In-context denoising with one-layer transformers: connections between attention and associative memory retrieval

Matthew Smart[1] **Alberto Bietti**[1] **Anirvan M. Sengupta**[1,2]
[1]Flatiron Institute, New York, NY, USA
[2]Department of Physics and Astronomy, Rutgers University, Piscataway, NJ, USA
{msmart, abietti, asengupta}@flatironinstitute.org

Abstract

We introduce in-context denoising, a task that refines the connection between attention-based architectures and dense associative memory (DAM) networks, also known as modern Hopfield networks. Using a Bayesian framework, we show theoretically and empirically that certain restricted denoising problems can be solved optimally even by a single-layer transformer. We demonstrate that a trained attention layer processes each denoising prompt by performing a single gradient descent update on a context-aware DAM energy landscape, where context tokens serve as associative memories and the query token acts as an initial state. This one-step update yields better solutions than exact retrieval of either a context token or a spurious local minimum, providing a concrete example of DAM networks extending beyond the standard retrieval paradigm. Overall, this work solidifies the link between associative memory and attention mechanisms first identified by Ramsauer et al., and demonstrates the relevance of associative memory models in the study of in-context learning.

1 Introduction

The most celebrated model for associative memories in systems neuroscience is the so-called Hopfield model (Amari, 1972; Nakano, 1972; Little, 1974; Hopfield, 1982). This model has a capacity to store "memories" (stable fixed points of a recurrent update rule) proportional to the number of nodes (Hopfield, 1982; Amit et al., 1985). In the last decade, new energy functions (Krotov & Hopfield, 2016; Demircigil et al., 2017) were proposed for dense associative associative memories with much higher capacities. These energy functions are often referred to as modern Hopfield models. Ramsauer et al. (2021) pointed out the similarity between the one-step update rule of a certain modern Hopfield network (Demircigil et al., 2017) and a particular one-layer transformer map (Vaswani et al., 2017), generating interest in the statistical physics and the systems neuroscience community (Krotov & Hopfield, 2021; Krotov, 2023; Lucibello & Mézard, 2024; Millidge et al., 2022). However, the construction in Ramsauer et al. (2021) appears to emphasize the specific task of exact retrieval (converging to a fixed point), while in practice transformers may tackle many other tasks (Devlin et al., 2019; Brown et al., 2020; Touvron et al., 2023; Dosovitskiy, 2020).

To explore this connection beyond retrieval, we introduce *in-context denoising*, a task that bridges the behavior of trained transformers and associative memory networks through the lens of in-context learning (ICL). While ICL has been extensively studied in supervised settings (Garg et al., 2022; Zhang et al., 2024; Akyürek et al., 2023; Reddy, 2024), recent work suggests that transformers may internally emulate gradient descent over a context-specific loss function during inference (Von Oswald et al., 2023; Dai et al., 2023; Ahn et al., 2023). This general perspective aligns with our findings.

2 Problem formulation: In-context denoising

In this section, we describe our general setup. Recurring common notation is described in Appendix A.1. Each task corresponds to a distribution D over the probability distribution of data: $p_X \sim D$. Let $X_1, \cdots, X_{L+1} \overset{\text{iid}}{\sim} p_X$, define the sampling of the tokens. Let the noise corruption be defined by $\tilde{X} \sim p_{\text{noise}}(\cdot|X_{L+1})$. The random sequence $E = (X_1, X_2, ..., X_L, \tilde{X})$ are given as "context" (input) to a sequence model $F(\cdot;\theta)$ which outputs an estimate $\hat{X}_{L+1}$ of the original $(L+1)$-th token . The task is to minimize the expected loss $\mathbb{E}[l(\hat{X}_{L+1}, X_{L+1})]$ for some loss function $l(\cdot,\cdot)$. Namely, our problem is to find

$$\min_{\theta} \mathbb{E}_{p_X \sim D, X_{1:L+1} \sim p_X^{L+1}, \tilde{X} \sim p_{\text{noise}}(\cdot|X_{L+1})}[l(F(E,\theta), X_{L+1})]. \tag{1}$$

In practice, we choose $\tilde{X} = X_{L+1} + Z$, a pure token corrupted by the addition of isotropic Gaussian noise $Z \sim \mathcal{N}(0, \sigma_Z^2 I_n)$, and our objective function to minimize is the mean squared error (MSE) $\mathbb{E}[||\hat{X}_{L+1} - X_{L+1}||^2]$.

The first L tokens in E are "pure samples" from p_X that should provide information about the distribution for our denoising task. Our performance is expected to be no better than that of the best method, in the case that the token distribution and also the corrupting process are exactly known. The following proposition formalizes baseline to which we expect to compare our results as $L \to \infty$. We seek a function $f : \mathbb{R}^n \to \mathbb{R}^n$ such that $\mathbb{E}_{X,\tilde{X}}\left[\|X - f(\tilde{X})\|^2\right]$ is minimized. As is well-known, the Bayes optimal predictor for l_2 loss is the posterior mean.

Proposition 1. *For each task, specified by the input distribution p_X, and the noise model $p_{\tilde{X}|X}$,*

$$\mathbb{E}_{X,\tilde{X}}\left[\|X - f(\tilde{X})\|^2\right] \geq \mathbb{E}_{\tilde{X}}\left[\operatorname{Tr}\operatorname{Cov}(X \mid \tilde{X})\right]. \tag{2}$$

This lower bound is met when $f(\tilde{X}) = \mathbb{E}[X \mid \tilde{X}]$.

For completeness, the proof is in Appendix B.1.

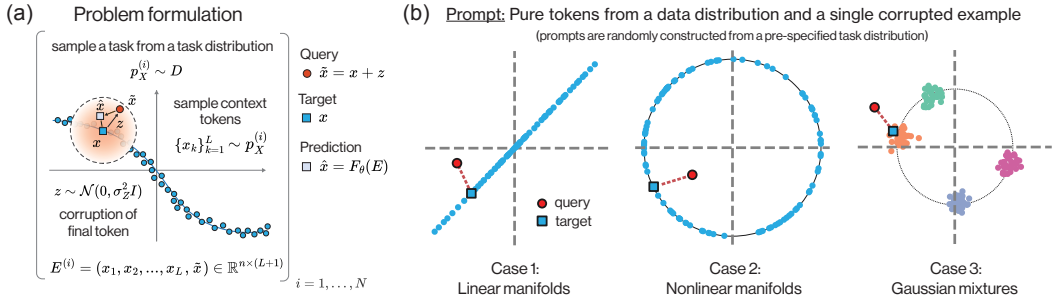


Figure 1: (a) Problem formulation for a general in-context denoising task. (b) The three denoising tasks considered here include instances of linear and non-linear manifolds as well as Gaussian mixtures. In each case, the context $E^{(i)}$ consists of a sequence of pure tokens from the data distribution $p_X^{(i)} \sim D$ where D denotes the task distribution, along with a single query token that has been corrupted by Gaussian noise. The objective is to predict the target (i.e. *denoise* the query) given information contained only in the prompt.

We consider three elementary in-context denoising tasks, where the data (vectors in $\mathbb{R}^n$) comes from:

1. Linear manifolds (d-dimensional subspaces): The data comes from d-dimensional random linear subspace of $\mathbb{R}^n$. Restricted to that space, p_X is an isotropic zero-centered Gaussian with variance of each component being σ_0^2.
2. Nonlinear manifolds (d-spheres): The data is uniformly sampled from from d-dimensional origin-centered sphere of radius R. The sphere is in a random $d+1$-dimensional linear subspace of $\mathbb{R}^n$. Restricted to that space, p_X is an isotropic zero-centered Gaussian with variance of each component being σ_0^2.

3. Small noise Gaussian mixtures (clusters): A mixture of isotropic Gaussians, where the component means have fixed norm and lives on an $n-1$ dimensional sphere or radius R. The variances of each component tend to zero.

The general setup and the three special cases are represented in Fig. 1. The details of the task-specific distributions p_X and the process for sampling tokens $\{x_t\}$ are described in Appendix B.2. The same corruption process applies to all cases: $\tilde{X} = X_{L+1} + Z, Z \sim \mathcal{N}(0, \sigma_Z^2 I_n)$.

3 In-context denoising with one-layer transformers

3.1 Theoretical results

In this section, we provide simple constructions of one-layer transformers that approximate well the Bayes optimal predictors. To motivate our choice of architecture, let us start by discussing the linear case. There, we have $f_{\text{opt}}(\tilde{X}) = \frac{\sigma_0^2}{\sigma_0^2+\sigma_Z^2} P\tilde{X}$. Note that, by the strong law of large numbers, $\hat{P} = \frac{1}{\sigma_0^2 L}\sum_{t=1}^{L} X_t X_t^T$ is a random matrix that almost surely converges component-by-component to the orthogonal projection P as $L \to \infty$, since, for each t, $X_t X_t^T$ has the expectation $\sigma_0^2 P$ and that X_t is a Gaussian random variable with zero mean and a finite covariance matrix.

