THE CAPACITY OF MODERN HOPFIELD NETWORKS UNDER THE DATA MANIFOLD HYPOTHESIS

Beatrice Achilli[1] **Luca Ambrogioni**[2] **Carlo Lucibello**[1] **Marc Mézard**[1]

Enrico Ventura[1]

[1]Department of Computing Sciences, BIDSA, Bocconi University, Milan, MI 20100, Italy.
[2]Donders Institute for Brain, Cognition and Behaviour, Radboud University,
6500 HD Nijmegen, The Netherlands.

ABSTRACT

We generalize the computation of the capacity of exponential Hopfield model from Lucibello and Mézard (2024) to more generic pattern ensembles, including binary patterns and patterns generated from a hidden manifold model.

1 INTRODUCTION

Since its introduction in 1982, the Hopfield Model (Hopfield, 1982) has been the standard model for associative memory. The retrieval of a full memory from partial information was shown to be possible for this model up to a critical capacity, which scales linearly with the system size (Amit et al., 1985). Hopfield-like models with super-linear capacity have been investigated for a long time Gardner (1987), and they achieved renewed popularity in recent years. In particular, Krotov and Hopfield (2016) discuss generalizations with polynomial capacity, while Demircigil et al. (2017) and Ramsauer et al. (2020b) introduce exponential capacity networks. In particular, these exponential models have attracted substantial interest in the past few years, partially due to their connections with transformer attention blocks (Ramsauer et al., 2020b; Hoover et al., 2024) and generative diffusion models (Ambrogioni, 2023; Hoover et al., 2023; Ambrogioni, 2024; Biroli et al., 2024). We will refer to the exponential capacity model of Ramsauer et al. (2020b) as to Modern Hopfield Network. In Ref. (Lucibello and Mézard, 2024), the authors proposed the use of an auxiliary Random Energy Model (REM) to study the retrieval transition in MHNs with a signal-to-noise argument. In this paper, we further develop this approach and extend it to more general pattern ensembles, beyond the rotationally invariant ones they considered. In particular, we study structured patterns, generated according to the Hidden Manifold Model (Goldt et al., 2020) with non-linear activations. We use the replica formalism to study the auxiliary REM associated to the problem and characterize the critical retrieval threshold as a function of the latent dimensionality of the patterns.

2 THE RANDOM ENERGY MODEL FORMALISM

MHNs are analogous to physical systems with randomly sampled energy levels, where each level corresponds to a pattern to be memorized (Lucibello and Mézard, 2024). In statistical physics, the Random Energy Model is a well-known analytically tractable disordered system whose thermodynamic limit can be studied both either large deviation theory or the replica method (Derrida, 1981; Mezard and Montanari, 2009; Fedrigo et al., 2007). In this section, we summarize the probabilistic tools needed to solve a generic REM. Let us consider $P = e^{\alpha N}$ i.i.d. energy levels $\varepsilon^\mu \sim p(\varepsilon \mid \omega)$, where α is a positive real number and we extend the typical REM setting allowing for a common source of quenched disorder $\omega \sim p_\omega$. In machine learning terms ω represents a common set of latent variables that influence the sampling of the observable patterns, or data points.

We are interested in the thermodynamic properties of the system for a large number of energy levels, that is the large N and P limit. These properties can be obtained from the asymptotic free energy of

the system at an inverse temperature λ, defined by

$$\phi_\alpha(\lambda) = \lim_{N\to\infty} \frac{1}{\lambda N} \mathbb{E} \log \sum_\mu e^{\lambda N\, \varepsilon^\mu}. \tag{1}$$

Notice that we adopt a convention where the energy has the opposite sign with respect to the usual physics convention. In order to study the thermodynamic limit, it is convenient to introduce the cumulant generating function and its Legendre transform:

$$\zeta(\lambda) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_\omega \log \mathbb{E}_{\varepsilon\,|\omega}\, e^{\lambda N\, \varepsilon}, \tag{2}$$

$$s(\varepsilon) = \sup_\lambda\ \varepsilon\,\lambda - \zeta(\lambda). \tag{3}$$

The generating function $\zeta(\lambda)$ is more tractable than $\phi_\alpha(\lambda)$ since the expectation with respect to ε is moved inside the logarithm, where it can often be computed explicitly.

