Long Sequence Hopfield Memory

Hamza Tahir Chaudhry[1,2], **Jacob A. Zavatone-Veth**[2,3],
Dmitry Krotov[5], **Cengiz Pehlevan**[1,2,4]
[1]John A. Paulson School of Engineering and Applied Sciences,
[2]Center for Brain Science, [3]Department of Physics,
[4]Kempner Institute for the Study of Natural and Artificial Intelligence,
Harvard University
Cambridge, MA 02138
[5]MIT-IBM Watson AI Lab, IBM Research,
Cambridge, MA 02142
hchaudhry@g.harvard.edu, jzavatoneveth@g.harvard.edu,
krotov@ibm.com, cpehlevan@seas.harvard.edu

Abstract

Sequence memory is an essential attribute of natural and artificial intelligence that enables agents to encode, store, and retrieve complex sequences of stimuli and actions. Computational models of sequence memory have been proposed where recurrent Hopfield-like neural networks are trained with temporally asymmetric Hebbian rules. However, these networks suffer from limited sequence capacity (maximal length of the stored sequence) due to interference between the memories. Inspired by recent work on Dense Associative Memories, we expand the sequence capacity of these models by introducing a nonlinear interaction term, enhancing separation between the patterns. We derive novel scaling laws for sequence capacity with respect to network size, significantly outperforming existing scaling laws for models based on traditional Hopfield networks, verify these theoretical results with numerical simulation, and demonstrate their usefulness in overlapping patterns. Finally, we describe a biologically-plausible implementation, with connections to motor neuroscience.

1 Introduction

The ability to recall sequences of memories is necessary for a large number of cognitive tasks with temporal or causal structure, including navigation, reasoning, and motor control [1–9]. Computational models have been proposed for how neural networks can encode sequence memory, ranging across a wide range of biological plausibility [10, 1, 3, 11–20, 2, 21]. Many of these are based on the concept of associative memory, where the Hopfield Network (HN) is the canonical model [22–24].

Unfortunately, a major limitation of the traditional Hopfield Network and related associative memory models is its capacity: the number of memories it can store and reliably retrieve scales linearly with the number of neurons in the network. This limitation is due to interference between different memories during recall, also known as crosstalk, which decreases the signal-to-noise ratio and thus the recall of incorrect local minima, which are undesired attractor states of the network commonly referred to as spin glass states [25–28]. Recent modifications of the Hopfield Network, known as Dense Associative Memories (DAMs) or Modern Hopfield Networks (MHNs), overcome this limitation by introducing a strong nonlinearity when computing the overlap between the network state and memories stored in the network [29, 30], leading to greater separation between partially overlapping memories and thereby reducing crosstalk [31].

Associative Memory & Hopfield Networks in 2023. NeurIPS 2023 workshop.

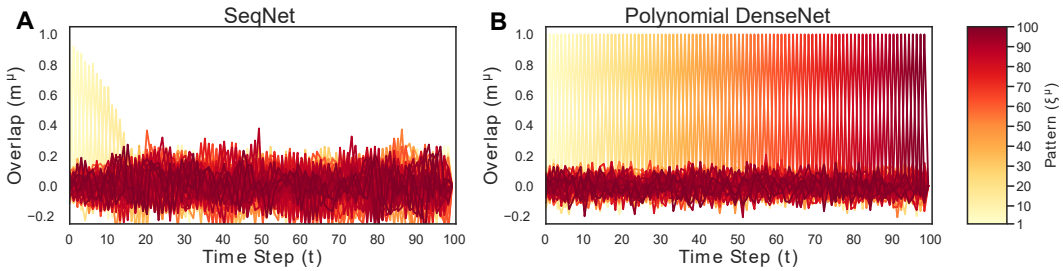

Figure 1: `SeqNet` and Polynomial `DenseNet` ($d = 2$) are simulated with $N = 300$ neurons and $P = 100$ patterns. One hundred curves are plotted as a function of time, each representing the overlap of the network state at time t with one of the patterns, $m^\mu = (1/N)\sum_{i=1}^{N} \xi_i^\mu S_i$. Patterns in the beginning and end of the sequence are shaded in yellow and red respectively). (A) `SeqNet` quickly loses the correct sequence, indicated by the lack of alignment of the network state with the correct pattern in the sequence ($m^\mu \ll 1$). (B) The Polynomial `DenseNet` faithfully recalls the entire sequence and maintains alignment with the correct pattern at any moment in time, $m^\mu \approx 1$.

In order to adapt the HN to store sequences, one must utilize asymmetric weights to drive the network from one memory to the next. Many models use temporally asymmetric Hebbian learning rules to strengthen synaptic connections between neural activity at times t_1 and t_2, thereby learning temporal association between patterns in a sequence [10, 1, 3, 11, 16, 17]. In this paper, we extend DAMs to the setting of asymmetric weights to store and recall long sequences of memories. We find a close match between theory and simulation, establish the ability of this model to store and recall sequences of correlated patterns, and demonstrate the ability to robustly recall highly correlated patterns. Finally, we describe applications of our network as a model of biological motor control.

2 `DenseNets` for Sequence Storage

We provide a high-level overview of the theory. Details, extensions, and comparisons with related models provided in the full conference submission [32]. Assume that we want to store a sequence of P patterns, $\{\boldsymbol{\xi}^1, \dots, \boldsymbol{\xi}^\mu\}$, where $\xi_j^\mu \in \{\pm 1\}$ is the j^{th} neuron of the μ^{th} pattern and the network transitions from pattern $\boldsymbol{\xi}^\mu$ to $\boldsymbol{\xi}^{\mu+1}$. Let N be the size of the network and $\mathbf{S}(t) \in \{-1, +1\}^N$ be the state of the network at time t. We want to design a network with dynamics such that when initialized in pattern ξ^1, it traverses the entire sequence. We define a network, `SeqNet`, which follows a discrete-time synchronous update rule:

$$T_{SN}(\mathbf{S})_i := \operatorname{sgn}\left[\sum_{j \neq i} J_{ij} S_j\right] = \operatorname{sgn}\left[\sum_{\mu=1}^{P} \xi_i^{\mu+1} m_i^\mu\right], \quad m_i^\mu := \frac{1}{(N-1)} \sum_{j \neq i} \xi_j^\mu S_j, \tag{1}$$

