6 Linear Recurrent Networks: Integrators and "Line Attractors"

6.1 A brief background of recurrent connections in brain networks

The storage of memories in the brain is an old and central issue in neuroscience. It was known from the world of digital logic that simple memory devices, like a flip-flop, could use feedback to hold electronic summing junctions in a particular state after their inputs had decayed away. These are bistable devices formed from threshold elements. It was conjectured that networks with many summing junctions, or neurons, might be able to store a multitude of states if the feedback was extended across all pairs of cells, *i.e.*, order N^2 connections across N neurons. What are the expected motifs for such circuits? By extension the idea of flip-flops, we might expect to find regions with neurons with axon collaterals that feed back to other neurons - all other in the best of worlds. These were highlighted in the perform cortex of the olfactory system by Haberly.

FIGURE - PIRIFORM CORTEX

They were also highlighted my many researchers for the CA3 region of hippocampus. This region is known for the occurrence of place cells. In their simplest substantiation these are neurons that fire only when the animal reaches a particular location in the local environment, like a box. Different cells in CA3 prefer to spike in different locations. Thus the animals builds up a map of the space, and in principle can use this map to determine a path to move from one location to another.

textbfFIGURE - HIPPOCAMPUS **FIGURE CA3 - FEW PLACE FIELDS**

So we have an idea - the use of feedback to form memories of places, or of anything by extrapolation, and we have biological motivation in terms of the anatomical evidence for recurrent networks. Let's start with the simplest system, one neuron. And a linear neuron to boot!

6.2 Positive feedback and the single neuron

Our formalism is in terms of the rate of spiking of the cell. We are dealing with linear modeling at this point so we can associate the spike rate as a linear function of the underlying potential. As such, we write differential equations directly in terms of the rate, which we denote$r(t)$,

$$\tau_0 \frac{dr(t)}{dt} + r(t) = h(t) \tag{6.6}$$

where $h(t)$ is an external input to the cell normalized in term of rate. This is the same equation for an "RC" circuit in electronics and can be readily solved, for which

$$r(t) \ = \ r(0)e^{-t/\tau_0} \ + \ \int_0^t dx e^{-(t-x)/\tau_0} h(x). \tag{6.7}$$

When the input is a constant, *i.e.*, $h(t) = h_0$, the rate will change toward that constant according to

$$r(t) \ = \ r(0)e^{-t/\tau_0} \ + \ h_0(1 - e^{-t/\tau_0}). \tag{6.8}$$

The problem is that this circuit has no memory of the initial rate, $r(0)$ or for that matter the rate at any past time, such as just after a transient input. How can we achieve memory? We consider the addition of positive feedback, where the strength of the feedback is set by w. Our rate equation is now

$$\begin{aligned}
\tau_0 \frac{dr(t)}{dt} \ + \ r(t) \ &= \ wr(t) \ + \ h(t) && (6.9)\\
\tau_0 \frac{dr(t)}{dt} \ + (1 \ - \ w) \, r(t) \ &= \ h(t) && (6.10)\\
\left(\frac{\tau_0}{1 \ - \ w}\right) \frac{dr(t)}{dt} \ + \ r(t) \ &= \ \frac{h(t)}{1 \ - \ w}
\end{aligned}$$

and we see that the time constant is no longer τ_0 but $\frac{\tau_0}{1-w}$. When w approaches a value of $w = 1$ from below, that is, from zero, we see that the effective time constant is very long. In fact, when $w = 1$ it is a perfect integrator with

$$r(t) \ = \ r(0) + \ h_0 \left(\frac{t}{\tau_0}\right). \tag{6.11}$$

Of course if the input is present for only a brief time, say T, the output just shifts from $r(t) \ = \ r(0)$ to $r(t) \ = \ r(0) + \ h_0 \left(\frac{T}{\tau_0}\right)$.

The good news is that we built an integrator - and thus a memory circuit - with linear components and positive feedback. The bad news is that w needs to be very close to $w \ = \ 1$ for the feedback to appreciably extend the time constant. Thus an extension from $\tau_0 = 100ms$ to $\tau = 10s$, as in the Robinson experiments on the stability of eye position, requires $w \ = \ 0.99$. A little variability that causes$w$ to creep up to $w \ = \ 1.01$ will lead to an unstable system.

FIGURE - AREA 1 PREMOTOR NUCLEUS

6.3 Stability in a rate based linear network

We learned that a single neuron can function as an integrator. Can we achieve the same behavior in a recurrent network? Is the stability requirement similar, or does the interconnectivity of the network somehow ease the issue. Unlikely, because a coupled linear system of N variables can be transformed to N uncoupled systems, but let's see.

$$\tau_0 \frac{dr_i(t)}{dt} \; + \; r_i(t) \; = \sum_{j=1}^{N} W_{i,j} \; r_j(t) \; + \; h_i(t) \tag{6.12}$$

In vector notion, this becomes

$$\tau_0 \frac{d\vec{\mathbf{r}}(t)}{dt} \; + \; \vec{\mathbf{r}}(t) \; = \mathbf{W}\vec{\mathbf{r}} \; + \; \vec{\mathbf{h}}(t) \tag{6.13}$$

and in steady state, for which $\vec{\mathbf{r}}^* \equiv \vec{\mathbf{r}}$,

$$0 = (\mathbf{I} - \mathbf{W}) \;\, \vec{\mathbf{r}}^* \; + \; \vec{\mathbf{h}}_\mathbf{0} \tag{6.14}$$

or

$$\vec{\mathbf{r}}^* \; = \; (\mathbf{I} - \mathbf{W})^{-1} \; \vec{\mathbf{h}}_\mathbf{0} \tag{6.15}$$

Is this a stable steady state solution? To address this, we consider a perturbation about $\vec{\mathbf{r}}^*$ and write

$$\vec{\mathbf{r}}(t) \; = \; \vec{\mathbf{r}}^* + \delta\vec{\mathbf{r}}(t) \tag{6.16}$$

