General Sequential Episodic Memory Model

Arjun Karuvally [1] **Terrence J. Sejnowski** [2] **Hava T. Siegelmann** [1]

Abstract

The state-of-the-art memory model is the General Associative Memory Model, a generalization of the classical Hopfield network. Like its ancestor, the general associative memory has a well-defined state-dependant energy surface, and its memories correlate with its fixed points. This is unlike human memories, which are commonly sequential rather than separated fixed points. In this paper, we introduce a class of General Sequential Episodic Memory Models (GSEMM) that, in the adiabatic limit, exhibit a dynamic energy surface, leading to a series of meta-stable states capable of encoding memory sequences. A multiple-timescale architecture enables the dynamic nature of the energy surface with newly introduced asymmetric synapses and signal propagation delays. We demonstrate its dense capacity under polynomial activation functions. GSEMM combines separate memories, short and long sequential episodic memories, under a unified theoretical framework, demonstrating how energy-based memory modeling can provide robust and scalable memory systems in static and dynamic memory cases.

Introduction

Episodic memory refers to the conscious recollection of facts or subjective past experiences and forms an essential component of long-term memory (Tulving, 2002; Duff et al., 2019; Renoult et al., 2019). The recollection process may have both singleton and sequence characteristics. Singleton retrieval is the associative recall of a single memory from a retrieval cue. This memory could be the description of a particular object of interest or important dates of events. Sequence retrieval leads to a recollection process that is not just a single memory but a trajectory of sequentially connected memories. Memories organized into these trajectories are called *episodes*. Memories may come together in episodes allowing us to link and retrieve sometimes distinct and representationally unrelated memories. The Sequence Episodic Memory (SEM) problem in Recurrent Neural Networks (RNNs) pertains to creating and manipulating these memories and their sequential relationships by encoding relevant information in some form in the synapses.

The energy paradigm plays a major role in singleton episodic memory modeling. The energy paradigm for memory was introduced by Hopfield (Hopfield, 1982; Amari, 2004), who defined energy as a quadratic function of the neural activity in symmetrically connected networks with binary neurons. A single memory is stored as a local minimum of the energy surface. The network states update such that it converges to one of the local minima and retrieves a stable activity state representing a single memory. The Hopfield network model has subsequently been generalized along two directions.

The first direction focuses on memory capacity. Capacity relates to the number of neurons required in the ensemble to store and retrieve a certain number of memories without corruption. The capacity of the original Hopfield model was 14% of the number of neurons, a small fraction of the number of neurons in the population(McEliece et al., 1987; Folli et al., 2016; Amit et al., 1985). A significant breakthrough in capacity came with the introduction of Dense Associative Memory (Krotov & Hopfield, 2016), which introduced a polynomial non-linearity to separate the contribution of each memory to the energy minimum. The non-linearity enabled the models to store more memories than the number of neurons (hence the term dense). Further studies extended these ideas to continuous state spaces, and exponential memory capacity (Demircigil et al., 2017). Currently, these models form the fundamental components of transformer architectures (Vaswani et al., 2017; Ramsauer et al., 2021) with high levels of performance on large-scale natural language processing tasks (Radford et al., 2018; Devlin et al., 2019) and computer vision (Carion et al., 2020) tasks. Recently, General Associative Memory Model (GAMM) (Krotov & Hopfield, 2021) unified these advances in associative mem-

[1]College of Information and Computer Sciences, University of Massachusetts Amherst [2]Computational Neurobiology Laboratory, The Salk Institute for Biological Studies. Correspondence to: Arjun Karuvally <akaruvally@umass.edu>, Hava T. Siegelmann <hava@umass.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