where $\mathbf{S}(t+1) = T_{SN}(\mathbf{S})$ and $J_{ij} = \frac{1}{N}\sum_{\mu=1}^{P} \xi_i^{\mu+1}\xi_j^\mu$ is an asymmetric matrix connecting pattern ξ^μ to $\xi^{\mu+1}$. Note that we are excluding self-interaction terms $i = j$. We rewrite the dynamics in terms of m_i^μ, the overlap of the network state $\mathbf{S}$ with pattern $\boldsymbol{\xi}^\mu$. When the network is aligned most closely with pattern ξ^μ, the overlap m_i^μ is the largest contribution in the sum and pushes the network to pattern $\xi^{\mu+1}$. Overlap between patterns reduces the signal-to-noise ratio and thus limits the capacity of the network, resulting in the `SeqNet`'s capacity to scale linearly relative to network size. To overcome the capacity limitations of the `SeqNet`, we define the `DenseNet` update rule:

$$T_{DN}(\mathbf{S})_i := \operatorname{sgn}\left[\sum_{\mu=1}^{P} \xi_i^{\mu+1} f\left(m_i^\mu\right)\right] \tag{2}$$

where f is a nonlinear, monotonically increasing interaction function.

To derive analytical results for the capacity, we must choose a distribution to generate the patterns. As in studies of the capacity of the classic HN [33–36, 25, 27, 26, 28], we choose this to be the Rademacher distribution, where $\xi_j^\mu \in \{-1, +1\}$ with equal probability for all neurons j in all patterns μ, and calculate the capacity for different update rules. We consider both the robustness of a single

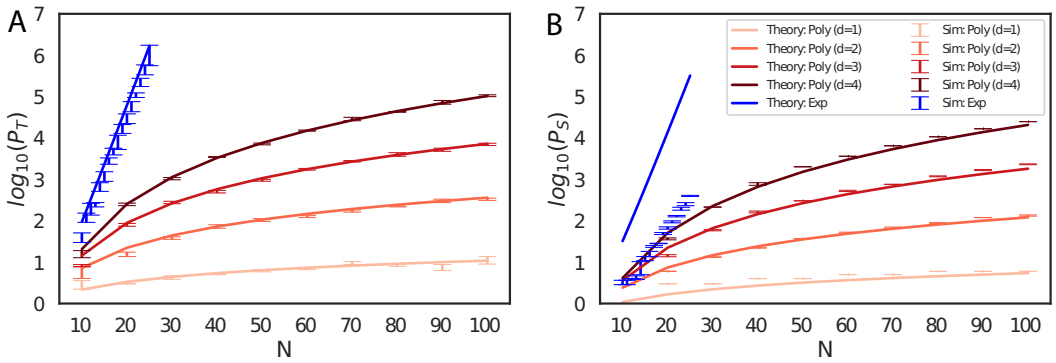

Figure 2: The transition and sequence capacities are tested for `DenseNets` with polynomial and exponential nonlinearities. 100 Sequences of P Rademacher-distributed patterns are generated, the update rule is applied, and the amount of errors is calculated. Smaller sequences are tested until there are no errors. Error bars are calculated by repeating this entire process for 20 different initializations. As network size increases, cross-talk variance decreases and the theory become more accurate, resulting in a tight match between theory (solid lines) and simulation (points with error bars). Transition capacity, $\log_{10}(P_T)$, is plotted on the left. Sequence capacity, $\log_{10}(P_S)$, is plotted on the right. The theory curves are given by Equations 4 and 5.

transition, and the robustness of propagation through the full sequence. Letting $P = P(N)$ such that $\lim_{N\to\infty} P(N) = \infty$, these capacities are defined by the conditions

$$\lim_{N\to\infty} \mathbb{P}\left[\mathbf{T}_{DN}(\boldsymbol{\xi}^\mu) = \boldsymbol{\xi}^{\mu+1}\right] \geq 1 - c, \quad \lim_{N\to\infty} \mathbb{P}\left[\bigcap_{\mu=1}^{P} \{\mathbf{T}_{DN}(\boldsymbol{\xi}^\mu) = \boldsymbol{\xi}^{\mu+1}\}\right] \geq 1 - c, \tag{3}$$

for a fixed constant $c \geq 0$. Note that the full sequence capacity is defined by demanding that all transitions are correct. For perfect recall, we want the threshold $c = 0$. We define the single-transition and full-sequence capacities, respectively, by the asymptotic threshold such that the left and right limit conditions hold for values of $P(N)$ that are asymptotically less than that threshold, and fail for values of $P(N)$ that are greater than the threshold.

2.1 Polynomial `DenseNet`

Consider the `DenseNet` with polynomial interaction function, $f(x) = x^d$, which we will call the Polynomial `DenseNet`. In Appendix A.1, we argue that this network's single-transition capacity scales as P_T while its full-sequence capacity scales as P_S:

$$P_T \sim \frac{N^d}{2(2d-1)!!\log(N)}, \quad P_S \sim \frac{N^d}{2(d+1)(2d-1)!!\log(N)}. \tag{4}$$

Note that the single-transition capacity scaling coincides with that of the symmetric MHN [29].

2.2 Exponential `DenseNet`

Consider the `DenseNet` with exponential interaction function, $f(x) = e^{(N-1)(x-1)}$, which we call the Exponential `DenseNet`. In Appendix A.2, we argue that this network's single-transition capacity scales as P_T while its full-sequence capacity scales as P_S:

$$P_T \sim \frac{\beta^{N-1}}{2\log N}, \quad P_S \sim \frac{\beta^{N-1}}{2\log(\beta)N}; \qquad \beta = \frac{\exp(2)}{\cosh(2)} \simeq 1.964\ldots \tag{5}$$

For smaller N, the cross-talk has non-negligible kurtosis in finite size networks leading to deviation from Gaussian approximation.

Note that these scaling laws were derived under the assumption of i.i.d. Rademacher random patterns. While theoretically convenient, this is unrealistic for real-world data. We test these models in a more realistic setting by storing correlated sequences of patterns. To do so, we concatenate the entire Moving MNIST dataset into a single sequence and simulate its recall with `DenseNet` with varying nonlinearities [37]. The results are shown in Figure 3.

3 Biologically-Plausible Implementation

Since biological neural networks must store sequence memories [5, 2, 6–8], one naturally asks if these results can be generalized to biologically-plausible neural networks. A straightforward biological interpretation of the `DenseNet` is problematic, as a network with polynomial interaction function of degree d is equivalent to having a neural network with many-body synapses between $d+1$ neurons. This is biologically unrealistic as synaptic connections usually occur between two neurons [38].

