Hopfield Networks is All You Need

Hubert Ramsauer[*] **Bernhard Schäfl**[*] **Johannes Lehner**[*]
Philipp Seidl[*] **Michael Widrich**[*]
Lukas Gruber[*] **Markus Holzleitner**[*]
Milena Pavlović[‡,§] **Geir Kjetil Sandve**[§] **Victor Greiff**[‡]
David Kreil[†] **Michael Kopp**[†]
Günter Klambauer[*] **Johannes Brandstetter**[*] **Sepp Hochreiter**[*,†]
[*]ELLIS Unit Linz and LIT AI Lab,
Institute for Machine Learning,
Johannes Kepler University Linz, Austria
[†]Institute of Advanced Research in Artificial Intelligence (IARAI)
[‡]Department of Immunology, University of Oslo, Norway
[§]Department of Informatics, University of Oslo, Norway

Abstract

The transformer and BERT models pushed the performance on NLP tasks to new levels via their attention mechanism. We show that this attention mechanism is the update rule of a modern Hopfield network with continuous states. This new Hopfield network can store exponentially (with the dimension) many patterns, converges with one update, and has exponentially small retrieval errors. The number of stored patterns must be traded off against convergence speed and retrieval error. The new Hopfield network has three types of energy minima (fixed points of the update): (1) global fixed point averaging over all patterns, (2) metastable states averaging over a subset of patterns, and (3) fixed points which store a single pattern. Transformers learn an attention mechanism by constructing an embedding of patterns and queries into an associative space. Transformer and BERT models operate in their first layers preferably in the global averaging regime, while they operate in higher layers in metastable states. The gradient in transformers is maximal in the regime of metastable states, is uniformly distributed when averaging globally, and vanishes when a fixed point is near a stored pattern. Based on the Hopfield network interpretation, we analyzed learning of transformer and BERT architectures. Learning starts with attention heads that average and then most of them switch to metastable states. However, the majority of heads in the first layers still averages and can be replaced by averaging operations like the Gaussian weighting that we propose. In contrast, heads in the last layers steadily learn and seem to use metastable states to collect information created in lower layers. These heads seem to be a promising target for improving transformers. Neural networks that integrate Hopfield networks, that are equivalent to attention heads, outperform other methods on immune repertoire classification, where the Hopfield net stores several hundreds of thousands of patterns. We provide a new PyTorch layer called "Hopfield" which allows to equip deep learning architectures with modern Hopfield networks as new powerful concept comprising pooling, memory, and attention. The implementation is available at: https://github.com/ml-jku/hopfield-layers

Introduction

The deep learning community has been looking for alternatives to recurrent neural networks (RNNs) for storing information. For example, linear memory networks use a linear autoencoder for sequences as a memory [16]. Additional memories for RNNs like holographic reduced representations [20] and classical associative memories [5, 6] have been suggested. The latter were generalized to learned matrices [79]. However, most approaches to new memories are based on attention. The neural Turing machine (NTM) is equipped with an external memory and an attention process [31]. Memory networks [69] use an $\arg\max$ attention by first mapping a query and patterns into a space, then computing scores like dot products between them, and finally retrieving the pattern with the largest dot product (score). End to end memory networks (EMN) make this attention scheme differentiable by replacing $\arg\max$ through a $\mathrm{softmax}$ [58, 59]. EMN with dot products became very popular and implement a key-value attention [21] for self-attention. An enhancement of EMN is the transformer [64, 65] and its extensions [22]. The transformer has had a great impact on the natural language processing (NLP) community as new records in NLP benchmarks have been achieved [64, 65]. MEMO uses the transformer attention mechanism for reasoning over longer distances [8]. The current state-of-the-art for language processing is a transformer architecture called "Bidirectional Encoder Representations from Transformers" (BERT) [24, 25].

We suggest using modern Hopfield networks to store information in neural networks. Binary Hopfield networks [37] seem to be an ancient technique, however, new energy functions improved the properties of Hopfield networks. The stability of spurious states or metastable states was sensibly reduced [9]. The largest and most impactful successes are reported on increasing the storage capacity of Hopfield networks. In a d-dimensional space, the standard Hopfield model can store d uncorrelated patterns without errors but only $Cd/\log(d)$ random patterns with $C < 1/2$ for a fixed stable pattern or $C < 1/4$ if all patterns are stable [45]. The same bound holds for nonlinear learning rules [44]. Using tricks-of-trade and allowing small retrieval errors, the storage capacity is about $0.138d$ [19, 33, 63]. If the learning rule is not related to the Hebb rule then up to d patterns can be stored [1]. Using a Hopfield network with non-zero diagonal matrices, the storage can be increased to $Cd\log(d)$ [28]. In contrast to the storage capacity, the number of energy minima (spurious states, stable states) of Hopfield networks is exponentially in d [61, 13, 66].

The standard binary Hopfield network has an energy function that can be expressed as the sum of interaction functions F with $F(x) = x^2$. Modern Hopfield networks called "dense associative memory" (DAM) models use an energy function with interaction functions of the form $F(x) = x^n$ and, thereby, achieve a storage capacity proportional to d^{n-1} [41, 42]. The energy function of DAMs makes them robust against adversarial attacks [42]. Modern binary Hopfield networks with energy functions based on interaction functions of the form $F(x) = \exp(x)$ even lead to storage capacity of $2^{d/2}$, where all stored binary patterns are fixed points but the radius of attraction vanishes [23].

In this publication we generalize the modern Hopfield networks with exponential interaction functions [23] to continuous patterns and states and obtain a new Hopfield network. We propose a new update rule for the new Hopfield network with continuous states. The new update rule ensures global convergence to stationary points of the energy (local minima or saddle points). We prove that our new Hopfield networks converge in one update step with exponentially low error and have storage capacity proportional to $c^{\frac{d-1}{4}}$, e.g. for $c = 1.37$ or $c = 3.15$. Surprisingly, our new update rule is also the key-value attention $\mathrm{softmax}$-update as used in the transformer and BERT. Using these insights, we modify transformer and BERT architectures to make them more efficient in learning and to obtain higher performances. In our companion paper [70], we propose and experimentally test a novel multiple instance learning method that is based on our new Hopfield network. The method is applied to immune repertoire classification, which is characterized by an unprecedentedly high number of instances per object. Our method outperforms all competitors in terms of predictive power in a large comparative study. In the experimental section we briefly report some outcomes of this study.

New Energy and Update Rule for Continuous State Modern Hopfield Nets

Overview: From binary modern Hopfield networks to the transformer. In the following we propose a new energy function that is a modification of the energy of modern Hopfield networks [23] to allow for continuous states. Consequently the new modern Hopfield networks can store

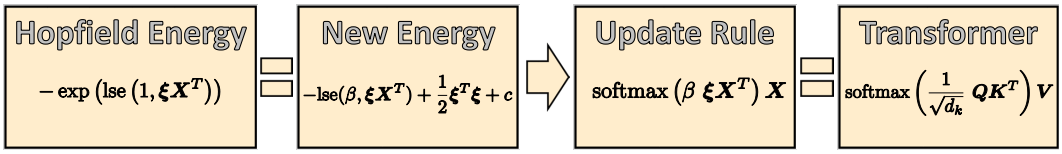

Figure 1: We generalized the energy of binary modern Hopfield networks for allowing continuous states while keeping convergence and storage capacity properties. We defined for the new energy also a new update rule that minimizes the energy. The new update rule is the attention mechanism of the transformer. Formulae are modified to express softmax as row vector as for transformers. "="-sign means "keeps the properties".

continuous patterns while keeping convergence and storage capacity properties of binary modern Hopfield networks. For the new energy we just take the logarithm of the negative energy of modern Hopfield networks and add a quadratic term of the current state. The quadratic term ensures that the norm of the state vector $\boldsymbol{\xi}$ remains finite and the energy is bounded. Classical Hopfield networks do not require to bound the norm of their state vector, since it is binary and has fixed length. We also propose a new update rule which can be proven to converge to stationary points of the energy (local minima or saddle points). We proof that a pattern that is well separated from other patterns can be retrieved with one update step and with an exponentially small error. Further we proof that our new Hopfield network has a storage capacity proportional to $c^{\frac{d-1}{4}}$, e.g. $c = 1.37$ or $c = 3.15$, for random patterns on the sphere. Most importantly, our new update rule is the attention mechanism of the transformer. Fig. 1 depicts the relation between binary modern Hopfield networks, our new Hopfield network with continuous states, our new update rule, and the transformer.

We have N patterns $\boldsymbol{x}_i \in \mathbb{R}^d$ represented by the matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ with the largest pattern $M = \max_i \|\boldsymbol{x}_i\|$. The query is $\boldsymbol{\xi} \in \mathbb{R}^d$. We need the *log-sum-exp function* (lse)

$$\text{lse}(\beta, \boldsymbol{x}) \ = \ \beta^{-1} \log \left(\sum_{i=1}^{N} \exp(\beta x_i) \right) \ , \tag{1}$$

which is convex (for the definition of the lse see appendix Eq. (A448), for convexity of the lse see appendix Lemma A22). The energy function E in the new type of Hopfield models of Krotov and Hopfield is $\text{E} = -\sum_{i=1}^{N} F\left(\boldsymbol{\xi}^T \boldsymbol{x}_i\right)$ for binary patterns $\boldsymbol{x}_i$ and binary state $\boldsymbol{\xi}$ with interaction function $F(x) = x^n$, where $n = 2$ gives the classical Hopfield model [41]. The storage capacity is proportional to d^{n-1} [41]. This model was generalized by Demircigil et al. [23] to exponential interaction functions $F(x) = \exp(x)$ which gives the energy $\text{E} = -\exp(\text{lse}(1, \boldsymbol{X}^T\boldsymbol{\xi}))$. This energy leads to an exponential storage capacity of $N = 2^{d/2}$ for binary patterns. Furthermore, with a single update, the fixed point is recovered with high probability. However, this modern Hopfield network has still binary states.

We generalize this energy function to continuous-valued patterns by using the logarithm of the negative energy and adding a quadratic term which ensures that the norm of the state vector $\boldsymbol{\xi}$ remains finite. We show that these modifications keep the properties of the modern Hopfield networks of exponential storage capacity and convergence after one update (see Fig. 1). We define the novel energy function E as

$$\text{E} \ = \ - \text{lse}(\beta, \boldsymbol{X}^T\boldsymbol{\xi}) \ + \ \frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\xi} \ + \ \beta^{-1} \log N \ + \ \frac{1}{2} M^2 \ . \tag{2}$$

We have $0 \leqslant \text{E} \leqslant 2M^2$ (see appendix Lemma A1). Using $\boldsymbol{p} = \text{softmax}(\beta \boldsymbol{X}^T\boldsymbol{\xi})$, we define a novel update rule (see Fig. 1):

$$\boldsymbol{\xi}^{\text{new}} \ = \ f(\boldsymbol{\xi}) \ = \ \boldsymbol{X}\boldsymbol{p} \ = \ \boldsymbol{X}\text{softmax}(\beta \boldsymbol{X}^T\boldsymbol{\xi}) \ . \tag{3}$$

The next theorem states that the update rule Eq. (3) converges globally.

Theorem 1. *The update rule Eq. (3) converges globally: For $\boldsymbol{\xi}^{t+1} = f(\boldsymbol{\xi}^t)$, the energy $\text{E}(\boldsymbol{\xi}^t) \rightarrow \text{E}(\boldsymbol{\xi}^*)$ for $t \rightarrow \infty$ and a fixed point $\boldsymbol{\xi}^*$.*

Proof. The update rule in Eq. (3) is the Concave-Convex Procedure (CCCP) [74, 75] for minimizing the energy E, which is the sum of the convex $1/2\boldsymbol{\xi}^T\boldsymbol{\xi}$ and concave $-\text{lse}$ (see details in appendix

Theorem 1). Theorem 2 in [74] states the global convergence of the theorem. Also, in Theorem 2 in [57] the global convergence of CCCP is proven via a rigorous analysis using Zangwill's global convergence theory of iterative algorithms. □

CCCP is equivalent to Legendre minimization [53, 54] algorithms [75].

The global convergence theorem only assures that for the energy $\mathrm{E}(\boldsymbol{\xi}^t) \to \mathrm{E}(\boldsymbol{\xi}^*)$ for $t \to \infty$ but not $\boldsymbol{\xi}^t \to \boldsymbol{\xi}^*$. The next theorem strengthens Zangwill's global convergence theorem [47] and gives convergence results similar to those known for expectation maximization [72].

Theorem 2. *For the iteration Eq. (3) we have* $\mathrm{E}\left(\boldsymbol{\xi}^t\right) \to \mathrm{E}\left(\boldsymbol{\xi}^*\right) = \mathrm{E}^*$ *as* $t \to \infty$*, for some stationary point* $\boldsymbol{\xi}^*$*. Furthermore,* $\left\|\boldsymbol{\xi}^{t+1} - \boldsymbol{\xi}^t\right\| \to 0$ *and either* $\{\boldsymbol{\xi}^t\}_{t=0}^{\infty}$ *converges or, in the other case, the set of limit points of* $\{\boldsymbol{\xi}^t\}_{t=0}^{\infty}$ *is a connected and compact subset of* $\mathcal{L}\left(\mathrm{E}^*\right)$*, where* $\mathcal{L}\left(a\right) = \{\boldsymbol{\xi} \in \mathcal{L} \mid \mathrm{E}\left(\boldsymbol{\xi}\right) = a\}$ *and* $\mathcal{L}$ *is the set of stationary points of the iteration Eq. (3). If* $\mathcal{L}\left(\mathrm{E}^*\right)$ *is finite, then any sequence* $\{\boldsymbol{\xi}^t\}_{t=0}^{\infty}$ *generated by the iteration Eq. (3) converges to some* $\boldsymbol{\xi}^* \in \mathcal{L}\left(\mathrm{E}^*\right)$*.*

See proof in appendix Theorem 2. Therefore, all the limit points of any sequence generated by the iteration Eq. (3) are the stationary points (local minima or saddle points) of the energy function E. Either the iteration converges or, in the second case, the set of limit points is a connected and compact set.

Next theorem gives the results on the storage capacity of our new continuous state modern Hopfield network. We first define what we mean by storing and retrieving patterns using a modern Hopfield network with continuous states.

Definition 1 (Pattern Stored and Retrieved)**.** *We assume that around every pattern* $\boldsymbol{x}_i$ *a sphere* S_i *is given. We say* $\boldsymbol{x}_i$ is stored *if there is a single fixed point* $\boldsymbol{x}_i^* \in \mathrm{S}_i$ *to which all points* $\boldsymbol{\xi} \in \mathrm{S}_i$ *converge, and* $\mathrm{S}_i \cap \mathrm{S}_j = \emptyset$ *for* $i \neq j$*. We say* $\boldsymbol{x}_i$ is retrieved *if the iteration Eq. (3) converged to the single fixed point* $\boldsymbol{x}_i^* \in \mathrm{S}_i$*. The retrieval error is* $\|\boldsymbol{x}_i - \boldsymbol{x}_i^*\|$*.*

As with classical Hopfield networks, we consider patterns on the sphere, i.e. patterns with a fixed norm. For randomly chosen patterns, the number of patterns that can be stored is exponential in the dimension d of the space of the patterns ($\boldsymbol{x}_i \in \mathbb{R}^d$).

Theorem 3. *We assume a failure probability* $0 < p \leqslant 1$ *and randomly chosen patterns on the sphere with radius* $M = K\sqrt{d-1}$*. We define* $a := \frac{2}{d-1}\left(1 + \ln(2\,\beta\,K^2\,p\,(d-1))\right)$*,* $b := \frac{2\,K^2\,\beta}{5}$*, and* $c = \frac{b}{W_0(\exp(a + \ln(b)))}$*, where* W_0 *is the upper branch of the Lambert W function and ensure* $c \geq \left(\frac{2}{\sqrt{p}}\right)^{\frac{4}{d-1}}$*. Then with probability* $1 - p$*, the number of random patterns that can be stored is*

$$N \geq \sqrt{p}\, c^{\frac{d-1}{4}} \ . \tag{4}$$

Examples are $c \geq 3.1546$ *for* $\beta = 1$*,* $K = 3$*,* $d = 20$ *and* $p = 0.001$ *(*$a + \ln(b) > 1.27$*) and* $c \geq 1.3718$ *for* $\beta = 1$*,* $K = 1$*,* $d = 75$*, and* $p = 0.001$ *(*$a + \ln(b) < -0.94$*).*

For a proof, see appendix Theorem A5.

The next theorem states that the update rule typically converges after one update if the patterns are well separated. First we need the concept of separation of a pattern. For pattern $\boldsymbol{x}_i$ we define its separation Δ_i to other patterns by:

$$\Delta_i := \min_{j, j \neq i} \left(\boldsymbol{x}_i^T \boldsymbol{x}_i - \boldsymbol{x}_i^T \boldsymbol{x}_j\right) = \boldsymbol{x}_i^T \boldsymbol{x}_i - \max_{j, j \neq i} \boldsymbol{x}_i^T \boldsymbol{x}_j \ . \tag{5}$$

The update rule converges after one update for well separated patterns:

Theorem 4. *With query* $\boldsymbol{\xi}$*, after one update the distance of the new point* $f(\boldsymbol{\xi})$ *to the fixed point* $\boldsymbol{x}_i^*$ *is exponentially small in the separation* Δ_i*. The precise bounds using the Jacobian* $\mathrm{J} = \frac{\partial f(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}$ *and its value* J^m *in the mean value theorem are:*

$$\|f(\boldsymbol{\xi}) - \boldsymbol{x}_i^*\| \leqslant \|\mathrm{J}^m\|_2 \, \|\boldsymbol{\xi} - \boldsymbol{x}_i^*\| \ , \tag{6}$$

$$\|\mathrm{J}^m\|_2 \leqslant 2\,\beta\,N\,M^2\,(N-1)\exp(-\beta\,(\Delta_i - 2\,\max\{\|\boldsymbol{\xi} - \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* - \boldsymbol{x}_i\|\}\,M)) \ . \tag{7}$$

See proof in appendix Theorem A8.

At the same time, the retrieval error decreases exponentially with the separation Δ_i.

Theorem 5. *The retrieval error* $\|\boldsymbol{x}_i - \boldsymbol{x}_i^*\|$ *of pattern* $\boldsymbol{x}_i$ *is bounded by*

$$\|\boldsymbol{x}_i - \boldsymbol{x}_i^*\| \leqslant 2\,(N-1)\,\exp(-\beta\,(\Delta_i - 2\,\|\boldsymbol{x}_i^* - \boldsymbol{x}_i\|\,M))\,M \tag{8}$$

and for $\|\boldsymbol{x}_i^* - \boldsymbol{x}_i\| \leqslant \frac{1}{\beta\,N\,M}$

$$\|\boldsymbol{x}_i - \boldsymbol{x}_i^*\| \leqslant 2\,(N-1)\,\exp(-\beta\,(\Delta_i - \frac{2}{\beta\,N}))\,M\,. \tag{9}$$

See proof in appendix Theorem A8.

Metastable states and one global fixed point. Above we considered the case where the patterns $\boldsymbol{x}_i$ are well separated, then the iterate converges to a fixed point which is near a pattern $\boldsymbol{x}_i$. If the patterns $\boldsymbol{x}_i$ are not well separated, the iterate converges to a global fixed point close to the arithmetic mean of the vectors. In this case the softmax vector $\boldsymbol{p}$ is close to uniform, that is, $p_i = 1/N$. If some vectors are similar to each other and well separated from all other vectors, then a metastable state near the similar vectors exists. Iterates that start near the metastable state converge to this metastable state (also if initialized by one of the similar patterns). For convergence proofs to one global fixed point and to metastable states see appendix Lemma A7 and Lemma A12, respectively.

Hopfield update rule is attention of the transformer. The Hopfield network update rule is the attention mechanism used in the transformer and BERT (see Fig. 1). To see this, we assume patterns $\boldsymbol{y}_i$ that are mapped to the Hopfield space of dimension d_k. We set $\boldsymbol{x}_i = \boldsymbol{W}_K^T \boldsymbol{y}_i$, $\boldsymbol{\xi}_i = \boldsymbol{W}_Q^T \boldsymbol{y}_i$, and multiply the result of our update rule with $\boldsymbol{W}_V$. The matrix $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)^T$ combines the $\boldsymbol{y}_i$ as row vectors. We define the matrices $\boldsymbol{X}^T = \boldsymbol{K} = \boldsymbol{Y}\boldsymbol{W}_K$, $\boldsymbol{Q} = \boldsymbol{Y}\boldsymbol{W}_Q$, and $\boldsymbol{V} = \boldsymbol{Y}\boldsymbol{W}_K\boldsymbol{W}_V = \boldsymbol{X}^T\boldsymbol{W}_V$, where $\boldsymbol{W}_K \in \mathbb{R}^{d_y \times d_k}$, $\boldsymbol{W}_Q \in \mathbb{R}^{d_y \times d_k}$, $\boldsymbol{W}_V \in \mathbb{R}^{d_k \times d_v}$. For combining all queries in matrix $\boldsymbol{Q}$, $\beta = 1/\sqrt{d_k}$, and softmax $\in \mathbb{R}^N$ changed to a row vector, we obtain for the update rule Eq. (3) multiplied by $\boldsymbol{W}_V$:

$$\text{softmax}\left(1/\sqrt{d_k}\,\boldsymbol{Q}\,\boldsymbol{K}^T\right)\,\boldsymbol{V}\,. \tag{10}$$

This formula is the transformer attention.

Analysis of transformer and BERT models

Our theoretical analysis of the attention mechanisms suggests three fixed point dynamics: a) If the patterns $\boldsymbol{x}_i$ are not well separated, the iterate goes to a fixed point close to the arithmetic mean of the vectors, *a global fixed point*. In this case, the softmax vector $\boldsymbol{p}$ is close to uniform ($p_i = 1/N$). b) If the patterns are well separated from each other, the iterate goes close to a pattern. If the initial $\boldsymbol{\xi}$ is similar to a pattern $\boldsymbol{x}_i$ then it will converge to a vector close to $\boldsymbol{x}_i$ and $\boldsymbol{p}$ will converge to a vector close to $\boldsymbol{e}_i$, which we call *a fixed point close to a single pattern*. c) If some vectors are similar to each other but well separated from all other vectors, then a so called *metastable state* between the similar vectors exists. Iterates that start near the metastable state converge to this metastable state.

Operating classes of transformer and BERT models. We observed that transformer and BERT models have attention heads that have all three kinds of fixed points (a)–(c) (see Fig. 2). A global fixed point of kind (a) and metastable states of kind (c) with many patterns have a similar dynamics: slow convergence and averaging over many patterns. Fixed points near single patterns of kind (b) and metastable states of kind (c) with few patterns also have similar dynamics: convergence after one update and averaging over few patterns. Therefore, we only consider metastable states, for which fixed points (a) and (b) are the extreme cases. We categorize the metastable states into four classes: (I) averaging over a very large number of patterns (very large metastable state or global fixed point), (II) averaging over a large number of patterns (large metastable state), (III) averaging over a medium number of patterns (medium metastable state), (IV) averaging over a small number of patterns (small metastable state or fixed point close to a single pattern). To investigate in which of the four classes the heads of BERT are predominately working, for each token in each head and layer, we calculated

Figure 2: Analysis of operating modes of the heads of a pre-trained BERT model. For each head in each layer, the distribution of the minimal number k of patterns required to sum up the $\mathrm{softmax}$ values to 0.90 is displayed as a violin plot in a panel. k indicates the size of a metastable state. The bold number in the center of each panel gives the median $\bar{k}$ of the distribution. The heads in each layer are sorted according to $\bar{k}$. Attention heads belong to the class they mainly operate in. **Class (IV) in blue:** Small metastable state or fixed point close to a single pattern, which is abundant in the middle layers (6, 7, and 8). **Class (II) in orange:** Large metastable state, which is prominent in middle layers (3, 4, and 5). **Class (I) in red:** Very large metastable state or global fixed point, which is predominant in the first layer. These heads can potentially be replaced by averaging operations. **Class (III) in green:** Medium metastable state, which is frequently observed in higher layers. We hypothesize that these heads are used to collect information required to perform the respective task. These heads should be the main target to improve transformer and BERT models.

the minimal number k of $\mathrm{softmax}$ values required to sum up to 0.90. Hence, k indicates the size of a metastable state. We computed the distribution of k across sequences processed by BERT models. Then we classified a head as mainly operating in one of the four classes. Concretely, for N tokens and for $\bar{k}$ as the median of the distribution, a head is classified as operating in class (I) if $1/2N < \bar{k}$, as operating in class (II) if $1/8N < \bar{k} \leqslant 1/2N$, as operating in class (III) if $1/32N < \bar{k} \leqslant 1/8N$, and as operating in class (IV) if $\bar{k} \leqslant 1/32N$. We analyzed pre-trained BERT models from Hugging Face Inc.[71] according to these operating classes. In Fig. 2 the distributions of the pre-trained *bert-base-cased* model is depicted (for plots of other models see appendix Section B1.3). Operating classes (II) (large metastable states) and (IV) (small metastable states or fixed point close to a single pattern) are often observed in the middle layers. Operating class (I) (averaging over a very large number of patterns) is abundant in lower layers. Operating class (III) (medium metastable states) is predominant in the last layers.

Attention heads in lower layers perform averaging. Figure 2 shows that many heads in the lower layers operate in class (I) (averaging over a very large number of patterns). Similar observations have been reported in other studies [62]. We performed analyses to investigate to which extent class (I) averaging is the predominant operating mode of heads in the respective layers of a pre-trained BERT. For one layer at a time, we forced the heads to compute the arithmetic mean by setting $\beta = 0$, which yields a $\mathrm{softmax}$ with $p_i = 1/N$. All other layers were left unchanged. The modified model was evaluated on ≈500 sequences to assess the difference in the loss with respect to the original model. In Figure 3d, the perplexity is plotted against the number of the modified layers (layer 1 is the lowest). The red line indicates the perplexity of the original model. We found that the performance is less affected in lower layers than in higher layers, and almost not affected in the first layer. This suggests that the heads in the first layer can be replaced by non-attention based averaging operations.

Motivated by our analyses, we replaced the heads in the first layer by a Gaussian weighting, where the mean and the variance of the Gaussians can be learned. Thus, we learned a positional Gaussian weighted averaging scheme with only two parameters per token instead of attention heads. Therefore, the heads perform always the same averaging independent of the input. Our Gaussian weighting is similar to the *Random Synthesizer* head, where the attention weights are learned directly [62]. We pre-trained a BERT-small model as specified in [18] on the *masked LM* task from the original publication [24] but omitted the *next sentence prediction* task. Like in the original publication [24] pre-training was done on the BookCorpus [80] and on English Wikipedia (details in appendix Section B1.1). As can be seen in Fig. 3 averaging in the first layer does not decrease the performance substantially. Further, it does not considerably change the learning dynamics of the model. Thus, we replaced heads with many weights in W_Q and W_K by Gaussians with only two parameters, therefore making BERT models more efficient without sacrificing performance.

Attention heads in the last layers operate in class (III) (medium metastable states) and seem to be important for BERT models. We found that attention heads in the last layers operate mainly in class (III) as seen in Fig. 2 (layer 10, 11, and 12). To investigate these layers, we replaced their attention heads by averaging operations as before (see previous paragraph). In contrast to replacing attention heads in the first layers and in the middle layers, the performance of the model dropped much more (see Fig. 3 in d). To further investigate the operating modes during learning in the last layers, we performed another experiment with the BERT-small from ELECTRA [18] that has twelve layers and four heads each. We trained a BERT-small model and saw that during the initial learning phase from 0 to 9,000 updates, attention heads operate in the averaging mode class (I). At around 9,000 updates, however, a strong decrease in the loss function (Fig. 3a) blue line) appears, which coincides with attention heads switching to other classes. See more details in Fig. B2(a) in the appendix. Most heads switch to class (II) or class (IV) and do no longer change. In contrast, the heads in the last layers still learn after the drop (Fig. 3c) and move more towards class (III). This switching behaviour is shown in Fig. 3, where in b) a head in layer 6 is shown that is committed after the drop of the error while the head in c) from layer 12 still learns. The gradients with respect to the weights are determined by the Jacobian J as can be seen in the appendix in Eq. (A409), Eq. (A420), and Eq. (A426). Fig. B2(b) in the appendix shows that the gradient in transformers is maximal in the regime of metastable states that is class (II) and (III), is uniformly distributed when averaging globally that is class (I), and vanishes when a fixed point is near a stored pattern that is class (IV). To summarize, the heads in the last layer operating in class (III) are presumably important in the BERT model, because they cannot be replaced by averaging operations and they still learn after the main drop in the loss curve. We hypothesize that these heads are used to collect information required

to perform the respective task. Therefore, the heads in the last layer should be the main target to improve BERT models.

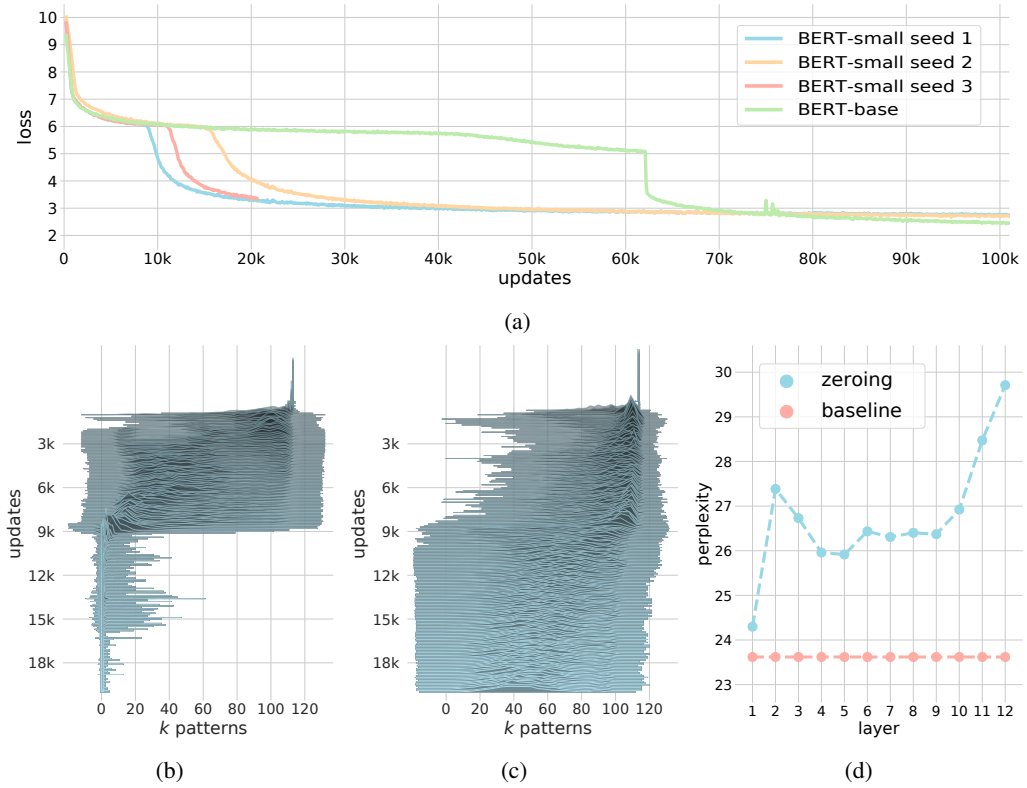

Figure 3: Learning dynamics of BERT models wrt. the operating mode of their attention heads (for details see appendix, Section B1.3 **a):** Learning curves of different BERT models. **b)** and **c):** Density plots of the minimal number k of $\mathrm{softmax}$ values required to sum up to 0.9 during learning. Shown are head 2 in layer 6 (b) and head 3 in layer 12 (c) of the BERT-small model trained with seed 1 (blue line in loss curves **a)**). Distributions with peaks at the right side indicate attention heads that operate in class (I) while distributions at the left side operate in class (IV). A strong change of the operating mode from class (I) to other classes occurs at update step 9,000 coinciding with a large drop in the error. **d)** Change in perplexity on the masked language modeling task, evaluated on WikiText-103 [46]. The values are obtained by zeroing out the attention weights W_Q and W_K in each respective layer. From subfigures (c) and (d) it can be hypothesized that heads in the last layer still learn and move more to operating class (III), while the other heads are already committed.

Layer normalization is highly relevant to adjust the most important parameters of the modern Hopfield network. We already identified β as a crucial parameter for the fixed point dynamics, and hence the operating mode of the attention heads (see above). In the appendix, e.g. in Lemma A7 or in Eq. A91 and Eq.A92, we showed that the characteristics of the fixed points of the new modern Hopfield network are determined by: β, M (maximal pattern norm), $m_{\max}$ (spread of the similar patterns), and $\|\boldsymbol{m}_{\boldsymbol{x}}\|$ (center of the similar patterns). Low values of β induce global averaging and higher values of β metastable states. Adjusting either β or M is equivalent to adjusting the (inverse) gain parameter of the layer normalization [5]. Furthermore, layer normalization can also move the center $\|\boldsymbol{m}_{\boldsymbol{x}}\|$ of the similar patterns. Thus, layer normalization controls the most important parameters of the Hopfield network: β, M,$m_{\max}$, and $\|\boldsymbol{m}_{\boldsymbol{x}}\|$. In experiments we switched off the layer normalization, which lead to unstable learning and we could not stabilize learning properly.

Experiments

Modern Hopfield networks for massive multiple instance learning and immune repertoire classification. We briefly summarize results from our companion paper [70], in which a new method based on our new Hopfield network is proposed. The new method relies on Theorem 3, which states that modern Hopfield networks possess an exponential storage capacity. This property enables to tackle massive multiple instance learning (MIL) problems [26] like immune repertoire classification [27]. This classification typically requires to extract a single or few patterns from a large set of sequences, the repertoire, that are indicative for the respective immune status. The new method in [70] is applied to this challenging task at which most MIL methods fail due the large number of instances.

In [70] we use experimentally observed immune receptor as well as simulated sequences, into which sequence motifs [3, 68] with low yet varying degrees of frequency are implanted. Four different categories of datasets are constructed: (a) Simulated immunosequencing data with implanted motifs, (b) immunosequencing data generated by long short-term memory (LSTM) with implanted motifs, (c) real-world immunosequencing data with implanted motifs, and (d) real-world immunosequencing data with known immune status [27]. Categories (a), (b), and (d) contain approx. 300,000 instances per immune repertoire. With over 30 billion sequences in total, this represents one of the largest multiple instance learning experiments ever conducted [15] (for details see [70], appendix, Section A2).Despite the massive number of instances as well as the low frequency of sequences indicative of the respective immune status, our novel deep learning architecture with modern Hopfield networks outperforms all competing methods with respect to average area under the ROC curve in all four categories, (a), (b), (c) and (d) (for details see [70], appendix, Section A2).

New Hopfield layer in PyTorch. We provide a PyTorch implementation of a new layer called "Hopfield" which allows to equip deep learning architectures with Hopfield networks as novel memory concepts. The Hopfield layer enables to associate two sets of vectors. This general functionality allows for transformer-like self-attention, for decoder-encoder attention, for time series prediction (maybe with positional encoding), for sequence analysis, for multiple instance learning, for learning with point sets, for combining data sources by associations, for constructing a memory, for averaging and pooling operations, and for many more. In particular, the new Hopfield layer can readily be used as plug-in replacement for existing layers like pooling layers (max-pooling or average pooling), permutation equivariant layers [32, 55], GRU [17] & LSTM [34, 35] layers, and attention layers [64, 65, 7]. The Hopfield layer is based on modern Hopfield networks with continuous states that have very high storage capacity and converge after one update. Code for our new Hopfield layer is provided in this github repo. For more details see appendix, Section C.

Conclusion. We have shown that the attention mechanism of transformers is equivalent to the update rule of a modern Hopfield network with continuous states. Furthermore, we have demonstrated that this new Hopfield network can store exponentially many patterns, converges with one update, and has exponentially small retrieval errors. Based on the theoretically identified fixed point dynamics and types of metastable states, we were able to analyze attention heads of BERT models according to their operation modes. We found that heads in the first layer mainly average and can be replaced by averaging operations like our Gaussian weighing scheme with only two parameters. Further we hypothesize that heads in the last layer are used to collect information produced in the lower layers. These heads are promising targets to improve BERT models. We successfully apply modern Hopfield networks to a real-world application, a massive multiple instance learning in computational biology.

Broader Impact

Impact on ML and related scientific fields. We envision, that the theoretical insights gained from this connection of modern Hopfield networks with the currently popular transformer architectures and the resulting improvements could lead to (a) further popularization and increased successful application of transformer-like methods, (b) continuing investigation and improvements that arise from this new theoretical foundation, and (c) more efficient and theoretically founded applications and analysis of transformer methods, possibly extending to new datasets and tasks.

Given the potential increase in efficiency of NLP applications, the automated usage of text-based databases could lead to a change in how datasets are acquired for the training of ML models, possibly opening up a tremendous source of knowledge to ML methods.

Impact on society. If the theoretical insights and resulting improvements provided in this paper prove to be successful in other practical applications, the usage of transformer-like methods could become more efficient and suitable for more challenging tasks. Given that the current transformer-like methods already have a noticeable effect on society in NLP applications, this effect might be amplified by the increased efficiency resulting from our analysis. Here the potential to understand and parse large amounts of text-based databases with increased efficiency could become an important factor for (a) research and development, e.g. for parsing the large amount of publications currently available for related work, (b) health-care, specifically treatment and diagnosis, e.g. enabling a doctor to match large databases for symptoms, treatments, and side-effect within seconds, and (c) communication and acquisition of information, e.g. more efficient language translation models which could foster the communication, exchange, and understanding between people of different languages but thereby also open up larger pools of information for individuals.

A potential negative effect in terms of more efficient NLP applications could be the creation of fake text, such as fake news articles, and the improvement of chatbots. Problems arise if such fake texts are met with a lack of knowledge or ignorance of the reader. Here possible solutions could be better education of the people, which could, in turn, result in an increased importance of verified texts, such as verified newspaper articles, and user accounts (see also next paragraph).

Consequences of failures of the method. Failures of the method might result in incorrect information provided to users, which could amplify information bubbles, offensive chatbots, or – in a medical context – incorrect information retrieved from biomedical literature, which in the worst case might lead to inappropriate decisions on therapies. One might also consider incorrect translations by a language model as a type of failure. Similarly, abstractive summarization methods can provide biased information. And for both methods, a failure can lead to the loss of important information [56].

Leveraging of biases in the data and potential discrimination. In terms of increased efficiency in NLP tasks, the naive large-scale mining of text-based data might be more susceptible to the biases introduced by humans in such non-curated datasets. As such, preexisting biases might be amplified.

Similarly, individuals which are more susceptible to manipulation by artificially generated texts and information bubbles, e.g. due to a lack of knowledge about the state of such methods, would be put at a further disadvantage w.r.t. being exploited for commercial, political, or other gains.

Acknowledgments

The ELLIS Unit Linz, the LIT AI Lab and the Institute for Machine Learning are supported by the Land Oberösterreich, LIT grants DeepToxGen (LIT-2017-3-YOU-003), and AI-SNN (LIT-2018-6-YOU-214), the Medical Cognitive Computing Center (MC3), Janssen Pharmaceutica, UCB Biopharma, Merck Group, Audi.JKU Deep Learning Center, Audi Electronic Venture GmbH, TGW, Primal, Silicon Austria Labs (SAL), FILL, EnliteAI, Google Brain, ZF Friedrichshafen AG, Robert Bosch GmbH, TÜV Austria, DCS, and the NVIDIA Corporation. Victor Greiff (VG) and Geir Kjetil Sandve (GKS) are supported by The Helmsley Charitable Trust (#2019PG-T1D011, to VG), UiO World-Leading Research Community (to VG), UiO:LifeSciences Convergence Environment Immunolingo (to VG and GKS), EU Horizon 2020 iReceptorplus (#825821, to VG) and Stiftelsen Kristian Gerhard Jebsen (K.G. Jebsen Coeliac Disease Research Centre, to GKS). IARAI is supported by Here Technologies.

Appendix

This appendix to the paper "Hopfield networks is all you need" consists of Section A, Section B, and Section C. Section A introduces the new modern Hopfield network with continuous states and its update rule. Furthermore, Section A provides a thorough and profound theoretical analysis of this new Hopfield network. Section B gives details on the experiments related to the new Hopfield network. Transformer and BERT architectures are investigated since they implement the Hopfield update rule via their attention heads. In Section C, a PyTorch implementation of a Hopfield layer is described, which allows a readily integration in any feed-forward or recurrent deep learning architecture (e.g. to construct Transformer architectures or sequence analysis networks). The Hopfield layer can be used to implement or to substitute (i) pooling layers, (ii) permutation equivariant layers, (iii) GRU & LSTM layers, (iv) self-attention cross-attention layers, and (v) layers operating on sets.

Contents of the appendix

List of theorems

List of definitions

List of figures

Continuous State Modern Hopfield Networks

Sepp Hochreiter Markus Holzleitner Lukas Gruber Hubert Ramsauer Günter Klambauer Johannes Brandstetter

A1 Introduction

In Section A2 our new modern Hopfield network is introduced. In Subsection A2.1 we present the new energy function. Then in Subsection A2.2, our new update rule is introduced. In Subsection A2.3, we show that this update rule ensures global convergence. We show that all the limit points of any sequence generated by the update rule are the stationary points (local minima or saddle points) of the energy function. In Section A2.4, we consider the local convergence of the update rule and see that it converges after one update. In Subsection A2.5, we consider the properties of the fixed points that are associated with the stored patterns. In Subsection A2.5.1, we show that exponentially many patterns can be stored. The main result is given in Theorem A5: For random patterns on a sphere we can store and retrieve exponentially (in the dimension of the Hopfield space) many patterns. Subsection A2.5.2 reports that the update converges after one update step and that the retrieval error is exponentially small.

In Subsection A2.6, we consider how associations for the new Hopfield networks can be learned. In Subsection A2.6.2, we analyze if the association is learned directly by a bilinear form. In Subsection A2.6.3, we analyze if stored patterns and query patterns are mapped to the space of the Hopfield network. Therefore, we treat the architecture of the transformer and BERT. In Subsection A2.7, we introduce a temporal component into the new Hopfield network that leads to a forgetting behavior. The forgetting allows us to treat infinite memory capacity in Subsection A2.7.1. In Subsection A2.7.2, we consider the controlled forgetting behavior.

In Section A3, we provide the mathematical background that is needed for our proofs. In particular we give lemmas on properties of the softmax, the log-sum-exponential, the Legendre transform, and the Lambert W function.

In Section A4, we review the new Hopfield network as introduced by Krotov and Hopfield in 2016. However in contrast to our new Hopfield network, the Hopfield network of Krotov and Hopfield is binary, that is, a network with binary states. In Subsection A4.1, we give an introduction to neural networks equipped with associative memories and new Hopfield networks. In Subsection A4.1.1, we discuss neural networks that are enhanced by an additional external memory and by attention mechanisms. In Subsection A4.1.2, we give an overview over the modern Hopfield networks. Finally, in Subsection A4.2, we present the energy function and the update rule for the modern, binary Hopfield networks.

A2 Modern Hopfield Networks: Continuous States (New Concept)

A2.1 New Energy Function

We have patterns $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ that are represented by the matrix

$$\boldsymbol{X} \ = \ (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \ . \tag{A1}$$

The largest norm of a pattern is

$$M \ = \ \max_i \| \boldsymbol{x}_i \| \ . \tag{A2}$$

The query or state of the Hopfield network is $\boldsymbol{\xi}$.

The energy function E in the new type of Hopfield models of Krotov and Hopfield is $\mathrm{E} = -\sum_{i=1}^{N} F\left(\boldsymbol{\xi}^T \boldsymbol{x}_i\right)$ for binary patterns $\boldsymbol{x}_i$ and binary state $\boldsymbol{\xi}$ with interaction function $F(x) = x^n$, where $n = 2$ gives classical Hopfield model [41]. The storage capacity is proportional to d^{n-1} [41]. This model was generalized by Demircigil et al. [23] to exponential interaction functions $F(x) = \exp(x)$, which gives the energy $\mathrm{E} = -\exp(\mathrm{lse}(1, \boldsymbol{X}^T \boldsymbol{\xi}))$. This energy leads to an exponential storage capacity of $N = 2^{d/2}$ for binary patterns. Furthermore, with a single update the fixed point is recovered with high probability. See more details in Section A4.

In contrast to the these binary modern Hopfield networks, we focus on modern Hopfield networks with *continuous state*s that can store *continuous patterns*. We generalize the energy of Demircigil et al. [23] to continuous states while keeping the lse properties which ensure high storage capacity and fast convergence. Our new energy E for a continuous query or state $\boldsymbol{\xi}$ is defined as

$$\mathrm{E} \ = \ - \ \mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) \ + \ \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \ + \ \beta^{-1} \ln N \ + \ \frac{1}{2} M^2 \tag{A3}$$

$$= \ - \ \beta^{-1} \ln \left(\sum_{i=1}^{N} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}) \right) \ + \ \beta^{-1} \ln N \ + \ \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \ + \ \frac{1}{2} M^2 \ . \tag{A4}$$

First let us collect and prove some properties of E. The next lemma gives bounds on the energy E.

