RESEARCH STATEMENT

MILTON LIN

Personal Experiences

My research in pure mathematics has centered on bridging discrete structures (number theory) with continuous spaces (topology) through algebraic formalism. I have attached them to the end of the document for those of interest. Building on my background, I aim to explore how **algebraic and categorical methods** can reveal qualitative information on the learning dynamics of modern language models and neural circuits. This involves examining **biological realism** in computational frameworks.

My Motivation from Biology. Biological neural circuits exhibit complex architectures and dynamics, including diverse synapse types, intricate connectivity patterns, and activity-dependent plasticity. Network models provide essential tools to bridge the biological substrates with computational properties observed in experiments. Simplified models-ranging from binary threshold neurons to Hodgkin-Huxley-type models-offer diverse insights, albeit with trade- offs [AK16; BS23]. My particular interest lies in variations of Hopfield network models [KH16; BF23], originally developed for studying content-addressable memory [Hop84]. Despite limited biological plausibility, these models capture fundamental aspects of memory storage and retrieval, and have connection to deep learning architectures such as transformers and recurrent neural networks, [Niu+24]. To address the limitations of backpropagation and explore biologically plausible credit assignment [Oro+24], I was drawn to the study of Minimum Energy Flow [HMK14] as a promising alternative method for network learning. These works highlight the need for a coherent framework, just like classical statistical learning theory, to study the learning dynamics in parameter spaces. Consequently, my research employs tools from tropical geometry and category theory to investigate the combinatorial and geometric structures inherent in parameter spaces of various networks. In Section 1, I detail my collaborative work with Chris Hillar (Redwood Research).

In Section 2, I discuss an extension of the previous project on the practical end. I aim to study modern deep-learning networks through the lens of (dense) associative memories. Combined, the above two research aim to study the limitations of synthetic memory networks and provide an algebraic framework to coherently study various families of models.

In Section 3, I discuss two future projects I wish to embark on that leverage my background in mathematics: Section 3.1 discuss categorical frameworks to study scaling properties, and Section 3.2 discuss reinforcement learning algorithm development in the baby context of mathematical research problems.

1. Geometric Structures in Parameter Spaces

Given a model architecture $\mathcal{A}$, such as transformers, CNN, multilayer perceptron, designed to interpolate a task $\mathcal{T}$, we investigate the encoded information within the parameter space $\mathrm{Par}_{\mathcal{A}}$. We define the mapping:

$$\mathcal{A}_{(-)} : \mathrm{Par}_{\mathcal{A}} \to \mathcal{T} \quad \Theta \mapsto \mathcal{A}_{\Theta}$$

which assigns each parameter Θ, $\mathcal{A}_{\Theta} \in \mathcal{T}$, an object of architecture $\mathcal{A}$, designed for a task, $\mathcal{T}$. For instance, let $\mathcal{A} := \mathrm{FF}\,[n, \sigma]$ be the class of L-layer feedforward neural network with hyperparameters:

width $n = (n_i)_{i=1}^{L+1}$ and collection of activation functions $(\sigma_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_{i+1}})_{i=1}^{L}$. For a set of a parameter

$$\Theta := \{A_i, b_i\}_{i=1}^{L} \in \mathbb{R}^{\sum_{i=1}^{L} n_i(n_i+1)}$$

we can associate a function

$$\mathcal{A}_\Theta := f_L \circ \cdots \circ f_1 : \mathbb{R}^{n_1} \to \mathbb{R}^{n_{L+1}}$$

where for each $i = 1, \dots, L$, and $f_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_{i+1}}$ is a linear function given by

$$f_i := \sigma_i \left(A_i x_i + b_i \right),$$

This induces our desired map, from space of parameters to the set of functions from $\mathbb{R}^{n_1}$ to $\mathbb{R}^{n_{L+1}}$.

$$\mathcal{A}_{(-)} : \mathrm{Par}_{\mathcal{A}} \to \mathrm{Fct}(\mathbb{R}^{n_1}, \mathbb{R}^{n_{L+1}})$$

$$\Theta \mapsto \mathcal{A}_\Theta$$

The pair $(\mathcal{A}(-), \mathcal{T})$, partitions the collection of parameters according to those which induces the same object $f \in \mathcal{T}$. We obtain a collection of subsets of parameter space

$$\Sigma_{\mathcal{T}}[\mathrm{Par}_{\mathcal{A}}] = \{\mathrm{Par}_f\}_{f \in I(\mathcal{D})} \quad \mathrm{Par}_f = \{\Theta \in \mathrm{Par}_{\mathcal{A}} : \mathcal{A}_\Theta = f\}$$

by considering regions inducing the same object under the mapping $\mathcal{A}_{(-)}$. Often, $\Sigma_{\mathcal{T}}[\mathrm{Par}_{\mathcal{A}}]$ is more than a *set*, but is a set with *structure*. For deep feedforward neural networks, this structure manifests as the face poset of hyperplane arrangement, encapsulating both model expressivity and decision boundaries.