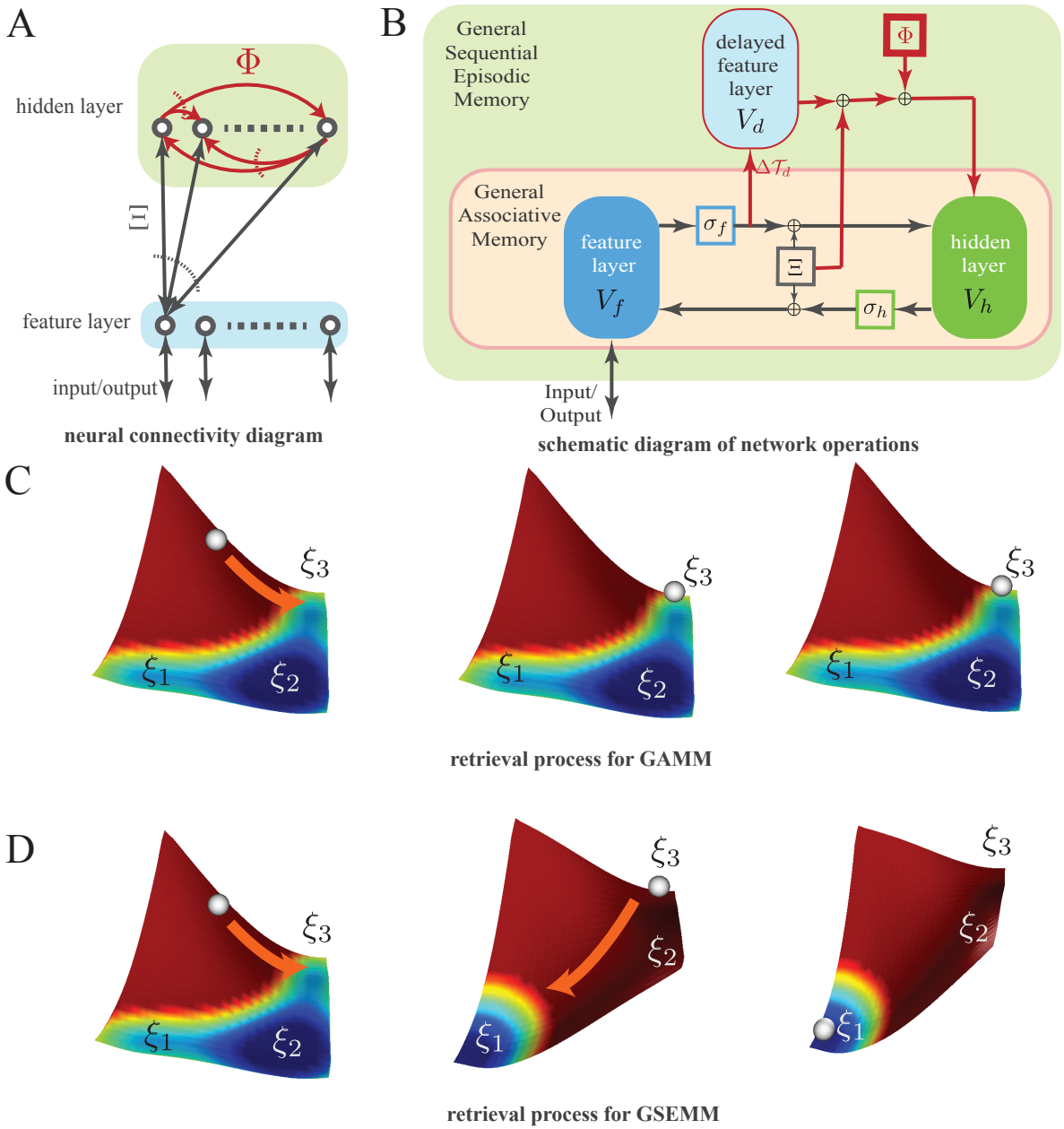

Figure 1: System architecture and schematic retrieval process for the General Sequential Episodic Memory Model(GSEMM) **A** The two-layer neural architecture with neural connectivity of GSEMM. The new delay-based synapses (Φ) we introduced (shown in red) are directed connections between neurons in the hidden layer of GAMM. Dotted curved lines in the figure indicate one-to-many connections. **B** The schematic representation of how the network performs its computations. The new synapses create a delayed signal of the feature layer neurons, which is provided as input to the hidden layer. **C** The typical retrieval process for GAMM with three stored memories (ξ_1, ξ_2, ξ_3,). The energy surface is shown by the colormap with high energy denoted by red and low energy blue. The system (shown by the white ball) flows to the basin corresponding to the nearest attractor. Due to the energy surface's static nature, GAMM stays in a low-energy memory. Consequently, the system cannot retrieve more than one memory (ξ_3 in the figure). **D** Similar retrieval process for GSEMM. Unlike GAMM, the system changes the energy surface so that a new minimum is formed that connects to a sequentially related memory ($\xi_3 \rightarrow \xi_1$ in the figure). The dynamic nature of the energy surface allows the system to adapt to the new minimum under the condition that the changes in the energy surface are adiabatic to changes in state. This feature of GSEMM enables it to retrieve more than one memory organized in sequence.

ory in a single theoretical framework. GAMM succeeded in explaining the capacity improvements through a simple energy function that characterized the long-term behavior of these models just like its predecessors. However, GAMM's state-parameterized fixed energy surface restricts it to singleton memories.

The second research direction focuses on extending energy-based models to handle sequences. In contrast to memories in singleton episodic memory, sequence memories are meta-stable states in the dynamical evolution (Rabinovich et al., 2008; Durstewitz & Deco, 2008; Camera et al., 2019). Some of the early works to produce sequential meta-stable memory (Kleinfeld, 1986; Sompolinsky & Kanter, 1986) used a combination of symmetric interactions, asymmetric interactions, and delay signals to produce stable sequential activation of memory patterns. However, these models required additional mechanisms to selectively raise the energies of states, which added complications to the use of the energy paradigm and demonstrated the difficulty of reconciling the static nature of the energy surface with the dynamical nature of models required for sequential memory retrieval. One way to alleviate this difficulty is the introduction of stochasticity (Miller & Katz, 2010; Jones et al., 2007) with sufficient noise to push the system's state beyond the basin of memory to another memory (Miller, 2016; Braun & Mattia, 2010). Models developed along these directions relaxed the symmetric constraints on the neural interactions of Hopfield Networks, resulting in a rich repertoire of dynamics (Asllani et al., 2018; Orhan & Pitkow, 2020). Theoretical proposals for meta-stable memory models used non-equilibrium landscapes where the energy function and a probability flux together determined the stability of memory states (Yan et al., 2013). In these models, stochasticity played a major role in determining the stability of meta-stable states. In contrast to the theoretical models, evidence from biology (Howard et al., 2014; Rolls & Mills, 2019; Umbach et al., 2020) has emphasized the importance of multiple timescales in SEM tasks. Empirical models (Kurikawa & Kaneko, 2021; Kurikawa, 2021) also use multiple timescales to generate SEM. However, current energy models do not take advantage of multiple timescales, and the SEM capacity is only about 12% of the number of neurons (Kurikawa & Kaneko, 2021).

In the GSEMM, we extend the static energy paradigm of the GAMM to the dynamic case using two timescales to define the dynamic behavior of the model. In the process, we discover mechanisms that significantly improve the memory capacity of sequence networks.

General Sequential Episodic Memory Model (GSEMM)

The General Sequential Episodic Memory Model unifies singleton and sequence memory retrieval. To facilitate this unification, we provide the mathematical description of GSEMM as a two layer system of interacting neurons organized according to the General Associative Memory Model (GAMM) (Krotov & Hopfield, 2021) with the addition of delay based intra-layer interactions between neurons in the hidden layer. The *feature layer* is mainly concerned with the input and output of the model. There are *no synaptic connections* between neurons in this layer. The *hidden layer* encodes abstract information about stored memories. Unlike the feature layer, the hidden layer neurons are connected using synapses that delay the signal from the feature layer. These intra-hidden layer connections enable *interactions between memories*. In the most general case, there is no restriction on the nature (in terms of symmetry) of the connections between neurons in this layer. In addition to these intra-layer connections, the neurons in the two layers are connected through symmetric synaptic interactions. The architecture for the model is shown in Figure 1B.