Thus

$$\tau_0 \frac{d\vec{\mathbf{r}}^*}{dt} \; + \; \tau_0 \frac{d\delta\vec{\mathbf{r}}(\mathbf{t})}{dt} \; + \; \vec{\mathbf{r}}^* + \delta \; \vec{\mathbf{r}}(t) \; = \mathbf{W}\vec{\mathbf{r}}^* \; + \mathbf{W}\delta\vec{\mathbf{r}}(t) \; + \; \vec{\mathbf{h}}_\mathbf{0} \tag{6.17}$$

so that

$$\tau_0 \frac{d\delta\vec{\mathbf{r}}(\mathbf{t})}{dt} \; = - \left(\mathbf{I} - \mathbf{W}\right) \; \delta\vec{\mathbf{r}}(t). \tag{6.18}$$

Let us solve this in terns of the eignevectors of $\mathbf{W}$ rather than in terms of the individual δr_i. In general,

$$\mathbf{W}\vec{\mu}_\mathbf{i} = \lambda_i \vec{\mu}_\mathbf{i} \tag{6.19}$$

where the $\vec{\mu}_\mathbf{i}$ are eigenvectors and the λ_i are the eigenvalues. Then

$$\delta\vec{\mathbf{r}}(t) = \sum_{i}^{N} \left[\delta\vec{\mathbf{r}}(\mathbf{t})\right]_i \; \vec{\mu}_\mathbf{i} \tag{6.20}$$

where the $\left[\delta\vec{\mathbf{r}}(\mathbf{t})\right]_i \equiv \delta\vec{\mathbf{r}}(t) \cdot \vec{\mu}_i$ are expansion coefficients. Then

$$\sum_{i=1}^{N} \left(\tau_0 \frac{d\left[\delta\vec{\mathbf{r}}(\mathbf{t})\right]_i}{dt} \; + \; (1 - \lambda_i) \left[\delta\vec{\mathbf{r}}(\mathbf{t})\right]_i \right) \vec{\mu}_\mathbf{i} \; = \; 0 \tag{6.21}$$

so that except for the trivial cases $\vec{\mu}_\mathbf{i} = 0$ we have

$$\left(\frac{\tau_0}{1 - \lambda_i} \right) \frac{d\left[\delta\vec{\mathbf{r}}(\mathbf{t})\right]_i}{dt} \; + \; \left[\delta\vec{\mathbf{r}}(\mathbf{t})\right]_i \; = \; 0 \tag{6.22}$$

for each term. The system is stable if $\lambda_i \leq 1\ \forall i$. The largest eigenvector, taken as λ_1 is the integration mode if it has the largest eigenvalue at $\lambda_1 \ = \ 1$. The other modes will decay away, and suggest the need for $\lambda_i << 1$ for $i \neq 1$.

We now return to the full system and write down a general solution for $\vec{\mathbf{r}}(t)$ in terms of the eigenmodes. Let

$$\vec{\mathbf{r}}(t) = \sum_i^N \left[\vec{\mathbf{r}}(\mathbf{t})\right]_i \ \vec{\mu}_{\mathbf{i}} \tag{6.23}$$

and

$$\vec{\mathbf{h}}(t) = \sum_i^N \left[\vec{\mathbf{h}}(\mathbf{t})\right]_i \ \vec{\mu}_{\mathbf{i}} \tag{6.24}$$

where $\left[\vec{\mathbf{r}}(\mathbf{t})\right]_i \ \equiv \ \vec{\mathbf{r}}(t) \cdot \vec{\mu}_i$ and $\left[\vec{\mathbf{h}}(\mathbf{t})\right]_i \ \equiv \ \delta\vec{\mathbf{h}}(t) \cdot \vec{\mu}_i$ are time dependent expansion coefficients. Then the original equation of motion

$$\tau_0 \frac{d\vec{\mathbf{r}}(t)}{dt} \ + \ \vec{\mathbf{r}}(t) \ - \mathbf{W}\vec{\mathbf{r}}(t) \ + \ \vec{\mathbf{h}}(t) \ = 0 \tag{6.25}$$

can be written in terms on a differential equation for each eigenmode, *i.e.*,

$$\sum_i^N \left(\tau_0 \frac{d\left[\vec{\mathbf{r}}(\mathbf{t})\right]_i}{dt} \ + \ \left[\vec{\mathbf{r}}(\mathbf{t})\right]_i \ - \ \lambda_i \left[\vec{\mathbf{r}}(\mathbf{t})\right]_i \ - \ \left[\vec{\mathbf{h}}(\mathbf{t})\right]_i \right) \vec{\mu}_i \ = \ 0 \tag{6.26}$$

for which each of the individual terms must go to zero. Thus the effective time constant for the $i = th$ mode is

$$\tau_i^{\text{effective}} \ = \ \frac{\tau_0}{1 - \lambda_i}. \tag{6.27}$$

We can immediately write down the solution for the coefficients for each mode as

$$\left[\vec{\mathbf{r}}(\mathbf{t})\right]_i \ = \ \left[\vec{\mathbf{r}}(\mathbf{0})\right]_i e^{-t(1-\lambda_i)/\tau_0} \ + \ \int_0^t dx \ e^{-(t-x)(1-\lambda_i)/\tau_i} \left[\vec{\mathbf{h}}(\mathbf{x})\right]_i. \tag{6.28}$$

For the special case of $\lambda_1 = 1$ and $Re\{\lambda_i\} < 1$ for $i > 1$, the dominate mode is also a stable mode, with a firing pattern proportional to $\vec{\mu}_1$, that acts as an integrator, *i.e.*,

$$\left[\vec{\mathbf{r}}(\mathbf{t})\right]_1 \ = \ \left[\vec{\mathbf{r}}(\mathbf{0})\right]_1 + \int_0^t dx \ \left[\vec{\mathbf{h}}(\mathbf{x})\right]_1. \tag{6.29}$$