The quantity $\Sigma_\alpha(\epsilon) = \alpha - s(\varepsilon)$ corresponds to the annealed entropy of the system (Mezard and Montanari, 2009). The annealed entropy corresponds to the correct (quenched) one, as long as it is positive. Let us define the condensation energy $\varepsilon_*(\alpha)$ as the largest root of $\Sigma_\alpha(\varepsilon_*) = 0$. We have $\Sigma(\varepsilon) < 0$ for $\varepsilon > \varepsilon^*(\alpha)$. Let us also call $\tilde{\varepsilon}(\lambda) = \zeta'(\lambda)$ the stationary energy.

For given λ and α, if $\Sigma(\tilde{\varepsilon}(\lambda)) = \alpha - s(\tilde{\varepsilon}(\lambda)) = \alpha + \zeta(\lambda) - \lambda\zeta'(\lambda)$ is larger than zero, the annealed description is correct, and an exponential number of energy levels with typical value $\tilde{\varepsilon}(\lambda)$ dominate the partition function. This is called the *uncondensed phase*.

Increasing λ, we select higher energy levels, until we reach the *condensation transition* where the entropy becomes zero, and will stay zero also for larger values of λ. The partition function is dominated by the ground state $\varepsilon^*(\alpha)$ in this region. Therefore we have what is called a *condensed phase* for $\lambda > \lambda^*(\alpha)$, with λ^* defined as the solution of:

$$\alpha + \zeta(\lambda_*) - \lambda_*\zeta'(\lambda_*) = 0. \tag{4}$$

Finally, the free energy is given by

$$\phi_\alpha(\lambda) = \begin{cases} \frac{\alpha+\zeta(\lambda)}{\lambda} & \lambda < \lambda_*(\alpha), \\ \varepsilon_*(\alpha) & \lambda \geq \lambda_*(\alpha). \end{cases} \tag{5}$$

3 Modern Hopfield Networks

Here we review and extend the formalism introduced in Lucibello and Mézard (2024) to compute the exact asymptotic thresholds for the capacity of the MHN introduced in Ramsauer et al. (2020a). While the study in Lucibello and Mézard (2024) was limited to rotational invariant pattern ensembles, we generalize it to a much larger class of ensembles for which the free energy of an auxiliary REM has the commonly found self-averaging property (Mézard et al., 1986). We will apply the REM formalism to the computation of the capacity in presence of binary patterns (Section 3.2) and then to patterns generated to live on a latent manifold (Section 3.3). Eventually, we consider the infinitive data-points scenario and prove that a linear manifold becomes the only attractor in this case.

We consider the generalization of the Hopfield model proposed in Ramsauer et al. (2020a). Given P patterns $\boldsymbol{\xi}^\mu \in \mathbb{R}^N$ and continuous configurations $x \in \mathbb{R}^N$, the energy of the model is defined as

$$E(\mathbf{x}) = -\frac{1}{\lambda} \log \sum_{\mu=1}^{P} e^{\lambda\, \mathbf{x}\cdot\boldsymbol{\xi}^\mu} + \frac{1}{2}\|\mathbf{x}\|^2, \tag{6}$$

where $\cdot$ denotes scalar product and λ plays the role of a fictitious inverse temperature. We study the retrieval of a a generic pattern, say $\boldsymbol{\xi}^1$, by analyzing under which conditions the energy has a local minimum in correspondence of $\boldsymbol{\xi}^1$, in particular in the limit of large systems size N and i.i.d. patterns. We shall assume that the components of $\mathbf{x}$ and of $\boldsymbol{\xi}^\mu$ are of order 1, so that the energy scales linearly with N at large N.

3.1 THE SIGNAL-TO-NOISE ARGUMENT

An exact capacity threshold can be obtained using a simple signal-to-noise argument. The idea is that, if we are evaluating the energy close to a given pattern $\boldsymbol{\xi}^1$, all the other patterns can be interpreted as 'unrelated noise' that interferes with the retrieval of the pattern. One can therefore split the energy into a signal and a noise contribution and approximate to exponential precision the energy density at large N by

$$\frac{1}{N}E(\mathbf{x}) \approx -\max\left(\frac{\mathbf{x}\cdot\boldsymbol{\xi}^1}{N}, \Phi(\mathbf{x})\right) + \frac{1}{2N}\|\mathbf{x}\|^2, \tag{7}$$

where the noise function Φ is defined by

$$\Phi(\mathbf{x}) = \frac{1}{\lambda N}\log\sum_{\mu\geq 2} e^{\lambda\,\mathbf{x}\cdot\boldsymbol{\xi}^\mu}. \tag{8}$$