Lemma A1. *The energy* E *is larger than zero:*

$$0 \ \leqslant \ \mathrm{E} \ . \tag{A5}$$

For $\boldsymbol{\xi}$ in the simplex defined by the patterns, the energy E *is upper bounded by:*

$$\mathrm{E} \ \leqslant \ \beta^{-1} \ln N \ + \ \frac{1}{2} \ M^2 \ , \tag{A6}$$

$$\mathrm{E} \ \leqslant \ 2 \ M^2 \ . \tag{A7}$$

Proof. We start by deriving the lower bound of zero. The pattern most similar to query or state $\boldsymbol{\xi}$ is $\boldsymbol{x}_{\boldsymbol{\xi}}$:

$$\boldsymbol{x}_{\boldsymbol{\xi}} \ = \ \boldsymbol{x}_k \ , \quad k \ = \ \arg\max_i \boldsymbol{\xi}^T \boldsymbol{x}_i \ . \tag{A8}$$

We obtain

$$
\begin{aligned}
\mathrm{E} \;&=\; -\,\beta^{-1} \ln \left(\sum_{i=1}^{N} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}) \right) \;+\; \beta^{-1} \ln N \;+\; \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \frac{1}{2}\, M^2 \\
&=\; -\,\beta^{-1} \ln \left(\frac{1}{N} \sum_{i=1}^{N} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}) \right) \;+\; \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \frac{1}{2}\, M^2 \\
&\geq\; -\,\beta^{-1} \ln \left(\frac{1}{N} \sum_{i=1}^{N} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}) \right) \;+\; \frac{1}{2}\, \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \frac{1}{2}\, \boldsymbol{x}_{\boldsymbol{\xi}}^T \boldsymbol{x}_{\boldsymbol{\xi}} \\
&\geq\; -\,\beta^{-1} \ln \left(\exp(\beta \boldsymbol{x}_{\boldsymbol{\xi}}^T \boldsymbol{\xi}) \right) \;+\; \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \frac{1}{2}\, \boldsymbol{x}_{\boldsymbol{\xi}}^T \boldsymbol{x}_{\boldsymbol{\xi}} \\
&=\; -\,\boldsymbol{x}_{\boldsymbol{\xi}}^T \boldsymbol{\xi} \;+\; \frac{1}{2}\, \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \frac{1}{2}\, \boldsymbol{x}_{\boldsymbol{\xi}}^T \boldsymbol{x}_{\boldsymbol{\xi}} \\
&=\; \frac{1}{2}\, \left(\boldsymbol{\xi} \;-\; \boldsymbol{x}_{\boldsymbol{\xi}} \right)^T \left(\boldsymbol{\xi} \;-\; \boldsymbol{x}_{\boldsymbol{\xi}} \right) \;=\; \frac{1}{2}\, \left\| \boldsymbol{\xi} \;-\; \boldsymbol{x}_{\boldsymbol{\xi}} \right\|^2 \;\geq\; 0 \,.
\end{aligned}
\tag{A9}
$$

The energy is zero and, therefore, the bound attained, if all $\boldsymbol{x}_i$ are equal, that is, $\boldsymbol{x}_i = \boldsymbol{x}$ for all i and $\boldsymbol{\xi} = \boldsymbol{x}$.

For deriving upper bounds on the energy E, we require the the query $\boldsymbol{\xi}$ to be in the simplex defined by the patterns, that is,

$$
\boldsymbol{\xi} \;=\; \sum_{i=1}^{N} p_i \, \boldsymbol{x}_i \,, \quad \sum_{i=1}^{N} p_i \;=\; 1 \,, \quad \forall_i : \; 0 \;\leqslant\; p_i \,. \tag{A10}
$$

The first upper bound is.

$$
\begin{aligned}
\mathrm{E} \;&=\; -\,\beta^{-1} \ln \left(\sum_{i=1}^{N} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}) \right) \;+\; \frac{1}{2}\, \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \beta^{-1} \ln N \;+\; \frac{1}{2}\, M^2 \\
&\leqslant\; -\sum_{i=1}^{N} p_i \, (\boldsymbol{x}_i^T \boldsymbol{\xi}) \;+\; \frac{1}{2}\, \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \beta^{-1} \ln N \;+\; \frac{1}{2}\, M^2 \\
&=\; -\,\frac{1}{2}\, \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \beta^{-1} \ln N \;+\; \frac{1}{2}\, M^2 \;\leqslant\; \beta^{-1} \ln N \;+\; \frac{1}{2}\, M^2 \,.
\end{aligned}
\tag{A11}
$$

For the first inequality we applied Lemma A19 to $-\mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi})$ with $\boldsymbol{z} = \boldsymbol{p}$ giving

$$
-\,\mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) \;\leqslant\; -\sum_{i=1}^{N} p_i \, (\boldsymbol{x}_i^T \boldsymbol{\xi}) \;+\; \beta^{-1} \sum_{i=1}^{N} p_i \ln p_i \;\leqslant\; -\sum_{i=1}^{N} p_i \, (\boldsymbol{x}_i^T \boldsymbol{\xi}) \,, \tag{A12}
$$

as the term involving the logarithm is non-positive.

Next we derive the second upper bound, for which we need the mean $\boldsymbol{m}_{\boldsymbol{x}}$ of the patterns

$$
\boldsymbol{m}_{\boldsymbol{x}} \;=\; \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i \,. \tag{A13}
$$

We obtain

$$
\begin{aligned}
\mathrm{E} \;&=\; -\,\beta^{-1} \ln \left(\sum_{i=1}^{N} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}) \right) \;+\; \frac{1}{2}\, \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \beta^{-1} \ln N \;+\; \frac{1}{2}\, M^2 \\
&\leqslant\; -\sum_{i=1}^{N} \frac{1}{N}\, \boldsymbol{x}_i^T \boldsymbol{\xi} \;+\; \frac{1}{2}\, \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \frac{1}{2}\, M^2 \\
&=\; -\,\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{\xi} \;+\; \frac{1}{2}\, \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \frac{1}{2}\, M^2 \\
&\leqslant\; \left\| \boldsymbol{m}_{\boldsymbol{x}} \right\| \left\| \boldsymbol{\xi} \right\| \;+\; \frac{1}{2}\, \left\| \boldsymbol{\xi} \right\|^2 \;+\; \frac{1}{2}\, M^2 \\
&\leqslant\; 2\, M^2 \,,
\end{aligned}
\tag{A14}
$$

where for the first inequality we again applied Lemma A19 with $\boldsymbol{z} = (1/N, \ldots, 1/N)$ and $\beta^{-1} \sum_i 1/N \ln(1/N) = -\beta^{-1} \ln(N)$. This inequality also follows from Jensen's inequality. The second inequality uses the Cauchy-Schwarz inequality. The last inequality uses

$$\|\boldsymbol{\xi}\| \;=\; \left\| \sum_i p_i \, \boldsymbol{x}_i \right\| \;\leqslant\; \sum_i p_i \, \|\boldsymbol{x}_i\| \;\leqslant\; \sum_i p_i M \;=\; M \tag{A15}$$

and

$$\|\boldsymbol{m}_{\boldsymbol{x}}\| \;=\; \left\| \sum_i (1/N) \, \boldsymbol{x}_i \right\| \;\leqslant\; \sum_i (1/N) \, \|\boldsymbol{x}_i\| \;\leqslant\; \sum_i (1/N) \, M \;=\; M \, . \tag{A16}$$

□

A2.2 New Update Rule

We now introduce an update rule for minimizing the energy function E. The new update rule is

$$\boldsymbol{\xi}^{\text{new}} \;=\; \boldsymbol{X}\boldsymbol{p} \;=\; \boldsymbol{X} \text{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) \, , \tag{A17}$$

where we used

$$\boldsymbol{p} \;=\; \text{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) \, . \tag{A18}$$

The new state $\boldsymbol{\xi}^{\text{new}}$ is in the simplex defined by the patterns, no matter what the previous state $\boldsymbol{\xi}$ was. For comparison, the synchronous update rule for the classical Hopfield network with threshold zero is

$$\boldsymbol{\xi}^{\text{new}} \;=\; \text{sgn}\left(\boldsymbol{X}\boldsymbol{X}^T \boldsymbol{\xi}\right) . \tag{A19}$$

Therefore, instead of using the vector $\boldsymbol{X}^T \boldsymbol{\xi}$ as in the classical Hopfield network, its softmax version $\text{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})$ is used.

In the next section (Section A2.3) we show that the update rule Eq. (A17) ensures global convergence. We show that all the limit points of any sequence generated by the update rule are the stationary points (local minima or saddle points) of the energy function E. In Section A2.4 we consider the local convergence of the update rule Eq. (A17) and see that it converges after one update.

A2.3 Global Convergence of the Update Rule

We are interested in the *global convergence*, that is, convergence from each initial point, of the iterate

$$\boldsymbol{\xi}^{\text{new}} \;=\; f(\boldsymbol{\xi}) \;=\; \boldsymbol{X}\boldsymbol{p} \;=\; \boldsymbol{X} \text{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) \, , \tag{A20}$$

where we used

$$\boldsymbol{p} \;=\; \text{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) \, . \tag{A21}$$

We defined the energy function

$$\text{E} \;=\; -\, \text{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) \;+\; \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \beta^{-1} \ln N \;+\; \frac{1}{2} M^2 \tag{A22}$$

$$\;=\; -\, \beta^{-1} \ln \left(\sum_{i=1}^{N} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}) \right) \;+\; \beta^{-1} \ln N \;+\; \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \frac{1}{2} M^2 \, . \tag{A23}$$

We will show that the update rule in Eq. (A20) is the Concave-Convex Procedure (CCCP) for minimizing the energy E. The CCCP is proven to converge globally.

Theorem A1 (Global Convergence (Zangwill): Energy)**.** *The update rule Eq. (A20) converges globally: For $\boldsymbol{\xi}^{t+1} = f(\boldsymbol{\xi}^t)$, the energy $\text{E}(\boldsymbol{\xi}^t) \to \text{E}(\boldsymbol{\xi}^*)$ for $t \to \infty$ and a fixed point $\boldsymbol{\xi}^*$.*

Proof. The Concave-Convex Procedure (CCCP) [74, 75] minimizes a function that is the sum of a concave function and a convex function. CCCP is equivalent to Legendre minimization [53, 54] algorithms [75]. The Jacobian of the softmax is positive semi-definite according to Lemma A22. The Jacobian of the softmax is the Hessian of the lse, therefore lse is a convex and $-$lse a concave function.

Therefore, the energy function $\mathrm{E}(\boldsymbol{\xi})$ is the sum of the convex function $\mathrm{E}_1(\boldsymbol{\xi}) = 1/2\boldsymbol{\xi}^T\boldsymbol{\xi} + C_1$ and the concave function $\mathrm{E}_2(\boldsymbol{\xi}) = -\mathrm{lse}$:

$$\mathrm{E}(\boldsymbol{\xi}) \;=\; \mathrm{E}_1(\boldsymbol{\xi}) \;+\; \mathrm{E}_2(\boldsymbol{\xi})\;, \tag{A24}$$

$$\mathrm{E}_1(\boldsymbol{\xi}) \;=\; \frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\xi} \;+\; \beta^{-1}\ln N \;+\; \frac{1}{2}M^2 \;=\; \frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\xi} \;+\; C_1\;, \tag{A25}$$

$$\mathrm{E}_2(\boldsymbol{\xi}) \;=\; -\,\mathrm{lse}(\beta, \boldsymbol{X}^T\boldsymbol{\xi})\;, \tag{A26}$$

where C_1 does not depend on $\boldsymbol{\xi}$.

The Concave-Convex Procedure (CCCP) [74, 75] applied to E is

$$\nabla_\xi \mathrm{E}_1\left(\boldsymbol{\xi}^{t+1}\right) \;=\; -\,\nabla_\xi \mathrm{E}_2\left(\boldsymbol{\xi}^{t}\right)\;, \tag{A27}$$

which is

$$\nabla_\xi\left(\frac{1}{2}\left(\boldsymbol{\xi}^{t+1}\right)^T\boldsymbol{\xi}^{t+1} \;+\; C_1\right) \;=\; \nabla_\xi \mathrm{lse}(\beta, \boldsymbol{X}^T\boldsymbol{\xi}^t)\;. \tag{A28}$$

The resulting update rule is

$$\boldsymbol{\xi}^{t+1} \;=\; \boldsymbol{X}\boldsymbol{p}^t \;=\; \boldsymbol{X}\mathrm{softmax}(\beta\boldsymbol{X}^T\boldsymbol{\xi}^t) \tag{A29}$$

using

$$\boldsymbol{p}^t \;=\; \mathrm{softmax}(\beta\boldsymbol{X}^T\boldsymbol{\xi}^t)\;. \tag{A30}$$

This is the update rule in Eq. (A20).

Theorem 2 in [74] and Theorem 2 in [75] state that the update rule Eq. (A20) is guaranteed to monotonically decrease the energy E as a function of time. See also Theorem 2 in [57]. □

Although the objective converges in all cases, it does not necessarily converge to a local minimum [43].

However the convergence proof of CCCP in [74, 75] was not as rigorous as required. In [57] a rigorous analysis of the convergence of CCCP is performed using Zangwill's global convergence theory of iterative algorithms.

In [57] the minimization problem

$$\begin{aligned} &\min_{\boldsymbol{\xi}} \;\mathrm{E}_1 \;+\; \mathrm{E}_2 \\ &\text{s.t.} \quad \boldsymbol{c}(\boldsymbol{\xi}) \leqslant \boldsymbol{0}\;, \quad \boldsymbol{d}(\boldsymbol{\xi}) \;=\; \boldsymbol{0} \end{aligned} \tag{A31}$$

is considered with E_1 convex, $-\mathrm{E}_2$ convex, $\boldsymbol{c}$ component-wise convex function, and $\boldsymbol{d}$ an affine function. The CCCP algorithm solves this minimization problem by linearization of the concave part and is defined in [57] as

$$\begin{aligned} \boldsymbol{\xi}^{t+1} \;\in\; \arg\min_{\boldsymbol{\xi}} \;&\mathrm{E}_1\left(\boldsymbol{\xi}\right) \;+\; \boldsymbol{\xi}^T\nabla_\xi \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) \\ &\text{s.t.} \quad \boldsymbol{c}(\boldsymbol{\xi}) \leqslant \boldsymbol{0}\;, \quad \boldsymbol{d}(\boldsymbol{\xi}) \;=\; \boldsymbol{0}\;. \end{aligned} \tag{A32}$$

We define the upper bound E_C on the energy:

$$\mathrm{E}_\mathrm{C}\left(\boldsymbol{\xi}, \boldsymbol{\xi}^t\right) \;:=\; \mathrm{E}_1\left(\boldsymbol{\xi}\right) \;+\; \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) \;+\; \left(\boldsymbol{\xi} \;-\; \boldsymbol{\xi}^t\right)^T\nabla_\xi \mathrm{E}_2\left(\boldsymbol{\xi}^t\right)\;. \tag{A33}$$

E_C is equal to the energy $\mathrm{E}\left(\boldsymbol{\xi}^t\right)$ for $\boldsymbol{\xi} = \boldsymbol{\xi}^t$:

$$\mathrm{E}_\mathrm{C}\left(\boldsymbol{\xi}^t, \boldsymbol{\xi}^t\right) \;=\; \mathrm{E}_1\left(\boldsymbol{\xi}^t\right) \;+\; \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) \;=\; \mathrm{E}\left(\boldsymbol{\xi}^t\right)\;. \tag{A34}$$

Since $-\mathrm{E}_2$ is convex, the first order characterization of convexity holds (Eq. 3.2 in [11]):

$$-\,\mathrm{E}_2\left(\boldsymbol{\xi}\right) \;\geq\; -\,\mathrm{E}_2\left(\boldsymbol{\xi}^t\right) \;-\; \left(\boldsymbol{\xi} \;-\; \boldsymbol{\xi}^t\right)^T\nabla_\xi \mathrm{E}_2\left(\boldsymbol{\xi}^t\right)\;, \tag{A35}$$

that is

$$\mathrm{E}_2\left(\boldsymbol{\xi}\right) \;\leqslant\; \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) \;+\; \left(\boldsymbol{\xi} \;-\; \boldsymbol{\xi}^t\right)^T\nabla_\xi \mathrm{E}_2\left(\boldsymbol{\xi}^t\right)\;. \tag{A36}$$

Therefore, for $\boldsymbol{\xi} \neq \boldsymbol{\xi}^t$ the function $\mathrm{E_C}$ is an upper bound on the energy:

$$\begin{aligned} \mathrm{E}\left(\boldsymbol{\xi}\right) \leqslant \mathrm{E_C}\left(\boldsymbol{\xi}, \boldsymbol{\xi}^t\right) &= \mathrm{E}_1\left(\boldsymbol{\xi}\right) + \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) + \left(\boldsymbol{\xi} - \boldsymbol{\xi}^t\right)^T \nabla_\xi \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) \\ &= \mathrm{E}_1\left(\boldsymbol{\xi}\right) + \boldsymbol{\xi}^T \nabla_\xi \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) + C_2 \,, \end{aligned} \tag{A37}$$

where C_2 does not depend on $\boldsymbol{\xi}$. Since we do not have constraints, $\boldsymbol{\xi}^{t+1}$ is defined as

$$\boldsymbol{\xi}^{t+1} \in \arg\min_{\boldsymbol{\xi}} \mathrm{E_C}\left(\boldsymbol{\xi}, \boldsymbol{\xi}^t\right) \,, \tag{A38}$$

hence $\mathrm{E_C}\left(\boldsymbol{\xi}^{t+1}, \boldsymbol{\xi}^t\right) \leqslant \mathrm{E_C}\left(\boldsymbol{\xi}^t, \boldsymbol{\xi}^t\right)$. Combining the inequalities gives:

$$\mathrm{E}\left(\boldsymbol{\xi}^{t+1}\right) \leqslant \mathrm{E_C}\left(\boldsymbol{\xi}^{t+1}, \boldsymbol{\xi}^t\right) \leqslant \mathrm{E_C}\left(\boldsymbol{\xi}^t, \boldsymbol{\xi}^t\right) = \mathrm{E}\left(\boldsymbol{\xi}^t\right) \,. \tag{A39}$$

Since we do not have constraints, $\boldsymbol{\xi}^{t+1}$ is the minimum of

$$\mathrm{E_C}\left(\boldsymbol{\xi}, \boldsymbol{\xi}^t\right) = \mathrm{E}_1\left(\boldsymbol{\xi}\right) + \boldsymbol{\xi}^T \nabla_\xi \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) + C_2 \tag{A40}$$

as a function of $\boldsymbol{\xi}$.

For a minimum not at the border, the derivative has to be the zero vector

$$\frac{\partial \mathrm{E_C}\left(\boldsymbol{\xi}, \boldsymbol{\xi}^t\right)}{\partial \boldsymbol{\xi}} = \boldsymbol{\xi} + \nabla_\xi \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) = \boldsymbol{\xi} - \boldsymbol{X}\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}^t) = \mathbf{0} \tag{A41}$$

and the Hessian must be positive semi-definite

$$\frac{\partial^2 \mathrm{E_C}\left(\boldsymbol{\xi}, \boldsymbol{\xi}^t\right)}{\partial \boldsymbol{\xi}^2} = \boldsymbol{I} \,. \tag{A42}$$

The Hessian is strict positive definite everywhere, therefore the optimization problem is strict convex (if the domain is convex) and there exist only one minimum, which is a global minimum. $\mathrm{E_C}$ can even be written as a quadratic form:

$$\mathrm{E_C}\left(\boldsymbol{\xi}, \boldsymbol{\xi}^t\right) = \frac{1}{2}\left(\boldsymbol{\xi} + \nabla_\xi \mathrm{E}_2\left(\boldsymbol{\xi}^t\right)\right)^T \left(\boldsymbol{\xi} + \nabla_\xi \mathrm{E}_2\left(\boldsymbol{\xi}^t\right)\right) + C_3 \,, \tag{A43}$$

where C_3 does not depend on $\boldsymbol{\xi}$.

Therefore, the minimum is

$$\boldsymbol{\xi}^{t+1} = -\nabla_\xi \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) = \boldsymbol{X}\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}^t) \tag{A44}$$

if it is in the domain as we assume.

Using $M = \max_i \|\boldsymbol{x}_i\|$, $\boldsymbol{\xi}^{t+1}$ is in the sphere $\mathrm{S} = \{\boldsymbol{x} \mid \|\boldsymbol{x}\| \leqslant M\}$ which is a convex and compact set. Hence, if $\boldsymbol{\xi}^0 \in \mathrm{S}$, then the iterate is a mapping from S to S. Therefore, the point-set-map defined by the iteration Eq. (A44) is uniformly compact on S according to Remark 7 in [57]. Theorem 2 and Theorem 4 in [57] states that all the limit points of the iteration Eq. (A44) are stationary points. These theorems follow from Zangwill's global convergence theorem: Convergence Theorem A, page 91 in [77] and page 3 in [72].

The global convergence theorem only assures that for the sequence $\boldsymbol{\xi}^{t+1} = f(\boldsymbol{\xi}^t)$ and a function Φ we have $\Phi(\boldsymbol{\xi}^t) \to \Phi(\boldsymbol{\xi}^*)$ for $t \to \infty$ but not $\boldsymbol{\xi}^t \to \boldsymbol{\xi}^*$. However, if f is strictly monotone with respect to Φ, then we can strengthen Zangwill's global convergence theorem [47]. We set $\Phi = \mathrm{E}$ and show $\mathrm{E}(\boldsymbol{\xi}^{t+1}) < \mathrm{E}(\boldsymbol{\xi}^t)$ if $\boldsymbol{\xi}^t$ is not a stationary point of E, that is, f is strictly monotone with respect to E. The following theorem is similar to the convergence results for the expectation maximization (EM) algorithm in [72] which are given in theorems 1 to 6 in [72]. The following theorem is also very similar to Theorem 8 in [57].

Theorem A2 (Global Convergence: Stationary Points)**.** *For the iteration Eq.* (A44) *we have* $\mathrm{E}\left(\boldsymbol{\xi}^t\right) \to \mathrm{E}\left(\boldsymbol{\xi}^*\right) = \mathrm{E}^*$ *as* $t \to \infty$*, for some stationary point* $\boldsymbol{\xi}^*$*. Furthermore* $\left\|\boldsymbol{\xi}^{t+1} - \boldsymbol{\xi}^t\right\| \to 0$ *and either* $\{\boldsymbol{\xi}^t\}_{t=0}^\infty$ *converges or, in the other case, the set of limit points of* $\{\boldsymbol{\xi}^t\}_{t=0}^\infty$ *is a connected and compact subset of* $\mathcal{L}\left(\mathrm{E}^*\right)$*, where* $\mathcal{L}\left(a\right) = \{\boldsymbol{\xi} \in \mathcal{L} \mid \mathrm{E}\left(\boldsymbol{\xi}\right) = a\}$ *and* $\mathcal{L}$ *is the set of stationary points of the iteration Eq.* (A44)*. If* $\mathcal{L}\left(\mathrm{E}^*\right)$ *is finite, then any sequence* $\{\boldsymbol{\xi}^t\}_{t=0}^\infty$ *generated by the iteration Eq.* (A44) *converges to some* $\boldsymbol{\xi}^* \in \mathcal{L}\left(\mathrm{E}^*\right)$*.*

Proof. We have $\mathrm{E}\left(\boldsymbol{\xi}^t\right) = \mathrm{E}_1\left(\boldsymbol{\xi}^t\right) + \mathrm{E}_2\left(\boldsymbol{\xi}^t\right)$. The gradient $\nabla_\xi \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) = -\nabla_\xi \mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi})$ is continuous. Therefore, Eq. (A40) has minimum in the sphere S, which is a convex and compact set. If $\boldsymbol{\xi}^{t+1} \neq \boldsymbol{\xi}^t$, then $\boldsymbol{\xi}^t$ was not the minimum of Eq. (A37) as the derivative at $\boldsymbol{\xi}^t$ is not equal to zero. Eq. (A42) shows that the optimization problem Eq. (A37) is strict convex, hence it has only one minimum, which is a global minimum. Eq. (A43) shows that the optimization problem Eq. (A37) is even a quadratic form. Therefore, we have

$$\mathrm{E}\left(\boldsymbol{\xi}^{t+1}\right) \leqslant \mathrm{E}_\mathrm{C}\left(\boldsymbol{\xi}^{t+1}, \boldsymbol{\xi}^t\right) < \mathrm{E}_\mathrm{C}\left(\boldsymbol{\xi}^t, \boldsymbol{\xi}^t\right) = \mathrm{E}\left(\boldsymbol{\xi}^t\right) . \tag{A45}$$

Therefore, the point-set-map defined by the iteration Eq. (A44) (for definitions see [57]) is strictly monotonic with respect to E. Therefore, we can apply Theorem 3 in [57] or Theorem 3.1 and Corollary 3.2 in [47], which give the statements of the theorem.

□

We showed global convergence of the iteration Eq. (A20). We have shown that all the limit points of any sequence generated by the iteration Eq. (A20) are the stationary points (critical points; local minima or saddle points) of the energy function E. Local maxima as stationary points are only possible if the iterations exactly hits a local maximum. However, convergence to a local maximum without being there is not possible because Eq. (A45) ensures a strict decrease of the energy E. Therefore, almost sure local maxima are not obtained as stationary points. Either the iteration converges or, in the second case, the set of limit points is a connected and compact set. But what happens if $\boldsymbol{\xi}^0$ is in an ϵ-neighborhood around a local minimum $\boldsymbol{\xi}^*$? Will the iteration Eq. (A20) converge to $\boldsymbol{\xi}^*$? What is the rate of convergence? These questions are about *local convergence* which will be treated in detail in next section.

A2.4 Local Convergence of the Update Rule: Fixed Point Iteration

For the proof of local convergence to a fixed point we will apply Banach fixed point theorem. For the rate of convergence we will rely on properties of a contraction mapping.

A2.4.1 General Bound on the Jacobian of the Iterate

We consider the iteration

$$\boldsymbol{\xi}^{\mathrm{new}} = f(\boldsymbol{\xi}) = \boldsymbol{X}\boldsymbol{p} = \boldsymbol{X}\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) \tag{A46}$$

using

$$\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) . \tag{A47}$$

The Jacobian J is symmetric and has the following form:

$$\mathrm{J} = \frac{\partial f(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \beta \, \boldsymbol{X} \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right) \boldsymbol{X}^T = \boldsymbol{X} \mathrm{J}_s \boldsymbol{X}^T , \tag{A48}$$

where J_s is Jacobian of the softmax.

To analyze the local convergence of the iterate, we distinguish between the following three cases (see also Fig. A1). Here we only provide an informal discussion to give the reader some intuition. A rigorous formulation of the results can be found in the corresponding subsections.

a) If the patterns $\boldsymbol{x}_i$ are not well separated, the iterate goes to a fixed point close to the arithmetic mean of the vectors. In this case $\boldsymbol{p}$ is close to $p_i = 1/N$.

b) If the patterns $\boldsymbol{x}_i$ are well separated, then the iterate goes to the pattern to which the initial $\boldsymbol{\xi}$ is similar. If the initial $\boldsymbol{\xi}$ is similar to a vector $\boldsymbol{x}_i$ then it will converge to a vector close to $\boldsymbol{x}_i$ and $\boldsymbol{p}$ will converge to a vector close to $\boldsymbol{e}_i$.

c) If some vectors are similar to each other but well separated from all other vectors, then a so called metastable state between the similar vectors exists. Iterates that start near the metastable state converge to this metastable state.

We begin with a bound on the Jacobian of the iterate, thereby heavily relying on the Jacobian of the softmax from Lemma A24.

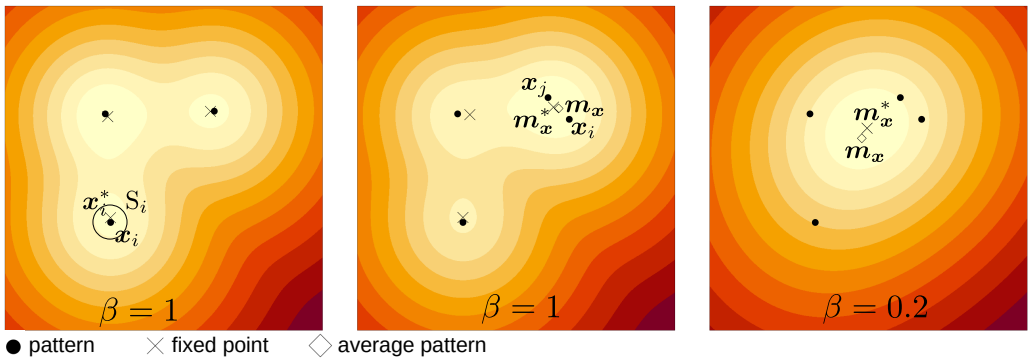

Figure A1: The three cases of fixed points. **a) Stored patterns (fixed point is single pattern)**: patterns are stored if they are well separated. Each pattern $\boldsymbol{x}_i$ has a single fixed point $\boldsymbol{x}_i^*$ close to it. In the sphere S_i, pattern $\boldsymbol{x}_i$ is the only pattern and $\boldsymbol{x}_i^*$ the only fixed point. **b) Metastable state (fixed point is average of similar patterns)**: $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are similar to each other and not well separated. The fixed point $\boldsymbol{m}_{\boldsymbol{x}}^*$ is a metastable state that is close to the mean $\boldsymbol{m}_{\boldsymbol{x}}$ of the similar patterns. **c) Global fixed point (fixed point is average of all patterns)**: no pattern is well separated from the others. A single global fixed point $\boldsymbol{m}_{\boldsymbol{x}}^*$ exists that is close to the arithmetic mean $\boldsymbol{m}_{\boldsymbol{x}}$ of all patterns.

Lemma A2. *For N patterns $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, $\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})$, $M = \max_i \|\boldsymbol{x}_i\|$, and $m = \max_i p_i (1 - p_i)$, the spectral norm of the Jacobian J of the fixed point iteration is bounded:*

$$\|\mathrm{J}\|_2 \leqslant 2 \beta \|\boldsymbol{X}\|_2^2 m \leqslant 2 \beta N M^2 m . \tag{A49}$$

If $p_{\max} = \max_i p_i \geq 1 - \epsilon$, then for the spectral norm of the Jacobian holds

$$\|\mathrm{J}\|_2 \leqslant 2 \beta N M^2 \epsilon - 2 \epsilon^2 \beta N M^2 < 2 \beta N M^2 \epsilon . \tag{A50}$$

Proof. With

$$\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) , \tag{A51}$$

the symmetric Jacobian J is

$$\mathrm{J} = \frac{\partial f(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \beta \boldsymbol{X} \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p} \boldsymbol{p}^T \right) \boldsymbol{X}^T = \boldsymbol{X} \mathrm{J}_s \boldsymbol{X}^T , \tag{A52}$$

where J_s is Jacobian of the softmax.

With $m = \max_i p_i (1 - p_i)$, Eq. (A461) from Lemma A24 is

$$\|\mathrm{J}_s\|_2 = \beta \left\| \mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p} \boldsymbol{p}^T \right\|_2 \leqslant 2 m \beta . \tag{A53}$$

Using this bound on $\|\mathrm{J}_s\|_2$, we obtain

$$\|\mathrm{J}\|_2 \leqslant \beta \left\| \boldsymbol{X}^T \right\|_2 \|\mathrm{J}_s\|_2 \|\boldsymbol{X}\|_2 \leqslant 2 m \beta \|\boldsymbol{X}\|_2^2 . \tag{A54}$$

The spectral norm $\|.\|_2$ is bounded by the Frobenius norm $\|.\|_F$ which can be expressed by the norm squared of its column vectors:

$$\|\boldsymbol{X}\|_2 \leqslant \|\boldsymbol{X}\|_F = \sqrt{\sum_i \|\boldsymbol{x}_i\|^2} . \tag{A55}$$

Therefore, we obtain the first statement of the lemma:

$$\|\mathrm{J}\|_2 \leqslant 2 \beta \|\boldsymbol{X}\|_2^2 m \leqslant 2 \beta N M^2 m . \tag{A56}$$

With $p_{\max} = \max_i p_i \geq 1 - \epsilon$ Eq. (A465) in Lemma A24 is

$$\|\mathrm{J}_s\|_2 \leqslant 2 \beta \epsilon - 2 \epsilon^2 \beta < 2 \beta \epsilon . \tag{A57}$$

Using this inequality, we obtain the second statement of the lemma:

$$\|\mathrm{J}\|_2 \leqslant 2 \beta N M^2 \epsilon - 2 \epsilon^2 \beta N M^2 < 2 \beta N M^2 \epsilon . \tag{A58}$$

□

We now define the "separation" Δ_i of a pattern $\boldsymbol{x}_i$ from data $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ here, since it has an important role for the convergence properties of the iteration.

Definition 2 (Separation of Patterns)**.** *We define Δ_i, i.e.* the separation of pattern $\boldsymbol{x}_i$ from data $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ *as:*

$$\Delta_i \ = \ \min_{j, j \neq i} \left(\boldsymbol{x}_i^T \boldsymbol{x}_i \ - \ \boldsymbol{x}_i^T \boldsymbol{x}_j \right) \ = \ \boldsymbol{x}_i^T \boldsymbol{x}_i \ - \ \max_{j, j \neq i} \boldsymbol{x}_i^T \boldsymbol{x}_j \ . \tag{A59}$$

The pattern is separated from the other data if $0 < \Delta_i$. Using the parallelogram identity, Δ_i can also be expressed as

$$\begin{aligned} \Delta_i \ &= \ \min_{j, j \neq i} \frac{1}{2} \ \left(\|\boldsymbol{x}_i\|^2 \ - \ \|\boldsymbol{x}_j\|^2 \ + \ \|\boldsymbol{x}_i \ - \ \boldsymbol{x}_j\|^2 \right) \\ &= \ \frac{1}{2} \|\boldsymbol{x}_i\|^2 \ - \ \frac{1}{2} \ \max_{j, j \neq i} \left(\|\boldsymbol{x}_j\|^2 \ - \ \|\boldsymbol{x}_i \ - \ \boldsymbol{x}_j\|^2 \right) \ . \end{aligned} \tag{A60}$$

For $\|\boldsymbol{x}_i\| = \|\boldsymbol{x}_j\|$ we have $\Delta_i = 1/2 \min_{j, j \neq i} \|\boldsymbol{x}_i \ - \ \boldsymbol{x}_j\|^2$.

Analog we say for a query $\boldsymbol{\xi}$ and data $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, that $\boldsymbol{x}_i$ is least separated from $\boldsymbol{\xi}$ while being separated from other $\boldsymbol{x}_j$ with $j \neq i$ if

$$i \ = \ \arg\max_k \min_{j, j \neq k} \left(\boldsymbol{\xi}^T \boldsymbol{x}_k \ - \ \boldsymbol{\xi}^T \boldsymbol{x}_j \right) \ = \ \arg\max_k \left(\boldsymbol{\xi}^T \boldsymbol{x}_k \ - \ \max_{j, j \neq k} \boldsymbol{\xi}^T \boldsymbol{x}_j \right) \tag{A61}$$

$$0 \ \leqslant \ c \ = \ \max_k \min_{j, j \neq k} \left(\boldsymbol{\xi}^T \boldsymbol{x}_k \ - \ \boldsymbol{\xi}^T \boldsymbol{x}_j \right) \ = \ \max_k \left(\boldsymbol{\xi}^T \boldsymbol{x}_k \ - \ \max_{j, j \neq k} \boldsymbol{\xi}^T \boldsymbol{x}_j \right) \ . \tag{A62}$$

Next we consider the case where the iteration has only one stable fixed point.

A2.4.2 One Stable State: Fixed Point Near the Mean of the Patterns

We start with the case where no pattern is well separated from the others.

Global fixed point near the global mean: Analysis using the data center. We revisit the bound on the Jacobian of the iterate by utilizing properties of pattern distributions. We begin with a probabilistic interpretation where we consider p_i as the probability of selecting the vector $\boldsymbol{x}_i$. Consequently, we define expectations as $\mathrm{E}_{\boldsymbol{p}}[f(\boldsymbol{x})] = \sum_{i=1}^N p_i f(\boldsymbol{x}_i)$. In this setting the matrix

$$\boldsymbol{X} \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T \right) \boldsymbol{X}^T \tag{A63}$$

is the covariance matrix of data $\boldsymbol{X}$ when its vectors are selected according to the probability $\boldsymbol{p}$:

$$\boldsymbol{X} \left(\mathrm{diag}(\boldsymbol{p}) \ - \ \boldsymbol{p}\boldsymbol{p}^T \right) \boldsymbol{X}^T \ = \ \boldsymbol{X} \mathrm{diag}(\boldsymbol{p}) \boldsymbol{X}^T \ - \ \boldsymbol{X}\boldsymbol{p}\boldsymbol{p}^T \boldsymbol{X}^T \tag{A64}$$

$$= \ \sum_{i=1}^N p_i \ \boldsymbol{x}_i \ \boldsymbol{x}_i^T \ - \ \left(\sum_{i=1}^N p_i \ \boldsymbol{x}_i \right) \left(\sum_{i=1}^N p_i \ \boldsymbol{x}_i \right)^T \tag{A65}$$

$$= \ \mathrm{E}_{\boldsymbol{p}}[\boldsymbol{x} \ \boldsymbol{x}^T] \ - \ \mathrm{E}_{\boldsymbol{p}}[\boldsymbol{x}] \ \mathrm{E}_{\boldsymbol{p}}[\boldsymbol{x}]^T \ = \ \mathrm{Var}_{\boldsymbol{p}}[\boldsymbol{x}] \ , \tag{A66}$$

therefore we have

$$\mathrm{J} \ = \ \beta \ \mathrm{Var}_{\boldsymbol{p}}[\boldsymbol{x}] \ . \tag{A67}$$

The largest eigenvalue of the covariance matrix (equal to the largest singular value) is the variance in the direction of the eigenvector associated with the largest eigenvalue.

We define:

$$\boldsymbol{m}_{\boldsymbol{x}} \ = \ \frac{1}{N} \ \sum_{i=1}^N \ \boldsymbol{x}_i \ , \tag{A68}$$

$$m_{\max} \ = \ \max_{1 \leqslant i \leqslant N} \|\boldsymbol{x}_i \ - \ \boldsymbol{m}_{\boldsymbol{x}}\|_2 \ . \tag{A69}$$

$\boldsymbol{m_x}$ is the arithmetic mean (the center) of the patterns. $m_{\max}$ is the maximal distance of the patterns to the center $\boldsymbol{m_x}$.

The variance of the patterns is

$$\begin{aligned} \mathrm{Var}_{\boldsymbol{p}}[\boldsymbol{x}] \ &= \ \sum_{i=1}^{N} p_i \ \boldsymbol{x}_i \ \boldsymbol{x}_i^T \ - \ \left(\sum_{i=1}^{N} p_i \ \boldsymbol{x}_i \right) \ \left(\sum_{i=1}^{N} p_i \ \boldsymbol{x}_i \right)^T \\ &= \ \sum_{i=1}^{N} p_i \ \left(\boldsymbol{x}_i \ - \ \sum_{i=1}^{N} p_i \boldsymbol{x}_i \right) \ \left(\boldsymbol{x}_i \ - \ \sum_{i=1}^{N} p_i \boldsymbol{x}_i \right)^T \ . \end{aligned} \tag{A70}$$

The maximal distance to the center $m_{\max}$ allows to derive a bound on the norm of the Jacobian.

Next lemma gives a condition for a global fixed point.

Lemma A3. *The following bound on the norm $\|\mathrm{J}\|_2$ of the Jacobian of the fixed point iteration f holds independent of $\mathbf{p}$ or the query $\boldsymbol{\xi}$.*

$$\|\mathrm{J}\|_2 \ \leqslant \ \beta \ m_{\max}^2 \ . \tag{A71}$$

For $\beta \ m_{\max}^2 < 1$ there exists a unique fixed point (global fixed point) of iteration f in each compact set.

Proof. In order to bound the variance we compute the vector $\boldsymbol{a}$ that minimizes

$$f(\boldsymbol{a}) \ = \ \sum_{i=1}^{N} p_i \|\boldsymbol{x}_i \ - \ \boldsymbol{a}\|^2 \ = \ \sum_{i=1}^{N} p_i (\boldsymbol{x}_i \ - \ \boldsymbol{a})^T (\boldsymbol{x}_i \ - \ \boldsymbol{a}) \ . \tag{A72}$$

The solution to

$$\frac{\partial f(\boldsymbol{a})}{\partial \boldsymbol{a}} \ = \ 2 \ \sum_{i=1}^{N} p_i (\boldsymbol{a} \ - \ \boldsymbol{x}_i) \ = \ 0 \tag{A73}$$

is

$$\boldsymbol{a} \ = \ \sum_{i=1}^{N} p_i \boldsymbol{x}_i \ . \tag{A74}$$

The Hessian of f is positive definite since

$$\frac{\partial^2 f(\boldsymbol{a})}{\partial \boldsymbol{a}^2} \ = \ 2 \ \sum_{i=1}^{N} p_i \ \boldsymbol{I} \ = \ 2 \ \boldsymbol{I} \tag{A75}$$

and f is a convex function. Hence, the mean

$$\bar{\boldsymbol{x}} \ := \ \sum_{i=1}^{N} p_i \ \boldsymbol{x}_i \tag{A76}$$

minimizes $\sum_{i=1}^{N} p_i \|\boldsymbol{x}_i - \boldsymbol{a}\|^2$. Therefore, we have

$$\sum_{i=1}^{N} p_i \|\boldsymbol{x}_i \ - \ \bar{\boldsymbol{x}}\|^2 \ \leqslant \ \sum_{i=1}^{N} p_i \|\boldsymbol{x}_i \ - \ \boldsymbol{m}_{\boldsymbol{x}}\|^2 \ \leqslant \ m_{\max}^2 \ . \tag{A77}$$

Let us quickly recall that the spectral norm of an outer product of two vectors is the product of the Euclidean norms of the vectors:

$$\left\| \boldsymbol{a}\boldsymbol{b}^T \right\|_2 \ = \ \sqrt{\lambda_{\max}(\boldsymbol{b}\boldsymbol{a}^T \boldsymbol{a}\boldsymbol{b}^T)} \ = \ \|\boldsymbol{a}\| \ \sqrt{\lambda_{\max}(\boldsymbol{b}\boldsymbol{b}^T)} \ = \ \|\boldsymbol{a}\| \ \|\boldsymbol{b}\| \ , \tag{A78}$$

since $\boldsymbol{b}\boldsymbol{b}^T$ has eigenvector $\boldsymbol{b}/\|\boldsymbol{b}\|$ with eigenvalue $\|\boldsymbol{b}\|^2$ and otherwise zero eigenvalues.

We now bound the variance of the patterns:

$$\begin{aligned} \left\| \mathrm{Var}_{\boldsymbol{p}}[\boldsymbol{x}] \right\|_2 &\leqslant \sum_{i=1}^{N} p_i \left\| \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}} \right) \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}} \right)^T \right\|_2 \\ &= \sum_{i=1}^{N} p_i \left\| \boldsymbol{x}_i - \bar{\boldsymbol{x}} \right\|^2 \leqslant \sum_{i=1}^{N} p_i \left\| \boldsymbol{x}_i - \boldsymbol{m}_{\boldsymbol{x}} \right\|^2 \leqslant m_{\max}^2 \, . \end{aligned} \tag{A79}$$

The bound of the lemma on $\left\| \mathrm{J} \right\|_2$ follows from Eq. (A67).

For $\left\| \mathrm{J} \right\|_2 \leqslant \beta \, m_{\max}^2 < 1$ we have a contraction mapping on each compact set. Banach fixed point theorem says there is a unique fixed point in the compact set.

□

Now let us further investigate the tightness of the bound on $\left\| \mathrm{Var}_{\boldsymbol{p}}[\boldsymbol{x}] \right\|_2$ via $\left\| \boldsymbol{x}_i - \bar{\boldsymbol{x}} \right\|^2$: we consider the trace, which is the sum $\sum_{k=1}^{d} e_k$ of the w.l.o.g. ordered nonnegative eigenvalues e_k of $\mathrm{Var}_{\boldsymbol{p}}[\boldsymbol{x}]$ The spectral norm is equal to the largest eigenvalue e_1, which is equal to the largest singular value, as we have positive semidefinite matrices. We obtain:

$$\begin{aligned} \left\| \mathrm{Var}_{\boldsymbol{p}}[\boldsymbol{x}] \right\|_2 &= \mathrm{Tr} \left(\sum_{i=1}^{N} p_i \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}} \right) \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}} \right)^T \right) - \sum_{k=2}^{d} e_k \\ &= \sum_{i=1}^{N} p_i \mathrm{Tr} \left(\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}} \right) \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}} \right)^T \right) - \sum_{k=2}^{d} e_k \\ &= \sum_{i=1}^{N} p_i \left\| \boldsymbol{x}_i - \bar{\boldsymbol{x}} \right\|^2 - \sum_{k=2}^{d} e_k \, . \end{aligned} \tag{A80}$$

Therefore, the tightness of the bound depends on eigenvalues which are not the largest. Hence variations which are not along the largest variation weaken the bound.

Next we investigate the location of fixed points which existence is ensured by the global convergence stated in Theorem A2. For N patterns $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, we consider the iteration

$$\boldsymbol{\xi}^{\mathrm{new}} = f(\boldsymbol{\xi}) = \boldsymbol{X} \boldsymbol{p} = \boldsymbol{X} \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) \tag{A81}$$

using

$$\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) \, . \tag{A82}$$

$\boldsymbol{\xi}^{\mathrm{new}}$ is in the simplex of the patterns, that is, $\boldsymbol{\xi}^{\mathrm{new}} = \sum_i p_i \boldsymbol{x}_i$ with $\sum_i p_i = 1$ and $0 \leqslant p_i$. Hence, after one update $\boldsymbol{\xi}$ is in the simplex of the pattern and stays there. If the center $\boldsymbol{m}_{\boldsymbol{x}}$ is the zero vector $\boldsymbol{m}_{\boldsymbol{x}} = \boldsymbol{0}$, that is, the data is centered, then the mean is a fixed point of the iteration. For $\boldsymbol{\xi} = \boldsymbol{m}_{\boldsymbol{x}} = \boldsymbol{0}$ we have

$$\boldsymbol{p} = 1/N \, \boldsymbol{1} \tag{A83}$$

and

$$\boldsymbol{\xi}^{\mathrm{new}} = 1/N \, \boldsymbol{X} \, \boldsymbol{1} = \boldsymbol{m}_{\boldsymbol{x}} = \boldsymbol{\xi} \, . \tag{A84}$$

In particular normalization methods like batch normalization would promote the mean as a fixed point.

