Assignment 8

G4360 Introduction to Theoretical Neuroscience
DUE: April 15, 2019 11:59 pm

As usual, recall that I tend to write long problem sets, but most of it is informational, there is not that much for you to actually do. So please don't be put off by the length; things you actually have to do are indicated in red. Also, the last problem in optional.
Notation: boldface small letters, like $\mathbf{r}$, represent column vectors; $\mathbf{r}^{\mathrm{T}}$ is a row vector, the transpose of $\mathbf{r}$; boldface capital letters, like $\mathbf{W}$, represent matrices; $\mathbf{W}^{\mathrm{T}}$ is the transpose of $\mathbf{W}$; non-boldface letters represent numbers, either scalars or the individual elements of vectors or matrices.

1 The inhibition-stabilized network (ISN)

First we'll do an extensive setup. The problem will be to demonstrate the paradoxical effect using nullclines.
Consider a two-population model of firing-rate neurons: one excitatory (E) population and one inhibitory (I) population. r_E and r_I are the firing rates of the excitatory and inhibitory populations, respectively, represented by the vector $\mathbf{r} = \begin{pmatrix} r_E \\ r_I \end{pmatrix}$. The matrix of connections between them is $\mathbf{W} = \begin{pmatrix} w_{EE} & -w_{EI} \\ w_{IE} & -w_{II} \end{pmatrix}$, where w_{XY} represents the (positive) strength of the connection from Y to X. We let the vector of external inputs to the two populations be $\mathbf{i}$. We let $f(\mathbf{v})$ be a nonlinear function applied element-wise to the elements of the vector $\mathbf{v}$, *i.e.* $f(\mathbf{v})$ is a vector with i^{th} element $f(\mathbf{v})_i \equiv f(v_i)$. We assume the steady-state firing rate $\mathbf{r}_{SS}$ for a given input is given by f applied to each unit's input: $\mathbf{r}_{SS} = f(\mathbf{W}\mathbf{r} + \mathbf{i})$. We assume the network approaches its instantaneous steady state with first-order dynamics: letting $\mathbf{T} = \begin{pmatrix} \tau_E & 0 \\ 0 & \tau_I \end{pmatrix}$ be the diagonal matrix of E and I time constants, we have

$$\mathbf{T}\frac{d}{dt}\mathbf{r} = -\mathbf{r} + f(\mathbf{W}\mathbf{r} + \mathbf{i}) \tag{1}$$

Suppose $\mathbf{r}_{SS}$ is a stable fixed point; we will linearize the dynamics about this fixed point. You know that, letting f'_E and f'_I be the derivative of f evaluated at the E and I components of $\mathbf{W}\mathbf{r}_{SS} + \mathbf{i}$, respectively, the linearized weights are $\begin{pmatrix} \partial f_E/\partial r_E & \partial f_E/\partial r_I \\ \partial f_I/\partial r_E & \partial f_I/\partial r_I \end{pmatrix} = \begin{pmatrix} f'_E w_{EE} & f'_E w_{EI} \\ f'_I w_{IE} & f'_I w_{II} \end{pmatrix}$; to make notation simpler, let's define this to be $\mathbf{J} = \begin{pmatrix} j_{EE} & -j_{EI} \\ j_{IE} & -j_{II} \end{pmatrix}$ (we'll assume $f(x)$ is a monotonically increasing function of x, so that all the f'_X's are positive and hence all the j_{XY}'s are positive). Let $\mathbf{i}_{SS}$ be the steady-state input that yields the fixed point $\mathbf{r}_{SS}$. If there is a deviation $\Delta\mathbf{i}$ from $\mathbf{i}_{SS}$, in the linearized equation this becomes $\delta\mathbf{i} \equiv \begin{pmatrix} f'_E \Delta i_E \\ f'_I \Delta i_I \end{pmatrix}$. Define small deviations in response from the steady state by $\mathbf{r} = \mathbf{r}_{SS} + \delta\mathbf{r}$. Then the

equation for the dynamics linearized about the fixed point is

$$\mathbf{T}\frac{d}{dt}\delta\mathbf{r} = -\delta\mathbf{r} + \mathbf{J}\delta\mathbf{r} + \delta\mathbf{i} = -(\mathbf{1}-\mathbf{J})\delta\mathbf{r} + \delta\mathbf{i} \tag{2}$$

where $\mathbf{1}$ is the identity matrix. This should all be familiar to you, but if it's not, satisfy yourself that this is all true.

Recall what we did in class to show the ISN paradoxical response: for a steady-state input perturbation $\delta\mathbf{i}$, we wrote down the equation for the steady-state response $\delta\mathbf{r}$: $\delta\mathbf{r} = (\mathbf{1}-\mathbf{J})^{-1}\delta\mathbf{i}$. For a 2×2 matrix $\mathbf{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, the inverse is given by $\mathbf{M}^{-1} = \frac{1}{\mathrm{Det}\,\mathbf{M}}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$. So $(\mathbf{1}-\mathbf{J})^{-1} = \frac{1}{\mathrm{Det}\,(\mathbf{1}-\mathbf{J})}\begin{pmatrix} 1+j_{II} & -j_{EI} \\ j_{IE} & 1-j_{EE} \end{pmatrix}$. Recall that, for the fixed point to be stable, we must have $\mathrm{Det}\,(\mathbf{1}-\mathbf{J}) > 0$. Thus, for a stable fixed point, if and only if $j_{EE} > 1$ (which means the E population alone would be unstable if I firing was frozen at its fixed point level; look at the equation for r_E with r_I fixed, to see why $j_{EE} > 1$ implies excitatory instability), the I cells show a "paradoxical" response. This means that, if an input is given only to I cells ($\delta\mathbf{i} \propto \begin{pmatrix} 0 \\ 1 \end{pmatrix}$), the steady-state response of the I cells is of opposite sign to the input, so that adding excitation to I cells paradoxically lowers their firing rate in the new steady state.

