SUPPLEMENT B: Modern Hopfield Networks and Attention for Immune Repertoire Classification

Sepp Hochreiter[†]
Markus Holzleitner **Lukas Gruber** **Hubert Ramsauer**
Günter Klambauer **Johannes Brandstetter**
ELLIS Unit Linz and LIT AI Lab
Institute for Machine Learning
Johannes Kepler University Linz, Austria
[†] also at Institute of Advanced Research in Artificial Intelligence (IARAI)

Abstract

Contents

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

List of theorems

List of definitions

B1 Introduction

This document is a supplement to the paper "Modern Hopfield Networks and Attention for Immune Repertoire Classification".
In the next section (Section B2) our new modern Hopfield network is introduced. In Subsection B2.1 we present the new energy function. Then in Subsection B2.2, our new update rule is introduced. In Subsection B2.3, 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 B2.4, we consider the local convergence of the update rule and see that it converges after one update. In Subsection B2.5, we consider the properties of the fixed points that are associated with the stored patterns. In Subsection B2.5.1, we show that exponentially many patterns can be stored. The main result is given in Theorem B5: for random patterns on a sphere we can store and retrieve exponentially (in the dimension of the space) many patterns. Subsection B2.5.2 reports that the update converges after one update step and that the retrieval error is exponentially small.
In Subsection B2.6, we consider how associations for the new Hopfield networks can be learned. In Subsection B2.6.1, we consider the initialization. In Subsection B2.6.2, we analyze if the association is learned directly by a bilinear form. In Subsection B2.6.3, 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 B2.7, 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 B2.7.1. In Subsection B2.7.2, we consider the controlled forgetting behavior.
In Section B3, 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 B4, we review the new Hopfield network as introduced by Krotov and Hopfield in 2016. However in contrast to our new Hopfield network, Krotov and Hopfield' new Hopfield network is a binary, that is, a network with binary states. In Subsection B4.1, we give an introduction to neural networks equipped with associative memories and new Hopfield networks. In Subsection B4.1.1, we discuss neural networks that are enhanced by an additional external memory and by attention mechanisms. In Subsection B4.1.2, we give an overview over the modern Hopfield networks. Finally, in Subsection B4.2, we present the energy function and the update rule for the modern, binary Hopfield networks.

B2 Modern Hopfield Networks: Continuous States (New Concept)

B2.1 New Energy Function

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

$$\boldsymbol{X} \ = \ (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \ . \tag{1}$$

The largest norm of a pattern is

$$M \ = \ \max_i \| \boldsymbol{x}_i \| \ . \tag{2}$$

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 [28]. The storage capacity is proportional to d^{n-1} [28]. This model was generalized by Demircigil et al. [18] 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 B4.
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. [18] 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

$$\mathrm{E} \;=\; -\,\mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) \;+\; \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \beta^{-1} \ln N \;+\; \frac{1}{2} M^2 \tag{3}$$

$$=\; -\,\beta^{-1} \ln \left(\sum_{i=1}^{N} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}) \right) \;+\; \beta^{-1} \ln N \;+\; \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \frac{1}{2} M^2 \,. \tag{4}$$

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

Lemma 1. *The energy* E *is larger than zero:*

$$0 \;\leqslant\; \mathrm{E} \,. \tag{5}$$

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

$$\mathrm{E} \;\leqslant\; \beta^{-1} \ln N \;+\; \frac{1}{2}\, M^2 \,, \tag{6}$$

$$\mathrm{E} \;\leqslant\; 2\, M^2 \,. \tag{7}$$

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

$$\boldsymbol{x}_{\boldsymbol{\xi}} \;=\; \boldsymbol{x}_k \,, \quad k \;=\; \arg\max_i \boldsymbol{\xi}^T \boldsymbol{x}_i \,. \tag{8}$$

