

In these simulations, the difference in time course between the two models is not visible when the amplification is about 2 (0.2 correlation distribution width in Fig. S3A, vs. 0.1 or less for control pattern in Fig. S3B), the level observed experimentally. This could mean that Hebbian and balanced mechanisms are not distinguishable by the speed of decay in this range of amplification. However, it is also possible, for example, that in the modified model balanced amplification dominates Hebbian amplification over this range of parameters. Unfortunately we have not studied this.

Neither network amplifies or slows the control map pattern from Fig. 5 of the main paper (Fig. S3B,D), at any level of recurrent strength. This rules out the possibility that nonlinearities cause the modified model to slow all patterns, rather than just those with positive eigenvalues.

S4.4 Constraints on models from the time scales observed in Kenet et al. [2003]

In the text, we briefly discuss the constraints imposed by the experiments of Kenet et al. [2003] on the amount of cortical slowing in V1. Here we provide a more detailed description of our reasoning. We refer to the time series or distribution of correlation coefficients of a pattern, meaning the correlation coefficients between the pattern and snapshots of spontaneous activity.

One expects the autocorrelation time of the time series of correlation coefficients of a pattern to be given roughly by the sum of the correlation time of the inputs and the time constant of the network activity for that pattern.[11] We expect the correlation time of inputs to upper layers to be many tens of milliseconds, based on the temporal kernels of inputs from lateral geniculate nucleus to layer 4 of V1 [Wolfe and Palmer 1998] or of simple cells in V1

[11] Consider a linear model in which the input is given by white noise filtered by an exponential kernel with time constant τ_n, and the response is given by the input filtered by an exponential kernel with time constant τ_λ. The autocorrelation function of the response with itself at time difference t is given by

$$\frac{\tau_n \exp(-t/\tau_n) - \tau_\lambda \exp(-t/\tau_\lambda)}{\tau_n^2 - \tau_\lambda^2} \tag{S38}$$

To see how this behaves, first consider the limit in which $\tau_n \to \tau_\lambda$. In this limit, the expression becomes $\frac{1}{2\tau_\lambda}\left(1 + \frac{t}{\tau_\lambda}\right)\exp(-t/\tau_\lambda)$. This becomes equal to $1/e$ of its peak height for $t \approx 2.15\tau_\lambda$, that is, t slightly larger than $\tau_n + \tau_\lambda$. Second, consider the limit in which $\tau_n \gg \tau_\lambda$ (the limit with $\tau_\lambda \gg \tau_n$ behaves identically since Eq. S38 is invariant under interchange $\tau_\lambda \leftrightarrow \tau_n$). The numerator peaks at $t = 0$ with value $\tau_n - \tau_\lambda$. To find the time when the numerator has decreased to $1/e$ of this value, note that the second term has become negligible at this time relative to the first, so we can approximate the condition as $\tau_n \exp(-t/\tau_n) = (\tau_n - \tau_\lambda)/e$ or $t = \tau_n(1 - \log(1 - \frac{\tau_\lambda}{\tau_n})) \approx \tau_n(1 + \tau_\lambda/\tau_n) = \tau_n + \tau_\lambda$.

[DeAngelis et al. 1993, 1999], which should provide the dominant input to V1 upper layers [*e.g.* Martinez et al. 2005]. The 73 ms time we used in the main paper seems reasonable based on the studies of simple cells, but shorter times are also reasonable, particularly if LGN rather than simple cells are considered. We take 30 ms as a lower bound of reasonable input correlation times.

In the experimental data, the autocorrelation time of the correlation coefficients of evoked maps, measured as the time for the autocorrelation to fall to $1/e$ of its maximum, is about 80 ms (M. Tsodyks, private communication; see also Kenet et al. [2003] for a different measure of the time course that also gives a time of about 80 ms). Thus, we take 50 ms to be an upper bound for the contribution of the network time constant.

In [Kenet et al. 2003], the width of the distribution of correlation coefficients of an evoked map was about 2 times the width of the distribution for a similar, control pattern. This suggests that input patterns corresponding to evoked maps are amplified about 2 times relative to input patterns corresponding to the control pattern, although there are uncertainties in this estimate, discussed in section S2. In a Hebbian-assembly model, eigenvectors are amplified by the factor $\frac{1}{1-\lambda}$, for steady state input, or $\frac{1}{\sqrt{1-\lambda}}$, for white noise, where λ is the corresponding eigenvalue of $\mathbf{W}$, and we suggested that values for correlated noise input will be bounded by these values (section S2). Goldberg et al. [2004] studied a Hebbian-assembly model with a threshold nonlinearity in the equation governing the firing rates, and with correlated noise inputs. They showed that $\lambda = 0.6$, which gives an amplification factor of 1.6 (white noise) to 2.5 (steady-state input) in a linear model, gave a widening of the distribution of correlation coefficients of the evoked map of 2X relative to the same model without recurrent connections, well within the predicted range. The dynamics of an eigenvector with eigenvalue λ are slowed by the factor $\frac{1}{1-\lambda}$, or 2.5 times in their model. If a slowing of 2.5 times is needed to achieve 2X amplification, the amplification seen by Kenet et al. [2003] (Section S2.2), then for the network time constant to be no more than 50 ms with an amplification factor of 2.0, the intrinsic decay time, τ, needs to be no greater than 20 ms.

S4.5 Differing excitatory and inhibitory timescales in the spiking model

We have used identical timecourses for excitatory and inhibitory synaptic conductances in our spiking model. Although we have not explored the issue extensively, we imagine that, so long as firing remains in the asynchronous regime, reasonable differences in excitatory and inhibitory timescales can be compensated by changes in the synaptic connectivity, as in a linear rate model. In the linear model, consider a 2×2 network with one excitatory and one inhibitory neuron, with time constants τ and $k\tau$ respectively:

$$\tau \begin{pmatrix} 1 & 0 \\ 0 & k \end{pmatrix} \frac{d}{dt} \begin{pmatrix} r_E \\ r_I \end{pmatrix} = - \begin{pmatrix} 1 - w_{EE} & w_{EI} \\ -w_{IE} & 1 + w_{II} \end{pmatrix} \begin{pmatrix} r_E \\ r_I \end{pmatrix} + \begin{pmatrix} I_E \\ I_I \end{pmatrix} \tag{S39}$$

This network is equivalent to a network with equal time constants and modified connectivity matrix and inputs:

$$\tau \frac{d}{dt} \begin{pmatrix} r_E \\ r_I \end{pmatrix} = - \begin{pmatrix} 1 - w_{EE} & w_{EI} \\ \frac{-w_{IE}}{k} & \frac{1+w_{II}}{k} \end{pmatrix} \begin{pmatrix} r_E \\ r_I \end{pmatrix} + \begin{pmatrix} I_E \\ \frac{I_I}{k} \end{pmatrix} \tag{S40}$$

In other words, suppose we begin with a network with equal excitatory and inhibitory time constants. If we then lengthen (shorten) the inhibitory time constant, but also compensate by appropriately increasing (decreasing) all of the inputs to I cells (the $E \to I$ and $I \to I$ weights and the external input to I cells), then the network behavior will be unchanged.