We now consider a simplified one-layer linear transformer (see Appendices E.1 and E.2 for more detailed discussions) which still has sufficient expressive power to capture our finite sample approximation to the Bayes optimal answer. We define

$$\hat{X} = F_{\text{Lin}}(E, \theta) := \frac{1}{L} W_{PV} X_{1:L} X_{1:L}^T W_{KQ} \tilde{X} = \frac{1}{L}\sum_{t=1}^{L} W_{PV} X_t \langle X_t, W_{KQ}\tilde{X}\rangle. \tag{3}$$

taking values in $\mathbb{R}^n$, where $X_{1:L} := [X_1, \ldots, X_L]$ taking values in $\mathbb{R}^{n\times L}$, with learnable weights $W_{KQ}, W_{PV} \in \mathbb{R}^{n\times n}$ abbreviated by θ.

Now, our argument could be formalized into the following theorem:

Theorem 3.1. *If we have a p_X from the linear case, then the function*

$$F_{Lin}((\{X_t\}_{t=1}^L, \tilde{x}), \theta^*) = \frac{1}{L(\sigma_0^2+\sigma_Z^2)}\sum_{t=1}^{L} X_t\langle X_t, \tilde{x}\rangle \tag{4}$$

converges almost surely to the Bayes optimal answer $f_{opt}(\tilde{x})$ for all $\tilde{x} \in \mathbb{R}^n$, as $L \to \infty$. The optimal parameter θ^* refers to $W_{PV} = \alpha I_n, W_{KQ} = \beta I_n$ with $\alpha\beta = \frac{1}{\sigma_0^2+\sigma_Z^2}$.

Similarly, we could argue that the other two problems, the d-dimensional spheres and the $\sigma_0 \to 0$ zero limit of the Gaussian mixtures could be addressed by softmax attention

$$\hat{X} = F(E, \theta) := W_{PV} X_{1:L} \text{softmax}(X_{1:L}^T W_{KQ}\tilde{X}) = \frac{\sum_{t=1}^{L} W_{PV} X_t e^{\langle X_t, W_{KQ}\tilde{X}\rangle}}{\sum_{t=1}^{L} e^{\langle X_t, W_{KQ}\tilde{X}\rangle}} \tag{5}$$

taking values in $\mathbb{R}^n$. The function $\text{softmax}(z) := \frac{1}{\sum_{i=1}^n e^{z_i}}(e^{z_1}, \ldots, e^{z_n})^T \in \mathbb{R}^n$ is applied column-wise.

For both problems, namely the spheres and the $\sigma_0 \to 0$ Gaussian mixtures, we could have $W_{PV} = \alpha I_n, W_{KQ} = \beta I_n$ with $\alpha = 1, \beta = 1/\sigma_Z^2$ providing Bayes optimal answers as $L \to \infty$. In fact, we could make a general statement about distributions supported on spheres.

Theorem 3.2. *If we have a task distribution D so that the support of each p_X is the subset of some sphere, centered around the origin, with a p_X-dependent radius R, then the function*

$$F((\{X_t\}_{t=1}^L, \tilde{x}), \theta^*) = \frac{\sum_{t=1}^{L} X_t e^{\langle X_t, \tilde{x}\rangle/\sigma_Z^2}}{\sum_{t=1}^{L} e^{\langle X_t, \tilde{x}\rangle/\sigma_Z^2}} \tag{6}$$

converges almost surely to the Bayes optimal answer $f_{opt}(\tilde{x})$ for all $\tilde{x} \in \mathbb{R}^n$, as $L \to \infty$. The optimal parameter θ^* refers to $W_{PV} = I_n, W_{KQ} = \frac{1}{\sigma_Z^2} I_n$.

The proof of the theorem is in Appendix E.3. Note that the condition of p_X being supported on a sphere is not artificial as, in many practical transformers, pre-norm with RMSNorm gives you inputs on the sphere, up to learned diagonal multipliers.

For the linear case, we use linear attention, but that may not be essential. Informally speaking, the softmax attention model has the capacity to subsume the linear attention model. See Appendix F for the details of small W_{KQ} expansion and Appendix F.1 for Proposition F.2. We therefore could use the softmax model for all three cases.

3.2 Empirical results

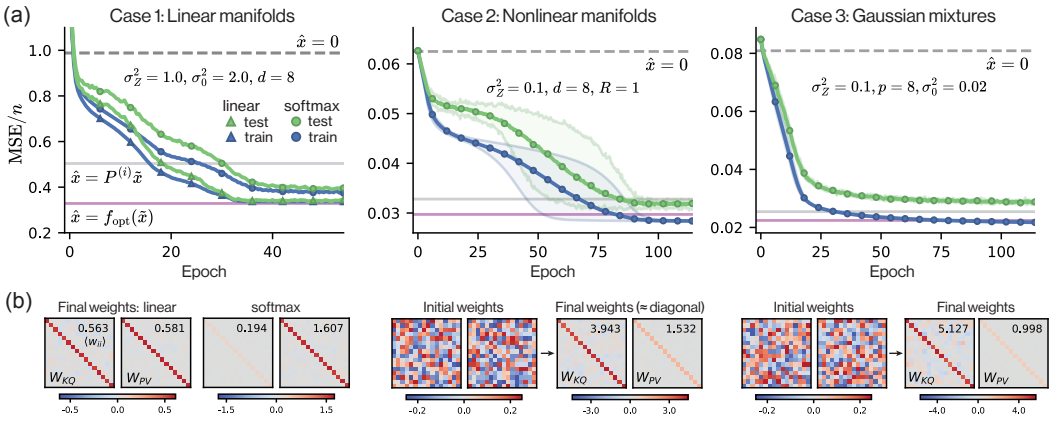


Figure 2: (a) Training dynamics for the studied cases using one-layer softmax attention (circles) as well as linear attention (triangles). Solid lines represent the average loss over six seeds, with the shaded area indicating the range for cases 2 and 3. For each case, the grey dashed baseline indicates the 0-predictor, and the pink line indicates the Bayes optimal predictor. All cases use a context length of $L = 500$, ambient dimension $n = 16$, and are trained with Adam on a dataset of size 800 with batch size 80 and standard weight initialization $w_{ij} \sim U[-1/\sqrt{n}, 1/\sqrt{n}]$. (b) Final attention weights W_{KQ} and W_{PV} are shown. For each, we indicate the mean of the diagonal elements. Initial weights are displayed for the second and third case.

Fig. 2 shows the training dynamics for the three cases showing that the qualitative nature of the trained model agrees with the theory above. The details of the training setup is in Appendix D. Further discussion of the empirical results are in Appendix F.2.

4 Connection to dense associative memory networks

In each of the denoising problems studied above, we have shown analytically and empirically that the optimal weights of the one-layer transformer are scaled identity matrices $W_{PV} \approx \alpha I, W_{KQ} \approx \beta I$. In the softmax case, the trained denoiser can be concisely expressed as

$$\hat{x} = g(X_{1:L}, \tilde{x}) := \alpha X_{1:L} \text{softmax}(\beta X_{1:L}^T \tilde{x}),$$

re-written such that $X \in \mathbb{R}^{n \times L}$ stores pure context tokens.

We now demonstrate that such denoising corresponds to one-step gradient descent (with specific step sizes) of energy models related to dense associative memory networks, also known as modern Hopfield networks (Ramsauer et al., 2021; Demircigil et al., 2017; Krotov & Hopfield, 2016).

Consider the energy function:

$$\mathcal{E}(X_{1:L}, s) = \frac{1}{2\alpha}\|s\|^2 - \frac{1}{\beta}\log\left(\sum_{t=1}^{L} e^{\beta X_t^T s}\right), \tag{7}$$

which mirrors the Ramsauer et al. (2021) construction but with a Lagrange multiplier added to the first term. An operation inherent to the associative memory perspective is the recurrent application of a denoising update. Gradient descent iteration $s(t+1) = s(t) - \gamma \nabla_s \mathcal{E}\big(X_{1:L}, s(t)\big)$ yields

$$s(t+1) = \left(1 - \frac{\gamma}{\alpha}\right) s(t) + \gamma X_{1:L} \text{softmax}\big(\beta X_{1:L}^T s(t)\big). \tag{8}$$

It is now transparent that initializing the state to the query $s(0) = \tilde{x}$ and taking a single step with size $\gamma = \alpha$ recovers the behavior of the trained attention model (Fig. 3). On the other hand, one could consider alternative step sizes and recurrent iteration. However, as Fig. 3 shows, this has the potential to degrade performance.