We now provide the mathematical description of GSEMM using the notations summarized in Table 1. We use standard linear algebra notations and indexed notations to denote states and synapses in our model. We use mainly indexed notation but switch to matrix and vector notation wherever convenient. Let $(V_f)_i$ be the current through the i^{th} neuron of the feature layer, $\sigma_f(V_f)$ be the activation function for the feature layer, $(V_h)_j$ be the current through the j^{th} neuron of the hidden layer, σ_h be the activation function for the hidden layer, and $(V_d)_i$ be the delayed feature neuron signal from the i^{th} feature neuron. The states V_f, V_h, V_d evolve with characteristic timescales $\mathcal{T}_f, \mathcal{T}_h, \mathcal{T}_d$ respectively. Let Ξ_{ij} be the strength of the synaptic connection between the neuron i in the feature layer to the neuron j in the hidden layer, Φ_{kj} be the strength of the synaptic connection from the k^{th} hidden neuron to the j^{th} hidden neuron. Similar to how memories are loaded in GAMM, each column of the matrix Ξ stores individual memories. We introduce two scalar parameters to control the strength of signals through the synapses. Let α_s, α_c be the strength of signals through the synapses Ξ and Φ respectively. The governing dynamics

Table 1: Mathematical Notation

	Symbol	Type	Meaning
Model Parameters	N_f	Z_+	number of feature layer neurons
	N_h	Z_+	number of hidden layer neurons
	α_s	R	parameter that controls the strength of feature-hidden interactions
	α_c	R	parameter that controls the strength of hidden-hidden interactions
	$\mathcal{T}_f$	R	timescale of feature layer neurons
	$\mathcal{T}_h$	R	timescale of hidden layer neurons
	$\mathcal{T}_d \gg \mathcal{T}_h, \mathcal{T}_f, 1$	R	timescale parameter attached to delay
State Definition	V_f	$R^{N_f \times 1}$	current through feature layer neurons
	$(V_f)_i$	R	current through the i^{th} feature neuron
	V_h	$R^{N_h \times 1}$	current through hidden layer neurons
	$(V_h)_i$	R	current through the i^{th} hidden neuron
	Ξ	$R^{N_f \times N_h}$	feature-hidden neurons interaction matrix
	Φ	$R^{N_h \times N_h}$	hidden-hidden neurons interaction matrix
	$I^{(n)}$	$R^{n \times n}$	Identity matrix of size n
Operators	σ_f	$R^{N_f \times 1} \to R^{N_f \times 1}$	activation function for feature layer
	σ_h	$R^{N_h \times 1} \to R^{N_h \times 1}$	activation function for hidden layer
	L_f	$R^{N_f \times 1} \to R$	Lagrangian of feature layer
	L_h	$R^{N_h \times 1} \to R$	Lagrangian of hidden layer
	$\mathcal{J}$	$R^{n \times 1} \to R^{m \times n}$	Jacobian operator on a vector valued function $f(X) : R^{n \times 1} \to R^{m \times 1}$
	$\mathcal{H}$	$R \to R^{n \times n}$	Hessian operator of a scalar valued function acting on a vector $f(X) : R^{n \times 1} \to R$

are given by:

$$\begin{cases} \mathcal{T}_f \dfrac{\mathrm{d}(V_f)_i}{\mathrm{d}t} = & \sqrt{\alpha_s} \displaystyle\sum_{j=1}^{N_h} \Xi_{ij} \left(\sigma_h(V_h)\right)_j - (V_f)_i \,, \\ \mathcal{T}_h \dfrac{\mathrm{d}(V_h)_j}{\mathrm{d}t} = & \sqrt{\alpha_s} \displaystyle\sum_{i=1}^{N_f} \Xi_{ij} \left(\sigma_f(V_f)\right)_i + \\ & \alpha_c \displaystyle\sum_{k=1}^{N_h} \sum_{i=1}^{N_f} \Phi_{kj} \, \Xi_{ik} \, (V_d)_i - (V_h)_j \,, \\ \mathcal{T}_d \dfrac{\mathrm{d}(V_d)_i}{\mathrm{d}t} = & \left(\sigma_f(V_f)\right)_i - (V_d)_i \,. \end{cases} \tag{1}$$

The dynamic evolution equations are analogous to GAMM (Krotov & Hopfield, 2021) except with the addition of delayed intra-layer synapses Φ, and two strength parameters α_s and α_c. The timescale $\mathcal{T}_d$ characterizes the timescale of delay and is assumed to be higher than the timescale of the feature and hidden layers. The delay signal is obtained by applying a continuous convolution operator (Kleinfeld, 1986) of the feature layer signal.

$$(V_d)_i = \frac{1}{\mathcal{T}_d} \int_0^\infty \left(\sigma_f(V_f(t-x))\right)_i \exp\left(-\frac{x}{\mathcal{T}_d}\right) dx. \tag{2}$$

We transformed the convolution operation to a dynamical state variable update $\frac{\mathrm{d}V_d}{\mathrm{d}t}$ to simplify the theoretical analysis of the system.

Without intra-layer synapses in the hidden layer, the GSEMM has properties of associative memory. This means that for certain conditions on the set of functions σ_f and σ_h, the long-term behavior of the state of the feature layer neurons converged to one of the stored memories. The necessary condition for convergence is that the dynamical trajectory of the system follows an energy function with minima near the stored memory states. The delay-based synapses we introduced enable the energy function to change with time, so the long-term behavior is not just a single memory but a sequence of related memories.