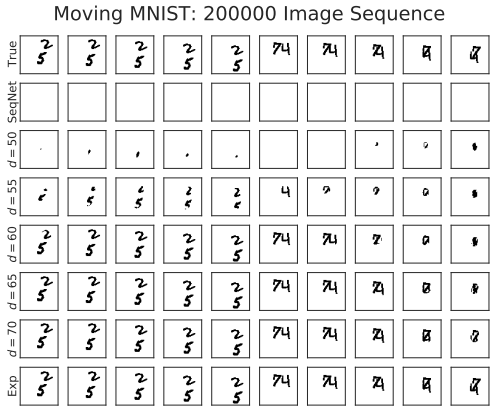

Figure 3: Simulation of correlated patterns using a 200000 image sequence from MovingMNIST.

We take inspiration from earlier work by Krotov and Hopfield [39] who reformulated a symmetric MHN using two-body synapses by partitioning the network into a bipartite graph with visible and hidden neurons [39]. Visible neurons corresponding to the neurons in our network dynamics, $\mathbf{S}_j$, are connected via weight matrix to hidden neurons corresponding to overlap with individual memories stored within the network. Since we asymmetric connections, we must instead define two sets of synaptic weights: $W_{j\mu}$ connects visible neuron v_j to hidden neuron h_μ, $M_{\mu j}$ connects hidden neuron h_μ to visible neuron v_j. This results in the same dynamics as in Equation (2), with the nonlinearity absorbed into the hidden neurons' dynamics.

Finally, we note that this network is reminiscent of recent computational models for motor action selection and control via the cortico-basal ganglia-thalamo-cortical loop, in which the basal ganglia inhibits thalamic neurons that are bidirectionally connected to a recurrent cortical network [40, 5]. This relates to our model as follows: the motor cortex (visible neurons) executes an action, each thalamic unit (hidden neurons) encodes a motor motif, and the basal ganglia silences thalamic neurons (external network modulating context). Thalamo-cortical loops have also been found to be important to song generation in zebra finches [41]. Thus, the biological implementation of the `DenseNet` can provide insight into how biological agents reliably store and generate complex sequences.

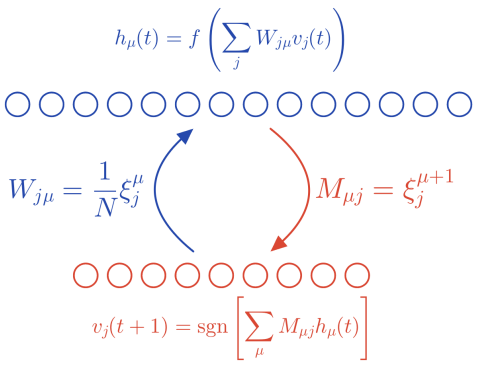

Figure 4: Biologically-plausible implementation of `DenseNet` as a bipartite network.

4 Discussion and Future Directions

We introduced the `DenseNet` for the reliable storage and recall of long sequences of patterns, derived the scaling of its single-transition and full-sequence capacity, and verified these results in numerical simulation. We found that depending on the choice of nonlinear interaction function, the `DenseNet` could scale polynomially or exponentially. For small Exponential `DenseNets`, we see that a large amount kurtosis in the cross-talk distribution, leading to significant deviation between simulation and theoretical results derived in the thermodynamic limit. We also tested the these models' ability to recall sequences of correlated patterns, by comparing the recall of a sequence of Moving MNIST images using `DenseNets` with different nonlinearities. As expected, the networks' reconstruction capabilities increased with degree d and best results were achieved with the exponential nonlinearity.

In this work, we limited ourselves to theoretical analysis of discrete-time networks storing binary patterns. An important direction for future research would be to go beyond the Gaussian theory in order to develop accurate predictions of the Exponential `DenseNet` capacity. There are also many potential avenues for extending these models and methods, including to continuous-time networks, continuous-valued patterns, computing capacity for correlated patterns, testing different weight functions, and examining different network topologies.

Acknowledgments and Disclosure of Funding

We thank Matthew Farrell, Shanshan Qin, and Sabarish Sainathan for useful discussions and comments on earlier versions of our manuscript. HC was supported by the GFSD Fellowship, Harvard GSAS Prize Fellowship, and Harvard James Mills Peirce Fellowship. JAZ-V and CP were supported by NSF Award DMS-2134157 and NSF CAREER Award IIS-2239780. CP received additional support from a Sloan Research Fellowship. This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence. The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

References

[1] D Kleinfeld and H Sompolinsky. Associative neural network model for the generation of temporal patterns. theory and application to central pattern generators. *Biophysical Journal*, 54 (6):1039–1051, 1988.

[2] Michael A. Long, Dezhe Z. Jin, and Michale S. Fee. Support for a synaptic chain model of neuronal sequence generation. *Nature*, 468(7322):394–399, Nov 2010. ISSN 1476-4687. doi:10.1038/nature09514. URL https://doi.org/10.1038/nature09514.

[3] Maxwell Gillett, Ulises Pereira, and Nicolas Brunel. Characteristics of sequential activity in networks with temporally asymmetric Hebbian learning. *Proceedings of the National Academy of Sciences*, 117(47):29948–29958, November 2020. doi:10.1073/pnas.1918674117.

[4] Stefano Recanatesi, Ulises Pereira-Obilinovic, Masayoshi Murakami, Zachary Mainen, and Luca Mazzucato. Metastable attractors explain the variable timing of stable behavioral action sequences. *Neuron*, 110(1):139–153, 2022.

[5] Luca Mazzucato. Neural mechanisms underlying the temporal organization of naturalistic animal behavior. *eLife*, 11:e76577, 2022.

[6] Edmund T Rolls and Patrick Mills. The generation of time in the hippocampal memory system. *Cell Reports*, 28(7):1649–1658, 2019.

[7] Alexander B. Wiltschko, Matthew J. Johnson, Giuliano Iurilli, Ralph E. Peterson, Jesse M. Katon, Stan L. Pashkovski, Victoria E. Abraira, Ryan P. Adams, and Sandeep Robert Datta. Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015. ISSN 0896-6273. doi:https://doi.org/10.1016/j.neuron.2015.11.031. URL https://www.sciencedirect.com/science/article/pii/S0896627315010375.

