Episodic Memory Theory of RNNs

Energy Analysis

Arjun Karuvally

1 System - Derivation from Continuous Time Dynamics

The general system of neurons is governed by the following equations. The state variables of the dynamical system are $V_f \in \mathbb{R}^{N_f \times 1}, V_h \in \mathbb{R}^{N_h \times 1}, V_d \in \mathbb{R}^{N_f \times 1}$. The interactions are represented by $\Xi \in \mathbb{R}^{N_f \times N_h}$ and $\Phi \in \mathbb{R}^{N_h \times N_h}$. Φ represents synaptic strength *from* V_h *to* V_h.

$$\begin{cases} \mathcal{T}_f \dfrac{\mathrm{d}V_f}{\mathrm{d}t} = & \sqrt{\alpha_s}\, \Xi\, \sigma_h(V_h) - V_f\,, \\ \mathcal{T}_h \dfrac{\mathrm{d}V_h}{\mathrm{d}t} = & \sqrt{\alpha_s}\, \Xi^\top \sigma_f(V_f) + \alpha_c \Phi^\top \Xi^\top V_d - V_h\,, \\ \mathcal{T}_d \dfrac{\mathrm{d}V_d}{\mathrm{d}t} = & \sigma_f(V_f) - V_d\,. \end{cases} \tag{1}$$

The system has an energy function given by

$$E = \left[V_f^\top \sigma_f(V_f) - L_f \right] + \left[V_h^\top \sigma_h(V_h) - L_h \right] - \left[\sqrt{\alpha_s}\, \sigma_f(V_f)^\top\, \Xi\, \sigma_h(V_h) \right] - \alpha_c \left[V_d^\top\, \Xi\, \Phi \sigma_h(V_h) \right] \tag{2}$$

Conditions:

- $\mathcal{T}_h \to 0$
- $\mathcal{T}_d \to 0$
- discretize time for V_f
- $\sigma_h(X) = X$
- $\alpha_s = \alpha_c = 1$
- $\mathcal{T}_f = 1$

$\sigma_h(X) = X \implies L_h = \frac{1}{2} V_h^\top V_h$

2 Deriving governing equations

From a given time t, the update equations are given as

$$\begin{cases} \mathcal{T}_f(V_f(t+1) - V_f(t)) = & \Xi\,\sigma_h(V_h(t)) - V_f(t)\,, \\ V_h(t) = & \Xi^\top\,\sigma_f(V_f(t)) + \Phi^\top\Xi^\top V_d(t)\,, \\ V_d(t) = & \sigma_f(V_f(t))\,. \end{cases} \tag{3}$$

$$\begin{cases} \mathcal{T}_f(V_f(t+1) - V_f(t)) = & \Xi\,\sigma_h(V_h) - V_f(t)\,, \\ V_h(t) = & \Xi^\top\,\sigma_f(V_f(t)) + \Phi^\top\Xi^\top\sigma_f(V_f)\,, \end{cases} \tag{4}$$

$$\begin{cases} \mathcal{T}_f(V_f(t+1) - V_f(t)) = & \Xi\,V_h - V_f(t)\,, \\ V_h(t) = & (I + \Phi^\top)\Xi^\top\sigma_f(V_f)\,, \end{cases} \tag{5}$$

$$\mathcal{T}_f(V_f(t+1) - V_f(t)) = \Xi\,(I + \Phi^\top)\Xi^\top\sigma_f(V_f) - V_f(t) \tag{6}$$

Final discrete upate equation

$$V_f(t+1) = \Xi\,(I + \Phi^\top)\Xi^\top\sigma_f(V_f) \tag{7}$$

Restrict the norm of matrix $||\Xi\,(I + \Phi^\top)\Xi^\top|| \leq 1$.
This allows us to consider the transformation $V'_f = \sigma_f(V_f)$, so for invertible σ_f,

$$\sigma_f^{-1}(V'_f(t+1)) = \Xi\,(I + \Phi^\top)\Xi^\top V'_f \tag{8}$$

$$\sigma_f^{-1}(V'_f(t+1)) = \Xi\,(I + \Phi^\top)\Xi^\top V'_f \tag{9}$$

$$V'_f(t+1) = \sigma_f(\Xi\,(I + \Phi^\top)\Xi^\top V'_f) \tag{10}$$

this is a general update equation for an RNN without bias.

3 Topological Conjugacy with RNNs

Proof that dynamical systems governed by Equations 7 and 10 are topological conjugates.