We consider the differences of dot products for $\boldsymbol{x}_i$: $\boldsymbol{x}_i^T \boldsymbol{x}_i - \boldsymbol{x}_i^T \boldsymbol{x}_j = \boldsymbol{x}_i^T (\boldsymbol{x}_i - \boldsymbol{x}_j)$, for fixed point $\boldsymbol{m}_{\boldsymbol{x}}^*$: $(\boldsymbol{m}_{\boldsymbol{x}}^*)^T \boldsymbol{x}_i - (\boldsymbol{m}_{\boldsymbol{x}}^*)^T \boldsymbol{x}_j = (\boldsymbol{m}_{\boldsymbol{x}}^*)^T (\boldsymbol{x}_i - \boldsymbol{x}_j)$, and for the center $\boldsymbol{m}_{\boldsymbol{x}}$: $\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_i - \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j = \boldsymbol{m}_{\boldsymbol{x}}^T (\boldsymbol{x}_i - \boldsymbol{x}_j)$. Using the Cauchy-Schwarz inequality, we get

$$\begin{aligned} \left| \boldsymbol{\xi}^T (\boldsymbol{x}_i - \boldsymbol{x}_j) \right| &\leqslant \left\| \boldsymbol{\xi} \right\| \left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\| \leqslant \left\| \boldsymbol{\xi} \right\| \left(\left\| \boldsymbol{x}_i - \boldsymbol{m}_{\boldsymbol{x}} \right\| + \left\| \boldsymbol{x}_j - \boldsymbol{m}_{\boldsymbol{x}} \right\| \right) \\ &\leqslant 2 \, m_{\max} \left\| \boldsymbol{\xi} \right\| \, . \end{aligned} \tag{A85}$$

This inequality gives:

$$\begin{aligned} \left|\boldsymbol{\xi}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\right| &\leqslant 2\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|)\,, \\ \left|\boldsymbol{\xi}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\right| &\leqslant 2\, m_{\max}\, M\,, \end{aligned} \tag{A86}$$

where we used $\|\boldsymbol{\xi} - \mathbf{0}\| \leqslant \|\boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}}\| + \|\boldsymbol{m}_{\boldsymbol{x}} - \mathbf{0}\|$, $\|\boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}}\| = \|\sum_i p_i \boldsymbol{x}_i - \boldsymbol{m}_{\boldsymbol{x}}\| \leqslant \sum_i p_i \|\boldsymbol{x}_i - \boldsymbol{m}_{\boldsymbol{x}}\| \leqslant m_{\max}$, and $M = \max_i \|\boldsymbol{x}_i\|$. In particular

$$\beta\, \left|\boldsymbol{m}_{\boldsymbol{x}}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\right| \leqslant 2\, \beta\, m_{\max}\, \|\boldsymbol{m}_{\boldsymbol{x}}\|\,, \tag{A87}$$

$$\beta\, \left|(\boldsymbol{m}_{\boldsymbol{x}}^*)^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\right| \leqslant 2\, \beta\, m_{\max}\, \|\boldsymbol{m}_{\boldsymbol{x}}^*\| \leqslant 2\, \beta\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|)\,, \tag{A88}$$

$$\beta\, \left|\boldsymbol{x}_i^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\right| \leqslant 2\, \beta\, m_{\max}\, \|\boldsymbol{x}_i\| \leqslant 2\, \beta\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|)\,. \tag{A89}$$

Let $i = \arg\max_j \boldsymbol{\xi}^T \boldsymbol{x}_j$, therefore the maximal softmax component is i. For the maximal softmax component i we have:

$$\begin{aligned} [\mathrm{softmax}(\beta\, \boldsymbol{X}^T \boldsymbol{\xi})]_i &= \frac{1}{1 + \sum_{j \neq i} \exp(-\beta\, (\boldsymbol{\xi}^T \boldsymbol{x}_i - \boldsymbol{\xi}^T \boldsymbol{x}_j))} \\ &\leqslant \frac{1}{1 + \sum_{j \neq i} \exp(-2\, \beta\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|))} \\ &= \frac{1}{1 + (N-1) \exp(-2\, \beta\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|))} \\ &= \frac{\exp(2\, \beta\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|))}{\exp(2\, \beta\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|)) + (N-1)} \\ &\leqslant 1/N\, \exp(2\, \beta\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|))\,. \end{aligned} \tag{A90}$$

Analogously we obtain for $i = \arg\max_j \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j$, a bound on the maximal softmax component i if the center is put into the iteration:

$$[\mathrm{softmax}(\beta\, \boldsymbol{X}^T \boldsymbol{m}_{\boldsymbol{x}})]_i \leqslant 1/N\, \exp(2\, \beta\, m_{\max}\, \|\boldsymbol{m}_{\boldsymbol{x}}\|)\,. \tag{A91}$$

Analog we obtain a bound for $i = \arg\max_j (\boldsymbol{m}_{\boldsymbol{x}}^*)^T \boldsymbol{x}_j$ on the maximal softmax component i of the fixed point:

$$\begin{aligned} [\mathrm{softmax}(\beta\, \boldsymbol{X}^T \boldsymbol{m}_{\boldsymbol{x}}^*)]_i &\leqslant 1/N\, \exp(2\, \beta\, m_{\max}\, \|\boldsymbol{m}_{\boldsymbol{x}}^*\|) \\ &\leqslant 1/N\, \exp(2\, \beta\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|))\,. \end{aligned} \tag{A92}$$

The two important terms are $m_{\max}$, the variance or spread of the data and $\|\boldsymbol{m}_{\boldsymbol{x}}\|$, which tells how well the data is centered. For a contraction mapping we already required $\beta m_{\max}^2 < 1$, therefore the first term in the exponent is $2\beta m_{\max}^2 < 2$. The second term $2\beta m_{\max} \|\boldsymbol{m}_{\boldsymbol{x}}\|$ is small if the data is centered.

Global fixed point near the global mean: Analysis using softmax values. If $\boldsymbol{\xi}^T \boldsymbol{x}_i \approx \boldsymbol{\xi}^T \boldsymbol{x}_j$ for all i and j, then $p_i \approx 1/N$ and we have $m = \max_i p_i(1 - p_i) < 1/N$. For $M \leqslant 1/\sqrt{2\beta}$ we obtain from Lemma A2:

$$\|\mathrm{J}\|_2 < 1\,. \tag{A93}$$

The local fixed point is $\boldsymbol{m}_{\boldsymbol{x}}^* \approx \boldsymbol{m}_{\boldsymbol{x}} = (1/N) \sum_{i=1}^N \boldsymbol{x}_i$ with $p_i \approx 1/N$.

We now treat this case more formally. First we discuss conditions that ensure that the iteration is a contraction mapping. We consider the iteration Eq. (A46) in the variable $\boldsymbol{p}$:

$$\boldsymbol{p}^{\mathrm{new}} = g(\boldsymbol{p}) = \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{p})\,. \tag{A94}$$

The Jacobian is

$$\mathrm{J}(\boldsymbol{p}) = \frac{\partial g(\boldsymbol{p})}{\partial \boldsymbol{p}} = \boldsymbol{X}^T \boldsymbol{X}\, \mathrm{J}_s \tag{A95}$$

with

$$\mathrm{J}_s(\boldsymbol{p}^{\mathrm{new}}) \;=\; \beta \left(\mathrm{diag}(\boldsymbol{p}^{\mathrm{new}}) \;-\; \boldsymbol{p}^{\mathrm{new}} (\boldsymbol{p}^{\mathrm{new}})^T\right) \;. \tag{A96}$$

The version of the mean value theorem in Lemma A32 states for $\mathrm{J}^m = \int_0^1 \mathrm{J}(\lambda \boldsymbol{p}) \, \mathrm{d}\lambda = \boldsymbol{X}^T \boldsymbol{X} \mathrm{J}_s^m$ with the symmetric matrix $\mathrm{J}_s^m = \int_0^1 \mathrm{J}_s(\lambda \boldsymbol{p}) \, \mathrm{d}\lambda$:

$$\boldsymbol{p}^{\mathrm{new}} \;=\; g(\boldsymbol{p}) \;=\; g(\boldsymbol{0}) \;+\; (\mathrm{J}^m)^T \boldsymbol{p} \;=\; g(\boldsymbol{0}) \;+\; \mathrm{J}_s^m \; \boldsymbol{X}^T \boldsymbol{X} \; \boldsymbol{p} \;=\; 1/N \; \boldsymbol{1} \;+\; \mathrm{J}_s^m \; \boldsymbol{X}^T \boldsymbol{X} \; \boldsymbol{p} \;. \tag{A97}$$

With $m = \max_i p_i(1-p_i)$, Eq. (A461) from Lemma A24 is

$$\left\| \mathrm{J}_s(\boldsymbol{p}) \right\|_2 \;=\; \beta \left\| \mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T \right\|_2 \;\leqslant\; 2 \, m \, \beta \;. \tag{A98}$$

First observe that $\lambda p_i(1-\lambda p_i) \leqslant p_i(1-p_i)$ for $p_i \leqslant 0.5$ and $\lambda \in [0,1]$, since $p_i(1-p_i) - \lambda p_i(1-\lambda p_i) = (1-\lambda) p_i (1-(1+\lambda)p_i) \geq 0$. For $\max_i p_i \leqslant 0.5$ this observation leads to the following bound for J_s^m:

$$\left\| \mathrm{J}_s^m \right\|_2 \;\leqslant\; 2 \, m \, \beta \;. \tag{A99}$$

Eq. (A464) in Lemma A24 states that every J_s is bounded by $1/2\beta$, therefore also the mean:

$$\left\| \mathrm{J}_s^m \right\|_2 \;\leqslant\; 0.5 \, \beta \;. \tag{A100}$$

Since $m = \max_i p_i(1-p_i) < \max_i p_i = p_{\max}$, the previous bounds can be combined as follows:

$$\left\| \mathrm{J}_s^m \right\|_2 \;\leqslant\; 2 \; \min\{0.25, p_{\max}\} \; \beta \;. \tag{A101}$$

Consequently,

$$\left\| \mathrm{J}^m \right\|_2 \;\leqslant\; N \; M^2 \; 2 \; \min\{0.25, p_{\max}\} \; \beta \;, \tag{A102}$$

where we used Eq. (A159). $\left\| \boldsymbol{X}^T \boldsymbol{X} \right\|_2 = \left\| \boldsymbol{X} \boldsymbol{X}^T \right\|_2$, therefore $\left\| \boldsymbol{X}^T \boldsymbol{X} \right\|_2$ is N times the maximal second moment of the data squared.

Obviously, $g(\boldsymbol{p})$ is a contraction mapping in compact sets, where

$$N \; M^2 \; 2 \; \min\{0.25, p_{\max}\} \; \beta \;<\; 1 \;. \tag{A103}$$

S is the sphere around the origin $\boldsymbol{0}$ with radius one. For

$$\boldsymbol{p}^{\mathrm{new}} \;=\; g(\boldsymbol{p}) \;=\; 1/N \; \boldsymbol{1} \;+\; \mathrm{J}^m \; \boldsymbol{p} \;, \tag{A104}$$

we have $\|\boldsymbol{p}\| \leqslant \|\boldsymbol{p}\|_1 = 1$ and $\|\boldsymbol{p}^{\mathrm{new}}\| \leqslant \|\boldsymbol{p}^{\mathrm{new}}\|_1 = 1$. Therefore, g maps points from S into S. g is a contraction mapping for

$$\left\| \mathrm{J}^m \right\|_2 \;\leqslant\; N \; M^2 \; 2 \; \min\{0.25, p_{\max}\} \; \beta \;=\; c \;<\; 1 \;. \tag{A105}$$

According to Banach fixed point theorem g has a fixed point in the sphere S.

Hölder's inequality gives:

$$\|\boldsymbol{p}\|^2 \;=\; \boldsymbol{p}^T \boldsymbol{p} \;\leqslant\; \|\boldsymbol{p}\|_1 \|\boldsymbol{p}\|_\infty \;=\; \|\boldsymbol{p}\|_\infty \;=\; p_{\max} \;. \tag{A106}$$

Alternatively:

$$\|\boldsymbol{p}\|^2 \;=\; \sum_i p_i^2 \;=\; p_{\max} \sum_i \frac{p_i}{p_{\max}} \, p_i \;\leqslant\; p_{\max} \sum_i p_i \;=\; p_{\max} \;. \tag{A107}$$

Let now S be the sphere around the origin $\boldsymbol{0}$ with radius $1/\sqrt{N} + \sqrt{p_{\max}}$ and let $\left\| \mathrm{J}^m(\boldsymbol{p}) \right\|_2 \leqslant c < 1$ for $\boldsymbol{p} \in \mathrm{S}$. The old $\boldsymbol{p}$ is in the sphere S ($\boldsymbol{p} \in \mathrm{S}$) since $p_{\max} < \sqrt{p_{\max}}$ for $p_{\max} < 1$. We have

$$\|\boldsymbol{p}^{\mathrm{new}}\| \;\leqslant\; 1/\sqrt{N} \;+\; \left\| \mathrm{J}^m \right\|_2 \|\boldsymbol{p}\| \;\leqslant\; 1/\sqrt{N} \;+\; \sqrt{p_{\max}} \;. \tag{A108}$$

Therefore, g is a mapping from S into S and a contraction mapping. According to Banach fixed point theorem, a fixed point exists in S.

For the 1-norm, we use Lemma A24 and $\|\boldsymbol{p}\|_1 = 1$ to obtain from Eq. (A104):

$$\|\boldsymbol{p}^{\rm new} - 1/N\,\mathbf{1}\|_1 \leqslant \|\mathrm{J}^m\|_1 \leqslant 2\,\beta\,m\,\|\boldsymbol{X}\|_\infty\,M_1\,, \tag{A109}$$

$$\|\boldsymbol{p}^{\rm new} - 1/N\,\mathbf{1}\|_1 \leqslant \|\mathrm{J}^m\|_1 \leqslant 2\,\beta\,m\,N\,M_\infty\,M_1\,, \tag{A110}$$

$$\|\boldsymbol{p}^{\rm new} - 1/N\,\mathbf{1}\|_1 \leqslant \|\mathrm{J}^m\|_1 \leqslant 2\,\beta\,m\,N\,M^2\,, \tag{A111}$$

where $m = \max_i p_i(1-p_i)$, $M_1 = \|\boldsymbol{X}\|_1 = \max_i \|\boldsymbol{x}_i\|_1$, $M = \max_i \|\boldsymbol{x}_i\|$, $\|\boldsymbol{X}\|_\infty = \left\|\boldsymbol{X}^T\right\|_1 = \max_i \left\|[X^T]_i\right\|_1$ (maximal absolute row sum norm), and $M_\infty = \max_i \|\boldsymbol{x}_i\|_\infty$. Let us quickly mention some auxiliary estimates related to $\boldsymbol{X}^T\boldsymbol{X}$:

$$\begin{aligned}\left\|\boldsymbol{X}^T\boldsymbol{X}\right\|_1 &= \max_i \sum_{j=1}^N \left|\boldsymbol{x}_i^T\boldsymbol{x}_j\right| \leqslant \max_i \sum_{j=1}^N \|\boldsymbol{x}_i\|_\infty\,\|\boldsymbol{x}_j\|_1 \\ &\leqslant M_\infty \sum_{j=1}^N M_1 = N\,M_\infty\,M_1\,,\end{aligned} \tag{A112}$$

where the first inequaltiy is from Hölder's inequality. We used

$$\begin{aligned}\left\|\boldsymbol{X}^T\boldsymbol{X}\right\|_1 &= \max_i \sum_{j=1}^N \left|\boldsymbol{x}_i^T\boldsymbol{x}_j\right| \leqslant \max_i \sum_{j=1}^N \|\boldsymbol{x}_i\|\,\|\boldsymbol{x}_j\| \\ &\leqslant M \sum_{j=1}^N M = N\,M^2\,,\end{aligned} \tag{A113}$$

where the first inequality is from Hölder's inequality (here the same as the Cauchy-Schwarz inequality). See proof of Lemma A24 for the 1-norm bound on J_s. Everything else follows from the fact that the 1-norm is sub-multiplicative as induced matrix norm.

We consider the minimal $\|\boldsymbol{p}\|$.

$$\begin{aligned}&\min_{\boldsymbol{p}} \ \|\boldsymbol{p}\|^2 \\ &\text{s.t.} \ \sum_i p_i = 1 \\ &\qquad \forall_i : \ p_i \geq 0\,.\end{aligned} \tag{A114}$$

The solution to this minimization problem is $\boldsymbol{p} = (1/N)\mathbf{1}$. Therefore, we have $1/\sqrt{N} \leqslant \|\boldsymbol{p}\|$ and $1/N \leqslant \|\boldsymbol{p}\|^2$ Using Eq. (A108) we obtain

$$1/\sqrt{N} \leqslant \|\boldsymbol{p}^{\rm new}\| \leqslant 1/\sqrt{N} + \sqrt{p_{\max}}\,. \tag{A115}$$

Moreover

$$\begin{aligned}\|\boldsymbol{p}^{\rm new}\|^2 &= (\boldsymbol{p}^{\rm new})^T\boldsymbol{p}^{\rm new} = 1/N + (\boldsymbol{p}^{\rm new})^T\mathrm{J}^m\,\boldsymbol{p} \leqslant 1/N + \|\mathrm{J}^m\|_2\,\|\boldsymbol{p}\| \\ &\leqslant 1/N + \|\mathrm{J}^m\|_2\,,\end{aligned} \tag{A116}$$

since $\boldsymbol{p}^{\rm new} \in \mathrm{S}$ and $\boldsymbol{p} \in \mathrm{S}$.

For the fixed point, we have

$$\|\boldsymbol{p}^*\|^2 = (\boldsymbol{p}^*)^T\boldsymbol{p}^* = 1/N + (\boldsymbol{p}^*)^T\mathrm{J}^m\,\boldsymbol{p}^* \leqslant 1/N + \|\mathrm{J}^m\|_2\,\|\boldsymbol{p}^*\|^2\,, \tag{A117}$$

and hence

$$1/N \leqslant \|\boldsymbol{p}^*\|^2 \leqslant 1/N\frac{1}{1 - \|\mathrm{J}^m\|_2} = 1/N\,(1 + \frac{\|\mathrm{J}^m\|_2}{1 - \|\mathrm{J}^m\|_2})\,. \tag{A118}$$

Therefore, for small $\|\mathrm{J}^m\|_2$ we have $\boldsymbol{p}^* \approx (1/N)\mathbf{1}$.

A2.4.3 Many Stable States: Fixed Points Near Stored Patterns

We move on to the next case, where the patterns $\boldsymbol{x}_i$ are well separated. In this case the iterate goes to the pattern to which the initial $\boldsymbol{\xi}$ is most similar. If the initial $\boldsymbol{\xi}$ is similar to a vector $\boldsymbol{x}_i$ then it will converge to $\boldsymbol{x}_i$ and $\boldsymbol{p}$ will be $\boldsymbol{e}_i$. The main ingredients are again Banach's Theorem and estimates on the Jacobian norm.

Proof of a fixed point by Banach Fixed Point Theorem. *Mapped Vectors Stay in a Compact Environment.* We show that if $\boldsymbol{x}_i$ is sufficient dissimilar to other $\boldsymbol{x}_j$ then there is an compact environment of $\boldsymbol{x}_i$ (a sphere) where the fixed point iteration maps this environment into itself. The idea of the proof is to define a sphere around $\boldsymbol{x}_i$ for which points from the sphere are mapped by f into the sphere.

We first need following lemma which bounds the distance $\|\boldsymbol{x}_i - f(\boldsymbol{\xi})\|$, where $\boldsymbol{x}_i$ is the pattern that is least separated from $\boldsymbol{\xi}$ but separated from other patterns.

Lemma A4. *For a query $\boldsymbol{\xi}$ and data $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, there exists a $\boldsymbol{x}_i$ that is least separated from $\boldsymbol{\xi}$ while being separated from other $\boldsymbol{x}_j$ with $j \neq i$:*

$$i = \arg\max_k \min_{j, j \neq k} \left(\boldsymbol{\xi}^T \boldsymbol{x}_k - \boldsymbol{\xi}^T \boldsymbol{x}_j \right) = \arg\max_k \left(\boldsymbol{\xi}^T \boldsymbol{x}_k - \max_{j, j \neq k} \boldsymbol{\xi}^T \boldsymbol{x}_j \right) \tag{A119}$$

$$0 \leqslant c = \max_k \min_{j, j \neq k} \left(\boldsymbol{\xi}^T \boldsymbol{x}_k - \boldsymbol{\xi}^T \boldsymbol{x}_j \right) = \max_k \left(\boldsymbol{\xi}^T \boldsymbol{x}_k - \max_{j, j \neq k} \boldsymbol{\xi}^T \boldsymbol{x}_j \right) . \tag{A120}$$

For $\boldsymbol{x}_i$, the following holds:

$$\|\boldsymbol{x}_i - f(\boldsymbol{\xi})\| \leqslant 2 \, \epsilon \, M \, , \tag{A121}$$

where

$$M = \max_i \|\boldsymbol{x}_i\| \, , \tag{A122}$$

$$\epsilon = (N-1) \, \exp(-\beta \, c) \, . \tag{A123}$$

Proof. For the softmax component i we have:

$$\begin{aligned} [\mathrm{softmax}(\beta \, \boldsymbol{X}^T \boldsymbol{\xi})]_i &= \frac{1}{1 + \sum_{j \neq i} \exp(\beta \, (\boldsymbol{\xi}^T \boldsymbol{x}_j - \boldsymbol{\xi}^T \boldsymbol{x}_i))} \geq \frac{1}{1 + \sum_{j \neq i} \exp(-\beta \, c)} \\ &= \frac{1}{1 + (N-1) \exp(-\beta \, c)} = 1 - \frac{(N-1) \exp(-\beta \, c)}{1 + (N-1) \exp(-\beta \, c)} \\ &\geq 1 - (N-1) \exp(-\beta \, c) = 1 - \epsilon \end{aligned} \tag{A124}$$

For softmax components $k \neq i$ we have

$$[\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})]_k = \frac{\exp(\beta \, (\boldsymbol{\xi}^T \boldsymbol{x}_k - \boldsymbol{\xi}^T \boldsymbol{x}_i))}{1 + \sum_{j \neq i} \exp(\beta \, (\boldsymbol{\xi}^T \boldsymbol{x}_j - \boldsymbol{\xi}^T \boldsymbol{x}_i))} \leqslant \exp(-\beta \, c) = \frac{\epsilon}{N-1} \, . \tag{A125}$$

The iteration f can be written as

$$f(\boldsymbol{\xi}) = \boldsymbol{X} \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) = \sum_{j=1}^{N} \boldsymbol{x}_j \, [\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})]_j \, . \tag{A126}$$

We now can bound $\|\boldsymbol{x}_i \ - \ f(\boldsymbol{\xi})\|$:

$$\begin{aligned}
\|\boldsymbol{x}_i \ - \ f(\boldsymbol{\xi})\| \ &= \ \left\| \boldsymbol{x}_i \ - \ \sum_{j=1}^{N} [\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})]_j \ \boldsymbol{x}_j \right\| \qquad \text{(A127)} \\
&= \ \left\| (1 - [\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})]_i) \ \boldsymbol{x}_i \ - \ \sum_{j=1, j \neq i}^{N} [\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})]_j \ \boldsymbol{x}_j \right\| \\
&\leqslant \ \epsilon \, \|\boldsymbol{x}_i\| \ + \ \frac{\epsilon}{N-1} \ \sum_{j=1, j \neq i}^{N} \|\boldsymbol{x}_j\| \\
&\leqslant \ \epsilon \, M \ + \ \frac{\epsilon}{N-1} \ \sum_{j=1, j \neq i}^{N} M \ = \ 2 \, \epsilon \, M \ .
\end{aligned}$$

□

We define Δ_i, i.e. the separation of pattern $\boldsymbol{x}_i$ from data $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ as:

$$\Delta_i \ = \ \min_{j, j \neq i} \left(\boldsymbol{x}_i^T \boldsymbol{x}_i \ - \ \boldsymbol{x}_i^T \boldsymbol{x}_j \right) \ = \ \boldsymbol{x}_i^T \boldsymbol{x}_i \ - \ \max_{j, j \neq i} \boldsymbol{x}_i^T \boldsymbol{x}_j \ . \qquad \text{(A128)}$$

The pattern is separated from the other data if $0 < \Delta_i$. Using the parallelogram identity, Δ_i can also be expressed as

$$\begin{aligned}
\Delta_i \ &= \ \min_{j, j \neq i} \frac{1}{2} \ \left(\|\boldsymbol{x}_i\|^2 \ - \ \|\boldsymbol{x}_j\|^2 \ + \ \|\boldsymbol{x}_i \ - \ \boldsymbol{x}_j\|^2 \right) \qquad \text{(A129)} \\
&= \ \frac{1}{2} \|\boldsymbol{x}_i\|^2 \ - \ \frac{1}{2} \ \max_{j, j \neq i} \left(\|\boldsymbol{x}_j\|^2 \ - \ \|\boldsymbol{x}_i \ - \ \boldsymbol{x}_j\|^2 \right) \ .
\end{aligned}$$

For $\|\boldsymbol{x}_i\| = \|\boldsymbol{x}_j\|$ we have $\Delta_i = 1/2 \min_{j, j \neq i} \|\boldsymbol{x}_i \ - \ \boldsymbol{x}_j\|^2$.

Next we define the sphere where we want to apply Banach fixed point theorem.

Definition 3 (Sphere S_i)**.** *The sphere* S_i *is defined as*

$$\mathrm{S}_i \ := \ \left\{ \boldsymbol{\xi} \mid \|\boldsymbol{\xi} \ - \ \boldsymbol{x}_i\| \ \leqslant \ \frac{1}{\beta \, N \, M} \right\} \ . \qquad \text{(A130)}$$

Lemma A5. *With* $\boldsymbol{\xi}$ *given, if the assumptions*

A1: $\boldsymbol{\xi}$ *is inside sphere:* $\boldsymbol{\xi} \in \mathrm{S}_i$*,*

A2: data point $\boldsymbol{x}_i$ *is well separated from the other data:*

$$\Delta_i \ \geq \ \frac{2}{\beta \, N} \ + \ \frac{1}{\beta} \ \ln \left(2 \, (N-1) \, N \, \beta \, M^2 \right) \qquad \text{(A131)}$$

hold, then $f(\boldsymbol{\xi})$ *is inside the sphere:* $f(\boldsymbol{\xi}) \in \mathrm{S}_i$*. Therefore, with assumption (A2),* f *is a mapping from* S_i *into* S_i*.*

Proof. We need the separation $\tilde{\Delta}_i$ of $\boldsymbol{\xi}$ from the data.

$$\tilde{\Delta}_i \ = \ \min_{j, j \neq i} \left(\boldsymbol{\xi}^T \boldsymbol{x}_i \ - \ \boldsymbol{\xi}^T \boldsymbol{x}_j \right) \ . \qquad \text{(A132)}$$

Using the Cauchy-Schwarz inequality, we obtain for $1 \leqslant j \leqslant N$:

$$\left| \boldsymbol{\xi}^T \boldsymbol{x}_j \ - \ \boldsymbol{x}_i^T \boldsymbol{x}_j \right| \leqslant \ \|\boldsymbol{\xi} \ - \ \boldsymbol{x}_i\| \, \|\boldsymbol{x}_j\| \ \leqslant \ \|\boldsymbol{\xi} \ - \ \boldsymbol{x}_i\| \, M \ . \qquad \text{(A133)}$$

We have the lower bound

$$\begin{aligned}
\tilde{\Delta}_i \ &\geq \ \min_{j, j \neq i} \left(\left(\boldsymbol{x}_i^T \boldsymbol{x}_i \ - \ \|\boldsymbol{\xi} \ - \ \boldsymbol{x}_i\| \, M \right) \ - \ \left(\boldsymbol{x}_i^T \boldsymbol{x}_j \ + \ \|\boldsymbol{\xi} \ - \ \boldsymbol{x}_i\| \, M \right) \right) \qquad \text{(A134)} \\
&= \ - \, 2 \, \|\boldsymbol{\xi} \ - \ \boldsymbol{x}_i\| \, M \ + \ \min_{j, j \neq i} \left(\boldsymbol{x}_i^T \boldsymbol{x}_i \ - \ \boldsymbol{x}_i^T \boldsymbol{x}_j \right) \ = \ \Delta_i \ - \ 2 \, \|\boldsymbol{\xi} \ - \ \boldsymbol{x}_i\| \, M \\
&\geq \ \Delta_i \ - \ \frac{2}{\beta \, N} \ ,
\end{aligned}$$

where we used the assumption (A1) of the lemma.

From the proof in Lemma A4 we have

$$p_{\max} \;=\; [\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})]_i \;\geq\; 1 \;-\; (N-1)\, \exp(-\,\beta\, \tilde{\Delta}_i) \;=\; 1 \;-\; \tilde{\epsilon}\,. \tag{A135}$$

Lemma A4 states that

$$\begin{aligned} \|\boldsymbol{x}_i \;-\; f(\boldsymbol{\xi})\| \;&\leqslant\; 2\, \tilde{\epsilon}\, M \;=\; 2\, (N-1)\, \exp(-\,\beta\, \tilde{\Delta}_i)\, M \\ &\leqslant\; 2\, (N-1)\, \exp(-\,\beta\, (\Delta_i \;-\; \frac{2}{\beta\, N}))\, M\,. \end{aligned} \tag{A136}$$

We have

$$\begin{aligned} &\|\boldsymbol{x}_i \;-\; f(\boldsymbol{\xi})\| \\ &\leqslant\; 2\, (N-1)\, \exp(-\,\beta\, (\frac{2}{\beta\, N} \;+\; \frac{1}{\beta}\, \ln\left(2\, (N-1)\, N\, \beta\, M^2\right) \;-\; \frac{2}{\beta\, N}))\, M \\ &=\; 2\, (N-1)\, \exp(-\,\ln\left(2\, (N-1)\, N\, \beta\, M^2\right))\, M \\ &=\; \frac{1}{N\, \beta\, M}\,, \end{aligned} \tag{A137}$$

where we used assumption (A2) of the lemma. Therefore, $f(\boldsymbol{\xi})$ is a mapping from the sphere S_i into the sphere S_i: If $\boldsymbol{\xi} \in \mathrm{S}_i$ then $f(\boldsymbol{\xi}) \in \mathrm{S}_i$. □

Contraction mapping. For applying Banach fixed point theorem we need to show that f is contraction in the compact environment S_i.

Lemma A6. *Assume that*

A1:

$$\Delta_i \;\geq\; \frac{2}{\beta\, N} \;+\; \frac{1}{\beta}\, \ln\left(2\, (N-1)\, N\, \beta\, M^2\right)\,, \tag{A138}$$

then f is a contraction mapping in S_i.

Proof. The version of the mean value theorem Lemma A32 states for $\mathrm{J}^m = \int_0^1 \mathrm{J}(\lambda \boldsymbol{\xi} + (1-\lambda) \boldsymbol{x}_i)\, \mathrm{d}\lambda$:

$$f(\boldsymbol{\xi}) \;=\; f(\boldsymbol{x}_i) \;+\; \mathrm{J}^m\, (\boldsymbol{\xi} \;-\; \boldsymbol{x}_i)\,. \tag{A139}$$

Therefore

$$\|f(\boldsymbol{\xi}) \;-\; f(\boldsymbol{x}_i)\| \;\leqslant\; \|\mathrm{J}^m\|_2\, \|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|\,. \tag{A140}$$

We define $\tilde{\boldsymbol{\xi}} = \lambda \boldsymbol{\xi} + (1-\lambda) \boldsymbol{x}_i$ for some $\lambda \in [0, 1]$. From the proof in Lemma A4 we have

$$p_{\max}(\tilde{\boldsymbol{\xi}}) \;=\; [\mathrm{softmax}(\beta\, \boldsymbol{X}^T\, \tilde{\boldsymbol{\xi}})]_i \;\geq\; 1 \;-\; (N-1)\, \exp(-\,\beta\, \tilde{\Delta}_i) \;=\; 1 \;-\; \tilde{\epsilon}\,, \tag{A141}$$

$$\tilde{\epsilon} \;=\; (N-1)\, \exp(-\,\beta\, \tilde{\Delta}_i)\,, \tag{A142}$$

$$\tilde{\Delta}_i \;=\; \min_{j, j \neq i} \left(\tilde{\boldsymbol{\xi}}^T \boldsymbol{x}_i \;-\; \tilde{\boldsymbol{\xi}}^T \boldsymbol{x}_j\right)\,. \tag{A143}$$

First we compute an upper bound on $\tilde{\epsilon}$. We need the separation $\tilde{\Delta}_i$ of $\boldsymbol{\xi}$ from the data. Using the Cauchy-Schwarz inequality, we obtain for $1 \leqslant j \leqslant N$:

$$\left|\tilde{\boldsymbol{\xi}}^T \boldsymbol{x}_j \;-\; \boldsymbol{x}_i^T \boldsymbol{x}_j\right| \leqslant \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{x}_i\right\|\, \|\boldsymbol{x}_j\| \;\leqslant\; \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{x}_i\right\|\, M\,. \tag{A144}$$

We have the lower bound on $\tilde{\Delta}_i$:

$$\begin{aligned} \tilde{\Delta}_i \;&\geq\; \min_{j, j \neq i} \left(\left(\boldsymbol{x}_i^T \boldsymbol{x}_i \;-\; \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{x}_i\right\|\, M\right) \;-\; \left(\boldsymbol{x}_i^T \boldsymbol{x}_j \;+\; \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{x}_i\right\|\, M\right)\right) \\ &=\; -\,2\, \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{x}_i\right\|\, M \;+\; \min_{j, j \neq i} \left(\boldsymbol{x}_i^T \boldsymbol{x}_i \;-\; \boldsymbol{x}_i^T \boldsymbol{x}_j\right) \;=\; \Delta_i \;-\; 2\, \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{x}_i\right\|\, M \\ &\geq\; \Delta_i \;-\; 2\, \|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|\, M\,, \end{aligned} \tag{A145}$$

where we used $\left\| \tilde{\boldsymbol{\xi}} - \boldsymbol{x}_i \right\| = \lambda \|\boldsymbol{\xi} - \boldsymbol{x}_i\| \leqslant \|\boldsymbol{\xi} - \boldsymbol{x}_i\|$. From the definition of $\tilde{\epsilon}$ in Eq. (A141) we have

$$\begin{aligned} \tilde{\epsilon} &= (N-1) \exp(-\beta \tilde{\Delta}_i) \\ &\leqslant (N-1) \exp\left(-\beta \left(\Delta_i - 2 \|\boldsymbol{\xi} - \boldsymbol{x}_i\| M\right)\right) \\ &\leqslant (N-1) \exp\left(-\beta \left(\Delta_i - \frac{2}{\beta N}\right)\right) , \end{aligned} \tag{A146}$$

where we used $\boldsymbol{\xi} \in \mathrm{S}_i$, therefore $\|\boldsymbol{\xi} - \boldsymbol{x}_i\| \leqslant \frac{1}{\beta N M}$.

Next we compute an lower bound on $\tilde{\epsilon}$. We start with an upper on $\tilde{\Delta}_i$:

$$\begin{aligned} \tilde{\Delta}_i &\leqslant \min_{j, j \neq i} \left(\left(\boldsymbol{x}_i^T \boldsymbol{x}_i + \left\| \tilde{\boldsymbol{\xi}} - \boldsymbol{x}_i \right\| M \right) - \left(\boldsymbol{x}_i^T \boldsymbol{x}_j - \left\| \tilde{\boldsymbol{\xi}} - \boldsymbol{x}_i \right\| M \right) \right) \\ &= 2 \left\| \tilde{\boldsymbol{\xi}} - \boldsymbol{x}_i \right\| M + \min_{j, j \neq i} \left(\boldsymbol{x}_i^T \boldsymbol{x}_i - \boldsymbol{x}_i^T \boldsymbol{x}_j \right) = \Delta_i + 2 \left\| \tilde{\boldsymbol{\xi}} - \boldsymbol{x}_i \right\| M \\ &\leqslant \Delta_i + 2 \|\boldsymbol{\xi} - \boldsymbol{x}_i\| M , \end{aligned} \tag{A147}$$

where we used $\left\| \tilde{\boldsymbol{\xi}} - \boldsymbol{x}_i \right\| = \lambda \|\boldsymbol{\xi} - \boldsymbol{x}_i\| \leqslant \|\boldsymbol{\xi} - \boldsymbol{x}_i\|$. From the definition of $\tilde{\epsilon}$ in Eq. (A141) we have

$$\begin{aligned} \tilde{\epsilon} &= (N-1) \exp(-\beta \tilde{\Delta}_i) \\ &\geq (N-1) \exp\left(-\beta \left(\Delta_i + 2 \|\boldsymbol{\xi} - \boldsymbol{x}_i\| M\right)\right) \\ &\geq (N-1) \exp\left(-\beta \left(\Delta_i + \frac{2}{\beta N}\right)\right) , \end{aligned} \tag{A148}$$

where we used $\boldsymbol{\xi} \in \mathrm{S}_i$, therefore $\|\boldsymbol{\xi} - \boldsymbol{x}_i\| \leqslant \frac{1}{\beta N M}$.

Now we bound the Jacobian. We can assume $\tilde{\epsilon} \leqslant 0.5$ otherwise $(1 - \tilde{\epsilon}) \leqslant 0.5$ in the following. From the proof of Lemma A24 we know for $p_{\max}(\tilde{\boldsymbol{\xi}}) \geq 1 - \tilde{\epsilon}$, then $p_i(\tilde{\boldsymbol{\xi}}) \leqslant \tilde{\epsilon}$ for $p_i(\tilde{\boldsymbol{\xi}}) \neq p_{\max}(\tilde{\boldsymbol{\xi}})$. Therefore, $p_i(\tilde{\boldsymbol{\xi}})(1 - p_i(\tilde{\boldsymbol{\xi}})) \leqslant m \leqslant \tilde{\epsilon}(1 - \tilde{\epsilon})$ for all i. Next we use the derived upper and lower bound on $\tilde{\epsilon}$ in previous Eq. (A50) in Lemma A2:

$$\begin{aligned} \left\| \mathrm{J}(\tilde{\boldsymbol{\xi}}) \right\|_2 &\leqslant 2 \beta N M^2 \tilde{\epsilon} - 2 \tilde{\epsilon}^2 \beta N M^2 \\ &\leqslant 2 \beta N M^2 (N-1) \exp\left(-\beta \left(\Delta_i - \frac{2}{\beta N}\right)\right) - \\ &\quad 2 (N-1)^2 \exp\left(-2 \beta \left(\Delta_i + \frac{2}{\beta N}\right)\right) \beta N M^2 . \end{aligned} \tag{A149}$$

The bound Eq. (A149) holds for the mean J^m, too, since it averages over $\mathrm{J}(\tilde{\boldsymbol{\xi}})$:

$$\begin{aligned} \|\mathrm{J}^m\|_2 &\leqslant 2 \beta N M^2 (N-1) \exp\left(-\beta \left(\Delta_i - \frac{2}{\beta N}\right)\right) - \\ &\quad 2 (N-1)^2 \exp\left(-2 \beta \left(\Delta_i + \frac{2}{\beta N}\right)\right) \beta N M^2 . \end{aligned} \tag{A150}$$

The assumption of the lemma is

$$\Delta_i \geq \frac{2}{\beta N} + \frac{1}{\beta} \ln\left(2 (N-1) N \beta M^2\right) , \tag{A151}$$

This is

$$\Delta_i - \frac{2}{\beta N} \geq \frac{1}{\beta} \ln\left(2 (N-1) N \beta M^2\right) , \tag{A152}$$

Therefore, the spectral norm $\|\mathrm{J}\|_2$ can be bounded by:

$$\begin{aligned}
&\|\mathrm{J}^m\|_2 \leqslant 2\,\beta\,(N-1)\,\exp\left(-\,\beta\,\frac{1}{\beta}\,\ln\left(2\,(N-1)\,N\,\beta\,M^2\right)\right)\,N\,M^2\,- \\
&2\,(N-1)^2\,\exp\left(-\,2\,\beta\,\left(\Delta_i\,+\,\frac{2}{\beta\,N}\right)\right)\,\beta\,N\,M^2 \\
&=\;2\,\beta\,(N-1)\,\frac{1}{2\,(N-1)\,N\,\beta\,M^2}\,N\,M^2\,- \\
&2\,(N-1)^2\,\exp\left(-\,2\,\beta\,\left(\Delta_i\,+\,\frac{2}{\beta\,N}\right)\right)\,\beta\,N\,M^2 \\
&=\;1\,-2\,(N-1)^2\,\exp\left(-\,2\,\beta\,\left(\Delta_i\,+\,\frac{2}{\beta\,N}\right)\right)\,\beta\,N\,M^2\;<1\,.
\end{aligned} \tag{A153}$$

Therefore, f is a contraction mapping in S_i. □

Banach Fixed Point Theorem. Now we have all ingredients to apply Banach fixed point theorem.

Lemma A7. *Assume that*

A1:

$$\Delta_i\;\geq\;\frac{2}{\beta\,N}\;+\;\frac{1}{\beta}\,\ln\left(2\,(N-1)\,N\,\beta\,M^2\right)\,, \tag{A154}$$

then f has a fixed point in S_i.

Proof. We use Banach fixed point theorem: Lemma A5 says that f maps from S_i into S_i. Lemma A6 says that f is a contraction mapping in S_i. □

Contraction mapping with a fixed point. We have shown that a fixed point exists. We want to know how fast the iteration converges to the fixed point. Let $\boldsymbol{x}_i^*$ be the fixed point of the iteration f in the sphere S_i. Using the mean value theorem Lemma A32, we have with $\mathrm{J}^m = \int_0^1 \mathrm{J}(\lambda\boldsymbol{\xi} + (1-\lambda)\boldsymbol{x}_i^*)\,\mathrm{d}\lambda$:

$$\|f(\boldsymbol{\xi})\;-\;\boldsymbol{x}_i^*\|\;=\;\|f(\boldsymbol{\xi})\;-\;f(\boldsymbol{x}_i^*)\|\;\leqslant\;\|\mathrm{J}^m\|_2\,\|\boldsymbol{\xi}\;-\;\boldsymbol{x}_i^*\| \tag{A155}$$

According to Lemma A24, if $p_{\max} = \max_i p_i \geq 1-\epsilon$ for all $\tilde{\boldsymbol{x}} = \lambda\boldsymbol{\xi} + (1-\lambda)\boldsymbol{x}_i^*$, then the spectral norm of the Jacobian is bounded by

$$\|\mathrm{J}_s(\tilde{\boldsymbol{x}})\|_2\;<\;2\,\epsilon\,\beta\,. \tag{A156}$$

The norm of Jacobian at $\tilde{\boldsymbol{x}}$ is bounded

$$\|\mathrm{J}(\tilde{\boldsymbol{x}})\|_2\;\leqslant\;2\,\beta\,\|\boldsymbol{X}\|_2^2\,\epsilon\;\leqslant\;2\,\beta\,NM^2\,\epsilon\,. \tag{A157}$$

We used that the spectral norm $\|.\|_2$ is bounded by the Frobenius norm $\|.\|_F$ which can be expressed by the norm squared of its column vectors:

$$\|\boldsymbol{X}\|_2\;\leqslant\;\|\boldsymbol{X}\|_F\;=\;\sqrt{\sum_i \|\boldsymbol{x}_i\|^2}\,. \tag{A158}$$

Therefore

$$\|\boldsymbol{X}\|_2^2\;\leqslant\;N\,M^2\,. \tag{A159}$$

The norm of Jacobian of the fixed point iteration is bounded

$$\|\mathrm{J}^m\|_2\;\leqslant\;2\,\beta\,\|\boldsymbol{X}\|_2^2\,\epsilon\;\leqslant\;2\,\beta\,NM^2\,\epsilon\,. \tag{A160}$$

The separation of pattern $\boldsymbol{x}_i$ from data $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ is

$$\Delta_i\;=\;\min_{j,j\neq i}\left(\boldsymbol{x}_i^T\boldsymbol{x}_i\;-\;\boldsymbol{x}_i^T\boldsymbol{x}_j\right)\;=\;\boldsymbol{x}_i^T\boldsymbol{x}_i\;-\;\max_{j,j\neq i}\boldsymbol{x}_i^T\boldsymbol{x}_j\,. \tag{A161}$$

We need the separation $\tilde{\Delta}_i$ of $\tilde{\boldsymbol{x}} = \lambda\boldsymbol{\xi} + (1-\lambda)\boldsymbol{x}_i^*$ from the data:

$$\tilde{\Delta}_i \;=\; \min_{j,j\neq i} \left(\tilde{\boldsymbol{x}}^T\boldsymbol{x}_i \;-\; \tilde{\boldsymbol{x}}^T\boldsymbol{x}_j\right) \;. \tag{A162}$$

We compute a lower bound on $\tilde{\Delta}_i$. Using the Cauchy-Schwarz inequality, we obtain for $1 \leqslant j \leqslant N$:

$$\left|\tilde{\boldsymbol{x}}^T\boldsymbol{x}_j \;-\; \boldsymbol{x}_i^T\boldsymbol{x}_j\right| \leqslant \; \|\tilde{\boldsymbol{x}} \;-\; \boldsymbol{x}_i\| \, \|\boldsymbol{x}_j\| \;\leqslant\; \|\tilde{\boldsymbol{x}} \;-\; \boldsymbol{x}_i\| \, M \;. \tag{A163}$$

We have the lower bound

$$\begin{aligned} \tilde{\Delta}_i \;&\geq\; \min_{j,j\neq i} \left(\left(\boldsymbol{x}_i^T\boldsymbol{x}_i \;-\; \|\tilde{\boldsymbol{x}} \;-\; \boldsymbol{x}_i\| \, M\right) \;-\; \left(\boldsymbol{x}_i^T\boldsymbol{x}_j \;+\; \|\tilde{\boldsymbol{x}} \;-\; \boldsymbol{x}_i\| \, M\right)\right) \\ &=\; -\, 2 \, \|\tilde{\boldsymbol{x}} \;-\; \boldsymbol{x}_i\| \, M \;+\; \min_{j,j\neq i} \left(\boldsymbol{x}_i^T\boldsymbol{x}_i \;-\; \boldsymbol{x}_i^T\boldsymbol{x}_j\right) \;=\; \Delta_i \;-\; 2 \, \|\tilde{\boldsymbol{x}} \;-\; \boldsymbol{x}_i\| \, M \;. \end{aligned} \tag{A164}$$

Since

$$\begin{aligned} \|\tilde{\boldsymbol{x}} \;-\; \boldsymbol{x}_i\| \;&=\; \|\lambda\boldsymbol{\xi} + (1-\lambda)\boldsymbol{x}_i^* \;-\; \boldsymbol{x}_i\| \\ &\leqslant\; \lambda \, \|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\| \;+\; (1-\lambda) \, \|\boldsymbol{x}_i^* \;-\; \boldsymbol{x}_i\| \\ &\leqslant\; \max\{\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \;-\; \boldsymbol{x}_i\|\} \;, \end{aligned} \tag{A165}$$

we have

$$\tilde{\Delta}_i \;\geq\; \Delta_i \;-\; 2 \, \max\{\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \;-\; \boldsymbol{x}_i\|\} \, M \;. \tag{A166}$$

For the softmax component i we have:

$$\begin{aligned} &[\mathrm{softmax}(\beta \, \boldsymbol{X}^T\tilde{\boldsymbol{\xi}})]_i \;=\; \frac{1}{1 \;+\; \sum_{j\neq i} \exp(\beta \, (\tilde{\boldsymbol{\xi}}^T\boldsymbol{x}_j \;-\; \tilde{\boldsymbol{\xi}}^T\boldsymbol{x}_i))} \\ &\geq\; \frac{1}{1 \;+\; \sum_{j\neq i} \exp(-\, \beta \, (\Delta_i \;-\; 2 \, \max\{\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \;-\; \boldsymbol{x}_i\|\} \, M))} \\ &=\; \frac{1}{1 \;+\; (N-1)\exp(-\, \beta \, (\Delta_i \;-\; 2 \, \max\{\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \;-\; \boldsymbol{x}_i\|\} \, M))} \\ &=\; 1 \;-\; \frac{(N-1)\exp(-\, \beta \, (\Delta_i \;-\; 2 \, \max\{\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \;-\; \boldsymbol{x}_i\|\} \, M))}{1 \;+\; (N-1)\exp(-\, \beta \, (\Delta_i \;-\; 2 \, \max\{\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \;-\; \boldsymbol{x}_i\|\} \, M))} \\ &\geq\; 1 \;-\; (N-1)\exp(-\, \beta \, (\Delta_i \;-\; 2 \, \max\{\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \;-\; \boldsymbol{x}_i\|\} \, M)) \\ &=\; 1 \;-\; \epsilon \;. \end{aligned} \tag{A167}$$

Therefore

$$\epsilon \;=\; (N-1)\exp(-\, \beta \, (\Delta_i \;-\; 2 \, \max\{\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \;-\; \boldsymbol{x}_i\|\} \, M)) \;. \tag{A168}$$

We can bound the spectral norm of the Jacobian, which upper bounds the Lipschitz constant:

$$\|\mathrm{J}^m\|_2 \;\leqslant\; 2 \, \beta \, N \, M^2 \, (N-1)\exp(-\, \beta \, (\Delta_i \;-\; 2 \, \max\{\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \;-\; \boldsymbol{x}_i\|\} \, M)) \;. \tag{A169}$$

For a contraction mapping we require

$$\|\mathrm{J}^m\|_2 \;<\; 1 \;, \tag{A170}$$

which can be ensured by

$$2 \, \beta \, NM^2 \, (N-1)\exp(-\, \beta \, (\Delta_i \;-\; 2 \, \max\{\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \;-\; \boldsymbol{x}_i\|\} \, M)) \;<\; 1 \;. \tag{A171}$$

Solving this inequality for Δ_i gives

$$\Delta_i \;>\; 2 \, \max\{\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \;-\; \boldsymbol{x}_i\|\} \, M \;+\; \frac{1}{\beta} \, \ln\left(2 \, (N-1) \, N \, \beta \, M^2\right) \;. \tag{A172}$$

In an environment around $\boldsymbol{x}_i^*$ in which Eq. (A172) holds, f is a contraction mapping and every point converges under the iteration f to $\boldsymbol{x}_i^*$ when the iteration stays in the environment. After every iteration the mapped point $f(\boldsymbol{\xi})$ is closer to the fixed point $\boldsymbol{x}_i^*$ than the original point $\boldsymbol{x}_i$:

$$\|f(\boldsymbol{\xi}) \;-\; \boldsymbol{x}_i^*\| \;\leqslant\; \|\mathrm{J}^m\|_2 \, \|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i^*\| \;<\; \|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i^*\| \;. \tag{A173}$$

Using

$$\|f(\boldsymbol{\xi}) - \boldsymbol{x}_i^*\| \leqslant \|\mathrm{J}^m\|_2 \|\boldsymbol{\xi} - \boldsymbol{x}_i^*\| \leqslant \|\mathrm{J}^m\|_2 \|\boldsymbol{\xi} - f(\boldsymbol{\xi})\| + \|\mathrm{J}^m\|_2 \|f(\boldsymbol{\xi}) - \boldsymbol{x}_i^*\| , \tag{A174}$$

we obtain

$$\|f(\boldsymbol{\xi}) - \boldsymbol{x}_i^*\| \leqslant \frac{\|\mathrm{J}^m\|_2}{1 - \|\mathrm{J}^m\|_2} \|\boldsymbol{\xi} - f(\boldsymbol{\xi})\| . \tag{A175}$$

For large Δ_i the iteration is close to the fixed point even after one update. This has been confirmed in several experiments.