Now show the same things using nullclines. Again assume that the function $f(x)$ is a monotonically increasing function of x. The equations for the E and I nullclines are the E and I components of the fixed-point equation: $\mathbf{r} = f(\mathbf{W}\mathbf{r}+\mathbf{i})$. We will draw the nullclines with r_E on the x axis and r_I on the y axis.

a For the I nullcline, compute its slope, dr_I/dr_E; you should find that it is given by $\frac{j_{IE}}{1+j_{II}}$. This means that the nullcline always has positive slope.

b Now for the E nullcline, compute the inverse of its slope, dr_E/dr_I; you should find that this inverse slope is $\frac{j_{EI}}{j_{EE}-1}$. This means that the slope is positive if the E subnetwork is unstable, and negative if the E subnetwork is stable.

c Show that the condition that $\mathrm{Det}\,(\mathbf{J}-\mathbf{1}) > 0$, which is necessary for stability, is equivalent to the I nullcline having a larger slope than the E nullcline. So for a fixed point to be stable, it is necessary that the I nullcline have a larger slope than the E nullcline at their crossing that defines the fixed point.

d So, we'll draw two versions of the nullclines: one that is an ISN, one that is not. First draw the I nullcline, which will be the same for both versions. Imagine that, for r_E small, the I-nullcline solution for r_I should be small, while for r_E large, r_I is large; so the nullcline could start toward the bottom left, and make, for example, a sigmoidal shape to the top right, with positive slope always.

e Now, draw the E nullclines, assuming a stable fixed point. Imagine that when r_I is high, r_E is low, so the nullcline starts in the upper left corner; while when r_I is low, r_E is high, so it ends up in the lower right corner. In the non-ISN version, it has a negative slope all the way. In the ISN version, it has a positive slope in a middle portion, so the nullcline looks like a sideways S; and the fixed point is on this positive-sloping middle portion (and the necessary condition for stability on E and I nullcline slopes is obeyed).

f Draw the arrows indicating the direction of flow in the different regions of the nullcline plane. Show that, in negative-sloping regions of the E nullcline, if r_I is kept fixed, small perturbations of r_E off

the E nullcline will flow back to the nullcline; while in positive-sloping regions, it will flow away. This also tells you that in positive-sloping regions, the E subnetwork alone is unstable, while in negative-sloping regions it is stable.

g Now, suppose you add a positive input to the I cells. Show that the resulting change in the I nullcline is to reduce r_E by the same amount for any given r_I, that is, to move the I nullcline leftward. There is no change in the E nullcline. Show that, for a stable fixed point, if the network is an ISN, the result is to decrease both r_E and r_I in moving to the new fixed point; while for a non-ISN, the result is to decrease r_E but increase r_I. (For the ISN, assume that the new fixed point, like the old one, is on the positive-sloping portion of the E nullcline.)

h In the ISN case, draw the dynamical path followed by r_E, r_I from the old fixed point to the new fixed point after adding the positive input to I. This addition of input instantaneously moves the I nullcline; the resulting derivative at the old fixed point (which is no longer on the I nullcline and so no longer a fixed point) has an upward component, becoming horizontal as the flow crosses the I nullcline, and then going downward to the new fixed point (it might spiral into the fixed point if there are complex eigenvalues, or go straight down to it if eigenvalues are real). Note, regarding the old fixed point as a perturbation from the new fixed point, that, even though the new fixed point is stable, the dynamics move further away from the new fixed point (the upward movement) before ultimately flowing back to it. This is an effect of non-normal dynamics. (Recall that biological weight matrices, of the form $\mathbf{J} = \begin{pmatrix} j_{EE} & -j_{EI} \\ j_{IE} & -j_{II} \end{pmatrix}$ with all j_{XY}'s positive, are non-normal, meaning that their eigenvectors are not orthogonal, because $\mathbf{J}\mathbf{J}^{\mathrm{T}} \neq \mathbf{J}^{\mathrm{T}}\mathbf{J}$, which is the necessary and sufficient condition for non-normality.)

2 Eigenvectors, Schur Vectors and Non-Normal Dynamics in Higher Dimensions

Here we'll see how the spatial structure of connectivity relates to the spatial structure of eigenvectors and Schur vectors in a simple case, and examine non-normal dynamics in this case.

Consider a 2D network of 32×32 E and I cells (one E cell and one I cell at each of the 32×32 grid positions). Assign each grid location a preferred orientation as follows: break the grid up into a 4×4 set of 8×8 grid positions. The top left 8×8 grid is assigned a preferred orientation according to its angle with its center. Orientation runs from 0 to 180^o, so that angle is divided by two. Thus, for grid position (x, y), with x and y both in the range 1 to 8, the preferred orientation is $\mathrm{ArcTan}(\frac{y-4.5}{x-4.5}) * 180^o/(2\pi)$ (and add 180^o if this is a negative number, so that the orientation runs from 0^o to 180^o instead of from -90^o to 90^o). Orientation maps of neighboring 8×8 grids are mirror images of each other, flipped across the border between them so that the orientations along their borders are matched. Thus, for y in the range 1 to 8, if x is in the range 9 to 16, (x, y) is assigned the same preferred orientation as $(17 - x, y)$; and then if x is in the range 17 to 32, (x, y) is assigned the same preferred orientation as $(x - 16, y)$. Then for y in the range 9 to 16, (x, y) is assigned the same orientation as $(x, 17 - y)$; and then for y in the range 17 to 32, (x, y) is given the same orientation as $(x, y - 16)$. Visualize the orientation map as a heat map, using a circular color map, *e.g.* one that goes from blue to red as orientation goes from 0^o to 180^o.