This gives us an idea for eye movement. Here we want all of th modes except the integrator mode to decay quickly, *i.e.*, $\lambda_1 = 1$ and $Re\{\lambda_i\} \ll 1$ for $i > 1$

FIGURE - EIGENSPECTRUM WITH A GAP BETWEEN STATE AT $\lambda = 1$ AND STATES WITH $\lambda << 1$

We assume that eye position, denoted $\theta(t)$, is proportional to a single firing pattern, which makes good sense when that pattern is stable and all others rapidly decay. In fact, this concept makes good sense for any motor act that requires extended stability, such as posture. With reference to angular position, we write

$$\begin{aligned}\theta(t) &= G\ [\vec{\mathbf{r}}(\mathbf{t})]\ \cdot\ \vec{\mu}_1\ +\ \theta_0 && (6.30)\\ &= G\ [\vec{\mathbf{r}}(\mathbf{t})]_1 +\ \theta_0 && (6.31)\\ &= G\ \int_0^t dx\ \left[\vec{\mathbf{h}}(\mathbf{x})\right]_1\ +\ G\ [\vec{\mathbf{r}}(\mathbf{0})]_1 +\ \theta_0\end{aligned}$$

where G is a gain factor and θ_0 is the baseline position of the eye. One could add all kinds of baseline rates, but this just ofiscates the story. The key is that the eye position now follows the integrator mode.

This model is called a line attractor. The name was coined since the stable output is proportional to a single vector, $\vec{\mu}_1$, but the continuum of points along that vector forms a line in the N-dimensional space of firing rates of the different cells. Changes to $[\vec{\mathbf{r}}(\mathbf{t})]_1$ that result from inputs along the direction of $\vec{\mu}_1$ are along the line. Inputs that are orthogonal to this line rapidly decay so the system returns to the line.

FIGURE - LINE ATTRACTOR LANDSCAPE

FIGURES - PREMOTOR NUCLEUS DYNAMICS

6.4 Absence of multistability in linear recurrent networks

Before we move on to non-linear network, we consider the question of how many stable patterns a linear network can support. The results of the integrator suggest only one memory, but let's see if we can get a general proof.

We consider a network with a symmetric weight matrix, $\mathbf{W}$, *i.e.*, a matrix of synaptic connections so that $W_{i,j}$ is the strength of the input to cell i from the output of cell j. The neurons act as linear devices, *i.e.*, the output of the cell is a linear function of the input. This clearly is not the case for cells that vary from quiescent to spiking as a function of their input, but could be the case for cells whose spike rate is uniformly and monotonically modulated up and down. It is also the case for networks of cells with solely graded synaptic release.

Since we are working in the linear regime, we again ignore the difference between cell potential and firing rat and write the input to the cell as

$$r_i(t)\ =\ \sum_{j=1}^{N} W_{ij} r_j(t) \qquad (6.32)$$

where N is the number of neurons. We assume a parallel, clocked updating scheme, in which we explicitly note the time steps, *i.e.*,

$$r_i(t+1)\ =\ \sum_{j=1}^{N} W_{ij} r_j(t). \qquad (6.33)$$

In vector notation, this is

$$\vec{\mathbf{r}}(t+1) = \mathbf{W} \cdot \vec{\mathbf{r}}(t) \tag{6.34}$$

Now we can iterate, the synchronous equivalent of recurrence, starting from time $t = 0$:

$$\begin{aligned} \vec{\mathbf{r}}(1) &= \mathbf{W} \cdot \vec{\mathbf{r}}(0) \\ \vec{\mathbf{r}}(2) &= \mathbf{W} \cdot \vec{\mathbf{r}}(1) \\ \vec{\mathbf{r}}(3) &= \mathbf{W} \cdot \vec{\mathbf{r}}(2) \\ &\cdot \\ &\cdot \\ \vec{\mathbf{r}}(n+1) &= \mathbf{W} \cdot \vec{\mathbf{r}}(n) \end{aligned} \tag{6.35}$$

This becomes

$$\vec{\mathbf{r}}(n) = \mathbf{W}^n \cdot \vec{\mathbf{r}}(0) \tag{6.36}$$

Now we recall that $\mathbf{W}$ satisfies an eigenvalue equation;

$$\mathbf{W} \cdot \vec{\mu}_k = \lambda_k \vec{\mu}_k \tag{6.37}$$

where k labels labels the eigenvalue and for a real symmetric $\mathbf{W}$ we have $1 < k < N$ if we ignore potential degenerate eigenvectors. We can rotate the symmetric matrix $\mathbf{W}$ by a unity transformation that preserves the eignenvalues, *i.e.*,

$$\mathbf{W} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathbf{T}} \tag{6.38}$$

where $\mathbf{U}$ is a unitary matrix defined through $\mathbf{U} \cdot \mathbf{U}^{\mathbf{T}} = \mathbf{I}$. The diagonal matrix $\mathbf{\Lambda}$ contains the eigenvalues along the diagonal, such that

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & 0 & \\ 0 & 0 & \lambda_3 & \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \end{pmatrix}$$

Clearly the rotated eigenvectors, $\mathbf{U}^{\mathbf{T}}\vec{\mu}$, are of the form

$$\mathbf{U}^{\mathbf{T}}\vec{\mu}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \end{pmatrix} \qquad \mathbf{U}^{\mathbf{T}}\vec{\mu}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \end{pmatrix} \cdots$$

since $\mathbf{W} \cdot \vec{\mu}_k = \lambda_k \vec{\mu}_k$ implies $\mathbf{\Lambda} \cdot \mathbf{U^T} \vec{\mu}_k = \lambda_k \mathbf{U^T} \vec{\mu}_k$ so the $\mathbf{U^T} \vec{\mu}_k$ are the eigenvalues of the diagonalized (rotated) system.
Now we can go back to the iterative expression for $\vec{\mathbf{r}}(n)$.