As we shall see, this function has a finite limit when $N \to +\infty$. The condition for retrieval is that the signal term in $\mathbf{x} = \boldsymbol{\xi}^1$ dominates the noise term, that is $\|\boldsymbol{\xi}^1\|^2 > \Phi(\boldsymbol{\xi}^1)$. We consider the case in which we have an exponential number $P = e^{\alpha N}$, $\alpha > 0$, of i.i.d. or conditionally i.i.d. patterns (as in the case of patterns produced by the hidden manifold model discussed later), normalized such that $\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\|\boldsymbol{\xi}\|^2 \equiv r_\xi^2$ is finite. For reasonable pattern distributions, $\|\boldsymbol{\xi}^1\|^2/N$ and $\Phi(\boldsymbol{\xi}^1)$ will jointly concentrate on their expected value for large N. Therefore, depending on the values on the coupling strength λ and the load α, the retrieval condition $\|\boldsymbol{\xi}^1\|^2/N > \Phi(\boldsymbol{\xi}^1)$ is either almost always true or almost always false. The critical line $\alpha_1(\lambda)$ separating the retrieval phase from the non-retrieval one when considering a uniformly chosen pattern, can be obtained solving the equation

$$r_\xi^2 = \lim_{N\to\infty}\frac{1}{N}\mathbb{E}\,\Phi(\boldsymbol{\xi}^1) = \phi_\alpha(\lambda). \tag{9}$$

Following Lucibello and Mézard (2024), we now make the crucial observation that $\phi_\alpha(\lambda)$ corresponds to the asymptotic free energy of a REM (Derrida, 1981), with distribution of energies induced by the distribution of patterns. This allows the computation of the free energy by using simple probabilistic techniques developed in this context.

The REM formalism introduced in Section 2 can be used to compute the single pattern retrieval threshold $\alpha_1(\lambda)$ obtained from Eq. (9). Specifically, one can employ the expression of the free energy of a REM presented in Eq. (5).

When the norm of the patterns is not bounded, we can have rare events in which the scalar product between $\boldsymbol{\xi}^1$ and another pattern with a very large norm can be dominating. In particular, when the typical norm is smaller than the maximum norm, there is a saturation effect in the capacity: even for very large λ, for α large enough you can always find a pattern that destabilizes the signal. This saturation of the single pattern retrieval threshold was already observed in Lucibello and Mézard (2024) in the case of Gaussian patterns.

3.2 MODERN HOPFIELD NETWORKS WITH BINARY PATTERNS

As a first application, let us evaluate the retrieval capacity $\alpha_1(\lambda)$ from Eq. (9) in the case of binary pattern $\boldsymbol{\xi}^\mu \sim \mathrm{Unif}(\{-1,+1\}^N)$. In order to compute $\mathbb{E}\Phi(\boldsymbol{\xi}^1)/N$, we first compute the generating and the rate function from Eqs. (2) and (3):

$$\zeta_{bin}(\lambda) = \lim_{N\to\infty}\frac{1}{N}\mathbb{E}_{\boldsymbol{\xi}_1}\log\mathbb{E}_{\boldsymbol{\xi}_2}\, e^{\lambda\,\boldsymbol{\xi}^1\cdot\boldsymbol{\xi}^2} = \log\cosh(\lambda), \tag{10}$$

$$s_{bin}(\epsilon) = \frac{1+\epsilon}{2}\log(1+\epsilon) + \frac{1-\epsilon}{2}\log(1-\epsilon). \tag{11}$$

Next, we compute numerically $\epsilon_*(\alpha)$ solving $s_{bin}(\epsilon) = \alpha$ for ε, and set $\lambda_*(\alpha) = s'_{bin}(\epsilon_*(\alpha))$. At this point, we can compute the REM free energy $\phi_\alpha(\lambda)$ from Eq. (5) and we can find the single pattern retrieval threshold $\alpha_1(\lambda)$ as the α for which $\phi_\alpha(\lambda) = 1$. The corresponding phase diagram is shown in Fig. 1.

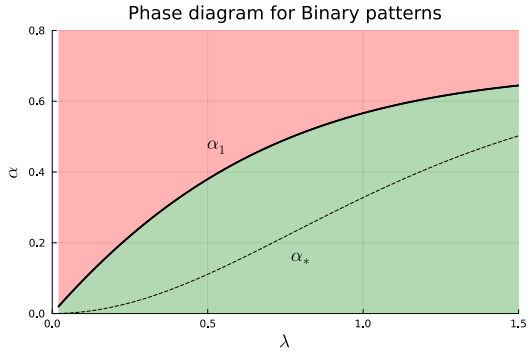

Figure 1: Phase diagram for binary patterns and variable. We also show the line $\alpha_*(\lambda)$ where the REM condensation occurs.

