Episodic Memory Theory of RNNs

Theoretical Concepts and Derivations

Arjun Karuvally, Peter Delmastro

1 Mathematical Preliminaries

Vector: An abstract mathematical object that is invariant to basis transformations. A vector v is represented using the Dirac notation: $|v\rangle$

Vector Space: An abstract vector space of n dimensions over the field $\mathbb{R}$ is the set of vectors obtained by the linear combination

$$|v\rangle = \sum_{i=1}^{n} v^i |e_i\rangle \tag{1}$$

The set of vectors $|e_i\rangle$ is called the **basis** of the space and the elements v^i are the **vector components** in that basis. From this definition, it can be easily seen that $|v\rangle$ is invariant to the basis but the vector components are basis dependent.

Vector Dual Space: Given a Vector space V, V^* is the set of associated linear forms/covectors $\phi : V \longrightarrow \mathbb{R}$ represented in Dirac notation as $\langle v|$. For the set of basis vectors $|e_i\rangle \in V$, the **basis dual** is the set of covectors $\left\langle e^j \right| \in V^*$ such that $\left\langle e^j \middle| e_i \right\rangle = \delta_{ij}$ where δ is the Kronecker delta.

Memory: persistent states in the evolution of a dynamical system. Memory can be **stable memory** if the system after reaching the state stays in the state for infinite time. Stable memories are fixed points of the system. Memory acn be **meta-stable memory** if the state stays for a non-trivial amount of time. Non-trivial meaning that the state is not transient.

Energy Function: of a dynamical system is any function $E(X)$ that if $\frac{\mathrm{d}X}{\mathrm{d}t} = f(X) \implies \frac{\mathrm{d}E(X)}{\mathrm{d}t} \leq 0$ in continuous dynamical system or $X(t+1) - X(t) \implies E(X(t+1)) - E(X(t)) \leq 0$ in discrete dynamical system.

2 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*.

$$\begin{cases} \mathcal{T}_f \dfrac{\mathrm{d}V_f}{\mathrm{d}t} = & \sqrt{\alpha_s}\,\Xi\,\sigma_h(V_h) - V_f\,, \\ \mathcal{T}_h \dfrac{\mathrm{d}V_h}{\mathrm{d}t} = & \sqrt{\alpha_s}\,\Xi^\top\,\sigma_f(V_f) + \alpha_c \Phi^\top \Xi^\top V_d - V_h\,, \\ \mathcal{T}_d \dfrac{\mathrm{d}V_d}{\mathrm{d}t} = & \sigma_f(V_f) - V_d\,. \end{cases} \tag{2}$$

The system has an energy function given by

$$E = \left[V_f^\top\,\sigma_f(V_f) - L_f\right] + \left[V_h^\top\,\sigma_h(V_h) - L_h\right] - \left[\sqrt{\alpha_s}\,\sigma_f(V_f)^\top\,\Xi\,\sigma_h(V_h)\right] - \alpha_c\left[V_d^\top\,\Xi\,\Phi\sigma_h(V_h)\right] \tag{3}$$

Conditions:

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

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

stored memories: These are the column vectors of the matrix Ξ.
spurious memories: These are memories in the system that does not belong to the set of the intended stored memories.
memory interactions: The matrix Φ denotes the intermemory interaction since it encodes the temporal relationships between the stored memories.

3 Deriving governing equations

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

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

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

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

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

Final discrete upate equation

$$V_f(t+1) = \Xi\,(I + \Phi^\top)\Xi^\top \sigma_f(V_f) \tag{8}$$

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,

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

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

$$V_f'(t+1) = \sigma_f(\Xi\,(I + \Phi^\top)\Xi^\top V_f') \tag{11}$$

this is a general update equation for an RNN without bias. The physical interpretation of this equation is that the columns of Ξ stores the individual *memories* of the system and the linear operator $(I + \Phi)$ is the temporal interaction between the stored *memories*. In the memory modeling literature, it is typical to consider memories as a fixed collection instead of a variable collection that shares a common interaction behavior. We will show how in the next sections how the dynamics as a result of fixed collection can be used to store variable information.