Additional details are provided in Appendix H. In particular, the energy model for linear attention is discussed in Appendix H.1.

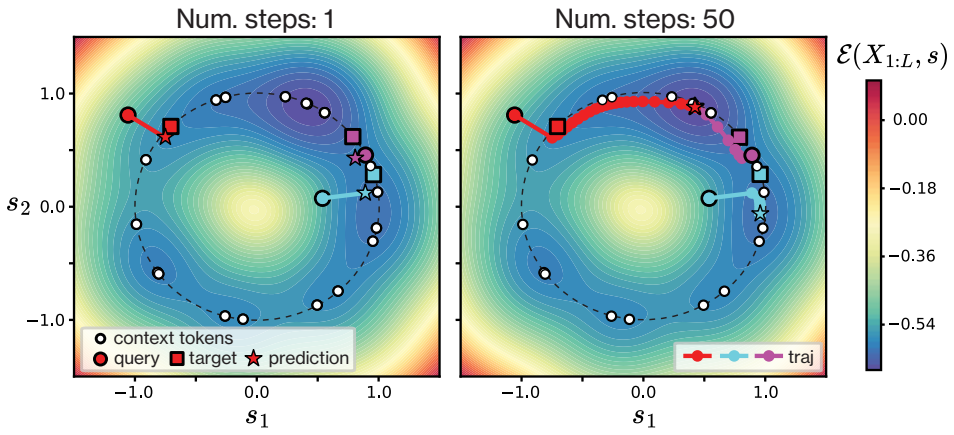


Figure 3: Gradient descent denoising for the nonlinear manifold case (spheres) in $n = 2$ with $d = 1$. A context-aware dense associative memory network $\mathcal{E}(X_{1:L}, s)$ is constructed whose gradient corresponds to the Bayes optimal update (trained attention layer). Note that the density of sampled context tokens sculpts the valleys of the energy landscape. Left: the attention step of a one-layer transformer trained on the denoising task corresponds to a single gradient descent step. Right: Iterating the denoising process—as is conventional for Hopfield networks—can potentially degrade the estimate by causing it to become query-independent (e.g. converging to a distant minimum). Here $R = 1, \sigma_Z^2 = 10, L = 20$ and $\alpha = 1, \beta = 1/\sigma_Z^2$.

5 Discussion

Overall, this work refines the connection between dense associative memories and attention layers first identified in Ramsauer et al. (2021). While we show that one energy minimization step of a particular DAM (associated with a trained attention layer) is optimal for the denoising tasks studied here, it remains an open question whether multilayer architectures with varying or tied weights could extend these results to more complex tasks by effectively performing multiple iterative steps. This aligns with recent studies on in-context learning, which have considered whether transformers with multiple layers emulate gradient descent updates on a context-specific objective (Von Oswald et al., 2023; Shen et al., 2023; Dai et al., 2023; Ahn et al., 2023), and may provide a bridge to work on emerging architectures guided by associative memory principles (Hoover et al., 2023). Investigating when and how multilayer attention architectures perform such gradient descent iterations in a manner that is both context-dependent and informed by a large training set represents an exciting direction for future research at the intersection of transformer mechanisms, associative memory retrieval, and in-context learning.

Acknowledgments

MS acknowledges M. Mézard for very useful feedback on an earlier version of this work. AS thanks D. Krotov and P. Mehta for enlightening discussions on related matters. Our early work also benefited from AS's participation in the deeplearning23 workshop at the Kavli Institute for Theoretical Physics (KITP), which was supported in part by grants NSF PHY-1748958 and PHY-2309135 to KITP.

References

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650, 2023.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=0g0X4H8yN4I`.

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

Daniel J. Amit, Hanoch Gutfreund, and H. Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007–1018, 1985. ISSN 10502947. doi: 10.1103/PhysRevA.32.1007.

D Bollé, Th M Nieuwenhuizen, I Pérez Castillo, and T Verbeiren. A spherical hopfield model. *Journal of Physics A: Mathematical and General*, 36(41):10269, 2003.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers, 2023. URL `https://arxiv.org/abs/2212.10559`.

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*, 168:288–299, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Konrad H Fischer and John A Hertz. *Spin glasses*. Number 1. Cambridge university press, 1993.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/c529dba08a146ea8d6cf715ae8930cbe-Paper-Conference.pdf`.

Izrail'S Gradstein, Iosif M Ryzhik, and Alan Jeffrey. Zwillinger, daniel (hrsg.): Table of integrals, series, and products, 2007.

Benjamin Hoover, Yuchen Liang, Bao Pham, Rameswar Panda, Hendrik Strobelt, Duen Horng Chau, Mohammed Zaki, and Dmitry Krotov. Energy transformer. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 27532–27559, 2023.

J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79(8):2554–2558, 1982. ISSN 00278424. doi: 10.1073/pnas.79.8.2554.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Dmitry Krotov. A new frontier for hopfield networks. *Nature Reviews Physics*, 5(7):366–367, 2023.

Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/eaae339c4d89fc102edd9dbdb6a28915-Paper.pdf.

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.

William A Little. The existence of persistent states in the brain. *Mathematical biosciences*, 19(1-2):101–120, 1974.

Carlo Lucibello and Marc Mézard. Exponential capacity of dense associative memories. *Phys. Rev. Lett.*, 132:077301, Feb 2024. doi: 10.1103/PhysRevLett.132.077301. URL https://link.aps.org/doi/10.1103/PhysRevLett.132.077301.

Beren Millidge, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz. Universal hopfield networks: A general framework for single-shot associative memory models. In *International Conference on Machine Learning*, pp. 15561–15583. PMLR, 2022.

Kaoru Nakano. Associatron-a model of associative memory. *IEEE Transactions on Systems, Man, and Cybernetics*, (3):380–388, 1972.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David P. Kreil, Michael K. Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=aN4Jf6Cx69.

Lingfeng Shen, Aayush Mishra, and Daniel Khashabi. Do pretrained transformers really learn in-context by gradient descent? *arXiv preprint arXiv:2310.08540*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 2017-December, 2017.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024. URL `http://jmlr.org/papers/v25/23-1042.html`.

Appendix

A Notation

A.1 Recurring notation

- n – ambient dimension of input tokens.
- $x_t \in \mathbb{R}^n$ – the value of the t-th random input token.
- $E = (X_1, ..., X_L, \tilde{X})$ – the random variable input to the sequence model. The "tilde" indicates that the final token has in some way been corrupted. E takes values $(x_1, ..., x_L, \tilde{x}) \in \mathbb{R}^{n\times(L+1)}$.
- L – context length = number of uncorrupted tokens.
- d – dimensionality of manifold S that x_t are sampled from
- N – number of training pairs

A.2 Bayes posterior notation

- $p_X(x)$ is task-dependent (the three scenarios considered here are introduced above).
- $p_{\tilde{X}}(\tilde{x})$ where $\tilde{x} = x + z$. For a sum of independent random variables, $Y = X_1 + X_2$, their pdf is a convolution $p_Y(y) = \int p_{X_1}(x)p_{X_2}(y - x)dx$. Thus:

$$\begin{aligned} p_{\tilde{X}}(\tilde{x}) &= \int p_Z(z)p_X(\tilde{x} - z)dz \\ &= C_Z \int e^{-\|z\|^2/2\sigma_Z^2} p_X(\tilde{x} - z)dz \end{aligned}$$

  where $C_Z = (2\pi\sigma_Z^2)^{-n/2}$ is a constant.
- $p_{\tilde{X}|X}(\tilde{x} \mid x)$: This is simply

$$p_Z(\tilde{x} - x) = C_Z e^{-\|\tilde{x}-x\|^2/2\sigma_Z^2}.$$

- $p_{X|\tilde{X}}(x \mid \tilde{x})$: By Bayes' theorem, this is

$$\begin{aligned} p_{X|\tilde{X}}(x \mid \tilde{x}) &= \frac{p_{\tilde{X}|X}(\tilde{x} \mid x)p_X(x)}{p_{\tilde{X}}(\tilde{x})} \\ &= \frac{e^{-\|\tilde{x}-x\|^2/2\sigma_Z^2} p_X(x)}{\int e^{-\|\tilde{x}-x'\|^2/2\sigma_Z^2} p_X(x')dx'}. \end{aligned}$$

- Posterior mean:

$$\begin{aligned} \mathbb{E}_{X|\tilde{X}}[X \mid \tilde{X}] &= \int x\, p_{X|\tilde{X}}(x \mid \tilde{x})dx \\ &= \frac{1}{p_{\tilde{X}}(\tilde{x})} \int x\, p_{X,\tilde{X}}(x, \tilde{x})dx. \end{aligned}$$