Energy Dynamics

The energy dynamics of the system is analyzed by considering the new delay variable V_d as a control parameter. We show that for a delay signal V_d that is changing *sufficiently slowly* compared to V_h and V_f, the energy function evaluated at the instantaneous state V_d can still be used to characterize the dynamical nature of V_f and V_h. The term *sufficiently slowly* means that V_f and V_h converge to their instantaneous attractor states before V_d changes the energy surface. To derive the energy function, we use two Lagrangian terms L_f and L_h for the feature and hidden neurons respectively (Krotov & Hopfield, 2021), defined as

$$\sigma_f(V_f) = \mathcal{J}(L_f)^\top, \text{ and } \sigma_h(V_h) = \mathcal{J}(L_h)^\top. \tag{3}$$

The new energy function (Appendix A.1) for GSEMM is derived as.

$$\begin{aligned} E = & \left[V_f^\top \, \sigma_f(V_f) - L_f\right] + \left[V_h^\top \, \sigma_h(V_h) - L_h\right] \\ & - \sqrt{\alpha_s} \left[\sigma_f(V_f) \, \Xi \, \sigma_h(V_h)\right] - \alpha_c \left[V_d^\top \, \Xi \, \Phi \, \sigma_h(V_h)\right]. \end{aligned} \tag{4}$$

At this point, it is instructive to note that without the additional synapses, Φ, and setting the strength parameter $\alpha_s = 1$, the system and the associated energy function reduce to GAMM energy with only singleton episodic memory.

In order to analyze how the dynamics of energy change with the introduction of delay based synapses, we take the time derivative of the GSEMM energy function along the dynamical trajectory of the system. We assume the conditions of positive semi-definite Hessians of the Lagrangian terms and bounded activation functions σ_f and σ_h (Krotov & Hopfield, 2021). It is to be noted that the entire state description of the system consists of three vectors V_f, V_d, and V_h. These states are grouped as a fast subsystem V_f and V_h, and a slow subsystem V_d. The analysis becomes easier when we consider the slow subsystem as a control variable of the fast subsystem. This allows the characterization of the state dynamics of the fast subsystem as instantaneous fixed point attractor dynamics modulated by input from the slow subsystem.

The dynamical evolution of the energy function after separating the slow and fast subsystems is given as (Appendix A.2),

$$\frac{\mathrm{d}E}{\mathrm{d}t} = F(\frac{\mathrm{d}V_f}{\mathrm{d}t}, \frac{\mathrm{d}V_h}{\mathrm{d}t}) + G(\frac{\mathrm{d}V_d}{\mathrm{d}t})\,. \tag{5}$$

$$\begin{aligned} F(\frac{\mathrm{d}V_f}{\mathrm{d}t}, \frac{\mathrm{d}V_h}{\mathrm{d}t}) = -\Bigg[& \mathcal{T}_f \left(\frac{\mathrm{d}V_f}{\mathrm{d}t}\right)^\top \mathcal{H}(L_f) \frac{\mathrm{d}V_f}{\mathrm{d}t} + \\ & \mathcal{T}_h \left(\frac{\mathrm{d}V_h}{\mathrm{d}t}\right)^\top \mathcal{H}(L_h) \frac{\mathrm{d}V_h}{\mathrm{d}t} \Bigg]\,. \end{aligned} \tag{6}$$

$$G(\frac{\mathrm{d}V_d}{\mathrm{d}t}) = -\alpha_c \left[\sigma_h(V_h)^\top \Phi^\top \Xi^\top \frac{\mathrm{d}V_d}{\mathrm{d}t} \right].$$

F and G separate the contributions of the two timescales - the fast ($\{\mathcal{T}_f, \mathcal{T}_h\}$) and slow ($\{\mathcal{T}_d\}$). It can be easily seen that among the two terms, only G is affected by the timescale of the delay signal. Just like in GAMM, under the assumption of positive semi-definite Hessian of the Lagrangian and bounded energy, we get,

$$F(\frac{\mathrm{d}V_f}{\mathrm{d}t}, \frac{\mathrm{d}V_h}{\mathrm{d}t}) \leq 0\,. \tag{7}$$

The inequality means that the fast subsystem can have two possible long-term behaviors when F eventually converges to zero. One behavior is convergence to a single stable state corresponding to minima of the energy function leading to fixed point attractor dynamics. The second possible behavior is when the system moves in an iso-energetic trajectory without convergence. In this paper, we focus only on the case of the fixed point attractor behavior of the system.

Like in GAMM, the fixed point attractor behavior of the system acts to stabilize the dynamics on the energy surface such that the energy is non-increasing and convergent, but unlike GAMM, delay based synapses lead to another term G.

$$G(\frac{\mathrm{d}V_d}{\mathrm{d}t}) = -\alpha_c \left[\sigma_h(V_h)^\top \Phi^\top \Xi^\top \frac{\mathrm{d}V_d}{\mathrm{d}t} \right]. \tag{8}$$

It may be difficult to specify the system's behavior for any general choice of Φ, σ_h, and V_d. However, in the adiabatic limit of the slow subsystem (under the condition that $\mathcal{T}_d \gg 1$ and $\mathcal{T}_d \gg \mathcal{T}_f, \mathcal{T}_h$), the system can still exhibit a non-increasing energy function because $\frac{\mathrm{d}V_d}{\mathrm{d}t} \to 0$ in this limit and $G \to 0$. This condition is especially true when analyzing the dynamic properties of the fast subsystem ($\frac{\mathrm{d}V_f}{\mathrm{d}t} \neq 0$ and $\frac{\mathrm{d}V_h}{\mathrm{d}t} \neq 0$), which is the property that seems to be relevant in dynamic memory models. The delay signals thus have two functions. The slow-changing nature of the delay signal helps to stabilize the dynamics of the fast subsystem on the energy surface. The second function is that the delay signal changes the energy surface to create new minima and destroy old minima. In our numerical simulations, we consider high enough settings of $\mathcal{T}_d$ such that V_d changes sufficiently slowly for the energy function to characterize the dynamics but not so high as to prevent the system from exhibiting state transitions in a reasonable time. It is to be noted that although we consider a specific case for the delay signal, the adiabatic theory is general enough to apply to any adiabatic control signal to the previously introduced general associate memory model (GAMM).