We obtain

$$\begin{aligned}
\mathrm{E} \;&=\; -\,\beta^{-1} \ln \left(\sum_{i=1}^{N} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}) \right) \;+\; \beta^{-1} \ln N \;+\; \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \frac{1}{2}\, M^2 \\
&=\; -\,\beta^{-1} \ln \left(\frac{1}{N} \sum_{i=1}^{N} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}) \right) \;+\; \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \frac{1}{2}\, M^2 \\
&\geq\; -\,\beta^{-1} \ln \left(\frac{1}{N} \sum_{i=1}^{N} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}) \right) \;+\; \frac{1}{2}\, \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \frac{1}{2}\, \boldsymbol{x}_{\boldsymbol{\xi}}^T \boldsymbol{x}_{\boldsymbol{\xi}} \\
&\geq\; -\,\beta^{-1} \ln \left(\exp(\beta \boldsymbol{x}_{\boldsymbol{\xi}}^T \boldsymbol{\xi}) \right) \;+\; \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \frac{1}{2}\, \boldsymbol{x}_{\boldsymbol{\xi}}^T \boldsymbol{x}_{\boldsymbol{\xi}} \\
&=\; -\,\boldsymbol{x}_{\boldsymbol{\xi}}^T \boldsymbol{\xi} \;+\; \frac{1}{2}\, \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \frac{1}{2}\, \boldsymbol{x}_{\boldsymbol{\xi}}^T \boldsymbol{x}_{\boldsymbol{\xi}} \\
&=\; \frac{1}{2}\, \left(\boldsymbol{\xi} \;-\; \boldsymbol{x}_{\boldsymbol{\xi}} \right)^T \left(\boldsymbol{\xi} \;-\; \boldsymbol{x}_{\boldsymbol{\xi}} \right) \;=\; \frac{1}{2}\, \left\| \boldsymbol{\xi} \;-\; \boldsymbol{x}_{\boldsymbol{\xi}} \right\|^2 \;\geq\; 0 \,.
\end{aligned} \tag{9}$$

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,

$$\boldsymbol{\xi} \;=\; \sum_{i=1}^{N} p_i\, \boldsymbol{x}_i \,, \quad \sum_{i=1}^{N} p_i \;=\; 1 \,, \quad \forall_i :\; 0 \;\leqslant\; p_i \,. \tag{10}$$

The first upper bound is.

$$\begin{aligned}
\mathrm{E} \;&=\; -\,\beta^{-1} \ln \left(\sum_{i=1}^{N} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}) \right) \;+\; \frac{1}{2}\, \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \beta^{-1} \ln N \;+\; \frac{1}{2}\, M^2 \\
&\leqslant\; -\sum_{i=1}^{N} p_i\, (\boldsymbol{x}_i^T \boldsymbol{\xi}) \;+\; \frac{1}{2}\, \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \beta^{-1} \ln N \;+\; \frac{1}{2}\, M^2 \\
&=\; -\,\frac{1}{2}\, \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \beta^{-1} \ln N \;+\; \frac{1}{2}\, M^2 \;\leqslant\; \beta^{-1} \ln N \;+\; \frac{1}{2}\, M^2 \,.
\end{aligned} \tag{11}$$

For the first inequality we applied Lemma 19 to $-\mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi})$ with $\boldsymbol{z} = \boldsymbol{p}$ giving

$$- \mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) \ \leqslant \ - \sum_{i=1}^{N} p_i \ (\boldsymbol{x}_i^T \boldsymbol{\xi}) \ + \ \beta^{-1} \sum_{i=1}^{N} p_i \ln p_i \ \leqslant \ - \sum_{i=1}^{N} p_i \ (\boldsymbol{x}_i^T \boldsymbol{\xi}) \ , \tag{12}$$

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

$$\boldsymbol{m}_{\boldsymbol{x}} \ = \ \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i \ . \tag{13}$$

We obtain

$$\begin{aligned}
\mathrm{E} \ &= \ - \beta^{-1} \ln \left(\sum_{i=1}^{N} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}) \right) \ + \ \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \ + \ \beta^{-1} \ln N \ + \ \frac{1}{2} M^2 \\
&\leqslant \ - \sum_{i=1}^{N} \frac{1}{N} \boldsymbol{x}_i^T \boldsymbol{\xi} \ + \ \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \ + \ \frac{1}{2} M^2 \\
&= \ - \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{\xi} \ + \ \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \ + \ \frac{1}{2} M^2 \\
&\leqslant \ \| \boldsymbol{m}_{\boldsymbol{x}} \| \ \| \boldsymbol{\xi} \| \ + \ \frac{1}{2} \| \boldsymbol{\xi} \|^2 \ + \ \frac{1}{2} M^2 \\
&\leqslant \ 2 \ M^2 \ ,
\end{aligned} \tag{14}$$

where for the first inequality we again applied Lemma 19 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

$$\| \boldsymbol{\xi} \| \ = \ \left\| \sum_i p_i \ \boldsymbol{x}_i \right\| \ \leqslant \ \sum_i p_i \ \| \boldsymbol{x}_i \| \ \leqslant \ \sum_i p_i M \ = \ M \tag{15}$$

and

$$\| \boldsymbol{m}_{\boldsymbol{x}} \| \ = \ \left\| \sum_i (1/N) \ \boldsymbol{x}_i \right\| \ \leqslant \ \sum_i (1/N) \ \| \boldsymbol{x}_i \| \ \leqslant \ \sum_i (1/N) \ M \ = \ M \ . \tag{16}$$