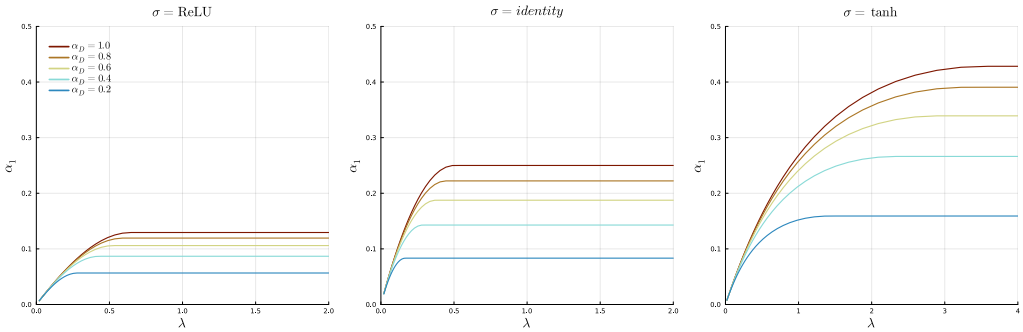

Figure 2: The retrieval lines $\alpha_1(\lambda)$ for patterns from an Hidden Manifold Model with ReLU (**left**), linear (**center**) and $\tanh$ (**right**) activations and different aspect ratios $\alpha_D = D/N$.

3.3 Patterns from the Hidden Manifold Model

Let us now consider patterns generated by a Hidden Manifold Model (HMM). The HMM is a simple synthetic generative process displaying the idea of the data manifold hypothesis, where data lies on D-dimensional sub-manifold of the ambient N-dimensional space. This generative process has been introduced and investigated in Goldt et al. (2020; 2022); Gerace et al. (2020) in the supervised learning context and studied in Negri et al. (2023) in the standard Hopfield case. Patterns are generated as $\boldsymbol{\xi}^\mu = \sigma\left(\frac{1}{\sqrt{D}} F\, \mathbf{z}^\mu\right)$, where the latent variables $\mathbf{z}^\mu$ are Gaussian, $\mathbf{z}^\mu \sim \mathcal{N}(0, I_D)$, σ is an element-wise non-linearity, and $F \in \mathbb{R}^{N\times D}$ is a random matrix with i.i.d. standard Gaussian entries. We define $\alpha_D = D/N$, and assume $D, N \to +\infty$ with α_D having a finite limit. The REM generating function (2) can be written as

$$\zeta_{HMM}(\lambda) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_{F,\mathbf{z}^1} \log \mathbb{E}_{\mathbf{z}^2}\, e^{\lambda\, \sigma\left(\frac{1}{\sqrt{D}} F\, \mathbf{z}^1\right) \cdot \sigma\left(\frac{1}{\sqrt{D}} F\, \mathbf{z}^2\right)} \tag{12}$$

We leverage the replica method to perform the computation, that is for a given integer $n > 0$ we write the replicated partition function

$$\mathbb{E} Z^n_{HMM} = \mathbb{E}_{F,\mathbf{z},\mathbf{z}^{1:n}}\, e^{\lambda \sum_{a=1}^n \sigma\left(\frac{1}{\sqrt{D}} F\, \mathbf{z}\right) \cdot \sigma\left(\frac{1}{\sqrt{D}} F\, \mathbf{z}^a\right)}, \tag{13}$$

we perform saddle point evaluation within a Replica Symmetric (RS) ansatz, and we obtain by analytical continuation the result for $\zeta_{HMM}(\lambda) = \lim_{n\to 0} \lim_{N\to\infty} \frac{1}{nN} \log \mathbb{E} Z^n_{HMM}$. The details of the computation are provided in A. The result is

$$\zeta_{HMM}(\lambda) = -\alpha_D m \hat{m} - \frac{\alpha_D}{2}(q_d \hat{q}_d - q_0 \hat{q}_0) + \alpha_D G_S + G_E, \tag{14}$$

with

$$G_S = -\frac{1}{2}\log\left(1-\hat{q}_d+\hat{q}_0\right)+\frac{1}{2}\frac{\hat{m}^2+\hat{q}_0}{1-\hat{q}_d+\hat{q}_0}, \tag{15}$$

$$G_E = \int Dt \int Du^0 \log \int Du\, e^{\lambda\sigma\left(u^0\right)\sigma\left(\sqrt{q_d-q_0}u+mu^0-\sqrt{q_0-m^2}t\right)}. \tag{16}$$