Now let's use vector notation, *e.g.* $\mathbf{x}$ or $\mathbf{y}$, to represent a two-dimensional grid position, *e.g.* $\mathbf{x} = (x_1, x_2)$ where x_1 and x_2 are locations along the two grid coordinates. Use periodic boundary conditions (treat the

top and bottom of the grid as next to each other, and similarly for the left and right of the grid). Let $d(\mathbf{x}, \mathbf{y})$ be the shortest distance across the periodic grid between grid positions $\mathbf{x}$ and $\mathbf{y}$:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\text{Min}[(x_1 - y_1)^2, (32 - (x_1 - y_1))^2] + \text{Min}[(x_2 - y_2)^2, (32 - (x_2 - y_2))^2]} \tag{3}$$

Let $\theta(\mathbf{x})$ be the preferred orientation at position $\mathbf{x}$, and let $\Theta(\mathbf{x}, \mathbf{y})$ be the shortest distance around a 180^o circle between $\theta(\mathbf{x})$ and $\theta(\mathbf{y})$:

$$\Theta(\mathbf{x}, \mathbf{y}) = \text{Min}\left(|\theta(x) - \theta(y)|, 180^o - |\theta(x) - \theta(y)|\right) \tag{4}$$

We'll use the simple case in which both E projections are identical, and both I projections are identical. Let the connection strength from the cell of type Y (E or I) at position $\mathbf{y}$ to the cell of either type at position $\mathbf{x}$ be proportional to a product of a Gaussian function of the distance between them and a Gaussian function of their difference in preferred orientation:

$$W^Y_{\mathbf{xy}} \propto e^{-\frac{d(\mathbf{x},\mathbf{y})^2}{2(\sigma^Y_d)^2}} e^{-\frac{\Theta(\mathbf{x},\mathbf{y})^2}{2\sigma^2_\theta}} \tag{5}$$

Here, σ_θ has been chosen identical for all connection types. Set $\sigma_\theta = 20^o$, in line with findings in V1 that the excitation and the inhibition received by middle and upper layer cells on average have the same orientation tuning. Take $\sigma^E_d = 23$ grid units and $\sigma^I_d = 2.3$ grid units, at least qualitatively in line with the observation that excitation but not inhibition makes long-range projections. Normalize (scale) the sum of excitatory weights to each cell to equal 20, and identically normalize the sum of inhibitory weights received by each cell.

Form the weight matrix. We regard the rate vector to be of the form $\begin{pmatrix} \mathbf{r}_E \\ \mathbf{r}_I \end{pmatrix}$, where $\mathbf{r}_E$ and r_I are each $N = 1024$-dimensional vectors of the firing rates of all E cells and all I cells respectively. The n^{th} element, $1 \le n \le N$, of $\mathbf{r}_E$ or $\mathbf{r}_I$ corresponds to the E or I cell, respectively, with (x, y) position $(\text{mod}(n, 32), \text{ceil}(n/32))$. Accordingly, the weight matrix can be written as $\mathbf{W} = \begin{pmatrix} \mathbf{W}^E & -\mathbf{W}^I \\ \mathbf{W}^E & -\mathbf{W}^I \end{pmatrix}$ where each of the four blocks is an $N \times N$ matrix, giving the projections of all N cells of one type to all N of the other type, where the types are $\begin{pmatrix} E \to E & I \to E \\ E \to I & I \to I \end{pmatrix}$. All the entries of all the $\mathbf{W}$'s are positive, and the minus sign for inhibitory synapses is added explicitly above.

For the special case of $\mathbf{W}$ of the form $\begin{pmatrix} \mathbf{W}^E & -\mathbf{W}^I \\ \mathbf{W}^E & -\mathbf{W}^I \end{pmatrix}$, we can write its eigenvectors and eigenvalues as follows. Let $\mathbf{e}^D_i$ be the eigenvectors of $\mathbf{W}^E - \mathbf{W}^I$, $i = 1, \ldots, N$, with corresponding eigenvalues λ^D_i (the 'D' is for 'difference'). Then $\begin{pmatrix} \mathbf{e}^D_i \\ \mathbf{e}^D_i \end{pmatrix}$ is an eigenvector of $\mathbf{W}$ with eigenvalue λ^D_i. This gives N of the $2N$ eigenvectors of $\mathbf{W}$. $\mathbf{W}$ is a $2N \times 2N$ matrix, but it only has N independent rows (the 2nd N rows are identical to the first N). Therefore it has rank N, meaning that it has N eigenvalues equal to zero. If either $\mathbf{W}^E$ or $\mathbf{W}^I$ is invertible, the corresponding eigenvectors can be written as $\begin{pmatrix} (\mathbf{W}^E)^{-1}\mathbf{W}^I\mathbf{b}_i \\ \mathbf{b}_i \end{pmatrix}$ or $\begin{pmatrix} \mathbf{b}_i \\ \mathbf{W}^I)^{-1}\mathbf{W}^E\mathbf{b}_i \end{pmatrix}$ respectively, where the $\mathbf{b}_i$, $i = 1, \ldots, N$ are any complete basis for N-dimensional space.