B Bayes optimal predictors for square loss

B.1 Proof of Proposition 1

Proof. Observe that

$$\begin{aligned}
\mathbb{E}\left[\|X - f(\tilde{X})\|^2\right] &= \mathbb{E}_{\tilde{X}}\left[\mathbb{E}_{X|\tilde{X}}\left[\|X - f(\tilde{X})\|^2 \mid \tilde{X}\right]\right] \\
&= \mathbb{E}_{\tilde{X}}\Big[\mathbb{E}_{X|\tilde{X}}\left[\|X - \mathbb{E}[X \mid \tilde{X}]\|^2 \mid \tilde{X}\right] \\
&\qquad + \|\mathbb{E}[X \mid \tilde{X}] - f(\tilde{X})\|^2\Big] \\
&\geq \mathbb{E}_{\tilde{X}}\left[\mathbb{E}_{X|\tilde{X}}\left[\|X - \mathbb{E}[X \mid \tilde{X}]\|^2 \mid \tilde{X}\right]\right] \\
&= \mathbb{E}_{\tilde{X}}\left[\operatorname{Tr}\operatorname{Cov}(X \mid \tilde{X})\right].
\end{aligned}$$

Note the final line is independent of f. This inequality becomes an equality when $f(\tilde{X}) = \mathbb{E}[X \mid \tilde{X}]$. □

B.2 Details of the distributions for the three individual tasks

B.2.1 Case 1 - Linear manifolds

A given training prompt consists of pure tokens sampled from a random d-dimensional subspace S of $\mathbb{R}^n$.

- Let P be the orthogonal projection operator to a random d-dim subspace S of $\mathbb{R}^n$, sampled according to the uniform measure, induced by the Haar measure on the coset space $O(n)/O(n-d) \times O(d)$, on the Grassmanian $G(d, n)$, the manifold of all d-dimensional subspaces of $\mathbb{R}^n$.
- Let $Y \sim \mathcal{N}(0, \sigma_0^2 I_n)$ and define $X = PY$; we use this process to construct the starting sequences $(X_1, ..., X_{L+1})$ of $L+1$ independent tokens.

We thus have $p_X = \mathcal{N}(0, \sigma_0^2 P)$, with the Haar distribution of P characterizing the task ensemble associated with D.

B.2.2 Case 2 - Nonlinear manifolds

We focus on the case of d-dimensional spheres of fixed radius R centered at the origin in $\mathbb{R}^n$.

- Choose a random $d+1$-dim subspace V of $\mathbb{R}^n$, sampled according to the uniform measure, as before, on the Grassmanian $G(d+1, n)$. The choice of this random subspace generates the distribution of tasks D.
- Inside V, sample uniformly from the radius R sphere (once more, a Haar induced measure on a coset space $O(d+1)/O(d)$). We use this process to construct input sequences $X_{1:L+1} = (x_1, ..., x_{L+1})$ of $L+1$ independent tokens.

In practice, we uniformly sample points with fixed norm in $\mathbb{R}^d$ and embed them in $\mathbb{R}^n$ by concatenating zeros. We then rotate the points by selecting a random orthogonal matrix $Q \in \mathbb{R}^{n \times n}$.

B.2.3 Case 3 - Gaussian mixtures (Clustering)

Pure tokens are sampled from a weighted mixture of isotropic Gaussians in n-dimensions, $\{w_a, (\mu_a, \sigma_a^2)\}_{a=1}^K$. The density is

$$p_X(x) = \sum_{a=1}^{K} w_a C_a e^{-\|x-\mu_a\|^2/2\sigma_a^2},$$

where $C_a = (2\pi\sigma_a^2)^{-n/2}$ are normalizing constants. The μ_a are independently chosen from a uniform distribution on the radius R sphere of dimension $n-1$, centered around zero. The distribution of tasks D, is decided by the choice of $\{\mu_a\}_{a=1}^K$.

For our ideal case, we will consider the limit that the variances go to zero. In that case, the density is simply

$$p_{X_0}(x) = \sum_{a=1}^{K} w_a \delta(x - \mu_a).$$

C Details of Bayes optimal denoising baselines for each case

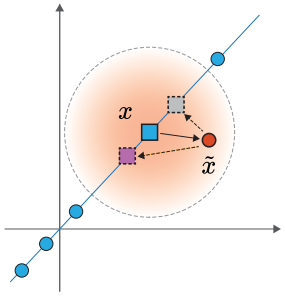
Linear case. For the linear denoising task, pure samples X are drawn from an isotropic Gaussian in a restricted subspace. The following result provides the Bayes optimal predictor in this case, the proof of which is in Appendix C.1.

Proposition 2. *For p_X corresponding to Subsection B.2.1, the Bayes optimal answer is*

$$f_{opt}(\tilde{X}) = \mathbb{E}[X|\tilde{X}] = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_Z^2} P\tilde{X}, \tag{A.1}$$

and the expected loss is

$$\mathbb{E}\left[\|P\tilde{X} - X_{L+1}\|^2\right] = d\sigma_0^2\sigma_Z^2/(\sigma_0^2 + \sigma_Z^2). \tag{A.2}$$



Projection

$\hat{x} = P^{(i)}\tilde{x}$

Projection (shrunk)

$\hat{x} = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_Z^2} P^{(i)}\tilde{x}$

Figure 4: Baseline estimators for the case of random linear manifolds with projection operator $P^{(i)}$.

Manifold case. In the nonlinear manifold denoising problem, we focus on the case of lower dimensional spheres S (e.g. the circle $S^1 \subset \mathbb{R}^2$). For such manifolds, the Bayes optimal answer is given by the following proposition.

Proposition 3. *For p_X defined as in Subsection B.2.2, with P being the orthogonal projection operator to V, the $d+1$ dimensional linear subspace, with R being the radius of sphere S, the Bayes optimal answer is*

$$\begin{aligned} f_{opt}(\tilde{X}) &= \mathbb{E}[X \mid \tilde{X}] \\ &= \frac{\int e^{\langle x, \tilde{X}_\parallel \rangle/\sigma_Z^2}\, x\, dS_x}{\int e^{\langle x, \tilde{X}_\parallel \rangle/\sigma_Z^2}\, dS_x} \qquad \text{(A.3)} \\ &= \frac{I_{\frac{d+1}{2}}\left(R\frac{\|\tilde{X}_\parallel\|}{\sigma_Z^2}\right)}{I_{\frac{d-1}{2}}\left(R\frac{\|\tilde{X}_\parallel\|}{\sigma_Z^2}\right)} R\frac{\tilde{X}_\parallel}{\|\tilde{X}_\parallel\|}, \qquad \text{(A.4)} \end{aligned}$$

where $\tilde{X}_\parallel = P\tilde{X}$ and I_ν is the modified Bessel function of the first kind.

Clustering case. For clustering with isotropic Gaussian mixtures $\{w_a, (\mu_a, \sigma_a^2)\}_{a=1}^p$, the Bayes optimal predictors for some important special cases are as follows. See Appendix C.3 for the general case.

Proposition 4. *For general isotropic Gaussian model with $\sigma_a = \sigma_0, ||\mu_a|| = R$ for all $a = 1, \dots, K$.*

$$f_{opt}(\tilde{X}) = \mathbb{E}[X|\tilde{X}] \\ = \frac{\sigma_0^2}{\sigma_0^2+\sigma_Z^2}\tilde{X} + \frac{\sigma_Z^2}{\sigma_0^2+\sigma_Z^2}\frac{\sum_a w_a e^{\langle \mu_a, \tilde{X}\rangle/(\sigma_0^2+\sigma_Z^2)}\ \mu_a}{\sum_a w_a e^{\langle \mu_a, \tilde{X}\rangle/(\sigma_0^2+\sigma_Z^2)}}. \tag{A.5}$$

If $\sigma_0 \to 0$,

$$f_{opt}(\tilde{X}) = \mathbb{E}[X \mid \tilde{X}] = \frac{\sum_a w_a e^{\langle \mu_a, \tilde{X}\rangle/\sigma_Z^2}\ \mu_a}{\sum_a w_a e^{\langle \mu_a, \tilde{X}\rangle/\sigma_Z^2}}. \tag{A.6}$$

In all three cases, we notice similarities between the form of the Bayes optimal predictor, and attention operations in transformers, a connection which we explore below.