A2.4.4 Metastable States: Fixed Points Near Mean of Similar Patterns

The proof concept is the same as for a single pattern but now for the arithmetic mean of similar patterns.

Bound on the Jacobian. The Jacobian of the fixed point iteration is

$$\mathrm{J} = \beta \boldsymbol{X} \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right) \boldsymbol{X}^T = \boldsymbol{X} \mathrm{J}_s \boldsymbol{X}^T . \tag{A176}$$

If we consider p_i as the probability of selecting the vector $\boldsymbol{x}_i$, then we can define expectations as $\mathrm{E}_{\boldsymbol{p}}[f(\boldsymbol{x})] = \sum_{i=1}^N p_i f(\boldsymbol{x}_i)$. In this setting the matrix

$$\boldsymbol{X} \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right) \boldsymbol{X}^T \tag{A177}$$

is the covariance matrix of data $\boldsymbol{X}$ when its vectors are selected according to the probability $\boldsymbol{p}$:

$$\boldsymbol{X} \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right) \boldsymbol{X}^T = \boldsymbol{X} \mathrm{diag}(\boldsymbol{p}) \boldsymbol{X}^T - \boldsymbol{X}\boldsymbol{p}\boldsymbol{p}^T\boldsymbol{X}^T \tag{A178}$$

$$= \sum_{i=1}^N p_i \boldsymbol{x}_i \boldsymbol{x}_i^T - \left(\sum_{i=1}^N p_i \boldsymbol{x}_i\right) \left(\sum_{i=1}^N p_i \boldsymbol{x}_i\right)^T \tag{A179}$$

$$= \mathrm{E}_{\boldsymbol{p}}[\boldsymbol{x}\, \boldsymbol{x}^T] - \mathrm{E}_{\boldsymbol{p}}[\boldsymbol{x}]\, \mathrm{E}_{\boldsymbol{p}}[\boldsymbol{x}]^T = \mathrm{Var}_{\boldsymbol{p}}[\boldsymbol{x}] , \tag{A180}$$

therefore we have

$$\mathrm{J} = \beta\, \mathrm{Var}_{\boldsymbol{p}}[\boldsymbol{x}] . \tag{A181}$$

We now elaborate more on this interpretation as variance. Specifically the singular values of J (or in other words: the covariance) should be reasonably small. The singular values are the key to ensure convergence of the iteration Eq. (A46). Next we present some thoughts.

1. It's clear that the largest eigenvalue of the covariance matrix (equal to the largest singular value) is the variance in the direction of the eigenvector associated with the largest eigenvalue.

2. Furthermore the variance goes to zero as one p_i goes to one, since only one pattern is chosen and there is no variance.

3. The variance is reasonable small if all patterns are chosen with equal probability.

4. The variance is small if few similar patterns are chosen with high probability. If the patterns are sufficient similar, then the spectral norm of the covariance matrix is smaller than one.

The first three issues have already been adressed. Now we focus on the last one in greater detail. We assume that the first l patterns are much more probable (and similar to one another) than the other

patterns. Therefore, we define:

$$M \ := \ \max_i \|\boldsymbol{x}_i\| \ , \tag{A182}$$

$$\gamma \ = \ \sum_{i=l+1}^{N} p_i \ \leqslant \ \epsilon \ , \tag{A183}$$

$$1-\gamma \ = \ \sum_{i=1}^{l} p_i \ \geq \ 1 \ - \ \epsilon \ , \tag{A184}$$

$$\tilde{p}_i \ := \ \frac{p_i}{1-\gamma} \ \leqslant \ p_i/(1-\epsilon) \ , \tag{A185}$$

$$\sum_{i=1}^{l} \tilde{p}_i \ = \ 1 \ , \tag{A186}$$

$$\boldsymbol{m}_{\boldsymbol{x}} \ = \ \frac{1}{l} \ \sum_{i=1}^{l} \ \boldsymbol{x}_i \ , \tag{A187}$$

$$m_{\max} \ = \ \max_{1 \leqslant i \leqslant l} \|\boldsymbol{x}_i \ - \ \boldsymbol{m}_{\boldsymbol{x}}\| \ . \tag{A188}$$

M is an upper bound on the Euclidean norm of the patterns, which are vectors. ϵ is an upper bound on the probability γ of not choosing one of the first l patterns, while $1-\epsilon$ is a lower bound the probability $(1-\gamma)$ of choosing one of the first l patterns. $\boldsymbol{m}_{\boldsymbol{x}}$ is the arithmetic mean (the center) of the first l patterns. $m_{\max}$ is the maximal distance of the patterns to the center $\boldsymbol{m}_{\boldsymbol{x}}$. $\tilde{\boldsymbol{p}}$ is the probability $\boldsymbol{p}$ normalized for the first l patterns.

The variance of the first l patterns is

$$\begin{aligned} \mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}] \ &= \ \sum_{i=1}^{l} \tilde{p}_i \ \boldsymbol{x}_i \ \boldsymbol{x}_i^T \ - \ \left(\sum_{i=1}^{l} \tilde{p}_i \ \boldsymbol{x}_i \right) \ \left(\sum_{i=1}^{l} \tilde{p}_i \ \boldsymbol{x}_i \right)^T \\ &= \ \sum_{i=1}^{l} \tilde{p}_i \ \left(\boldsymbol{x}_i \ - \ \sum_{i=1}^{l} \tilde{p}_i \boldsymbol{x}_i \right) \ \left(\boldsymbol{x}_i \ - \ \sum_{i=1}^{l} \tilde{p}_i \boldsymbol{x}_i \right)^T \ . \end{aligned} \tag{A189}$$

Lemma A8. *With the definitions in Eq. (A182) to Eq. (A189), the following bounds on the norm $\|\mathrm{J}\|_2$ of the Jacobian of the fixed point iteration hold. The γ-bound for $\|\mathrm{J}\|_2$ is*

$$\|\mathrm{J}\|_2 \ \leqslant \ \beta \left((1-\gamma) \ m_{\max}^2 \ + \ \gamma \ 2 \ (2 \ - \ \gamma) \ M^2 \right) \tag{A190}$$

and the ϵ-bound for $\|\mathrm{J}\|_2$ is:

$$\|\mathrm{J}\|_2 \ \leqslant \ \beta \left(\ m_{\max}^2 \ + \ \epsilon \ 2 \ (2 \ - \ \epsilon) \ M^2 \right) \ . \tag{A191}$$

Proof. The variance $\mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}]$ can be expressed as:

$$(1-\gamma) \ \mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}] \ = \ \sum_{i=1}^{l} p_i \ \left(\boldsymbol{x}_i \ - \ \frac{1}{1-\gamma} \ \sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right) \ \left(\boldsymbol{x}_i \ - \ \frac{1}{1-\gamma} \ \sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right)^T \tag{A192}$$

$$\begin{aligned} &= \ \sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \ \boldsymbol{x}_i^T \ - \ \left(\sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right) \ \frac{1}{1-\gamma} \ \left(\sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right)^T \ - \ \frac{1}{1-\gamma} \ \left(\sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right) \ \left(\sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right)^T \\ &+ \ \frac{\sum_{i=1}^{l} p_i}{(1-\gamma)^2} \ \left(\sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right) \ \left(\sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right)^T \ = \ \sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \ \boldsymbol{x}_i^T \ - \ \frac{1}{1-\gamma} \ \left(\sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right) \ \left(\sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right)^T \\ &= \ \sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \ \boldsymbol{x}_i^T \ - \ \left(\sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right) \ \left(\sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right)^T \ + \ \left(1 \ - \ \frac{1}{1-\gamma} \right) \ \left(\sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right) \ \left(\sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right)^T \\ &= \ \sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \ \boldsymbol{x}_i^T \ - \ \left(\sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right) \ \left(\sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right)^T \ - \ \frac{\gamma}{1-\gamma} \ \left(\sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right) \ \left(\sum_{i=1}^{l} p_i \ \boldsymbol{x}_i \right)^T \ . \end{aligned}$$

Therefore, we have

$$\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T \; - \; \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right)^T \tag{A193}$$

$$= \; (1-\gamma) \, \mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}] \; + \; \frac{\gamma}{1-\gamma} \, \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right)^T \; .$$

We now can reformulate the Jacobian J:

$$\begin{aligned}
\mathrm{J} \; &= \; \beta \, \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T \; + \; \sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T \right. \\
&- \; \left. \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \; + \; \sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \; + \; \sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right)^T \right) \\
&= \; \beta \, \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T \; - \; \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right)^T \right. \\
&+ \; \sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T \; - \; \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right)^T \\
&- \; \left. \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right)^T \; - \; \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right)^T \right) \\
&= \; \beta \, \left((1-\gamma) \, \mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}] \; + \; \frac{\gamma}{1-\gamma} \, \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right)^T \right. \\
&+ \; \sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T \; - \; \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right)^T \\
&- \; \left. \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right)^T \; - \; \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right)^T \right) \; .
\end{aligned} \tag{A194}$$

The spectral norm of an outer product of two vectors is the product of the Euclidean norms of the vectors:

$$\left\| \boldsymbol{a} \boldsymbol{b}^T \right\|_2 \; = \; \sqrt{\lambda_{\max}(\boldsymbol{b} \boldsymbol{a}^T \boldsymbol{a} \boldsymbol{b}^T)} \; = \; \| \boldsymbol{a} \| \, \sqrt{\lambda_{\max}(\boldsymbol{b} \boldsymbol{b}^T)} \; = \; \| \boldsymbol{a} \| \, \| \boldsymbol{b} \| \; , \tag{A195}$$

since $\boldsymbol{b} \boldsymbol{b}^T$ has eigenvector $\boldsymbol{b} / \| \boldsymbol{b} \|$ with eigenvalue $\| \boldsymbol{b} \|^2$ and otherwise zero eigenvalues.

We now bound the norms of some matrices and vectors:

$$\left\| \sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right\| \; \leqslant \; \sum_{i=1}^{l} p_i \, \| \boldsymbol{x}_i \| \; \leqslant \; (1-\gamma) \, M \; , \tag{A196}$$

$$\left\| \sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right\| \; \leqslant \; \sum_{i=l+1}^{N} p_i \, \| \boldsymbol{x}_i \| \; \leqslant \; \gamma \, M \; , \tag{A197}$$

$$\left\| \sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T \right\|_2 \; \leqslant \; \sum_{i=l+1}^{N} p_i \left\| \boldsymbol{x}_i \, \boldsymbol{x}_i^T \right\|_2 \; = \; \sum_{i=l+1}^{N} p_i \, \| \boldsymbol{x}_i \|^2 \; \leqslant \; \sum_{i=l+1}^{N} p_i \, M^2 \; = \; \gamma \, M^2 \; . \tag{A198}$$

In order to bound the variance of the first l patterns, we compute the vector $\boldsymbol{a}$ that minimizes

$$f(\boldsymbol{a}) \;=\; \sum_{i=1}^{l} p_i \|\boldsymbol{x}_i \;-\; \boldsymbol{a}\|^2 \;=\; \sum_{i=1}^{l} p_i (\boldsymbol{x}_i \;-\; \boldsymbol{a})^T (\boldsymbol{x}_i \;-\; \boldsymbol{a}) \,. \tag{A199}$$

The solution to

$$\frac{\partial f(\boldsymbol{a})}{\partial \boldsymbol{a}} \;=\; 2 \sum_{i=1}^{N} p_i (\boldsymbol{a} \;-\; \boldsymbol{x}_i) \;=\; 0 \tag{A200}$$

is

$$\boldsymbol{a} \;=\; \sum_{i=1}^{N} p_i \boldsymbol{x}_i \,. \tag{A201}$$

The Hessian of f is positive definite since

$$\frac{\partial^2 f(\boldsymbol{a})}{\partial \boldsymbol{a}^2} \;=\; 2 \sum_{i=1}^{N} p_i \, \boldsymbol{I} \;=\; 2 \, \boldsymbol{I} \tag{A202}$$

and f is a convex function. Hence, the mean

$$\bar{\boldsymbol{x}} \;:=\; \sum_{i=1}^{N} p_i \, \boldsymbol{x}_i \tag{A203}$$

minimizes $\sum_{i=1}^{N} p_i \|\boldsymbol{x}_i - \boldsymbol{a}\|^2$. Therefore, we have

$$\sum_{i=1}^{l} p_i \|\boldsymbol{x}_i \;-\; \bar{\boldsymbol{x}}\|^2 \;\leqslant\; \sum_{i=1}^{l} p_i \|\boldsymbol{x}_i \;-\; \boldsymbol{m}_{\boldsymbol{x}}\|^2 \;\leqslant\; (1 \;-\; \gamma) \, m_{\max}^2 \,. \tag{A204}$$

We now bound the variance on the first l patterns:

$$\begin{aligned} (1-\gamma) \, \|\mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}]\|_2 \;&\leqslant\; \sum_{i=1}^{l} p_i \Big\| (\boldsymbol{x}_i \;-\; \bar{\boldsymbol{x}}) \; (\boldsymbol{x}_i \;-\; \bar{\boldsymbol{x}})^T \Big\|_2 \qquad \text{(A205)} \\ &=\; \sum_{i=1}^{l} p_i \|\boldsymbol{x}_i \;-\; \bar{\boldsymbol{x}}\|^2 \;\leqslant\; \sum_{i=1}^{l} p_i \|\boldsymbol{x}_i \;-\; \boldsymbol{m}_{\boldsymbol{x}}\|^2 \;\leqslant\; (1 \;-\; \gamma) \, m_{\max}^2 \,. \end{aligned}$$

We obtain for the spectral norm of J:

$$\begin{aligned} \|\mathrm{J}\|_2 \;&\leqslant\; \beta \left((1-\gamma) \, \|\mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}]\|_2 \right. \qquad \text{(A206)} \\ &+\; \frac{\gamma}{1-\gamma} \left\| \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right)^T \right\|_2 \\ &+\; \left\| \sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T \right\|_2 \;+\; \left\| \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right)^T \right\|_2 \\ &+\; \left. \left\| \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right)^T \right\|_2 \;+\; \left\| \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right)^T \right\|_2 \right) \\ &\leqslant\; \beta \left((1-\gamma) \, \|\mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}]\|_2 \;+\; \gamma \, (1-\gamma) \, M^2 \;+\; \gamma \, M^2 \;+\; \gamma^2 \, M^2 \;+ \right. \\ &\left. \gamma \, (1-\gamma) \, M^2 \;+\; \gamma \, (1-\gamma) \, M^2 \right) \\ &=\; \beta \left((1-\gamma) \;\; \|\mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}]\|_2 \;+\; \gamma \, 2 \, (2 \;-\; \gamma) \, M^2 \right) \,. \end{aligned}$$

Combining the previous two estimates immediately leads to Eq. (A190).

The function $h(x) = x2(2-x)$ has the derivative $h'(x) = 4(1-x)$. Therefore, $h(x)$ is monotone increasing for $x < 1$. For $0 \leqslant \gamma \leqslant \epsilon < 1$, we can immediately deduce that $\gamma 2(2-\gamma) \leqslant \epsilon 2(2-\epsilon)$. Since ϵ is larger than γ, we obtain the following ϵ-bound for $\|\mathrm{J}\|_2$:

$$\|\mathrm{J}\|_2 \ \leqslant \ \beta \left(\, m_{\max}^2 \ + \ \epsilon \, 2 \, (2 \ - \ \epsilon) \, M^2 \right) \ . \tag{A207}$$

□

We revisit the bound on $(1-\gamma)\ \mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}]$. The trace $\sum_{k=1}^{d} e_k$ is the sum of the eigenvalues e_k. The spectral norm is equal to the largest eigenvalue e_1, that is, the largest singular value. We obtain:

$$\begin{aligned} \left\| \mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}] \right\|_2 \ &= \ \mathrm{Tr} \left(\sum_{i=1}^{l} p_i \, (\boldsymbol{x}_i \ - \ \bar{\boldsymbol{x}}) \ (\boldsymbol{x}_i \ - \ \bar{\boldsymbol{x}})^T \right) \ - \ \sum_{k=2}^{d} e_k \\ &= \ \sum_{i=1}^{l} p_i \mathrm{Tr} \left((\boldsymbol{x}_i \ - \ \bar{\boldsymbol{x}}) \ (\boldsymbol{x}_i \ - \ \bar{\boldsymbol{x}})^T \right) \ - \ \sum_{k=2}^{d} e_k \\ &= \ \sum_{i=1}^{l} p_i \| \boldsymbol{x}_i \ - \ \bar{\boldsymbol{x}} \|^2 \ - \ \sum_{k=2}^{d} e_k \ . \end{aligned} \tag{A208}$$

Therefore, the tightness of the bound depends on eigenvalues which are not the largest. That is variations which are not along the strongest variation weaken the bound.

Proof of a fixed point by Banach Fixed Point Theorem. Without restricting the generality, we assume that the first l patterns are much more probable (and similar to one another) than the other patterns. Therefore, we define:

$$M \ := \ \max_i \| \boldsymbol{x}_i \| \ , \tag{A209}$$

$$\gamma \ = \ \sum_{i=l+1}^{N} p_i \ \leqslant \ \epsilon \ , \tag{A210}$$

$$1 - \gamma \ = \ \sum_{i=1}^{l} p_i \ \geq \ 1 \ - \ \epsilon \ , \tag{A211}$$

$$\tilde{p}_i \ := \ \frac{p_i}{1-\gamma} \ \leqslant \ p_i / (1-\epsilon) \ , \tag{A212}$$

$$\sum_{i=1}^{l} \tilde{p}_i \ = \ 1 \ , \tag{A213}$$

$$\boldsymbol{m}_{\boldsymbol{x}} \ = \ \frac{1}{l} \sum_{i=1}^{l} \boldsymbol{x}_i \ , \tag{A214}$$

$$m_{\max} \ = \ \max_{1 \leqslant i \leqslant l} \| \boldsymbol{x}_i \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \ . \tag{A215}$$

M is an upper bound on the Euclidean norm of the patterns, which are vectors. ϵ is an upper bound on the probability γ of not choosing one of the first l patterns, while $1-\epsilon$ is a lower bound the probability $(1-\gamma)$ of choosing one of the first l patterns. $\boldsymbol{m}_{\boldsymbol{x}}$ is the arithmetic mean (the center) of the first l patterns. $m_{\max}$ is the maximal distance of the patterns to the center $\boldsymbol{m}_{\boldsymbol{x}}$. $\tilde{\boldsymbol{p}}$ is the probability $\boldsymbol{p}$ normalized for the first l patterns.

Mapped vectors stay in a compact environment. We show that if $\boldsymbol{m}_{\boldsymbol{x}}$ is sufficient dissimilar to other $\boldsymbol{x}_j$ with $l < j$ then there is an compact environment of $\boldsymbol{m}_{\boldsymbol{x}}$ (a sphere) where the fixed point iteration maps this environment into itself. The idea of the proof is to define a sphere around $\boldsymbol{m}_{\boldsymbol{x}}$ for which the points from the sphere are mapped by f into the sphere.

We first need following lemma which bounds the distance $\| \boldsymbol{m}_{\boldsymbol{x}} \ - \ f(\boldsymbol{\xi}) \|$ of a $\boldsymbol{\xi}$ which is close to $\boldsymbol{m}_{\boldsymbol{x}}$.

Lemma A9. *For a query $\boldsymbol{\xi}$ and data $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, we define*

$$0 \leqslant c = \min_{j,l<j} \left(\boldsymbol{\xi}^T \boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{\xi}^T \boldsymbol{x}_j \right) = \boldsymbol{\xi}^T \boldsymbol{m}_{\boldsymbol{x}} - \max_{j,l<j} \boldsymbol{\xi}^T \boldsymbol{x}_j \,. \tag{A216}$$

The following holds:

$$\| \boldsymbol{m}_{\boldsymbol{x}} - f(\boldsymbol{\xi}) \| \leqslant m_{\max} + 2\,\gamma\,M \leqslant m_{\max} + 2\,\epsilon\,M \,, \tag{A217}$$

where

$$M = \max_i \| \boldsymbol{x}_i \| \,, \tag{A218}$$

$$\epsilon = (N-l)\, \exp(-\beta\, c) \,. \tag{A219}$$

Proof. Let $s = \arg\max_{j, j \leqslant l} \boldsymbol{\xi}^T \boldsymbol{x}_j$, therefore $\boldsymbol{\xi}^T \boldsymbol{m}_{\boldsymbol{x}} = \frac{1}{l} \sum_{i=1}^l \boldsymbol{\xi}^T \boldsymbol{x}_i \leqslant \frac{1}{l} \sum_{i=1}^l \boldsymbol{\xi}^T \boldsymbol{x}_s = \boldsymbol{\xi}^T \boldsymbol{x}_s$. For softmax components j with $l < j$ we have

$$[\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})]_j = \frac{\exp(\beta\, (\boldsymbol{\xi}^T \boldsymbol{x}_j - \boldsymbol{\xi}^T \boldsymbol{x}_s))}{1 + \sum_{k, k \neq s} \exp(\beta\, (\boldsymbol{\xi}^T \boldsymbol{x}_k - \boldsymbol{\xi}^T \boldsymbol{x}_s))} \leqslant \exp(-\beta\, c) = \frac{\epsilon}{N-l} \,, \tag{A220}$$

since $\boldsymbol{\xi}^T \boldsymbol{x}_s - \boldsymbol{\xi}^T \boldsymbol{x}_j \geq \boldsymbol{\xi}^T \boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{\xi}^T \boldsymbol{x}_j$ for each j with $l < j$, therefore $\boldsymbol{\xi}^T \boldsymbol{x}_s - \boldsymbol{\xi}^T \boldsymbol{x}_j \geq c$

The iteration f can be written as

$$f(\boldsymbol{\xi}) = \boldsymbol{X} \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) = \sum_{j=1}^N \boldsymbol{x}_j\, [\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})]_j \,. \tag{A221}$$

We set $p_i = [\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})]_i$, therefore $\sum_{i=1}^l p_i = 1 - \gamma \geq 1 - \epsilon$ and $\sum_{i=l+1}^N p_i = \gamma \leqslant \epsilon$. Therefore

$$\left\| \boldsymbol{m}_{\boldsymbol{x}} - \sum_{j=1}^l \frac{p_j}{1-\gamma}\, \boldsymbol{x}_j \right\|^2 = \left\| \sum_{j=1}^l \frac{p_j}{1-\gamma}\, (\boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{x}_j) \right\|^2 \tag{A222}$$

$$= \sum_{j=1,k=1}^l \frac{p_j}{1-\gamma} \frac{p_k}{1-\gamma}\, (\boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{x}_j)^T\, (\boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{x}_k)$$

$$= \frac{1}{2} \sum_{j=1,k=1}^l \frac{p_j}{1-\gamma} \frac{p_k}{1-\gamma} \left(\| \boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{x}_j \|^2 + \| \boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{x}_k \|^2 - \| \boldsymbol{x}_j - \boldsymbol{x}_k \|^2 \right)$$

$$= \sum_{j=1}^l \frac{p_j}{1-\gamma} \| \boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{x}_j \|^2 - \frac{1}{2} \sum_{j=1,k=1}^l \frac{p_j}{1-\gamma} \frac{p_k}{1-\gamma} \| \boldsymbol{x}_j - \boldsymbol{x}_k \|^2$$

$$\leqslant \sum_{j=1}^l \frac{p_j}{1-\gamma} \| \boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{x}_j \|^2 \leqslant m_{\max}^2 \,.$$

It follows that

$$\left\| \boldsymbol{m}_{\boldsymbol{x}} - \sum_{j=1}^l \frac{p_j}{1-\gamma}\, \boldsymbol{x}_j \right\| \leqslant m_{\max} \tag{A223}$$

We now can bound $\|\boldsymbol{m}_{\boldsymbol{x}} \,-\, f(\boldsymbol{\xi})\|$:

$$
\begin{aligned}
\|\boldsymbol{m}_{\boldsymbol{x}} \,-\, f(\boldsymbol{\xi})\| \;&=\; \left\| \boldsymbol{m}_{\boldsymbol{x}} \,-\, \sum_{j=1}^{N} p_j \, \boldsymbol{x}_j \right\| \\
&=\; \left\| \boldsymbol{m}_{\boldsymbol{x}} \,-\, \sum_{j=1}^{l} p_j \, \boldsymbol{x}_j \,-\, \sum_{j=l+1}^{N} p_j \, \boldsymbol{x}_j \right\| \\
&=\; \left\| \boldsymbol{m}_{\boldsymbol{x}} \,-\, \sum_{j=1}^{l} \frac{p_j}{1-\gamma} \, \boldsymbol{x}_j \,+\, \frac{\gamma}{1-\gamma} \sum_{j=1}^{l} p_j \, \boldsymbol{x}_j \,-\, \sum_{j=l+1}^{N} p_j \, \boldsymbol{x}_j \right\| \\
&\leqslant\; \left\| \boldsymbol{m}_{\boldsymbol{x}} \,-\, \sum_{j=1}^{l} \frac{p_j}{1-\gamma} \, \boldsymbol{x}_j \right\| \,+\, \frac{\gamma}{1-\gamma} \left\| \sum_{j=1}^{l} p_j \, \boldsymbol{x}_j \right\| \,+\, \left\| \sum_{j=l+1}^{N} p_j \, \boldsymbol{x}_j \right\| \\
&\leqslant\; \left\| \boldsymbol{m}_{\boldsymbol{x}} \,-\, \sum_{j=1}^{l} \frac{p_j}{1-\gamma} \, \boldsymbol{x}_j \right\| \,+\, \frac{\gamma}{1-\gamma} \sum_{j=1}^{l} p_j \, M \,+\, \sum_{j=l+1}^{N} p_j \, M \\
&\leqslant\; \left\| \boldsymbol{m}_{\boldsymbol{x}} \,-\, \sum_{j=1}^{l} \frac{p_j}{1-\gamma} \, \boldsymbol{x}_j \right\| \,+\, 2 \, \gamma \, M \\
&\leqslant\; m_{\max} \,+\, 2 \, \gamma \, M \;\leqslant\; m_{\max} \,+\, 2 \, \epsilon \, M \,,
\end{aligned}
\tag{A224}
$$

where we applied Eq. (A222) in the penultimate inequality. This is the statement of the lemma. □

The separation of the center (the arithmetic mean) $\boldsymbol{m}_{\boldsymbol{x}}$ of the first l from data $\boldsymbol{X} = (\boldsymbol{x}_{l+1}, \ldots, \boldsymbol{x}_N)$ is Δ_m, defined as

$$
\Delta_m \;=\; \min_{j, l<j} \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} \,-\, \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \right) \;=\; \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} \,-\, \max_{j, l<j} \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \,. \tag{A225}
$$

The center is separated from the other data $\boldsymbol{x}_j$ with $l < j$ if $0 < \Delta_m$. By the same arguments as in Eq. (A129), Δ_m can also be expressed as

$$
\begin{aligned}
\Delta_m \;&=\; \min_{j, l<j} \frac{1}{2} \left(\|\boldsymbol{m}_{\boldsymbol{x}}\|^2 \,-\, \|\boldsymbol{x}_j\|^2 \,+\, \|\boldsymbol{m}_{\boldsymbol{x}} \,-\, \boldsymbol{x}_j\|^2 \right) \\
&=\; \frac{1}{2} \|\boldsymbol{m}_{\boldsymbol{x}}\|^2 \,-\, \frac{1}{2} \max_{j, l<j} \left(\|\boldsymbol{x}_j\|^2 \,-\, \|\boldsymbol{m}_{\boldsymbol{x}} \,-\, \boldsymbol{x}_j\|^2 \right) \,.
\end{aligned}
\tag{A226}
$$

For $\|\boldsymbol{m}_{\boldsymbol{x}}\| = \|\boldsymbol{x}_j\|$ we have $\Delta_m = 1/2 \min_{j, l<j} \|\boldsymbol{m}_{\boldsymbol{x}} \,-\, \boldsymbol{x}_j\|^2$.

Next we define the sphere where we want to apply Banach fixed point theorem.

Definition 4 (Sphere S_m)**.** *The sphere S_m is defined as*

$$
\mathrm{S}_m \;:=\; \left\{ \boldsymbol{\xi} \mid \|\boldsymbol{\xi} \,-\, \boldsymbol{m}_{\boldsymbol{x}}\| \;\leqslant\; \frac{1}{\beta \, m_{\max}} \right\} \,. \tag{A227}
$$

Lemma A10. *With $\boldsymbol{\xi}$ given, if the assumptions*

A1: $\boldsymbol{\xi}$ is inside sphere: $\boldsymbol{\xi} \in \mathrm{S}_m$,

A2: the center $\boldsymbol{m}_{\boldsymbol{x}}$ is well separated from other data $\boldsymbol{x}_j$ with $l < j$:

$$
\Delta_m \;\geq\; \frac{2 \, M}{\beta \, m_{\max}} \,-\, \frac{1}{\beta} \, \ln \left(\frac{1 \,-\, \beta \, m_{\max}^2}{2 \, \beta \, (N-l) \, M \, \max\{m_{\max} \,,\, 2 \, M\}} \right) \,, \tag{A228}
$$

A3: the distance $m_{\max}$ of similar patterns to the center is sufficient small:

$$
\beta \, m_{\max}^2 \;\leqslant\; 1 \tag{A229}
$$

hold, then $f(\boldsymbol{\xi}) \in \mathrm{S}_m$. Therefore, under conditions (A2) and (A3), f is a mapping from S_m into S_m.

Proof. We need the separation $\tilde{\Delta}_m$ of $\boldsymbol{\xi}$ from the rest of the data, which is the last $N-l$ data points $\boldsymbol{X} = (\boldsymbol{x}_{l+1}, \ldots, \boldsymbol{x}_N)$.

$$\tilde{\Delta}_m \ = \ \min_{j,l<j} \left(\boldsymbol{\xi}^T \boldsymbol{m}_{\boldsymbol{x}} \ - \ \boldsymbol{\xi}^T \boldsymbol{x}_j \right) \ . \tag{A230}$$

Using the Cauchy-Schwarz inequality, we obtain for $l+1 \leqslant j \leqslant N$:

$$\left| \boldsymbol{\xi}^T \boldsymbol{x}_j \ - \ \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \right| \leqslant \ \| \boldsymbol{\xi} \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \, \| \boldsymbol{x}_j \| \ \leqslant \ \| \boldsymbol{\xi} \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \, M \ . \tag{A231}$$

We have the lower bound

$$\begin{aligned} \tilde{\Delta}_m \ &\geq \ \min_{j,l<j} \left(\left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} \ - \ \| \boldsymbol{\xi} \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \, M \right) \ - \ \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \ + \ \| \boldsymbol{\xi} \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \, M \right) \right) \\ &= \ - \, 2 \, \| \boldsymbol{\xi} \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \, M \ + \ \min_{j,l<j} \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} \ - \ \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \right) \ = \ \Delta_m \ - \ 2 \, \| \boldsymbol{\xi} \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \, M \\ &\geq \ \Delta_m \ - \ 2 \, \frac{M}{\beta \, m_{\max}} \ , \end{aligned} \tag{A232}$$

where we used the assumption (A1) of the lemma.

From the proof in Lemma A9 we have

$$\sum_{i=1}^{l} p_i \ \geq \ 1 \ - \ (N-l) \, \exp(- \, \beta \, \tilde{\Delta}_m) \ = \ 1 \ - \ \tilde{\epsilon} \, , \tag{A233}$$

$$\sum_{i=l+1}^{N} p_i \ \leqslant \ (N-l) \, \exp(- \, \beta \, \tilde{\Delta}_m) \ = \ \tilde{\epsilon} \, . \tag{A234}$$

Lemma A9 states that

$$\begin{aligned} \| \boldsymbol{m}_{\boldsymbol{x}} \ - \ f(\boldsymbol{\xi}) \| \ &\leqslant \ m_{\max} \ + \ 2 \, \tilde{\epsilon} \, M \\ &\leqslant \ m_{\max} \ + \ 2 \, (N-l) \, \exp(- \, \beta \, \tilde{\Delta}_m) \, M \, . \\ &\leqslant \ m_{\max} \ + \ 2 \, (N-l) \, \exp(- \, \beta \, (\Delta_m \ - \ 2 \, \frac{M}{\beta \, m_{\max}})) \, M \, . \end{aligned} \tag{A235}$$

Therefore, we have

$$\begin{aligned} &\| \boldsymbol{m}_{\boldsymbol{x}} \ - \ f(\boldsymbol{\xi}) \| \ \leqslant \ m_{\max} \ + \ 2 \, (N-l) \, \exp \left(- \, \beta \, (\Delta_m \ - \ 2 \, \frac{M}{\beta \, m_{\max}}) \right) \, M \\ &\leqslant \ m_{\max} \ + \ 2 \, (N-l) \, \exp \left(- \, \beta \, \left(\frac{2 \, M}{\beta \, m_{\max}} \ - \right. \right. \\ &\left. \left. \frac{1}{\beta} \, \ln \left(\frac{1 \ - \ \beta \, m_{\max}^2}{2 \, \beta \, (N-l) \, M \ \max\{m_{\max} \, , \, 2 \, M\}} \right) \ - \ 2 \, \frac{M}{\beta \, m_{\max}} \right) \right) \, M \\ &= \ m_{\max} \ + \ 2 \, (N-l) \, \frac{1 \ - \ \beta \, m_{\max}^2}{2 \, \beta \, (N-l) \, M \ \max\{m_{\max} \, , \, 2 \, M\}} \, M \\ &\leqslant \ m_{\max} \ + \ \frac{1 \ - \ \beta \, m_{\max}^2}{\beta \, m_{\max}} \ = \ \frac{1}{\beta \, m_{\max}} \, , \end{aligned} \tag{A236}$$

where we used assumption (A2) of the lemma. Therefore, $f(\boldsymbol{\xi})$ is a mapping from the sphere S_m into the sphere S_m.

$$m_{\max} = \max_{1 \leqslant i \leqslant l} \|\boldsymbol{x}_i - \boldsymbol{m}_{\boldsymbol{x}}\| \tag{A237}$$

$$= \max_{1 \leqslant i \leqslant l} \left\| \boldsymbol{x}_i - 1/l \sum_{j=1}^{l} \boldsymbol{x}_j \right\| \tag{A238}$$

$$= \max_{1 \leqslant i \leqslant l} \left\| 1/l \sum_{j=1}^{l} (\boldsymbol{x}_i - \boldsymbol{x}_j) \right\| \tag{A239}$$

$$\leqslant \max_{1 \leqslant i,j \leqslant l} \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \tag{A240}$$

$$\leqslant \max_{1 \leqslant i \leqslant l} \|\boldsymbol{x}_i\| + \max_{1 \leqslant j \leqslant l} \|\boldsymbol{x}_i\| \tag{A241}$$

$$\leqslant 2M \tag{A242}$$

□

Contraction mapping. For applying Banach fixed point theorem we need to show that f is contraction in the compact environment S_m.

Lemma A11. *Assume that*

A1:

$$\Delta_m \ \geq \ \frac{2\,M}{\beta\,m_{\max}} \ - \ \frac{1}{\beta} \ \ln\left(\frac{1 \ - \ \beta\,m_{\max}^2}{2\,\beta\,(N-l)\,M \ \max\{m_{\max}\,,\ 2\,M\}} \right) , \tag{A243}$$

and

A2:

$$\beta\,m_{\max}^2 \ \leqslant \ 1\,, \tag{A244}$$

then f is a contraction mapping in S_m.