There is a limit to such compensation: the new w_{II}, $w_{II}^{\text{new}} = \frac{1+w_{II}}{k} - 1$, cannot be negative. This will become negative if $k > 1 + w_{II}$, so the I time constant cannot be larger than the E time constant by a factor of more than $1 + w_{II}$ for an analysis in terms of an equivalent network with equal E and I time constants to apply.

More generally, we can qualitatively say the following: for the mechanism of balanced amplification to work, inhibition must have a combination of speed and strength that allows excitation to grow transiently yet stabilizes the system. If inhibition is too fast and/or strong relative to excitation, it will quench growth of a sum mode so quickly that balanced amplification will be very weak. If inhibition is too slow and/or weak relative to excitation, and excitation is strong enough to be unstable by itself, the network will lose stability, and a sum mode will grow without bound. In between, there is a reasonable range of parameters for which inhibition provides stability without instantly quenching growth.

S5 Supplementary Figures

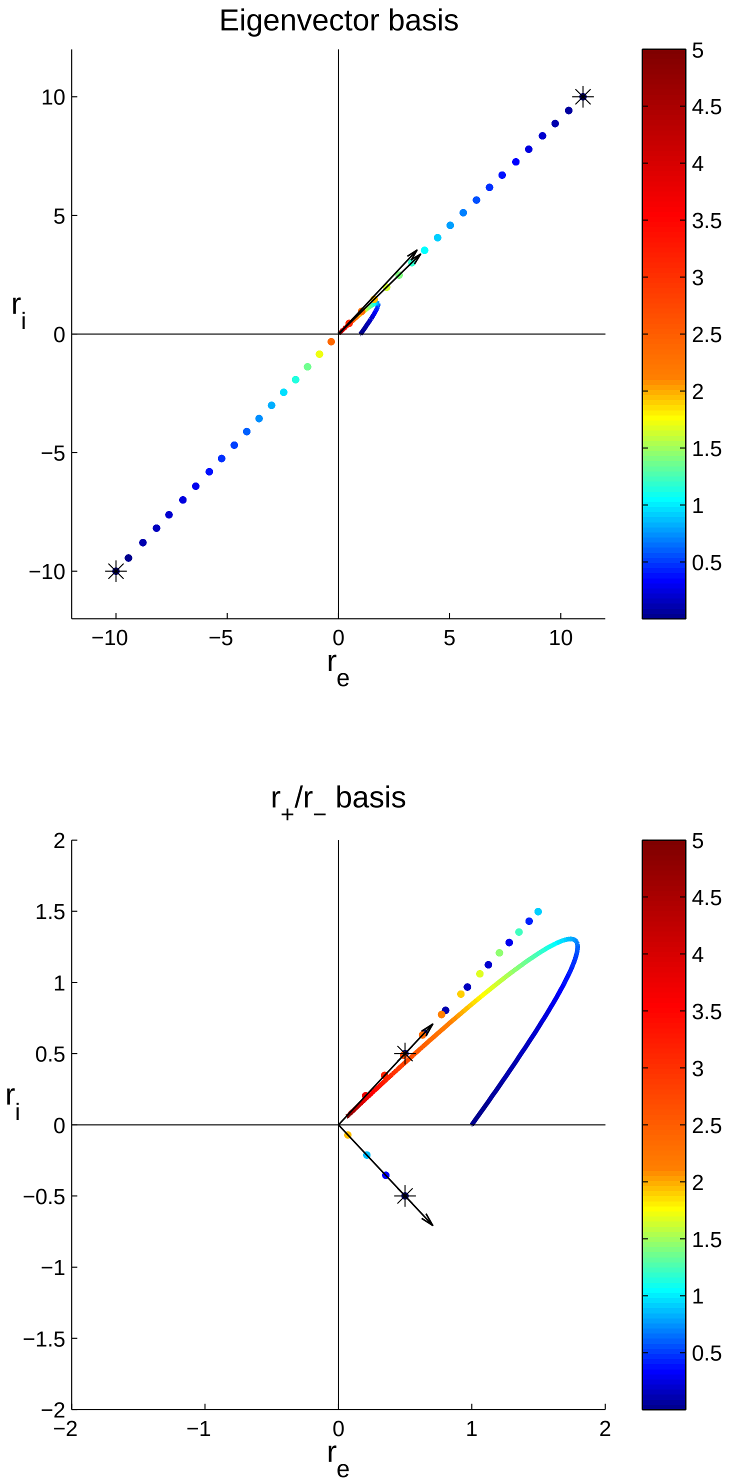

Figure S1: (See caption on next page)

Figure S1: **Comparison of eigenvector basis and sum/difference basis**. The dynamics of Fig. 1B are shown in the r_E/r_I plane, along with their decomposition into basis patterns consisting of (**A**) the eigenvectors of $\mathbf{W}$ or (**B**) the orthogonal sum and difference modes, $\mathbf{p}^+$ and $\mathbf{p}^-$, all normalized to have unit length. Time is color-coded, from 0 (blue) to 5τ (dark red), as shown in color bar. Solid line shows trajectory of r_E and r_I. Trajectory at any time is decomposed into a weighted sum of the two basis vectors; the dots show the corresponding weights or amplitudes of the two basis vectors. Asterisks indicate amplitudes at time 0, which add up to the initial value $r_E = 1, r_I = 0$. (**A**) In eigenvector basis, amplitudes are very large relative to the trajectory, and monotonically decay to the origin. Eigenvectors (black lines with arrows) are shown normalized to length 5 for visibility. (**B**) In $\mathbf{p}^+/\mathbf{p}^-$ basis, amplitudes directly reflect the dynamics both in size and non-monotonicity. Sum mode grows and then shrinks, due to feedforward connection from difference mode (Fig. 1C), while difference mode monotonically shrinks.