Here $\int Dx$ denotes standard Gaussian integration. Eq. (14) has to be evaluated on its stationary point with respect to the auxiliary parameters $q_d, \hat{q}_d, q_0, \hat{q}_0$. Solving numerically the saddle point equations and then using the usual REM formalism we can obtain the single pattern retrieval threshold $\alpha_1(\lambda)$. Results are shown for patterns from a linear manilfold, ReLU and tanh activations in Fig. 2. The capacity decreases as the latent dimension D shrinks. Notice that this is not straightforward, since the expected distance between the patterns does not depend on D. Another phenomena is the saturation effect that we can observe in all the three studied cases, though for $\sigma = \tanh$ it takes place at larger λ values. This was anticipated in Section 3.1: if the typical norm of the patterns is smaller than the maximum norm, even at large λ, for α big enough we will find patterns with large scalar product in the noise term.

4 Generalization in Modern Hopfield Networks

We hereby consider data generated on a linear manifold defined by a feature matrix F. In the previous sections we have proved the existence of a maximum amount of patterns that can be retrieved as local minima of the energy function defined in Eq. (6). We will now show that, by increasing the number of patterns, we can bring the model closer to the condition of having the entire latent manifold as the only attractor in the landscape. We call this specific scenario *generalization* limit of the MHN. Consider data generated on a linear hidden manifold as $\sigma(F\,\mathbf{z}) = F\,\mathbf{z}$, where we have re-absorbed the factor $1/\sqrt{D}$ into F for simplicity of the notation. Then we can compute the energy gradient-descent dynamics driving the memory retrieval as

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \nabla_{\mathbf{x}} E(\mathbf{x}^{(t)}), \tag{17}$$

Let us rewrite the energy function of the MHN as

$$E(\mathbf{x}) = -\frac{1}{\lambda}\log\int D\,\mathbf{z}\, e^{\lambda\,\mathbf{x}\,F\,\mathbf{z}} + \frac{1}{2}\|\mathbf{x}\|^2 \tag{18}$$

$$= -\frac{1}{\lambda}\log\left(e^{\frac{\lambda^2}{2}\,\mathbf{x}^\top\,FF^\top\,\mathbf{x}}\right) + \frac{1}{2}\|\mathbf{x}\|^2 \tag{19}$$

$$= \frac{1}{2}\,\mathbf{x}^\top\left(\mathbf{1}_N - \lambda FF^\top\right)\mathbf{x}, \tag{20}$$

By choosing $\lambda = 1$ we obtain the following updating rule for the dynamics

$$\mathbf{x}^{(t+1)} = FF^\top\mathbf{x}^{(t)}, \tag{21}$$

which projects any configuration $\mathbf{x}^{(t)}$ on the data-manifold in one single iteration. The entire manifold is now attractive for the gradient descent dynamics.

5 Conclusions

In this paper, we extended the formalism of the Random Energy Model applied to study the retrieval transition in dense associative memory to more general pattern ensembles. In particular, we analyzed a simple case of structured patterns, the Hidden Manifold Model, considering the effect that the latent dimension and the non-linearity have on the retrieval of a pattern. Even if the typical distance between patterns does not depend on the hidden dimension, we found that it affects the retrieval threshold, which decreases with a smaller latent dimension. It would be interesting to extend this study to what happens beyond the retrieval threshold. For instance, one possible research direction is to analyze the stability of mixtures of patterns as in Kalaj et al. (2024), to see how generalization extends to the exponential regime. We finally stress the importance of studying MHNs because of their formal equivalence with empirical-score-driven Diffusion Models (DMs), state-of-the-art generative models (Ventura et al., 2025; Achilli et al., 2024; Biroli et al., 2024; Ambrogioni, 2024). Specifically, the energy function of a MHN with spherical patterns (as treated in Lucibello and Mézard (2024)) coincides with the time-dependent potential of a DM by substituting $1/\lambda$ with the diffusion time t.

ACKNOWLEDGEMENTS

This publication benefited from European Union - Next Generation EU funds, component M4.C2, investment 1.1. - CUP J53D23001330001.