Proof. The version of the mean value theorem Lemma A32 states for the symmetric $\mathrm{J}^m = \int_0^1 \mathrm{J}(\lambda\boldsymbol{\xi} + (1-\lambda)\boldsymbol{m}_{\boldsymbol{x}})\,\mathrm{d}\lambda$:

$$f(\boldsymbol{\xi}) \ = \ f(\boldsymbol{m}_{\boldsymbol{x}}) \ + \ \mathrm{J}^m \ (\boldsymbol{\xi} \ - \ \boldsymbol{m}_{\boldsymbol{x}})\,. \tag{A245}$$

In complete analogy to Lemma A6, we get:

$$\|f(\boldsymbol{\xi}) \ - \ f(\boldsymbol{m}_{\boldsymbol{x}})\| \ \leqslant \ \|\mathrm{J}^m\|_2 \ \|\boldsymbol{\xi} \ - \ \boldsymbol{m}_{\boldsymbol{x}}\|\,. \tag{A246}$$

We define $\tilde{\boldsymbol{\xi}} = \lambda\boldsymbol{\xi} + (1-\lambda)\boldsymbol{m}_{\boldsymbol{x}}$ for some $\lambda \in [0,1]$. We need the separation $\tilde{\Delta}_m$ of $\tilde{\boldsymbol{\xi}}$ from the rest of the data, which is the last $N-l$ data points $\boldsymbol{X} = (\boldsymbol{x}_{l+1}, \ldots, \boldsymbol{x}_N)$.

$$\tilde{\Delta}_m \ = \ \min_{j,l<j} \left(\tilde{\boldsymbol{\xi}}^T \boldsymbol{m}_{\boldsymbol{x}} \ - \ \tilde{\boldsymbol{\xi}}^T \boldsymbol{x}_j \right) . \tag{A247}$$

From the proof in Lemma A9 we have

$$\tilde{\epsilon} \ = \ (N-l) \ \exp(-\,\beta\,\tilde{\Delta}_m)\,, \tag{A248}$$

$$\sum_{i=1}^{l} p_i(\tilde{\boldsymbol{\xi}}) \ \geq \ 1 \ - \ (N-l) \ \exp(-\,\beta\,\tilde{\Delta}_m) \ = \ 1 \ - \ \tilde{\epsilon}\,, \tag{A249}$$

$$\sum_{i=l+1}^{N} p_i(\tilde{\boldsymbol{\xi}}) \ \leqslant \ (N-l) \ \exp(-\,\beta\,\tilde{\Delta}_m) \ = \ \tilde{\epsilon}\,. \tag{A250}$$

We first compute an upper bound on $\tilde{\epsilon}$. Using the Cauchy-Schwarz inequality, we obtain for $l+1 \leqslant j \leqslant N$:

$$\left| \tilde{\boldsymbol{\xi}}^T \boldsymbol{x}_j \, - \, \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \right| \leqslant \left\| \tilde{\boldsymbol{\xi}} \, - \, \boldsymbol{m}_{\boldsymbol{x}} \right\| \| \boldsymbol{x}_j \| \, \leqslant \, \left\| \tilde{\boldsymbol{\xi}} \, - \, \boldsymbol{m}_{\boldsymbol{x}} \right\| M \, . \tag{A251}$$

We have the lower bound on $\tilde{\Delta}_m$:

$$\begin{aligned} \tilde{\Delta}_m \, &\geq \, \min_{j,l<j} \left(\left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} \, - \, \left\| \tilde{\boldsymbol{\xi}} \, - \, \boldsymbol{m}_{\boldsymbol{x}} \right\| M \right) \, - \, \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \, + \, \left\| \tilde{\boldsymbol{\xi}} \, - \, \boldsymbol{m}_{\boldsymbol{x}} \right\| M \right) \right) \\ &= \, - \, 2 \left\| \tilde{\boldsymbol{\xi}} \, - \, \boldsymbol{m}_{\boldsymbol{x}} \right\| M \, + \, \min_{j,l<j} \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} \, - \, \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \right) \, = \, \Delta_m \, - \, 2 \left\| \tilde{\boldsymbol{\xi}} \, - \, \boldsymbol{m}_{\boldsymbol{x}} \right\| M \\ &\geq \, \Delta_m \, - \, 2 \, \| \boldsymbol{\xi} \, - \, \boldsymbol{m}_{\boldsymbol{x}} \| \, M \, . \end{aligned} \tag{A252}$$

where we used $\left\| \tilde{\boldsymbol{\xi}} - \boldsymbol{m}_{\boldsymbol{x}} \right\| = \lambda \| \boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}} \| \leqslant \| \boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}} \|$. We obtain the upper bound on $\tilde{\epsilon}$:

$$\begin{aligned} \tilde{\epsilon} \, &\leqslant \, (N-l) \, \exp \left(- \, \beta \, \left(\Delta_m \, - \, 2 \, \| \boldsymbol{\xi} \, - \, \boldsymbol{m}_{\boldsymbol{x}} \| \, M \right) \right) \\ &\leqslant \, (N-l) \, \exp \left(- \, \beta \, \left(\Delta_m \, - \, \frac{2 \, M}{\beta \, m_{\max}} \right) \right) \, . \end{aligned} \tag{A253}$$

where we used that in the sphere S_i holds:

$$\| \boldsymbol{\xi} \, - \, \boldsymbol{m}_{\boldsymbol{x}} \| \, \leqslant \, \frac{1}{\beta \, m_{\max}} \, , \tag{A254}$$

therefore

$$2 \, \| \boldsymbol{\xi} \, - \, \boldsymbol{m}_{\boldsymbol{x}} \| \, M \, \leqslant \, \frac{2 \, M}{\beta \, m_{\max}} \, . \tag{A255}$$

Next we compute a lower bound on $\tilde{\epsilon}$ and to this end start with the upper bound on $\tilde{\Delta}_m$ using the same arguments as in Eq. (A147) in combination with Eq. (A255).

$$\begin{aligned} \tilde{\Delta}_m \, &\geq \, \min_{j,l<j} \left(\left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} \, + \, \left\| \tilde{\boldsymbol{\xi}} \, - \, \boldsymbol{m}_{\boldsymbol{x}} \right\| M \right) \, - \, \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \, - \, \left\| \tilde{\boldsymbol{\xi}} \, - \, \boldsymbol{m}_{\boldsymbol{x}} \right\| M \right) \right) \\ &= \, 2 \left\| \tilde{\boldsymbol{\xi}} \, - \, \boldsymbol{m}_{\boldsymbol{x}} \right\| M \, + \, \min_{j,l<j} \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} \, - \, \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \right) \, = \, \Delta_m \, + \, 2 \left\| \tilde{\boldsymbol{\xi}} \, - \, \boldsymbol{m}_{\boldsymbol{x}} \right\| M \\ &\geq \, \Delta_m \, + \, 2 \, \| \boldsymbol{\xi} \, - \, \boldsymbol{m}_{\boldsymbol{x}} \| \, M \, . \end{aligned} \tag{A256}$$

where we used $\left\| \tilde{\boldsymbol{\xi}} - \boldsymbol{m}_{\boldsymbol{x}} \right\| = \lambda \| \boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}} \| \leqslant \| \boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}} \|$. We obtain the lower bound on $\tilde{\epsilon}$:

$$\tilde{\epsilon} \, \geq \, (N-l) \, \exp \left(- \, \beta \, \left(\Delta_m \, + \, \frac{2 \, M}{\beta \, m_{\max}} \right) \right) \, , \tag{A257}$$

where we used that in the sphere S_i holds:

$$\| \boldsymbol{\xi} \, - \, \boldsymbol{m}_{\boldsymbol{x}} \| \, \leqslant \, \frac{1}{\beta \, m_{\max}} \, , \tag{A258}$$

therefore

$$2 \, \| \boldsymbol{\xi} \, - \, \boldsymbol{m}_{\boldsymbol{x}} \| \, M \, \leqslant \, \frac{2 \, M}{\beta \, m_{\max}} \, . \tag{A259}$$

From Lemma A8 we have

$$\begin{aligned} \left\| \mathrm{J}(\tilde{\boldsymbol{\xi}}) \right\|_2 \, &\leqslant \, \beta \left(\, m_{\max}^2 \, + \, \tilde{\epsilon} \, 2 \, (2 \, - \, \tilde{\epsilon}) \, M^2 \right) \\ &= \, \beta \left(m_{\max}^2 \, + \, \tilde{\epsilon} 4 \, M^2 \, - \, 2 \, \tilde{\epsilon}^2 \, M^2 \right) \\ &\leqslant \, \beta \left(m_{\max}^2 \, + \, (N-l) \, \exp \left(- \, \beta \, \left(\Delta_m \, - \, \frac{2 \, M}{\beta \, m_{\max}} \right) \right) 4 \, M^2 \, - \right. \\ &\left. 2 \, (N-l)^2 \, \exp \left(- \, 2 \, \beta \, \left(\Delta_m \, + \, \frac{2 \, M}{\beta \, m_{\max}} \right) \right) \, M^2 \right) \, . \end{aligned} \tag{A260}$$

The bound Eq. (A260) holds for the mean J^m, too, since it averages over $\mathrm{J}(\tilde{\boldsymbol{\xi}})$:

$$\|\mathrm{J}^m\|_2 \leqslant \beta \left(m_{\max}^2 + (N-l) \exp\left(-\beta \left(\Delta_m - \frac{2\,M}{\beta\, m_{\max}} \right) \right) 4\,M^2 - 2\,(N-l)^2 \exp\left(-2\,\beta \left(\Delta_m + \frac{2\,M}{\beta\, m_{\max}} \right) \right) M^2 \right) . \tag{A261}$$

The assumption of the lemma is

$$\Delta_m \geq \frac{2\,M}{\beta\, m_{\max}} - \frac{1}{\beta} \ln\left(\frac{1 - \beta\, m_{\max}^2}{2\,\beta\,(N-l)\,M\, \max\{m_{\max}\,,\, 2\,M\}} \right) , \tag{A262}$$

Therefore, we have

$$\Delta_m - \frac{2\,M}{\beta\, m_{\max}} \geq -\frac{1}{\beta} \ln\left(\frac{1 - \beta\, m_{\max}^2}{2\,\beta\,(N-l)\,M\, \max\{m_{\max}\,,\, 2\,M\}} \right) . \tag{A263}$$

Therefore, the spectral norm $\|\mathrm{J}^m\|_2$ can be bounded by:

$$\begin{aligned}
&\|\mathrm{J}^m\|_2 \leqslant \qquad\qquad (\text{A264})\\
&\beta \left(m_{\max}^2 + (N-l) \exp\left(-\beta \left(-\frac{1}{\beta} \ln\left(\frac{1 - \beta\, m_{\max}^2}{2\,\beta\,(N-l)\,M\, \max\{m_{\max}\,,\, 2\,M\}} \right) \right) \right) \right.\\
&\left. 4\,M^2 - 2\,(N-l)^2 \exp\left(-2\,\beta \left(\Delta_m + \frac{2\,M}{\beta\, m_{\max}} \right) \right) M^2 \right)\\
&= \beta \left(m_{\max}^2 + (N-l) \exp\left(\ln\left(\frac{1 - \beta\, m_{\max}^2}{2\,\beta\,(N-l)\,M\, \max\{m_{\max}\,,\, 2\,M\}} \right) \right) \right.\\
&\left. 4\,M^2 - 2\,(N-l)^2 \exp\left(-2\,\beta \left(\Delta_m + \frac{2\,M}{\beta\, m_{\max}} \right) \right) M^2 \right)\\
&= \beta \left(m_{\max}^2 + (N-l) \frac{1 - \beta\, m_{\max}^2}{2\,\beta\,(N-l)\,M\, \max\{m_{\max}\,,\, 2\,M\}} 4\,M^2 - \right.\\
&\left. 2\,(N-l)^2 \exp\left(-2\,\beta \left(\Delta_m + \frac{2\,M}{\beta\, m_{\max}} \right) \right) M^2 \right)\\
&= \beta m_{\max}^2 + \frac{1 - \beta\, m_{\max}^2}{\max\{m_{\max}\,,\, 2\,M\}} 2\,M -\\
&\beta\, 2\,(N-l)^2 \exp\left(-2\,\beta \left(\Delta_m + \frac{2\,M}{\beta\, m_{\max}} \right) \right) M^2\\
&\leqslant \beta m_{\max}^2 + 1 - \beta\, m_{\max}^2 - \beta\, 2\,(N-l)^2 \exp\left(-2\,\beta \left(\Delta_m + \frac{2\,M}{\beta\, m_{\max}} \right) \right) M^2\\
&= 1 - \beta\, 2\,(N-l)^2 \exp\left(-2\,\beta \left(\Delta_m + \frac{2\,M}{\beta\, m_{\max}} \right) \right) M^2 < 1\,.
\end{aligned}$$

For the last but one inequality we used $2M \leqslant \max\{m_{\max}, 2M\}$.

Therefore, f is a contraction mapping in S_m. □

Banach Fixed Point Theorem. Now we have all ingredients to apply Banach fixed point theorem.

Lemma A12. *Assume that*

A1:

$$\Delta_m \geq \frac{2\,M}{\beta\, m_{\max}} - \frac{1}{\beta} \ln\left(\frac{1 - \beta\, m_{\max}^2}{2\,\beta\,(N-l)\,M\, \max\{m_{\max}\,,\, 2\,M\}} \right) , \tag{A265}$$

and

A2:

$$\beta\, m_{\max}^2 \leqslant 1\,, \tag{A266}$$

then f has a fixed point in S_m.

Proof. We use Banach fixed point theorem: Lemma A10 says that f maps from the compact set S_m into the same compact set S_m. Lemma A11 says that f is a contraction mapping in S_m. □

Contraction mapping with a fixed point. We assume that the first l patterns are much more probable (and similar to one another) than the other patterns. Therefore, we define:

$$M := \max_i \|\boldsymbol{x}_i\| , \tag{A267}$$

$$\gamma = \sum_{i=l+1}^{N} p_i \leqslant \epsilon , \tag{A268}$$

$$1-\gamma = \sum_{i=1}^{l} p_i \geq 1 - \epsilon , \tag{A269}$$

$$\tilde{p}_i := \frac{p_i}{1-\gamma} \leqslant p_i/(1-\epsilon) , \tag{A270}$$

$$\sum_{i=1}^{l} \tilde{p}_i = 1 , \tag{A271}$$

$$\boldsymbol{m}_{\boldsymbol{x}} = \frac{1}{l} \sum_{i=1}^{l} \boldsymbol{x}_i , \tag{A272}$$

$$m_{\max} = \max_{1 \leqslant i \leqslant l} \|\boldsymbol{x}_i - \boldsymbol{m}_{\boldsymbol{x}}\| . \tag{A273}$$

M is an upper bound on the Euclidean norm of the patterns, which are vectors. ϵ is an upper bound on the probability γ of not choosing one of the first l patterns, while $1-\epsilon$ is a lower bound the probability $(1-\gamma)$ of choosing one of the first l patterns. $\boldsymbol{m}_{\boldsymbol{x}}$ is the arithmetic mean (the center) of the first l patterns. $m_{\max}$ is the maximal distance of the patterns to the center $\boldsymbol{m}_{\boldsymbol{x}}$. $\tilde{\boldsymbol{p}}$ is the probability $\boldsymbol{p}$ normalized for the first l patterns.

The variance of the first l patterns is

$$\begin{aligned} \mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}] &= \sum_{i=1}^{l} \tilde{p}_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T - \left(\sum_{i=1}^{l} \tilde{p}_i \, \boldsymbol{x}_i \right) \left(\sum_{i=1}^{l} \tilde{p}_i \, \boldsymbol{x}_i \right)^T \\ &= \sum_{i=1}^{l} \tilde{p}_i \left(\boldsymbol{x}_i - \sum_{i=1}^{l} \tilde{p}_i \boldsymbol{x}_i \right) \left(\boldsymbol{x}_i - \sum_{i=1}^{l} \tilde{p}_i \boldsymbol{x}_i \right)^T . \end{aligned} \tag{A274}$$

We have shown that a fixed point exists. We want to know how fast the iteration converges to the fixed point. Let $\boldsymbol{m}_{\boldsymbol{x}}^*$ be the fixed point of the iteration f in the sphere S_m. Using the mean value theorem Lemma A32, we have with $\mathrm{J}^m = \int_0^1 \mathrm{J}(\lambda \boldsymbol{\xi} + (1-\lambda)\boldsymbol{m}_{\boldsymbol{x}}^*) \, \mathrm{d}\lambda$:

$$\|f(\boldsymbol{\xi}) - \boldsymbol{m}_{\boldsymbol{x}}^*\| = \|f(\boldsymbol{\xi}) - f(\boldsymbol{m}_{\boldsymbol{x}}^*)\| \leqslant \|\mathrm{J}^m\|_2 \, \|\boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}}^*\| \tag{A275}$$

According to Lemma A8 the following bounds on the norm $\|\mathrm{J}\|_2$ of the Jacobian of the fixed point iteration hold. The γ-bound for $\|\mathrm{J}\|_2$ is

$$\|\mathrm{J}\|_2 \leqslant \beta \left((1-\gamma) \, m_{\max}^2 + \gamma \, 2 \, (2 - \gamma) \, M^2 \right) , \tag{A276}$$

while the ϵ-bound for $\|\mathrm{J}\|_2$ is:

$$\|\mathrm{J}\|_2 \leqslant \beta \left(m_{\max}^2 + \epsilon \, 2 \, (2 - \epsilon) \, M^2 \right) . \tag{A277}$$

From the last condition we require for a contraction mapping:

$$\beta \, m_{\max}^2 < 1 . \tag{A278}$$

We want to see how large ϵ is. The separation of center $\boldsymbol{m}_{\boldsymbol{x}}$ from data $\boldsymbol{X} = (\boldsymbol{x}_{l+1}, \ldots, \boldsymbol{x}_N)$ is

$$\Delta_m \ = \ \min_{j,l<j} \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} \ - \ \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \right) \ = \ \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} \ - \ \max_{j,l<j} \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \ . \tag{A279}$$

We need the separation $\tilde{\Delta}_m$ of $\tilde{\boldsymbol{x}} = \lambda \boldsymbol{\xi} + (1-\lambda) \boldsymbol{m}_{\boldsymbol{x}}^*$ from the data.

$$\tilde{\Delta}_m \ = \ \min_{j,l<j} \left(\tilde{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} \ - \ \tilde{\boldsymbol{x}}^T \boldsymbol{x}_j \right) \ . \tag{A280}$$

We compute a lower bound on $\tilde{\Delta}_m$. Using the Cauchy-Schwarz inequality, we obtain for $1 \leqslant j \leqslant N$:

$$\left| \tilde{\boldsymbol{x}}^T \boldsymbol{x}_j \ - \ \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \right| \leqslant \ \| \tilde{\boldsymbol{x}} \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \ \| \boldsymbol{x}_j \| \ \leqslant \ \| \tilde{\boldsymbol{x}} \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \ M \ . \tag{A281}$$

We have the lower bound

$$\begin{aligned} \tilde{\Delta}_m \ &\geq \ \min_{j,l<j} \left(\left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} \ - \ \| \tilde{\boldsymbol{x}} \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \ M \right) \ - \ \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \ + \ \| \tilde{\boldsymbol{x}} \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \ M \right) \right) \\ &= \ - \ 2 \ \| \tilde{\boldsymbol{x}} \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \ M \ + \ \min_{j,l<j} \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} \ - \ \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \right) \ = \ \Delta_m \ - \ 2 \ \| \tilde{\boldsymbol{x}} \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \ M \ . \end{aligned} \tag{A282}$$

Since

$$\begin{aligned} \| \tilde{\boldsymbol{x}} \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \ &= \ \| \lambda \boldsymbol{\xi} + (1-\lambda) \boldsymbol{m}_{\boldsymbol{x}}^* \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \\ &\leqslant \ \lambda \ \| \boldsymbol{\xi} \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \ + \ (1-\lambda) \ \| \boldsymbol{m}_{\boldsymbol{x}}^* \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \\ &\leqslant \ \max\{ \| \boldsymbol{\xi} \ - \ \boldsymbol{m}_{\boldsymbol{x}} \|, \| \boldsymbol{m}_{\boldsymbol{x}}^* \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \} \ , \end{aligned} \tag{A283}$$

we have

$$\tilde{\Delta}_m \ \geq \ \Delta_m \ - \ 2 \ \max\{ \| \boldsymbol{\xi} \ - \ \boldsymbol{m}_{\boldsymbol{x}} \|, \| \boldsymbol{m}_{\boldsymbol{x}}^* \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \} \ M \ . \tag{A284}$$

$$\epsilon \ = \ (N-l) \exp(- \beta \ (\Delta_m \ - \ 2 \ \max\{ \| \boldsymbol{\xi} \ - \ \boldsymbol{m}_{\boldsymbol{x}} \|, \| \boldsymbol{m}_{\boldsymbol{x}}^* \ - \ \boldsymbol{m}_{\boldsymbol{x}} \| \} \ M)) \ . \tag{A285}$$

A2.5 Properties of Fixed Points Near Stored Pattern

In Subsection A2.4.3 many stable states that are fixed points near the stored patterns are considered. We now consider this case. In the fist subsection we investigate the storage capacity if all patterns are sufficiently separated so that metastable states do not appear. In the next subsection we look into the convergence speed and error when retrieving the stored patterns. For metastable states we can do the same analyses if each metastable state is treated as one state like one pattern.

We see a trade-off that is known from classical Hopfield networks and for modern Hopfield networks. Small separation Δ_i of the pattern $\boldsymbol{x}_i$ from the other patterns gives high storage capacity. However the convergence speed is lower and the retrieval error higher. In contrast, large separation Δ_i of the pattern $\boldsymbol{x}_i$ from the other pattern gives exponentially fast convergence (one update is sufficient) and exponentially low retrieval error.

A2.5.1 Exponentially Many Patterns can be Stored

From Subsection A2.4.3 need some definitions. We assume to have N patterns, the separation of pattern $\boldsymbol{x}_i$ from the other patterns $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, \ldots, \boldsymbol{x}_N\}$ is Δ_i, defined as

$$\Delta_i \ = \ \min_{j,j \neq i} \left(\boldsymbol{x}_i^T \boldsymbol{x}_i \ - \ \boldsymbol{x}_i^T \boldsymbol{x}_j \right) \ = \ \boldsymbol{x}_i^T \boldsymbol{x}_i \ - \ \max_{j,j \neq i} \boldsymbol{x}_i^T \boldsymbol{x}_j \ . \tag{A286}$$

The pattern is separated from the other data if $0 < \Delta_i$. The separation Δ_i can also be expressed as

$$\begin{aligned} \Delta_i \ &= \ \min_{j,j \neq i} \frac{1}{2} \ \left(\| \boldsymbol{x}_i \|^2 \ - \ \| \boldsymbol{x}_j \|^2 \ + \ \| \boldsymbol{x}_i \ - \ \boldsymbol{x}_j \|^2 \right) \\ &= \ \frac{1}{2} \| \boldsymbol{x}_i \|^2 \ - \ \frac{1}{2} \ \max_{j,j \neq i} \left(\| \boldsymbol{x}_j \|^2 \ - \ \| \boldsymbol{x}_i \ - \ \boldsymbol{x}_j \|^2 \right) \ . \end{aligned} \tag{A287}$$

For $\| \boldsymbol{x}_i \| = \| \boldsymbol{x}_j \|$ we have $\Delta_i = 1/2 \min_{j,j \neq i} \| \boldsymbol{x}_i \ - \ \boldsymbol{x}_j \|^2$. The sphere S_i with center $\boldsymbol{x}_i$ is defined as

$$\mathrm{S}_i \ = \ \left\{ \boldsymbol{\xi} \mid \| \boldsymbol{\xi} \ - \ \boldsymbol{x}_i \| \ \leqslant \ \frac{1}{\beta \ N \ M} \right\} \ . \tag{A288}$$

The maximal length of a pattern is $M = \max_i \| \boldsymbol{x}_i \|$.

We next define what we mean with storing and retrieving a pattern.

Definition 5 (Pattern Stored and Retrieved)**.** *We assume that around every pattern $\boldsymbol{x}_i$ a sphere S_i is given. We say $\boldsymbol{x}_i$* is stored *if there is a single fixed point $\boldsymbol{x}_i^* \in \mathrm{S}_i$ to which all points $\boldsymbol{\xi} \in \mathrm{S}_i$ converge, and $\mathrm{S}_i \cap \mathrm{S}_j = \emptyset$ for $i \neq j$. We say $\boldsymbol{x}_i$* is retrieved *if iteration (update rule) Eq. (A81) converged to the single fixed point $\boldsymbol{x}_i^* \in \mathrm{S}_i$. The retrieval error is $\| \boldsymbol{x}_i - \boldsymbol{x}_i^* \|$.*

For a query $\boldsymbol{\xi} \in \mathrm{S}_i$ to converge to a fixed point $\boldsymbol{x}_i^* \in \mathrm{S}_i$ we required for the application of Banach fixed point theorem and for ensuring a contraction mapping the following inequality:

$$\Delta_i \ \geq \ \frac{2}{\beta \, N} \ + \ \frac{1}{\beta} \, \ln \left(2 \, (N-1) \, N \, \beta \, M^2 \right) \ . \tag{A289}$$

This is the assumption in Lemma A7 to ensure a fixed point in sphere S_i. Since replacing $(N-1)N$ by N^2 gives

$$\frac{2}{\beta \, N} \ + \ \frac{1}{\beta} \, \ln \left(2 \, N^2 \, \beta \, M^2 \right) \ > \ \frac{2}{\beta \, N} \ + \ \frac{1}{\beta} \, \ln \left(2 \, (N-1) \, N \, \beta \, M^2 \right) \ , \tag{A290}$$

the inequality follows from following master inequality

$$\Delta_i \ \geq \ \frac{2}{\beta \, N} \ + \ \frac{1}{\beta} \, \ln \left(2 \, N^2 \, \beta \, M^2 \right) \ , \tag{A291}$$

If we assume that $\mathrm{S}_i \cap \mathrm{S}_j \neq \emptyset$ with $i \neq j$, then the triangle inequality with a point from the intersection gives

$$\| \boldsymbol{x}_i \ - \ \boldsymbol{x}_j \| \ \leqslant \ \frac{2}{\beta \, N \, M} \ . \tag{A292}$$

Therefore, we have using the Cauchy-Schwarz inequality:

$$\Delta_i \ \leqslant \ \boldsymbol{x}_i^T \left(\boldsymbol{x}_i \ - \ \boldsymbol{x}_j \right) \ \leqslant \ \| \boldsymbol{x}_i \| \, \| \boldsymbol{x}_i \ - \ \boldsymbol{x}_j \| \ \leqslant M \, \frac{2}{\beta \, N \, M} \ = \ \frac{2}{\beta \, N} \ . \tag{A293}$$

The last inequality is a contraction to Eq. (A291) if we assume that

$$1 \ < \ 2 \, (N-1) \, N \, \beta \, M^2 \ . \tag{A294}$$

With this assumption, the spheres S_i and S_j do not intersect. Therefore, each $\boldsymbol{x}_i$ has its separate fixed point in S_i. We define

$$\Delta_{\min} \ = \ \min_{1 \leqslant i \leqslant N} \Delta_i \tag{A295}$$

to obtain the master inequality

$$\Delta_{\min} \ \geq \ \frac{2}{\beta \, N} \ + \ \frac{1}{\beta} \, \ln \left(2 \, N^2 \, \beta \, M^2 \right) \ . \tag{A296}$$

Patterns on a sphere. For simplicity and in accordance with the results of the classical Hopfield network, we assume all *patterns being on a sphere* with radius M:

$$\forall_i : \ \| \boldsymbol{x}_i \| \ = \ M \ . \tag{A297}$$

Under assumption Eq. (A294) we have only to show that the master inequality Eq. (A296) is fulfilled for each $\boldsymbol{x}_i$ to have a separate fixed point near each $\boldsymbol{x}_i$.

We defined α_{ij} as the angle between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. The minimal angle $\alpha_{\min}$ between two data points is

$$\alpha_{\min} \ = \ \min_{1 \leqslant i < j \leqslant N} \alpha_{ij} \ . \tag{A298}$$

On the sphere with radius M we have

$$\Delta_{\min} \ = \ \min_{1 \leqslant i < j \leqslant N} M^2 (1 \ - \ \cos(\alpha_{ij})) \ = \ M^2 (1 \ - \ \cos(\alpha_{\min})) \ , \tag{A299}$$

therefore it is sufficient to show the master inequality on the sphere:

$$M^2 (1 \ - \ \cos(\alpha_{\min})) \ \geq \ \frac{2}{\beta \, N} \ + \ \frac{1}{\beta} \, \ln \left(2 \, N^2 \, \beta \, M^2 \right) \ . \tag{A300}$$

Under assumption Eq. (A294) we have only to show that the master inequality Eq. (A296) is fulfilled for $\Delta_{\min}$. We consider patterns on the sphere, therefore the master inequality Eq. (A296) becomes Eq. (A300). First we show results when pattern positions on the sphere are constructed and $\Delta_{\min}$ is ensured. Then we move on to random patterns on a sphere, where $\Delta_{\min}$ becomes a random variable.

Storage capacity for patterns placed on the sphere. Next theorem says how many patterns we can stored (fixed point with attraction basin near pattern) if we are allowed to place them on the sphere.

Theorem A3 (Storage Capacity (M=2): Placed Patterns)**.** *We assume $\beta = 1$ and patterns on the sphere with radius M. If $M = 2\sqrt{d-1}$ and the dimension d of the space is $d \geq 4$ or if $M = 1.7\sqrt{d-1}$ and the dimension d of the space is $d \geq 50$, then the number of patterns N that can be stored (fixed point with attraction basin near pattern) is at least*

$$N \ = \ 2^{2(d-1)} \ . \tag{A301}$$

Proof. For random patterns on the sphere, we have to show that the master inequality Eq. (A300) holds:

$$M^2(1 \ - \ \cos(\alpha_{\min})) \ \geq \ \frac{2}{\beta \, N} \ + \ \frac{1}{\beta} \, \ln\left(2 \, N^2 \, \beta \, M^2\right) \ . \tag{A302}$$

We now place the patterns equidistant on the sphere where the pattern are separated by an angle $\alpha_{\min}$:

$$\forall_i : \ \min_{j, j \neq i} \alpha_{ij} \ = \ \alpha_{\min} \ , \tag{A303}$$

In a d-dimensional space we can place

$$N \ = \ \left(\frac{2\pi}{\alpha_{\min}}\right)^{d-1} \tag{A304}$$

points on the sphere. In a spherical coordinate system a pattern differs from its most closest patterns by an angle $\alpha_{\min}$ and there are $d-1$ angles. Solving for $\alpha_{\min}$ gives

$$\alpha_{\min} \ = \ \frac{2\pi}{N^{1/(d-1)}} \ . \tag{A305}$$

The number of patterns that can be stored is determined by the largest N that fulfils

$$M^2 \left(1 \ - \ \cos\left(\frac{2\pi}{N^{1/(d-1)}}\right)\right) \ \geq \ \frac{2}{\beta \, N} \ + \ \frac{1}{\beta} \, \ln\left(2 \, N^2 \, \beta \, M^2\right) \ . \tag{A306}$$

We set $N = 2^{2(d-1)}$ and obtain for Eq. (A306):

$$M^2 \left(1 \ - \ \cos\left(\frac{\pi}{2}\right)\right) \ \geq \ \frac{2}{\beta \, 2^{3(d-1)}} \ + \ \frac{1}{\beta} \, \ln\left(2 \, \beta \, M^2\right) \ + \ \frac{1}{\beta} \, 4 \, (d-1) \ln 2 \, . \tag{A307}$$

This inequality is equivalent to

$$\beta \, M^2 \ \geq \ \frac{1}{2^{2(d-1)-1}} \ + \ \ln\left(2 \, \beta \, M^2\right) \ + \ 4 \, (d-1) \ln 2 \, . \tag{A308}$$

The last inequality can be fulfilled with $M = K\sqrt{d-1}$ and proper K. For $\beta = 1$, $d = 4$ and $K = 2$ the inequality is fulfilled. The left hand side minus the right hand side is $4(d-1) - 1/2^{2(d-1)-1} - \ln(8(d-1)) - 4(d-1)\ln 2$. Its derivative with respect to d is strict positive. Therefore, the inequality holds for $d \geq 4$.

For $\beta = 1$, $d = 50$ and $K = 1.7$ the inequality is fulfilled. The left hand side minus the right hand side is $2.89(d-1) - 1/2^{2(d-1)-1} - \ln(5.78(d-1)) - 4(d-1)\ln 2$. Its derivative with respect to d is strict positive. Therefore, the inequality holds for $d \geq 50$.

□

If we want to store considerably more patterns, then we have to increase the length of the vectors or the dimension of the space where the vectors live. The next theorem shows results for the number of patterns N with $N = 2^{3(d-1)}$.

Theorem A4 (Storage Capacity (M=5): Placed Patterns)**.** *We assume $\beta = 1$ and patterns on the sphere with radius M. If $M = 5\sqrt{d-1}$ and the dimension d of the space is $d \geq 3$ or if $M = 4\sqrt{d-1}$ and the dimension d of the space is $d \geq 13$, then the number of patterns N that can be stored (fixed point with attraction basin near pattern) is at least*

$$N \ = \ 2^{3(d-1)} \ . \tag{A309}$$

Proof. We set $N = 2^{3(d-1)}$ and obtain for Eq. (A306):

$$M^2 \left(1 - \cos\left(\frac{\pi}{4}\right)\right) \geq \frac{2}{\beta\, 2^{3(d-1)}} + \frac{1}{\beta} \ln\left(2\,\beta\, M^2\right) + \frac{1}{\beta}\, 6\,(d-1) \ln 2\,. \tag{A310}$$

This inequality is equivalent to

$$\beta\, M^2 \left(1 - \frac{\sqrt{2}}{2}\right) \geq \frac{1}{2^{3(d-1)-1}} + \ln\left(2\,\beta\, M^2\right) + 6\,(d-1) \ln 2\,. \tag{A311}$$

The last inequality can be fulfilled with $M = K\sqrt{d-1}$ and proper K. For $\beta = 1$, $d = 13$ and $K = 4$ the inequality is fulfilled. The left hand side minus the right hand side is $4.686292(d-1) - 1/2^{3(d-1)-1} - \ln(32(d-1)) - 6(d-1)\ln 2$. Its derivative with respect to d is strict positive. Therefore, the inequality holds for $d \geq 13$.

For $\beta = 1$, $d = 3$ and $K = 5$ the inequality is fulfilled. The left hand side minus the right hand side is $7.32233(d-1) - 1/2^{3(d-1)-1} - \ln(50(d-1)) - 6(d-1)\ln 2$. Its derivative with respect to d is strict positive. Therefore, the inequality holds for $d \geq 3$.

□

Storage capacity for random patterns on the sphere. Next we investigate random points on the sphere. Under assumption Eq. (A294) we have to show that the master inequality Eq. (A300) is fulfilled for $\alpha_{\min}$, where now $\alpha_{\min}$ is now a random variable. We use results on the distribution of the minimal angles between random patterns on a sphere according to [14] and [12]. Theorem 2 in [14] gives the distribution of the minimal angle for random patterns on the unit sphere. Proposition 3.5 in [12] gives a lower bound on the probability of the minimal angle being larger than a given constant. We require this proposition to derive the probability of pattern having a minimal angle $\alpha_{\min}$. Proposition 3.6 in[12] gives the expectation of the minimal angle.

We will prove high probability bounds for the expected storage capacity. We need the following tail-bound on $\alpha_{\min}$ (the minimal angle of random patterns on a sphere):

Lemma A13 ([12])**.** *Let d be the dimension of the pattern space,*

$$\kappa_d := \frac{1}{d\,\sqrt{\pi}} \frac{\Gamma((d+1)/2)}{\Gamma(d/2)}\,. \tag{A312}$$

and $\delta > 0$ such that $\frac{\kappa_{d-1}}{2}\delta^{(d-1)} \leqslant 1$. Then

$$\Pr(N^{\frac{2}{d-1}} \alpha_{\min} \geq \delta) \geq 1 - \frac{\kappa_{d-1}}{2}\, \delta^{d-1}\,. \tag{A313}$$

Proof. The statement of the lemma is Eq. (3-6) from Proposition 3.5 in [12]. □

Next we derive upper and lower bounds on the constant κ_d since we require them later for proving storage capacity bounds.

Lemma A14. *For κ_d defined in Eq. (A312) we have the following bounds for every $d \geq 1$:*

$$\frac{1}{\exp(1/6)\,\sqrt{e\,\pi\,d}} \leqslant \kappa_d \leqslant \frac{\exp(1/12)}{\sqrt{2\,\pi\,d}} < 1\,. \tag{A314}$$

Proof. We use for $x > 0$ the following bound related to Stirling's approximation formula for the gamma function, c.f. [48, (5.6.1)]:

$$1 < \Gamma(x)\,(2\,\pi)^{-\frac{1}{2}} x^{\frac{1}{2}-x} \exp(x) < \exp\left(\frac{1}{12\,x}\right)\,. \tag{A315}$$

Using Stirling's formula Eq. (A315), we upper bound κ_d:

$$\kappa_d = \frac{1}{d\sqrt{\pi}} \frac{\Gamma((d+1)/2)}{\Gamma(d/2)} < \frac{1}{d\sqrt{\pi}} \frac{\exp\left(\frac{1}{6(d+1)}\right) \exp\left(-\frac{d+1}{2}\right) \left(\frac{d+1}{2}\right)^{\frac{d}{2}}}{\exp\left(-\frac{d}{2}\right) \left(\frac{d}{2}\right)^{\frac{d}{2}-\frac{1}{2}}} \tag{A316}$$

$$= \frac{1}{d\sqrt{\pi\, e}} \exp\left(\frac{1}{6(d+1)}\right) \left(1 + \frac{1}{d}\right)^{\frac{d}{2}} \sqrt{\frac{d}{2}} \leqslant \frac{\exp\left(\frac{1}{12}\right)}{\sqrt{2\,\pi}\,\sqrt{d}} .$$

For the first inequality, we applied Eq. (A315), while for the second we used $(1+\frac{1}{d})^d < e$ for $d \geq 1$.

Next, we lower bound κ_d by again applying Stirling's formula Eq. (A315):

$$\kappa_d = \frac{1}{d\sqrt{\pi}} \frac{\Gamma((d+1)/2)}{\Gamma(d/2)} > \frac{1}{d\sqrt{\pi}} \frac{\exp\left(-\frac{d+1}{2}\right) \left(\frac{d+1}{2}\right)^{\frac{d}{2}}}{\exp\left(\frac{1}{6\,d}\right) \exp\left(-\frac{d}{2}\right) \left(\frac{d}{2}\right)^{\frac{d}{2}-\frac{1}{2}}} \tag{A317}$$

$$= \frac{1}{d\sqrt{\pi\, e}\, \exp\left(\frac{1}{6\,d}\right)} \left(1+\frac{1}{d}\right)^{\frac{d}{2}} \sqrt{\frac{d}{2}} \geq \frac{1}{\exp\left(\frac{1}{6}\right)\, \sqrt{e\,\pi\, d}} ,$$

where the last inequality holds because of monotonicity of $(1+\frac{1}{d})^d$ and using the fact that for $d=1$ it takes on the value 2. □

We require a bound on cos to bound the master inequality Eq. (A300).

Lemma A15. *For $0 \leqslant x \leqslant \pi$ the function* cos *can be upper bounded by:*

$$\cos(x) = 1 - \frac{x^2}{5} . \tag{A318}$$

Proof. We use the infinite product representation of cos from [48, (4.22.2)]:

$$\cos(x) = \prod_{n=1}^{\infty} \left(1 - \frac{4\,x^2}{(2n-1)^2\,\pi^2}\right) . \tag{A319}$$

It holds

$$1 - \frac{4\,x^2}{(2n-1)^2\,\pi^2} \leqslant 1 \tag{A320}$$

for $|x| \leqslant \pi$ and $n \geq 2$, we can get the following upper bound on Eq. (A319):

$$\cos(x) \leqslant \prod_{n=1}^{2} \left(1 - \frac{4\,x^2}{(2n-1)^2\pi^2}\right) = \left(1 - \frac{4\,x^2}{\pi^2}\right) \left(1 - \frac{4\,x^2}{9\,\pi^2}\right) \tag{A321}$$

$$= 1 - \frac{40\,x^2}{9\,\pi^2} + \frac{16\,x^4}{9\,\pi^4} \leqslant 1 - \frac{40\,x^2}{9\,\pi^2} + \frac{16\,x^2}{9\,\pi^2}$$

$$= 1 - \frac{24\,x^2}{9\,\pi^2} \leqslant 1 - \frac{x^2}{5} .$$

The last but one inequality uses $x \leqslant \pi$, which implies $x/\pi \leqslant 1$. Thus Eq. (A318) is proven.

□

Exponential storage capacity: the base c as a function of the parameter β, the radius of the sphere M, the probability p, and the dimension d of the space. We express the number N of stored patterns by an exponential function with base $c > 1$ and an exponent linear in d. We derive constraints on he base c as a function of β, the radius of the sphere M, the probability p that all patterns can be stored, and the dimension d of the space. With $\beta > 0$, $K > 0$, and $d \geq 2$ (to ensure a sphere), the following theorem gives our main result.

Theorem A5 (Storage Capacity (Main): Random Patterns)**.** *We assume a failure probability* $0 < p \leqslant 1$ *and randomly chosen patterns on the sphere with radius* $M = K\sqrt{d-1}$*. We define*

$$a := \frac{2}{d-1}\left(1 + \ln(2\,\beta\,K^2\,p\,(d-1))\right), \quad b := \frac{2\,K^2\,\beta}{5},$$
$$c = \frac{b}{W_0(\exp(a + \ln(b))}, \tag{A322}$$

where W_0 *is the upper branch of the Lambert* W *function and ensure*

$$c \geq \left(\frac{2}{\sqrt{p}}\right)^{\frac{4}{d-1}}. \tag{A323}$$

Then with probability $1 - p$*, the number of random patterns that can be stored is*

$$N \geq \sqrt{p}\,c^{\frac{d-1}{4}}. \tag{A324}$$

Examples are $c \geq 3.1546$ *for* $\beta = 1$*,* $K = 3$*,* $d = 20$ *and* $p = 0.001$ *(*$a + \ln(b) > 1.27$*) and* $c \geq 1.3718$ *for* $\beta = 1$*,* $K = 1$*,* $d = 75$*, and* $p = 0.001$ *(*$a + \ln(b) < -0.94$*).*

Proof. We consider the probability that the master inequality Eq. (A300) is fulfilled:

$$\Pr\left(M^2(1 - \cos(\alpha_{\min})) \geq \frac{2}{\beta\,N} + \frac{1}{\beta}\ln\left(2\,N^2\,\beta\,M^2\right)\right) \geq 1 - p\,. \tag{A325}$$

Using Eq. (A318), we have:

$$1 - \cos(\alpha_{\min}) \geq \frac{1}{5}\,\alpha_{\min}^2\,. \tag{A326}$$

Therefore, with probability $1 - p$ the storage capacity is largest N that fulfills

$$\Pr\left(M^2\frac{\alpha_{min}^2}{5} \geq \frac{2}{\beta\,N} + \frac{1}{\beta}\ln\left(2\,N^2\,\beta\,M^2\right)\right) \geq 1 - p\,. \tag{A327}$$

This inequality is equivalent to

$$\Pr\left(N^{\frac{2}{d-1}}\,\alpha_{min} \geq \frac{\sqrt{5}\,N^{\frac{2}{d-1}}}{M}\left(\frac{2}{\beta\,N} + \frac{1}{\beta}\ln\left(2\,N^2\,\beta\,M^2\right)\right)^{\frac{1}{2}}\right) \geq 1 - p\,. \tag{A328}$$

We use Eq. (A313) to obtain:

$$\Pr\left(N^{\frac{2}{d-1}}\,\alpha_{min} \geq \frac{\sqrt{5}\,N^{\frac{2}{d-1}}}{M}\left(\frac{2}{\beta\,N} + \frac{1}{\beta}\ln\left(2\,N^2\,\beta\,M^2\right)\right)^{\frac{1}{2}}\right) \tag{A329}$$
$$\geq 1 - \frac{\kappa_{d-1}}{2}\,5^{\frac{d-1}{2}}\,N^2\,M^{-(d-1)}\left(\frac{2}{\beta\,N} + \frac{1}{\beta}\ln\left(2\,N^2\,\beta\,M^2\right)\right)^{\frac{d-1}{2}}.$$

For Eq. (A328) to be fulfilled, it is sufficient that

$$\frac{\kappa_{d-1}}{2}\,5^{\frac{d-1}{2}}\,N^2\,M^{-(d-1)}\left(\frac{2}{\beta\,N} + \frac{1}{\beta}\ln\left(2\,N^2\,\beta M^2\right)\right)^{\frac{d-1}{2}} - p \leqslant 0\,. \tag{A330}$$

If we insert the assumption Eq. (A323) of the theorem into Eq. (A324), then we obtain $N \geq 2$. We now apply the upper bound $\kappa_{d-1}/2 < \kappa_{d-1} < 1$ from Eq. (A314) and the upper bound $\frac{2}{\beta N} \leqslant \frac{1}{\beta}$ from $N \geq 2$ to inequality Eq. (A330). In the resulting inequality we insert $N = \sqrt{p}c^{\frac{d-1}{4}}$ to check whether it is fulfilled with this special value of N and obtain:

$$5^{\frac{d-1}{2}}\,p\,c^{\frac{d-1}{2}}\,M^{-(d-1)}\left(\frac{1}{\beta} + \frac{1}{\beta}\ln\left(2\,p\,c^{\frac{d-1}{2}}\,\beta M^2\right)\right)^{\frac{d-1}{2}} \leqslant p\,. \tag{A331}$$

Dividing by p, inserting $M = K\sqrt{d-1}$, and exponentiation of the left and right side by $\frac{2}{d-1}$ gives:

$$\frac{5\,c}{K^2\,(d-1)}\left(\frac{1}{\beta}\;+\;\frac{1}{\beta}\,\ln\left(2\,\beta\,c^{\frac{d-1}{2}}\,p\,K^2\,(d-1)\right)\right)\;-\;1\;\leqslant\;0\,. \tag{A332}$$

After some algebraic manipulation, this inequality can be written as

$$a\,c\;+\;c\,\ln(c)\;-\;b\;\leqslant\;0\,, \tag{A333}$$

where we used

$$a\;:=\;\frac{2}{d-1}\,(1\;+\;\ln(2\,\beta\,K^2\,p\,(d-1)))\,,\quad b\;:=\;\frac{2\,K^2\,\beta}{5}\,.$$

We determine the value $\hat{c}$ of c which makes the inequality Eq. (A333) equal to zero. We solve

$$a\,\hat{c}\;+\;\hat{c}\,\ln(\hat{c})\;-\;b\;=\;0 \tag{A334}$$

for $\hat{c}$:

$$\begin{aligned}
&a\,\hat{c}\;+\;\hat{c}\,\ln(\hat{c})\;-\;b\;=\;0 \qquad\qquad (A335)\\
\Leftrightarrow\;&a\;+\;\ln(\hat{c})\;=\;b/\hat{c}\\
\Leftrightarrow\;&a\;+\;\ln(b)\;+\;\ln(\hat{c}/b)\;=\;b/\hat{c}\\
\Leftrightarrow\;&b/\hat{c}\;+\;\ln(b/\hat{c})\;=\;a\;+\;\ln(b)\\
\Leftrightarrow\;&b/\hat{c}\,\exp(b/\hat{c})\;=\;\exp(a\;+\;\ln(b))\\
\Leftrightarrow\;&b/\hat{c}\;=\;W_0(\exp(a\;+\;\ln(b)))\\
\Leftrightarrow\;&\hat{c}\;=\;\frac{b}{W_0(\exp(a\;+\;\ln(b))}\,,
\end{aligned}$$

where W_0 is the upper branch of the Lambert W function (see Def. A6). Hence, the solution is

$$\hat{c}\;=\;\frac{b}{W_0(\exp(a\;+\;\ln(b))}\,. \tag{A336}$$

The solution exist, since the Lambert function $W_0(x)$ is defined for $-1/e < x$ and we have $0 < \exp(a + \ln(b))$.

Since $\hat{c}$ fulfills inequality Eq. (A333) and therefore also Eq. (A331), we have a lower bound on the storage capacity N:

$$N\;\geq\;\sqrt{p}\,\hat{c}^{\,\frac{d-1}{4}}\,. \tag{A337}$$

□

Next we aim at a lower bound on c which does not use the Lambert W function. Therefore, we upper bound $W_0(\exp(a + \ln(b))$ to obtain a lower bound on c, therefore, also a lower bound on the storage capacity N. The lower bound is given in the next corollary.

Corollary A1. *We assume a failure probability $0 < p \leqslant 1$ and randomly chosen patterns on the sphere with radius $M = K\sqrt{d-1}$. We define*

$$a\;:=\;\frac{2}{d-1}\,(1\;+\;\ln(2\,\beta\,K^2\,p\,(d-1)))\,,\quad b\;:=\;\frac{2\,K^2\,\beta}{5}\,.$$

Using the omega constant $\Omega \approx 0.56714329$ we set

$$c\;=\;\begin{cases} b\,\ln\left(\frac{\Omega\,\exp(a\,+\,\ln(b))\,+\,1}{\Omega\,(1\,+\,\Omega)}\right)^{-1} & \text{for } a\;+\;\ln(b)\;\leqslant\;0\,,\\ b\,(a\;+\;\ln(b))^{-\frac{a\,+\,\ln(b)}{a\,+\,\ln(b)\,+\,1}} & \text{for } a\;+\;\ln(b)\;>\;0 \end{cases} \tag{A338}$$

and ensure

$$c\;\geq\;\left(\frac{2}{\sqrt{p}}\right)^{\frac{4}{d-1}}\,. \tag{A339}$$

Then with probability $1 - p$*, the number of random patterns that can be stored is*

$$N \geq \sqrt{p}\, c^{\frac{d-1}{4}} . \tag{A340}$$

Examples are $c \geq 3.1444$ *for* $\beta = 1$*,* $K = 3$*,* $d = 20$ *and* $p = 0.001$ *(*$a + \ln(b) > 1.27$*) and* $c \geq 1.2585$ *for* $\beta = 1$ $K = 1$*,* $d = 75$*, and* $p = 0.001$ *(*$a + \ln(b) < -0.94$*).*

Proof. We lower bound the c defined in Theorem A5. According to [36, Theorem 2.3] we have for any real u and $y > \frac{1}{e}$:

$$W_0(\exp(u)) \leqslant \ln\left(\frac{\exp(u) + y}{1 + \ln(y)}\right) . \tag{A341}$$

To upper bound $W_0(x)$ for $x \in [0, 1]$, we set

$$y = 1/W_0(1) = 1/\Omega = \exp\Omega = -1/\ln\Omega \approx 1.76322 , \tag{A342}$$

where the Omega constant Ω is

$$\Omega = \left(\int_{-\infty}^{\infty} \frac{\mathrm{d}t}{(e^t - t)^2 + \pi^2}\right)^{-1} - 1 \approx 0.56714329 . \tag{A343}$$

See for these equations the special values of the Lambert W function in Lemma A31. We have the upper bound on W_0:

$$W_0(\exp(u)) \leqslant \ln\left(\frac{\exp(u) + 1/\Omega}{1 + \ln(1/\Omega)}\right) = \ln\left(\frac{\Omega\, \exp(u) + 1}{\Omega(1 + \Omega)}\right) . \tag{A344}$$

At the right hand side of interval $[0, 1]$, we have $u = 0$ and $\exp(u) = 1$ and get:

$$\ln\left(\frac{\Omega\, 1 + 1}{\Omega(1 + \Omega)}\right) = \ln\left(\frac{1}{\Omega}\right) = -\ln(\Omega) = \Omega = W_0(1) . \tag{A345}$$

Therefore, the bound is tight at the right hand side of of interval $[0, 1]$, that is for $\exp(u) = 1$, i.e. $u = 0$. We have derived an bound for $W_0(\exp(u))$ with $\exp(u) \in [0, 1]$ or, equivalently, $u \in [-\infty, 0]$. We obtain from [36, Corollary 2.6] the following bound on $W_0(\exp(u))$ for $1 < \exp(u)$, or, equivalently $0 < u$:

$$W_0(\exp(u)) \leqslant u^{\frac{u}{1+u}} . \tag{A346}$$

A lower bound on $\hat{c}$ is obtained via the upper bounds Eq. (A346) and Eq. (A344) on W_0 as $W_0 > 0$. We set $u = a + \ln(b)$ and obtain

$$W_0(\exp(a + \ln(b))) \leqslant \begin{cases} \ln\left(\frac{\Omega\, \exp(a + \ln(b)) + 1}{\Omega\,(1 + \Omega)}\right)^{-1} & \text{for } a + \ln(b) \leqslant 0 , \\ (a + \ln(b))^{-\frac{a + \ln(b)}{a + \ln(b) + 1}} & \text{for } a + \ln(b) > 0 \end{cases} \tag{A347}$$

We insert this bound into Eq. (A336), the solution for $\hat{c}$, to obtain the statement of the theorem.

