6 Linear recurrent networks: Integrators, i.e., line attractors, and the absence of multi-stability

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 since the 1930's that bistable devices formed from threshold elements, like a digital flip-flop, could be built using feedback to hold electronic summing junctions in a particular state after their inputs had decayed away. By the 1970’s, 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 of the idea of flip-flops, we might expect to find regions of the brain with neurons whose axon collaterals feed back onto other neurons. This anatomical arrangement was highlighted for the perform cortex of the olfactory system in a ca 1980’s paper by Haberly and by other researchers for the CA3 region of hippocampus. It was explored theoretically starting in the 1970s and culminated with a pivotal contribution by Hopfield in 1982 and an analysis of Hopfield’s model by Amit, Gutfriend and Sompolinsky, by Gardner, and by others in the mid 1980s.

The hippocampus seemed a particularly valuable region to consider feedback, as it 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. Different cells 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.

So we have an idea - the use of feedback connections to form memories of many places, or of anything by extrapolation, and we have biological motivation, in terms of the anatomical evidence, to understand the dynamics of such networks as well as search for them in real nervous systems.

6.2 Absence of multistability in linear recurrent networks

The simplest question is how many stable patterns a linear network can support. 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. Let’s see if we can get a general proof of how many
states such a network can support and, if there is something interesting, derive the design rule that relates the desired output to the underlying connectivity.

To begin, we consider a network with a symmetric weight matrix, $W$, i.e., a matrix of synaptic connections so that $W_{ij}$ 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. Since we are working in the linear regime, we can ignore the difference between cell potential and firing rate and write the input to the cell as

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

(6.6)

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

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

(6.7)

In vector notation, this is

$$ \vec{r}(t) = W \cdot \vec{r}(t-1). $$

(6.8)

We now iterate, the synchronous equivalent of recurrence, starting from time $t = 0$:

$$ \vec{r}(1) = W \cdot \vec{r}(0) $$

(6.9)

$$ \vec{r}(2) = W \cdot \vec{r}(1) $$

$$ \vec{r}(3) = W \cdot \vec{r}(2) $$

$$ \vdots $$

$$ \vec{r}(n) = W \cdot \vec{r}(n-1). $$

This becomes

$$ \vec{r}(n) = W^n \cdot \vec{r}(0). $$

(6.10)

Now we recall that a matrix $W$ satisfies an eigenvalue equation

$$ W \cdot \vec{\mu}_k = \lambda_k \vec{\mu}_k $$

(6.11)

where $k$ labels the eigenvalue with $k = 1, \ldots, N$ and includes the case of potential degenerate eigenvectors. The eigenvalues are real numbers when $W$ is symmetric matrix whose elements are real. The spectral theorem states that a symmetric matrix whose elements are real can be diagonalized by an orthogonal matrix. Thus we can rotate $W$ by a unity transformation that preserves the eigenvalues, i.e.,

$$ W = U \Lambda U^T $$

(6.12)
where \( \mathbf{U} \) is a unitary matrix defined through \( \mathbf{U}\mathbf{U}^T = \mathbf{I} \) and \( \det(\mathbf{U}) = 1 \). Each column in \( \mathbf{U} \) is one of the eigenvectors \( \vec{\mu}_k \), i.e.,

\[
\mathbf{U} = \begin{pmatrix}
\cdots & \cdots & \cdots \\
\vec{\mu}_1 & \vec{\mu}_2 & \cdots \vec{\mu}_N \\
\cdots & \cdots & \cdots \\
\end{pmatrix}
\quad \text{and} \quad
\mathbf{U}^T = \begin{pmatrix}
\cdots & \cdots & \cdots \\
\vec{\mu}_1 & \cdots & \cdots \\
\vec{\mu}_2 & \cdots & \cdots \\
\vec{\mu}_3 & \cdots & \cdots \\
\cdots & \cdots & \cdots \\
\vec{\mu}_N & \cdots & \cdots \\
\end{pmatrix}
\]

and the rotated eigenvectors, \( \mathbf{U}^T \vec{\mu} \), are of the form

\[
\mathbf{U}^T \vec{\mu}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \end{pmatrix}, \quad \mathbf{U}^T \vec{\mu}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \end{pmatrix}, \quad \mathbf{U}^T \vec{\mu}_3 = \begin{pmatrix} \ddots \end{pmatrix}
\]

since \( \mathbf{W} \vec{\mu}_k = \lambda_k \vec{\mu}_k \) implies \( \mathbf{A} \mathbf{U}^T \vec{\mu}_k = \lambda_k \mathbf{U}^T \vec{\mu}_k \) so the \( \mathbf{U}^T \vec{\mu}_k \) are the eigenvectors of the diagonalized (rotated) system. The diagonal matrix \( \mathbf{A} \) contains the eigenvalues along the diagonal, such that

\[
\mathbf{A} = \begin{pmatrix}
\lambda_1 & 0 & 0 & \cdots \\
0 & \lambda_2 & 0 & \cdots \\
0 & 0 & \lambda_3 & \cdots \\
\vdots & \vdots & \vdots & \ddots \\
\end{pmatrix}
\]