Consider $f(x) = \Xi\,(I+\Phi^\top)\Xi^\top\sigma_f(x)$ for Equation 7 and $g(x) = \sigma_f(\Xi\,(I+\Phi^\top)\Xi^\top x)$ for Equation 10. Consider a homeomorphism $h(y) = \sigma_f(y)$ on g. Then,

$$\begin{aligned} (h^{-1} \circ g \circ h)(x) &= \sigma_f^{-1}(\sigma_f(\Xi\,(I + \Phi^\top)\Xi^\top\sigma_f(x))) \\ &= \Xi\,(I + \Phi^\top)\Xi^\top\sigma_f(x) \\ &= f(x) \end{aligned} \tag{11}$$

So, for the homeomorphism h on g, we get that $h^{-1} \circ g \circ h = f$ proving that f and g are topological conjugates. Therefore all dynamical properties of f and g are shared.

4 Deriving Energy

From the proof of topological conjugacy, we can show energy properties on only f. Careful about the time in discretization. Keep V_d as is only to convert in the end

$$E(t) = \left[V_f(t)^\top \sigma_f(V_f(t)) - L_f(t)\right] + \frac{1}{2}V_h(t)^\top V_h(t) - \left[\sigma_f(V_f(t))^\top \Xi V_h(t)\right] - \left[V_d(t)^\top \Xi \Phi V_h(t)\right] \tag{12}$$

$$\begin{aligned} E(t+1) = &\left[V_f(t+1)^\top \sigma_f(V_f(t+1)) - L_f(t+1)\right] + \frac{1}{2}V_h(t+1)^\top V_h(t+1) \\ &- \left[\sigma_f(V_f(t+1))^\top \Xi V_h(t+1)\right] - \left[V_d(t+1)^\top \Xi \Phi V_h(t+1)\right] \end{aligned} \tag{13}$$

Working with $V_h(t) = \Xi^\top \sigma_f(V_f) + \Phi^\top \Xi^\top V_d(t)$. The next subsections will try to compute $E(t+1) - E(t)$. Instead of directly discretizing E, lets try to use the continuous formulation first using $V_h = \Xi^\top \sigma_f(V_f) + \Phi^\top \Xi^\top V_d$

$$E(t) = \left[V_f^\top \sigma_f(V_f) - L_f\right] + \frac{1}{2}V_h^\top V_h - \left[\sigma_f(V_f)^\top \Xi V_h\right] - \left[V_d^\top \Xi \Phi V_h\right] \tag{14}$$

$$E_2(t) = \frac{1}{2}\left[\left(\sigma_f(V_f)^\top \Xi + V_d^\top \Xi \Phi\right)\left(\Xi^\top \sigma_f(V_f) + \Phi^\top \Xi^\top V_d\right)\right] \tag{15}$$

$$E_2(t) = \frac{1}{2}\left[\sigma_f(V_f)^\top \Xi \Xi^\top \sigma_f(V_f) + \sigma_f(V_f)^\top \Xi \Phi^\top \Xi^\top V_d + V_d^\top \Xi \Phi \Xi^\top \sigma_f(V_f) + V_d^\top \Xi \Phi \Phi^\top \Xi^\top V_d\right] \tag{16}$$

$$E_2(t) = \frac{1}{2}\sigma_f(V_f)^\top \Xi \Xi^\top \sigma_f(V_f) + \sigma_f(V_f)^\top \Xi \Phi^\top \Xi^\top V_d + \frac{1}{2}V_d^\top \Xi \Phi \Phi^\top \Xi^\top V_d \tag{17}$$

$$E_3(t) = -\sigma_f(V_f)^\top \Xi V_h \tag{18}$$

$$E_3(t) = -\sigma_f(V_f)^\top \Xi \left(\Xi^\top \sigma_f(V_f) + \Phi^\top \Xi^\top V_d\right) \tag{19}$$

$$E_3(t) = -\left[\sigma_f(V_f)^\top \Xi \Xi^\top \sigma_f(V_f) + \sigma_f(V_f)^\top \Xi \Phi^\top \Xi^\top V_d\right] \tag{20}$$

$$E_4(t) = -\left[V_d^\top \Xi\Phi\left(\Xi^\top \sigma_f(V_f) + \Phi^\top \Xi^\top V_d\right)\right] \tag{21}$$

$$E_4(t) = -\left[V_d^\top \Xi\Phi\Xi^\top \sigma_f(V_f) + V_d^\top \Xi\Phi\Phi^\top \Xi^\top V_d\right] \tag{22}$$