□

Exponential storage capacity: the dimension d of the space as a function of the parameter β, the radius of the sphere M, and the probability p. We express the number N of stored patterns by an exponential function with base $c > 1$ and an exponent linear in d. We derive constraints on the dimension d of the space as a function of β, the radius of the sphere M, the probability p that all patterns can be stored, and the base of the exponential storage capacity. The following theorem gives this result.

Theorem A6 (Storage Capacity (d computed): Random Patterns)**.** *We assume a failure probability* $0 < p \leqslant 1$ *and randomly chosen patterns on the sphere with radius* $M = K\sqrt{d-1}$*. We define*

$$\begin{aligned} a &:= \frac{\ln(c)}{2} - \frac{K^2\,\beta}{5\,c}\,, \quad b := 1 + \ln\left(2\,p\,\beta\,K^2\right)\,, \\ d &= \begin{cases} 1 + \frac{1}{a}\,W(a\,\exp(-b)) & \text{for } a \neq 0\,, \\ 1 + \exp(-b) & \text{for } a = 0\,, \end{cases} \end{aligned} \tag{A348}$$

where W *is the Lambert W function. For* $0 < a$ *the function* W *is the upper branch* W_0 *and for* $a < 0$ *we use the lower branch* W_{-1}*. If we ensure that*

$$c \geq \left(\frac{2}{\sqrt{p}}\right)^{\frac{4}{d-1}}\,, \quad -\frac{1}{e} \leqslant a\,\exp(-b)\,, \tag{A349}$$

then with probability $1 - p$*, the number of random patterns that can be stored is*

$$N \geq \sqrt{p}\,c^{\frac{d-1}{4}}\,. \tag{A350}$$

Proof. We consider the probability that the master inequality Eq. (A300) is fulfilled:

$$\Pr\left(M^2(1 - \cos(\alpha_{\min}))) \geq \frac{2}{\beta\,N} + \frac{1}{\beta}\,\ln\left(2\,N^2\,\beta\,M^2\right)\right) \geq 1 - p\,. \tag{A351}$$

Using Eq. (A318), we have:

$$1 - \cos(\alpha_{\min}) \geq \frac{1}{5}\,\alpha_{\min}^2\,. \tag{A352}$$

Therefore, with probability $1 - p$ the storage capacity is largest N that fulfills

$$\Pr\left(M^2\frac{\alpha_{min}^2}{5} \geq \frac{2}{\beta\,N} + \frac{1}{\beta}\,\ln\left(2\,N^2\,\beta\,M^2\right)\right) \geq 1 - p\,. \tag{A353}$$

This inequality is equivalent to

$$\Pr\left(N^{\frac{2}{d-1}}\,\alpha_{min} \geq \frac{\sqrt{5}\,N^{\frac{2}{d-1}}}{M}\,\left(\frac{2}{\beta\,N} + \frac{1}{\beta}\,\ln\left(2\,N^2\,\beta\,M^2\right)\right)^{\frac{1}{2}}\right) \geq 1 - p\,. \tag{A354}$$

We use Eq. (A313) to obtain:

$$\begin{aligned} &\Pr\left(N^{\frac{2}{d-1}}\,\alpha_{min} \geq \frac{\sqrt{5}\,N^{\frac{2}{d-1}}}{M}\,\left(\frac{2}{\beta\,N} + \frac{1}{\beta}\,\ln\left(2\,N^2\,\beta\,M^2\right)\right)^{\frac{1}{2}}\right) \\ &\geq 1 - \frac{\kappa_{d-1}}{2}\,5^{\frac{d-1}{2}}\,N^2\,M^{-(d-1)}\left(\frac{2}{\beta\,N} + \frac{1}{\beta}\,\ln\left(2\,N^2\,\beta\,M^2\right)\right)^{\frac{d-1}{2}}\,. \end{aligned} \tag{A355}$$

For Eq. (A354) to be fulfilled, it is sufficient that

$$\frac{\kappa_{d-1}}{2}\,5^{\frac{d-1}{2}}\,N^2\,M^{-(d-1)}\left(\frac{2}{\beta\,N} + \frac{1}{\beta}\,\ln\left(2\,N^2\,\beta M^2\right)\right)^{\frac{d-1}{2}} - p \leqslant 0\,. \tag{A356}$$

If we insert the assumption Eq. (A349) of the theorem into Eq. (A350), then we obtain $N \geq 2$. We now apply the upper bound $\kappa_{d-1}/2 < \kappa_{d-1} < 1$ from Eq. (A314) and the upper bound $\frac{2}{\beta N} \leqslant \frac{1}{\beta}$ from $N \geq 2$ to inequality Eq. (A356). In the resulting inequality we insert $N = \sqrt{p}c^{\frac{d-1}{4}}$ to check whether it is fulfilled with this special value of N and obtain:

$$5^{\frac{d-1}{2}}\,p\,c^{\frac{d-1}{2}}\,M^{-(d-1)}\left(\frac{1}{\beta} + \frac{1}{\beta}\,\ln\left(2\,p\,c^{\frac{d-1}{2}}\,\beta M^2\right)\right)^{\frac{d-1}{2}} \leqslant p\,. \tag{A357}$$

Dividing by p, inserting $M = K\sqrt{d-1}$, and exponentiation of the left and right side by $\frac{2}{d-1}$ gives:

$$\frac{5\,c}{K^2\,(d-1)}\left(\frac{1}{\beta} + \frac{1}{\beta}\,\ln\left(2\,\beta\,c^{\frac{d-1}{2}}\,p\,K^2\,(d-1)\right)\right) - 1 \leqslant 0\,. \tag{A358}$$

This inequality Eq. (A358) can be reformulated as:

$$1 \, + \, \ln\left(2 \, p \, \beta \, c^{\frac{d-1}{2}} \, K^2 \, (d-1)\right) \, - \, \frac{(d-1) \, K^2 \, \beta}{5 \, c} \, \leqslant \, 0 \, . \tag{A359}$$

Using

$$a \, := \, \frac{\ln(c)}{2} \, - \, \frac{K^2 \, \beta}{5 \, c} \, , \quad b \, := \, 1 \, + \, \ln\left(2 \, p \, \beta \, K^2\right) \, , \tag{A360}$$

we write inequality Eq. (A359) as

$$\ln(d-1) \, + \, a \, (d-1) \, + \, b \, \leqslant \, 0 \, . \tag{A361}$$

We determine the value $\hat{d}$ of d which makes the inequality Eq. (A361) equal to zero. We solve

$$\ln(\hat{d}-1) \, + \, a \, (\hat{d}-1) \, + \, b \, = \, 0 \, . \tag{A362}$$

for $\hat{d}$

For $a \neq 0$ we have

$$\begin{aligned}
& \ln(\hat{d}-1) \, + \, a \, (\hat{d}-1) \, + \, b \, = \, 0 \qquad \text{(A363)} \\
\Leftrightarrow \; & a \, (\hat{d}-1) \, + \, \ln(\hat{d}-1) \, = \, -\, b \\
\Leftrightarrow \; & (\hat{d}-1) \exp(a \, (\hat{d}-1)) \, = \, \exp(-b) \\
\Leftrightarrow \; & a \, (\hat{d}-1) \exp(a \, (\hat{d}-1)) \, = \, a \, \exp(-b) \\
\Leftrightarrow \; & a \, (\hat{d}-1) \, = \, W(a \, \exp(-b)) \\
\Leftrightarrow \; & \hat{d} \, - \, 1 \, = \, \frac{1}{a} \, W(a \, \exp(-b)) \\
\Leftrightarrow \; & \hat{d} \, = \, 1 \, + \, \frac{1}{a} \, W(a \, \exp(-b)) \, ,
\end{aligned}$$

where W is the Lambert W function (see Def. A6). For $a > 0$ we have to use the upper branch W_0 of the Lambert W function and for $a < 0$ we use the lower branch W_{-1} of the Lambert W function. We have to ensure that $-1/e \leqslant a \exp(-b)$ for a solution to exist. For $a = 0$ we have $\hat{d} = 1 + \exp(-b)$.

Hence, the solution is

$$\hat{d} \, = \, 1 \, + \, \frac{1}{a} \, W(a \exp(-b)) \, . \tag{A364}$$

Since $\hat{d}$ fulfills inequality Eq. (A358) and therefore also Eq. (A357), we have a lower bound on the storage capacity N:

$$N \, \geq \, \sqrt{p} \, \hat{c}^{\, \frac{d-1}{4}} \, . \tag{A365}$$

□

Corollary A2. *We assume a failure probability $0 < p \leqslant 1$ and randomly chosen patterns on the sphere with radius $M = K\sqrt{d-1}$. We define*

$$\begin{aligned}
a \, &:= \, \frac{\ln(c)}{2} \, - \, \frac{K^2 \, \beta}{5 \, c} \, , \quad b \, := \, 1 \, + \, \ln\left(2 \, p \, \beta \, K^2\right) \, , \\
d \, &= \, 1 \, + \, \frac{1}{a} \, \left(-\, \ln(-a) \, + \, b\right) \, , \qquad \text{(A366)}
\end{aligned}$$

and ensure

$$c \, \geq \, \left(\frac{2}{\sqrt{p}}\right)^{\frac{4}{d-1}} \, , \quad -\, \frac{1}{e} \, \leqslant \, a \, \exp(-b) \, , \quad a \, < \, 0 \, , \tag{A367}$$

then with probability $1 - p$, the number of random patterns that can be stored is

$$N \, \geq \, \sqrt{p} \, c^{\, \frac{d-1}{4}} \, . \tag{A368}$$

Setting $\beta = 1$, $K = 3$, $c = 2$ and $p = 0.001$ yields $d < 24$.

Proof. For $a < 0$ the Eq. (A348) from Theorem (A6) can be written as

$$d \;=\; 1 \;+\; \frac{W_{-1}(a\exp(-b))}{a} \;=\; 1 \;+\; \frac{W_{-1}(-\exp\left(-(-\ln(-a)+b-1)-1\right))}{a} \tag{A369}$$

From [4, Theorem 3.1] we get the following bound on W_{-1}:

$$-\frac{e}{e-1}\,(u+1) \;<\; W_{-1}(-\exp(-u-1)) \;<\; -\,(u+1)\,. \tag{A370}$$

for $u > 0$. We apply Eq. (A370) to Eq. (A369) with $u = -\ln(-a) + b - 1$.

Since $a < 0$ we get

$$d \;>\; 1 \;+\; \frac{-\ln(-a)+b}{a}\,. \tag{A371}$$

□

Storage capacity for the expected minimal separation instead of the probability that all patterns can be stored. In contrast to the previous paragraph, we want to argue about the storage capacity for the expected minimal separation. Therefore, we will use the following bound on the expectation of $\alpha_{\min}$ (minimal angle), which gives also a bound on the expected of $\Delta_{\min}$ (minimal separation):

Lemma A16 (Proposition 3.6 in [12])**.** *We have the following lower bound on the expectation of* $\alpha_{\min}$*:*

$$\mathrm{E}\left[N^{\frac{2}{d-1}}\;\alpha_{\min}\right] \;\geq\; \left(\frac{\Gamma(\frac{d}{2})}{2(d-1)\,\sqrt{\pi}\,\Gamma(\frac{d-1}{2})}\right)^{-\frac{1}{d-1}} \Gamma(1+\frac{1}{d-1})\,\frac{d^{-\frac{1}{d-1}}}{\Gamma(2+\frac{1}{d-1})} \;:=\; C_{d-1}. \tag{A372}$$

The bound is valid for all $N \geq 2$ and $d \geq 2$.

Let us start with some preliminary estimates. First of all we need some asymptotics for the constant C_{d-1} in Eq. (A372):

Lemma A17. *The following estimate holds for $d \geq 2$:*

$$C_d \;\geq\; 1 \;-\; \frac{\ln(d+1)}{d}\,. \tag{A373}$$

Proof. The recursion formula for the Gamma function is [48, (5.5.1)]:

$$\Gamma(x+1) \;=\; x\,\Gamma(x)\,. \tag{A374}$$

We use Eq. (A314) and the fact that $d^{\frac{1}{d}} \geq 1$ for $d \geq 1$ to obtain:

$$\begin{aligned} C_d \;&\geq\; (2\sqrt{d})^{\frac{1}{d}}\Gamma(1+\frac{1}{d})\,\frac{(d+1)^{-\frac{1}{d}}}{\Gamma(2+\frac{1}{d})} \;=\; (2\sqrt{d})^{\frac{1}{d}}\,\frac{(d+1)^{-\frac{1}{d}}}{1-\frac{1}{d}} \;>\; (d+1)^{\frac{1}{d}} \\ &=\; \exp(-\frac{1}{d}\,\ln(d+1)) \;\geq\; 1 \;-\; \frac{1}{d}\,\ln(d+1)\,, \end{aligned} \tag{A375}$$

where in the last step we used the elementary inequality $\exp(x) \geq 1 + x$, which follows from the mean value theorem. □

The next theorem states the number of stored patterns for the expected minimal separation.

Theorem A7 (Storage Capacity (expected separation): Random Patterns)**.** *We assume patterns on the sphere with radius $M = K\sqrt{d-1}$ that are randomly chosen. Then for all values $c \geq 1$ for which*

$$\frac{1}{5}\,(d-1)\,K^2\,c^{-1}(1\;-\;\frac{\ln(d-1)}{(d-1)})^2 \;\geq\; \frac{2}{\beta\,c^{\frac{d-1}{4}}} \;+\; \frac{1}{\beta}\,\ln\left(2\,c^{\frac{d-1}{2}}\,\beta\,(d-1)\,K^2\right) \tag{A376}$$

holds, the number of stored patterns for the expected minimal separation is at least

$$N \;=\; c^{\frac{d-1}{4}}\,. \tag{A377}$$

The inequality Eq. (A376) is e.g. fulfilled with $\beta = 1$, $K = 3$, $c = 2$ and $d \geq 17$.

Proof. Instead of considering the probability that the master inequality Eq. (A300) is fulfilled we now consider whether this inequality is fulfilled for the expected minimal distance. We consider the expectation of the minimal distance $\Delta_{\min}$:

$$\mathrm{E}[\Delta_{\min}] \;=\; \mathrm{E}[M^2(1 \;-\; \cos(\alpha_{\min})))] \;=\; M^2(1 \;-\; \mathrm{E}[\cos(\alpha_{\min}))]) \,. \tag{A378}$$

For this expectation, the master inequality Eq. (A300) becomes

$$M^2(1 \;-\; \mathrm{E}[\cos(\alpha_{\min}))]) \;\geq\; \frac{2}{\beta\, N} \;+\; \frac{1}{\beta}\, \ln\left(2\, N^2\, \beta\, M^2\right) \,. \tag{A379}$$

We want to find the largest N that fulfills this inequality.

We apply Eq. (A318) and Jensen's inequality to deduce the following lower bound:

$$1 \;-\; \mathrm{E}[\cos(\alpha_{\min})] \;\geq\; \frac{1}{5}\, \mathrm{E}\left[\alpha_{\min}^2\right] \;\geq\; \frac{1}{5}\, \mathrm{E}[\alpha_{\min}]^2 \,. \tag{A380}$$

Now we use Eq. (A372) and Eq. (A373) to arrive at

$$\mathrm{E}[\alpha_{\min}]^2 \;\geq\; N^{-\frac{4}{d-1}}\, \mathrm{E}[N^{\frac{2}{d-1}}\, \alpha_{\min}]^2 \;\geq\; N^{-\frac{4}{d-1}}\, C_{d-1}^2 \;\geq\; N^{-\frac{4}{d-1}}\, (1 - \frac{\ln(d-1)}{(d-1)})^2 \,, \tag{A381}$$

for sufficiently large d. Thus in order to fulfill Eq. (A379), it is enough to find values that satisfy Eq. (A376).

□

A2.5.2 Convergence after One Update and Small Retrieval Error

Theorem A8 (Convergence After One Update)**.** *With query $\boldsymbol{\xi}$, after one update the distance of the new point $f(\boldsymbol{\xi})$ to the fixed point $\boldsymbol{x}_i^*$ is exponentially small in the separation Δ_i. The precise bounds are:*

$$\|f(\boldsymbol{\xi}) \;-\; \boldsymbol{x}_i^*\| \;\leqslant\; \|\mathrm{J}^m\|_2\, \|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i^*\| \,, \tag{A382}$$

$$\|\mathrm{J}^m\|_2 \;\leqslant\; 2\, \beta\, N\, M^2\, (N-1) \exp(-\,\beta\, (\Delta_i \;-\; 2\, \max\{\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \;-\; \boldsymbol{x}_i\|\}\, M)) \,. \tag{A383}$$

Proof. From Eq. (A169) we have

$$\|\mathrm{J}^m\|_2 \;\leqslant\; 2\, \beta\, N\, M^2\, (N-1) \exp(-\,\beta\, (\Delta_i \;-\; 2\, \max\{\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \;-\; \boldsymbol{x}_i\|\}\, M)) \,. \tag{A384}$$

After every iteration the mapped point $f(\boldsymbol{\xi})$ is closer to the fixed point $\boldsymbol{x}_i^*$ than the original point $\boldsymbol{x}_i$:

$$\|f(\boldsymbol{\xi}) \;-\; \boldsymbol{x}_i^*\| \;\leqslant\; \|\mathrm{J}^m\|_2\, \|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i^*\| \,. \tag{A385}$$

□

We want to estimate how large Δ_i is. For $\boldsymbol{x}_i$ we have:

$$\Delta_i \;=\; \min_{j, j \neq i} \left(\boldsymbol{x}_i^T \boldsymbol{x}_i \;-\; \boldsymbol{x}_i^T \boldsymbol{x}_j\right) \;=\; \boldsymbol{x}_i^T \boldsymbol{x}_i \;-\; \max_{j, j \neq i} \boldsymbol{x}_i^T \boldsymbol{x}_j \,. \tag{A386}$$

To estimate how large Δ_i is, assume vectors $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{y} \in \mathbb{R}^d$ that have as components standard normally distributed values. The expected value of the separation of two points with normally distributed components is

$$\mathrm{E}\left[\boldsymbol{x}^T \boldsymbol{x} \;-\; \boldsymbol{x}^T \boldsymbol{y}\right] \;=\; \sum_{j=1}^{d} \mathrm{E}\left[x_j^2\right] \;+\; \sum_{j=1}^{d} \mathrm{E}\left[x_j\right] \sum_{j=1}^{d} \mathrm{E}\left[y_j\right] \;=\; d \,. \tag{A387}$$

The variance of the separation of two points with normally distributed components is

$$\begin{aligned}
&\mathrm{Var}\left[\boldsymbol{x}^T\boldsymbol{x} \,-\, \boldsymbol{x}^T\boldsymbol{y}\right] \,=\, \mathrm{E}\left[\left(\boldsymbol{x}^T\boldsymbol{x} \,-\, \boldsymbol{x}^T\boldsymbol{y}\right)^2\right] \,-\, d^2 \qquad \text{(A388)}\\
&= \sum_{j=1}^{d} \mathrm{E}\left[x_j^4\right] \,+\, \sum_{j=1,k=1,k\neq j}^{d} \mathrm{E}\left[x_j^2\right]\, \mathrm{E}\left[x_k^2\right] \,-\, 2\, \sum_{j=1}^{d} \mathrm{E}\left[x_j^3\right] \mathrm{E}\left[y_j\right] \,-\\
&2 \sum_{j=1,k=1,k\neq j}^{d} \mathrm{E}\left[x_j^2\right] \mathrm{E}\left[x_k\right] \mathrm{E}\left[y_k\right] \,+\, \sum_{j=1}^{d} \mathrm{E}\left[x_j^2\right]\, \mathrm{E}\left[y_j^2\right] \,+\\
&\sum_{j=1,k=1,k\neq j}^{d} \mathrm{E}\left[x_j\right] \mathrm{E}\left[y_j\right] \mathrm{E}\left[x_k\right] \mathrm{E}\left[y_k\right] \,-\, d^2\\
&= 3\, d \,+\, d\,(d-1) \,+\, d \,-\, d^2 \,=\, 3\, d\,.
\end{aligned}$$

The expected value for the separation of two random vectors gives:

$$\|\mathrm{J}^m\|_2 \,\leqslant\, 2\,\beta\, N\, M^2\,(N-1) \exp(-\,\beta\,(d \,-\, 2\,\max\{\|\boldsymbol{\xi} \,-\, \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \,-\, \boldsymbol{x}_i\|\}\, M))\,. \qquad \text{(A389)}$$

For the exponential storage we set $M = 2\sqrt{d-1}$. We see the Lipschitz constant $\|\mathrm{J}^m\|_2$ decreases exponentially with the dimension. Therefore, $\|f(\boldsymbol{\xi}) \,-\, \boldsymbol{x}_i^*\|$ is exponentially small after just one update. Therefore, the fixed point is well retrieved after one update.

The retrieval error decreases exponentially with the separation Δ_i.

Theorem A9 (Exponentially Small Retrieval Error)**.** *The retrieval error* $\|\boldsymbol{x}_i \,-\, \boldsymbol{x}_i^*\|$ *of pattern* $\boldsymbol{x}_i$ *is bounded by*

$$\|\boldsymbol{x}_i \,-\, \boldsymbol{x}_i^*\| \,\leqslant\, 2\,(N-1)\, \exp(-\,\beta\,(\Delta_i \,-\, 2\,\|\boldsymbol{x}_i^* \,-\, \boldsymbol{x}_i\|\, M))\, M \qquad \text{(A390)}$$

and for $\|\boldsymbol{x}_i - \boldsymbol{x}_i^*\| \leqslant \frac{1}{2\,\beta\,M}$ *by*

$$\|\boldsymbol{x}_i \,-\, \boldsymbol{x}_i^*\| \,\leqslant\, e\,(N-1)\, M\, \exp(-\,\beta\,\Delta_i)\,. \qquad \text{(A391)}$$

Proof. We compute the retrieval error which is just $\|\boldsymbol{x}_i \,-\, \boldsymbol{x}_i^*\|$. From Lemma A4 we have

$$\|\boldsymbol{x}_i \,-\, f(\boldsymbol{\xi})\| \,\leqslant\, 2\,\epsilon\, M\,, \qquad \text{(A392)}$$

From Eq. (A168) we have

$$\epsilon \,=\, (N-1) \exp(-\,\beta\,(\Delta_i \,-\, 2\,\max\{\|\boldsymbol{\xi} \,-\, \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \,-\, \boldsymbol{x}_i\|\}\, M))\,. \qquad \text{(A393)}$$

We use $\boldsymbol{\xi} = \boldsymbol{x}_i^*$ and get

$$\epsilon \,=\, (N-1) \exp(-\,\beta\,(\Delta_i \,-\, 2\,\|\boldsymbol{x}_i^* \,-\, \boldsymbol{x}_i\|\, M))\,. \qquad \text{(A394)}$$

We obtain

$$\|\boldsymbol{x}_i \,-\, \boldsymbol{x}_i^*\| \,\leqslant\, 2\,(N-1)\, \exp(-\,\beta\,(\Delta_i \,-\, 2\,\|\boldsymbol{x}_i^* \,-\, \boldsymbol{x}_i\|\, M))\, M\,. \qquad \text{(A395)}$$

For $\|\boldsymbol{x}_i - \boldsymbol{x}_i^*\| \leqslant \frac{1}{2\,\beta\,M}$ inequality Eq. (A395) gives

$$\|\boldsymbol{x}_i \,-\, \boldsymbol{x}_i^*\| \,\leqslant\, e\,(N-1)\, M\, \exp(-\,\beta\,\Delta_i)\,. \qquad \text{(A396)}$$

□

A2.6 Learning Associations

We consider three cases of learning associations, i.e. three cases of how sets are associated. (i) Non of the sets is mapped in an associative space. The raw state pattern $\boldsymbol{r}_n$ is the state (query) pattern $\boldsymbol{\xi}_n$, i.e. $\boldsymbol{\xi}_n = \boldsymbol{r}_n$, and the raw stored pattern $\boldsymbol{y}_s$ is the stored pattern (key), i.e. $\boldsymbol{x}_s = \boldsymbol{y}_s$. (ii) Either one of the sets is mapped to the space of the other set or an association matrix is learned. (iia) The state patterns are equal to the raw patterns, i.e. $\boldsymbol{\xi}_n = \boldsymbol{r}_n$, and raw stored patterns are mapped via $\boldsymbol{W}$ to the space of the state patterns, i.e. $\boldsymbol{x}_s = \boldsymbol{W}\boldsymbol{y}_s$. (iib) The stored patterns are equal to the raw patterns, i.e. $\boldsymbol{x}_s = \boldsymbol{y}_s$, and raw state patterns are mapped via $\boldsymbol{W}$ to the space of the stored patterns, i.e. $\boldsymbol{\xi}_n = \boldsymbol{W}^T \boldsymbol{r}_n$. (iic) The matrix $\boldsymbol{W}$ is an association matrix. We will compute the derivative of the new state pattern with respect to $\boldsymbol{W}$, which is valid for all sub-cases (iib)–(iic). (iii) Both set of patterns are mapped in a common associative space. A raw state pattern $\boldsymbol{r}_n$ is mapped by $\boldsymbol{W}_Q$ to a state pattern (query) $\boldsymbol{\xi}_n$, that is $\boldsymbol{\xi}_n = \boldsymbol{W}_Q \boldsymbol{r}_n$. A raw stored pattern $\boldsymbol{y}_s$ is mapped via $\boldsymbol{W}_K$ to stored pattern (key) $\boldsymbol{x}_s$, that is $\boldsymbol{x}_s = \boldsymbol{W}_K \boldsymbol{y}_s$. We will compute the derivative of the new state pattern with respect to both $\boldsymbol{W}_Q$ and $\boldsymbol{W}_K$.

A2.6.1 Association of Raw Patterns – No Mapping in an Associative Space

The sets are associated via their raw patterns, i.e. the raw state pattern $\boldsymbol{r}_n$ is the state (query) pattern $\boldsymbol{\xi}_n$, i.e. $\boldsymbol{\xi}_n = \boldsymbol{r}_n$, and raw stored pattern $\boldsymbol{y}_s$ is the stored pattern (key), i.e. $\boldsymbol{x}_s = \boldsymbol{y}_s$. There is no mapping in an associative space.

The update rule is

$$\boldsymbol{\xi}^{\text{new}} \ = \ \boldsymbol{X} \, \boldsymbol{p} \, , \tag{A397}$$

where we used

$$\boldsymbol{p} \ = \ \text{softmax}(\beta \, \boldsymbol{X}^T \boldsymbol{\xi}) \, . \tag{A398}$$

The derivative with respect to $\boldsymbol{\xi}$ is

$$\frac{\partial \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{\xi}} \ = \ \beta \, \boldsymbol{X} \, \left(\text{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T \right) \, \boldsymbol{X}^T \tag{A399}$$

The derivative with respect to $\boldsymbol{X}$ is

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{X}} \ = \ \boldsymbol{a} \, \boldsymbol{p}^T \ + \ \beta \, \boldsymbol{X} \, \left(\text{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T \right) \, (\boldsymbol{\xi}^T \boldsymbol{a}) \, . \tag{A400}$$

These derivatives allow to apply the chain rule if a Hopfield layer is integrated into a deep neural network.

A2.6.2 Learning an Association Matrix – Only One Set is Mapped in an Associative Space

Only one of the sets $\boldsymbol{R}$ or $\boldsymbol{Y}$ is mapped in the space of the patterns of the other set. Case (a): the state patterns are equal to the raw patterns $\boldsymbol{\xi}_n = \boldsymbol{r}_n$ and raw stored patterns are mapped via $\boldsymbol{W}$ to the space of the state patterns, i.e. $\boldsymbol{x}_s = \boldsymbol{W}\boldsymbol{y}_s$. Case (b): the stored patterns are equal to the raw patterns $\boldsymbol{x}_s = \boldsymbol{y}_s$ and raw state patterns are mapped via $\boldsymbol{W}$ to the space of the stored patterns, i.e. $\boldsymbol{\xi}_n = \boldsymbol{W}^T \boldsymbol{r}_n$. Case (c): the matrix $\boldsymbol{W}$ associates the sets $\boldsymbol{R}$ and $\boldsymbol{Y}$. This case also includes that $\boldsymbol{W}^T = \boldsymbol{W}_K^T \boldsymbol{W}_Q$, which is treated in next subsection. The next subsection focuses on a low rank approximation of $\boldsymbol{W}$ by defining the dimension d_k of associative space and use the matrices $\boldsymbol{W}_K^T$ and $\boldsymbol{W}_Q$ to define $\boldsymbol{W}$, or equivalently to map $\boldsymbol{R}$ and $\boldsymbol{Y}$ into the associative space.

From a mathematical point of view all these case are equal as they lead to the same update rule. Therefore, we consider in the following Case (a) with $\boldsymbol{x}_s = \boldsymbol{W}\boldsymbol{y}_s$ and $\boldsymbol{\xi}_n = \boldsymbol{r}_n$. Still, the following formula are valid for all three cases (a)–(c).

The update rule is

$$\boldsymbol{\xi}^{\text{new}} \ = \ \boldsymbol{W} \, \boldsymbol{Y} \, \boldsymbol{p} \, , \tag{A401}$$

where we used

$$\boldsymbol{p} \ = \ \text{softmax}(\beta \, \boldsymbol{Y}^T \boldsymbol{W}^T \boldsymbol{\xi}) \, . \tag{A402}$$

We consider the state (query) pattern $\boldsymbol{\xi}$ with result $\boldsymbol{\xi}^{\text{new}}$:

$$\boldsymbol{\xi}^{\text{new}} \ = \ \boldsymbol{W} \ \boldsymbol{Y} \ \boldsymbol{p} \ = \ \boldsymbol{W} \ \boldsymbol{Y} \ \text{softmax}(\beta \ \boldsymbol{Y}^T \boldsymbol{W}^T \boldsymbol{\xi}) \tag{A403}$$

For multiple updates this update rule has to be used. However for a single update, or the last update we consider a simplified update rule.

Since new state vector $\boldsymbol{\xi}^{\text{new}}$ is projected by a weight matrix $\boldsymbol{W}_V$ to another vector, we consider the simplified update rule:

$$\boldsymbol{\xi}^{\text{new}} \ = \ \boldsymbol{Y} \ \boldsymbol{p} \ = \ \boldsymbol{Y} \ \text{softmax}(\beta \ \boldsymbol{Y}^T \boldsymbol{W}^T \boldsymbol{\xi}) \tag{A404}$$

The derivative with respect to $\boldsymbol{W}$ is

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{W}} \ = \ \frac{\partial \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{W}} \ \frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{\xi}^{\text{new}}} \ = \ \frac{\partial \boldsymbol{\xi}^{\text{new}}}{\partial (\boldsymbol{W}^T \boldsymbol{\xi})} \ \frac{\partial (\boldsymbol{W}^T \boldsymbol{\xi})}{\partial \boldsymbol{W}} \ \frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{\xi}^{\text{new}}} \ . \tag{A405}$$

$$\frac{\partial \boldsymbol{\xi}^{\text{new}}}{\partial (\boldsymbol{W}^T \boldsymbol{\xi})} \ = \ \beta \ \boldsymbol{Y} \ \left(\text{diag}(\boldsymbol{p}) - \boldsymbol{p} \boldsymbol{p}^T \right) \ \boldsymbol{Y}^T \tag{A406}$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{\xi}^{\text{new}}} \ = \ \boldsymbol{a} \ . \tag{A407}$$

We have the product of the 3-dimensional tensor $\frac{\partial (\boldsymbol{W}^T \boldsymbol{\xi})}{\partial \boldsymbol{W}}$ with the vector $\boldsymbol{a}$ which gives a 2-dimensional tensor, i.e. a matrix:

$$\frac{\partial (\boldsymbol{W}^T \boldsymbol{\xi})}{\partial \boldsymbol{W}} \ \frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{\xi}^{\text{new}}} \ = \ \frac{\partial (\boldsymbol{W}^T \boldsymbol{\xi})}{\partial \boldsymbol{W}} \ \boldsymbol{a} \ = \ \boldsymbol{\xi}^T \boldsymbol{a} \boldsymbol{I} \ . \tag{A408}$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{W}} \ = \ \beta \ \boldsymbol{Y} \ \left(\text{diag}(\boldsymbol{p}) - \boldsymbol{p} \boldsymbol{p}^T \right) \ \boldsymbol{Y}^T (\boldsymbol{\xi}^T \boldsymbol{a}) \ = \ \text{J} \ (\boldsymbol{\xi}^T \boldsymbol{a}) \ , \tag{A409}$$

where J is the Jacobian of the update rule defined in Eq. (A48).

To obtain the derivative of the full update rule Eq. (A403) we have to add the term

$$\boldsymbol{a} \ \boldsymbol{p}^T \boldsymbol{Y}^T \tag{A410}$$

and include the factor $\boldsymbol{W}$ to get

$$\begin{aligned} \frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{W}} \ &= \ \boldsymbol{a} \ \boldsymbol{p}^T \boldsymbol{Y}^T \ + \ \beta \ \boldsymbol{W} \ \boldsymbol{Y} \ \left(\text{diag}(\boldsymbol{p}) - \boldsymbol{p} \boldsymbol{p}^T \right) \ \boldsymbol{Y}^T (\boldsymbol{\xi}^T \boldsymbol{a}) \\ &= \ \boldsymbol{a} \ \boldsymbol{p}^T \boldsymbol{Y}^T \ + \ \boldsymbol{W} \ \text{J} \ (\boldsymbol{\xi}^T \boldsymbol{a}) \ . \end{aligned} \tag{A411}$$

A2.6.3 Learning Two Association Mappings – Both Sets are Mapped in an Associative Space

Both sets $\boldsymbol{R}$ and $\boldsymbol{Y}$ are mapped in an associative space. Every raw state pattern $\boldsymbol{r}_n$ is mapped via $\boldsymbol{W}_Q$ to a state pattern (query) $\boldsymbol{\xi}_n = \boldsymbol{W}_Q \boldsymbol{r}_n$. Every raw stored pattern $\boldsymbol{y}_s$ is mapped via $\boldsymbol{W}_K$ to a stored pattern (key) $\boldsymbol{x}_s = \boldsymbol{W}_K \boldsymbol{y}_s$. In the last subsection we considered a single matrix $\boldsymbol{W}$. For $\boldsymbol{W}^T = \boldsymbol{W}_K^T \boldsymbol{W}_Q$ we have the case of the last subsection. However in this subsection we are looking for a low rank approximation of $\boldsymbol{W}$. Toward this end we define the dimension d_k of associative space and use the matrices $\boldsymbol{W}_K^T$ and $\boldsymbol{W}_Q$ to map to the associative space.

The update rule is

$$\boldsymbol{\xi}^{\text{new}} \ = \ \boldsymbol{X} \ \boldsymbol{p} \ , \tag{A412}$$

where we used

$$\boldsymbol{p} \ = \ \text{softmax}(\beta \ \boldsymbol{X}^T \boldsymbol{\xi}) \ . \tag{A413}$$

We consider raw state patterns $\boldsymbol{r}_n$ that are mapped to state patterns $\boldsymbol{\xi}_n = \boldsymbol{W}_Q \boldsymbol{r}_n$ with $\boldsymbol{Q}^T = \boldsymbol{\Xi} = \boldsymbol{W}_Q \boldsymbol{R}$ and raw stored pattern $\boldsymbol{y}_s$ that are mapped to stored patterns $\boldsymbol{x}_s = \boldsymbol{W}_K \boldsymbol{y}_s$ with $\boldsymbol{K}^T = \boldsymbol{X} = \boldsymbol{W}_K \boldsymbol{Y}$. The update rule is

$$\boldsymbol{\xi}^{\text{new}} = \boldsymbol{W}_K \, \boldsymbol{Y} \, \boldsymbol{p} = \boldsymbol{W}_K \, \boldsymbol{Y} \, \text{softmax}(\beta \, \boldsymbol{Y}^T \boldsymbol{W}_K^T \boldsymbol{W}_Q \, \boldsymbol{r}) \, . \tag{A414}$$

Since new state vector $\boldsymbol{\xi}^{\text{new}}$ is projected by a weight matrix $\boldsymbol{W}_V$ to another vector, we consider the simplified update rule:

$$\boldsymbol{\xi}^{\text{new}} = \boldsymbol{Y} \, \boldsymbol{p} = \boldsymbol{Y} \, \text{softmax}(\beta \, \boldsymbol{Y}^T \boldsymbol{W}_K^T \boldsymbol{W}_Q \, \boldsymbol{r}) \, . \tag{A415}$$

For the simplified update rule, the vector $\boldsymbol{\xi}^{\text{new}}$ does not live in the associative space but in the space of raw stored pattern $\boldsymbol{y}$. However $\boldsymbol{W}_K$ would map it to the associative space.

Derivative with respect to $\boldsymbol{W}_Q$. The derivative with respect to $\boldsymbol{W}_Q$ is

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{W}_Q} = \frac{\partial \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{W}_Q} \frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{\xi}^{\text{new}}} = \frac{\partial \boldsymbol{\xi}^{\text{new}}}{\partial (\boldsymbol{W}_Q \, \boldsymbol{r})} \frac{\partial (\boldsymbol{W}_Q \, \boldsymbol{r})}{\partial \boldsymbol{W}_Q} \frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{\xi}^{\text{new}}} \, . \tag{A416}$$

$$\frac{\partial \boldsymbol{\xi}^{\text{new}}}{\partial (\boldsymbol{W}_Q \, \boldsymbol{r})} = \beta \, \boldsymbol{Y} \, \left(\text{diag}(\boldsymbol{p}) - \boldsymbol{p} \boldsymbol{p}^T \right) \, \boldsymbol{Y}^T \boldsymbol{W}_K^T \tag{A417}$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{\xi}^{\text{new}}} = \boldsymbol{a} \, . \tag{A418}$$

We have the product of the 3-dimensional tensor $\frac{\partial (\boldsymbol{W}_Q \boldsymbol{r})}{\partial \boldsymbol{W}_Q}$ with the vector $\boldsymbol{a}$ which gives a 2-dimensional tensor, i.e. a matrix:

$$\frac{\partial (\boldsymbol{W}_Q \, \boldsymbol{r})}{\partial \boldsymbol{W}_Q} \frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{\xi}^{\text{new}}} = \frac{\partial (\boldsymbol{W}_Q \, \boldsymbol{r})}{\partial \boldsymbol{W}_Q} \, \boldsymbol{a} = \boldsymbol{r}^T \boldsymbol{a} \, \boldsymbol{I} \, . \tag{A419}$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{W}_Q} = \beta \, \boldsymbol{Y} \, \left(\text{diag}(\boldsymbol{p}) - \boldsymbol{p} \boldsymbol{p}^T \right) \, \boldsymbol{Y}^T \, \boldsymbol{W}_K^T (\boldsymbol{r}^T \boldsymbol{a}) = \text{J} \, \boldsymbol{W}_K^T (\boldsymbol{r}^T \boldsymbol{a}) \, , \tag{A420}$$

where J is the Jacobian of the update rule defined in Eq. (A48).

To obtain the derivative of the full update rule Eq. (A414) we have to include the factor $\boldsymbol{W}_K$, then get

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{W}_Q} = \beta \, \boldsymbol{W}_K \, \boldsymbol{Y} \left(\text{diag}(\boldsymbol{p}) - \boldsymbol{p} \boldsymbol{p}^T \right) \, \boldsymbol{Y}^T \, \boldsymbol{W}_K^T (\boldsymbol{r}^T \boldsymbol{a}) = \boldsymbol{W}_K \, \text{J} \, \boldsymbol{W}_K^T (\boldsymbol{r}^T \boldsymbol{a}) \, . \tag{A421}$$

Derivative with respect to $\boldsymbol{W}_K$. The derivative with respect to $\boldsymbol{W}_K$ is

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{W}_K} = \frac{\partial \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{W}_K} \frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{\xi}^{\text{new}}} = \frac{\partial \boldsymbol{\xi}^{\text{new}}}{\partial (\boldsymbol{W}_K^T \boldsymbol{W}_Q \, \boldsymbol{r})} \frac{\partial (\boldsymbol{W}_K^T \boldsymbol{W}_Q \, \boldsymbol{r})}{\partial \boldsymbol{W}_K} \frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{\xi}^{\text{new}}} \, . \tag{A422}$$

$$\frac{\partial \boldsymbol{\xi}^{\text{new}}}{\partial (\boldsymbol{W}_K^T \boldsymbol{W}_Q \, \boldsymbol{r})} = \beta \, \boldsymbol{Y} \, \left(\text{diag}(\boldsymbol{p}) - \boldsymbol{p} \boldsymbol{p}^T \right) \, \boldsymbol{Y}^T \tag{A423}$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{\xi}^{\text{new}}} = \boldsymbol{a} \, . \tag{A424}$$

We have the product of the 3-dimensional tensor $\frac{\partial (\boldsymbol{W} \boldsymbol{r})}{\partial \boldsymbol{W}_K}$ with the vector $\boldsymbol{a}$ which gives a 2-dimensional tensor, i.e. a matrix:

$$\frac{\partial (\boldsymbol{W}_K^T \boldsymbol{W}_Q \, \boldsymbol{r})}{\partial \boldsymbol{W}_K} \frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{\xi}^{\text{new}}} = \frac{\partial (\boldsymbol{W}_K^T \boldsymbol{W}_Q \, \boldsymbol{r})}{\partial \boldsymbol{W}_K} \, \boldsymbol{a} = \boldsymbol{W}_Q^T \boldsymbol{r}^T \boldsymbol{a} \, \boldsymbol{I} \, . \tag{A425}$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{W}_K} \ = \ \beta \, \boldsymbol{Y} \, \left(\text{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right) \, \boldsymbol{Y}^T \, (\boldsymbol{W}_Q^T \boldsymbol{r}^T \boldsymbol{a}) \ = \ \text{J} \, (\boldsymbol{W}_Q^T \boldsymbol{r}^T \boldsymbol{a}) \, , \tag{A426}$$

where J is the Jacobian of the update rule defined in Eq. (A48).

To obtain the derivative of the full update rule Eq. (A414) we have to add the term

$$\boldsymbol{a} \, \boldsymbol{p}^T \boldsymbol{Y}^T \tag{A427}$$

and to include the factor $\boldsymbol{W}_K$, then get

$$\begin{aligned} \frac{\partial \boldsymbol{a}^T \boldsymbol{\xi}^{\text{new}}}{\partial \boldsymbol{W}_K} \ &= \ \boldsymbol{a} \, \boldsymbol{p}^T \boldsymbol{Y}^T \ + \ \beta \, \boldsymbol{W}_K \, \boldsymbol{Y} \, \left(\text{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right) \, \boldsymbol{Y}^T (\boldsymbol{W}_Q^T \boldsymbol{r}^T \boldsymbol{a}) \\ &= \ \boldsymbol{a} \, \boldsymbol{p}^T \boldsymbol{Y}^T \ + \ \boldsymbol{W}_K \, \text{J} \, (\boldsymbol{W}_Q^T \boldsymbol{r}^T \boldsymbol{a}) \, . \end{aligned} \tag{A428}$$

A2.7 Infinite Many Patterns and Forgetting Patterns

In the next subsection we show how the new Hopfield networks can be used for auto-regressive tasks by causal masking. In the following subsection, we introduce forgetting to the new Hopfield networks by adding a negative value to the softmax which is larger if the pattern was observed more in the past.

A2.7.1 Infinite Many Patterns

The new Hopfield networks can be used for auto-regressive tasks, that is time series prediction and similar. Causal masking masks out the future by a large negative value in the softmax.

We assume to have infinite many stored patterns (keys) $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ that are represented by the infinite matrix

$$\boldsymbol{X} \ = \ (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots,) \ . \tag{A429}$$

The pattern index is now a time index, that is, we observe $\boldsymbol{x}_t$ at time t.

The pattern matrix at time t is

$$\boldsymbol{X}_t \ = \ (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t) \ . \tag{A430}$$

The query at time t is $\boldsymbol{\xi}_t$.