We studied the case when $\mathcal{A}$ is a Hopfield network, and the number of top dimensional faces in $\Sigma_{\mathcal{T}}[\mathcal{A}]$ reduces to a problem of counting regions induced by hyperplane arrangements, and is explicitly given by Zaslavsky's theorem, [Sta07]. A simple corollary is:

Corollary 1.1 (Hillar, Lin)**.** *A Hopfield network of two nodes with asynchronous (or synchronous) updates cannot express XOR functions.*

We are currently focusing on parameter estimation. How is the dynamics of parameter estimation reflected in the parameter space? [Koh+22] has studied a similar story for linear convolutional networks. Our particular estimation method of interest in *minimum probability flow*, [SBD20], which has been used by joint author Chris Hillar in the case of Hopfield networks, [HMK14]. We summarize here where the future research would focus on:

(1) Understanding a suitable notion of "equivalence class of networks". What can we say about pairs $(\mathcal{A}, \mathcal{T}), (\mathcal{A}', \mathcal{T}')$ and the differences in their induced face posets, $\Sigma_{\mathcal{T}}[\mathrm{Par}_{\mathcal{A}}], \Sigma_{\mathcal{T}'}[\mathrm{Par}_{\mathcal{A}'}]$ Similarly, can one describe distributions on the $\mathrm{Par}_{\mathcal{A}}$ that corresponds to networks with certain properties?[1] We hope this gives a coarse-grained comparisons of various networks.

(2) Extend our analysis of architectures *higher order networks*, which includes transformer networks, and simplicial hopfield network [BF23]. $\mathrm{Par}_{\mathcal{A}}$ decomposes into semi-algebraic sets, and one may approach with the theory of splines, [LLL24].

[1] A similar question was asked in [Mon+14].

2. Understanding modern networks through memory networks

Recent developments in associative memory networks have significantly advanced these models along two fronts: i) *Improved storage capacity*, progressing from polynomial [KH16], to exponential [Dem+17], and in other point of views, [HT14] ii) *Integration into modern deep learning architectures*, such as attention mechanisms [Ram+21], energy-based transformers [Hoo+23], and higher-order models like simplicial Hopfield networks [BF23]. Their relations with, and their potential to explain, modern transformer-based decoder models are under explored.

Joint with Chris Hillar, Muhan Gao (Johns Hopkins University), and Tenzin Chan (Algebraic) we evaluate dense assocative memories, [KH16] beyond the theoretical memory capacity, see Equation (5) and (6) of *op. cit.*. While much effort has been focused on designing networks that extends the memory capacity, there is little work on studying such regimes. Our first empirical results show that storage capacity is not a hard constraint to task performance. Such insensitivity to memory capacity echoes trends seen in scaling laws of deep learning. Moving forward, we are exploring

(1) *Generalization and catastrophic forgetting:* The behavior of stored memory patterns appears highly sensitive to the nature of the task. How does task variability influence memory retrieval, and could this sensitivity offer insights into *catastrophic forgetting*? Understanding this phenomenon, especially in the context of continual learning, could bridge memory networks with advances in lifelong machine learning [Kem+17].

(2) *Correlated data and memory convergence* Experimental evidence shows that correlated datasets significantly alter convergence behavior to stored memory patterns. Can these observations be formalized theoretically? A deeper understanding of how data structure impacts memory retrieval could inform both theoretical bounds and practical applications.

The end goal is to provide both empirical and theoretical comparison with modern networks; works along these lines include, [ND21], [Niu+24], and [CDB24].

3. Future research projects

3.1. **Categorical Models and Homotopy Theory.** The following project extends previous project Section 1. Categorical approaches have gained momentum as a systematic framework for studying network structures [Gav+24]. This has been particularly successful in the field of *geometric deep learning* [Bro+21], where abstract mathematical structures help describe complex neural networks. We propose to explore memory networks using a recent formalism by Manin et al. [MM24], which uses summing functors and Gamma spaces to model the allocation of resources in neural networks. These concepts will allow us to understand how the complexity of memory networks scales as network size increases. The formalism allows us to study a *homotopy type* - a mathematical construct at a deeper level than *homology*[2]. Homotopy captures invariants of network up to continuous deformations. Previous studies have shown that stimulus space can be reconstructed up to homotopy [Man15].