$$E_4(t) = -\left[\sigma_f(V_f)\,\Xi\Phi^\top\Xi^\top V_d + V_d^\top \Xi\Phi\Phi^\top \Xi^\top V_d\right] \tag{23}$$

$$E = E_1 + E_2 + E_3 + E_4 \tag{24}$$

Simplifying the colored terms

$$E = V_f^\top \sigma_f(V_f) - L_f - \frac{1}{2}\sigma_f(V_f)^\top \,\Xi\,\Xi^\top \sigma_f(V_f) - \sigma_f(V_f)^\top \,\Xi\Phi^\top\Xi^\top V_d - \frac{1}{2}V_d^\top \Xi\Phi\Phi^\top \Xi^\top V_d \tag{25}$$

$$\boxed{E = V_f^\top \sigma_f(V_f) - \left[L_f + \frac{1}{2}\sigma_f(V_f)^\top \,\Xi\,\Xi^\top \sigma_f(V_f) + \sigma_f(V_f)^\top \,\Xi\Phi^\top\Xi^\top V_d + \frac{1}{2}V_d^\top \Xi\Phi\Phi^\top \Xi^\top V_d\right]} \tag{26}$$

Discrete Energy Update - Not complete (non linear terms issue)

4.1 Trial 1: Adiabatic V_d

Case of direct substitution of time.

$$\begin{cases} V_f(t+1) = & \Xi\,\Xi^\top \sigma_f(V_f(t)) + \Xi\,\Phi^\top \,\Xi^\top V_d(t), \\ V_d(t) = & \sigma_f(V_f(t))\,. \end{cases} \tag{27}$$

Consider the discrete energy equation at time t

$$\begin{aligned} E(t) = V_f(t)^\top \sigma_f(V_f(t)) - L_f(t) + \frac{1}{2}\sigma_f(V_f(t))^\top \,\Xi\,\Xi^\top \sigma_f(V_f(t)) \\ - \sigma_f(V_f(t))\,\Xi\Phi^\top\Xi^\top V_d(t) - \frac{1}{2}V_d(t)^\top \Xi\Phi\Phi^\top \Xi^\top V_d(t) \end{aligned} \tag{28}$$

and

$$\begin{aligned} E(t+1) = V_f(t+1)^\top \sigma_f(V_f(t+1)) - L_f(t+1) + \frac{1}{2}\sigma_f(V_f(t+1))^\top \,\Xi\,\Xi^\top \sigma_f(V_f(t+1)) \\ - \sigma_f(V_f(t+1))\,\Xi\Phi^\top\Xi^\top V_d(t) - \frac{1}{2}V_d(t)^\top \Xi\Phi\Phi^\top \Xi^\top V_d(t) \end{aligned} \tag{29}$$

4.1.1 E(t+1) in terms of V_f(t)

$$E_1(t+1) = \left(\sigma_f(V_f(t))^\top\, \Xi\, \Xi^\top + V_d^\top(t)\, \Xi\, \Phi\, \Xi^\top\right) \sigma_f(V_f(t+1)) \tag{30}$$

$$E_3(t+1) = \left(\sigma_f(V_f(t))^\top\, \Xi\, \Xi^\top + V_d^\top(t)\, \Xi\, \Phi\, \Xi^\top\right) \sigma_f(V_f(t+1)) \tag{31}$$

How to deal with these nonlinear terms?

5 Tanh Case - only Elmann RNN w/o bias - nonlinear terms issue

$$\begin{aligned} E(t) = V_f(t)^\top \tanh\big(V_f(t)\big)^\top - \sum \log|\cosh\big(V_f(t)\big)| - \frac{1}{2} \tanh\big(V_f(t)\big)^\top\, \Xi\, \Xi^\top\, \sigma_f(V_f(t)) \\ - \tanh\big(V_f(t)\big)\, \Xi \Phi^\top \Xi^\top V_d(t) - \frac{1}{2} V_d(t)^\top \Xi \Phi \Phi^\top\, \Xi^\top V_d(t) \end{aligned} \tag{32}$$

$$\begin{aligned} E(t+1) = V_f(t+1)^\top \tanh\big(V_f(t+1)\big)^\top - \sum \log|\cosh\big(V_f(t+1)\big)| - \frac{1}{2} \tanh\big(V_f(t+1)\big)^\top\, \Xi\, \Xi^\top\, \sigma_f(V_f(t+1)) \\ - \tanh\big(V_f(t+1)\big)\, \Xi \Phi^\top \Xi^\top V_d(t+1) - \frac{1}{2} V_d(t+1)^\top \Xi \Phi \Phi^\top\, \Xi^\top V_d(t+1) \end{aligned} \tag{33}$$