C.1 The linear case - Proof of Proposition 2

Proof. The linear denoising task is a special case of the result in Proposition 1. Here, X is an isotropic Gaussian in a restricted subspace,

$$p_{X|\tilde{X}}(x \mid \tilde{x}) = C(\tilde{x}) p_X(x) e^{-\frac{\|x-\tilde{x}\|^2}{2\sigma_Z^2}}$$

where $C(\tilde{x})$ is a normalizing factor. The noise can be decomposed into parallel and perpendicular parts using the projection P onto S, i.e.

$$\tilde{X} = \tilde{X}_\parallel + \tilde{X}_\perp = P\tilde{X} + (I-P)\tilde{X},$$

so that

$$e^{-\frac{\|x-\tilde{x}\|^2}{2\sigma_Z^2}} = e^{-\frac{\|x-\tilde{x}_\parallel\|^2}{2\sigma_Z^2}}\ e^{-\frac{\|\tilde{x}_\perp\|^2}{2\sigma_Z^2}}.$$

Only the first factor matters for $p_{X|\tilde{X}}(x \mid \tilde{x})$ since it depends on x. Then, for $x \in S$, the linear subspace supporting p_X, dropping the x independent $\tilde{x}_\perp$ contribution,

$$p_X(x) e^{-\frac{\|x-\tilde{x}_\parallel\|^2}{2\sigma_Z^2}} \propto e^{-\frac{\|x\|^2}{2\sigma_0^2} - \frac{\|x-\tilde{x}_\parallel\|^2}{2\sigma_Z^2}} \\ \propto \exp\left(-\frac{\|x - \frac{\sigma_0^2}{\sigma_0^2+\sigma_Z^2}\tilde{x}_\parallel\|^2}{2\frac{\sigma_0^2\sigma_Z^2}{\sigma_0^2+\sigma_Z^2}}\right).$$

Thus, $f(\tilde{X}) = \frac{\sigma_0^2}{\sigma_0^2+\sigma_Z^2}\tilde{X}_\parallel = \frac{\sigma_0^2}{\sigma_0^2+\sigma_Z^2}P\tilde{X}$.

Using $\tilde{X} = X + Z$, $X = PX$, and the independence of X and Z

$$\mathbb{E}\Big[\|X - \frac{\sigma_0^2}{\sigma_0^2+\sigma_Z^2}P\tilde{X}\|^2\Big] = \mathbb{E}\Big[\|\frac{\sigma_Z^2}{\sigma_0^2+\sigma_Z^2}PX\|^2\Big] + \mathbb{E}\Big[\|\frac{\sigma_0^2}{\sigma_0^2+\sigma_Z^2}PZ\|^2\Big] = \frac{\sigma_Z^4 d\sigma_0^2 + \sigma_0^4 d\sigma_Z^2}{(\sigma_0^2+\sigma_Z^2)^2} = \frac{d\sigma_0^2\sigma_Z^2}{\sigma_0^2+\sigma_Z^2}.$$

□

C.2 The manifold case - Proof of Proposition 3

Proof. In the nonlinear manifold denoising problem, we focus on the case of lower dimensional spheres S (e.g. the circle $S^1 \subset \mathbb{R}^2$). For such manifolds, we have

$$\mathbb{E}[X \mid \tilde{X}] = \frac{\int e^{-\frac{\|x-\tilde{x}_\parallel\|^2}{2\sigma_Z^2}}\ x\, p_X(x) dx}{\int e^{-\frac{\|x-\tilde{x}_\parallel\|^2}{2\sigma_Z^2}}\ p_X(x) dx} \\ = \frac{\int e^{\langle x, \tilde{x}_\parallel\rangle/\sigma_Z^2}\ x\, dS_x}{\int e^{\langle x, \tilde{x}_\parallel\rangle/\sigma_Z^2}\ dS_x}.$$

We have used the fact that $\|x - \tilde{x}_\parallel\|^2 = \|x\|^2 + \|\tilde{x}_\parallel\|^2 - 2\langle x, \tilde{x}_\parallel\rangle$ and that $\|x\|$ is fixed on the sphere.

The integrals can be evaluated directly once the parameters are specified. If S is a d–sphere of radius R, then the optimal predictor is again a shrunk projection of $\tilde{x}$ onto S,

$$\frac{\int_0^\pi e^{R\|\tilde{x}_\parallel\|\cos\theta/\sigma_Z^2}\,\cos\theta\sin^{(d-1)}\theta\,d\theta}{\int_0^\pi e^{R\|\tilde{x}_\parallel\|\cos\theta/\sigma_Z^2}\,\sin^{(d-1)}\theta\,d\theta} R\frac{\tilde{x}_\parallel}{\|\tilde{x}_\parallel\|} = \frac{I_{\frac{d+1}{2}}\left(R\frac{\|\tilde{x}_\parallel\|}{\sigma_Z^2}\right)}{I_{\frac{d-1}{2}}\left(R\frac{\|\tilde{x}_\parallel\|}{\sigma_Z^2}\right)} R\frac{\tilde{x}_\parallel}{\|\tilde{x}_\parallel\|},$$

where we used identities involving $I_\nu(y)$, modified Bessel function of the first kind of order ν (Gradstein et al., 2007). The vector $R\frac{\tilde{x}_\parallel}{\|\tilde{x}_\parallel\|}$ is the point on S in the direction of $x_\parallel$.

□

C.3 The clustering case - Proof of Proposition 4

Proof. For the clustering case involving isotropic Gaussian mixtures with parameters $\{w_a, (\mu_a, \sigma_a^2)\}_{a=1}^p$,

$$\mathbb{E}[X \mid \tilde{X}] = \frac{\int e^{-\frac{\|x-\tilde{x}\|^2}{2\sigma_Z^2}} \sum_a \left(w_a C_a e^{-\frac{\|x-\mu_\alpha\|^2}{2\sigma_a^2}}\right) x\,dx}{\int e^{-\frac{\|x-\tilde{x}\|^2}{2\sigma_Z^2}} \sum_a \left(w_a C_a e^{-\frac{\|x-\mu_a\|^2}{2\sigma_a^2}}\right) dx},$$

where $C_a = (2\pi\sigma_a^2)^{-\frac{n}{2}}$.

We can simplify this expression by completing the square in the exponent and using the fact that the integral of a Gaussian about its mean is zero. This yields

$$\mathbb{E}[X \mid \tilde{X}] = \frac{\sum_a w_a C_a m_a \int \exp(-g_a)\,dx}{\sum_a w_a C_a \int \exp(-g_a)\,dx}$$

where we have introduced

$$g_a = \frac{1}{2}\Big(\frac{\sigma_Z^2 + \sigma_a^2}{\sigma_Z^2\sigma_a^2}\Big)\|x - m_\alpha\|^2 + \frac{1}{2(\sigma_Z^2 + \sigma_a^2)}\|\tilde{x} - \mu_a\|^2,$$

with

$$m_a = \frac{\sigma_a^2\,\tilde{x} + \sigma_Z^2\,\mu_a}{\sigma_a^2 + \sigma_Z^2}.$$

Doing the integrals and using the expressions for C_a, m_a

$$\mathbb{E}[X \mid \tilde{X}] = \frac{\sum_a w_a \big(\frac{\sigma_Z^2+\sigma_a^2}{\sigma_a^2}\big)^{n/2} \exp\big(-\frac{\|\tilde{x}-\mu_a\|^2}{2(\sigma_Z^2+\sigma_a^2)}\big)\big(\frac{\sigma_a^2\,\tilde{x}+\sigma_Z^2\,\mu_a}{\sigma_a^2+\sigma_Z^2}\big)}{\sum_a w_a \big(\frac{\sigma_Z^2+\sigma_a^2}{\sigma_a^2}\big)^{n/2} \exp\big(-\frac{\|\tilde{x}-\mu_a\|^2}{2(\sigma_Z^2+\sigma_a^2)}\big)}$$

In the case that the center norms $\|\mu_a\|$ are independent of a and variances $\sigma_a^2 = \sigma_0$, we have

$$\mathbb{E}[X \mid \tilde{X}] = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_Z^2}\,\tilde{x} + \frac{\sigma_Z^2}{\sigma_0^2 + \sigma_Z^2}\frac{\sum_a w_a \mu_a \exp\left(\frac{\langle\tilde{x},\mu_a\rangle}{\sigma_Z^2+\sigma_0^2}\right)}{\sum_a w_a \exp\left(\frac{\langle\tilde{x},\mu_a\rangle}{\sigma_Z^2+\sigma_0^2}\right)}.$$

Note that in the limit that $\sigma_0 \to 0$, this becomes expressible by one-layer self-attention, since one can simply replace the matrix of cluster centers $M = [\mu_1 \dots \mu_p]$ implicit in the expression with the context $X_{1:L}$ itself,

$$\mathbb{E}[X \mid \tilde{X}] = \frac{\sum_a w_a e^{\langle \mu_\alpha, \tilde{X} \rangle / \sigma_Z^2} \mu_a}{\sum_a w_a e^{\langle \mu_\alpha, \tilde{X} \rangle / \sigma_Z^2}}.$$

□

D Training setup

Input: Let $p_X^{(1)}, \dots, p_X^{(N)} \overset{\text{iid}}{\sim} D$, be distributions sampled for one of the tasks. For each distribution $p_X^{(i)}$, we sample $E^{(i)} := (X_1^{(i)}, \dots, X_L^{(i)}, \tilde{X}^{(i)})$ taking value in $\mathbb{R}^{n \times (L+1)}$ be an input to a sequence model. We also retain the true $(L+1)$-th token $X_{L+1}^{(i)}$ for each i.