For $M_t = \max_{1 \leqslant i \leqslant t} \|\boldsymbol{x}_t\|$, the energy function at time t is E_t

$$\text{E}_t \ = \ - \, \text{lse}(\beta, \boldsymbol{X}_t^T \boldsymbol{\xi}_t) \ + \ \frac{1}{2} \boldsymbol{\xi}_t^T \boldsymbol{\xi}_t \ + \ \beta^{-1} \ln t \ + \ \frac{1}{2} M_t^2 \tag{A431}$$

$$= \ - \, \beta^{-1} \ln \left(\sum_{i=1}^{t} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}_t) \right) \ + \ \frac{1}{2} \boldsymbol{\xi}_t^T \boldsymbol{\xi}_t \ + \ \beta^{-1} \ln t \ + \ \frac{1}{2} M_t^2 \, . \tag{A432}$$

The update rule is

$$\boldsymbol{\xi}_t^{\text{new}} \ = \ \boldsymbol{X}_t \, \boldsymbol{p}_t \ = \ \boldsymbol{X}_t \, \text{softmax}(\beta \, \boldsymbol{X}_t^T \boldsymbol{\xi}_t) \, , \tag{A433}$$

where we used

$$\boldsymbol{p}_t \ = \ \text{softmax}(\beta \, \boldsymbol{X}_t^T \boldsymbol{\xi}_t) \, . \tag{A434}$$

We can use an infinite pattern matrix with an infinite softmax when using causal masking. The pattern matrix at time t is

$$\boldsymbol{X}_t \ = \ (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t, -\alpha \boldsymbol{\xi}_t, -\alpha \boldsymbol{\xi}_t, \ldots) \ , \tag{A435}$$

with the query $\boldsymbol{\xi}_t$ and $\alpha \to \infty$. The energy function at time t is E_t

$$\text{E}_t \ = \ - \, \text{lse}(\beta, \boldsymbol{X}_t^T \boldsymbol{\xi}_t) \ + \ \frac{1}{2} \boldsymbol{\xi}_t^T \boldsymbol{\xi}_t \ + \ \beta^{-1} \ln t \ + \ \frac{1}{2} M_t^2 \tag{A436}$$

$$= \ - \, \beta^{-1} \ln \left(\sum_{i=1}^{t} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}_t) \ + \ \sum_{i=t+1}^{\lfloor \alpha \rfloor} \exp(-\beta \alpha \|\boldsymbol{\xi}_t\|^2) \right) \ + \ \frac{1}{2} \boldsymbol{\xi}_t^T \boldsymbol{\xi}_t \ + \tag{A437}$$

$$\beta^{-1} \ln t \ + \ \frac{1}{2} M_t^2 \, .$$

For $\alpha \to \infty$ and $\|\boldsymbol{\xi}_t\| > 0$ this becomes

$$\mathrm{E}_t \ = \ - \ \mathrm{lse}(\beta, \boldsymbol{X}_t^T \boldsymbol{\xi}_t) \ + \ \frac{1}{2} \boldsymbol{\xi}_t^T \boldsymbol{\xi}_t \ + \ \beta^{-1} \ln t \ + \ \frac{1}{2} M_t^2 \tag{A438}$$

$$= \ - \ \beta^{-1} \ln \left(\sum_{i=1}^{t} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}_t) \right) \ + \ \frac{1}{2} \boldsymbol{\xi}_t^T \boldsymbol{\xi}_t \ + \ \beta^{-1} \ln t \ + \ \frac{1}{2} M_t^2 \ . \tag{A439}$$

A2.7.2 Forgetting Patterns

We introduce forgetting to the new Hopfield networks by adding a negative value in the softmax which increases with patterns that are more in the past.

We assume to have infinite many patterns $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ that are represented by the infinite matrix

$$\boldsymbol{X} \ = \ (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots,) \ . \tag{A440}$$

The pattern index is now a time index, that is, we observe $\boldsymbol{x}_t$ at time t.

The pattern matrix at time t is

$$\boldsymbol{X}_t \ = \ (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t) \ . \tag{A441}$$

The query at time t is $\boldsymbol{\xi}_t$.

The energy function with forgetting parameter γ at time t is E_t

$$\mathrm{E}_t \ = \ - \ \mathrm{lse}(\beta, \boldsymbol{X}_t^T \boldsymbol{\xi}_t \ - \ \gamma (t-1, t-2, \ldots, 0)^T) \ + \ \frac{1}{2} \boldsymbol{\xi}_t^T \boldsymbol{\xi}_t \ + \ \beta^{-1} \ln t \ + \ \frac{1}{2} M_t^2 \tag{A442}$$

$$= \ - \ \beta^{-1} \ln \left(\sum_{i=1}^{T} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}_t \ - \ \gamma (t-i)) \right) \ + \ \frac{1}{2} \boldsymbol{\xi}_t^T \boldsymbol{\xi}_t \ + \ \beta^{-1} \ln t \ + \ \frac{1}{2} M_t^2 \ . \tag{A443}$$

The update rule is

$$\boldsymbol{\xi}_t^{\mathrm{new}} \ = \ \boldsymbol{X}_t \, \boldsymbol{p}_t \ = \ \boldsymbol{X}_t \, \mathrm{softmax}(\beta \boldsymbol{X}_t^T \boldsymbol{\xi}_t) \ , \tag{A444}$$

where we used

$$\boldsymbol{p}_t \ = \ \mathrm{softmax}(\beta \boldsymbol{X}_t^T \boldsymbol{\xi}_t) \ . \tag{A445}$$

A3 Properties of Softmax, Log-Sum-Exponential, Legendre Transform, Lambert W Function

For $\beta > 0$, the *softmax* is defined as

Definition A1 (Softmax)**.**

$$\boldsymbol{p} \ = \ \mathrm{softmax}(\beta \boldsymbol{x}) \tag{A446}$$

$$p_i \ = \ [\mathrm{softmax}(\beta \boldsymbol{x})]_i \ = \ \frac{\exp(\beta x_i)}{\sum_k \exp(\beta x_k)} \ . \tag{A447}$$

We also need the *log-sum-exp function* (lse), defined as

Definition A2 (Log-Sum-Exp Function)**.**

$$\mathrm{lse}(\beta, \boldsymbol{x}) \ = \ \beta^{-1} \ln \left(\sum_{i=1}^{N} \exp(\beta x_i) \right) \ . \tag{A448}$$

Next, we give the relation between the softmax and the lse function.

Lemma A18. *The softmax is the gradient of the* lse*:*

$$\mathrm{softmax}(\beta \boldsymbol{x}) \ = \ \nabla_{\boldsymbol{x}} \mathrm{lse}(\beta, \boldsymbol{x}) \ . \tag{A449}$$

In the next lemma we report some important properties of the lse function.

Lemma A19. *We define*

$$\mathrm{L} \ := \ \boldsymbol{z}^T \boldsymbol{x} \ - \ \beta^{-1} \sum_{i=1}^{N} z_i \ln z_i \tag{A450}$$

with $\mathrm{L} \geq \boldsymbol{p}^T \boldsymbol{x}$. *The* lse *is the maximum of* L *on the* N*-dimensional simplex* D *with* $D = \{\boldsymbol{z} \mid \sum_i z_i = 1, 0 \leqslant z_i\}$*:*

$$\mathrm{lse}(\beta, \boldsymbol{x}) \ = \ \max_{\boldsymbol{z} \in D} \boldsymbol{z}^T \boldsymbol{x} \ - \ \beta^{-1} \sum_{i=1}^{N} z_i \ln z_i \ . \tag{A451}$$

The softmax $\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{x})$ *is the argument of the maximum of* L *on the* N*-dimensional simplex* D *with* $D = \{\boldsymbol{z} \mid \sum_i z_i = 1, 0 \leqslant z_i\}$*:*

$$\boldsymbol{p} \ = \ \mathrm{softmax}(\beta \boldsymbol{x}) \ = \ \arg\max_{\boldsymbol{z} \in D} \boldsymbol{z}^T \boldsymbol{x} \ - \ \beta^{-1} \sum_{i=1}^{N} z_i \ln z_i \ . \tag{A452}$$

Proof. Eq. (A451) is obtained from Equation (8) in [29] and Eq. (A452) from Equation (11) in [29]. □

From a physical point of view, the lse function represents the "free energy" in statistical thermodynamics [29].

Next we consider the Jacobian of the softmax and its properties.

Lemma A20. *The Jacobian* J_s *of the softmax* $\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{x})$ *is*

$$\mathrm{J}_s \ = \ \frac{\partial \mathrm{softmax}(\beta \boldsymbol{x})}{\partial \boldsymbol{x}} \ = \ \beta \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p} \boldsymbol{p}^T \right) \ , \tag{A453}$$

which gives the elements

$$[\mathrm{J}_s]_{ij} \ = \ \begin{cases} \beta p_i (1 - p_i) & \text{for } i = j \\ -\beta p_i p_j & \text{for } i \neq j \end{cases} \ . \tag{A454}$$

Next we show that J_s has eigenvalue 0.

Lemma A21. *The Jacobian* J_s *of the softmax function* $\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{x})$ *has a zero eigenvalue with eigenvector* $\mathbf{1}$*.*

Proof.

$$[\mathrm{J}_s \mathbf{1}]_i \ = \ \beta \left(p_i (1 - p_i) \ - \ \sum_{j, j \neq i} p_i p_j \right) \ = \ \beta \, p_i (1 \ - \ \sum_j p_j) \ = 0 \ . \tag{A455}$$

□

Next we show that 0 is the smallest eigenvalue of J_s, therefore J_s is positive semi-definite but not (strict) positive definite.

Lemma A22. *The Jacobian* J_s *of the softmax* $\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{\xi})$ *is symmetric and positive semi-definite.*

Proof. For an arbitrary $\boldsymbol{z}$, we have

$$\begin{aligned} \boldsymbol{z}^T \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p} \boldsymbol{p}^T \right) \boldsymbol{z} \ &= \ \sum_i p_i z_i^2 - \left(\sum_i p_i z_i \right)^2 \\ &= \ \left(\sum_i p_i z_i^2 \right) \left(\sum_i p_i \right) - \left(\sum_i p_i z_i \right)^2 \ \geq \ 0 \ . \end{aligned} \tag{A456}$$

The last inequality hold true because the Cauchy-Schwarz inequality says $(\boldsymbol{a}^T\boldsymbol{a})(\boldsymbol{b}^T\boldsymbol{b}) \geq (\boldsymbol{a}^T\boldsymbol{b})^2$, which is the last inequality with $a_i = z_i\sqrt{p_i}$ and $b_i = \sqrt{p_i}$. Consequently $\left(\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)$ is positive semi-definite.

Alternatively $\sum_i p_i z_i^2 - \left(\sum_i p_i z_i\right)^2$ can be viewed as the expected second moment minus the mean squared which gives the variance that is larger equal to zero.

The Jacobian is $0 < \beta$ times a positive semi-definite matrix, which is a positive semi-definite matrix. □

Moreover, the softmax is a monotonic map, as described in the next lemma.

Lemma A23. *The softmax $\boldsymbol{p} = \operatorname{softmax}(\beta\boldsymbol{x})$ is monotone, that is,*

$$\left(\operatorname{softmax}(\beta\boldsymbol{x}) \;-\; \operatorname{softmax}(\beta\boldsymbol{x}')\right)^T (\boldsymbol{x} \;-\; \boldsymbol{x}') \;\geq\; 0\,. \tag{A457}$$

Proof. We use the version of mean value theorem Lemma A32 with the symmetric matrix $\mathrm{J}_s^m = \int_0^1 \mathrm{J}_s(\lambda\boldsymbol{x} \;+\; (1-\lambda)\boldsymbol{x}')\,\mathrm{d}\lambda$:

$$\operatorname{softmax}(\boldsymbol{x}) \;-\; \operatorname{softmax}(\boldsymbol{x}') \;=\; \mathrm{J}_s^m \;(\boldsymbol{x} \;-\; \boldsymbol{x}')\;. \tag{A458}$$

Therefore

$$\left(\operatorname{softmax}(\boldsymbol{x}) \;-\; \operatorname{softmax}(\boldsymbol{x}')\right)^T (\boldsymbol{x} \;-\; \boldsymbol{x}') \;=\; (\boldsymbol{x} \;-\; \boldsymbol{x}')^T\, \mathrm{J}_s^m \;(\boldsymbol{x} \;-\; \boldsymbol{x}') \;\geq\; 0\,, \tag{A459}$$

since J_s^m is positive semi-definite. For all λ the Jacobians $\mathrm{J}_s(\lambda\boldsymbol{x} \;+\; (1-\lambda)\boldsymbol{x}')$ are positive semi-definite according to Lemma A22. Since

$$\boldsymbol{x}^T \mathrm{J}_s^m \boldsymbol{x} \;=\; \int_0^1 \boldsymbol{x}^T \mathrm{J}_s(\lambda\boldsymbol{x} \;+\; (1-\lambda)\boldsymbol{x}')\;\boldsymbol{x}\,\mathrm{d}\lambda \;\geq\; 0 \tag{A460}$$

is an integral over positive values for every $\boldsymbol{x}$, J_s^m is positive semi-definite, too. □

Next we give upper bounds on the norm of J_s.

Lemma A24. *For a softmax $\boldsymbol{p} = \operatorname{softmax}(\beta\boldsymbol{x})$ with $m = \max_i p_i(1-p_i)$, the spectral norm of the Jacobian J_s of the softmax is bounded:*

$$\|\mathrm{J}_s\|_2 \;\leqslant\; 2\,m\,\beta\,, \tag{A461}$$
$$\|\mathrm{J}_s\|_1 \;\leqslant\; 2\,m\,\beta\,, \tag{A462}$$
$$\|\mathrm{J}_s\|_\infty \;\leqslant\; 2\,m\,\beta\,. \tag{A463}$$

In particular everywhere holds

$$\|\mathrm{J}_s\|_2 \;\leqslant\; \frac{1}{2}\,\beta\,. \tag{A464}$$

If $p_{\max} = \max_i p_i \geq 1 - \epsilon \geq 0.5$, then for the spectral norm of the Jacobian holds

$$\|\mathrm{J}_s\|_2 \;\leqslant\; 2\,\epsilon\,\beta \;-\; 2\,\epsilon^2\,\beta \;<\; 2\,\epsilon\,\beta\,. \tag{A465}$$

Proof. We consider the maximum absolute column sum norm

$$\|\boldsymbol{A}\|_1 \;=\; \max_j \sum_i |a_{ij}| \tag{A466}$$

and the maximum absolute row sum norm

$$\|\boldsymbol{A}\|_\infty \;=\; \max_i \sum_j |a_{ij}|\,. \tag{A467}$$

We have for $\boldsymbol{A} = \mathrm{J}_s = \beta \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)$

$$\begin{aligned}\sum_j |a_{ij}| &= \beta \left(p_i(1-p_i) + \sum_{j, j \neq i} p_i p_j \right) = \beta\, p_i \,(1 - 2p_i + \sum_j p_j) \\ &= 2\,\beta\, p_i\,(1-p_i) \leqslant 2\,m\,\beta\,, \end{aligned} \tag{A468}$$

$$\begin{aligned}\sum_i |a_{ij}| &= \beta \left(p_j\,(1-p_j) + \sum_{i, i \neq j} p_j p_i \right) = \beta\, p_j \,(1 - 2p_j + \sum_i p_i) \\ &= 2\,\beta\, p_j\,(1-p_j) \leqslant 2\,m\,\beta\,. \end{aligned} \tag{A469}$$

Therefore, we have

$$\|\mathrm{J}_s\|_1 \leqslant 2\,m\,\beta\,, \tag{A470}$$

$$\|\mathrm{J}_s\|_\infty \leqslant 2\,m\,\beta\,, \tag{A471}$$

$$\|\mathrm{J}_s\|_2 \leqslant \sqrt{\|\mathrm{J}_s\|_1 \|\mathrm{J}_s\|_\infty} \leqslant 2\,m\,\beta\,. \tag{A472}$$

The last inequality is a direct consequence of Hölder's inequality.

For $0 \leqslant p_i \leqslant 1$, we have $p_i(1-p_i) \leqslant 0.25$. Therefore, $m \leqslant 0.25$ for all values of p_i.

If $p_{\max} \geq 1 - \epsilon \geq 0.5$ ($\epsilon \leqslant 0.5$), then $1 - p_{\max} \leqslant \epsilon$ and for $p_i \neq p_{\max}$ $p_i \leqslant \epsilon$. The derivative $\partial x(1-x)/\partial x = 1 - 2x > 0$ for $x < 0.5$, therefore $x(1-x)$ increases with x for $x < 0.5$. Using $x = 1 - p_{\max}$ and for $p_i \neq p_{\max}$ $x = p_i$, we obtain $p_i(1-p_i) \leqslant \epsilon(1-\epsilon)$ for all i. Consequently, we have $m \leqslant \epsilon(1-\epsilon)$. □

Using the bounds on the norm of the Jacobian, we give some Lipschitz properties of the softmax function.

Lemma A25. *The softmax function $\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{x})$ is $(\beta/2)$-Lipschitz. The softmax function $\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{x})$ is $(2\beta m)$-Lipschitz in a convex environment U for which $m = \max_{\boldsymbol{x} \in U} \max_i p_i(1 - p_i)$. For $p_{\max} = \min_{\boldsymbol{x} \in U} \max_i p_i = 1 - \epsilon$, the softmax function $\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{x})$ is $(2\beta\epsilon)$-Lipschitz. For $\beta < 2m$, the softmax $\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{x})$ is contractive in U on which m is defined.*

Proof. The version of mean value theorem Lemma A32 states for the symmetric matrix $\mathrm{J}_s^m = \int_0^1 \mathrm{J}(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{x}')\,\mathrm{d}\lambda$:

$$\mathrm{softmax}(\boldsymbol{x}) - \mathrm{softmax}(\boldsymbol{x}') = \mathrm{J}_s^m\,(\boldsymbol{x} - \boldsymbol{x}')\,. \tag{A473}$$

According to Lemma A24 for all $\tilde{\boldsymbol{x}} = \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{x}'$)

$$\|\mathrm{J}_s(\tilde{\boldsymbol{x}})\|_2 \leqslant 2\,\tilde{m}\,\beta\,, \tag{A474}$$

where $\tilde{m} = \max_i \tilde{p}_i(1-\tilde{p}_i)$. Since $\boldsymbol{x} \in U$ and $\boldsymbol{x}' \in U$ we have $\tilde{\boldsymbol{x}} \in U$, since U is convex. For $m = \max_{\boldsymbol{x} \in U} \max_i p_i(1-p_i)$ we have $\tilde{m} \leqslant m$ for all $\tilde{m}$. Therefore, we have

$$\|\mathrm{J}_s(\tilde{\boldsymbol{x}})\|_2 \leqslant 2\,m\,\beta \tag{A475}$$

which also holds for the mean:

$$\|\mathrm{J}_s^m\|_2 \leqslant 2\,m\,\beta\,. \tag{A476}$$

Therefore,

$$\|\mathrm{softmax}(\boldsymbol{x}) - \mathrm{softmax}(\boldsymbol{x}')\| \leqslant \|\mathrm{J}_s^m\|_2\, \|\boldsymbol{x} - \boldsymbol{x}'\| \leqslant 2\,m\,\beta\,\|\boldsymbol{x} - \boldsymbol{x}'\|\,. \tag{A477}$$

From Lemma A24 we know $m \leqslant 1/4$ globally. For $p_{\max} = \min_{\boldsymbol{x} \in U} \max_i p_i = 1 - \epsilon$ we have according to Lemma A24: $m \leqslant \epsilon$. □

For completeness we present a result about cocoercivity of the softmax:

Lemma A26. *For $m = \max_{\boldsymbol{x}\in U} \max_i p_i(1-p_i)$, softmax function $\boldsymbol{p} = \mathrm{softmax}(\beta\boldsymbol{x})$ is $1/(2m\beta)$-cocoercive in U, that is,*

$$\left(\mathrm{softmax}(\boldsymbol{x}) - \mathrm{softmax}(\boldsymbol{x}')\right)^T (\boldsymbol{x} - \boldsymbol{x}') \geq \frac{1}{2\,m\,\beta} \|\mathrm{softmax}(\boldsymbol{x}) - \mathrm{softmax}(\boldsymbol{x}')\|. \quad \text{(A478)}$$

In particular the softmax function $\boldsymbol{p} = \mathrm{softmax}(\beta\boldsymbol{x})$ is $(2/\beta)$-cocoercive everywhere. With $p_{\max} = \min_{\boldsymbol{x}\in U} \max_i p_i = 1-\epsilon$, the softmax function $\boldsymbol{p} = \mathrm{softmax}(\beta\boldsymbol{x})$ is $1/(2\beta\epsilon)$-cocoercive in U.

Proof. We apply the Baillon-Haddad theorem (e.g. Theorem 1 in [29]) together with Lemma A25. □

Finally, we introduce the Legendre transform and use it to describe further properties of the lse. We start with the definition of the convex conjugate.

Definition A3 (Convex Conjugate)**.** *The* Convex Conjugate (Legendre-Fenchel transform) *of a function f from a Hilbert Space X to $[-\infty,\infty]$ is f^* which is defined as*

$$f^*(\boldsymbol{x}^*) = \sup_{\boldsymbol{x}\in X} (\boldsymbol{x}^T\boldsymbol{x}^* - f(\boldsymbol{x})), \quad \boldsymbol{x}^* \in X \quad \text{(A479)}$$

See page 219 Def. 13.1 in [10] and page 134 in [30]. Next we define the Legendre transform, which is a more restrictive version of the convex conjugate.

Definition A4 (Legendre Transform)**.** *The* Legendre transform *of a convex function f from a convex set $X \subset \mathbb{R}^n$ to $\mathbb{R}$ ($f: X \to \mathbb{R}$) is f^*, which is defined as*

$$f^*(\boldsymbol{x}^*) = \sup_{\boldsymbol{x}\in X} (\boldsymbol{x}^T\boldsymbol{x}^* - f(\boldsymbol{x})), \quad \boldsymbol{x}^* \in X^*, \quad \text{(A480)}$$

$$X^* = \left\{ \boldsymbol{x}^* \in \mathbb{R}^n \mid \sup_{\boldsymbol{x}\in X} (\boldsymbol{x}^T\boldsymbol{x}^* - f(\boldsymbol{x})) < \infty \right\}. \quad \text{(A481)}$$

See page 91 in [11].

Definition A5 (Epi-Sum)**.** *Let f and g be two functions from X to $(-\infty,\infty]$, then the infimal convolution (or epi-sum) of f and g is*

$$f \square g : X \to [-\infty,\infty], \ \boldsymbol{x} \mapsto \inf_{\boldsymbol{y}\in X} (f(\boldsymbol{y}) + g(\boldsymbol{x}-\boldsymbol{y})) \quad \text{(A482)}$$

See Def. 12.1 in [10].

Lemma A27. *Let f and g be functions from X to $(-\infty,\infty]$. Then the following hold:*

1. *Convex Conjugate of norm squared*

$$\left(\frac{1}{2}\|.\|^2\right)^* = \frac{1}{2}\|.\|^2. \quad \text{(A483)}$$

2. *Convex Conjugate of a function multiplied by scalar $0 < \alpha \in \mathbb{R}$*

$$(\alpha\, f)^* = \alpha\, f^*(./\alpha). \quad \text{(A484)}$$

3. *Convex Conjugate of the sum of a function and a scalar $\beta \in \mathbb{R}$*

$$(f + \beta)^* = f^* - \beta. \quad \text{(A485)}$$

4. *Convex Conjugate of affine transformation of the arguments. Let $\boldsymbol{A}$ be a non-singular matrix and $\boldsymbol{b}$ a vector*

$$(f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}))^* = f^*\left(\boldsymbol{A}^{-T}\boldsymbol{x}^*\right) - \boldsymbol{b}^T\boldsymbol{A}^{-T}\boldsymbol{x}^*. \quad \text{(A486)}$$

5. *Convex Conjugate of epi-sums*

$$(f \square g)^* = f^* + g^*. \quad \text{(A487)}$$

Proof. 1. Since $h(t) := \frac{t^2}{2}$ is a non-negative convex function and $h(t) = 0 \iff t = 0$ we have because of Proposition 11.3.3 in [30] that $h\left(\|x\|\right)^* = h^*\left(\|x^*\|\right)$. Additionally, by example (a) on page 137 we get for $1 < p < \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$ that $\left(\frac{|t|^p}{p}\right)^* = \frac{|t^*|^q}{q}$. Putting all together we get the desired result. The same result can also be deduced from page 222 Example 13.6 in [10].

2. Follows immediately from the definition since

$$\alpha f^*\left(\frac{\boldsymbol{x}^*}{\alpha}\right) = \alpha \sup_{\boldsymbol{x}\in X}\left(\boldsymbol{x}^T\frac{\boldsymbol{x}^*}{\alpha} - f(\boldsymbol{x})\right) = \sup_{\boldsymbol{x}\in X}(\boldsymbol{x}^T\boldsymbol{x}^* - \alpha f(\boldsymbol{x})) = (\alpha f)^*(\boldsymbol{x}^*)$$

3. $(f+\beta)^* := \sup_{\boldsymbol{x}\in X}\left(\boldsymbol{x}^T\boldsymbol{x}^* - f(\boldsymbol{x}) - \beta\right) =: f^* - \beta$

4.

$$\begin{aligned}
\left(f\left(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}\right)\right)^*\left(\boldsymbol{x}^*\right) &= \sup_{\boldsymbol{x}\in X}\left(\boldsymbol{x}^T\boldsymbol{x}^* - f\left(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}\right)\right)\\
&= \sup_{\boldsymbol{x}\in X}\left(\left(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}\right)^T\boldsymbol{A}^{-T}\boldsymbol{x}^* - f\left(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}\right)\right) - \boldsymbol{b}^T\boldsymbol{A}^{-T}\boldsymbol{x}^*\\
&= \sup_{\boldsymbol{y}\in X}\left(\boldsymbol{y}^T\boldsymbol{A}^{-T}\boldsymbol{x}^* - f\left(\boldsymbol{y}\right)\right) - \boldsymbol{b}^T\boldsymbol{A}^{-T}\boldsymbol{x}^*\\
&= f^*\left(\boldsymbol{A}^{-T}\boldsymbol{x}^*\right) - \boldsymbol{b}^T\boldsymbol{A}^{-T}\boldsymbol{x}^*
\end{aligned}$$

5. From Proposition 13.24 (i) in [10] and Proposition 11.4.2 in [30] we get

$$\begin{aligned}
\left(f\square g\right)^*\left(\boldsymbol{x}^*\right) &= \sup_{\boldsymbol{x}\in X}\left(\boldsymbol{x}^T\boldsymbol{x}^* - \inf_{\boldsymbol{y}\in X}\left(f(\boldsymbol{y}) - g(\boldsymbol{x}-\boldsymbol{y})\right)\right)\\
&= \sup_{\boldsymbol{x},\boldsymbol{y}\in X}\left(\boldsymbol{x}^T\boldsymbol{x}^* - f(\boldsymbol{y}) - g(\boldsymbol{x}-\boldsymbol{y})\right)\\
&= \sup_{\boldsymbol{x},\boldsymbol{y}\in X}\left(\left(\boldsymbol{y}^T\boldsymbol{x}^* - f(\boldsymbol{y})\right) + \left(\left(\boldsymbol{x}-\boldsymbol{y}\right)^T\boldsymbol{x}^* - g(\boldsymbol{x}-\boldsymbol{y})\right)\right)\\
&= f^*(\boldsymbol{x}^*) + g^*(\boldsymbol{x}^*)
\end{aligned}$$

□

Lemma A28. *The Legendre transform of the* lse *is the negative entropy function, restricted to the probability simplex and vice versa. For the log-sum exponential*

$$f(\boldsymbol{x}) \;=\; \ln\left(\sum_{i=1}^n \exp(x_i)\right)\;, \tag{A488}$$

the Legendre transform is the negative entropy function, restricted to the probability simplex:

$$f^*(\boldsymbol{x}^*) \;=\; \begin{cases}\sum_{i=1}^n x_i^*\ln(x_i^*) & \text{for } 0 \leqslant x_i^* \text{ and } \sum_{i=1}^n x_i^* = 1\\ \infty & \text{otherwise}\end{cases}\;. \tag{A489}$$

For the negative entropy function, restricted to the probability simplex:

$$f(\boldsymbol{x}) \;=\; \begin{cases}\sum_{i=1}^n x_i\ln(x_i) & \text{for } 0 \leqslant x_i \text{ and } \sum_{i=1}^n x_i = 1\\ \infty & \text{otherwise}\end{cases}\;. \tag{A490}$$

the Legendre transform is the log-sum exponential

$$f^*(\boldsymbol{x}^*) \;=\; \ln\left(\sum_{i=1}^n \exp(x_i^*)\right)\;, \tag{A491}$$

Proof. See page 93 Example 3.25 in [11] and [29]. If f is a regular convex function (lower semi-continuous convex function), then $f^{**} = f$ according to page 135 Exercise 11.2.3 in [30]. If f is lower semi-continuous and convex, then $f^{**} = f$ according to Theorem 13.37 (Fenchel-Moreau) in [10]. The log-sum-exponential is continuous and convex. □

Lemma A29. *Let $\boldsymbol{X}\boldsymbol{X}^T$ be non-singular and X a Hilbert space. We define*

$$X^* = \left\{ \boldsymbol{a} \mid 0 \leqslant \boldsymbol{X}^T \left(\boldsymbol{X}\boldsymbol{X}^T \right)^{-1} \boldsymbol{a} , \ \mathbf{1}^T \boldsymbol{X}^T \left(\boldsymbol{X}\boldsymbol{X}^T \right)^{-1} \boldsymbol{a} = 1 \right\} . \tag{A492}$$

and

$$X^v = \left\{ \boldsymbol{a} \mid \boldsymbol{a} = \boldsymbol{X}^T \boldsymbol{\xi} , \ \boldsymbol{\xi} \in X \right\} . \tag{A493}$$

The Legendre transform of $\mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi})$ *with* $\boldsymbol{\xi} \in X$ *is*

$$\left(\mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) \right)^* (\boldsymbol{\xi}^*) = (\mathrm{lse}(\beta, \boldsymbol{v}))^* \left(\boldsymbol{X}^T \left(\boldsymbol{X}\boldsymbol{X}^T \right)^{-1} \boldsymbol{\xi}^* \right) , \tag{A494}$$

with $\boldsymbol{\xi}^* \in X^*$ *and* $\boldsymbol{v} \in X^v$*. The domain of* $\left(\mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) \right)^*$ *is* X^**.*

Furthermore we have

$$\left(\mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) \right)^{**} = \mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) . \tag{A495}$$

Proof. We use the definition of the Legendre transform:

$$\begin{aligned}
\left(\mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) \right)^* (\boldsymbol{\xi}^*) &= \sup_{\boldsymbol{\xi} \in X} \boldsymbol{\xi}^T \boldsymbol{\xi}^* - \mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) \\
&= \sup_{\boldsymbol{\xi} \in X} \left(\boldsymbol{X}^T \boldsymbol{\xi} \right)^T \left(\boldsymbol{X}\boldsymbol{X}^T \right)^{-1} \boldsymbol{\xi}^* - \mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) \\
&= \sup_{\boldsymbol{v} \in X^v} \boldsymbol{v}^T \boldsymbol{X}^T \left(\boldsymbol{X}\boldsymbol{X}^T \right)^{-1} \boldsymbol{\xi}^* - \mathrm{lse}(\beta, \boldsymbol{v}) \\
&= \sup_{\boldsymbol{v} \in X^v} \boldsymbol{v}^T \boldsymbol{v}^* - \mathrm{lse}(\beta, \boldsymbol{v}) \\
&= (\mathrm{lse}(\beta, \boldsymbol{v}))^* (\boldsymbol{v}^*) = (\mathrm{lse}(\beta, \boldsymbol{v}))^* \left(\boldsymbol{X}^T \left(\boldsymbol{X}\boldsymbol{X}^T \right)^{-1} \boldsymbol{\xi}^* \right) ,
\end{aligned} \tag{A496}$$

where we used $\boldsymbol{v}^* = \boldsymbol{X}^T \left(\boldsymbol{X}\boldsymbol{X}^T \right)^{-1} \boldsymbol{\xi}^*$.

According to page 93 Example 3.25 in [11], the equations for the maximum $\max_{\boldsymbol{v} \in X^v} \boldsymbol{v}^T \boldsymbol{v}^* - \mathrm{lse}(\beta, \boldsymbol{v})$ are solvable if and only if $0 < \boldsymbol{v}^* = \boldsymbol{X}^T \left(\boldsymbol{X}\boldsymbol{X}^T \right)^{-1} \boldsymbol{\xi}^*$ and $\mathbf{1}^T \boldsymbol{v}^* = \mathbf{1}^T \boldsymbol{X}^T \left(\boldsymbol{X}\boldsymbol{X}^T \right)^{-1} \boldsymbol{\xi}^* = 1$. Therefore, we assumed $\boldsymbol{\xi}^* \in X^*$.

The domain of $\left(\mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) \right)^*$ is X^*, since on page 93 Example 3.25 in [11] it was shown that outside X^* the $\sup_{\boldsymbol{v} \in X^v} \boldsymbol{v}^T \boldsymbol{v}^* - \mathrm{lse}(\beta, \boldsymbol{v})$ is not bounded.

Using

$$\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) , \tag{A497}$$

the Hessian of $\mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi})$

$$\frac{\partial^2 \mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi})}{\partial \boldsymbol{\xi}^2} = \beta \, \boldsymbol{X} \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T \right) \boldsymbol{X}^T \tag{A498}$$

is positive semi-definite since $\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T$ is positive semi-definite according to Lemma A22. Therefore, $\mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi})$ is convex and continuous.

If f is a regular convex function (lower semi-continuous convex function), then $f^{**} = f$ according to page 135 Exercise 11.2.3 in [30]. If f is lower semi-continuous and convex, then $f^{**} = f$ according to Theorem 13.37 (Fenchel-Moreau) in [10]. Consequently we have

$$\left(\mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) \right)^{**} = \mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) . \tag{A499}$$

□

We introduce the Lambert W function and some of its properties, since it is needed to derive bounds on the storage capacity of our new Hopfield networks.

Definition A6 (Lambert Function)**.** *The* Lambert W function *is the inverse function of*

$$f(y) \;=\; y e^{y} \,. \tag{A500}$$

The Lambert W function has an upper branch W_0 for $-1 \leqslant y$ and a lower branch W_{-1} for $y \leqslant -1$. We use W if a formula holds for both branches. We have

$$W(x) \;=\; y \;\Rightarrow y e^{y} \;=\; x \,. \tag{A501}$$

We present some identities for the Lambert W function:

Lemma A30. *Identities for the Lambert W function are*

$$W(x)\, e^{W(x)} \;=\; x \,, \tag{A502}$$

$$W(x e^{x}) \;=\; x \,, \tag{A503}$$

$$e^{W(x)} \;=\; \frac{x}{W(x)} \,, \tag{A504}$$

$$e^{-W(x)} \;=\; \frac{W(x)}{x} \,, \tag{A505}$$

$$e^{nW(x)} \;=\; \left(\frac{x}{W(x)}\right)^{n} \,, \tag{A506}$$

$$W_0\left(x\, \ln x\right) \;=\; \ln x \quad \textit{for}\;\; x \;\geq\; \frac{1}{e} \,, \tag{A507}$$

$$W_{-1}\left(x\, \ln x\right) \;=\; \ln x \quad \textit{for}\;\; x \;\leqslant\; \frac{1}{e} \,, \tag{A508}$$

$$W(x) \;=\; \ln \frac{x}{W(x)} \quad \textit{for}\;\; x \;\geq\; -\frac{1}{e} \,, \tag{A509}$$

$$W\left(\frac{n\, x^{n}}{W\left(x\right)^{n-1}}\right) \;=\; n\, W(x) \quad \textit{for}\;\; n, x \;>\; 0 \,, \tag{A510}$$

$$W(x) \;+\; W(y) \;=\; W\left(x\, y\, \left(\frac{1}{W(x)} \;+\; \frac{1}{W(y)}\right)\right) \quad \textit{for}\;\; x, y \;>\; 0 \,, \tag{A511}$$

$$W_0\left(-\,\frac{\ln x}{x}\right) \;=\; -\,\ln x \quad \textit{for}\;\; 0 \;<\; x \;\leqslant\; e \,, \tag{A512}$$

$$W_{-1}\left(-\,\frac{\ln x}{x}\right) \;=\; -\,\ln x \quad \textit{for}\;\; x \;>\; e \,, \tag{A513}$$

$$e^{-\,W(-\,\ln x)} \;=\; \frac{W(-\,\ln x)}{-\,\ln x} \quad \textit{for}\;\; x \;\neq\; 1 \,. \tag{A514}$$

We also present some special values for the Lambert W function:

Lemma A31.

$$W(0) \;=\; 0 \,, \tag{A515}$$

$$W(e) \;=\; 1 \,, \tag{A516}$$

$$W\left(-\frac{1}{e}\right) \;=\; -1 \,, \tag{A517}$$

$$W\left(e^{1+e}\right) \;=\; e \,, \tag{A518}$$

$$W\left(2 \ln 2\right) \;=\; \ln 2 \,, \tag{A519}$$

$$W(1) \;=\; \Omega \,, \tag{A520}$$

$$W(1) \;=\; e^{-W(1)} \;=\; \ln\left(\frac{1}{W(1)}\right) \;=\; -\,\ln W(1) \,, \tag{A521}$$

$$W\left(-\frac{\pi}{2}\right) \;=\; \frac{i\pi}{2} \,, \tag{A522}$$

$$W(-1) \;\approx\; -0.31813 + 1.33723 i \,, \tag{A523}$$

where the Omega constant Ω *is*

$$\Omega \;=\; \left(\int_{-\infty}^{\infty} \frac{\mathrm{d}t}{\left(e^t \;-\; t\right)^2 \;+\; \pi^2} \right)^{-1} \;-\; 1 \;\approx\; 0.56714329\,. \tag{A524}$$

We need in some proofs a version of the mean value theorem as given in the next lemma.

Lemma A32 (Mean Value Theorem)**.** *Let* $U \subset \mathbb{R}^n$ *be open,* $f : U \to \mathbb{R}^m$ *continuously differentiable, and* $\boldsymbol{x} \in U$ *as well as* $\boldsymbol{h} \in \mathbb{R}^n$ *vectors such that the line segment* $\boldsymbol{x} + t\boldsymbol{h}$ *for* $0 \leqslant t \leqslant 1$ *is in* U*. Then the following holds:*

$$f(\boldsymbol{x} \;+\; \boldsymbol{h}) \;-\; f(\boldsymbol{x}) \;=\; \left(\int_0^1 J(\boldsymbol{x} \;+\; t\,\boldsymbol{h})\,\mathrm{d}t \right) \boldsymbol{h}\,, \tag{A525}$$

where J *is the Jacobian of* f *and the integral of the matrix is component-wise.*

Proof. Let $f_1, \ldots, f_m$ denote the components of f and define $g_i : [0,1] \to \mathbb{R}$ by

$$g_i(t) \;=\; f_i(\boldsymbol{x} \;+\; t\,\boldsymbol{h})\,, \tag{A526}$$

then we obtain

$$f_i(\boldsymbol{x} \;+\; \boldsymbol{h}) \;-\; f_i(\boldsymbol{x}) \;=\; g_i(1) \;-\; g_i(0) \;=\; \int_0^1 g'(t)\,\mathrm{d}t \tag{A527}$$

$$\int_0^1 \left(\sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j}(\boldsymbol{x} \;+\; t\,\boldsymbol{h})\,h_j \right) \mathrm{d}t \;=\; \sum_{j=1}^{n} \left(\int_0^1 \frac{\partial f_i}{\partial x_j}(\boldsymbol{x} \;+\; t\,\boldsymbol{h})\,\mathrm{d}t \right) h_j\,.$$

The statement follows since the Jacobian J has as entries $\frac{\partial f_i}{\partial x_j}$. □

A4 Modern Hopfield Networks: Binary States (Krotov and Hopfield)

A4.1 Modern Hopfield Networks: Introduction

A4.1.1 Additional Memory and Attention for Neural Networks

Modern Hopfield networks may serve as additional memory for neural networks. Different approaches have been suggested to equip neural networks with an additional memory beyond recurrent connections. The neural Turing machine (NTM) is a neural network equipped with an external memory and an attention process [31]. The NTM can write to the memory and can read from it. A memory network [69] consists of a memory together with the components: (1) input feature map (converts the incoming input to the internal feature representation) (2) generalization (updates old memories given the new input), (3) output feature map (produces a new output), (4) response (converts the output into the response format). Memory networks are generalized to an end-to-end trained model, where the $\arg\max$ memory call is replaced by a differentiable $\mathrm{softmax}$ [58, 59]. Linear Memory Network use a linear autoencoder for sequences as a memory [16].

To enhance RNNs with additional associative memory like Hopfield networks have been proposed [5, 6]. The associative memory stores hidden states of the RNN, retrieves stored states if they are similar to actual ones, and has a forgetting parameter. The forgetting and storing parameters of the RNN associative memory have been generalized to learned matrices [79]. LSTMs with associative memory via Holographic Reduced Representations have been proposed [20].

Recently most approaches to new memories are based on attention. The neural Turing machine (NTM) is equipped with an external memory and an attention process [31]. End to end memory networks (EMN) make the attention scheme of memory networks [69] differentiable by replacing $\arg\max$ through a $\mathrm{softmax}$ [58, 59]. EMN with dot products became very popular and implement a key-value attention [21] for self-attention. An enhancement of EMN is the transformer [64, 65] and its extensions [22]. The transformer had great impact on the natural language processing (NLP) community as new records in NLP benchmarks have been achieved [64, 65]. MEMO uses the transformer attention mechanism for reasoning over longer distances [8]. Current state-of-the-art for language processing is a transformer architecture called "the Bidirectional Encoder Representations from Transformers" (BERT) [24, 25].

A4.1.2 Modern Hopfield networks: Overview

The storage capacity of classical binary Hopfield networks [37] has been shown to be very limited. In a d-dimensional space, the standard Hopfield model can store d uncorrelated patterns without errors but only $Cd/\ln(d)$ random patterns with $C < 1/2$ for a fixed stable pattern or $C < 1/4$ if all patterns are stable [45]. The same bound holds for nonlinear learning rules [44]. Using tricks-of-trade and allowing small retrieval errors, the storage capacity is about $0.138d$ [19, 33, 63]. If the learning rule is not related to the Hebb rule then up to d patterns can be stored [1]. Using Hopfield networks with non-zero diagonal matrices, the storage can be increased to $Cd\ln(d)$ [28]. In contrast to the storage capacity, the number of energy minima (spurious states, stable states) of Hopfield networks is exponentially in d [61, 13, 66].

Recent advances in the field of binary Hopfield networks [37] led to new properties of Hopfield networks. The stability of spurious states or metastable states was sensibly reduced by a Hamiltonian treatment for the new relativistic Hopfield model [9]. Recently the storage capacity of Hopfield networks could be increased by new energy functions. Interaction functions of the form $F(x) = x^n$ lead to storage capacity of $\alpha_n d^{n-1}$, where α_n depends on the allowed error probability [41, 42, 23] (see [42] for the non-binary case). Interaction functions of the form $F(x) = x^n$ lead to storage capacity of $\alpha_n \frac{d^{n-1}}{c_n \ln d}$ for $c_n > 2(2n-3)!!$ [23].

Interaction functions of the form $F(x) = \exp(x)$ lead to *exponential* storage capacity of $2^{d/2}$ where all stored patterns are fixed points but the radius of attraction vanishes [23]. It has been shown that the network converges even after one update [23].

A4.2 Energy and Update Rule for Binary Modern Hopfield Networks

We follow [23] where the goal is to store a set of input data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ that are represented by the matrix

$$\boldsymbol{X} \ = \ (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \ . \tag{A528}$$

The $\boldsymbol{x}_i$ is pattern with binary components $x_{ij} \in \{-1, +1\}$ for all i and j. $\boldsymbol{\xi}$ is the actual state of the units of the Hopfield model. Krotov and Hopfield [41] defined the energy function E with the interaction function F that evaluates the dot product between patterns $\boldsymbol{x}_i$ and the actual state $\boldsymbol{\xi}$:

$$\mathrm{E} \ = \ - \sum_{i=1}^{N} F\left(\boldsymbol{\xi}^T \boldsymbol{x}_i\right) \tag{A529}$$

with $F(a) = a^n$, where $n = 2$ gives the energy function of the classical Hopfield network. This allows to store $\alpha_n d^{n-1}$ patterns [41]. Krotov and Hopfield [41] suggested for minimizing this energy an asynchronous updating dynamics $T = (T_j)$ for component ξ_j:

$$T_j(\boldsymbol{\xi}) \ := \ \mathrm{sgn}\left[\sum_{i=1}^{N}\left(F\left(x_{ij} \ + \ \sum_{l \neq j} x_{il} \ \xi_l\right) \ - \ F\left(- \ x_{ij} \ + \ \sum_{l \neq j} x_{il} \ \xi_l\right)\right)\right] \tag{A530}$$

While Krotov and Hopfield used $F(a) = a^n$, Demircigil et al. [23] went a step further and analyzed the model with the energy function $F(a) = \exp(a)$, which leads to an exponential storage capacity of $N = 2^{d/2}$. Furthermore with a single update the final pattern is recovered with high probability. These statements are given in next theorem.

Theorem A10 (Storage Capacity for Binary Modern Hopfield Nets (Demircigil et al. 2017)). *Consider the generalized Hopfield model with the dynamics described in Eq. (A530) and interaction function F given by $F(x) = e^x$. For a fixed $0 < \alpha < \ln(2)/2$ let $N = \exp(\alpha d) + 1$ and let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ be N patterns chosen uniformly at random from $\{-1, +1\}^d$. Moreover fix $\varrho \in [0, 1/2)$. For any i and any $\widetilde{\boldsymbol{x}}_i$ taken uniformly at random from the Hamming sphere with radius ϱd centered in $\boldsymbol{x}_i$, $\mathcal{S}(\boldsymbol{x}_i, \varrho d)$, where ϱd is assumed to be an integer, it holds that*

$$\Pr\left(\exists i \ \exists j : \ T_j\left(\widetilde{\boldsymbol{x}}_i\right) \ \neq \ x_{ij}\right) \ \to \ 0 \ ,$$

if α is chosen in dependence of ϱ such that

$$\alpha \ < \ \frac{I(1 - 2\varrho)}{2}$$

with

$$I : \ a \ \mapsto \ \frac{1}{2}\left((1 + a)\ln(1 + a) \ + \ (1 - a)\ln(1 - a)\right) \ .$$

Proof. The proof can be found in [23]. □

The number of patterns $N = \exp(\alpha d) + 1$ is exponential in the number d of components. The result

$$\Pr\left(\exists i \ \exists j : \ T_j\left(\widetilde{\boldsymbol{x}}_i\right) \ \neq \ x_{ij}\right) \ \to \ 0$$

means that one update for each component is sufficient to recover the pattern with high probability. The constraint $\alpha < \frac{I(1-2\varrho)}{2}$ on α gives the trade-off between the radius of attraction ϱd and the number $N = \exp(\alpha d) + 1$ of pattern that can be stored.