[8] Jeffrey E. Markowitz, Winthrop F. Gillis, Maya Jay, Jeffrey Wood, Ryley W. Harris, Robert Cieszkowski, Rebecca Scott, David Brann, Dorothy Koveal, Tomasz Kula, Caleb Weinreb, Mohammed Abdal Monium Osman, Sandra Romero Pinto, Naoshige Uchida, Scott W. Linderman, Bernardo L. Sabatini, and Sandeep Robert Datta. Spontaneous behaviour is structured by reinforcement without explicit reward. *Nature*, 614(7946):108–117, Feb 2023. ISSN 1476-4687. doi:10.1038/s41586-022-05611-2. URL https://doi.org/10.1038/s41586-022-05611-2.

[9] Cengiz Pehlevan, Farhan Ali, and Bence P Ölveczky. Flexibility in motor timing constrains the topology and dynamics of pattern generator circuits. *Nature communications*, 9(1):977, 2018. doi:https://doi.org/10.1038/s41467-018-03261-5.

[10] H. Sompolinsky and I. Kanter. Temporal Association in Asymmetric Neural Networks. *Physical Review Letters*, 57(22):2861–2864, December 1986. doi:10.1103/PhysRevLett.57.2861.

[11] Zijian Jiang, Ziming Chen, Tianqi Hou, and Haiping Huang. Spectrum of non-Hermitian deep-Hebbian neural networks. *Physical Review Research*, 5:013090, Feb 2023. doi:10.1103/PhysRevResearch.5.013090. URL https://link.aps.org/doi/10.1103/PhysRevResearch.5.013090.

[12] Ulises Pereira and Nicolas Brunel. Unsupervised learning of persistent and sequential activity. *Frontiers in Computational Neuroscience*, 13:97, 2020.

[13] Christian Leibold and Richard Kempter. Memory capacity for sequences in a recurrent network with biological constraints. *Neural Computation*, 18(4):904–941, 2006.

[14] Jeff Hawkins, Dileep George, and Jamie Niemasik. Sequence memory for prediction, inference and behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364 (1521):1203–1209, 2009.

[15] Jeff Hawkins and Subutai Ahmad. Why neurons have thousands of synapses, a theory of sequence memory in neocortex. *Frontiers in Neural Circuits*, page 23, 2016.

[16] Daniel J Amit. Neural networks counting chimes. *Proceedings of the National Academy of Sciences*, 85(7):2141–2145, 1988.

[17] H. Gutfreund and M. Mezard. Processing of temporal sequences in neural networks. *Phys. Rev. Lett.*, 61:235–238, Jul 1988. doi:10.1103/PhysRevLett.61.235. URL https://link.aps.org/doi/10.1103/PhysRevLett.61.235.

[18] Kanaka Rajan, Christopher D. Harvey, and David W. Tank. Recurrent network models of sequence generation and memory. *Neuron*, 90(1):128–142, 2016. ISSN 0896-6273. doi:https://doi.org/10.1016/j.neuron.2016.02.009. URL https://www.sciencedirect.com/science/article/pii/S0896627316001021.

[19] Markus Diesmann, Marc-Oliver Gewaltig, and Ad Aertsen. Stable propagation of synchronous spiking in cortical neural networks. *Nature*, 402(6761):529–533, Dec 1999. ISSN 1476-4687. doi:10.1038/990101. URL https://doi.org/10.1038/990101.

[20] Nicholas F Hardy and Dean V Buonomano. Neurocomputational models of interval and pattern timing. *Current Opinion in Behavioral Sciences*, 8:250–257, 2016. ISSN 2352-1546. doi:https://doi.org/10.1016/j.cobeha.2016.01.012. URL https://www.sciencedirect.com/science/article/pii/S2352154616300195. Time in perception and action.

[21] Dina Obeid, Jacob A. Zavatone-Veth, and Cengiz Pehlevan. Statistical structure of the trial-to-trial timing variability in synfire chains. *Phys. Rev. E*, 102:052406, Nov 2020. doi:10.1103/PhysRevE.102.052406. URL https://link.aps.org/doi/10.1103/PhysRevE.102.052406.

[22] S-I Amari. Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions on computers*, 100(11):1197–1206, 1972.

[23] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[24] John J Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092, 1984.

[25] John Hertz, Anders Krogh, and Richard G Palmer. *Introduction to the theory of neural computation*. CRC Press, 2018.

[26] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007, 1985.

[27] Daniel J. Amit, Hanoch Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55:1530–1533, Sep 1985. doi:10.1103/PhysRevLett.55.1530. URL https://link.aps.org/doi/10.1103/PhysRevLett.55.1530.

[28] Daniel J Amit, Hanoch Gutfreund, and H Sompolinsky. Statistical mechanics of neural networks near saturation. *Annals of Physics*, 173(1):30–67, 1987. ISSN 0003-4916. doi:https://doi.org/10.1016/0003-4916(87)90092-3. URL https://www.sciencedirect.com/science/article/pii/0003491687900923.

[29] Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. *Advances in Neural Information Processing Systems*, 29, 2016.

[30] Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2): 288–299, July 2017. ISSN 0022-4715, 1572-9613. doi:10.1007/s10955-017-1806-y.

[31] Dmitry Krotov. A new frontier for hopfield networks. *Nature Reviews Physics*, pages 1–2, 2023.

[32] Hamza Tahir Chaudhry, Jacob A Zavatone-Veth, Dmitry Krotov, and Cengiz Pehlevan. Long sequence hopfield memory. *arXiv preprint arXiv:2306.04532*, 2023.

[33] Dimitri Petritis. Thermodynamic formalism of neural computing. In Eric Goles and Servet Martínez, editors, *Dynamics of Complex Interacting Systems*, pages 81–146. Springer Netherlands, Dordrecht, 1996. doi:10.1007/978-94-017-1323-8_3. URL https://doi.org/10.1007/978-94-017-1323-8_3.

[34] Anton Bovier. Sharp upper bounds on perfect retrieval in the Hopfield model. *Journal of Applied Probability*, 36(3):941–950, 1999. doi:10.1239/jap/1032374647.

[35] R. McEliece, E. Posner, E. Rodemich, and S. Venkatesh. The capacity of the Hopfield associative memory. *IEEE Transactions on Information Theory*, 33(4):461–482, July 1987. ISSN 0018-9448. doi:10.1109/TIT.1987.1057328.

[36] G. Weisbuch and F. Fogelman-Soulié. Scaling laws for the attractors of Hopfield networks. *J. Physique Lett.*, 46(14):623–630, 1985. doi:10.1051/jphyslet:019850046014062300. URL https://doi.org/10.1051/jphyslet:019850046014062300.