We return to the iterative expression for \( \vec{r}(n) \), i.e.,

\[
\vec{r}(n) = \mathbf{W}^n \vec{r}(0) = (\mathbf{U}\mathbf{A}\mathbf{U}^T)^n \vec{r}(0) = \mathbf{U}\mathbf{A}^n\mathbf{U}^T \vec{r}(0)
\]

where we used

\[
(\mathbf{U}\mathbf{A}\mathbf{U}^T)^n = \mathbf{U}\mathbf{A}^n\mathbf{U}^T \mathbf{U}\mathbf{A}\mathbf{U}^T \cdots \mathbf{U}\mathbf{A}\mathbf{U}^T = \mathbf{U}\mathbf{A}^n\mathbf{U}^T.
\]

But the diagonal matrix \( \mathbf{A}^n \), when rank ordered so that \( \lambda_1 \) is the dominant eigenvalue, becomes

\[
\mathbf{A}^n = \begin{pmatrix}
\lambda_1^n & 0 & 0 & \cdots \\
0 & \lambda_2^n & 0 & \cdots \\
0 & 0 & \lambda_3^n & \cdots \\
\vdots & \vdots & \vdots & \ddots \\
\end{pmatrix} = \lambda_1^n \begin{pmatrix}
1 & 0 & 0 & \cdots \\
0 & (\lambda_2/\lambda_1)^n & 0 & \cdots \\
0 & 0 & (\lambda_3/\lambda_1)^n & \cdots \\
\vdots & \vdots & \vdots & \ddots \\
\end{pmatrix} \xrightarrow{n \to \infty} \lambda_1^n \begin{pmatrix}
1 & 0 & 0 & \cdots \\
0 & 0 & 0 & \cdots \\
0 & 0 & 0 & \cdots \\
\vdots & \vdots & \vdots & \ddots \\
\end{pmatrix}
\]
Thus the system converges to a numerical factor times the dominant eigenvector of $W$, i.e.,

$$
\vec{r}(n) \overset{n \to \infty}{\longrightarrow} \lambda_1^n U \begin{pmatrix}
1 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & \ddots \\
\vdots \\
\end{pmatrix} U^T \vec{r}(0) = \lambda_1^n [\vec{\mu}_1 \cdot \vec{r}(0)] \vec{\mu}_1.
$$

Thus only a single state is supported in an iterative network with linear neurons. The stability of this state depends of the sign of $\lambda_1$.

The essential issue is that neurons that function as linear transducers can support only a single stable state. As such, a design rule is to pick a desired stable state, which we will call $\vec{\zeta}$, and compute the weight matrix for our linear network as the outer product

$$W = \vec{\zeta} \vec{\zeta}^T. \quad (6.15)$$

While linear networks will not be useful as associative networks that intrinsically store many patterns, they can be shown to be useful for the particular problem of making a circuit that integrates an input. This comes up for the case of motor systems and was proposed for ocular motor control, first by Robinson and later, in a more general form, by Seung.

### 6.3 Positive feedback and the single neuron

We start with the case of one cell to learn about the importance of integration. Our formalism is in terms of the rate of spiking of the cell. Since we are dealing with linear modeling at this point, we can associate the spike rate with the underlying potential. As such, we write differential equations directly in terms of the rate, which we denote $r(t)$,

$$
\tau_0 \frac{dr(t)}{dt} + r(t) = h(t) \quad (6.16)
$$

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

$$
r(t) = r(0)e^{-t/\tau_0} + \int_0^t \frac{dx}{\tau_0} e^{-(t-x)/\tau_0} h(x). \quad (6.17)
$$

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

$$
r(t) = r(0)e^{-t/\tau_0} + h_0(1 - e^{-t/\tau_0}). \quad (6.18)
$$

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 the scalar constant $w$. Our rate equation is now

$$
\tau_0 \frac{dr(t)}{dt} + r(t) = wr(t) + h(t) \tag{6.19}
$$

$$
\tau_0 \frac{dr(t)}{dt} + (1 - w) r(t) = h(t)
$$

$$
\left( \frac{\tau_0}{1 - w} \right) \frac{dr(t)}{dt} + r(t) = \frac{h(t)}{1 - w}
$$

and we see that the time constant is no longer $\tau_0$ 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

$$
r(t) = r(0) + h_0 \left( \frac{t}{\tau_0} \right). \tag{6.20}
$$

Thus 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 - which is 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 $\tau_0 = 100$ ms to $\tau = 10$ s, 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.