REFERENCES

Beatrice Achilli, Enrico Ventura, Gianluigi Silvestri, Bao Pham, Gabriel Raya, Dmitry Krotov, Carlo Lucibello, and Luca Ambrogioni. Losing dimensions: Geometric memorization in generative diffusion. *arXiv:2410.08727*, 2024.

Luca Ambrogioni. In search of dispersed memories: Generative diffusion models are associative memory networks. *arXiv:2309.17290*, 2023.

Luca Ambrogioni. The statistical thermodynamics of generative diffusion models: Phase transitions, symmetry breaking and critical instability. *arXiv:2310.17467*, 2024.

Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007–1018, 1985.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

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, 2017.

Bernard Derrida. Random-energy model: An exactly solvable model of disordered systems. *Physical Review B*, 24(5):2613–2626, 1981.

Mattia Fedrigo, Franco Flandoli, and Francesco Morandin. A large deviation principle for the free energy of random gibbs measures with application to the rem. *Annali di Matematica Pura ed Applicata*, 186:381–417, 2007.

Elizabeth Gardner. Multiconnected neural network models. *Journal of Physics A: Mathematical and General*, 20(11):3453, 1987.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462, 2020.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471, 2022.

Benjamin Hoover, Hendrik Strobelt, Dmitry Krotov, Judy Hoffman, Zsolt Kira, and Duen Horng Chau. Memory in Plain Sight: A Survey of the Uncanny Resemblances between Diffusion Models and Associative Memories. *arXiv:2309.16750*, 2023.

Benjamin Hoover, Yuchen Liang, Bao Pham, Rameswar Panda, Hendrik Strobelt, Duen Horng Chau, Mohammed Zaki, and Dmitry Krotov. Energy transformer. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

Silvio Kalaj, Clarissa Lauditi, Gabriele Perugini, Carlo Lucibello, Enrico M. Malatesta, and Matteo Negri. Random features hopfield networks generalize retrieval to previously unseen examples. *arXiv:2407.05658*, 2024.

Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, 2016.

Carlo Lucibello and Marc Mézard. Exponential capacity of dense associative memories. *Phys. Rev. Lett.*, 132:077301, 2024.

Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin Glass Theory and Beyond*. World Scientific, 1986.

Matteo Negri, Clarissa Lauditi, Gabriele Perugini, Carlo Lucibello, and Enrico Malatesta. Storage and learning phase transitions in the random-features hopfield model. *Physical Review Letters*, 131:257301, 2023.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield Networks is All You Need. *arXiv: 2008.02217*, 2020a.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv:2008.02217*, 2020b.

Enrico Ventura, Beatrice Achilli, Gianluigi Silvestri, Carlo Lucibello, and Luca Ambrogioni. Manifolds, random matrices and spectral gaps: The geometric phases of generative diffusion. In *International Conference on Learning Representations*, 2025.

A Replica calculation for the MHN capacity with HMM patterns

Setting $n = 0$ fixes the reference-only overlaps, i.e. $q_{00} = 1$ and $\hat{q}_{00} = 0$.

The RS ansatz is

$$q_{ab} = \begin{pmatrix} 1 & m & \dots & m \\ m & q_d & & q_0 \\ \vdots & & \ddots & \\ m & q_0 & & q_d \end{pmatrix}; \; \hat{q}_{ab} = \begin{pmatrix} 0 & \hat{m} & \dots & \hat{m} \\ \hat{m} & \hat{q}_d & & \hat{q}_0 \\ \vdots & & \ddots & \\ \hat{m} & \hat{q}_0 & & \hat{q}_d \end{pmatrix} \tag{22}$$

The entropic term is

$$G_S = \frac{1}{n} \log \int \prod_{a=0}^{n} \frac{dz^a}{\sqrt{2\pi}} \, e^{-\frac{1}{2}\sum_a (z^a)^2} e^{\frac{1}{2}\sum_{ab} \hat{q}_{ab} z^a z^b} \tag{23}$$

$$= \frac{1}{n} \log \int \frac{dz^0}{\sqrt{2\pi}} \prod_{a'=1}^{n} \frac{dz^{a'}}{\sqrt{2\pi}} \, e^{-\frac{1}{2}\left(z^0\right)^2 + \hat{m} z^0 \sum_{a'} z^{a'}} e^{-\frac{1}{2}(1-(\hat{q}_d - \hat{q}_0)) \sum_{a'} \left(z^{a'}\right)^2 + \frac{1}{2}\hat{q}_0 \left(\sum_{a'} z^{a'}\right)^2} \tag{24}$$