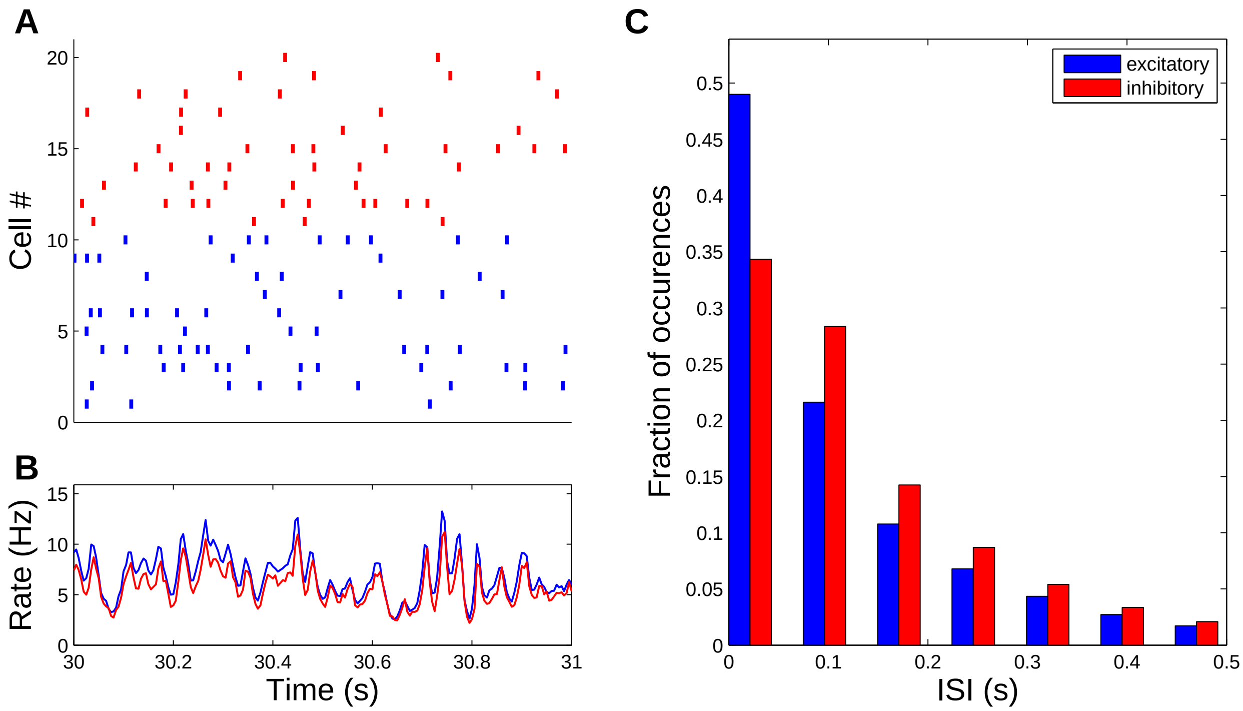

Figure S2: **Asynchronous, irregular activity in the spiking model.** During spontaneous activity excitatory neurons fired irregularly with a mean firing rate of 15 Hz and an average coefficient of variation (CV) for inter-spike intervals (ISI) of 1.0. Inhibitory neurons were similar with mean firing rates of 14.5 Hz and a CV of .95. ISIs for both types of neuron have a roughly exponential distribution. **A**) Spike raster plots over a one second long interval for 10 randomly selected excitatory (blue) and inhibitory (red) neurons. **B**) The average firing rate computed in 5ms bins of the entire population of excitatory (blue) and inhibitory (red) neurons for the same period. **C** Histogram showing the relative frequencies of different ISIs for excitatory (blue) and inhibitory (red) neurons.