□

B2.2 New Update Rule

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

$$\boldsymbol{\xi}^{\mathrm{new}} \ = \ \boldsymbol{X} \boldsymbol{p} \ = \ \boldsymbol{X} \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) \ , \tag{17}$$

where we used

$$\boldsymbol{p} \ = \ \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) \ . \tag{18}$$

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

$$\boldsymbol{\xi}^{\mathrm{new}} \ = \ \mathrm{sgn} \left(\boldsymbol{X} \boldsymbol{X}^T \boldsymbol{\xi} \right) . \tag{19}$$

Therefore instead of using the vector $\boldsymbol{X}^T \boldsymbol{\xi}$ as in the classical Hopfield network, its softmax version $\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})$ is used.
In the next section (Section B2.3) we show that the update rule Eq. (17) 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 B2.4 we consider the local convergence of the update rule Eq. (17) and see that it converges after one update.

B2.3 Global Convergence of the Update Rule

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

$$\boldsymbol{\xi}^{\text{new}} \;=\; f(\boldsymbol{\xi}) \;=\; \boldsymbol{X}\boldsymbol{p} \;=\; \boldsymbol{X}\text{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) \;, \tag{20}$$

where we used

$$\boldsymbol{p} \;=\; \text{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) \;. \tag{21}$$

We defined the energy function

$$\text{E} \;=\; -\,\text{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) \;+\; \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \beta^{-1} \ln N \;+\; \frac{1}{2} M^2 \tag{22}$$

$$=\; -\,\beta^{-1} \ln \left(\sum_{i=1}^{N} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}) \right) \;+\; \beta^{-1} \ln N \;+\; \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \frac{1}{2} M^2 \;. \tag{23}$$

We will show that the update rule in Eq. (20) is the Concave-Convex Procedure (CCCP) for minimizing the energy E. The CCCP is proven to converge globally.

Theorem B1 (Global Convergence (Zangwill): Energy)**.** *The update rule Eq. (20) 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) [51, 52] minimizes a function that is the sum of a concave function and a convex function. CCCP is equivalent to Legendre minimization [36, 37] algorithms [52]. The Jacobian of the softmax is positive semi-definite according to Lemma 22. 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 $\text{E}(\boldsymbol{\xi})$ is the sum of the convex function $\text{E}_1(\boldsymbol{\xi}) = 1/2 \boldsymbol{\xi}^T \boldsymbol{\xi} + C_1$ and the concave function $\text{E}_2(\boldsymbol{\xi}) = -\text{lse}$:

$$\text{E}(\boldsymbol{\xi}) \;=\; \text{E}_1(\boldsymbol{\xi}) \;+\; \text{E}_2(\boldsymbol{\xi}) \;, \tag{24}$$

$$\text{E}_1(\boldsymbol{\xi}) \;=\; \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; \beta^{-1} \ln N \;+\; \frac{1}{2} M^2 \;=\; \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \;+\; C_1 \;, \tag{25}$$

$$\text{E}_2(\boldsymbol{\xi}) \;=\; -\,\text{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) \;, \tag{26}$$

where C_1 does not depend on $\boldsymbol{\xi}$.
The Concave-Convex Procedure (CCCP) [51, 52] applied to E is

$$\nabla_\xi \text{E}_1 \left(\boldsymbol{\xi}^{t+1} \right) \;=\; -\,\nabla_\xi \text{E}_2 \left(\boldsymbol{\xi}^t \right) \;, \tag{27}$$

which is

$$\nabla_\xi \left(\frac{1}{2} \left(\boldsymbol{\xi}^{t+1} \right)^T \boldsymbol{\xi}^{t+1} \;+\; C_1 \right) \;=\; \nabla_\xi \text{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}^t) \;. \tag{28}$$

The resulting update rule is

$$\boldsymbol{\xi}^{t+1} \;=\; \boldsymbol{X} \boldsymbol{p}^t \;=\; \boldsymbol{X} \text{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}^t) \tag{29}$$

using

$$\boldsymbol{p}^t \;=\; \text{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}^t) \;. \tag{30}$$

This is the update rule in Eq. (20).
Theorem 2 in [51] and Theorem 2 in [52] state that the update rule Eq. (20) is guaranteed to monotonically decrease the energy E as a function of time. See also Theorem 2 in [39]. □