$$\begin{aligned} \vec{\mathbf{r}}(n) \quad &= \mathbf{W}^n \cdot \vec{\mathbf{r}}(0) \\ &= \left(\mathbf{U\Lambda U^T}\right)^n \cdot \vec{\mathbf{r}}(0) \\ &= \mathbf{U\Lambda^n U^T} \cdot \vec{\mathbf{r}}(0) \end{aligned} \tag{6.39}$$

where we used

$$\begin{aligned} \left(\mathbf{U\Lambda U^T}\right)^n \quad &= \mathbf{U\Lambda U^T U\Lambda U^T} \cdots \mathbf{U\Lambda U^T} \\ &= \mathbf{U\Lambda^n U^T} \end{aligned} \tag{6.40}$$

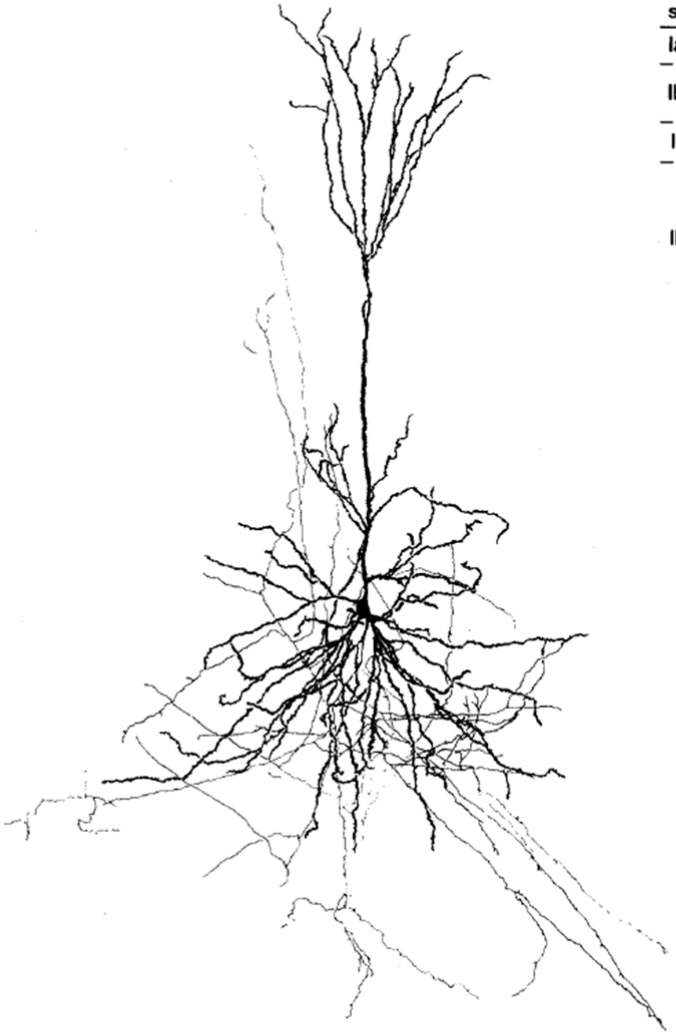
Thus

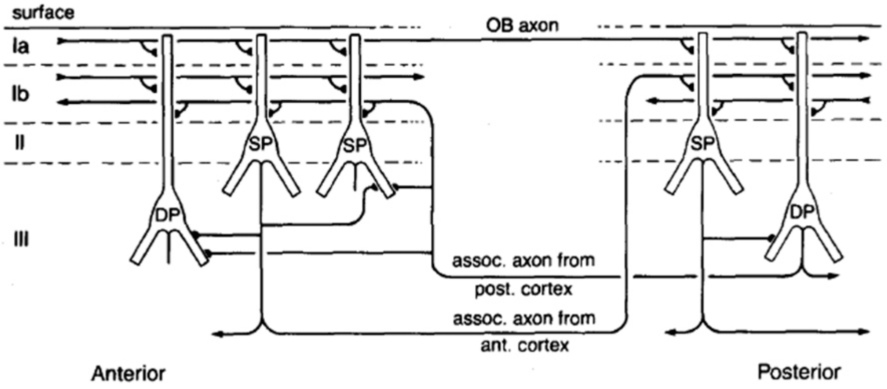
$$\mathbf{U^T}\ \vec{\mathbf{r}}(n) = \mathbf{\Lambda}^n\ \mathbf{U^T}\vec{\mathbf{r}}(0) \tag{6.41}$$