a Compute and plot the 5 eigenvectors $\mathbf{e}^D_i$ with the largest eigenvalues (the eigenvalues will be real, because $\mathbf{W}^E - \mathbf{W}^I$ is symmetric) and note their corresponding eigenvalues. Plot them as

2-dimensional heat maps, unpacking the N-dimensional vector into values on a 32×32 grid. The corresponding 5 eigenvectors of $\mathbf{W}$ have identical E and I activity patterns of the shape of the given $\mathbf{e}_i^D$.

b Can you say anything about how the spatial structure of these eigenvectors reflects the spatial structure of the connectivity? For comparison, compute and plot the steady-state response of linear rate dynamics to an oriented full-field input of orientation $\theta_0 = 0^o$, and the steady-state response to input of orientation $\theta_0 = 90^o$, where the input to the E and I neurons at position $\mathbf{x}$ is given by $4e^{-\frac{(\theta_0 - \theta(\mathbf{x}))^2}{2(20^o)^2}}$. If the $2N$-dimensional input vector is $\mathbf{i}$, the steady-state response $\mathbf{r}$ is given by $\mathbf{r} = (\mathbf{1} - \mathbf{W})^{-1}\mathbf{i}$ where $\mathbf{1}$ is the $2N$-dimensional identity matrix.

c Compute the correlation coefficient of each eigenvector $\mathbf{e}_i^D$ with each orientation response: to find the correlation coefficient, make each vector zero-mean by subtracting off the mean element from each element of the vector; then take the dot product of these two zero-mean vectors, and divide by the product of the norms of the two zero-mean vectors.

d If $\mathbf{e}_i^S$ ('S' for sum) are the eigenvectors of $\mathbf{W}^E + \mathbf{W}^I$, with eigenvalue λ_i^S, then $\mathbf{W}\begin{pmatrix} \mathbf{e}_i^S \\ -\mathbf{e}_i^S \end{pmatrix} = \lambda_i^S \begin{pmatrix} \mathbf{e}_i^S \\ \mathbf{e}_i^S \end{pmatrix}$. That is, there is an effective feedforward weight from the difference vectors $\begin{pmatrix} \mathbf{e}_i^S \\ -\mathbf{e}_i^S \end{pmatrix}$ to the corresponding sum vector $\begin{pmatrix} \mathbf{e}_i^S \\ \mathbf{e}_i^S \end{pmatrix}$. Small E/I differences in various spatial patterns evoke large (if λ_i^S is large) responses of both E and I in the same pattern. (The fact that a difference in a given spatial pattern evokes a sum response in exactly the *same* spatial pattern is a result of our simplifying assumption that the two E projections are identical and the two I projections are identical; the idea that patterns representing roughly opposite patterns of E and I response (difference patterns) evoke responses in which E and I have similar patterns of response (sum patterns) is more general.)

Now, note that assuming the $\mathbf{e}_i^D$ and the $\mathbf{e}_i^S$ are each complete orthonormal bases for the N-dimensional space (which they are, because $\mathbf{W}_E$ and $\mathbf{W}_I$ are both symmetric); then, the vectors $\left\{ \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_i^D \\ \mathbf{e}_i^D \end{pmatrix}, \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_j^S \\ -\mathbf{e}_j^S \end{pmatrix} \right\}$ form an orthonormal basis for the $2N$-dimensional space (a given vector's dot product with itself is 1, and the dot product of any two different vectors is 0). In this case, we have found a Schur basis – an orthonormal basis in which $\mathbf{W}$ is upper triangular with the eigenvalues on the diagonal. In particular, note that the vectors $\begin{pmatrix} \mathbf{e}_i^S \\ \mathbf{e}_i^S \end{pmatrix}$ can be written as linear combinations of the vectors $\begin{pmatrix} \mathbf{e}_i^D \\ \mathbf{e}_i^D \end{pmatrix}$ – the latter are a complete basis for the N-dimensional set of $2N$-dimensional activity patterns in which E and I have identical activity – so we can write $\frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_i^S \\ \mathbf{e}_i^S \end{pmatrix} = \sum_j w_{ij} \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_j^D \\ \mathbf{e}_j^D \end{pmatrix}$ for some weights w_{ij}. So, in the basis ordered with first all of the vectors $\left\{ \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_1^D \\ \mathbf{e}_1^D \end{pmatrix}, \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_2^D \\ \mathbf{e}_2^D \end{pmatrix}, \ldots, \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_N^D \\ \mathbf{e}_N^D \end{pmatrix} \right\}$ and then all of the vectors $\left\{ \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_1^S \\ -\mathbf{e}_1^S \end{pmatrix}, \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_2^S \\ -\mathbf{e}_2^S \end{pmatrix}, \ldots, \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_N^S \\ -\mathbf{e}_N^S \end{pmatrix} \right\}$, the matrix $\mathbf{W}$ can be written as follows: on the diagonal are the entries $\lambda_1^D, \lambda_2^D, \ldots, \lambda_N^D$, followed by N 0's; the weights $\lambda_i^S w_{ij}$ are in the upper right, in the i^{th} row and $(N+j)^{th}$ column; and all other entries are zero. (To see this, note that, given

some $2N$-dimensional basis $\mathbf{b}_i$ and some $2N \times 2N$ matrix $\mathbf{M}$, if $\mathbf{M}\mathbf{b}_i = \sum_j m_{ij}\mathbf{b}_j$, then the m_{ij} are the elements of the matrix $\mathbf{M}$ in the basis $\mathbf{b}_i$.) (Note that the general form of a matrix in a Schur basis allows nonzero entries anywhere above the diagonal, and nonzero eigenvalues anywhere along the diagonal; this particular matrix has the lower half of rows = 0 because of our special assumption that the two E projections are identical and the two I projections are identical.)