Objective: Given an input sequence $E^{(i)}$, return the uncorrupted final token $X_{L+1}^{(i)}$. We consider the mean-squared error loss over a collection of N training pairs, $\{E^{(i)}, X_{L+1}^{(i)}\}_{i=1}^N$,

$$C(\theta) = \sum_{i=1}^{N} \| F(E^{(i)}, \theta) - x_{L+1}^{(i)} \|^2, \tag{A.7}$$

where $F(E^{(i)}, \theta)$ denotes the parametrized function predicting the target final token based on input sequence $E^{(i)}$.

E Notes on attention and softmax expansion

E.1 Standard self-attention

Given a sequence of L_{seq} input tokens $x_i \in \mathbb{R}^n$ represented as a matrix $X \in \mathbb{R}^{n \times L_{\text{seq}}}$, standard self-attention defines query, key, and value matrices

$$K = W_K X, Q = W_Q X, V = W_V X \tag{A.8}$$

where $W_K, W_Q \in \mathbb{R}^{n_{\text{attn}} \times n}$ and $W_V \in \mathbb{R}^{n_{\text{out}} \times n}$. The softmax self-attention map (Vaswani et al., 2017) is then

$$\text{Attn}(X, W_V, W_K^T W_Q) := V \text{softmax}(K^T Q) \in \mathbb{R}^{n_{\text{out}} \times L_{\text{seq}}}. \tag{A.9}$$

On merging W_K, W_Q into $W_{KQ} = W_K^T W_Q$: The simplification $W_{KQ} = W_K^T W_Q$ (made here and elsewhere) is general only when $n_{\text{attn}} \geq n$; in that case, the product W_{KQ} can have rank n and thus it is reasonable to work with the combined matrix. On the other hand, if $n_{\text{attn}} < n$, then the rank of their product is at most n_{attn} and thus there are matrices in $\mathbb{R}^{n \times n}$ that cannot be expressed as $W_K^T W_Q$. A similar point can be made about W_{PV}. We note that while $n_{\text{attn}} < n$ may be used in practical settings, one often also uses multiple heads which when concatenated could be (roughly) viewed as a single higher-rank head.

We will also use the simplest version of linear attention Katharopoulos et al. (2020),

$$\text{Attn}_{\text{Lin}}(X, W_V, W_K^T W_Q) := \frac{1}{L_{\text{seq}}} V(K^T Q) \in \mathbb{R}^{n_{\text{out}} \times L_{\text{seq}}}. \tag{A.10}$$

E.2 Minimal transformer architecture for denoising

We now consider a simplified one-layer linear transformer in term of our variable $E = (X_{1:L}, \tilde{X})$ taking values in $\mathbb{R}^{n \times (L+1)}$ and start with the linear transformer which still has

sufficient expressive power to capture our finite sample approximation to the Bayes optimal answer in the linear case. Inspired by Zhang et al. (2024), we define

$$\mathrm{Attn}_{\mathrm{Lin}}(E, W_{PV}, W_{KQ}) := \frac{1}{L} W_{PV} E M_{\mathrm{Lin}} E^T W_{KQ} E \tag{A.11}$$

taking values in $\mathbb{R}^{n\times(L+1)}$. The additional aspect compared to the last subsection is the masking matrix $M_{\mathrm{Lin}} \in \mathbb{R}^{(L+1)\times(L+1)}$ which is of the form

$$M_{\mathrm{Lin}} = \begin{bmatrix} I_L & 0_{L\times 1} \\ 0_{1\times L} & 0 \end{bmatrix}, \tag{A.12}$$

preventing $W_{PV}\tilde{X}$ from being added to the output.

Note that this more detailed expression is equivalent to the form used in the main text.

$$\hat{X} = F_{\mathrm{Lin}}(E, \theta) := \frac{1}{L} W_{PV} X_{1:L} X_{1:L}^T W_{KQ} \tilde{X}$$

With learnable weights $W_{KQ}, W_{PV} \in \mathbb{R}^{n\times n}$ abbreviated by θ, we define

$$F(E, \theta) := [\mathrm{Attn}_{\mathrm{Lin}}(E, W_{PV}, W_{KQ})]_{:,L+1}. \tag{A.13}$$

Note that, when $W_{PV} = \alpha I_n, W_{KQ} = \beta I_n$, and $\alpha\beta = \frac{1}{\sigma_0^2+\sigma_Z^2}$, $F(E, \theta)$ should approximate the Bayes optimal answer $f_{\mathrm{opt}}(\tilde{X})$ as $L \to \infty$.

Similarly, we could argue that the second two problems, the d-dimesional spheres and the $\sigma_0 \to 0$ zero limit of the Gaussian mixtures could be addressed by the full softmax attention

$$\mathrm{Attn}(E, W_{PV}, W_{KQ}) = W_{PV} E \mathrm{softmax}(E^T W_{KQ} E + M) \tag{A.14}$$

taking values in $\mathbb{R}^{n\times(L+1)}$ where $M \in \bar{\mathbb{R}}^{(L+1)\times(L+1)}$ is a masking matrix of the form

$$M = \begin{bmatrix} 0_{L\times(L+1)} \\ (-\infty)1_{1\times L+1} \end{bmatrix}, \tag{A.15}$$

once more, preventing the contribution of $\tilde{X}$ value to the output. The function $\mathrm{softmax}(z) := \frac{1}{\sum_{i=1}^n e^{z_i}}(e^{z_1}, \ldots, e^{z_n})^T \in \mathbb{R}^n$ is applied column-wise.

We then define

$$F(E, \theta) := [\mathrm{Attn}(E, W_{PV}, W_{KQ})]_{:,L+1}, \tag{A.16}$$

which is equivalent to the simplified form used in the main text:

$$\hat{X} = F(E, \theta) := W_{PV} X_{1:L} \mathrm{softmax}(X_{1:L}^T W_{KQ} \tilde{X}).$$

E.3 Proof of Theorem 3.2

Proof. Let the support of p_X be a subset of a sphere, centered around the origin, of radius R. Then the function

$$g(\{X_t\}_{t=1}^L, \tilde{x}) = \frac{\sum_{t=1}^L X_t e^{\langle X_t, \tilde{x}\rangle/\sigma_Z^2}}{\sum_{t=1}^L e^{\langle X_t, \tilde{x}\rangle/\sigma_Z^2}} = \frac{\frac{1}{L}\sum_{t=1}^L X_t e^{\langle X_t, \tilde{x}\rangle/\sigma_Z^2}}{\frac{1}{L}\sum_{t=1}^L e^{\langle X_t, \tilde{x}\rangle/\sigma_Z^2}}. \tag{A.17}$$

Both the numerator $\frac{1}{L}\sum_{t=1}^L X_t e^{\langle X_t, \tilde{x}\rangle/\sigma_Z^2}$ and the denominator $\frac{1}{L}\sum_{t=1}^L e^{\langle X_t, \tilde{x}\rangle/\sigma_Z^2}$ are averages of independent and identically distributed bounded random variables. By the strong law of large numbers, as $L \to \infty$, the average vector in the numerator converges to almost surely to $\int e^{\langle x, \tilde{x}_\parallel\rangle/\sigma_Z^2}\, x\, dp_X(x)$ for each component, while the average in the denominator almost surely converges $\int e^{\langle x, \tilde{x}_\parallel\rangle/\sigma_Z^2}\, dp_X(x)$, which is positive. So, as $L \to \infty$, the ratio in Eq. A.17 converges almost surely to

$$\frac{\int e^{\langle x, \tilde{x}_\parallel\rangle/\sigma_Z^2}\, x\, dp_X(x)}{\int e^{\langle x, \tilde{x}_\parallel\rangle/\sigma_Z^2}\, dp_X(x)}$$

which is the Bayes optimal answer $f_{\mathrm{opt}}(\tilde{x})$ for all $\tilde{x} \in \mathbb{R}^n$. □

F Limiting behaviors of the softmax function and attention

For small argument

A Taylor expansion of the softmax function at zero gives

$$\text{softmax}(\beta v) = \frac{1}{Z}\left(\mathbb{1} + \beta v + O(\beta^2)\right),$$

where $Z = \sum_i \left(1 + \beta v_i + O(\beta^2)\right)\big) = L(1 + \beta\bar{v} + O(\beta^2))$ is a normalizing factor, with $\bar{v} = \frac{1}{L}\sum_i v_i$. The notation $\mathbb{1}$ stands for a column vector of ones with the same dimension as v.