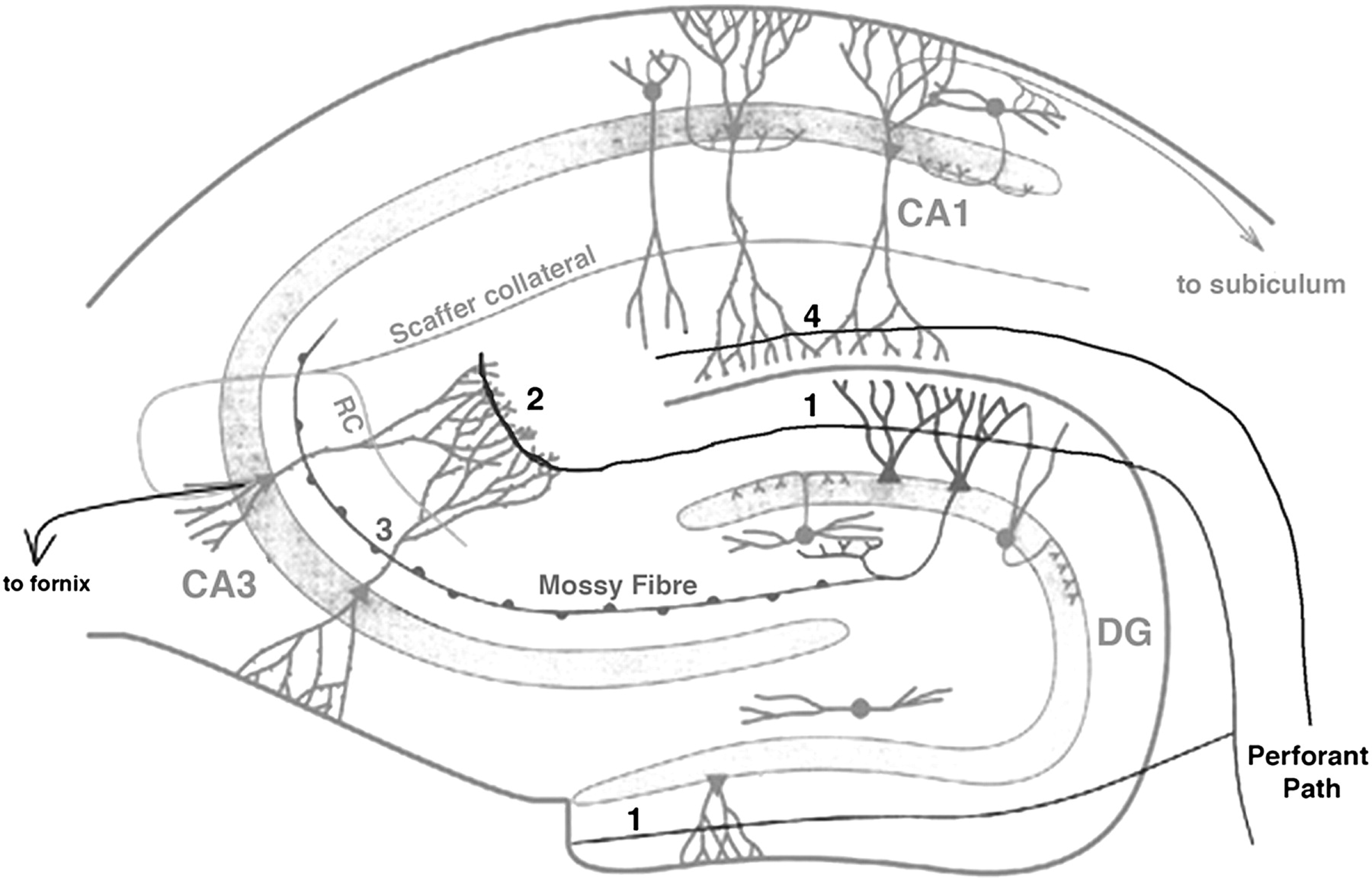
But the diagonal matrix $\mathbf{\Lambda^n}$, when rank ordered so that λ_1 is the dominant eigenvalue, becomes,

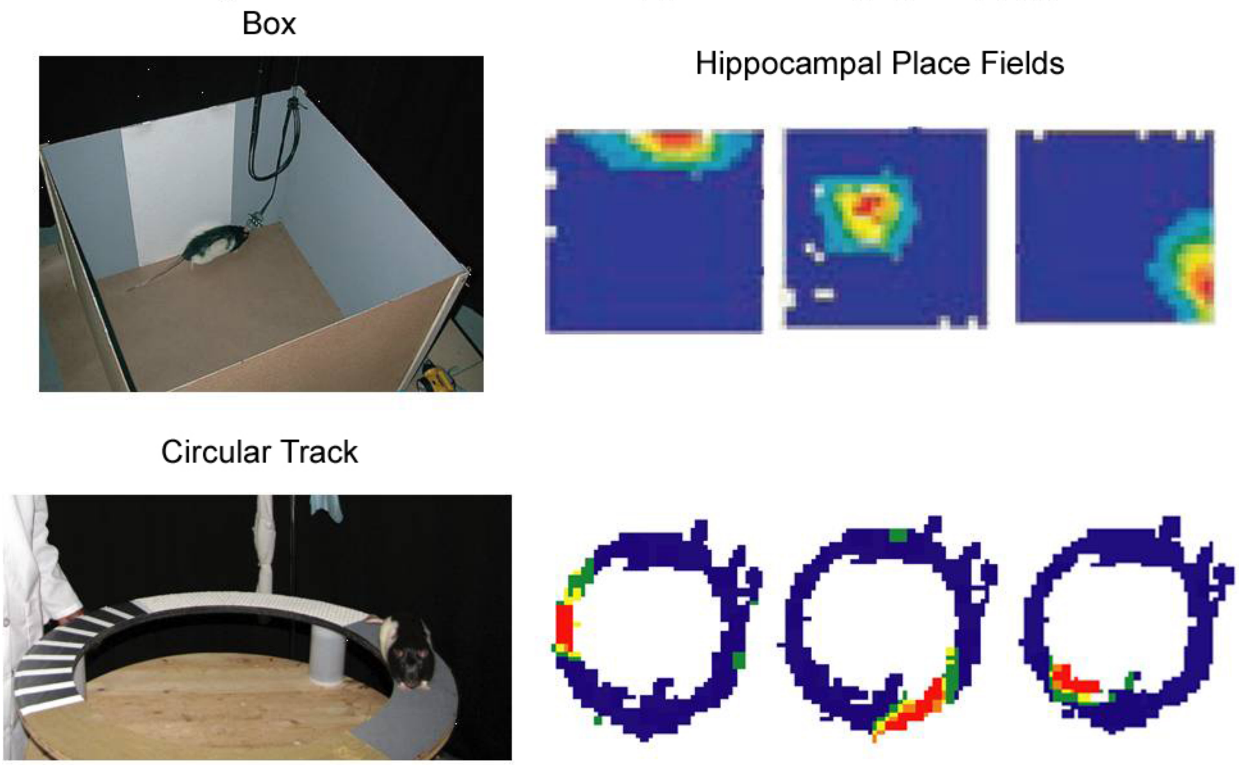
$$\mathbf{\Lambda}^n = \begin{pmatrix} \lambda_1^n & 0 & 0 & \cdots \\ 0 & \lambda_2^n & 0 & \\ 0 & 0 & \lambda_3^n & \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \end{pmatrix} = \lambda_1^n \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 0 & \left(\frac{\lambda_2}{\lambda_1}\right)^n & 0 & \\ 0 & 0 & \left(\frac{\lambda_3}{\lambda_1}\right)^n & \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \end{pmatrix} \rightarrow \lambda_1^n \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \\ 0 & 0 & 0 & \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \end{pmatrix}$$