(Note, just FYI: more generally, a Schur basis for a matrix is found by (1) putting the eigenvectors of the matrix in some order and then (2) performing Gram-Schmidt orthonormalization on these ordered eigenvectors to produce an orthonormal basis, which will be a Schur basis. Because the ordering of the eigenvectors is arbitrary, there are many possible Schur bases for a given matrix. However, there is an invariant: because the Schur basis is orthonormal, the transformation to the Schur basis is a unitary transformation, and under unitary transformations, the sum of the absolute squares of the matrix entries are preserved. Since the diagonal is always the eigenvalues in any Schur basis, and the entries below the diagonal are all zero, then if the matrix entries in a given Schur basis are m_{ij}, this means that the sum $\sum_{i<j} |m_{ij}|^2$ of the absolute squares of the feedforward weights – the nonzero entries above the diagonal – is the same in any Schur basis. This sum, relative to the sum of the absolute squares of all the matrix entries, is an invariant – the same for all Schur bases for a given matrix – that gives some sense of how much "power" is in the feedforward connections relative to the eigenvalues, which are the "self-loops" of Schur-basis activity patterns onto themselves.)

Now plot the 5 leading $\mathbf{e}_i^S$ (those with the largest λ_i^S), and note their corresponding λ_i^S. Note that, if λ_i^S is large, small activity patterns in which E and I are both in the shape $\mathbf{e}_i^S$ but with opposite signs evoke large activity patterns in which E and I are both in this shape with the same sign.

e Again, what can you say about how the spatial structure of the $\mathbf{e}_i^S$ relates to the spatial structure of the connectivity and the evoked orientation responses.

f Compute the correlation coefficient of each eigenvector $\mathbf{e}_i^S$ with each orientation response.

g Finally, examine the dynamics of the non-normal amplification in this model. For each of the 5 leading $\mathbf{e}_i^S$, start the dynamics with an initial condition $\begin{pmatrix} \mathbf{r}_E \\ \mathbf{r}_I \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{e}_i^S \\ -\mathbf{e}_i^S \end{pmatrix}$. The dynamics is given by $\tau \frac{d\mathbf{r}}{dt} = -\mathbf{r} + \mathbf{W}\mathbf{r}$ (the external input is zero). Plot the time course of $|\mathbf{r}|$. To compute the dynamics, you could use simple Euler, found by replacing $\tau \frac{d\mathbf{r}}{dt}$ with $\frac{\mathbf{r}(t+\Delta t)-\mathbf{r}(t)}{\Delta t/\tau}$:
$\mathbf{r}(t+\Delta t) = (1-\Delta t/\tau)\mathbf{r}(t) + (\Delta t/\tau)\mathbf{W}\mathbf{r}(t)$. Using $\Delta t/\tau = 0.1$ should be sufficient (you could check that cutting Δt in half doesn't noticeably change the outcome, which is a decent check that your time step is small enough). Continue the simulation until both components of $\mathbf{r}$ are small, say < 0.01.

h For each of the 5 $\mathbf{e}_i^S$, plot the time course of $|\mathbf{r}|$ along with the functions $\lambda_i^S e^{-t/\tau}$ and $\lambda_i^S t e^{-t/\tau}$. You should find that the time course of $|\mathbf{r}|$ is roughly sandwiched between these two, looking more like the first for lower components (those with largest λ_i^S) and moving toward the latter for later components. For these later components, you might want to scale down the $\lambda_i^S e^{-t/\tau}$ so its peak is the same height of the peak of the given sum vector component. The reasons why these time courses should bracket the time course of $|\mathbf{r}|$ are explained in this footnote.[1]

Some context: it was observed that, in V1 spontaneous activity in anesthetized cat, patterns of activity across the cortical surface that looked like responses to oriented stimuli occurred with larger amplitude

[1]Recall that, if there are two patterns, where the first pattern connects to itself with eigenvalue λ_1 and connects to the second pattern with feedforward weight w_{FF}, and the second pattern connects to itself with eigenvalue λ_2; then the response

than expected by chance. That is, looking at the correlation coefficient between spontaneous activity and the evoked response to a drifting oriented grating, the average correlation coefficient was zero, but the distribution of correlation coefficients was wider than to a control pattern with similar spatial-frequency characteristics as the evoked response, suggesting larger spontaneous excursions in the directions of evoked responses than in control responses [1]. It was suggested that this might mean there was the equivalent of a bump attractor underlying this activity, so that activity intrinsically wanted to look like the cortex was seeing a particular orientation even if the input didn't statistically favor any orientation. A model was built of a bump attractor network [2]: it concluded that either the bump attractor had to represent position in a space of a high number of features (> 10), not just orientation; or the results had to come from a regime without a bump attractor. This was based on such considerations as the slow dynamics of the ring bump attractor, which become faster with multiple features; the single-peaked structure of the distribution of correlation coefficients, which became multi-peaked for bump attractors with too few features; and the width of the distribution of correlation coefficients, which was two wide for bump attractors with too few features. On the other hand, in their hands, the non-bump-attractor scenario (1) used Mexican hat connectivity and (2) predicted much too narrow a width of the distribution of correlation coefficients, which could be fixed by assuming sufficient spatial correlations in the LGN input.