[37] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 843–852, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/srivastava15.html.

[38] Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, A James Hudspeth, Sarah Mack, et al. *Principles of neural science*. McGraw-hill New York, 6 edition, 2021.

[39] Dmitry Krotov and John J. Hopfield. Large associative memory problem in neurobiology and machine learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=X4y_10OX-hX.

[40] Laureline Logiaco, LF Abbott, and Sean Escola. Thalamic control of cortical dynamics in a model of flexible motor sequencing. *Cell Reports*, 35(9):109090, 2021.

[41] Felix W. Moll, Devorah Kranz, Ariadna Corredera Asensio, Margot Elmaleh, Lyn A. Ackert-Smith, and Michael A. Long. Thalamus drives vocal onsets in the zebra finch courtship song. *Nature*, 616(7955):132–136, Apr 2023. ISSN 1476-4687. doi:10.1038/s41586-023-05818-x. URL https://doi.org/10.1038/s41586-023-05818-x.

[42] John E. Kolassa. *Series Approximation Methods in Statistics*. Springer New York, 1997. doi:10.1007/978-1-4757-4277-0. URL https://doi.org/10.1007/978-1-4757-4277-0.

A `DenseNet` Capacity

In this Appendix, we analyze the capacity of the `DenseNet`. As introduced in Section 2 of the main text, there are two notions of robustness to consider: the robustness of a single transition and the robustness of the full sequence, which we determine based on the conditions

$$\lim_{N,P\to\infty} \mathbb{P}\left[\mathbf{T}_{DN}(\boldsymbol{\xi}^\mu) = \boldsymbol{\xi}^{\mu+1}\right] \geq 1 - c \tag{A.1}$$

and

$$\lim_{N,P\to\infty} \mathbb{P}\left[\bigcap_{\mu=1}^{P} \{\mathbf{T}_{DN}(\boldsymbol{\xi}^\mu) = \boldsymbol{\xi}^{\mu+1}\}\right] \geq 1 - c, \tag{A.2}$$

respectively, for a fixed constant $c \geq 0$.

Following Petritis [33]'s approach to the HN, to make analytical progress, we can use a union bound to control the single-step error probability in terms of the probability of a single bitflip:

$$\begin{aligned}
&\mathbb{P}\left[\mathbf{T}_{DN}(\boldsymbol{\xi}^\mu) = \boldsymbol{\xi}^{\mu+1}\right] \\
&= 1 - \mathbb{P}\left[\bigcup_{i=1}^{N} \{T_{DN}(\boldsymbol{\xi}^\mu)_i \neq \xi_i^{\mu+1}\}\right] && \text{(A.3)} \\
&\geq 1 - \sum_{i=1}^{N} \mathbb{P}\left[T_{DN}(\boldsymbol{\xi}^\mu)_i \neq \xi_i^{\mu+1}\right] && \text{(A.4)} \\
&= 1 - N\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_2^1]. && \text{(A.5)}
\end{aligned}$$

where we use the fact that all elements of all patterns are i.i.d. by assumption. We use a similar approach to control the sequence error probability in terms of the probability of a single bitflip:

$$\begin{aligned}
&\mathbb{P}\left[\bigcap_{\mu=1}^{P} \{\mathbf{T}_{DN}(\boldsymbol{\xi}^\mu) = \boldsymbol{\xi}^{\mu+1}\}\right] \\
&= 1 - \mathbb{P}\left[\bigcup_{\mu=1}^{P}\bigcup_{i=1}^{N} \{T_{DN}(\boldsymbol{\xi}^\mu)_i \neq \xi_i^{\mu+1}\}\right] && \text{(A.6)} \\
&\geq 1 - \sum_{\mu=1}^{P}\sum_{i=1}^{N} \mathbb{P}\left[T_{DN}(\boldsymbol{\xi}^\mu)_i \neq \xi_i^{\mu+1}\right] && \text{(A.7)} \\
&= 1 - NP\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_2^1]. && \text{(A.8)}
\end{aligned}$$

Thus, if

$$\lim_{N,P\to\infty} N\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_1^2] \leq c \implies \lim_{N,P\to\infty} \mathbb{P}[\mathbf{T}_{DN}(\boldsymbol{\xi}^\mu) = \boldsymbol{\xi}^{\mu+1}] \geq 1 - c, \tag{A.9}$$

while the stronger condition guarantees

$$\lim_{N,P\to\infty} NP\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_1^2] \leq c \implies \lim_{N,P\to\infty} \mathbb{P}[\cup_{\mu=1}^{P}\{\mathbf{T}_{DN}(\boldsymbol{\xi}^\mu) = \boldsymbol{\xi}^{\mu+1}\}] \geq 1 - c. \tag{A.10}$$

As introduced in the main text, for perfect recall, we want to take the threshold $c = 0$. If the condition (A.9) holds for all $P = P(N)$ such that $\lim_{N\to\infty} P(N)/P_T(N) \leq 1$ and fails for all $P(N)$ such that $\lim_{N\to\infty} P(N)/P_T(N) > 1$, we say that $P_T = P_T(N)$ is the single-transition capacity of the network. Similarly, if (A.10) holds for all $P = P(N)$ such that $\lim_{N\to\infty} P(N)/P_S(N) \leq 1$ and fails for all $P(N)$ such that $\lim_{N\to\infty} P(N)/P_S(N) > 1$, we say that $P_S = P_S(N)$ is the full sequence capacity of the network. The capacities estimated through this argument are lower bounds on the true capacities, as they are obtained from lower bounds on the true recall probability. However, we expect for these bounds to in fact be tight in the thermodynamic limit [33, 34].