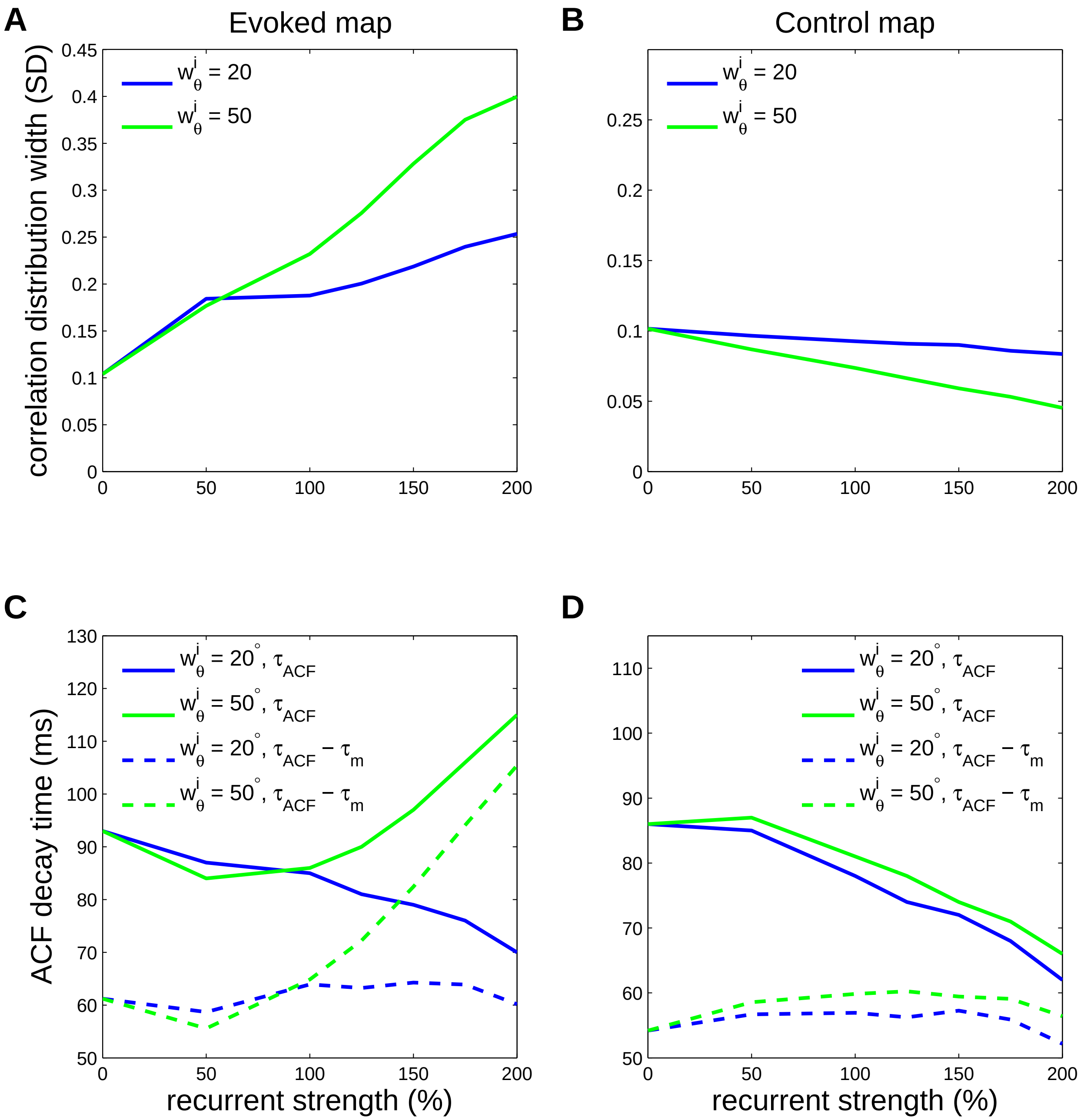

Figure S3: (See caption on next page)

Figure S3: **Effects of changing strength of recurrent synapses in spiking models.** A comparison of the effects of increasing the strength of recurrent synapses in a network with equal excitatory and inhibitory tuning widths (blue lines; simulations of Fig. 5) or wider inhibitory tuning (green lines). Orientation tuning of excitatory and inhibitory neurons are proportional to a Gaussian with standard deviation $w_\theta^e/\sqrt{2}$ and $w_\theta^i/\sqrt{2}$, respectively (more details in Methods), with $w_\theta^e = 20°$; $w_\theta^i = 20°$ (blue lines) or $w_\theta^i = 50°$ (green lines). **A,B)** The effect of increasing recurrent strength on the width (standard deviation) of the distribution of correlation coefficients with the 0° evoked map and the control map used in Fig. 5. **C,D)** The effect of increasing recurrent strength on the time constant of network activity, as measured by the time required for the autocorrelation function of the correlation coefficient timeseries to decay to 1/e of its maximum value (τ_{ACF}). The membrane time constant of the neurons (τ_m), taking into account the average synaptic conductance associated with ongoing spontaneous activity, decreases with increasing recurrent strength. The blue and green dashed lines plot $\tau_{\mathrm{ACF}} - \tau_m$ for $w_\theta^i = 20°$ and $w_\theta^i = 50°$ respectively. In all panels a strength of 100% corresponds to the synaptic strengths in the network presented in Fig. 5.

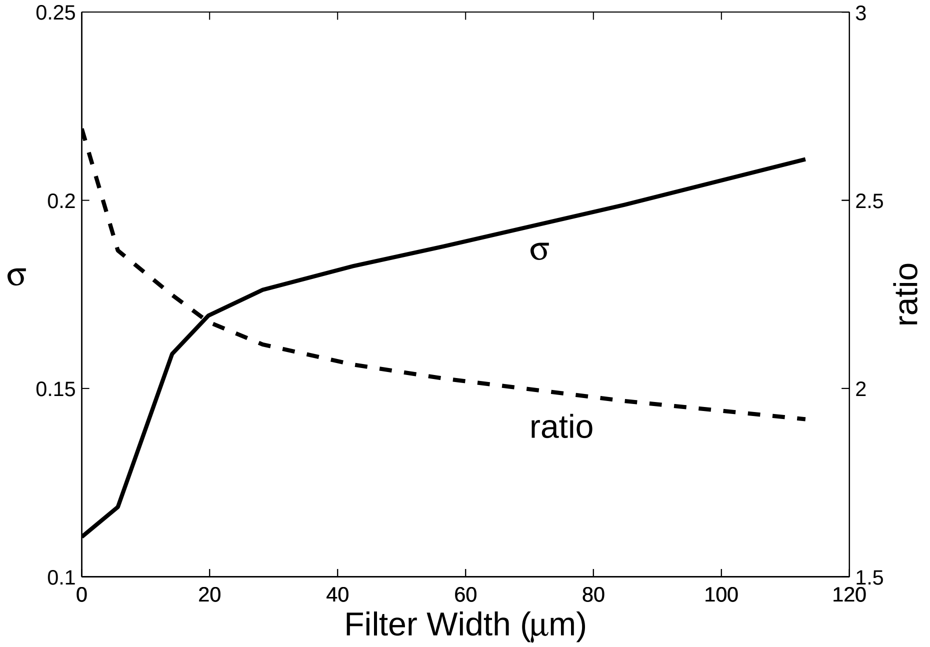

Figure S4: **Amplification of evoked maps vs. control patterns varies slowly with filter width over a broad range of filter widths.** X-axis is the standard deviation of the Gaussian filter applied to the voltage map before computing correlation coefficients, see Methods. Solid line is the standard deviation of the correlation coefficient distribution for the 0° map (left axis), which increases with filter width. Dashed line is the ratio of the standard deviation of the correlation coefficient distribution for the evoked map to that for the control pattern (right axis). This ratio serves as a measure of the degree to which input patterns corresponding to evoked maps are amplified relative to similar control patterns.

References

T. Kenet, D. Bibitchkov, M. Tsodyks, A. Grinvald, and A. Arieli. Spontaneously emerging cortical representations of visual attributes. *Nature*, 425:954–956, 2003.

M. V. Tsodyks, W. E. Skaggs, and B. L. Sejnowski, T. J.and McNaughton. Paradoxical effects of external modulation of inhibitory interneurons. *J. Neurosci.*, 17:4382–4388, 1997.

H. Ozeki, I. M. Finn, E. S. Schaffer, K. D. Miller, and D. Ferster. Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62:578–592, 2009.

J. A. Goldberg, U. Rokni, and H. Sompolinsky. Patterns of ongoing activity and the functional architecture of the primary visual cortex. *Neuron*, 13:489–500, 2004.

J.R. Polimeni, D. Granquist-Fraser, R.J. Wood, and E.L. Schwartz. Physical limits to spatial resolution of optical recording: Clarifying the spatial structure of cortical hypercolumns. *Proceedings of the National Academy of Sciences*, 102(11):4158–4163, 2005.

N. Spruston. Pyramidal neurons: dendritic structure and synaptic integration. *Nat. Rev. Neurosci.*, 9:206–221, 2008.

M. Beierlein, J.R. Gibson, and B.W. Connors. Two dynamically distinct inhibitory networks in layer 4 of the neocortex. *J. Neurophysiol.*, 90:2987–3000, 2003.