Specifically, we will examine how *memory capacity* correlates with *homotopical invariants* like *Betti numbers* (which measure the number of independent cycles in a space) and *simplicial complexes* (which provide a higher-dimensional generalization of networks). Burns and Fukai have already done early work in this direction [BF23], but much remains to be explored.

[2] which is commonly used in topological data analysis (TDA). For a short survey of topology and neural code, see [Cur16].

3.2. **Reinforcement Learning Approaches in Mathematical Research.** Building on a recent study of applications of reinforcement learning to mathematics - a research problem in combinatorial group theory, [She+24] - I hope to continue further and develop algorithms that expand the action space dynamically. As an initial step, I am implementing a variant of the options framework combined with hindsight experience replay [Lev+19] in the controlled experimental setting of *op. cit.*. The options framework of Markov decision processes stands as a first step in temporal planning, [SB18]. Future approaches could also include various methods of multi-agent reinforcement learning.

References

[AK16] Almog, Mara and Korngreen, Alon. "Is realistic neuronal modeling realistic?" In: *Journal of neurophysiology* 116 5 (2016), pp. 2180–2209. URL: https://api.semanticscholar.org/CorpusID:19117050 (cit. on p. 1).

[BF23] Burns, Thomas F and Fukai, Tomoki. "Simplicial Hopfield networks". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=_QLsH8gatwx (cit. on pp. 1–3).

[Bro+21] Bronstein, Michael M., Bruna, Joan, Cohen, Taco, and Veličković, Petar. *Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges*. 2021. arXiv: 2104.13478 [cs.LG]. URL: https://arxiv.org/abs/2104.13478 (cit. on p. 3).

[BS23] Bredenberg, Colin and Savin, Cristina. *Desiderata for normative models of synaptic plasticity*. 2023. arXiv: 2308.04988 [q-bio.NC]. URL: https://arxiv.org/abs/2308.04988 (cit. on p. 1).

[CDB24] Cabannes, Vivien, Dohmatob, Elvis, and Bietti, Alberto. *Scaling Laws for Associative Memories*. 2024. arXiv: 2310.02984 [stat.ML]. URL: https://arxiv.org/abs/2310.02984 (cit. on p. 3).

[Cur16] Curto, Carina. *What can topology tell us about the neural code?* 2016. arXiv: 1605.01905 [q-bio.NC]. URL: https://arxiv.org/abs/1605.01905 (cit. on p. 3).

[Dem+17] Demircigil, Mete, Heusel, Judith, Löwe, Matthias, Upgang, Sven, and Vermet, Franck. "On a Model of Associative Memory with Huge Storage Capacity". In: 168 (May 2017), pp. 288–299 (cit. on p. 3).

[Gav+24] Gavranović, Bruno, Lessard, Paul, Dudzik, Andrew, Glehn, Tamara von, Araújo, João G. M., and Veličković, Petar. *Position: Categorical Deep Learning is an Algebraic Theory of All Architectures*. 2024. arXiv: 2402.15332 [cs.LG]. URL: https://arxiv.org/abs/2402.15332 (cit. on p. 3).

[HMK14] Hillar, Christopher, Mehta, Ram, and Koepsell, Kilian. "A hopfield recurrent neural network trained on natural images performs state-of-the-art image compression". In: 2014, pp. 4092–4096 (cit. on pp. 1, 2).

[Hoo+23] Hoover, Benjamin, Liang, Yuchen, Pham, Bao, Panda, Rameswar, Strobelt, Hendrik, Chau, Duen Horng, Zaki, Mohammed J., and Krotov, Dmitry. *Energy Transformer*. 2023. arXiv: 2302.07253 [cs.LG]. URL: https://arxiv.org/abs/2302.07253 (cit. on p. 3).

[Hop84] Hopfield, John J. "Neurons with graded response have collective computational properties like those of two-state neurons." In: *Proceedings of the national academy of sciences* 81.10 (1984), pp. 3088–3092 (cit. on p. 1).

[HT14] Hillar, Christopher J. and Tran, Ngoc Mai. "Robust Exponential Memory in Hopfield Networks". In: *Journal of Mathematical Neuroscience* 8 (2014). URL: https://api.semanticscholar.org/CorpusID:11295055 (cit. on p. 3).

[Kem+17] Kemker, Ronald, Abitino, Angelina, McClure, Marc, and Kanan, Christopher. "Measuring Catastrophic Forgetting in Neural Networks". In: *ArXiv* abs/1708.02072 (2017). URL: https://api.semanticscholar.org/CorpusID:22910766 (cit. on p. 3).