By the definition of the `DenseNet` update rule with interaction function f given in Equation (2), we have

$$T_{DN}(\boldsymbol{\xi}^1)_1 = \operatorname{sgn}\left[\sum_{\mu=1}^{P} \xi_1^{\mu+1} f\left(\frac{1}{N-1}\sum_{j=2}^{N} \xi_j^\mu \xi_j^1\right)\right] \tag{A.11}$$

and therefore the single-bitflip probability is

$$\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_1^2] = \mathbb{P}\left[\operatorname{sgn}\left[\sum_{\mu=1}^{P} \xi_1^{\mu+1} f\left(\frac{1}{N-1}\sum_{j=2}^{N} \xi_j^\mu \xi_j^1\right)\right] \neq \xi_1^2\right] \tag{A.12}$$

$$= \mathbb{P}\left[\xi_1^2 \sum_{\mu=1}^{P} \xi_1^{\mu+1} f\left(\frac{1}{N-1}\sum_{j=2}^{N} \xi_j^\mu \xi_j^1\right) < 0\right] \tag{A.13}$$

$$= \mathbb{P}\left[f(1) + \xi_1^2 \sum_{\mu=2}^{P} \xi_1^{\mu+1} f\left(\frac{1}{N-1}\sum_{j=2}^{N} \xi_j^\mu \xi_j^1\right) < 0\right] \tag{A.14}$$

For both the polynomial ($f(x) = x^d$) and exponential ($f(x) = e^{(N-1)(x-1)}$) interaction functions, $f(1) = 1$, and so

$$\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_1^2] = \mathbb{P}\left[\sum_{\mu=2}^{P} \xi_1^2 \xi_1^{\mu+1} f\left(\frac{1}{N-1}\sum_{j=2}^{N} \xi_j^\mu \xi_j^1\right) < -1\right]. \tag{A.15}$$

We refer to the random variable

$$C = \sum_{\mu=2}^{P} \xi_1^2 \xi_1^{\mu+1} f\left(\frac{1}{N-1}\sum_{j=2}^{N} \xi_j^\mu \xi_j^1\right) \tag{A.16}$$

on the left-hand-side of this inequality as the *crosstalk*, because it represents the effect of interference between the first pattern and all other patterns [25, 36].

We now observe that, as we have excluded self-interactions (i.e., the sum over neurons inside the interaction function does not include $j = 1$), we can use the periodic boundary conditions to shift indices as $\xi_1^\mu \leftarrow \xi_1^{\mu+1}$ for all μ, yielding

$$C \overset{d}{=} \sum_{\mu=2}^{P} \xi_1^1 \xi_1^\mu f\left(\frac{1}{N-1}\sum_{j=2}^{N} \xi_j^\mu \xi_j^1\right) \tag{A.17}$$

Thus, the single-bitflip probability for this `DenseNet` is identical to that for the corresponding MHN with symmetric interactions. Then, we can use the fact that $\xi_j^\mu \xi_j^1 \overset{d}{=} \xi_j^\mu$ for all $\mu = 2, \ldots, P$ to obtain

$$C \overset{d}{=} \sum_{\mu=2}^{P} \xi_1^\mu f\left(\frac{1}{N-1}\sum_{j=2}^{N} \xi_j^\mu\right) < -1. \tag{A.18}$$

Now, define the $P - 1$ random variables

$$\chi^\mu = \xi_1^\mu f\left(\frac{1}{N-1}\sum_{j=2}^{N} \xi_j^\mu\right) \tag{A.19}$$

for $\mu = 2, \ldots, P$, such that the crosstalk is their sum,

$$C = \sum_{\mu=2}^{P} \chi^\mu. \tag{A.20}$$

As the patterns ξ_j^μ are i.i.d., χ^μ are i.i.d. random variables of mean

$$\mathbb{E}[\chi^\mu] = \mathbb{E}[\xi_1^\mu]\mathbb{E}\left[f\left(\frac{1}{N-1}\sum_{j=2}^{N} \xi_j^\mu\right)\right] = 0 \tag{A.21}$$

and variance

$$\mathrm{var}(\chi^\mu) = \mathbb{E}\left[f\left(\frac{1}{N-1} \sum_{j=2}^{N} \xi_j^\mu \right)^2 \right], \tag{A.22}$$

which is bounded from above for any sensible interaction function. We observe also that the distribution of each χ^μ is symmetric because of the symmetry of the distribution of ξ_1^μ. We will therefore simply write χ for any given χ^μ.

Then, the classical central limit theorem implies that the crosstalk tends in distribution to a Gaussian of mean zero and variance $(P-1)\,\mathrm{var}(\chi)$ as $P \to \infty$, at lease for any fixed N. However, we are interested in the joint limit in which $N, P \to \infty$ together. We will proceed by approximating the distribution of C as Gaussian, and will not attempt to rigorously control its behavior in the joint limit.

Approximating the distribution of the crosstalk for $N, P \gg 1$ by a Gaussian, we then have

$$\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_1^2] \approx H\left(\frac{1}{\sqrt{(P-1)\,\mathrm{var}(\chi)}} \right) \tag{A.23}$$

where $H(x) = \mathrm{erfc}(x/\sqrt{2})/2$ is the Gaussian tail distribution function. We want to have $\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_1^2] \to 0$, so we must have $(P-1)\,\mathrm{var}(\chi) \to 0$. Then, we can use the asymptotic expansion [25]

$$H(\sqrt{z}) = \frac{1}{\sqrt{2\pi z}} \exp\left(-\frac{z}{2}\right) \left[1 + \mathcal{O}\left(\frac{1}{z}\right) \right] \quad \text{as} \quad z \to \infty \tag{A.24}$$

to obtain

$$\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_1^2] \approx \sqrt{\frac{(P-1)\,\mathrm{var}(\chi)}{2\pi}} \exp\left(-\frac{1}{2(P-1)\,\mathrm{var}(\chi)} \right). \tag{A.25}$$

For each model, we can evaluate $\mathrm{var}(\chi)$ and then determine the resulting predicted capacity.

Our first check on the accuracy of the Gaussian approximation will be comparison of the resulting predictions for capacity with numerical experiment. As another diagnostic, we will consider the excess kurtosis $\varkappa = \kappa_4(C)/\kappa_2(C)$ for $\kappa_n(C)$ the n-th cumulant of C. If the distribution is indeed Gaussian, the excess kurtosis vanishes, while large values of the excess kurtosis indicate deviations from Gaussianity [42]. By the additivity of cumulants, we have

$$\kappa_n(C) = (P-1)\kappa_n(\chi). \tag{A.26}$$

By symmetry, all odd cumulants of χ—and therefore all odd cumulants of C—are identically zero. As noted above, we have

$$\mathrm{var}(\chi) = \kappa_2(\chi) = \mathbb{E}\left[f\left(\frac{1}{N-1} \sum_{j=2}^{N} \xi_j^\mu \right)^2 \right]. \tag{A.27}$$

If C is indeed Gaussian, then all cumulants above the second should vanish. As the third cumulant vanishes by symmetry, the leading possible correction to Gaussianity is the fourth cumulant, which as χ has zero mean is given by

$$\kappa_4(\chi) = \mathbb{E}[(\chi)^4] - 3\mathbb{E}[(\chi)^2] \tag{A.28}$$

$$= \mathbb{E}\left[f\left(\frac{1}{N-1} \sum_{j=2}^{N} \xi_j^\mu \right)^4 \right] - 3\mathbb{E}\left[f\left(\frac{1}{N-1} \sum_{j=2}^{N} \xi_j^\mu \right)^2 \right]^2. \tag{A.29}$$

Rather than considering the fourth cumulant directly, we will consider the excess kurtosis

$$\varkappa = \frac{\kappa_4(C)}{\kappa_2(C)^2} = \frac{1}{P-1} \frac{\kappa_4(\chi)}{\kappa_2(\chi)^2}, \tag{A.30}$$

which is a more useful metric because it is normalized.