Dense GSEMM

We next derive the activation functions which will make GSEMM dense and thus most useful both as human memory and for AI. Analogous to how practical models are derived from GAMM, we consider the diabatic limit of hidden neurons, $\mathcal{T}_h \to 0$, such that.

$$\begin{aligned} (V_h)_j = & \sqrt{\alpha_s} \sum_{i=1}^{N_f} \Xi_{ij} \left(\sigma_f(V_f)\right)_i \\ & + \alpha_c \sum_{k=1}^{N_h} \sum_{i}^{N_f} \Phi_{kj}\, \Xi_{ik}\, (V_d)_i\,. \end{aligned} \tag{9}$$

Substituting this in the dynamical evolution of feature neurons we get,

$$\begin{aligned} \mathcal{T}_f \frac{\mathrm{d}(V_f)_i}{\mathrm{d}t} = & \sqrt{\alpha_s} \sum_{j=1}^{N_h} \Xi_{ij}\, \sigma_h(\Xi^\top \sigma_f(V_f) + \\ & \alpha_c \Phi^\top \Xi^\top V_d)_j - (V_f)_i\,. \end{aligned} \tag{10}$$

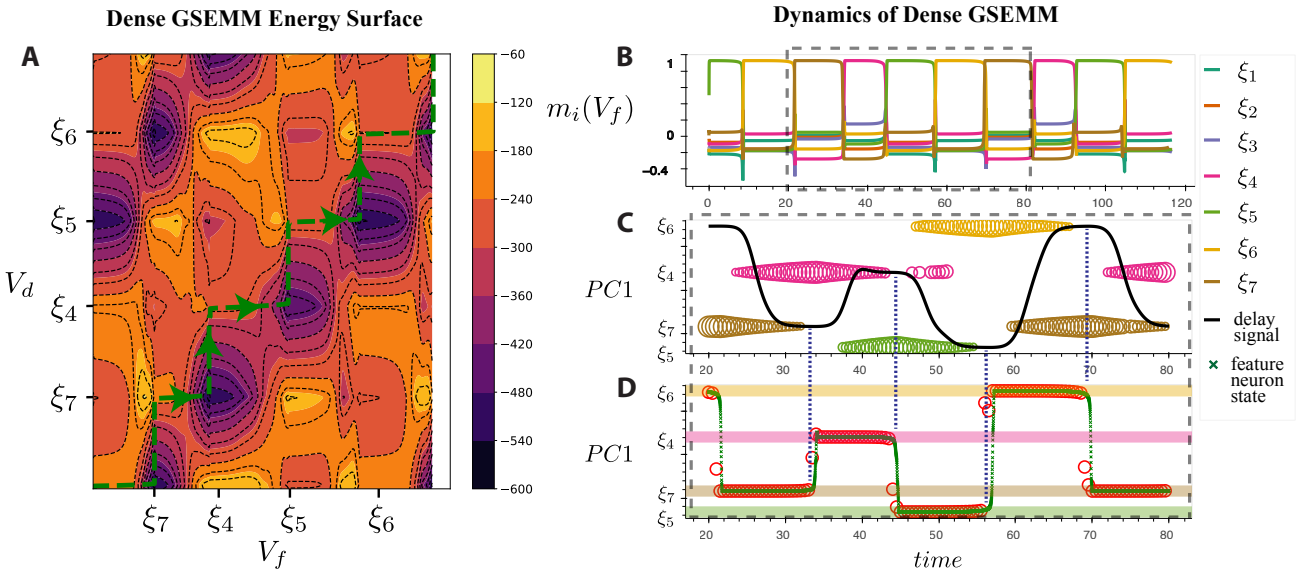

Figure 2: Energy dynamics of the Dense GSEMM shows the instantaneous fixed point attractor behavior. **A** The energy surface for the feature layer states V_f (x-axis), and the delayed feature layer states V_d (y-axis) visited by it in one simulation during the time range $20 - 80$. The plot shows that V_d changes the energy surface so that a new state for V_f becomes a new basin of attraction (darker shade) while the previous basin increases energy. The trajectory (green line) of the system during retrieval shows the movement of V_f to the *instantaneous* minima of the energy surface. **B** The dynamics of V_f on a SEM task. The time range 20-80 over which energy behavior is plotted is shown in the gray dotted box. The y-axis denotes the overlap between the feature neuron state and each stored memory. **C** The first principal component ($PC1$) of V_d and how it influences the fixed points of the energy function near *all* stored memories (depicted by colored circles). The size of the circle is inversely proportional to the fixed points' energy. The absence of circles near the state transition point shows the momentary loss in the fixed point stability. **D** The first principal component ($PC1$) of the dynamical evolution of the nearest fixed point (red circle) of the energy surface and the current state of the fast sub-system (green cross). The evolution shows how the system is attracted to the nearest fixed point from the current state of the energy surface at each point in time.

It can be seen from the dynamical evolution of the feature neurons that depending on the setting of σ_h, the feature-hidden synapses may interact linearly with hidden-feature synapses and hidden-hidden synapses. In Dense GSEMM, we choose the feature layer activation function to be the $\tanh$ non-linearity - $\sigma_f(x) = \tanh(x)$, and the hidden layer activation to be the polynomial activation function of degree n - $\sigma_h(x) = x^n$ for $n \geq 0$. When $n = 1$, the parametric model has linear memory interactions, and $n > 1$ progressively increases the order of these interactions. These settings are analogous to polynomial interactions in associative memory models that improved their capacity significantly (Krotov & Hopfield, 2016).

According to the theoretical analysis above, the system follows the *instantaneous* minima of the energy function and flows from memory to memory via the slow updates to the energy function. We validate this for the case of $n = 1$ with simulation in Figure 2. We plot the energy function and the state of the system as time progresses for a simulated episode of the system. The momentary loss in stability of fixed points near memories that allow for state transition can be clearly observed in the figure. The addition of the adiabatic control signal provides a way to adaptively change the energy surface once the system converges so that dynamic properties can emerge from static energy models.

In Figure 3, we show that increasing the polynomial degree of the hidden layer activation increases the sequence memory capacity, similar to what is observed in associative memory models with polynomial activations. This indicates that the proposed polynomial non-linearity in GSEMM can effectively improve the capacity of sequence memories.