4 Topological Conjugacy with RNNs

Proof that dynamical systems governed by Equations 8 and 11 are topological conjugates.

Consider $f(x) = \Xi\,(I+\Phi^\top)\Xi^\top \sigma_f(x)$ for Equation 8 and $g(x) = \sigma_f(\Xi\,(I+\Phi^\top)\Xi^\top x)$ for Equation 11. Consider a homeomorphism $h(y) = \sigma_f(y)$ on g. Then,

$$\begin{aligned}(h^{-1} \circ g \circ h)(x) &= \sigma_f^{-1}(\sigma_f(\Xi\,(I + \Phi^\top)\Xi^\top \sigma_f(x))) \\ &= \Xi\,(I + \Phi^\top)\Xi^\top \sigma_f(x) \\ &= f(x)\end{aligned} \tag{12}$$

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.

5 Tensor Formalism

Since I and Φ are linear operators, they can be combined into a single linear operator. Lets call this combined linaer operator Φ. Further theory will use the new Φ definition. Let $|e_i\rangle$ be the i^{th} standard basis *vector*. $\left\langle \epsilon^j \right|$ be the *covector* of the j^{th} standard basis vector such that $\left\langle \epsilon_j \middle| e_i \right\rangle = \delta_{ij}$ where δ is the Kronecker delta. Consider the matrix formulation of the discrete RNN system analogous to pseudoinverse associative memories.

$$h(t+1) = \Xi\,\Phi\,\Xi^{\dagger} g(h(t)) \tag{13}$$

The same system represented in tensor format is given by (using Einstein Summation Convention) - greek indices iterate over memory space indices $\{1, 2, \ldots, N_h\}$, alpha numeric indices iterate over feature space indices $\{1, 2, \ldots, N_f\}$

$$\begin{aligned} |h(t+1)\rangle &= \left(\xi^i_\mu \, \Phi^\mu_\nu \, (\xi^\dagger)^\nu_j \;\; |e_i\rangle \left\langle \epsilon^j \right| \right) |h(t)\rangle \\ &= W_{hh} \vec{h}(t) \end{aligned} \tag{14}$$

In this view, W_{hh} can be thought of a linear operator acting on the current state vector $\vec{h}(t)$. This abstract linear algebra view of the update equation will be useful when we transform the basis of the system. In the tensor formalism, $|h\rangle$ is defined as.

$$|h(t)\rangle = \left\langle \epsilon^j \middle| h(t) \right\rangle |e_i\rangle \tag{15}$$

since $|e_i\rangle$ is the standard basis vectors, $\left\langle \epsilon^j \middle| h(t) \right\rangle$ are the *vector components* of $|h(t)\rangle$ we obtain from simulations. Now, consider a new set of basis vectors given by $\left\langle \psi_\mu \right| = \xi^i_\mu \, |e_i\rangle$. It can be seen that the dual of the basis vectors is $\langle \psi^\nu | = (\xi^\dagger)^\nu_j \left\langle \epsilon^j \right|$. In this new basis, the update equations transform to

$$|h(t+1)\rangle = \left(\Phi^\mu_\nu \, \left| \psi_\mu \right\rangle \, \langle \psi^\nu | \right) |h(t)\rangle \tag{16}$$

In this new basis $|\psi\rangle$, the update equations can be easily interpreted as applying a single linear operation Φ on $h(t)$ represented in the the memory basis ψ. Since the new bases simplifies the interpretation of the dynamics by projecting onto the space spanned by the stored *memories*, we name the set of $\left|\psi_\mu\right\rangle$ the **memory bases**. Using these conventions, we design our theoretical model.