Theorem A10 in particular implies that

$$\Pr\left(\exists i \ \exists j : \ T_j\left(\boldsymbol{x}_i\right) \ \neq \ x_{ij}\right) \ \to \ 0$$

as $d \to \infty$, i.e. with a probability converging to 1, all the patterns are fixed points of the dynamics. In this case we can have $\alpha \to \frac{I(1)}{2} = \ln(2)/2$.

Krotov and Hopfield define the update dynamics $T_j(\boldsymbol{\xi})$ in Eq. (A530) via energy differences of the energy in Eq. (A529). First we express the energy in Eq. (A529) with $F(a) = \exp(a)$ [23] by the lse

function. Then we use the mean value theorem to express the update dynamics $T_j(\boldsymbol{\xi})$ in Eq. (A530) by the softmax function. For simplicity, we set $\beta = 1$ in the following. There exists a $v \in [-1, 1]$ with

$$T_j(\boldsymbol{\xi}) = \operatorname{sgn}\left[\mathrm{E}(\xi_j = 1) - \mathrm{E}(\xi_j = -1)\right] = \operatorname{sgn}\left[-\exp(\operatorname{lse}(\xi_j = 1)) + \exp(\operatorname{lse}(\xi_j = -1))\right] \tag{A531}$$

$$= \operatorname{sgn}\left[(2\boldsymbol{e}_j)^T \nabla_{\boldsymbol{\xi}} \mathrm{E}(\xi_j = v)\right] = \operatorname{sgn}\left[\exp(\operatorname{lse}(\xi_j = v))\, (2\boldsymbol{e}_j)^T \frac{\operatorname{lse}(\xi_j = v)}{\partial \boldsymbol{\xi}}\right]$$

$$= \operatorname{sgn}\left[\exp(\operatorname{lse}(\xi_j = 1))\, (2\boldsymbol{e}_j)^T \boldsymbol{X} \operatorname{softmax}(\boldsymbol{X}^T \boldsymbol{\xi}(\xi_j = v))\right]$$

$$= \operatorname{sgn}\left[[\boldsymbol{X} \operatorname{softmax}(\boldsymbol{X}^T \boldsymbol{\xi}(\xi_j = v))]_j\right] = \operatorname{sgn}\left[[\boldsymbol{X} \boldsymbol{p}(\xi_j = v)]_j\right],$$

where $\boldsymbol{e}_j$ is the Cartesian unit vector with a one at position j and zeros elsewhere, $[.]_j$ is the projection to the j-th component, and

$$\boldsymbol{p} = \operatorname{softmax}(\boldsymbol{X}^T \boldsymbol{\xi}) . \tag{A532}$$

A5 Hopfield Update Rule is Attention of The Transformer

The Hopfield network update rule is the attention mechanism used in the transformer and BERT (see Fig. A2). To see this, we assume patterns $\boldsymbol{y}_i$ that are mapped to the Hopfield space of dimension d_k. We set $\boldsymbol{x}_i = \boldsymbol{W}_K^T \boldsymbol{y}_i$, $\boldsymbol{\xi}_i = \boldsymbol{W}_Q^T \boldsymbol{y}_i$, and multiply the result of our update rule with $\boldsymbol{W}_V$. The matrix $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)^T$ combines the $\boldsymbol{y}_i$ as row vectors. We define the matrices $\boldsymbol{X}^T = \boldsymbol{K} = \boldsymbol{Y}\boldsymbol{W}_K$, $\boldsymbol{Q} = \boldsymbol{Y}\boldsymbol{W}_Q$, and $\boldsymbol{V} = \boldsymbol{Y}\boldsymbol{W}_K\boldsymbol{W}_V = \boldsymbol{X}^T\boldsymbol{W}_V$, where $\boldsymbol{W}_K \in \mathbb{R}^{d_y \times d_k}$, $\boldsymbol{W}_Q \in \mathbb{R}^{d_y \times d_k}$, $\boldsymbol{W}_V \in \mathbb{R}^{d_k \times d_v}$. For combining all queries in matrix $\boldsymbol{Q}$, $\beta = 1/\sqrt{d_k}$, and $\operatorname{softmax} \in \mathbb{R}^N$ changed to a row vector, we obtain for the update rule Eq. (A17) multiplied by $\boldsymbol{W}_V$:

$$\operatorname{softmax}\left(1/\sqrt{d_k}\, \boldsymbol{Q}\, \boldsymbol{K}^T\right) \boldsymbol{V} . \tag{A533}$$

This formula is the transformer attention.

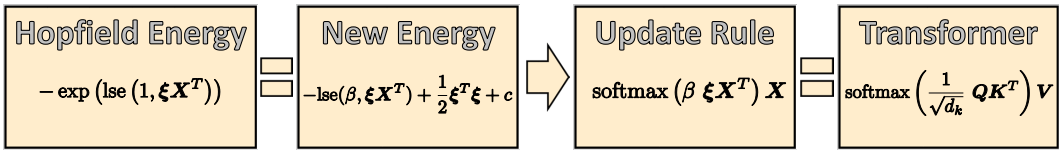

Figure A2: We generalized the energy of binary modern Hopfield networks for allowing continuous states while keeping convergence and storage capacity properties. We defined for the new energy also a new update rule that minimizes the energy. The new update rule is the attention mechanism of the transformer. Formulae are modified to express softmax as row vector as for transformers. "="-sign means "keeps the properties".

Experiments

Hubert Ramsauer Bernhard Schäfl Johannes Lehner Philipp Seidl Michael Widrich Günter Klambauer Johannes Brandstetter Sepp Hochreiter

B1 Experiment 1: Attention in Transformers described by Hopfield dynamics

B1.1 Experimental Setup

Transformer architectures are known for their high computational demands. To investigate the learning dynamics of such a model and at the same time keeping training time manageable, we adopted the BERT-small setting from ELECTRA [18]. It has 12 layers, 4 heads and a reduced hidden size, the sequence length is shortened from 512 to 128 tokens and the batch size is reduced from 256 to 128. Additionally, the hidden dimension is reduced from 768 to 256 and the embedding dimension is reduced from 768 to 128 [18]. The training of such a BERT-small model for 1.45 million update steps takes roughly four days on a single NVIDIA V100 GPU.

As the code base we use the *transformers* repository from Hugging Face, Inc [71]. We aim to reproduce the dataset of [25] as close as possible, which consists of the English Wikipedia dataset and the Toronto BookCorpus dataset [80]. Due to recent copyright claims the later is not publicly available anymore. Therefore, the pre-training experiments use an uncased snapshot of the original BookCorpus dataset.

B1.2 Hopfield Operating Classes of Transformer and BERT Models

To better understand how operation modes in attention heads develop, we tracked the distribution of counts k (see main paper) over time in a BERT-small model. At the end of training we visualized the count distribution, grouped into four classes (see Figure B1). The thresholds for the classes were chosen according to the thresholds of Figure 2 in the main paper. However, they are divided by a factor of 4 to adapt to the shorter sequence length of 128 compared to 512. From this plot it is clear, that the attention in heads of **Class IV** commit very early to the operating class of small metastable states.

B1.3 Learning Dynamics of Transformer and BERT Models

To observe this behavior in the early phase of training, we created a ridge plot of the distributions of counts k for the first $20,000$ steps (see Figure B2 (a)). This plot shows that the attention in heads of middle layers often change the operation mode to **Class IV** around $9,000$ to $10,000$ steps. At the same time the second big drop in the loss occurs. The question arises whether this is functionally important or whether it is an artefact which could be even harmful. To check if the attention mechanism is still able to learn after the change in the operation mode we analyzed the gradient flow through the $\mathrm{softmax}$ function. For every token we calculate the Frobenius norm of the Jacobian of the $\mathrm{softmax}$ over multiple samples. Then, for every head we plot the distribution of the norm (see

Figure B2(b)). The gradients with respect to the weights are determined by the Jacobian J defined in Eq. (A48) as can be seen in Eq. (A409), Eq. (A420), and Eq. (A426). We can see that the attention in heads of **Class IV** remain almost unchanged during the rest of the training.

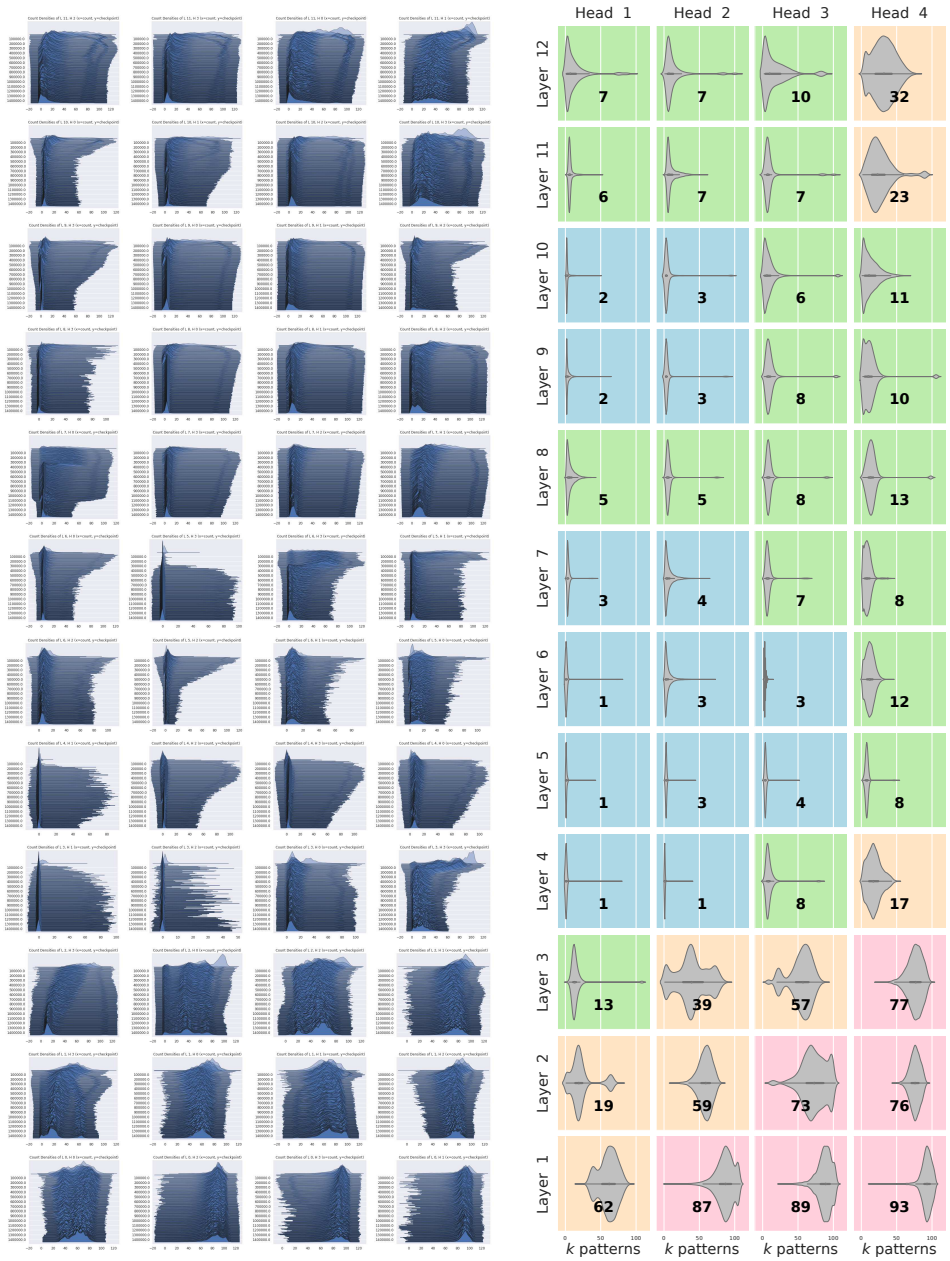

Figure B1: **Left**: Ridge plots of the distribution of counts k over time for BERT-small **Right**: Violin plot of counts k after $1,450000$ steps, divided into the four classes from the main paper. The thresholds were adapted to the shorter sequence length.

B1.4 Attention Heads Replaced by Gaussian Averaging Layers

The self-attention mechanism proposed in [64] utilizes the $\mathrm{softmax}$ function to compute the coefficients of a convex combination over the embedded tokens, where the $\mathrm{softmax}$ is conditioned on the input. However, our analysis showed that especially in lower layers many heads perform averaging over a very large number of patterns. This suggests that at this level neither the dependency on the input nor a fine grained attention to individual positions is necessary. As an alternative to the

(a) Densities (b) Norm of Jacobian

Figure B2: **(a)**: change of count density during training is depicted for the first $20,000$ steps. **(b)**: the corresponding distribution of the Frobenius norm of the Jacobian of the $\mathrm{softmax}$ function is depicted. The gradients with respect to the weights are determined by the Jacobian J defined in Eq. (A48) as can be seen in Eq. (A409), Eq. (A420), and Eq. (A426).

original mechanism we propose Gaussian averaging heads which are computationally more efficient. Here, the $\mathrm{softmax}$ function is replaced by a discrete Gaussian kernel, where the location μ and the scale σ are learned. In detail, for a sequence length of N tokens we are given a vector of location parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)^T$ and a vector of corresponding scale parameters $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)^T$. We subdivide the interval $[-1, 1]$ into N equidistant supporting points $\{s_j\}_{j=1}^N$, where

$$s_j = \frac{(j-1) - 0.5\,(N-1)}{0.5\,(N-1)}.$$

The attention $[A]_{i,j}$ from the i-th token to the j-th position is calculated as

$$[A]_{i,j} = \frac{1}{z_i} \exp\left\{-\frac{1}{2}\left(\frac{s_j - \mu_i}{\sigma_i}\right)^2\right\},$$

where z_i normalizes the i-th row of the attention matrix A to sum up to one:

$$z_i = \sum_{j=1}^{N} \exp\left\{-\frac{1}{2}\left(\frac{s_j - \mu_i}{\sigma_i}\right)^2\right\}.$$

For initialization we uniformly sample a location vector $\boldsymbol{\mu} \in [-1, 1]^N$ and a scale vector $\boldsymbol{\sigma} \in [0.75, 1.25]^N$ per head. A simple way to consider the individual position of each token at initialization is to use the supporting points $\mu_i = s_i$ (see Figure B3). In practice no difference to the random initialization was observed.

Number of parameters. Gaussian averaging heads can reduce the number of parameters significantly. For an input size of N tokens, there are $2 \cdot N$ parameters per head. In contrast, a standard self-attention head with word embedding dimension d_y and projection dimension d_k has two weight matrices $W_Q, W_K \in \mathbb{R}^{d_k \times d_y}$, which together amount to $2 \cdot d_k \cdot d_y$ parameters. As a concrete example, the BERT-base model from [25] has an embedding dimension $d_y = 768$, a projection dimension $d_k = 64$ and a sequence length of $N = 512$. Compared to the Gaussian head, in this case $(2 \cdot 768 \cdot 64)/(2 \cdot 512) = 95.5$ times more parameters are trained for the attention mechanism itself. Only for very long sequences (and given that the word embedding dimension stays the same) the dependence on N may become a disadvantage. But of course, due to the independence from the input the Gaussian averaging head is less expressive in comparison to the original attention mechanism. A recently proposed input independent replacement for self-attention is the so called Random Synthesizer [62]. Here the $\mathrm{softmax}$-attention is directly parametrized with an $N \times N$ matrix. This amounts to $0.5 \cdot N$ more parameters than Gaussian averaging.

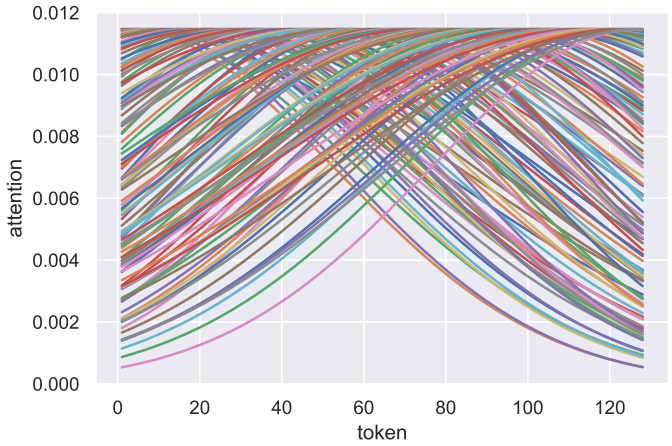

Figure B3: Attentions of a Gaussian averaging head at initialization for sequence length $N = 128$. Every line depicts one Gaussian kernel. Here, the location parameters are initialized with the value of the supporting points $\mu_i = s_i$.

PyTorch Implementation of a Hopfield Layer

Bernhard Schäfl Michael Widrich Hubert Ramsauer Johannes Lehner
Sepp Hochreiter Günter Klambauer Johannes Brandstetter

C1 Introduction

In this section, we describe the implementation of a Hopfield layer in PyTorch [49, 50] and, additionally, provide a brief usage manual. Possible applications for a Hopfield layer in a deep network architecture comprise:

- multiple instance learning (MIL) [26],
- processing of and learning with point sets [51, 52, 73],
- set-based and permutation invariant learning [32, 55, 76, 40, 38, 78],
- attention-based learning [64],
- associative learning,
- natural language processing,
- sequence analysis and time series prediction, and
- storing and retrieving reference or experienced data, e.g. to store training data and retrieve it by the model or to store experiences for reinforcement learning.

The Hopfield layer in a deep neural network architecture can implement:

- a memory (storage) with associative retrieval [20, 5],
- conditional pooling and averaging operations [67, 39],
- combining data by associations [2],
- associative credit assignment (e.g. Rescorla-Wagner model or value estimation) [60], and
- attention mechanisms [64, 7].

In particular, a Hopfield layer can substitute attention layers in architectures of transformer and BERT models. The Hopfield layer is designed to be used as plug-in replacement for existing layers like

- pooling layers (max-pooling or average pooling),
- permutation equivariant layers [32, 55],
- GRU & LSTM layers, and
- attention layers.

In contrast to classical Hopfield networks, the Hopfield layer is based on the modern Hopfield networks with continuous states that have increased storage capacity, as discussed in the main paper. Like classical Hopfield networks, the dynamics of the single heads of a Hopfield layer follow a energy minimization dynamics. The energy minimization empowers our Hopfield layer with several advantages over other architectural designs like memory cells, associative memory, or attention mechanisms. For example, the Hopfield layer has more functionality than a transformer self-attention layer [64] as described in Sec. C2. Possible use cases are given in Sec. C3.

Source code is provided at https://github.com/ml-jku/hopfield-layers.

C2 Functionality

Non-standard functionalities that are added by a Hopfield layer are

- *Association of two sets*,
- *Variable Beta* that determines the kind of fixed points,
- *Multiple Updates* for precise fixed points,
- *Dimension of the associative space* for controlling the storage capacity,
- *Static Patterns* for fixed pattern search, and
- *Pattern Normalization* to control the fixed point dynamics by norm of the patterns and shift of the patterns.

A functional sketch of our Hopfield layer is shown in Fig. C1.

Association of two sets. The Hopfield layer makes it possible to associate two sets of vectors. This general functionality allows

- for transformer-like self-attention,
- for decoder-encoder attention,
- for time series prediction (maybe with positional encoding),
- for sequence analysis,
- for multiple instance learning,
- for learning with point sets,
- for combining data sources by associations,
- for constructing a memory,
- for averaging and pooling operations, and
- for many more.

The first set of vectors consists of N *raw state patterns* $\boldsymbol{R} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N)^T$ with $\boldsymbol{r}_n \in \mathbb{R}^{d_r}$ and the second set of vectors consists of S *raw stored patterns* $\boldsymbol{Y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_S)^T$ with $\boldsymbol{y}_s \in \mathbb{R}^{d_y}$. Both the N raw state patterns and S raw stored patterns are mapped to an associative space in $\mathbb{R}^{d_k}$ via the matrices $\boldsymbol{W}_Q \in \mathbb{R}^{d_r \times d_k}$ and $\boldsymbol{W}_K \in \mathbb{R}^{d_y \times d_k}$, respectively. We define a matrix $\boldsymbol{Q}$ ($\boldsymbol{\Xi}^T$) of *state patterns* $\boldsymbol{\xi}_n = \boldsymbol{W}_Q \boldsymbol{r}_n$ in an associative space $\mathbb{R}^{d_k}$ and a matrix $\boldsymbol{K}$ ($\boldsymbol{X}^T$) of *stored patterns* $\boldsymbol{x}_i = \boldsymbol{W}_K \boldsymbol{y}_s$ in the associative space $\mathbb{R}^{d_k}$:

$$\boldsymbol{Q} = \boldsymbol{\Xi}^T = \boldsymbol{R}\, \boldsymbol{W}_Q \,, \tag{C1}$$

$$\boldsymbol{K} = \boldsymbol{X}^T = \boldsymbol{Y}\, \boldsymbol{W}_K \,. \tag{C2}$$

In the main paper, Eq. (3) defines the novel update rule:

$$\boldsymbol{\xi}^{\text{new}} = f(\boldsymbol{\xi}) = \boldsymbol{X}\, \text{softmax}(\beta\, \boldsymbol{X}^T \boldsymbol{\xi}) \,, \tag{C3}$$

For multiple patterns, Eq. (3) becomes:

$$\boldsymbol{\Xi}^{\text{new}} = f(\boldsymbol{\Xi}) = \boldsymbol{X}\, \text{softmax}(\beta\, \boldsymbol{X}^T \boldsymbol{\Xi}) \,, \tag{C4}$$

where $\boldsymbol{\Xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N)$ is the matrix of N state (query) patterns, $\boldsymbol{X}$ is the matrix of stored (key) patterns, and $\boldsymbol{\Xi}^{\text{new}}$ is the matrix of new state patterns, which are averages over stored patterns. A new state pattern can also be very similar to a single stored pattern, in which case we call the stored pattern to be retrieved.

These matrices allow to rewrite Eq. (C4) as:

$$(\boldsymbol{Q}^{\text{new}})^T \ = \ \boldsymbol{K}^T \text{softmax}(\beta \ \boldsymbol{K} \ \boldsymbol{Q}^T) \ . \tag{C5}$$

For $\beta = 1/\sqrt{d_k}$ and changing in Eq. (C5) $\text{softmax} \in \mathbb{R}^N$ to a row vector (and evaluating a row vector), we obtain:

$$\boldsymbol{Q}^{\text{new}} \ = \ \text{softmax}(1/\sqrt{d_k} \ \boldsymbol{Q} \ \boldsymbol{K}^T) \ \boldsymbol{K} \ , \tag{C6}$$

where $\boldsymbol{Q}^{\text{new}}$ is again the matrix of new state patterns. The new state patterns $\boldsymbol{\Xi}^{\text{new}}$ are projected via $\boldsymbol{W}_V$ to the result patterns $\boldsymbol{Z} = \boldsymbol{\Xi}^{\text{new}} \boldsymbol{W}_V$, where $\boldsymbol{W}_V \in \mathbb{R}^{d_k \times d_v}$. With the pattern projection $\boldsymbol{V} = \boldsymbol{K} \boldsymbol{W}_V$, we obtain the update rule Eq. (10) from the main paper:

$$\boldsymbol{Z} \ = \ \text{softmax}(1/\sqrt{d_k} \ \boldsymbol{Q} \ \boldsymbol{K}^T) \ \boldsymbol{V} \ . \tag{C7}$$

Multiple Updates. The update Eq. (C5) can be iteratively applied to the initial state $\boldsymbol{\xi}$ of every Hopfield layer head. After the last update, the new states $\boldsymbol{\Xi}^{\text{new}}$ are projected via $\boldsymbol{W}_V$ to the result patterns $\boldsymbol{Z} = \boldsymbol{\Xi}^{\text{new}} \boldsymbol{W}_V$. Therefore, the Hopfield layer allows multiple update steps in the forward pass without changing the number of parameters. The number of update steps can be given for every Hopfield head individually. Furthermore, it is possible to set a threshold for the number of updates of every Hopfield head based on $\| \boldsymbol{\xi} - \boldsymbol{\xi}^{\text{new}} \|_2$. In the general case of multiple initial states $\boldsymbol{\Xi}$, the maximum over the individual norms is taken.

Variable Beta. In the main paper, we have identified β as a crucial parameter for the fixed point dynamics of the Hopfield network, which governs the operating mode of the attention heads. In appendix, e.g. in Lemma A7 or in Eq. (A91) and Eq. (A92), we showed that the characteristics of the fixed points of the new modern Hopfield network are determined by: β, M (maximal pattern norm), $m_{\max}$ (spread of the similar patterns), and $\| \boldsymbol{m}_{\boldsymbol{x}} \|$ (center of the similar patterns). Low values of β induce global averaging and higher values of β metastable states. In the transformer attention, the β parameter is set to $\beta = 1/\sqrt{d_k}$ as in Eq. (C7). The Hopfield layer, however, allows to freely choose $\beta > 0$, since the fixed point dynamics does not only depend on the dimension of the associative space d_k. Additionally, β heavily influences the gradient flow to the matrices $\boldsymbol{W}_Q$ and $\boldsymbol{W}_K$. Thus, finding the right β for the respective application can be crucial.

Variable dimension of the associative space. Theorem A5 says that the storage capacity of the modern Hopfield network grows exponentially with the dimension of the associative space. However higher dimension of the associative space also means less averaging and smaller metastable states. The dimension of the associative space trades off storage capacity against the size of metastable states, e.g. over how many pattern is averaged. In Eq. (C2) and in Eq. (C1), we assumed N raw state patterns $\boldsymbol{R} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N)^T$ and S raw stored patterns $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_S)^T$ that are mapped to a d_k-dimensional associative space via the matrices $\boldsymbol{W}_Q \in \mathbb{R}^{d_r \times d_k}$ and $\boldsymbol{W}_K \in \mathbb{R}^{d_y \times d_k}$, respectively. In the associative space $\mathbb{R}^{d_k}$, we obtain the state patterns $\boldsymbol{Q} = \boldsymbol{\Xi}^T = \boldsymbol{R} \boldsymbol{W}_Q$ and the stored patterns $\boldsymbol{K} = \boldsymbol{X}^T = \boldsymbol{Y} \ \boldsymbol{W}_K$. The Hopfield view relates the dimension d_k to the number of input patterns N that have to be processed. The storage capacity depends exponentially on the dimension d_k (the dimension of the associative space) and the size to metastable states is governed by this dimension, too. Consequently, d_k should be chosen with respect to the number N of patterns one wants to store and the desired size of metastable states, which is the number of patterns one wants to average over. For example, if the input consists of many low dimensional input patterns, it makes sense to project the patterns into a higher dimensional space to allow a proper fixed point dynamics. Intuitively, this coincides with the construction of a richer feature space for the patterns.

Static Patterns. In Eq. (C2) and Eq. (C1), the N raw state patterns $\boldsymbol{R} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N)^T$ and S raw stored patterns $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_S)^T$ are mapped to an associative space via the matrices $\boldsymbol{W}_Q \in \mathbb{R}^{d_r \times d_k}$ and $\boldsymbol{W}_K \in \mathbb{R}^{d_y \times d_k}$, which gives the state patterns $\boldsymbol{Q} = \boldsymbol{\Xi}^T = \boldsymbol{R} \boldsymbol{W}_Q$ and the stored patterns $\boldsymbol{K} = \boldsymbol{X}^T = \boldsymbol{Y} \ \boldsymbol{W}_K$. We allow for static state and static stored patterns. Static pattern

means that the pattern does not depend on the network input, i.e. it is determined by the bias weights and remains constant across different network inputs. Static state patterns allow to determine whether particular fixed patterns are among the stored patterns and vice versa. The static pattern functionality is typically needed if particular patterns must be identified in the data, e.g. as described for immune repertoire classification in the main paper, where a fixed d_k-dimensional state vector $\boldsymbol{\xi}$ is used.

Pattern Normalization. In the appendix, e.g. in Lemma A7 or in Eq. (A91) and Eq. (A92), we showed that the characteristics of the fixed points of the new modern Hopfield network are determined by: β, M (maximal pattern norm), $m_{\max}$ (spread of the similar patterns), and $\|\boldsymbol{m_x}\|$ (center of the similar patterns). We already discussed the parameter β while the spread of the similar patterns $m_{\max}$ is given by the data. The remaining variables M and $\boldsymbol{m_x}$ that both control the fixed point dynamics are adjusted pattern normalization. M is the maximal pattern norm and $\boldsymbol{m_x}$ the center of the similar patterns. Theorem A5 says that larger M allows for more patterns to be stored. However, the size of metastable states will decrease with increasing M. The vector $\boldsymbol{m_x}$ says how well the (similar) patterns are centered. If the norm $\|\boldsymbol{m_x}\|$ is large, then this leads to smaller metastable states. The two parameters M and $\boldsymbol{m_x}$ are controlled by pattern normalization and determine the size and convergence properties of metastable states. These two parameters are important for creating large gradients if heads start with global averaging which has small gradient. These two parameters can shift a head towards small metastable states which have largest gradient as shown in Fig. B2(b). We allow for three different pattern normalizations:

- pattern normalization of the input patterns,
- pattern normalization after mapping into the associative space,
- no pattern normalization.

The default setting is a pattern normalization of the input patterns.

C3 Usage

As outlined in Sec. C1, there are a variety of possible use cases for the Hopfield layer, e.g. to build memory networks or transformer models. The goal of the implementation is therefore to provide an easy to use Hopfield module that can be used in a wide range of applications, be it as part of a larger architecture or as a standalone module. Consequently, the focus of the Hopfield layer interface is set on its core parameters: the association of two sets, the scaling parameter β, the maximum number of updates, the dimension of the associative space, the possible usage of static patterns, and the pattern normalization. The integration into the PyTorch framework is built such that with all the above functionalities disabled, the "HopfieldEncoderLayer" and the "HopfieldDecoderLayer", both extensions of the Hopfield module, can be used as a one-to-one plug-in replacement for the *TransformerEncoderLayer* and the *TransformerDecoderLayer*, respectively, of the PyTorch transformer module.

The Hopfield layer can be used to implement or to substitute different layers:

- **Pooling layers:** We consider the Hopfield layer as a pooling layer if only one static state (query) pattern exists. Then, it is de facto a pooling over the sequence, which results from the softmax values applied on the stored patterns. Therefore, our Hopfield layer can act as a pooling layer.
- **Permutation equivariant layers:** Our Hopfield layer can be used as a plug-in replacement for permutation equivariant layers. Since the Hopfield layer is an associative memory it assumes no dependency between the input patterns.
- **GRU & LSTM layers:** Our Hopfield layer can be used as a plug-in replacement for GRU & LSTM layers. Optionally, for substituting GRU & LSTM layers, positional encoding might be considered.
- **Attention layers:** Our Hopfield layer can act as an attention layer, where state (query) and stored (key) patterns are different, and need to be associated.
- Finally, the extensions of the Hopfield layer are able to operate as a self-attention layer (HopfieldEncoderLayer) and as cross-attention layer (HopfieldDecoderLayer), as described in [64]. As such, it can be used as building block of transformer-based or general architectures.

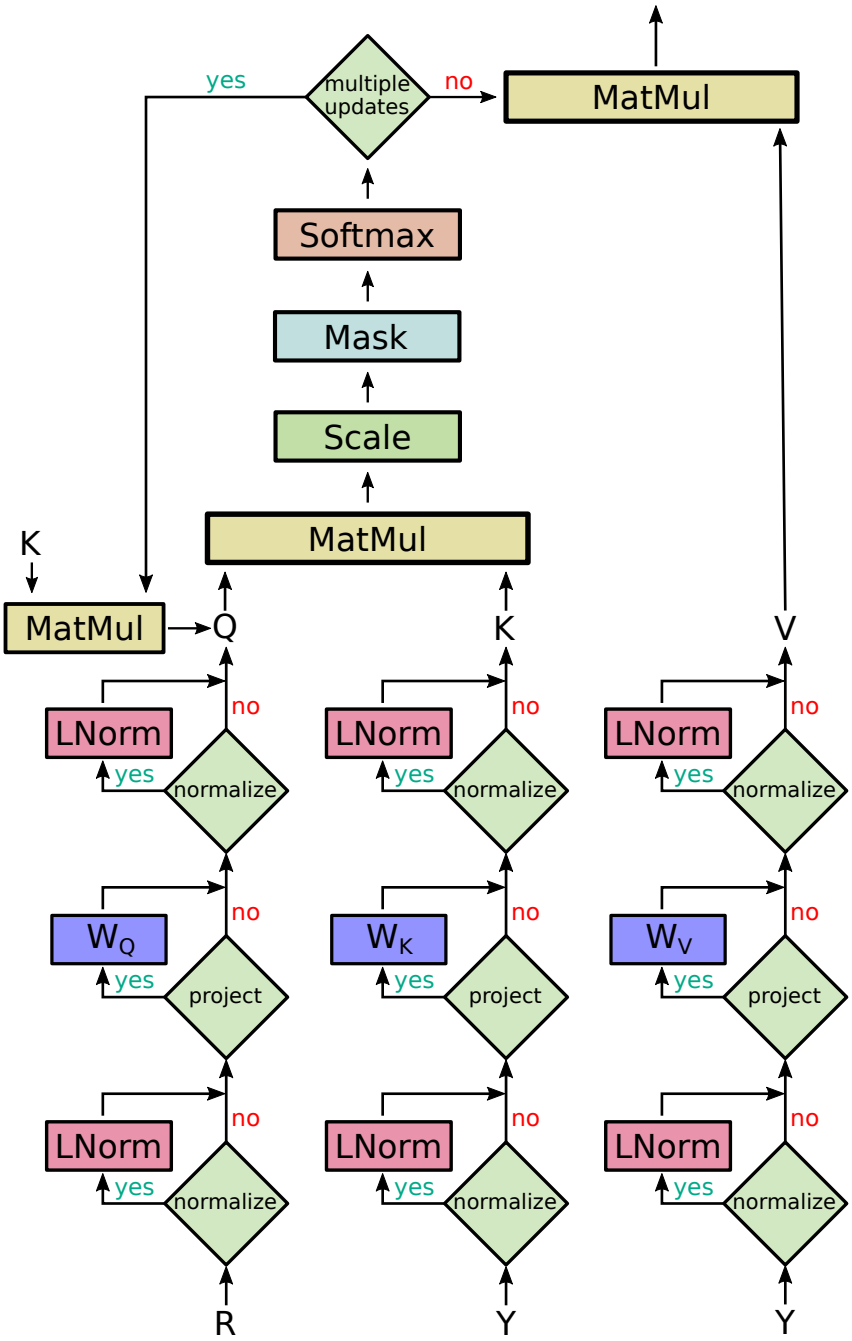

Figure C1: A flowchart of the Hopfield layer. First, the raw state (query) patterns $\boldsymbol{R}$ and the raw stored (key) patterns $\boldsymbol{Y}$ are optionally normalized (with layer normalization), projected and optionally normalized (with layer normalization) again. The default setting is a layer normalization of the input patterns, and no layer normalization of the projected patterns. The raw stored patterns $\boldsymbol{Y}$ can in principle be also two different input tensors. Optionally, multiple updates take place in the projected space of $\boldsymbol{Q}$ and $\boldsymbol{K}$. This update rule is obtained e.g. from the full update Eq. (A414) or the simplified update Eq. (A415) in the appendix.

Bibliography

[1] Y. Abu-Mostafa and J.-M-StJacques. Information capacity of the Hopfield model. *IEEE Transactions on Information Theory*, 31, 1985.

[2] R. Agrawal, T. Imieliundefinedski, and A. Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, 1993.

[3] R. Akbar, P. A. Robert, M. Pavlović, J. R. Jeliazkov, I. Snapkov, A. Slabodkin, C. R. Weber, L. Scheffer, E. Miho, I. H. Haff, et al. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *bioRxiv*, 2019.

[4] F. Alzahrani and A. Salem. Sharp bounds for the lambert w function. *Integral Transforms and Special Functions*, 29(12):971–978, 2018.

[5] J. Ba, G. E. Hinton, V. Mnih, J. Z. Leibo, and C. Ionescu. Using fast weights to attend to the recent past. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4331–4339. Curran Associates, Inc., 2016.

[6] J. Ba, G. E. Hinton, V. Mnih, J. Z. Leibo, and C. Ionescu. Using fast weights to attend to the recent past. *ArXiv*, 1610.06258, 2016.

[7] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *ArXiv*, 1409.0473, 2014. appeared in ICRL 2015.

[8] A. Banino, A. P. Badia, R. Köster, M. J. Chadwick, V. Zambaldi, D. Hassabis, C. Barry, M. Botvinick, D. Kumaran, and C. Blundell. MEMO: a deep network for flexible combination of episodic memories. *ArXiv*, 2001.10913, 2020.

[9] A. Barra, M. Beccaria, and A. Fachechi. A new mechanical approach to handle generalized Hopfield neural networks. *Neural Networks*, 106:205–222, 2018.

[10] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Cham: Springer International Publishing, 2nd edition, 2017.

[11] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 7th edition, 2009.

[12] J. S. Brauchart, A. B. Reznikov, E. B. Saff, I. H. Sloan, Y. G. Wang, and R. S. Womersley. Random point sets on the sphere - hole radii, covering, and separation. *Experimental Mathematics*, 27(1):62–81, 2018.

[13] J. Bruck and V. P. Roychowdhury. On the number of spurious memories in the Hopfield model. *IEEE Transactions on Information Theory*, 36(2):393–397, 1990.

[14] T. Cai, J. Fan, and T. Jiang. Distributions of angles in random packing on spheres. *Journal of Machine Learning Research*, 14(21):1837–1864, 2013.

[15] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon. Multiple instance learning: a survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018.

[16] A. Carta, A. Sperduti, and D. Bacciu. Encoding-based memory modules for recurrent neural networks. *ArXiv*, 2001.11771, 2020.

[17] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 1724–1734, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.

[18] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. *ArXiv*, 2003.10555, 2020.

[19] A. Crisanti, D. J. Amit, and H. Gutfreund. Saturation level of the Hopfield model for neural network. *Europhysics Letters (EPL)*, 2(4):337–341, 1986.

[20] I. Danihelka, G. Wayne, B. Uria, N. Kalchbrenner, and A. Graves. Associative long short-term memory. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1986–1994, New York, USA, 2016.

[21] M. Daniluk, T. Rocktäschel, J. Welbl, and S. Riedel. Frustratingly short attention spans in neural language modeling. *ArXiv*, 1702.04521, 2017. appeared in ICRL 2017.

[22] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and L. Kaiser. Universal transformers. *ArXiv*, 1807.03819, 2018. Published at ICLR 2019.

[23] M. Demircigil, J. Heusel, M. Löwe, S. Upgang, and F. Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2):288–299, 2017.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *ArXiv*, 1810.04805, 2018.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. 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)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[26] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

[27] R. O. Emerson, W. S. DeWitt, M. Vignali, J. Gravley, J. K. Hu, E. J. Osborne, C. Desmarais, M. Klinger, C. S. Carlson, J. A. Hansen, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nature Genetics*, 49(5):659, 2017.

[28] V. Folli, M. Leonetti, and G. Ruocco. On the maximum storage capacity of the Hopfield model. *Frontiers in Computational Neuroscience*, 10(144), 2017.

[29] B. Gao and L. Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *ArXiv*, 1704.00805, 2017.

[30] D. J. H. Garling. *Analysis on Polish Spaces and an Introduction to Optimal Transportation*. London Mathematical Society Student Texts. Cambridge University Press, 2017.

[31] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *ArXiv*, 1410.5401, 2014.

[32] N. Guttenberg, N. Virgo, O. Witkowski, H. Aoki, and R. Kanai. Permutation-equivariant neural networks applied to dynamics prediction. *arXiv*, 1612.04530, 2016.

[33] J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation*. Addison-Wesley Longman Publishing Co., Inc., Redwood City, CA, 1991.

[34] S. Hochreiter. Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, Institut für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München, 1991. Advisor: J. Schmidhuber.

[35] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[36] A. Hoorfar and M. Hassani. Inequalities on the Lambert w function and hyperpower function. *Journal of Inequalities in Pure and Applied Mathematics*, 9(2):1–5, 2008.

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

[38] M. Ilse, J. M. Tomczak, and M. Welling. Attention-based deep multiple instance learning. *International Conference on Machine Learning (ICML)*, 2018.

[39] M. Ilse, J. M. Tomczak, and M. Welling. Deep multiple instance learning for digital histopathology. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, pages 521–546. Elsevier, 2020.

[40] I. Korshunova, J. Degrave, F. Huszar, Y. Gal, A. Gretton, and J. Dambre. BRUNO: A deep recurrent model for exchangeable data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7190–7198. Curran Associates, Inc., 2018.

[41] D. Krotov and J. J. Hopfield. Dense associative memory for pattern recognition. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 1172–1180. Curran Associates, Inc., 2016.

[42] D. Krotov and J. J. Hopfield. Dense associative memory is robust to adversarial inputs. *Neural Computation*, 30(12):3151–3167, 2018.

[43] T. Lipp and S. Boyd. Variations and extension of the convex–concave procedure. *Optimization and Engineering*, 17(2):263–287, 2016.

[44] C. Mazza. On the storage capacity of nonlinear neural networks. *Neural Networks*, 10(4):593–597, 1997.

[45] R. J. McEliece, E. C. Posner, E. R. Rodemich, and S. S. Venkatesh. The capacity of the Hopfield associative memory. *IEEE Trans. Inf. Theor.*, 33(4):461–482, 1987.

[46] S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer sentinel mixture models. *ArXiv*, 2003.10555, 2016.

[47] R. R. Meyer. Sufficient conditions for the convergence of monotonic mathematical programming algorithms. *Journal of Computer and System Sciences*, 12(1):108–121, 1976.

[48] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark. *NIST handbook of mathematical functions*. Cambridge University Press, 1 pap/cdr edition, 2010.

[49] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *Workshop in Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[50] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[51] C. R. Qi, H. Su, M. Kaichun, and L. J. Guibas. PointNet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2017.

[52] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *31st International Conference on Neural Information Processing Systems*, page 5105–5114. Curran Associates Inc., 2017.

[53] A. Rangarajan, S. Gold, and E. Mjolsness. A novel optimizing network architecture with applications. *Neural Computation*, 8(5):1041–1060, 1996.

[54] A. Rangarajan, A. Yuille, and E. E. Mjolsness. Convergence properties of the softassign quadratic assignment algorithm. *Neural Computation*, 11(6):1455–1474, 1999.

[55] S. Ravanbakhsh, J. Schneider, and B. Poczos. Deep learning with sets and point clouds. *arXiv*, 1611.04500, 2016.

[56] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, and J. Wang. Release strategies and the social impacts of language models. *arXiv*, 1908.09203, 2019.

[57] B. K. Sriperumbudur and G. R. Lanckriet. On the convergence of the concave-convex procedure. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1759–1767. Curran Associates, Inc., 2009.

[58] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. End-to-end memory networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2440–2448. Curran Associates, Inc., 2015.

[59] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. End-to-end memory networks. *ArXiv*, 1503.08895, 2015.

[60] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2 edition, 2018.

[61] F. Tanaka and S. F. Edwards. Analytic theory of the ground state properties of a spin glass. I. Ising spin glass. *Journal of Physics F: Metal Physics*, 10(12):2769–2778, 1980.

[62] Y. Tay, D. Bahri, D. Metzler, D.-C. Juan, Z. Zhao, and C. Zheng. Synthesizer: Rethinking self-attention in transformer models. *ArXiv*, 2005.00743, 2020.

[63] J. J. Torres, L. Pantic, and H. H. J. Kappen. Storage capacity of attractor neural networks with depressing synapses. *Phys. Rev. E*, 66:061910, 2002.

[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[65] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *ArXiv*, 1706.03762, 2017.

[66] G. Wainrib and J. Touboul. Topological and dynamical complexity of random neural networks. *Phys. Rev. Lett.*, 110:118101, 2013.

[67] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018.

[68] C. R. Weber, R. Akbar, A. Yermanos, M. Pavlović, I. Snapkov, G. K. Sandve, S. T. Reddy, and V. Greiff. immuneSIM: tunable multi-feature simulation of B- and T-cell receptor repertoires for immunoinformatics benchmarking. *Bioinformatics*, 03 2020.

[69] J. Weston, S. Chopra, and A. Bordes. Memory networks. *ArXiv*, 1410.3916, 2014.

[70] M. Widrich, B. Schäfl, M. Pavlović, H. Ramsauer, L. Gruber, M. Holzleitner, J. Brandstetter, G. K. Sandve, V. Greiff, S. Hochreiter, and G. Klambauer. Modern Hopfield networks and attention for immune repertoire classification. *ArXiv*, 2020.

[71] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. HuggingFace's transformers: State-of-the-art natural language processing. *ArXiv*, 1910.03771, 2019.

[72] J. C. F. Wu. On the convergence properties of the em algorithm. *Ann. Statist.*, 11(1):95–103, 1983.

[73] Y. Xu, T. Fan, M. Xu, L. Zeng, and Y. Qiao. SpiderCNN: Deep learning on point sets with parameterized convolutional filters. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *European Conference on Computer Vision (ECCV)*, pages 90–105. Springer International Publishing, 2018.

[74] A. L. Yuille and A. Rangarajan. The concave-convex procedure (CCCP). In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 1033–1040. MIT Press, 2002.

[75] A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.

[76] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola. Deep sets. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3391–3401. Curran Associates, Inc., 2017.

[77] W. I. Zangwill. *Nonlinear programming: a unified approach*. Prentice-Hall international series in management. Englewood Cliffs, N.J., 1969.

[78] S. Zhai, W. Talbott, M. A. Bautista, C. Guestrin, and J. M. Susskind. Set distribution networks: a generative model for sets of images. *arXiv*, 2006.10705, 2020.

[79] W. Zhang and B. Zhou. Learning to update auto-associative memory in recurrent neural networks for improving sequence memorization. *ArXiv*, 1709.06493, 2017.

[80] Y. Zhu, R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015. arXiv 1506.06724.