Discussion

The General Sequential Episodic Memory Model (GSEMM) introduced in this paper is an approach to encoding memories along with their sequential relationships. The key to this capability is the slow-fast timescale dynamics created by adding delay-based synapses in the hidden layer of the General Associative Memory Model. These delay-based

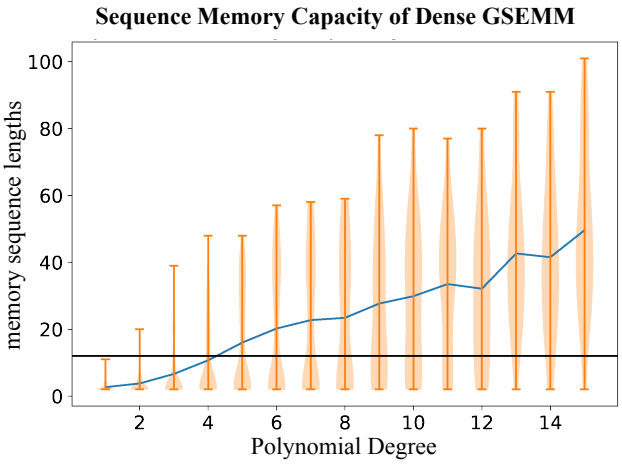

Figure 3: Memory capacity grows with the polynomial degree in the hidden layer. The plot shows the distribution of model capacities in 100 neuron network for different sets of random binary vectors along with the mean of the distributions (blue line) for 30 trials. The distributions can be compared against the baseline of 12 length sequence (black horizontal line). The baseline performance seems to be maximum number of memories that can be stored in a polynomial degree=1 Dense GSEMM. It can be seen that for polynomial degree=15, there are models (with the highest sequence length) capable of storing more than 100 memories making the SEM capacity dense in the number of neurons.

synapses lead to an energy surface that slowly changes with time, allowing the system to adapt to its instantaneous fixed points. This shift from a single fixed point to a sequence of instantaneous fixed points enables the storage of memory sequences in meta-stable states during the dynamic evolution of the system. Numerical simulations revealed that increasing the degree of the polynomial interactions in the hidden layer directly increases the sequence memory capacity of the system. These structural modifications provide a roadmap for capacity improvements in sequence episodic memory models.

We suggest GSEMM as the most relevant energy-based memory model to date in that it incorporates sequences. Although the current version of GSEMM is limited in its biological plausibility due to the presence of symmetric interactions and the specialized use of delays, the fact that GSEMM utilizes mechanisms observed in episodic memory experiments shows that it may be possible to build future models that close the gap between artificial neural networks and neuroscience. In addition to the potential impact on neuroscience, the capacity experiments suggest that Dense GSEMM may be used in machine learning applications requiring robust, high-capacity sequence memory storage and retrieval.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112190041. We thank Peter Delmastro for valuable insights leading to the development of the model. We thank Edward Rietman, Andre Pietrzykowski, Joshua Russel, Adam Kohan, and the reviewers for valuable feedback to improve the manuscript

References

Amari, S. Neural theory of association and concept-formation. *Biological Cybernetics*, 26:175–185, 2004.

Amit, Gutfreund, and Sompolinsky. Spin-glass models of neural networks. *Physical review. A, General physics*, 32 2:1007–1018, 1985.

Asllani, M., Lambiotte, R., and Carletti, T. Structure and dynamical behavior of non-normal networks. *Science Advances*, 4, 2018.

Braun, J. and Mattia, M. Attractors and noise: Twin drivers of decisions and multistability. *NeuroImage*, 52:740–751, 2010.

Camera, G. L., Fontanini, A., and Mazzucato, L. Cortical computations via metastable activity. *Current Opinion in Neurobiology*, 58:37–45, 2019.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. *ArXiv*, abs/2005.12872, 2020.

Demircigil, M., Heusel, J., Löwe, M., Upgang, S., and Vermet, F. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168:288–299, 2017.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

Duff, M. C., Covington, N. V., Hilverman, C., and Cohen, N. J. Semantic memory and the hippocampus: Revisiting, reaffirming, and extending the reach of their critical relationship. *Frontiers in Human Neuroscience*, 13, 2019.

Durstewitz, D. and Deco, G. Computational significance of transient dynamics in cortical networks. *European Journal of Neuroscience*, 27, 2008.

Folli, V., Leonetti, M., and Ruocco, G. On the maximum storage capacity of the hopfield model. *Frontiers in Computational Neuroscience*, 10, 2016.

Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554. URL https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554.

Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., and Eichenbaum, H. A unified mathematical framework for coding time, space, and sequences in the hippocampal region. *The Journal of Neuroscience*, 34:4692 – 4707, 2014.

Jones, L. M., Fontanini, A., Sadacca, B. F., Miller, P. I., and Katz, D. B. Natural stimuli evoke dynamic sequences of states in sensory cortical ensembles. *Proceedings of the National Academy of Sciences*, 104:18772 – 18777, 2007.

Kleinfeld, D. Sequential state generation by model neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 83 24:9469–73, 1986.

Krotov, D. and Hopfield, J. J. Dense associative memory for pattern recognition. In *Proceedings of Thirtieth Conference on Neural Information Processing Systems*, 2016.

Krotov, D. and Hopfield, J. J. Large associative memory problem in neurobiology and machine learning. *ArXiv*, abs/2008.06996, 2021.

Kurikawa, T. Transitions among metastable states underlie context-dependent working memories in a multiple timescale network. In *ICANN*, 2021.

Kurikawa, T. and Kaneko, K. Multiple-timescale neural networks: Generation of history-dependent sequences and inference through autonomous bifurcations. *Frontiers in Computational Neuroscience*, 15, 2021.

McEliece, R. J., Posner, E. C., Rodemich, E. R., and Venkatesh, S. S. The capacity of the hopfield associative memory. *IEEE Trans. Inf. Theory*, 33:461–482, 1987.

Miller, P. I. Itinerancy between attractor states in neural systems. *Current Opinion in Neurobiology*, 40:14–22, 2016.