We believed that there was no "Mexican hat", based on findings that, at least in upper and middle layers of V1, the excitation and the inhibition that cells receive have the same orientation tuning [3–5]. We wanted to know if the amplification could happen without a Mexican hat. We explored a simulation with E and I cells with connectivity like that above, and found that patterns like evoked responses were being amplified more than control patterns, as in the data of **(author?)** [1]. Figuring out why this was happening then led us to understand the role of non-normal dynamics and effective feedforward connections [6]: strong effective feedforward connections from small E/I differences in orientation-like

---

to an initial condition $r_1(0) = 1$, $r_2(0) = 0$, with no external input, is

$$r_1(t) = r_1(0)e^{-(1-\lambda_1)t/\tau} \tag{6}$$

$$r_2(t) = w_{FF}r_1(0)\frac{e^{-(1-\lambda_1)t/\tau} - e^{-(1-\lambda_2)t/\tau}}{\lambda_1 - \lambda_2} \tag{7}$$

In our case, pattern 1 is the difference pattern, and it has $\lambda_1 = 0$. When $\lambda_1 = \lambda_2$, then $\frac{e^{-(1-\lambda_1)t/\tau} - e^{-(1-\lambda_2)t/\tau}}{\lambda_1 - \lambda_2}$ becomes $(t/\tau)e^{-(1-\lambda_1)t/\tau}$, as can be seen by taking $\lambda_2 = \lambda_1 + \epsilon$ and taking the limit $\epsilon \to 0$:

$$\frac{e^{-(1-\lambda_1)t/\tau} - e^{-(1-\lambda_2)t/\tau}}{\lambda_1 - \lambda_2} = \frac{e^{-(1-\lambda_1)t/\tau}(1 - e^{\epsilon t/\tau})}{-\epsilon} \tag{8}$$

$$\to \frac{e^{-(1-\lambda_1)t/\tau}(-\epsilon t/\tau)}{-\epsilon} = (t/\tau)e^{-(1-\lambda_1)t/\tau} \tag{9}$$

where we used $e^{\epsilon t/\tau} = 1 + \epsilon t/\tau + O(\epsilon^2)$ where $O(\epsilon^2)$ means terms that depend on 2nd or higher powers of ϵ.

For us, r_2 would be the sum pattern, but it does not correspond to a single eigenvalue; rather it is a linear combination of different patterns $\begin{pmatrix} \mathbf{e}_j^D \\ \mathbf{e}_j^D \end{pmatrix}$, each of which has a different eigenvalue λ_j^D and thus decays at a different rate. One of these different orthogonal components will decay as $\lambda_i^S w_{ij} \frac{e^{-t/\tau} - e^{-(1-\lambda_j^D)t/\tau}}{-\lambda_j^D}$. Since $\sum_j w_{ij}^2 = 1$, then the norm of the sum of these different components would have a size λ_i^S times an effective time course which represents some sort of mixture of these time courses for different λ_j^D.

The sum patterns with larger λ_i^S will likely be composed of patterns $\begin{pmatrix} \mathbf{e}_i^D \\ \mathbf{e}_i^D \end{pmatrix}$ with larger (less negative) λ_i^D. Assuming all the λ_i^D are ≤ 0, then the most slowly the sum pattern could decay is if it was dominated by a pattern that decayed at rate 0, giving a decay of $\lambda_i^S(t/\tau)e^{-t/\tau}$. On the other hand, later patterns should decay at rates determined by more negative eigenvalues. With $\lambda_1 = 0$, as $\lambda_2 \to -\infty$, the response time course $\frac{e^{-(1-\lambda_1)t/\tau} - e^{-(1-\lambda_2)t/\tau}}{\lambda_1 - \lambda_2}$ goes to $\frac{e^{-t/\tau}}{\infty}$, that is, to a time course approaching $e^{-t/\tau}$ with shrinking amplitude. The actual decays, as a mix of different finite decay rates, should cover some range of time courses between $\lambda_i^S(t/\tau)e^{-t/\tau}$ and $\lambda_i^S e^{-t/\tau}$, with the latter's amplitude likely needing to be adjusted.

patterns to strong E/I sums in orientation-like patterns could explain the result. Given noise that would include random fluctuations in many patterns including those of the small E/I differences, the fact that the strongest feedforward connections are for patterns resembling evoked orientation maps would explain why such patterns would appear with larger amplitude than alternative patterns. (Note: the *most* strongly amplified pattern is the spatially uniform pattern. **(author?)** [1] saw plenty of variability in this pattern, but because this could result artifactually, *e.g.* from changes in power to the light source (they were doing intrinsic signal imaging based on shining a light of a given wavelength on cortex and looking at the strength of the reflected signal), they filtered out any spatially uniform signals and did not study them.)