Thus the system converges to the dominant eigenvector, $\vec{\mathbf{r}}_1 = \mathbf{U^T}\vec{\mu}_1$, and eigenvalue, λ_1, independent of the initial starting state. Thus only a single state is supported in an iterative network with linear neurons. The stability of this state depends of the sign of λ_1. Nonetheless, the essential issue is that neurons that function as linear transducers can support a single stable state. This can still make these useful as an integrator, as proposed for a model of the ocular motor system. But linear networks will not be useful as associative networks that store many patterns.

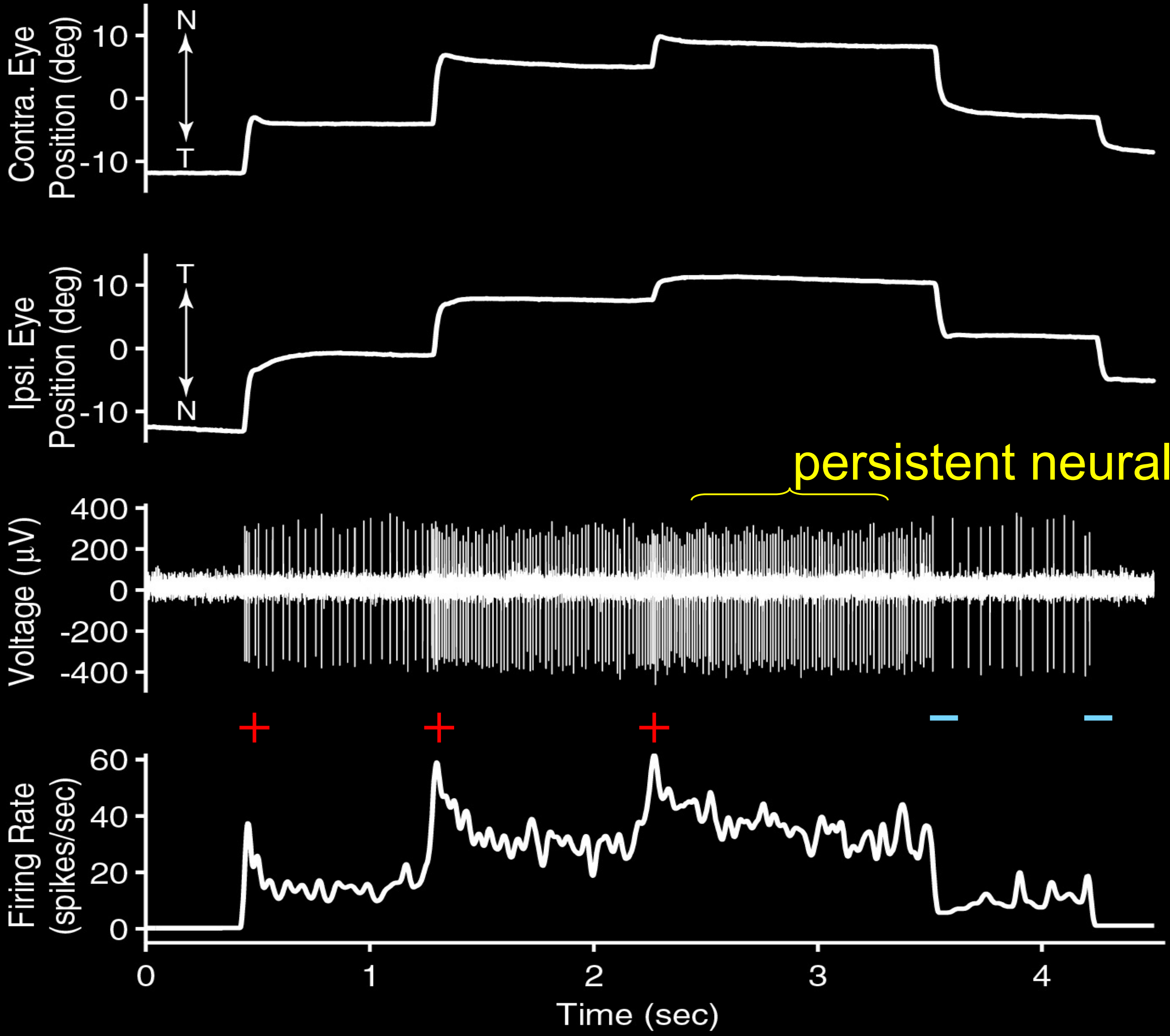



surface
OB axon
Ia
Ib
II
III
SP
SP
SP
DP
DP
assoc. axon from
post. cortex
assoc. axon from
ant. cortex
Anterior
Posterior


CA1
to subiculum
Scaffer collateral
4
2
RC
1
3
to fornix
CA3
Mossy Fibre
DG
Perforant
Path
1