$$\Delta E_1 = \tanh\big(V_f(t)\big)^\top\, \Xi\, \Xi^\top \tanh\big(V_f(t+1)\big) + V_d(t)^\top \Xi\, \Phi\, \Xi^\top \tanh\big(V_f(t+1)\big) - V_f(t)^\top \tanh\big(V_f(t)\big) \tag{34}$$

$$\Delta E_2 = \sum \log \frac{|\cosh\big(V_f(t+1)\big)|}{|\cosh\big(V_f(t)\big)|} \tag{35}$$

$$\Delta E_3 = -\frac{1}{2}\left[\tanh\big(V_f(t+1)\big)^\top \Xi \Xi^\top \tanh\big(V_f(t+1)\big) - \tanh\big(V_f(t)\big)^\top \Xi \Xi^\top \tanh\big(V_f(t)\big)\right] \tag{36}$$

$$\Delta E_4 = -\left[\tanh\big(V_f(t+1)\big)^\top\, \Xi \Phi^\top \Xi^\top V_d(t+1) - \tanh\big(V_f(t)\big)^\top\, \Xi \Phi^\top \Xi^\top V_d(t)\right] \tag{37}$$

Suppose adiabatic, then $V_d(t+1) = V_d(t)$

$$\Delta E_4 = -\left[\tanh\left(V_f(t+1)\right) - \tanh\left(V_f(t)\right)\right]^\top \Xi \Phi^\top \Xi^\top V_d(t) \tag{38}$$

Trying simplification using identities

$$\Delta E_4 = -\left[\tanh\left(V_f(t+1) - V_f(t)\right)(1 - \tanh\left(V_f(t+1)\right) \tanh\left(V_f(t)\right))\right] \Xi \Phi^\top \Xi^\top V_d(t) \tag{39}$$

Suppose $V_d(t+1) = V_d(t) = \sigma_f(V_f(t))$

$$\Delta E_4 = -\left[\tanh\left(V_f(t+1)\right)^\top \Xi \Phi^\top \Xi^\top \tanh\left(V_f(t)\right) - \tanh\left(V_f(t)\right)^\top \Xi \Phi^\top \Xi^\top \tanh\left(V_f(t)\right)\right] \tag{40}$$

$$\Delta E = \Delta E_1 + \Delta E_2 + \Delta E_3 + \Delta E_4 + \Delta E_5 \tag{41}$$

$$= \tanh\left(V_f(t)\right)^\top \Xi \Xi^\top \tanh\left(V_f(t+1)\right) - V_f(t)^\top \tanh\left(V_f(t)\right) + \sum \log \frac{|\cosh\left(V_f(t+1)\right)|}{|\cosh\left(V_f(t)\right)|} \tag{42}$$
$$-\frac{1}{2} \tanh\left(V_f(t+1)\right)^\top \Xi \Xi^\top \tanh\left(V_f(t+1)\right) + \frac{1}{2} \tanh\left(V_f(t)\right)^\top \Xi \Xi^\top \tanh\left(V_f(t)\right) + \tanh\left(V_f(t)\right)^\top \Xi \Phi^\top \Xi^\top \tanh\left(V_f(t)\right)$$

$$\begin{aligned} &= \frac{1}{2}\left(\tanh\left(V_f(t)\right)^\top - \tanh\left(V_f(t+1)^\top\right)\right) \Xi \Xi^\top \tanh\left(V_f(t+1)\right) \\ &\quad + \frac{1}{2} \tanh\left(V_f(t)\right)^\top \Xi \Xi^\top \left(\tanh\left(V_f(t)\right) + \tanh\left(V_f(t+1)^\top\right)\right) - V_f(t)^\top \tanh\left(V_f(t)\right) \\ &\quad + \sum \log \frac{|\cosh\left(V_f(t+1)\right)|}{|\cosh\left(V_f(t)\right)|} + \tanh\left(V_f(t)\right)^\top \Xi \Phi^\top \Xi^\top \tanh\left(V_f(t)\right) \end{aligned} \tag{43}$$