6 Problem Setup - Variable Binding

We formalize the variable binding problem in the following manner. Let the RNN be defined by a task containing two phases - the input phase and the output phase. At each timestep of the the input phase, external information is provided to the network. In the output phase, at each time step, the network needs to utilize this external information to synthesize novel outputs. Formally, let the input phase consist of s timesteps where at each time step t, a vector of d dimensions $|u(t)\rangle = u^i(t) \, |e_i\rangle$ is provided as input to the model. We call the vector components $u^i(s)$, the external information that needs to be potentially *stored* in the RNN for future computation. After the input phase is complete, the zero vector is passed as input to the model. The RNN thus evolves autonomously (without any external input) during the output phase. During training, the RNN output is compared to the ground truth sequence to obtain a loss function to be used for backpropagation. Generally, the task for the RNN is to estimate a dynamical system of x given by the following equation

$$|x(t+1)\rangle = f(|x(t)\rangle, |x(t-1)\rangle, \ldots |x(1)\rangle, z) \tag{17}$$

where z can be a latent variable not directly observable in the system but may be infered from the history. The RNN is trained to approximate this dynamical system. The Elman RNN has an update equation (in matrix notation) given by

$$\begin{cases} h(t+1) = \tanh(W_{hh}h(t) + W_{uh}u(t)) \\ o(t) = W_r \; h(t+1) \end{cases} \tag{18}$$

where W_{hh}, W_{uh}, W_r are linear operators, $h(t)$ is the hidden state, $u(t)$ is the input, and $o(t)$ is the output. To simplify the theory we assume that W_{hh} has sufficient capacity to represent all the variables required to estimate the dynamical system. We further assume that $|h(0)\rangle = 0\,|e_i\rangle$ - the zero vector.

7 Theoretical model of variable binding

Instead of directly analyzing the non-linear system, we define the variable binding mechanisms on a linear system defined by

$$|h(t+1)\rangle = \Phi^{\mu}_{\nu} \left|\psi_{\mu}\right\rangle \langle\psi^{\nu}|h(t)\rangle \tag{19}$$

Consider that $|h(t)\rangle$ is defined in terms of the subspaces that are each spanned by subsets of vectors in the collection $\left|\psi_{\mu}\right\rangle$. The components of the i^{th} subspace can be extracted from $|h(t)\rangle$ in the standard basis by the linear operator Ψ^*_i variable defined as $\Psi^*_i = \sum_{\mu=(i-1)\kappa+1}^{i\kappa} \left|e_{\mu}\right\rangle \langle\psi^{\mu}|$, κ is the number of dimensions in each of the variable subspaces. The linear operator Φ is defined in the theoretical model as:

$$\Phi = \sum_{\mu=1}^{(N-1)\kappa} \left|\psi_{\mu}\right\rangle \left\langle\psi^{\mu+\kappa}\right| + \sum_{\mu=(N-1)\kappa}^{N\kappa} \Phi^{\mu}_{\nu} \left|\psi_{\mu}\right\rangle \langle\psi^{\nu}| \tag{20}$$

Here the N^{th} subspace activity is a linear composition operation acting on all the variable subspaces.

7.1 Writing Variables

We will describe how external information can be written to the hidden state of the RNN within the framework. Typically, RNNs have W_{uh} which facilitates the interaction of external information with the RNN. In our framework, W_{uh} has the following equation when the input $|u\rangle = u^i\,|e_i\rangle$.

$$\begin{aligned} W_{uh} &= \Psi_N \\ &= \left|\psi_{(N-1)\kappa+j}\right\rangle \left\langle e^j\right| \end{aligned} \tag{21}$$

It can be easily seen that the loading operation "inserts" the input variable into the N^{th} subspace. Due to the circulant nature of the Φ operator, this external input will get moved around to the sequentially connected subspaces over time.