Box
Hippocampal Place Fields
Circular Track

Neural Recording from the Oculomotor Integrator

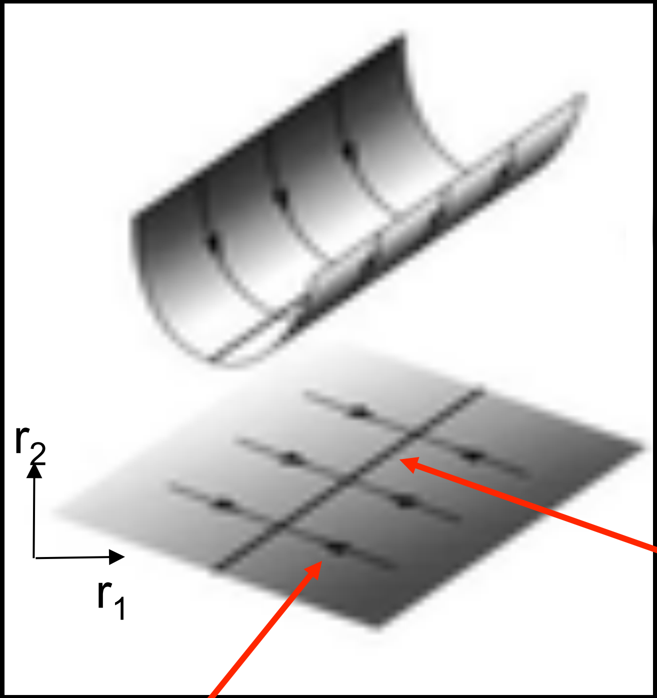


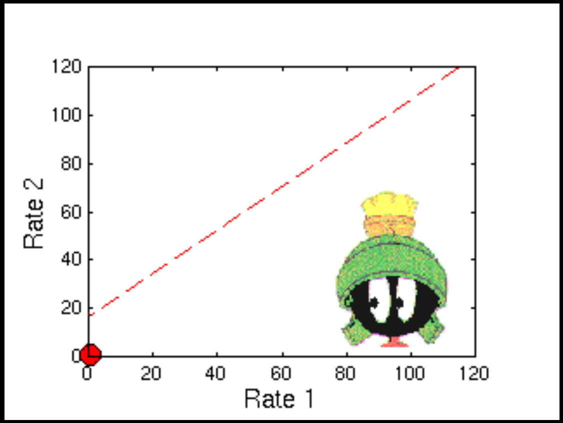
persistent neural activity

On/off
burst
input

Line Attractor Picture of the Neural Integrator

Geometrical picture of eigenvectors:



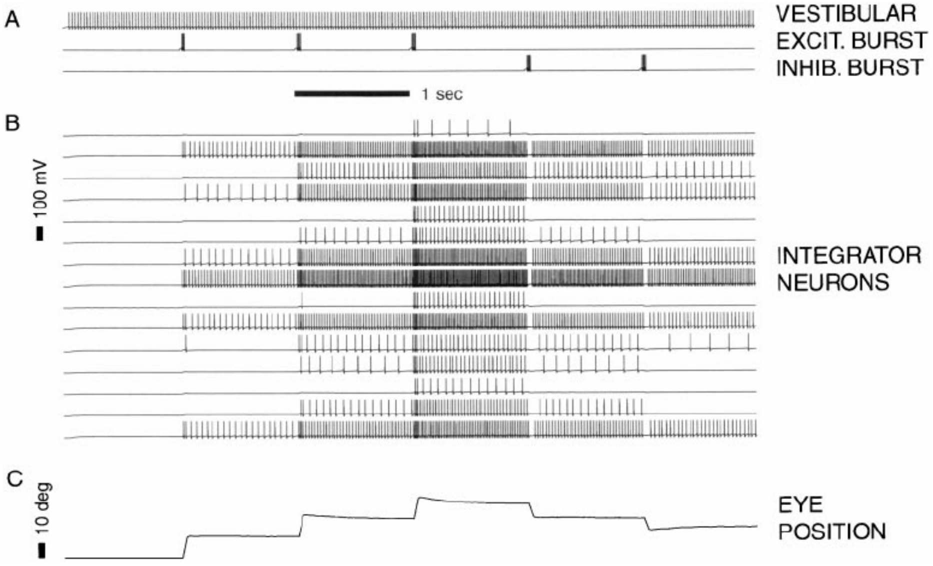


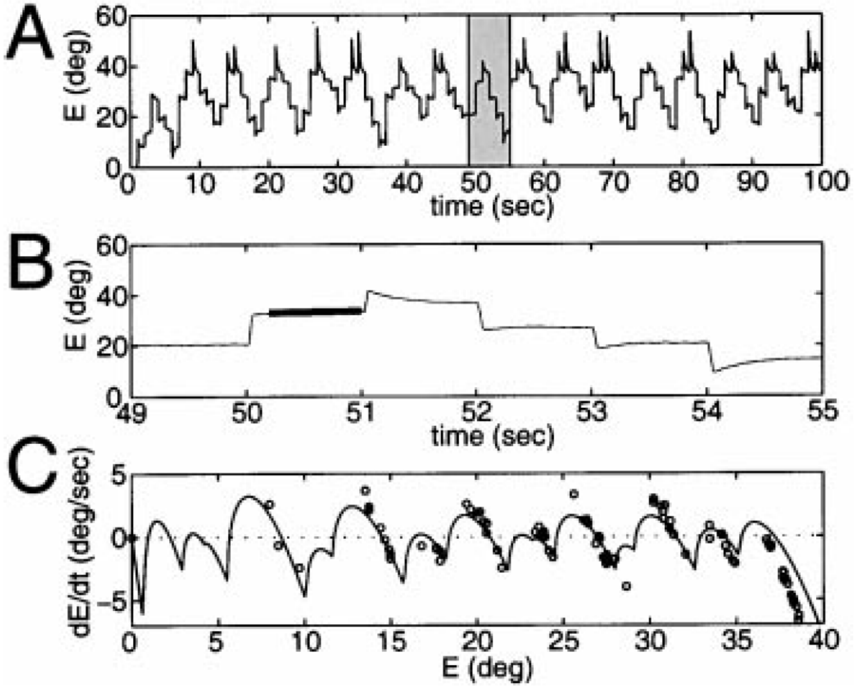
No decay along direction of eigenvector with eigenvalue = 1