A.1 Polynomial `DenseNet` Capacity

We first consider the Polynomial `DenseNet`, with interaction function $f(x) = x^d$ for $d \in \mathbb{N}_{>0}$. To compute the capacity, our goal is then to evaluate

$$\mathrm{var}(\chi) = \mathbb{E}\left[\left(\frac{1}{N-1}\sum_{j=2}^{N}\xi_j^1\right)^{2d}\right] \tag{A.31}$$

at large N. From the central limit theorem, we expect

$$\mathbb{E}\left[\left(\frac{1}{N-1}\sum_{j=2}^{N}\xi_j^1\right)^{2d}\right] \sim \frac{(2d-1)!!}{(N-1)^d}. \tag{A.32}$$

We can make this quantitatively precise through the following straightforward argument. Let

$$\Xi = \frac{1}{\sqrt{N-1}}\sum_{j=2}^{N}\xi_j^2. \tag{A.33}$$

We then have immediately that the moment generating function of Ξ is

$$M(t) = \mathbb{E}[e^{t\Xi}] = \cosh\left(\frac{t}{\sqrt{N-1}}\right)^{N-1}, \tag{A.34}$$

hence the cumulant generating function is

$$K(t) = \log M(t) = (N-1)\log\cosh\left(\frac{t}{\sqrt{N-1}}\right). \tag{A.35}$$

The function $x \mapsto \log\cosh(x)$ is an even function of x, and is analytic near the origin, with the first few orders of its MacLaurin series being

$$\log\cosh(x) = \frac{x^2}{2} - \frac{x^4}{12} + \mathcal{O}(x^6). \tag{A.36}$$

Then, the odd cumulants of Ξ vanish—as we expect from symmetry—while the even cumulants obey

$$\kappa_{2k} = \frac{C_{2k}}{(N-1)^{k-1}} \tag{A.37}$$

for combinatorial factors C_{2k} that do not scale with N. We have, in particular, $C_2 = 1$ and $C_4 = -2$. By the moments-cumulants formula, we have

$$\mathbb{E}[\Xi^{2k}] = B_{2k}(0, \kappa_2, 0, \kappa_4, \cdots, \kappa_{2k}) \tag{A.38}$$

for B_{2k} the $2k$-th complete exponential Bell polynomial. From this, it follows that

$$\mathbb{E}[\Xi^{2k}] = (2k-1)!! + \mathcal{O}(N^{-1}), \tag{A.39}$$

as all cumulants other than $\kappa_2 = 1$ are $\mathcal{O}(N^{-1})$. Therefore, neglecting subleading terms, we have

$$\mathrm{var}(\chi) = \mathbb{E}\left[\left(\frac{1}{N-1}\sum_{j=2}^{N}\xi_j^1\right)^{2d}\right] = \frac{(2d-1)!!}{N^d}\left[1 + \mathcal{O}\left(\frac{1}{N}\right)\right]. \tag{A.40}$$

Following the general arguments above, we then approximate

$$\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_1^2] \sim \sqrt{\frac{P(2d-1)!!}{2\pi N^d}}\exp\left(-\frac{N^d}{2P(2d-1)!!}\right). \tag{A.41}$$

To determine the single-transition capacity following the argument in Section 2, we must determine how large we can take $P = P(N)$ such that $N\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_1^2] \to 0$. Following the requirement that $P\,\mathrm{var}(\chi) \to 0$, we make the *Ansatz*

$$P \sim \frac{N^d}{\alpha(2d-1)!!\,\log N} \tag{A.42}$$

for some α. We then have

$$N\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_1^2] \sim \sqrt{\frac{1}{2\pi\alpha\log N}}N^{1-\alpha/2}. \tag{A.43}$$

This tends to zero if $\alpha \geq 2$, meaning that the predicted capacity in this case is

$$P_T \sim \frac{N^d}{2(2d-1)!!\,\log N}. \tag{A.44}$$

We now want to determine the sequence capacity, which requires the stronger condition $NP\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_1^2] \to 0$. Again making the *Ansatz*

$$P \sim \frac{N^d}{\alpha(2d-1)!!\,\log N} \tag{A.45}$$

for some α, we then have

$$NP\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_1^2] \sim \frac{1}{\sqrt{2\pi}(2d-1)!!\,(\alpha\log N)^{3/2}}N^{d+1-\alpha/2}, \tag{A.46}$$

which tends to zero if $\alpha \geq 2d+2$. Then, the predicted sequence capacity is

$$P_S \sim \frac{N^d}{2(d+1)(2d-1)!!\,\log N}. \tag{A.47}$$

Using the Gaussian approximation for moments of χ given above, we can easily work out that

$$\kappa_4(\chi) = \mathbb{E}[(\chi)^4] - 3\mathbb{E}[(\chi)^2] \tag{A.48}$$

$$= \mathbb{E}\left[f\left(\frac{1}{N-1}\sum_{j=2}^{N}\xi_j^\mu\right)^4\right] - 3\mathbb{E}\left[f\left(\frac{1}{N-1}\sum_{j=2}^{N}\xi_j^\mu\right)^2\right]^2 \tag{A.49}$$

$$= \frac{1}{N^{2d}}\{(4d-1)!! - 3[(2d-1)!!]^2\}\left[1+\mathcal{O}\left(\frac{1}{N}\right)\right]. \tag{A.50}$$

Then, the excess kurtosis of the Polynomial `DenseNet`'s crosstalk is

$$\varkappa = \frac{1}{P-1}\left[\frac{(4d-1)!!}{[(2d-1)!!]^2} - 3\right]\left[1+\mathcal{O}\left(\frac{1}{N}\right)\right]. \tag{A.51}$$