B. Pfeuty, G. Mato, D. Golomb, and D. Hansel. The combined effects of inhibitory and electrical synapses in synchrony. *Neural Comput*, 17:633–670, 2005.

R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, Cambridge, 1985.

K. Rajan and L. Abbott. Eigenvalue spectra of random matrices for neural networks. *Phys Rev Lett*, 97:188104, 2006.

N. Brunel. Dynamics of networks of randomly connected excitatory and inhibitory spiking neurons. *J Physiol Paris*, 94:445–463, 2000.

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

A. Lerchner, G. Sterner, J. Hertz, and M. Ahmadi. Mean field theory for a balanced hypercolumn model of orientation selectivity in primary visual cortex. *Network*, 17:131–150, 2006.

O. Shriki, D. Hansel, and H. Sompolinsky. Rate models for conductance-based cortical neuronal networks. *Neural Comput.*, 15:1809–1841, 2003.

H. Sompolinsky and O. L. White. Theory of large recurrent networks: From spikes to behavior. In C. Chow., B. Gutkin, D. Hansel, C. Meunier, and J. Dalibard, editors, *Methods and Models in Neurophysics, Volume Session LXXX Lecture Notes of the Les Houches Summer School 2003*, pages 267–340. Elsevier, 2005.

J. Wolfe and L. A. Palmer. Temporal diversity in the lateral geniculate nucleus of cat. *Vis. Neurosci.*, 15:653–675, 1998.

G. C. DeAngelis, I. Ohzawa, and R. D. Freeman. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development. *J. Neurophysiol.*, 69:1091–1117, 1993.

G. C. DeAngelis, G. M. Ghose, I. Ohzawa, and R. D. Freeman. Functional micro-organization of primary visual cortex: receptive field analysis of nearby neurons. *J. Neurosci.*, 19:4046–4064, 1999.

L.M. Martinez, Q. Wang, R.C. Reid, C. Pillai, J.M. Alonso, F.T. Sommer, and J.A. Hirsch. Receptive field structure varies with layer in the primary visual cortex. *Nat. Neurosci.*, 8: 372–9, 2005.

S1.2 Multi-Neuron Model

We consider the weight matrix $\mathbf{W} = \begin{pmatrix} \mathbf{W}_E & -\mathbf{W}_I \\ \mathbf{W}_E & -\mathbf{W}_I \end{pmatrix}$, an example of which was studied in Fig. 3.

We first characterize the eigenvectors and eigenvalues of $\mathbf{W}$. Let $\mathbf{W}_E$ and $\mathbf{W}_I$ be $N \times N$, and let the normalized eigenvectors of $\mathbf{W}_E - \mathbf{W}_I$ be $\mathbf{e}_i^D$ with eigenvalues $-\lambda_i^D$, $(\mathbf{W}_E - \mathbf{W}_I)\mathbf{e}_i^D = -\lambda_i^D \mathbf{e}_i^D$, $i = 1, \ldots, N$.[1] We will imagine that inhibition balances or dominates excitation in such a manner that no pattern can excite itself – all the eigenvalues of $(\mathbf{W}_E - \mathbf{W}_I)$ have real part ≤ 0 – so we have taken the eigenvalue to be $-\lambda_i^D$ so that λ_i^D will have positive real part. Then $\mathbf{W}$ has N eigenvalues equal to the $-\lambda_i^D$, with corresponding normalized eigenvectors $\mathbf{p}_i^{D+} = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_i^D \\ \mathbf{e}_i^D \end{pmatrix}$ (the + is used to indicate that these are sum modes), as can be seen directly by applying $\mathbf{W}$ to $\mathbf{p}_i^{D+}$. An additional N eigenvalues of $\mathbf{W}$ are equal to zero, because the top N rows are identical to the bottom N rows. If either $\mathbf{W}_E$ or $\mathbf{W}_I$ are invertible, the corresponding eigenvectors can be written as proportional to $\begin{pmatrix} \mathbf{W}_E{}^{-1}\mathbf{W}_I\mathbf{v} \\ \mathbf{v} \end{pmatrix}$ or ~~$\begin{pmatrix} \mathbf{v} \\ \mathbf{W}_I{}^{-1}A\mathbf{v} \end{pmatrix}$~~ $\begin{pmatrix} \mathbf{v} \\ \mathbf{W}_I{}^{-1}\mathbf{W}_E\mathbf{v} \end{pmatrix}$ for any N-dimensional basis $\mathbf{v}$. Note that, with the assumption that inhibition appropriately balances or dominates excitation, $\mathbf{W}$ has no eigenvalues with positive real part.

We now consider the feedforward connectivity. We let $\mathbf{e}_i^S$ be the normalized eigenvectors of $\mathbf{W}_E + \mathbf{W}_I$ with eigenvalues λ_i^S, and note that $\mathbf{W}_E + \mathbf{W}_I$ is a nonnegative matrix with large entries (if excitation and inhibition are large) so that some of these eigenvalues will be large and positive. We define the difference modes $\mathbf{p}_i^{S-} = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_i^S \\ -\mathbf{e}_i^S \end{pmatrix}$ and the sum modes $\mathbf{p}_i^{S+} = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_i^S \\ \mathbf{e}_i^S \end{pmatrix}$ and find that $\mathbf{W}\mathbf{p}_i^{S-} = \lambda_i^S \mathbf{p}_i^{S+}$. Thus, each pair $\mathbf{p}_i^{S-}$, $\mathbf{p}_i^{S+}$ behaves much like the difference and sum modes, $\mathbf{p}^-$ and $\mathbf{p}^+$, in the simpler, two-neuron model we studied previously, with feedforward weight $w_i^{FF} = \lambda_i^S$.

There is one difference, however. Each $\mathbf{p}_i^{S+}$ is a linear combination[2] of the $\mathbf{p}_i^{D+}$, each of which in turn decays at its own rate (determined by its λ_j^D). So the decay of $\mathbf{p}_i^{S+}$ is actually a mix of decays at different rates, rather than a decay at a single rate as before. Instead of thinking in terms of $\mathbf{p}_i^{S-}$ making a single feedforward connection to $\mathbf{p}_i^{S+}$, which

[1] In the main text we used the convention for basis vectors of denoting both which basis vector (i) and which type of basis vector (+) as superscripts, $\mathbf{p}^{i+}$, so that subscripts could be used to designate elements of the vector. In the supplement we will revert to the more usual convention $\mathbf{p}_i^+$; should we need to refer to the j^{th} element, we would write $(\mathbf{p}_i^+)_j$.