Although the objective converges in all cases, it does not necessarily converge to a local minimum [30].
However the convergence proof of CCCP in [51, 52] was not as rigorous as required. In [39] a rigorous analysis of the convergence of CCCP is performed using Zangwill's global convergence theory of iterative algorithms.
In [39] the minimization problem

$$\begin{aligned} &\min_{\boldsymbol{\xi}} \;\; \text{E}_1 \;+\; \text{E}_2 \\ &\text{s.t.} \;\; \boldsymbol{c}(\boldsymbol{\xi}) \leqslant \boldsymbol{0} \;, \quad \boldsymbol{d}(\boldsymbol{\xi}) \;=\; \boldsymbol{0} \end{aligned} \tag{31}$$

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 [39] as

$$\begin{aligned} \boldsymbol{\xi}^{t+1} \in \arg\min_{\boldsymbol{\xi}} \; & \mathrm{E}_1\left(\boldsymbol{\xi}\right) + \boldsymbol{\xi}^T \nabla_{\xi} \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) \\ \text{s.t.} \quad & \boldsymbol{c}(\boldsymbol{\xi}) \leqslant \mathbf{0}\,, \quad \boldsymbol{d}(\boldsymbol{\xi}) = \mathbf{0}\,. \end{aligned} \tag{32}$$

We define the upper bound E_C on the energy:

$$\mathrm{E}_\mathrm{C}\left(\boldsymbol{\xi}, \boldsymbol{\xi}^t\right) := \mathrm{E}_1\left(\boldsymbol{\xi}\right) + \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) + \left(\boldsymbol{\xi} - \boldsymbol{\xi}^t\right)^T \nabla_{\xi} \mathrm{E}_2\left(\boldsymbol{\xi}^t\right)\,. \tag{33}$$

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

$$\mathrm{E}_\mathrm{C}\left(\boldsymbol{\xi}^t, \boldsymbol{\xi}^t\right) = \mathrm{E}_1\left(\boldsymbol{\xi}^t\right) + \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) = \mathrm{E}\left(\boldsymbol{\xi}^t\right)\,. \tag{34}$$

Since $-\mathrm{E}_2$ is convex, the first order characterization of convexity holds (Eq. 3.2 in [9]):

$$-\mathrm{E}_2\left(\boldsymbol{\xi}\right) \geq -\mathrm{E}_2\left(\boldsymbol{\xi}^t\right) - \left(\boldsymbol{\xi} - \boldsymbol{\xi}^t\right)^T \nabla_{\xi} \mathrm{E}_2\left(\boldsymbol{\xi}^t\right)\,, \tag{35}$$

that is

$$\mathrm{E}_2\left(\boldsymbol{\xi}\right) \leqslant \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) + \left(\boldsymbol{\xi} - \boldsymbol{\xi}^t\right)^T \nabla_{\xi} \mathrm{E}_2\left(\boldsymbol{\xi}^t\right)\,. \tag{36}$$

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

$$\begin{aligned} \mathrm{E}\left(\boldsymbol{\xi}\right) &\leqslant \mathrm{E}_\mathrm{C}\left(\boldsymbol{\xi}, \boldsymbol{\xi}^t\right) = \mathrm{E}_1\left(\boldsymbol{\xi}\right) + \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) + \left(\boldsymbol{\xi} - \boldsymbol{\xi}^t\right)^T \nabla_{\xi} \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) \\ &= \mathrm{E}_1\left(\boldsymbol{\xi}\right) + \boldsymbol{\xi}^T \nabla_{\xi} \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) + C_2\,, \end{aligned} \tag{37}$$

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

$$\boldsymbol{\xi}^{t+1} \in \arg\min_{\boldsymbol{\xi}} \; \mathrm{E}_\mathrm{C}\left(\boldsymbol{\xi}, \boldsymbol{\xi}^t\right)\,, \tag{38}$$

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

$$\mathrm{E}\left(\boldsymbol{\xi}^{t+1}\right) \leqslant \mathrm{E}_\mathrm{C}\left(\boldsymbol{\xi}^{t+1}, \boldsymbol{\xi}^t\right) \leqslant \mathrm{E}_\mathrm{C}\left(\boldsymbol{\xi}^t, \boldsymbol{\xi}^t\right) = \mathrm{E}\left(\boldsymbol{\xi}^t\right)\,. \tag{39}$$