Decay along direction of eigenvectors with eigenvalue < 1

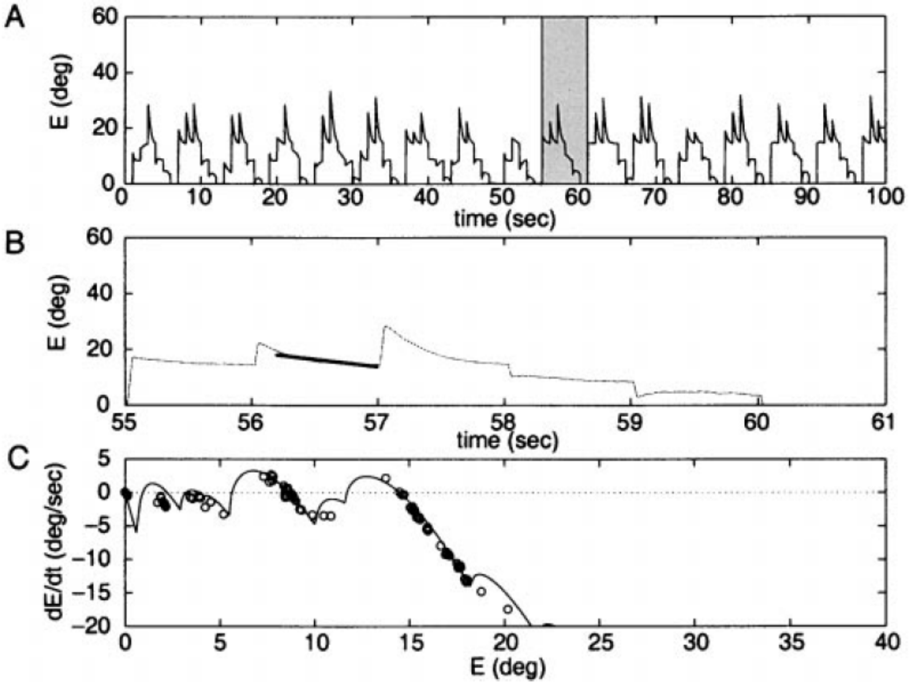
“Line Attractor” or “Line of Fixed Points”

Seung Integrator Model (15 neurons)

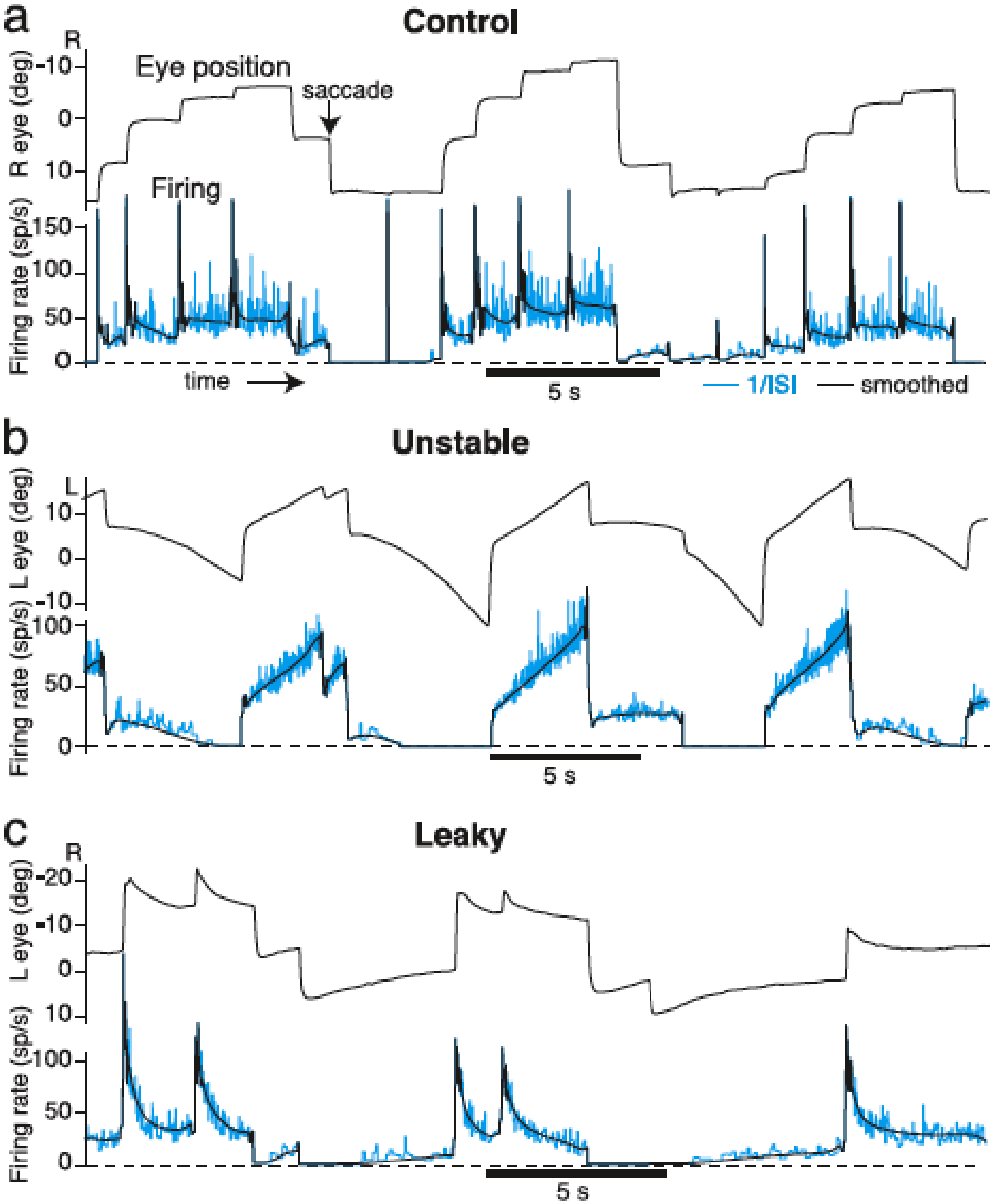




Death of a neuron!



Stability versus time constant



TC ~ 20 s

TC ~ 1 s

TC ~ 1 s