[2] This is true because the $\mathbf{p}_i^{S+}$ and the $\mathbf{p}_i^{D+}$ each span the N-dimensional space of vectors that have identical patterns of activity in the excitatory and the inhibitory neurons

then decays as a mixture of modes, one can alternatively think of $\mathbf{p}_i^{S-}$ making a set of feedforward connections to the different $\mathbf{p}_i^{D+}$'s, each of which decays at its own rate. If $\mathbf{p}_i^{S+} = \sum_j c_{ij}\mathbf{p}_j^{D+}$, then the feedforward connection from $\mathbf{p}_i^{S-}$ to $\mathbf{p}_j^{D+}$ is equal to $\lambda_i^S c_{ij}$. If the $\mathbf{e}_j^D$ and thus the $\mathbf{p}_j^{D+}$ are mutually orthogonal (see below), then $c_{ij} = \mathbf{p}_j^{D+} \cdot \mathbf{p}_i^{S+} = \mathbf{e}_j^D \cdot \mathbf{e}_i^S$.

There is one other slight wrinkle. If the matrix $\mathbf{W}_E + \mathbf{W}_I$ is not normal, then the $\mathbf{p}_i^{S-}$ will not be mutually orthogonal, nor will the $\mathbf{p}_i^{S+}$, though each $\mathbf{p}_i^{S-}$ will be orthogonal to each $\mathbf{p}_j^{S+}$. Similarly, if $\mathbf{W}_E - \mathbf{W}_I$ is not normal, the $\mathbf{p}_j^{D+}$ will not be mutually orthogonal. If this is true, this description, while correct, could be misleading in the same way that the solution in the eigenvector basis is misleading when the eigenvectors are not orthogonal, namely the size or dynamics of the basis pattern amplitudes may not directly reflect the size or dynamics of the rates. The $\mathbf{W}_E$ and $\mathbf{W}_I$ matrices we used in Fig. 3 are slightly nonnormal, because the normalization of total excitatory and inhibitory weights onto each neuron (see Methods) results in small asymmetries. However, this non-normality is very small, as assessed by measures such as $f^{\mathbf{M}}$ (see section S3.2), so the vast majority of the non-normality of the overall matrix $\mathbf{W}$ is the result of the arrangement of the submatrices, not the non-normality of the submatrices themselves. In other words, these basis patterns should be close to orthogonal to one another, if not orthogonal, so distortions, if any, should be small. Our guess is that this will be typical of biological connection matrices.

We can write down the solution in a basis of the $\mathbf{p}_i^{S-}$ and either of the group of sum modes; we choose to use $\mathbf{p}_j^{D+}$. Each $\mathbf{p}_i^{S-}$ is orthogonal to each $\mathbf{p}_j^{D+}$, and if $\mathbf{W}_E + \mathbf{W}_I$ and $\mathbf{W}_E - \mathbf{W}_I$ are normal (or close to normal), this is an orthonormal (or close to orthonormal) basis. We let $\mathbf{C}$ be the matrix with elements $C_{ij} = c_{ji}\lambda_j^S$, and let $\mathbf{L}^D$ be the diagonal matrix of the the $-\lambda_i^D$. Then in the basis $\{\mathbf{p}_1^{D+}, \ldots, \mathbf{p}_N^{D+}, \mathbf{p}_1^{S-}, \ldots, \mathbf{p}_N^{S-}\}$, the matrix $\mathbf{W}$ becomes $\begin{pmatrix} \mathbf{L}^D & \mathbf{C} \\ 0 & 0 \end{pmatrix}$.

The solution to $\tau\frac{d}{dt}\mathbf{r} = -\mathbf{r} + \mathbf{W}\mathbf{r} + \mathbf{I}$ for time-independent $\mathbf{I}$ can be formally written ~~$\mathbf{r}(t) = e^{-\frac{t}{\tau}(\mathbf{1}-\mathbf{W})}(\mathbf{r}(0) - \mathbf{I}) + \mathbf{I}$~~$\mathbf{r}(t) = e^{-\frac{t}{\tau}(\mathbf{1}-\mathbf{W})}\mathbf{r}(0) + (\mathbf{1} - e^{-(\mathbf{1}-\mathbf{W})\frac{t}{\tau}})(\mathbf{1} - \mathbf{W})^{-1}\mathbf{I}$, where, for a matrix $\mathbf{M}$, the matrix $e^{\mathbf{M}}$ is defined by the same power series as for the ordinary exponential, $e^{\mathbf{M}} = \mathbf{1} + \mathbf{M} + \mathbf{M}^2/2! + \mathbf{M}^3/3! + \ldots$. Thus, calculating $e^{-\frac{t}{\tau}(\mathbf{1}-\mathbf{W})} = e^{-\frac{t}{\tau}}e^{\frac{t}{\tau}\mathbf{W}}$ amounts to solving the equation. This turns out to be easy to do, and we can write the solution as follows. Let $\mathcal{L}^D$ be the diagonal matrix of $e^{-\lambda_i^D\frac{t}{\tau}}$, and define $\mathbf{K}$ as the matrix with entries $K_{ij} = c_{ji}\lambda_j^S(1 - e^{-\lambda_i^D\frac{t}{\tau}})/\lambda_i^D$. Then $e^{-\frac{t}{\tau}(\mathbf{1}-\mathbf{W})} = e^{-\frac{t}{\tau}}\begin{pmatrix} \mathcal{L}^D & \mathbf{K} \\ 0 & \mathbf{1} \end{pmatrix}$.

This solution tells us that an initial condition of size 1 of $\mathbf{p}_j^{S-}$ causes a response in the sum pattern $\mathbf{p}_i^{D+}$ equal to $e^{\frac{t}{\tau}}\mathbf{K}_{ij} = \lambda_j^S c_{ji}\left(e^{-\frac{t}{\tau}} - e^{-(1+\lambda_i^D)\frac{t}{\tau}}\right)/\lambda_i^D = w_{FF}g_{\lambda_i^D(t)}$ with $w_{FF} = \lambda_j^S c_{ji}$ and $g_{\lambda_j^D}(t) = g_{w_+}(t)$ (defined in Section S1.1.1) for $w_+ = \lambda_j^D$. This is precisely the response we derived for the sum mode amplitude $r_+(t)$ in the two-population model in response to an when the matrix is normal, so that $|\mathbf{e}_1 \cdot \mathbf{e}_2| = 0$. (It also becomes zero if $\lambda_1 = \lambda_2$, but this also means that the matrix is normal, because, assuming that there are two distinct eigenvectors, then when the two eigenvalues are equal, any linear combination of the two eigenvectors is also an eigenvector so we can always choose the eigenvectors to be orthonormal.)