Thus, we have

Lemma F.1 (Small argument expansion of softmax)**.** *As $\beta \to 0$,*

$$\text{softmax}(\beta v) = \frac{1}{L\left(1 + \beta\bar{v} + O(\beta^2)\right)}\left(\mathbb{1} + \beta v + O(\beta^2)\right) = \frac{1}{L}\left(\mathbb{1} + \beta(v - \bar{v}\mathbb{1}) + O(\beta^2)\right).$$

Proposition F.2. *As $\epsilon \to 0$,*

$$F\Big(E, \big(\frac{1}{\epsilon}W_{PV}, \epsilon W_{KQ}\big)\Big) = \frac{1}{\epsilon}W_{PV}\bar{X} + \frac{1}{L}W_{PV}\sum_{t=1}^{L} X_t(X_t - \bar{X})^T W_{KQ}\tilde{X} + O(\epsilon), \quad \text{(A.18)}$$

where $\bar{X} = \frac{1}{L}\sum_{t=1}^{L} X_t$ is the empirical mean.

See Appendix F for the details of small W_{KQ} expansion and Appendix F.1 for the proof of Proposition F.2.

For case 1, note that $\mathbb{E}[X_t] = 0$ and covariance of X_t is finite, $E[\bar{X}] = 0$, and $E[||\bar{X}||^2] = O(\frac{1}{L})$, allowing us to drop $\bar{X}$ as $L \to \infty$. If, in addition, ϵ is small, only the second term survives. Thus, $F\big(E, (\frac{1}{\epsilon}W_{PV}, \epsilon W_{KQ})\big)$ starts to approximate $F_{\text{Lin}}\big(E, (W_{PV}, W_{KQ})\big)$ when L is large and ϵ is small, with $\epsilon\sqrt{L}$ large.

F.1 Proof of Proposition F.2

Proof.

$$F\Big(E, \big(\frac{1}{\epsilon}W_{PV}, \epsilon W_{KQ}\big)\Big) := \frac{1}{\epsilon}W_{PV}X_{1:L}\text{softmax}(\epsilon X_{1:L}^T W_{KQ}\tilde{X}).$$

Using Lemma F.1, as $\epsilon \to 0$,

$$F\Big(E, \big(\frac{1}{\epsilon}W_{PV}, \epsilon W_{KQ}\big)\Big) = \frac{1}{\epsilon}W_{PV}X_{1:L}\left[\frac{1}{L}\left(\mathbb{1}_L + \epsilon\big(X_{1:L}^T W_{KQ}\tilde{X} - (\frac{1}{L}\sum_t X_t^T W_{KQ}\tilde{X})\mathbb{1}_L\big) + O(\epsilon^2)\right)\right]$$

$$= \frac{1}{\epsilon}W_{PV}\bar{X} + \frac{1}{L}W_{PV}\sum_{t=1}^{L} X_t(X_t - \bar{X})^T W_{KQ}\tilde{X} + O(\epsilon), \quad \text{(A.19)}$$

where $\bar{X} = \frac{1}{L}\sum_{t=1}^{L} X_t$ is the empirical mean and the notation $\mathbb{1}_L$ emphasizes that it is a column vector of ones with dimension L.

□

For large argument

As $\beta \to \infty$, the softmax function simply selects the maximum over its inputs (as long as the the maximum is unique):

$$\text{softmax}(\beta v) \approx \begin{cases} 1 & \text{if } i = \arg\max_j v_j, \\ 0 & \text{otherwise.} \end{cases}$$

In this case, all attention weight is given to a single element, and the others are effectively ignored.

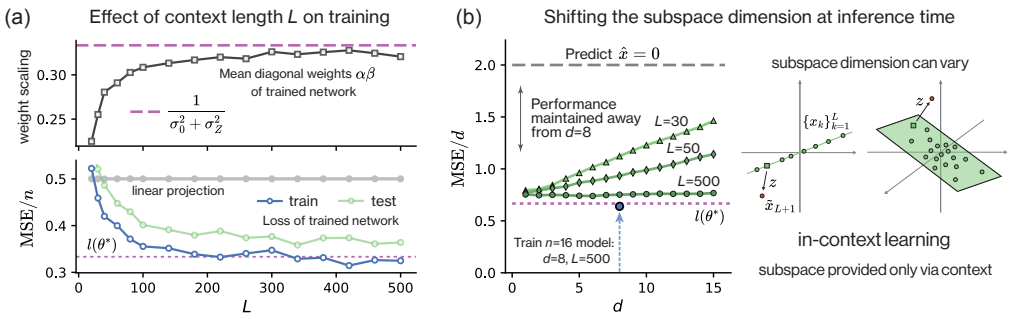


Figure 5: (a) Trained linear attention network converges to Bayes optimal estimator as context length increases ($n = 16$, $d = 8$, $\sigma_0^2 = 2, \sigma_z^2 = 1$). (b) A network trained to denoise subspaces of dimension $d = 8$ can accurately denoise subspaces of different dimensions presented at inference time, given sufficient context.

F.2 Details of the empirical results

F.3 Case 1 - Linear manifolds

The Bayes optimal predictor for the linear denoising task from Section **??** suggests that the linear attention weights should be scaled identity matrices with their product satisfying $\alpha\beta = \frac{1}{\sigma_0^2+\sigma_Z^2}$. Fig. 2 shows that a one-layer network of size $n = 16$ trained on tasks with $\sigma_Z^2 = 1, \sigma_0^2 = 2, d = 8, L = 500$ indeed achieves this bound, training to nearly diagonal weights with the appropriate scale $\langle w_{KQ}^{(ii)} \rangle \langle w_{PV}^{(ii)} \rangle = 0.327 \approx 1/3$ (similar weights are learned for each seed, up to a sign flip).

Fig. 5(a) displays how this bound is approached as the context length L of training samples is increased. In Fig. 5(b) we study how the performance of a model trained to denoise random subspaces of dimension $d = 8$ is affected by shifts in the subspace dimension at inference time. We find that when provided sufficient context, such models can adapt with mild performance loss to solve more challenging tasks not present in the training set.

It is evident from Fig. 2(a) that the softmax network performs similarly to the linear one for this task. We can understand this through the small argument expansion of the softmax function mentioned above. The learned weights displayed in Fig. 2(b) indicate that $\beta^{\text{softmax}} \approx 0.194$ becomes small (note it decreases by a factor $\epsilon \approx 0.344$ relative to β^{linear}), while the value scale $\alpha^{\text{softmax}} \approx 1.607$ becomes larger by a similar factor $\sim 1/\epsilon$ to compensate. Thus, although the optimal denoiser for this case is intuitively expressed through linear self-attention, it can also be achieved with softmax self-attention in the appropriate limit.

F.4 Case 2 - Nonlinear manifolds

Fig. 2 (case 2) shows networks of size $n = 16$ trained to denoise subspheres of dimension $d = 8$ and radius $R = 1$, with corruption $\sigma_Z^2 = 0.1$ and context length $L = 500$. Once again, the network trains to have scaled identity weights.

We note that although the network nearly achieves the optimal MSE on the test set, the weights appear at first glance to deviate slightly from the Bayes optimal predictor of Subsection 3.1, which indicated $W_{PV} = \alpha I$, $W_{KQ} = \beta I$ with $\alpha = 1, \beta = 1/\sigma_Z^2$. To better understand this, we consider a coarse-grained MSE loss landscape by scanning over α and β. See Fig. 6(a) in Appendix G. We find that the 2D loss landscape has roughly hyperbolic level sets which is suggestive of the linear attention limit, where the weight scales become constrained by their product $\alpha\beta$. Reflecting the symmetry of the problem, we also note mirrored negative solutions (i.e. one could also identify $\alpha = -1$, $\beta = -1/\sigma_Z^2$ from the analysis in Subection 3.1). Importantly, the plot shows that the trained network lies in the same valley of the loss landscape as the optimal predictor, in agreement with Fig. 2.