Thus, for the Polynomial `DenseNet`, we expect the excess kurtosis to be small for any fixed d so long as P and N are both fairly large, without any particular requirement on their relationship. In particular, under the Gaussian approximation we predicted above that the transition and sequence capacities should both scale as

$$P \sim \frac{N^d}{\alpha_d\log N}, \tag{A.52}$$

where α_d depends on d but not on N. This gives an excess kurtosis of

$$\varkappa = \frac{\alpha_d\log N}{N^d}\left[\frac{(4d-1)!!}{[(2d-1)!!]^2} - 3\right]\left[1+\mathcal{O}\left(\frac{1}{N}\right)\right] \tag{A.53}$$

which for any fixed d rapidly tends to zero with increasing N. This suggests that the Gaussian approximation should be reasonably accurate even at modest N, but of course does not constitute a proof of its accuracy because we have not considered higher cumulants. However, this matches the results of numerical simulations shown in Figure 2.

A.2 Exponential `DenseNet` capacity

We now turn our attention to the Exponential `DenseNet`, with separation function $f(x) = e^{(N-1)(x-1)}$. In this case, we have

$$\mathrm{var}(\chi) = \exp[-2(N-1)]\mathbb{E}\left[\exp\left(2\sum_{j=2}^{N}\xi_j^2\right)\right] \tag{A.54}$$

$$= \exp[-2(N-1)]\prod_{j=2}^{N}\mathbb{E}\left[\exp\left(2\xi_j^2\right)\right] \tag{A.55}$$

$$= \exp[-2(N-1)]\cosh(2)^{N-1} \tag{A.56}$$

$$= \frac{1}{\beta^{N-1}}, \tag{A.57}$$

where we have defined the constant

$$\beta = \frac{\exp(2)}{\cosh(2)} \simeq 1.96403. \tag{A.58}$$

Then, we have the Gaussian approximation

$$\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_1^2] \sim \sqrt{\frac{P}{2\pi\beta^{N-1}}}\exp\left(-\frac{\beta^{N-1}}{2P}\right). \tag{A.59}$$

As in the polynomial case, we first determine the single-transition capacity by demanding that $N\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_1^2] \to 0$. We plug in the *Ansatz*

$$P \sim \frac{\beta^{N-1}}{\alpha\log N} \tag{A.60}$$

for some α, which yields

$$N\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_1^2] \sim \sqrt{\frac{1}{2\pi\alpha\log N}}N^{1-\alpha/2}. \tag{A.61}$$

This tends to zero if $\alpha \geq 2$, which gives a predicted capacity of

$$P_T \sim \frac{\beta^{N-1}}{2\log N}. \tag{A.62}$$

Considering the sequence capacity, which again requires that $NP\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_1^2] \to 0$, we plug in the *Ansatz*

$$P \sim \frac{\beta^{N-1}}{\alpha N}, \tag{A.63}$$

which yields

$$NP\mathbb{P}[T_{DN}(\boldsymbol{\xi}^1)_1 \neq \xi_1^2] \sim \frac{1}{\alpha\beta}\sqrt{\frac{1}{2\pi\alpha N}}\exp\left[\left(\log\beta - \frac{\alpha}{2}\right)N\right]. \tag{A.64}$$

This tends to zero for $\alpha \geq 2\log\beta$, meaning that the predicted capacity is in this case

$$P_S \sim \frac{\beta^{N-1}}{2\log(\beta)N}. \tag{A.65}$$

Therefore, while the ratio of the predicted single-transition to sequence capacities is finite for the Polynomial `DenseNet`—it is simply $P_S/P_T \sim d+1$—for the Exponential `DenseNet` it tends to zero as $P_S/P_T \sim \log N/[\log(\beta)N]$.

Now considering the fourth cumulant, we can easily compute

$$\kappa_4(\chi) = \left(\frac{\cosh(4)}{\exp(4)}\right)^{N-1} - 3\left(\frac{\cosh(2)^2}{\exp(4)}\right)^{N-1}, \tag{A.66}$$

which yields an excess kurtosis of

$$\varkappa = \frac{1}{P-1}\left[\left(\frac{\cosh(4)}{\cosh(2)^2}\right)^{N-1} - 3\right]. \tag{A.67}$$

For this to be small, P must be exponentially large in N, which contrasts with the situation for the Polynomial `DenseNet`, in which the excess kurtosis is small for any reasonably large P. If we consider taking

$$P \sim \frac{\beta^{N-1}}{\alpha \log N}, \tag{A.68}$$

for a constant α, as the Gaussian theory predicts for the Exponential `DenseNet` transition capacity, we have

$$\varkappa \sim \frac{\alpha \log N}{\beta^{N-1}}\left[\left(\frac{\cosh(4)}{\cosh(2)^2}\right)^{N-1} - 3\right] \tag{A.69}$$

$$\sim \alpha \log N \left(\frac{\cosh(4)}{\exp(2)\cosh(2)}\right)^{N-1} \tag{A.70}$$

$$\simeq \alpha \log(N)(0.9823)^{N-1}. \tag{A.71}$$

This tends to zero as N increases, but only very slowly. In particular, $\log(N)(0.9823)^{N-1}$ increases with N up to around $N \simeq 19$, where it attains a maximum value around 2, before decreasing towards zero. The situation is even worse for the sequence capacity, for which the Gaussian theory predicts

$$P \sim \frac{\beta^{N-1}}{\alpha N}, \tag{A.72}$$

yielding

$$\varkappa \sim \frac{\alpha N}{\beta^{N-1}}\left[\left(\frac{\cosh(4)}{\cosh(2)^2}\right)^{N-1} - 3\right] \tag{A.73}$$

$$\sim \alpha N \left(\frac{\cosh(4)}{\exp(2)\cosh(2)}\right)^{N-1} \tag{A.74}$$

$$\simeq \alpha N(0.9823)^{N-1}. \tag{A.75}$$

$N(0.9823)^{N-1}$ increases with N up to around $N \simeq 56$, where it attains a value of approximately 21.

Taken together, these results suggest that we might expect substantial finite-size corrections to the Gaussian theory's prediction for the capacity. In particular, as the excess kurtosis of the crosstalk is positive, the tails of the crosstalk distribution should be heavier-than-Gaussian, suggesting that the Gaussian theory should overestimate the true capacity. This holds provided that the lower bound on the memorization probability resulting from the union bound is reasonably tight.