7.2 Reading Variables

We will describe how inputs at each timestep is read from the RNN. RNNs have a linear operator W_r which facilitates the reading of information from $|h(t)\rangle$ at each time step. In our framework, W_r has the following equation when the output $|o\rangle = o^i |e_i\rangle$.

$$\begin{aligned} W_r &= \Psi_N^* \\ &= \sum_{\mu=(N-1)\kappa+1}^{N\kappa} \left|e_{\mu-(N-1)\kappa}\right\rangle \langle\psi^\mu| \end{aligned} \tag{22}$$

It can be easily seen that the reading operation reads the contents of the r^{th} subspace.

8 Sample Tasks

We will demonstrate the behavior of the RNN according to the theory on two tasks that needs some variable binding mechanisms to learn and generalize.

9 Repeat Copy Task

Repeat Copy is a task typically used to evaluate the memory storage characteristics of RNNs since the task has a deterministic evolution represented by a simple algorithm that stores all input vectors in memory for later retrieval. Although elementary, repeat copy provides a simple framework to imagine the variable binding mechanisms we theorized in action. For the repeat copy task, the linear operators of the RNN has the following equations.

$$\begin{cases} \Phi = \Sigma_{\mu=1}^{(s-1)\kappa} \left|\psi_\mu\right\rangle \langle\psi^{\mu+\kappa}| + \Sigma_{\mu=(s-1)\kappa+1}^{s\kappa} \left|\psi_\mu\right\rangle \left\langle\psi^{\mu-(s-1)\kappa}\right| \\ W_{uh} = \Psi_s \\ W_r = \Psi_s^* \end{cases} \tag{23}$$

This ϕ can be imagined as copying the contents of the subspaces in a cyclic fashion. That is, the content of the i^{th} subspace goes to $(i-1)^{\text{th}}$ subspace with the first subspace being copied to the N^{th} subspace. The dynamical evolution of the RNN is represented at the time step 1 as,

$$|h(1)\rangle = \left|\psi_{(s-1)\kappa+j}\right\rangle \left\langle e^j\right| u^i(1) \left|e_i\right\rangle \tag{24}$$

$$|h(1)\rangle = u^i(1) \left|\psi_{(s-1)\kappa+j}\right\rangle \left\langle e^j \middle| e_i\right\rangle \tag{25}$$

$$|h(1)\rangle = u^i(1) \left|\psi_{(s-1)\kappa+j}\right\rangle \delta_{ij} \tag{26}$$

Kronecker delta index cancellation

$$|h(1)\rangle = u^i(1) \left|\psi_{(s-1)\kappa+i}\right\rangle \tag{27}$$

At time step 2,

$$|h(2)\rangle = u^i(1)\,\Phi\, \left|\psi_{(s-1)\kappa+i}\right\rangle + u^i(2)\, \left|\psi_{(s-1)\kappa+i}\right\rangle \tag{28}$$

Expanding Φ

$$|h(2)\rangle = u^i(1) \left(\sum_{\mu=1}^{(s-1)\kappa} \left|\psi_\mu\right\rangle \left\langle \psi^{\mu+\kappa}\right| + \sum_{\mu=(s-1)\kappa+1}^{s\kappa} \left|\psi_\mu\right\rangle \left\langle \psi^{\mu-(s-1)\kappa}\right| \right) \left|\psi_{(s-1)\kappa+i}\right\rangle + u^i(2) \left|\psi_{(s-1)\kappa+i}\right\rangle \tag{29}$$

$$|h(2)\rangle = u^i(1) \left|\psi_{(s-2)\kappa+i}\right\rangle + u^i(2) \left|\psi_{(s-1)\kappa+i}\right\rangle \tag{30}$$

At the final step of the input phase when $t = s$, $|h(s)\rangle$ is defined as:

$$|h(s)\rangle = \sum_{\mu=1}^{s} u^i(\mu) \left|\psi_{(\mu-1)\kappa+i}\right\rangle \tag{31}$$

For t timesteps after s, the general equation for $|h(s+t)\rangle$ is:

$$|h(s+t)\rangle = \sum_{\mu=1}^{s} u^i(\mu) \left|\psi_{[((\mu-t-1 \mod s)+1)\kappa+i]}\right\rangle \tag{32}$$

From this equation for the hidden state vector, it can be easily seen that the μ^{th} variable is stored in the $[(\mu - t - 1 \mod s) + 1]^{\text{th}}$ subspace at time step t. The readout weights $W_r = \Psi_s^*$ reads out the contents of the s^{th} subspace.