[KH16] Krotov, Dmitry and Hopfield, John J. *Dense Associative Memory for Pattern Recognition*. 2016. arXiv: 1606.01164 [cs.NE]. URL: https://arxiv.org/abs/1606.01164 (cit. on pp. 1, 3).

[Koh+22] Kohn, Kathlén, Merkh, Thomas, Montúfar, Guido, and Trager, Matthew. *Geometry of Linear Convolutional Networks*. 2022. arXiv: 2108.01538 [cs.LG]. URL: https://arxiv.org/abs/2108.01538 (cit. on p. 2).

[Lev+19] Levy, Andrew, Konidaris, George, Platt, Robert, and Saenko, Kate. *Learning Multi-Level Hierarchies with Hindsight*. 2019. arXiv: 1712.00948 [cs.AI]. URL: https://arxiv.org/abs/1712.00948 (cit. on p. 4).

[LLL24] Lai, Zehua, Lim, Lek-Heng, and Liu, Yucong. *Attention is a smoothed cubic spline*. 2024. arXiv: 2408.09624 [cs.AI]. URL: https://arxiv.org/abs/2408.09624 (cit. on p. 2).

[Man15] Manin, Yuri I. *Neural codes and homotopy types: mathematical models of place field recognition*. 2015. arXiv: 1501.00897 [math.HO]. URL: https://arxiv.org/abs/1501.00897 (cit. on p. 3).

[MM24] Manin, Yuri and Marcolli, Matilde. “Homotopy Theoretic and Categorical Models of Neural Information Networks”. In: *Compositionality* Volume 6 (2024) (Sept. 2024). ISSN: 2631-4444. URL: http://dx.doi.org/10.46298/compositionality-6-4 (cit. on p. 3).

[Mon+14] Montúfar, Guido, Pascanu, Razvan, Cho, Kyunghyun, and Bengio, Yoshua. *On the Number of Linear Regions of Deep Neural Networks*. 2014. arXiv: 1402.1869 [stat.ML]. URL: https://arxiv.org/abs/1402.1869 (cit. on p. 2).

[ND21] Nichol, Alex and Dhariwal, Prafulla. *Improved Denoising Diffusion Probabilistic Models*. 2021. arXiv: 2102.09672 [cs.LG]. URL: https://arxiv.org/abs/2102.09672 (cit. on p. 3).

[Niu+24] Niu, Xueyan, Bai, Bo, Deng, Lei, and Han, Wei. *Beyond Scaling Laws: Understanding Transformer Performance with Associative Memory*. 2024. arXiv: 2405.08707 [cs.LG]. URL: https://arxiv.org/abs/2405.08707 (cit. on pp. 1, 3).

[Oro+24] Ororbia, Alexander, Mali, Ankur, Kohan, Adam, Millidge, Beren, and Salvatori, Tommaso. *A Review of Neuroscience-Inspired Machine Learning*. 2024. arXiv: 2403.18929 [cs.NE]. URL: https://arxiv.org/abs/2403.18929 (cit. on p. 1).

[Ram+21] Ramsauer, Hubert, Schäfl, Bernhard, Lehner, Johannes, Seidl, Philipp, Widrich, Michael, Gruber, Lukas, Holzleitner, Markus, Adler, Thomas, Kreil, David, Kopp, Michael K, Klambauer, Günter, Brandstetter, Johannes, and Hochreiter, Sepp. “Hopfield Networks is All You Need”. In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=tL89RnzIiCd (cit. on p. 3).

[SB18] Sutton, Richard S and Barto, Andrew G. *Reinforcement learning: An introduction*. MIT press, 2018 (cit. on p. 4).

[SBD20] Sohl-Dickstein, Jascha, Battaglino, Peter, and DeWeese, Michael R. *A new method for parameter estimation in probabilistic models: Minimum probability flow*. 2020. arXiv: 2007.09240 [cs.LG]. URL: https://arxiv.org/abs/2007.09240 (cit. on p. 2).

[She+24] Shehper, Ali, Medina-Mardones, Anibal M., Lewandowski, Bartłomiej, Gruen, Angus, Kucharski, Piotr, and Gukov, Sergei. *What makes math problems hard for reinforcement learning: a case study*. 2024. arXiv: 2408.15332 [cs.LG]. URL: https://arxiv.org/abs/2408.15332 (cit. on p. 4).

[Sta07] Stanley, Richard P. “An Introduction to Hyperplane Arrangements”. In: 2007 (cit. on p. 2).