### 6.4 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 linear network? There will be only a single attractor, since linear networks only support one stable state, and we wish to make this the integrator mode. Thus we expect that the path forward will be to transform the set of $N$ coupled linear variables into $N$ uncoupled systems. One of these will be the integrator mode and will must have the largest eigenvalue. Let’s see what other constraints arise. We start with

$$
\tau_0 \frac{dr_i(t)}{dt} + r_i(t) = \sum_{j=1}^{N} W_{i,j} r_j(t) + h_i(t). \tag{6.21}
$$

In vector notion, this becomes

$$
\tau_0 \frac{d\vec{r}(t)}{dt} + \vec{r}(t) = W \vec{r}(t) + \vec{h}(t) \tag{6.22}
$$

and in steady state, for which $\vec{r}(t) \equiv \vec{r}_0$,

$$
0 = (I - W) \vec{r}_0 - \vec{h}_0 \tag{6.23}
$$
or
\[ \mathbf{r}_0 = (\mathbf{I} - \mathbf{W})^{-1} \mathbf{h}_0 \] (6.24)

Is this a stable steady state solution? To address this, we consider a perturbation about \( \mathbf{r}_0 \) and write
\[ \mathbf{r}(t) = \mathbf{r}_0 + \delta\mathbf{r}(t) \] (6.25)

Thus
\[ 0 + \tau_0 \frac{d\delta\mathbf{r}(t)}{dt} + \mathbf{r}_0 + \delta\mathbf{r}(t) = \mathbf{W}\mathbf{r}_0 + \mathbf{W}\delta\mathbf{r}(t) + \mathbf{h}_0 \] (6.26)

so that
\[ \tau_0 \frac{d\delta\mathbf{r}(t)}{dt} = -(\mathbf{I} - \mathbf{W}) \delta\mathbf{r}(t). \] (6.27)

Let us solve this in terms of the eigenvectors of \( \mathbf{W} \) rather than in terms of the individual \( \delta r_i(t) \). In general,
\[ \mathbf{W}\mathbf{\tilde{u}}_i = \lambda_i \mathbf{\tilde{u}}_i \] (6.28)

where the \( \mathbf{\tilde{u}}_i \) are eigenvectors and the \( \lambda_i \) are the eigenvalues. Then
\[ \delta\mathbf{r}(t) = \sum_{i=1}^{N} [\delta\mathbf{r}(t)]_i \mathbf{\tilde{u}}_i \] (6.29)

where the \([\delta\mathbf{r}(t)]_i \equiv \delta\mathbf{r}(t) \cdot \mathbf{\tilde{u}}_i \) are expansion coefficients. Then
\[ \sum_{i=1}^{N} \left( \tau_0 \frac{d [\delta\mathbf{r}(t)]_i}{dt} + (1 - \lambda_i) [\delta\mathbf{r}(t)]_i \right) \mathbf{\tilde{u}}_i = 0 \] (6.30)

so that except for the trivial cases \( \mathbf{\tilde{u}}_i = 0 \) we have
\[ \left( \frac{\tau_0}{1 - \lambda_i} \right) \frac{d [\delta\mathbf{r}(t)]_i}{dt} + [\delta\mathbf{r}(t)]_i = 0 \] (6.31)

for each term. The system is stable if \( \lambda_i \leq 1 \ \forall i \). The largest eigenvector, taken without loss of generality as \( \lambda_1 \), is the integration mode if it has the largest possible eigenvalue \( \lambda_1 = 1 \). 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 \( \mathbf{r}(t) \) in terms of the eigenmodes. Let
\[ \mathbf{r}(t) = \sum_{i=1}^{N} [\mathbf{r}(t)]_i \mathbf{\tilde{u}}_i \] (6.32)

and
\[ \mathbf{h}(t) = \sum_{i=1}^{N} [\mathbf{h}(t)]_i \mathbf{\tilde{u}}_i \] (6.33)