Now, for our particular matrix, we wish to compute β. ~~We do this[7], and find that , in the orthonormal Schur basis~~To begin, we compute $|\beta|^2$ by using the fact, discussed in the last paragraph of the previous section, that the sum of the absolute squares of the matrix elements is a unitary invariant, and hence is the same in the original basis as in the Schur basis. Therefore,

$$|\beta|^2 = w_{EE}^2 + w_{EI}^2 + w_{IE}^2 + w_{II}^2 - |\lambda_+|^2 - |\lambda_-|^2 \tag{S23}$$

When the eigenvalues are real ($|Y|^2 = X^2 - w_{IE}/w_{EI}$), the sum of their absolute squares is $2w_{EI}^2(Z^2 + Y^2) = w_{EE}^2 + w_{II}^2 - 2w_{IE}w_{EI}$, so

$$|\beta|^2 = (w_{EI} + w_{IE})^2 \qquad \text{(eigenvalues real)} \tag{S24}$$

When the eigenvalues are complex ($|Y|^2 = w_{IE}/w_{EI} - X^2$), the sum of their absolute squares is $2w_{EI}^2(Z^2 + |Y|^2) = -2w_{EE}w_{II} + 2w_{IE}w_{EI}$ so

$$|\beta|^2 = (w_{EI} - w_{IE})^2 + (w_{EE} + w_{II})^2 \qquad \text{(eigenvalues complex)} \tag{S25}$$

Note that, when eigenvalues are real, β is a measure of the deviation of $\mathbf{W}$ from symmetry (a symmetric matrix would have $w_{IE} = -w_{EI}$), while when eigenvalues are complex, β is a measure of the deviation of $\mathbf{W}$ from antisymmetry (an antisymmetric matrix would have $w_{EE} = w_{II} = 0$ and $w_{IE} = w_{EI}$). Symmetric and antisymmetric real matrices are both normal matrices with real or imaginary eigenvalues, respectively. Thus, β could be thought of as a measure of distance from these "canonical" normal matrix classes whose eigenvalues are real or imaginary, respectively.[7]

To determine the phase of β, we must explicitly compute its value. We find[8] that, in the

[7] ~~Note: Footnote 7 in the original Supplement is identical to Footnote 8 in the new Supplement, except last sentence of Footnote 8 is new; original footnote 7 not shown here.~~

[7] We thank Yashar Ahmadian for pointing out to us this simple derivation and interpretation.

[8] First we note that $(\lambda_- - \lambda_+) = 2w_{EI}Y$. Next, $\mathbf{e}_- \cdot \mathbf{e}_+ = \frac{1+x_-^* x_+}{\sqrt{(1+|x_+|^2)(1+|x_-|^2)}}$. Write this as A/B where A, the numerator, may be complex, and B is real. Then the term $\frac{(\mathbf{e}_- \cdot \mathbf{e}_+)}{\sqrt{1-|\mathbf{e}_- \cdot \mathbf{e}_+|^2}} = \frac{A}{B\sqrt{1-|A|^2/B^2}} = \frac{A}{\sqrt{B^2-|A|^2}}$. Now $A = 1 + (X - Y^*)(X + Y) = 1 + X^2 - |Y|^2 + X(Y - Y*)$, with $X(Y - Y*) = 0$ if Y is real and $= 2Y$ if Y is imaginary. After some manipulation, we find that $\sqrt{B^2 - |A|^2} = 2|Y|$. Putting this all together, we find $\beta = w_{EI}(Y/|Y|)(1 + X^2 - |Y|^2 + X(Y - Y*)$. $Y/|Y| = 1$, Y real; $= i$, Y imaginary (where $i = \sqrt{-1}$). So we arrive at

$$\beta \quad = \quad w_{EI}(1 + X^2 - Y^2), \qquad Y \text{ real}$$

orthonormal Schur basis $\{\mathbf{e}_+, \mathbf{q}\}$:

$$\beta = w_{EI} + w_{IE},\ \text{Y real}\ (\xi \geq 1);$$

$$= i\left(w_{EI} + w_{IE}(2\xi^2 - 1) + i w_{IE}\xi\sqrt{1-\xi^2}\right),\ \text{Y imaginary}\ (\xi < 1);\ \text{and}$$

$$\xi = \frac{w_{EE} + w_{II}}{2\sqrt{w_{EI}w_{IE}}}$$

~~We can determine the size of the feedforward weight β when $\xi < 1$, by computing for this case~~

$$|\beta| = \left((w_{EI} - w_{IE})^2 + 4w_{EI}w_{IE}\xi^2 - 3w_{IE}^2\xi^2(1-\xi^2)\right)^{\frac{1}{2}}$$

$$= \left((w_{EI} - w_{IE})^2 + (w_{EE} + w_{II})^2\left(1 - \frac{3w_{IE}}{4w_{EI}}\right) + \frac{3(w_{EE}+w_{II})^4}{(4w_{EI})^2}\right)^{\frac{1}{2}},\ \xi < 1\,(\text{Y imaginary})$$

$$\beta = w_{EI} + w_{IE},\ \text{Y real}; \tag{S26}$$

$$= -\frac{1}{2w_{EI}}\left((w_{EE} + w_{II})\sqrt{4w_{EI}w_{IE} - (w_{EE}+w_{II})^2} + i\left(2w_{EI}(w_{EI} - w_{IE}) + (w_{EE}+w_{II})^2\right)\right),\ \text{Y imaginary} \tag{S27}$$

We have noted (section S1.2) that the solution to the dynamical equation for $\mathbf{r}$ with time-independent input $\mathbf{I}$ can be written in terms of the matrix $e^{-(\mathbf{1}-\mathbf{W})t/\tau} = e^{-t/\tau}e^{\mathbf{W}t/\tau}$ as ~~$\mathbf{r}(t) = e^{-(\mathbf{1}-\mathbf{W})t/\tau}(\mathbf{r}(0) - \mathbf{I}) + \mathbf{I}$~~ $\mathbf{r}(t) = e^{-(\mathbf{1}-\mathbf{W})t/\tau}\mathbf{r}(0) + (\mathbf{1} - e^{-(\mathbf{1}-\mathbf{W})t/\tau})(\mathbf{1}-\mathbf{W})^{-1}\mathbf{I}$ (or with time-dependent input $\mathbf{I}(t)$, $\mathbf{r}(t) = e^{-(\mathbf{1}-\mathbf{W})t/\tau}\mathbf{r}(0) + \frac{1}{\tau}\int_0^t dt' e^{-(\mathbf{1}-\mathbf{W})(t-t')/\tau}\mathbf{I}(t')$). We can compute this matrix:

$$e^{-(\mathbf{1}-\mathbf{W})t/\tau} = e^{-t/\tau}\begin{pmatrix} e^{\lambda_+ t/\tau} & \beta\frac{e^{\lambda_- t/\tau} - e^{\lambda_+ t/\tau}}{\lambda_- - \lambda_+} \\ 0 & e^{\lambda_- t/\tau}\end{pmatrix} \tag{S28}$$