6 Proof using Concave Convex Procedure

$$V_f(t+1) = \Xi \Xi^\top \sigma_f(V_f(t)) + \Xi \Phi^\top \Xi^\top V_d(t) \tag{44}$$

$$E = V_f^\top \sigma_f(V_f) - L_f + \frac{1}{2}\sigma_f(V_f)^\top \Xi \Xi^\top \sigma_f(V_f) + \sigma_f(V_f)^\top \Xi \Phi^\top \Xi^\top V_d + \frac{1}{2} V_d^\top \Xi \Phi \Phi^\top \Xi^\top V_d \tag{45}$$

Let $x = V_f andy = V_d, \sigma = \sigma_f = \frac{\partial L}{\partial x}, L = L_f$

$$x(t+1) = \Xi \Xi^\top \sigma_f(x(t)) + \Xi \Phi^\top \Xi^\top y(t) \tag{46}$$

$$E(x) = x^\top \sigma(x) - L + \frac{1}{2}\sigma(x)^\top \Xi \Xi^\top \sigma(x) + \sigma(x)^\top \Xi \Phi^\top \Xi^\top y + \frac{1}{2} y^\top \Xi \Phi \Phi^\top \Xi^\top y \tag{47}$$

Assume that E is bounded think of proving this

Function convexity

Energy Term 1, 2

$$\mathcal{J}(x^\top \sigma(x) - L) = \sigma(x)^\top + x^\top.\mathcal{J}(\sigma(x)) - \mathcal{J}(L) \tag{48}$$

$$\mathcal{J}(x^\top \sigma(x) - L) = x^\top.\mathcal{J}(\sigma(x)) \tag{49}$$

$$\mathcal{H}(x^\top \sigma(x) - L) = \mathcal{J}(\mathcal{J}(\sigma(x))^\top.x) \tag{50}$$

Let $\mathcal{J}(\sigma(x)) = (J_1\, J_2\, J_3 ...)$. note that $x^\top.\mathcal{H}(\sigma)$ is a tensor product

$$\mathcal{H}(x^\top \sigma(x) - L) = \begin{pmatrix} J_1^\top + x^\top.\mathcal{H}(\sigma) \\ J_2^\top + x^\top.\mathcal{H}(\sigma) \\ \vdots \end{pmatrix} = \mathcal{J}(\sigma(x))^\top + x^\top.\mathcal{H}(\sigma) \tag{51}$$

$$\mathcal{H}(x^\top \sigma(x) - L) = \mathcal{H}(L)^\top + x^\top.\mathcal{H}(\sigma) \tag{52}$$

Energy Term 3

$$\begin{aligned} \frac{1}{2}\mathcal{J}(\sigma(x)^\top\, \Xi\Xi^\top\, \sigma(x)) &= \frac{1}{2}2\,\Xi\Xi^\top.\sigma(x).\mathcal{J}(\sigma) \\ &= \sigma(x)^\top\Xi\Xi^\top.\mathcal{J}(\sigma) \end{aligned} \tag{53}$$

$$\frac{1}{2}\mathcal{H}(\sigma(x)^\top\, \Xi\Xi^\top\, \sigma(x)) = \mathcal{J}((\sigma(x)^\top\Xi\Xi^\top.\mathcal{J}(\sigma))^\top) \tag{54}$$

$$\frac{1}{2}\mathcal{H}(\sigma(x)^\top\, \Xi\Xi^\top\, \sigma(x)) = \mathcal{J}(\mathcal{J}(\sigma)^\top\Xi\Xi^\top\, \sigma(x)) \tag{55}$$

$$\frac{1}{2}\mathcal{H}(\sigma(x)^\top\, \Xi\Xi^\top\, \sigma(x)) = \mathcal{J}(\sigma)^\top\Xi\Xi^\top\, \mathcal{J}(\sigma) + \mathcal{J}(\mathcal{J}(\sigma)^\top\Xi\Xi^\top)\,\sigma(x) \tag{56}$$

$$\frac{1}{2}\mathcal{H}(\sigma(x)^\top\, \Xi\Xi^\top\, \sigma(x)) = \mathcal{J}(\sigma)^\top\Xi\Xi^\top\, \mathcal{J}(\sigma) + \mathcal{H}(\sigma)^\top.\Xi\Xi^\top\, \sigma(x) \tag{57}$$

Energy Term 4

$$\mathcal{J}(\sigma(x)^\top\, \Xi\, \Phi^\top \Xi^\top y) = y^\top\, \Xi\, \Phi\, \Xi^\top\, \mathcal{H}(\sigma) \tag{58}$$