3 [Optional] Ring networks: Bump attractors or an SSN

- **Bump attractor:** We'll consider a ring of 180 grid position, representing preferred direction from 0 to 2π (2π radians, *i.e.* 360^o) by $\Delta\theta = 2\pi/180$. There is a single unit at each position, which projects both positive and negative synapses. We consider linear-rectified input/output functions. We will construct discrete dynamics on the grid from a continuous model, in which θ is a continuous variable from 0 to 2π, $r(\theta)$ is the response of the unit preferring orientation θ and $\mathbf{i}(\theta)$ is its input, $\mathbf{W}(\theta - \theta')$ is the connection between the units preferring θ and θ', and v_{th} is a threshold for firing:

$$\frac{dr(\theta)}{dt} = -r(\theta) + \left[\int_0^{2\pi} \frac{d\theta'}{2\pi} W(\theta - \theta')r(\theta') + i(\theta) - v_{th}\right]_+ \tag{10}$$

  Here, $[x]_+$ is rectification: $= x$ if $x > 0$, $= 0$ otherwise.

  We move this to a grid with positions θ_i, $i = 1, \ldots, 180$; $r_i = r(\theta_i)$, and $\mathbf{r}$ is the resulting vector of rates, and similarly for i_i and $\mathbf{i}$; $\mathbf{v}_{th}$ is the vector all of whose elements are v_{th}; and $W_{ij} = W(\theta_i - \theta_j)\frac{\Delta\theta}{2\pi}$ and $\mathbf{W}$ the resulting matrix, giving dynamics

$$\frac{d\mathbf{r}}{dt} = -\mathbf{r} + [\mathbf{W}\mathbf{r} + \mathbf{i} - \mathbf{v}_{th}]_+ \tag{11}$$

  We define $\mathbf{W}$ and $\mathbf{i}$ from

$$W(\theta) = W_0 + 2W_1\cos(\theta) \tag{12}$$

$$i(\theta) = i_0 + 2i_1\cos(\theta - \theta_i) \tag{13}$$

  Choose $1 < W_1 < 2$ and $W_0 + W_1 < 2$ with $0 < W_0 < 1$ (the dynamics lose stability if $W_1 > 2$ or $W_0 > 1$ or, approximately, $W_0 + W_1 > 2$; for $W_1 < 1$, there is no bump solution).

  a First consider a uniform input, $i_1 = 0$. Verify that for $i_0 < v_{th}$, even if you start with a random initial condition of positive activations, the dynamics will decay to $\mathbf{r} = 0$. Simulate for a couple of values of $i_0 > v_{th}$, say $i_0 = v_{th} + 1$ and $i_0 = v_{th} + 10$. Verify that if your initial condition has any nonzero (positive) noise, no matter how small, the dynamics will evolve to a bump solution (they will probably evolve to a bump solution even for an initial condition $\mathbf{r} = 0$, due to numerical noise in the simulation). There is "dynamical symmetry breaking": the dynamics and the input are circularly symmetric, but the circularly symmetric activity pattern (the uniform pattern) is unstable to any small perturbation, and the bump solution, which breaks the circular symmetry by choosing a particular location on the circle, is stable. For noisy initial conditions the bump should appear at a random location (probably selected by where some weighted sum

over the initial noise in a local region is largest, but likely to appear random to you), with a common shape and height for a given i_0. How do the shape and height change for the different values of i_0?

b Analytically, the bump activity should reach 0 at an angle ψ from the bump center, where $2W_1G_1(\psi) = 1$ and $G_1(\psi) = \frac{1}{2\pi}\left(\psi - \frac{\sin(2\psi)}{2}\right)$, with $0 < \psi < \pi$ (this is the analytic solution for continuous θ; might be slightly changed by going to a discrete grid). Does this appear to agree with your simulations? (I will place in the course directory a file, ring-model.pdf, that gives the analytics for those who are interested.)

c Now add a weak tuned input i_1, say $i_1 = 0.1(i_0 - v_{th})$. Does this choose the bump location? Does the bump appear to be otherwise similar or identical?

d Finally, simulate with the same parameters except $0 < W_1 < 1$. Now you should find that the uniform solution is stable, and there is no bump solution to a uniform input. What steady state do you arrive at for a non-uniform input (nonzero i_1), and how does it compare to the bump solution for $W_1 > 1$?

Additional things you might try (optional): explore the dynamics of the bumps in one or both of two ways:

- For the case with $i_1 = 0$: Add noise to the simulation, say adding some small i.i.d. noise to each i_i at each timestep. You should find that the steady-state bump will drift in location, roughly as a random walk meaning the distance the bump travels over some time will grow as the squareroot of the time;
- For a case with $i_1 > 0$: After the steady state is reached in response to a tuned stimulus centered at θ_i, instantaneously turn that stimulus off and turn on another tuned stimulus of the same strength at a different location. How does the bump move from one location to the other – does one bump shrink while the other grows, or does the bump rotate from one position to the other? Does this depend on whether the 2nd bump is relatively near to or far from the first? How long does the change take?

- **SSN network** We'll use the same grid of 180 positions on a ring, but now there is an E and an I cell at each position. We'll consider the ring to span 180^o, representing a preferred orientation, so the grid points have spacing 1^o. We use a power-law input/output function. We use connectivity with no "Mexican hat" – as in the model of non-normal dynamics above, we take the four connectivity functions ($E \to E$, $E \to I$, $I \to E$, $I \to I$) to have the same width, differing only in their strengths. We define these functions on the grid: the connection between the unit of type Y (E or I) at position θ_j to the unit of type X at position θ_i is

$$W_{ij}^{XY} = J^{XY} e^{-\frac{\Theta(\theta_i,\theta_j)}{2\sigma_W^2}} \tag{14}$$

Here, $\Theta()$ is the shortest distance around a circle defined in Eq. 4, above.

For parameters, use $J^{EE} = 0.044$, $J^{IE} = 0.042$, $J^{EI} = 0.023$, $J^{II} = 0.018$, $\sigma_W = 32^o$.