where \([\mathbf{r}(t)]_i \equiv \mathbf{r}(t) \cdot \mathbf{\tilde{u}}_i \) and \([\mathbf{h}(t)]_i \equiv \mathbf{h}(t) \cdot \mathbf{\tilde{u}}_i \) are time dependent expansion coefficients. Then the original equation of motion
\[ \tau_0 \frac{d\mathbf{r}(t)}{dt} + \mathbf{r}(t) - \mathbf{W}\mathbf{r}(t) + \mathbf{h}(t) = 0 \] (6.34)
can be written in terms on a differential equation for each eigenmode, i.e.,

\[
\sum_i^N \left( \tau_0 \frac{d[\mathbf{r}(t)]_i}{dt} + [\mathbf{r}(t)]_i - \lambda_i [\mathbf{r}(t)]_i - [\mathbf{h}(t)]_i \right) \mu_i = 0 \tag{6.35}
\]

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

\[
\tau_{i_{\text{effective}}} = \frac{\tau_0}{1 - \lambda_i}. \tag{6.36}
\]

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

\[
[\mathbf{r}(t)]_i = [\mathbf{r}(0)]_i e^{-t(1-\lambda_i)/\tau_0} + \int_0^t \frac{dx}{\tau_0} e^{-(t-x)(1-\lambda_i)/\tau_0} [\mathbf{h}(x)]_i. \tag{6.37}
\]

For the special case of \(\lambda_1 = 1\) and \(\text{Re}\{\lambda_i\} < 1\) for \(i > 1\), the dominate mode is also a stable mode, with a firing pattern proportional to \(\mu_1\), that acts as an integrator, i.e.,

\[
[\mathbf{r}(t)]_1 = [\mathbf{r}(0)]_1 + \int_0^t \frac{dx}{\tau_0} [\mathbf{h}(x)]_1. \tag{6.38}
\]

How do we make a connection matrix, \(\mathbf{W}\)? Our constraint is on the eigenvalues, i.e.,

\[
\Lambda = \begin{pmatrix}
1 & 0 & 0 & \cdots \\
0 & \lambda_2 & 0 & \\
0 & 0 & \lambda_3 & \\
& & & \\
& & & \\
& & & \\
& & & & 1 \\
& & & & 0 & \lambda \\
\end{pmatrix}
\]

where \(1 >> \lambda_2 > \lambda_3 > \cdots \lambda_N \geq 0\). As a concrete example, we take a system with two neurons, i.e.,

\[
\Lambda = \begin{pmatrix}
1 & 0 \\
0 & \lambda \\
\end{pmatrix}
\]

with \(1 >> \lambda \geq 0\) and \(\mathbf{U}\) as the rotation matrix

\[
\mathbf{U} = \begin{pmatrix}
\cos \eta & \sin \eta \\
-\sin \eta & \cos \eta \\
\end{pmatrix}
\]

Then

\[
\mathbf{W} = \mathbf{U} \Lambda \mathbf{U}^T = \begin{pmatrix}
1 - (1 - \lambda) \sin^2 \eta & -\left(\frac{1-\lambda}{2}\right) \sin 2\eta \\
-\left(\frac{1-\lambda}{2}\right) \sin 2\eta & 1 - (1 - \lambda) \cos^2 \eta
\end{pmatrix},
\]

which is in the form of two neurons that have excitatory self-feedback, as in the single-cell integrator, but mutual inhibition. There are an infinity of such networks since \(\eta\) is a continuous variable, but the signs of the synaptic connections are fixed.

Next, we need the eigenvector \(\mu_i\) to form the expansion coefficients \([\mathbf{r}(t)]_1\) and \([\mathbf{h}(x)]_1\). Thus

\[
\mu_1 = \mathbf{U} \begin{pmatrix}
1 \\
0
\end{pmatrix} = \begin{pmatrix}
\cos \eta \\
-\sin \eta
\end{pmatrix}.
\]
As a last step, we relate the neuronal output to eye position. 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

$$\theta(t) = G \bar{r}(t) \cdot \bar{\mu}_1 + \theta_0$$

(6.39)

where $G$ is a gain factor and $\theta_0$ is the baseline position of the eye. The key is that the eye position now follows the integrator mode.

This model is called a line attractor. The stable output is proportional to a single vector, $\bar{\mu}_1$, 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 $\bar{r}(t)$ that result from inputs along the direction of $\bar{\mu}_1$ are along the line. Inputs that are orthogonal to this line have eigenvalues closer to 0 than 1 and rapidly decay so the system returns to the line.