Miller, P. I. and Katz, D. B. Stochastic transitions between neural states in taste processing and decision-making. *The Journal of Neuroscience*, 30:2559 – 2570, 2010.

Orhan, A. E. and Pitkow, X. Improved memory in recurrent neural networks with sequential non-normal dynamics. *ArXiv*, abs/1905.13715, 2020.

Rabinovich, M. I., Huerta, R., Varona, P., and Afraimovich, V. S. Transient cognitive dynamics, metastability, and decision making. *PLoS Computational Biology*, 4, 2008.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training (2018), 2018.

Ramsauer, H., Schafl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Pavlovi'c, M., Sandve, G. K., Greiff, V., Kreil, D. P., Kopp, M., Klambauer, G., Brandstetter, J., and Hochreiter, S. Hopfield networks is all you need. *ArXiv*, abs/2008.02217, 2021.

Renoult, L., Irish, M., Moscovitch, M., and Rugg, M. D. From knowing to remembering: The semantic–episodic distinction. *Trends in Cognitive Sciences*, 23:1041–1057, 2019.

Rolls, E. T. and Mills, P. The generation of time in the hippocampal memory system. *Cell reports*, 28 7:1649–1658.e6, 2019.

Sompolinsky and Kanter. Temporal association in asymmetric neural networks. *Physical review letters*, 57 22: 2861–2864, 1986.

Tulving, E. Episodic memory: from mind to brain. *Annual review of psychology*, 53:1–25, 2002.

Umbach, G. S., Kantak, P. A., Jacobs, J., Kahana, M. J., Pfeiffer, B. E., Sperling, M. R., and Lega, B. C. Time cells in the human hippocampus and entorhinal cortex support episodic memory. *Proceedings of the National Academy of Sciences of the United States of America*, 117: 28463 – 28474, 2020.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NIPS*, 2017.

Yan, H., Zhao, L., Hu, L., Wang, X., Wang, E., and Wang, J. Nonequilibrium landscape theory of neural networks. *Proceedings of the National Academy of Sciences*, 110: E4185 – E4194, 2013.

Appendix

A. Energy

The most important aspect of the model we discussed is the energy function. We use the function to show the behavior of the system in the adiabatic case and compute instantaneuos attractors.

A.1. Energy Function for GSEMM

Here, we will derive the Energy function of GSEMM starting from a previously derived energy function used for associative memory. Assume a signal $\mathcal{I}_h$ applied to the neurons in the hidden layer.

$$E = \left[V_f^\top \, \sigma_f(V_f) - L_f \right] + \left[(V_h - \mathcal{I}_H)^\top \, \sigma_h(V_h) - L_h \right] - \left[\sqrt{\alpha_s} \, \sigma_f(V_f) \, \Xi \, \sigma_h(V_h) \right] \tag{11}$$

In our case, the input signal comes from the delay signal activity V_d and is given as $\mathcal{I}_h = \alpha_c \Phi^\top \, \Xi^\top \, V_d$ from our governing dynamics. Substituting this in the energy equation

$$E = \left[V_f^\top \, \sigma_f(V_f) - L_f \right] + \left[(V_h - \alpha_c \Phi^\top \, \Xi^\top \, V_d)^\top \, \sigma_h(V_h) - L_h \right] - \left[\sqrt{\alpha_s} \, \sigma_f(V_f)^\top \, \Xi \, \sigma_h(V_h) \right] \tag{12}$$

Expanding this equation, we get

$$E = \left[V_f^\top \, \sigma_f(V_f) - L_f \right] + \left[V_h^\top \, \sigma_h(V_h) - L_h \right] - \left[\sqrt{\alpha_s} \, \sigma_f(V_f)^\top \, \Xi \, \sigma_h(V_h) \right] - \alpha_c \left[V_d^\top \, \Xi \, \Phi \sigma_h(V_h) \right] \tag{13}$$

A.2. Energy Function Dynamics

To find how the energy function behaves along the dynamical trajectory of the system. Taking the derivative of the energy function with respect time

$$\begin{aligned} \frac{\mathrm{d}E}{\mathrm{d}t} = & \left[V_f^\top \, \mathcal{J}(\sigma_f) \, \frac{\mathrm{d}V_f}{\mathrm{d}t} + \sigma_f(V_f)^\top \frac{\mathrm{d}V_f}{\mathrm{d}t} - \frac{\mathrm{d}L_f}{\mathrm{d}t} \right] + \left[V_h^\top \mathcal{J}(\sigma_h) \, \frac{\mathrm{d}V_h}{\mathrm{d}t} \right. \\ & \left. + \sigma_h(V_h)^\top \, \frac{\mathrm{d}V_h}{\mathrm{d}t} - \frac{\mathrm{d}L_h}{\mathrm{d}t} \right] - \left[\sqrt{\alpha_s} \, \sigma_f(V_f)^\top \, \Xi \, \mathcal{J}(\sigma_h) \, \frac{\mathrm{d}V_h}{\mathrm{d}t} \right. \\ & \left. + \sqrt{\alpha_s} \, \sigma_h(V_h)^\top \, \Xi^\top \, \mathcal{J}(\sigma_f) \, \frac{\mathrm{d}V_f}{\mathrm{d}t} \right] - \alpha_c \, \frac{\mathrm{d}}{\mathrm{d}t} \left[V_d^\top \, \Xi \, \Phi \mathcal{J}(\sigma_h) \frac{\mathrm{d}V_h}{\mathrm{d}t} + \sigma_h(V_h)^\top \, \Xi^\top \, \Phi^\top \frac{\mathrm{d}V_d}{\mathrm{d}t} \right] \end{aligned} \tag{14}$$

The derivatives of the lagrangian terms can be converted as $\frac{\mathrm{d}L_f}{\mathrm{d}t} = \sigma_f(V_f)^\top \frac{\mathrm{d}V}{\mathrm{d}t}$ and $\frac{\mathrm{d}L_h}{\mathrm{d}t} = f(V_h)^\top \frac{\mathrm{d}V_h}{\mathrm{d}t}$. Substituting these.