We take $\mathbf{r} = \begin{pmatrix} \mathbf{r}_E \\ \mathbf{r}_I \end{pmatrix}$, where $\mathbf{r}_E$ and $\mathbf{r}_I$ are the firing rates of the E and I cells respectively, both ordered in the same way around the ring (*e.g.*, from 1^o to 180^o). Our dynamical equations are

$$\mathbf{T}\tau_E \frac{d\mathbf{r}}{dt} = -\mathbf{r} + k(\mathbf{W}\mathbf{r} + \mathbf{i})_+^n \tag{15}$$

where $(\mathbf{x})_+^n$ is applied element by element and, for a given element x_i, $= x_i^n$ if $x_i > 0$; $= 0$, otherwise. We'll take $k = 0.04$ and $n = 2$. Take $\tau_E = 20ms$ and take $\mathbf{T}$ to be a diagonal element with entries 1 for the E cells and 1/2 for the I cells, *i.e.* $\tau_I = 10ms$. (The faster τ_I may not be necessary but helps ensure stability).

For an input $\mathbf{i}$ of a stimulus with direction θ_0, the input to both the E and the I units at θ_i is $i_i = ce^{-\frac{\Theta(\theta_i,\theta_0)}{2\sigma_i^2}}$. Here, c is a constant (c for 'contrast') that you will vary to vary the strength of the stimulus. Take $\sigma_i = 30^o$.

a First, for a single stimulus of orientation of your choice θ_0, simulate the response, starting from an initial condition $\mathbf{r} = 0$, for $c = \{1.25, 2.5, 5, 10, 20, 40\}$. Again, use first-order Euler, a time constant of $1ms$ should be fine. For each c, simulate until a steady state is reached by some criterion (change per timestep gets sufficiently small). For the steady state, for the E unit and the I unit at the stimulus center, plot, as a function of c:

- Their firing rate;
- Their feedforward input, their net recurrent input ($E - I$, where E is the recurrent excitatory input and I is the recurrent inhibitory input, taken to have a positive sign), and their total input (feedforward + net recurrent).
- The percent of the unit's input that is feedforward or is recurrent, counting recurrent input now as $E + I$ and total input as $FF + E + I$
- For the recurrent input, the percent of it that is excitatory: $\frac{E}{E+I}$

You should see: saturation of excitatory firing rates; a transition from a feedforward-dominated regime for weak input, to a recurrent-dominated regime for stronger input; that for stronger input, the recurrent input largely cancels or 'balances' the feedforward input; and that the recurrent input becomes more inhibition-dominated for stronger stimuli.

b Now consider adding a 2nd stimulus 90^o away from the first. By symmetry, that stimulus by itself should produce a response exactly like the response to the θ_0 stimulus, except shifted by 90^o. So you don't need to simulate response to that stimulus alone; but simulate response to the two stimuli shown at the same time, again for the given values of c (same c for both stimuli). You know by symmetry that the responses must be identical at each stimulus center. So, choosing the units at one of the stimulus centers, for the E and for the I units, plot the ratio of their steady-state response when both stimuli are shown together, to their steady-state response when only one stimulus is shown. You should find that this ratio is > 1, representing supralinear summation, for weaker inputs but < 1, representing sublinear summation, for stronger inputs.

c For at least some, if not all, of the c values, you probably want to plot, with preferred orientation from 0^o to 180^o on the x axis, the sum of the responses of the E unit to each stimulus shown alone, and its response to the two stimuli shown together; and the same for the I unit. This will allow you to directly see the supralinear and sublinear summation.

Other things you might want to try (optional):

- Give a uniform input of varying strengths to the network; do you ever see non-uniform solutions emerge? (you shouldn't)
- Consider adding the two stimuli with different c values; you should see the emergence of "winner-take-all" behavior, where the greater the difference between the c values for the two

different stimuli, the more the response to the weaker stimulus is suppressed (relative to its response if shown alone with that c value) and the more the response to the stronger stimulus approaches the response if it were shown by itself.

References

[1] Tal Kenet, Dmitri Bibitchkov, Misha Tsodyks, Amiram Grinvald, and Amos Arieli. Spontaneously emerging cortical representations of visual attributes. *Nature*, 425(6961):954–956, 2003.

[2] Joshua A. Goldberg, Uri Rokni, and Haim Sompolinsky. Patterns of ongoing activity and the functional architecture of the primary visual cortex. *Neuron*, 42(3):489 – 500, 2004.

[3] Jeffrey S. Anderson, Matteo Carandini, and David Ferster. Orientation tuning of input conductance, excitation, and inhibition in cat primary visual cortex. *Journal of Neurophysiology*, 84(2):909–926, 2000. PMID: 10938316.

[4] Luis Martinez, Jose-Manuel Alonso, R. Clay Reid, and Judith Hirsch. Laminar processing of stimulus orientation in cat visual cortex. *The Journal of physiology*, 540:321–33, 05 2002.

[5] Jorge Mariño, James Schummers, David C. Lyon, Lars Schwabe, Oliver Beck, Peter Wiesing, Klaus Obermayer, and Mriganka Sur. Invariant computations in local cortical networks with balanced excitation and inhibition. *Nature Neuroscience*, 8(2):194–201, 2005.

[6] Brendan K. Murphy and Kenneth D. Miller. Balanced amplification: A new mechanism of selective amplification of neural activity patterns. *Neuron*, 61(4):635 – 648, 2009.