F.5 Case 3 - Gaussian mixtures

Figure 2 (case 3) shows networks of size $n = 16$ trained to denoise balanced Gaussian mixtures with $p = 8$ components that have isotropic variance $\sigma_0^2 = 0.02$ and centers randomly placed on the unit sphere in $\mathbb{R}^n$. The corruption magnitude is $\sigma_Z^2 = 0.1$ and context length is $L = 500$. The baselines show the zero predictor (dashed grey line) as well as the optimum from Proposition (4) (pink) and its $\sigma_0^2 \to 0$ approximation Eq. (A.6) (grey).

The trained weights qualitatively approach the optimal estimator for the zero-variance limit but with a slightly different scaling: while the scale of W_{PV} is $\alpha \approx 1$, the W_{KQ} scale is $\beta \approx 5.127 < 1/\sigma_Z^2$. To study this, we provide a corresponding plot of the 2D loss landscape in Fig. 6(a) in Appendix G. While the symmetry of the previous case has been broken (the context cluster centers $\{\mu_a\}$ will not satisfy $\langle \mu \rangle = 0$), we again find that the trained network lies in the anticipated global valley of the MSE loss landscape.

G MSE Loss landscape for scaled identity weights

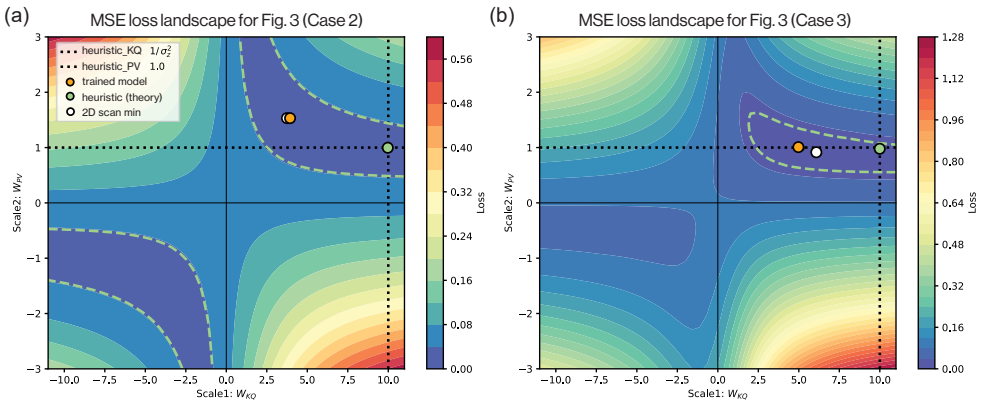


Figure 6: Loss landscape corresponding to case 2 and case 3 of Fig. 2. The MSE is numerically evaluated by assuming scaled identity weights $W_{KQ} = \beta I_n$ (x-axis) and $W_{PV} = \alpha I_n$ (y-axis) and scanning over a 50×50 grid. The green point corresponds to the heuristic minimizer identified from the posterior mean. In case 2 it is exact, while in case 3 it is an approximation that neglects the residual term (see Proposition 4). The orange point corresponds to the learned weights displayed in Fig. 2(b), while the white point corresponds to the numerically identified minimum from this 2D scan. These can fluctuate due to the finite context ($L = 500$) and sampling ($N = 800$ here). In both panels, it is apparent that the trained weights and the heuristic estimator co-occur in a broad valley (contour) of the loss landscape.

H Additional comments on the mapping from attention to associative memory models

H.1 Linear attention and traditional Hopfield model

We have considered a trained network with linear attention, relating the query $\tilde{X}$ and the estimate of the target $\hat{X}$, of the form

$$\hat{X} = f(\tilde{X}) := \frac{\gamma}{L} \sum_{t=1}^{L} X_t \langle X_t, \tilde{X} \rangle \tag{A.20}$$

with $\gamma = \frac{1}{\sigma_0^2 + \sigma_Z^2}$.

With

$$\mathcal{E}(X_{1:L}, s) := \frac{1}{2\gamma}\|s\|^2 - \frac{1}{2L}s^T(\sum_{t=1}^{L} X_t X_t^T)s \tag{A.21}$$

gradient descent iteration $s(t+1) = s(t) - \gamma\, \nabla_s \mathcal{E}\big(X_{1:L}, s(t)\big)$ gives us

$$s(t+1) = \frac{\gamma}{L}\sum_t X_t \langle X_t, s(t)\rangle$$

making the one-step iteration our denoising operation.

We will call this energy function the Naive Spherical Hopfield model for the following reason. For random memory patterns $X_{1:L}$, and the query denoting Ising spins $s \in \{-1, 1\}^n$, the so-called Hopfield energy is

$$\mathcal{E}_{\text{Hopfield}}(X_{1:L}, s) := -\frac{1}{2L}s^T(\sum_{t=1}^{L} X_t X_t^T)s. \tag{A.22}$$

We could relax the Ising nature of the spins by letting $s \in \mathbb{R}^n$, with a constraint $||s||^2 = n$. This is the spherical model (Fischer & Hertz, 1993) since the spin vector s lives on a sphere. If we minimize this energy the optimal s would be aligned with the dominant eigenvector of the matrix $\frac{1}{L}(\sum_{t=1}^{L} X_t X_t^T)$ (Fischer & Hertz, 1993), and the model will not have a retrieval phase (see Bollé et al. (2003) for a similar model that does). A soft-constrained variant can also be found in Section 3.3, Model C of Krotov & Hopfield (2021).

We could reformulate the optimization problem of minimizing the Hopfield energy, subject to $||s||^2 = R^2$, as

$$\arg\min_{s\in\mathbb{R}^n}\Big[\max_{\lambda}\big\{-\frac{1}{2L}s^T(\sum_{t=1}^{L} X_t X_t^T)s + \lambda(s^T s - R^2)\big\}\Big].$$

The s-dependent part of the Lagrangian, with λ replaced by $\frac{1}{2\gamma}$ gives us the energy function in Eq. A.21 which we have called the Naive Spherical Hopfield model.

$$\mathcal{E}(X_{1:L}, s) := \frac{1}{2\gamma}\|s\|^2 - \frac{1}{2L}s^T(\sum_{t=1}^{L} X_t X_t^T)s = \frac{1}{2}s^T\Big[(\sigma_0^2 + \sigma_Z^2)I_n - \frac{1}{L}(\sum_{t=1}^{L} X_t X_t^T)\Big]s. \tag{A.23}$$

For L much larger than n, $\frac{1}{L}\sum_{t=1}^{L} X_t X_t^T \approx \sigma_0^2 P$, so its eigenvalues are either 0 or are very close to σ_0^2. Hence, for large L and $\sigma_Z > 0$, this quadratic function is very likely to be positive definite. One-step gradient descent brings s down to the d-dimensional linear subspace S spanned by the patterns, but repeated gradient descent steps would take s towards zero.

H.2 Remarks on the softmax attention case (mapping to dense associative memory networks)

Regarding the mapping discussed in the main text, we note that there is a symmetry condition on the weights W_{KQ}, W_{PV} that is necessary for the softmax update to be interpreted as a gradient descent (i.e. a conservative flow). In general, a flow $ds/dt = f(s)$ is conservative if it can be written as the gradient of a potential, i.e. $f(s) = \nabla_s V(s)$ for some V. For this to hold, the Jacobian of the dynamics $J_f(s) = \nabla_s f$ must be symmetric.

Let $z(s) = X^T W_{KQ}\, s$ and $g(s) = \text{softmax}(z(s))$. Then the Jacobian of the softmax layer presented in the main text is

$$J(s) = W_{PV} X \frac{\partial g}{\partial s} = W_{PV} X \left(\text{diag}(g) - gg^T\right) X^T W_{KQ}. \tag{A.24}$$

Observe that $Y = X\left(\text{diag}(g) - gg^T\right)X^T$ is symmetric (keeping in mind that $g(s)$ depends on W_{KQ}). The Jacobian symmetry requirement $J = J^T$ therefore places a type of adjoint constraint on feasible W_{KQ}, W_{PV}:

$$W_{PV}\, Y\, W_{KQ}^T = W_{KQ}\, Y\, W_{PV}^T. \tag{A.25}$$

It is clear that this condition holds for the scaled identity attention weights discussed in the main text. Potentially, it could allow for more general weights that might arise from non-isotropic denoising tasks to be cast as gradient descent updates.

The mapping discussed in the main text involves discrete gradient descent steps, Eq. (8). In general, this update rule retains a "residual" term in $s(t)$ if we choose a different descent step size $\gamma \neq \alpha$. Thus, taking K recurrent updates could be viewed as the depthwise propagation of query updates through a K-layer architecture if one were to use tied weights. Analogous residual streams are commonly utilized in more elaborate transformer architectures to help propagate information to downstream attention heads.