10 Compose Copy

Repeat copy require only the storage and retrieval of external information without any novel synthesis. The compose copy synthesizes novel output from the given inputs. The input to the compose copy task is defined with s vectors, each of of dimension s, $\{x(1), x(2), \ldots, x(s)\}$. For any time $t > s$, $x(t)$ is defined as

$$|x(t)\rangle = \sum_{i=1}^{s} \left\langle e^i \middle| x(t-s-2+i)\right\rangle |e_i\rangle \tag{33}$$

Analogous to compose copy, RNN linear opeartors can be written as

$$\begin{cases} \Phi = \sum_{\mu=1}^{(s-1)\kappa} \left|\psi_\mu\right\rangle \langle\psi^{\mu+\kappa}| + \sum_{\mu=(s-1)\kappa+1}^{s\kappa} \left|\psi_\mu\right\rangle \left\langle\psi^{(\mu-(s-1)\kappa-1)\kappa+\mu-(s-1)\kappa}\right| \\ W_{uh} = \Psi_s \\ W_r = \Psi_s^* \end{cases} \tag{34}$$

The linear operator Φ has a similar structure as repeat copy but now instead of just copying the 1st subspace to the s^{th} subspace in timestep $s + 1$, the contents are composed to obtain a new vector which is read out by the readout operator.

11 Non-linear RNN

The linear RNNs we discussed are powerful in terms of the content of variables that can be stored and reliably retrieved. The variable contents, u^i, can be any real number and this information can be reliably retrieved in the end using the appropriate readout weights. However, learning such a system is difficult using gradient descent procedures. To see this, setting the components of Φ to anything other than unity might result in dynamics that is eventually converging or diverging resulting in a loss of information in these variables. Additionally, linear systems are not used in the practical design of RNNs. The main difference is now the presence of the nonlinearity. In this case, our theory can still be used. To illustrate this, consider a general RNN evolving according to $h(t+1) = g(W_{hh}h(t) + b)$ where b is a bias term. Suppose $h(t) = h^*$ is a fixed point of the system. We can then linearize the system around the fixed point to obtain the linearized dynamics in a small region around the fixed point.

$$h(t+1) - h^* = \mathcal{J}(g)|_{h^*}\, W_{hh}\, (h(t+1) - h^*) + O((h(t+1) - h^*)^2) \tag{35}$$

where $\mathcal{J}$ is the jacobian of the activation function g. If the RNN had an additional input, this can also be incorporated into the linearized system by treating the external input as a control variable

$$h(t+1) - h^* = \mathcal{J}(g)|_{h^*}\, W_{hh}\, (h(t) - h^*) + \mathcal{J}(g)|_{h^*}\, W_{uh} u(t) \tag{36}$$

Substituting $h(t) - h^* = h'(t)$

$$h'(t+1) = \mathcal{J}(g)|_{h^*}\, W_{hh}\, h'(t) + \mathcal{J}(g)|_{h^*}\, W_{uh} u(t) \tag{37}$$

which is exactly the linear system which we studied where instead of $W_{hh} = \Xi\Phi\Xi^\dagger$, we have $J(g)|_{h^*}W_h h = \Xi\Phi\Xi^\dagger$. With this result, we will analyse Elman RNN models that have the general update equations $h(t+1) = \tanh(W_{hh}h(t) + W_{uh}u(t) + b)$.