$= \mathrm{iw}_{EI}(1 + X^2 - |Y|^2 + 2XY)$, $\quad Y$ imaginary

For the case Y real, which means $\left(\frac{w_{EE}+w_{II}}{2}\right)^2 \geq w_{EI}w_{IE}$, we have $X^2 - Y^2 = X^2 - (X^2 - w_{IE}/w_{EI}) = w_{IE}/w_{EI}$. Thus $\beta = w_{EI}(1 + w_{IE}/w_{EI}) = w_{EI} + w_{IE}$. In other words, the feedforward weight is just given by the sum of the two feedback inhibition terms, so if feedback inhibition is strong, there is strong balanced amplification. For the case Y imaginary, which means $w_{EI}w_{IE} > \left(\frac{w_{EE}+w_{II}}{2}\right)^2$, $X^2 - |Y|^2 = X^2 - (w_{IE}/w_{EI} - X^2) = 2X^2 - w_{IE}/w_{EI}$. Thus $\beta = iw_{EI}(1 - w_{IE}/w_{EI} + 2X(X+Y)) = i(w_{EI} - w_{IE} + 2w_{EI}X(X+Y))$. The expression simplifies somewhat if we define $\xi = \sqrt{X^2/(w_{IE}/w_{EI})} = \frac{w_{EE}+w_{II}}{2\sqrt{w_{EI}w_{IE}}}$, and note Y is imaginary if and only if $\xi < 1$. Then $X = \sqrt{w_{IE}/w_{EI}}\xi$, $Y = i\sqrt{(w_{IE}/w_{EI})(1-\xi^2)}$, and $X(X+Y) = (w_{IE}/w_{EI})(\xi^2 + i\xi\sqrt{1-\xi^2})$. Thus we can write $\beta = i(w_{EI} - w_{IE} + 2w_{IE}\xi(\xi + i\sqrt{1-\xi^2})$. Substituting back for ξ and simplifying yields Eq. S27.

of position and of preferred orientation would be translation invariant. To the extent that this latter form of model is adequate to understand many features of V1, the conclusions drawn from considering the behavior of translation-invariant matrices will apply.

Let $\mathbf{e}_i$ be the orthonormal basis of the $N \times N$ subspace in which all of the submatrices are diagonal. Let $D_{EE}(i)$ be the eigenvalue of $\mathbf{W}_{EE}$ corresponding to $\mathbf{e}_i$, and similarly for the other submatrices. For a translation-invariant matrix in which connectivity depends only on space, i corresponds to a spatial frequency, and $D_{EE}(i)$ is the Fourier transform of the excitatory connectivity at frequency i. For a translation-invariant matrix that depends on multiple spatial or feature dimensions, i represents a particular set of frequencies, one for each dimension, and $D_{EE}(i)$ is the product of the Fourier transforms of the excitatory connectivity along each dimension at the corresponding frequency for that dimension.

Define orthonormal basis vectors of the full space by the excitatory cell vector $\mathbf{e}^E_i = \begin{pmatrix} \mathbf{e}_i \\ \mathbf{0} \end{pmatrix}$ and inhibitory cell vector $\mathbf{e}^I_i = \begin{pmatrix} \mathbf{0} \\ \mathbf{e}_i \end{pmatrix}$, where $\mathbf{0}$ is the N-dimensional vector of all 0's, and work in the basis $\{\mathbf{e}^E_1, \mathbf{e}^I_1, \mathbf{e}^E_2, \mathbf{e}^I_2, \ldots, \mathbf{e}^E_N, \mathbf{e}^I_N\}$. In this basis, the matrix $\mathbf{W}$ becomes a set of N 2×2 matrices arrayed along the diagonal, with the k^{th} such matrix corresponding to the basis vectors $\mathbf{e}^E_k, \mathbf{e}^I_k$ and being of the form $D(k) = \begin{pmatrix} D_{EE}(k) & -D_{EI}(k) \\ D_{IE}(k) & -D_{II}(k) \end{pmatrix}$. Thus, the dynamics break up into independent two-dimensional subspaces, one for each N-dimensional eigenvector. E and I amplitudes for a given eigenvector interact with one another by the corresponding 2×2 matrix, but do not interact with the amplitudes for any other eigenvector.

In section S3.3, we computed the Schur decomposition for this 2×2 matrix. We showed that, if all of the D_{XY}'s were positive, the Schur basis showed a feedforward connection of size β from a difference-like mode to a sum-like mode. Here, we cannot be certain that all the D_{XY}'s will be positive, but if the connection strengths decrease smoothly with distance (in all the dimensions on which they depend), then they are likely to be, particularly when they are large. We also showed (Eqs. ~~??~~S24-~~??~~S25), on the assumption that the $D_{XY}(k)$ are real (as they will be *e.g.* if the submatrices $\mathbf{W}_{XY}$ are symmetric), that, when the eigenvalues of $D(k)$ are real, the feedforward connection strength is $\beta = D_{EI}(k) + D_{IE}(k)$; while when the eigenvalues are complex, ~~β has a more complicated form in which $|\beta|$ depends upon $D_{EE}(k) + D_{II}(k)$ and also on $|D_{EI}(k) - D_{IE}(k)|$~~$|\beta| = \left((D_{EE}(k) + D_{II}(k))^2 + (D_{EI}(k) - D_{IE}(k))^2\right)^{1/2}$. Assuming each submatrix individually has large elements, each of the D_{XY}'s must take large values for some k's (*e.g.*, the sum of the absolute squares of the $D_{EI}(k)$'s is equal to the sum of the absolute squares of the elements of $\mathbf{W}_{EI}$, etc.). If they are positive (for example, the Fourier transform of a Gaussian connectivity function is a Gaussian, and similar results are expected for any connectivity that falls off gradually with distance in the the relevant real or feature spaces that define connectivity), or more generally if there is no conspiracy by which