$$\begin{aligned} \frac{\mathrm{d}E}{\mathrm{d}t} = & \left[V_f^\top \, \mathcal{J}(\sigma_f) \, \frac{\mathrm{d}V_f}{\mathrm{d}t} + V_h^\top \mathcal{J}(\sigma_h) \, \frac{\mathrm{d}V_h}{\mathrm{d}t} \right] - \left[\sqrt{\alpha_s} \, \sigma_f(V_f)^\top \, \Xi \, \mathcal{J} \, \frac{\mathrm{d}V_h}{\mathrm{d}t} \right. \\ & \left. + \sqrt{\alpha_s} \, \sigma_h(V_h)^\top \, \Xi^\top \, \mathcal{J}(\sigma_f) \, \frac{\mathrm{d}V_f}{\mathrm{d}t} \right] - \alpha_c \, \frac{\mathrm{d}}{\mathrm{d}t} \left[V_d^\top \, \Xi \, \Phi \mathcal{J}(\sigma_h) \frac{\mathrm{d}V_h}{\mathrm{d}t} + \sigma_h(V_h)^\top \, \Xi^\top \, \Phi^\top \frac{\mathrm{d}V_d}{\mathrm{d}t} \right] \end{aligned} \tag{15}$$

Rearranging terms

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -\Bigg[\left(\sqrt{\alpha_s}\,\sigma_h(V_h)^\top \Xi^\top - V_f^\top\right) \mathcal{J}(\sigma_f) \frac{\mathrm{d}V_f}{\mathrm{d}t} + \left(\sqrt{\alpha_s}\,\sigma_f(V_f)^\top \Xi + \alpha_c V_d^\top \Xi \Phi - V_h^\top\right) \mathcal{J}(\sigma_h) \frac{\mathrm{d}V_h}{\mathrm{d}t} \Bigg] - \alpha_c \left[\sigma_h(V_h)^\top \Xi^\top \Phi^\top \frac{\mathrm{d}V_d}{\mathrm{d}t} \right] \tag{16}$$

Substituting from dynamical equations

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -\left[\mathcal{T}_f \frac{\mathrm{d}V_f}{\mathrm{d}t}^\top \mathcal{H}(L_f) \frac{\mathrm{d}V_f}{\mathrm{d}t} + \mathcal{T}_h \frac{\mathrm{d}V_h}{\mathrm{d}t}^\top \mathcal{H}(L_h) \frac{\mathrm{d}V_h}{\mathrm{d}t} \right] - \left[\sigma_h(V_h)^\top \Xi^\top \Phi^\top \frac{\mathrm{d}V_d}{\mathrm{d}t} \right] \tag{17}$$

B. Simulation Procedures - Figure 1

We used the fourth order Runge-Kutta numerical procedure with step size 0.01 for numerical simulations. The output for the Dense GSEMM is the state of the feature neurons. The similarity between the output and memory is evaluated using the overlap of the feature neuron state with each memory in the system, which is defined as $m_i(V_f) = (1/N_f) \sum_j^{N_f} (\xi_i)_j \, (\sigma_f(V_f))_j$ where $\xi^{(i)}$ is the i^{th} memory in the system. Each memory in the model is a random binary vector such that $\Pr\left[\xi_j^{(i)} = +1\right] = \Pr\left[\xi_j^{(i)} = -1\right] = 1/2$. These memories are organized as 2 separate cyclical episodes: $\xi_1 \to \xi_2 \to \xi_3 \to \xi_1$ and $\xi_4 \to \xi_5 \to \xi_6 \to \xi_7 \to \xi_4$ with their sequential relationships stored as an adjacency matrix in Φ. Two key factors were considered when we used these two episodes for evaluation. One factor is to demonstrate the ability of the model to extract only the related stored episode, even in the presence of other episodes. The second factor is that successfully generating the stored episode requires long-term non-equilibrium behavior with no stable states in the dynamics. The input to the system is the memory ξ_5 which retrieves the 4 length cycle.

B.1. Parameter Settings

We simulated Dense GSEMM with $N_f = 100, \alpha_s = 0.05, \alpha_c = 0.007, \mathcal{T}_f = 1.0$, and $\mathcal{T}_d = 20.0$. The power of the polynomial non-linearity is 1 for the energy surface shown in the figure. The code for the simulations are available in the repository: https://github.com/arjun23496/gsemm.

B.2. Fixed point analysis

We used a fixed point finding algorithm to find the fixed points of the energy surface in Figure 1. The algorithm uses an iterative process to find the fixed points of the energy surface evaluated from a certain point in the state space. Starting from the neuron state on the energy landscape, the state is updated to follow the direction of the energy gradient till no more updates are possible, indicating convergence to a fixed point on the energy surface. This fixed point is also one meta-stable point in the network dynamics since it loses stability over time.

C. Simulation Procedures - Figure 3

We used the fourth order Runge-Kutta numerical procedure with step size 0.01 for numerical simulations. Each memory in the model is a random binary vector such that $\Pr\left[\xi_j^{(i)} = +1\right] = \Pr\left[\xi_j^{(i)} = -1\right] = 1/2$. These memories are organized in a single cyclical episode with K memories such that: $\xi_1 \to \xi_2 ... \xi_K \to \xi_1$ with their sequential relationships stored as an adjacency matrix in Φ. Φ is thus a circulant matrix with a single cycle.

We simulated using this procedure, multiple instantiations of Dense GSEMM with polynomial degrees ranging from $1 - 15$ with 30 different seeds for each polynomial degree. The parameters $N_f = 100, \mathcal{T}_f = 1.0, \mathcal{T}_d = 20.0, \alpha_s = 1.0$ are fixed. The parameter $\alpha_c \in \{0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ is treated as hyperparameter and optimized to find the maximum number of memories that can be stored for each [seed, polynomial] pair.

We define that the model successfully retrieves the K size memory cycle if the output of the model can successfully retrieve the memories in the cycle in the correct order at least 3 times with no errors.