$$= \frac{1}{n} \log \int \prod_{a'=1}^{n} \frac{dz^{a'}}{\sqrt{2\pi}} \, e^{\frac{1}{2}\hat{m}^2 \left(\sum_{a'} z^{a'}\right)^2} e^{-\frac{1}{2}(1-\hat{q}_d + \hat{q}_0) \sum_{a'} \left(z^{a'}\right)^2 + \frac{1}{2}\hat{q}_0 \left(\sum_{a'} z^{a'}\right)^2} \tag{25}$$

$$= \frac{1}{n} \log \int Dt \int \prod_{a'=1}^{n} \frac{dz^{a'}}{\sqrt{2\pi}} \, e^{-\frac{1}{2}(1-\hat{q}_d + \hat{q}_0) \sum_{a'} \left(z^{a'}\right)^2} e^{\sqrt{\hat{m}^2 + \hat{q}_0} t \sum_{a'} z^{a'}} \tag{26}$$

$$= \frac{1}{n} \log \int Dt \left(\frac{1}{\sqrt{1 - (\hat{q}_d - \hat{q}_0)}} e^{\frac{1}{2} \frac{\left(\hat{m}^2 + \hat{q}_0\right)}{(1-\hat{q}_d + \hat{q}_0)} t^2} \right)^n \tag{27}$$

so that

$$\lim_{n \to 0} G_S = -\frac{1}{2} \log \left(1 - \hat{q}_d + \hat{q}_0\right) + \frac{1}{2} \frac{\hat{m}^2 + \hat{q}_0}{1 - \hat{q}_d + \hat{q}_0} \tag{28}$$

while the energetic part is

$$G_E = \frac{1}{n} \log \int \prod_{a=0}^{n} \frac{du^a d\hat{u}^a}{2\pi} \, e^{\lambda \sum_{a'=1}^{n} \sigma\left(u^0\right) \sigma\left(u^{a'}\right) - \sum_a i\hat{u}^a u^a - \frac{1}{2}\sum_{ab} \hat{u}^a \hat{u}^b q_{ab}} \tag{29}$$

$$= \frac{1}{n} \log \int \frac{du^0 d\hat{u}^0}{2\pi} \prod_{a'=1}^{n} \frac{du^{a'} d\hat{u}^{a'}}{2\pi} \, e^{\lambda \sum_{a'=1}^{n} \sigma\left(u^0\right) \sigma\left(u^{a'}\right) - \sum_{a'} i\hat{u}^{a'} u^{a'} - i\hat{u}^0 u^0}$$
$$\times e^{-\frac{1}{2}\left(\hat{u}^0\right)^2 - m\hat{u}^0 \sum_{a'} \hat{u}^{a'} - \frac{1}{2}(q_d - q_0) \sum_{a'} \left(\hat{u}^{a'}\right)^2 - \frac{1}{2} q_0 \left(\sum_{a'} \hat{u}^{a'}\right)^2} \tag{30}$$

$$= \frac{1}{n} \log \int \frac{du^0}{\sqrt{2\pi}} \prod_{a'=1}^{n} \frac{du^{a'} d\hat{u}^{a'}}{2\pi} \, e^{\frac{1}{2}\left(m \sum_{a'} \hat{u}^{a'} + iu^0\right)^2} e^{\lambda \sum_{a'=1}^{n} \sigma\left(u^0\right) \sigma\left(u^{a'}\right) - \sum_{a'} i\hat{u}^{a'} u^{a'}}$$
$$\times e^{-\frac{1}{2}(q_d - q_0) \sum_{a'} \left(\hat{u}^{a'}\right)^2 - \frac{1}{2} q_0 \left(\sum_{a'} \hat{u}^{a'}\right)^2} \tag{31}$$

$$= \frac{1}{n} \log \int Dt \int \frac{du^0}{\sqrt{2\pi}} \prod_{a'=1}^{n} \frac{du^{a'} d\hat{u}^{a'}}{2\pi} \, e^{-\frac{\left(u^0\right)^2}{2}} e^{\lambda \sum_{a'=1}^{n} \sigma\left(u^0\right) \sigma\left(u^{a'}\right)}$$
$$\times e^{-i \sum_{a'} \hat{u}^{a'} \left(u^{a'} - m u_0 + \sqrt{q_0 - m^2} t\right) - \frac{1}{2}(q_d - q_0) \sum_{a'} \left(\hat{u}^{a'}\right)^2} \tag{32}$$

$$= \frac{1}{n} \log \int Dt \int Du^0 \tag{33}$$

$$\times \left(\int \frac{du}{\sqrt{2\pi}} \, e^{\lambda \sigma\left(u^0\right) \sigma(u)} \frac{1}{\sqrt{q_d - q_0}} e^{-\frac{1}{2(q_d - q_0)} \left(u - m u^0 + \sqrt{q_0 - m^2} t\right)^2} \right)^n \tag{34}$$

So that

$$\lim_{n\to 0} G_E = \int Dt \int Du^0 \log\left(\int \frac{du}{\sqrt{2\pi(q_d-q_0)}}\, e^{\lambda\sigma\left(u^0\right)\sigma(u)-\frac{1}{2(q_d-q_0)}\left(u-mu^0+\sqrt{q_0-m^2}t\right)^2}\right). \tag{35}$$

We can then write

$$\zeta_\lambda(q_d,q_0,m,\hat{q}_d,\hat{q}_0,\hat{m}) = -\alpha_D m\hat{m} - \frac{\alpha_D}{2}(q_d\hat{q}_d - q_0\hat{q}_0) - \frac{\alpha_D}{2}\log\left(1-\hat{q}_d+\hat{q}_0\right) + \frac{\alpha_D}{2}\frac{\hat{m}^2+\hat{q}_0}{1-\hat{q}_d+\hat{q}_0} + \int Dt \int Du^0 \log\left(\int Du\, e^{\lambda\sigma\left(u^0\right)\sigma\left(\sqrt{q_d-q_0}u+mu^0-\sqrt{q_0-m^2}t\right)}\right) \tag{36}$$

and take its derivatives to obtain the saddle point equations

$$\hat{q}_d = \frac{2}{\alpha_D}\int Dt \int Du^0 \int Du\, e^{\lambda\sigma\left(u^0\right)\sigma\left(\sqrt{q_d-q_0}u+mu^0-\sqrt{q_0-m^2}t\right)} \times\left[\frac{\frac{\left(u-mu^0+\sqrt{q_0-m^2}t\right)^2-(q_d-q_0)}{2(q_d-q_0)^2}}{\int Du\, e^{\lambda\sigma(u^0)\sigma\left(\sqrt{q_d-q_0}u+mu^0-\sqrt{q_0-m^2}t\right)}}\right] \tag{37}$$

$$\hat{q}_0 = -\frac{2}{\alpha_D}\int Dt \int Du^0 \int Du\, e^{\lambda\sigma\left(u^0\right)\sigma\left(\sqrt{q_d-q_0}u+mu^0-\sqrt{q_0-m^2}t\right)} \times\left[\frac{\left(\frac{\left(u-mu^0+\sqrt{q_0-m^2}t\right)^2}{2(q_d-q_0)^2}+\frac{\left(u-mu^0+\sqrt{q_0-m^2}t\right)\sqrt{q_0-m^2}t+1}{2(q_d-q_0)}\right)}{\int Du\, e^{\lambda\sigma(u^0)\sigma\left(\sqrt{q_d-q_0}u+mu^0-\sqrt{q_0-m^2}t\right)}}\right] \tag{38}$$

$$\hat{m} = \frac{1}{\alpha_D}\int Dt \int Du^0 \int Du\, e^{\lambda\sigma\left(u^0\right)\sigma\left(\sqrt{q_d-q_0}u+mu^0-\sqrt{q_0-m^2}t\right)} \times\left[\frac{\frac{\left(u-mu^0+\sqrt{q_0-m^2}t\right)\left(u^0+mt(q_0-m^2)^{-1/2}\right)}{(q_d-q_0)}}{\int Du\, e^{\lambda\sigma(u^0)\sigma\left(\sqrt{q_d-q_0}u+mu^0-\sqrt{q_0-m^2}t\right)}}\right] \tag{39}$$

$$q_d = \frac{1}{1-\hat{q}_d+\hat{q}_0} + \frac{\hat{m}^2+\hat{q}_0}{\left(1-\hat{q}_d+\hat{q}_0\right)^2} \tag{40}$$

$$q_0 = \frac{1}{1-\hat{q}_d+\hat{q}_0} + \frac{\hat{m}^2+\hat{q}_d-1}{\left(1-\hat{q}_d+\hat{q}_0\right)^2} \tag{41}$$

$$m = \frac{\hat{m}}{1-\hat{q}_d+\hat{q}_0} \tag{42}$$