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

$$\mathrm{E}_\mathrm{C}\left(\boldsymbol{\xi}, \boldsymbol{\xi}^t\right) = \mathrm{E}_1\left(\boldsymbol{\xi}\right) + \boldsymbol{\xi}^T \nabla_{\xi} \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) + C_2 \tag{40}$$

as a function of $\boldsymbol{\xi}$.
For a minimum not at the border, the derivative has to be the zero vector

$$\frac{\partial \mathrm{E}_\mathrm{C}\left(\boldsymbol{\xi}, \boldsymbol{\xi}^t\right)}{\partial \boldsymbol{\xi}} = \boldsymbol{\xi} + \nabla_{\xi} \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) = \boldsymbol{\xi} - \boldsymbol{X}\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}^t) = \mathbf{0} \tag{41}$$

and the Hessian must be positive semi-definite

$$\frac{\partial^2 \mathrm{E}_\mathrm{C}\left(\boldsymbol{\xi}, \boldsymbol{\xi}^t\right)}{\partial \boldsymbol{\xi}^2} = \boldsymbol{I}\,. \tag{42}$$

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. E_C can even be written as a quadratic form:

$$\mathrm{E}_\mathrm{C}\left(\boldsymbol{\xi}, \boldsymbol{\xi}^t\right) = \frac{1}{2}\left(\boldsymbol{\xi} + \nabla_{\xi} \mathrm{E}_2\left(\boldsymbol{\xi}^t\right)\right)^T \left(\boldsymbol{\xi} + \nabla_{\xi} \mathrm{E}_2\left(\boldsymbol{\xi}^t\right)\right) + C_3\,, \tag{43}$$

where C_3 does not depend on $\boldsymbol{\xi}$.
Therefore the minimum is

$$\boldsymbol{\xi}^{t+1} = -\nabla_{\xi} \mathrm{E}_2\left(\boldsymbol{\xi}^t\right) = \boldsymbol{X}\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}^t) \tag{44}$$

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. (44) is uniformly compact on S according to Remark 7 in [39]. Theorem 2 and

Theorem 4 in [39] states that all the limit points of the iteration Eq. (44) are stationary points. These theorems follow from Zangwill's global convergence theorem: Convergence Theorem A, page 91 in [53] and page 3 in [49].
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 [34]. 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 [49] which are given in theorems 1 to 6 in [49]. The following theorem is also very similar to Theorem 8 in [39].

Theorem B2 (Global Convergence: Stationary Points)**.** *For the iteration Eq. (44) 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. (44). If* $\mathcal{L}\left(\mathrm{E}^*\right)$ *is finite, then any sequence* $\{\boldsymbol{\xi}^t\}_{t=0}^{\infty}$ *generated by the iteration Eq. (44) 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. (40) 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. (37) as the derivative at $\boldsymbol{\xi}^t$ is not equal to zero. Eq. (42) shows that the optimization problem Eq. (37) is strict convex, hence it has only one minimum, which is a global minimum. Eq. (43) shows that the optimization problem Eq. (37) is even a quadratic form. Therefore we have

$$\mathrm{E}\left(\boldsymbol{\xi}^{t+1}\right) \leqslant \mathrm{E}_{\mathrm{C}}\left(\boldsymbol{\xi}^{t+1}, \boldsymbol{\xi}^t\right) < \mathrm{E}_{\mathrm{C}}\left(\boldsymbol{\xi}^t, \boldsymbol{\xi}^t\right) = \mathrm{E}\left(\boldsymbol{\xi}^t\right) . \tag{45}$$

Therefore the point-set-map defined by the iteration Eq. (44) (for definitions see [39]) is strictly monotonic with respect to E. Therefore we can apply Theorem 3 in [39] or Theorem 3.1 and Corollary 3.2 in [34], which give the statements of the theorem.

□

We showed global convergence of the iteration Eq. (20). We have shown that all the limit points of any sequence generated by the iteration Eq. (20) are the stationary 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 maximum. However, a local maximum as an accumulation of different iteration points is not possible because Eq. (45) 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. (20) 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.

B2.4 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.

B2.4.1 General Bound on the Jacobian of the Iterate

We consider the iteration

$$\boldsymbol{\xi}^{\mathrm{new}} = f(\boldsymbol{\xi}) = \boldsymbol{X}\boldsymbol{p} = \boldsymbol{X}\mathrm{softmax}(\beta\boldsymbol{X}^T\boldsymbol{\xi}) \tag{46}$$

using

$$\boldsymbol{p} = \mathrm{softmax}(\beta\boldsymbol{X}^T\boldsymbol{\xi}) . \tag{47}$$

The Jacobian J is symmetric and has the following form:

$$\mathrm{J} = \frac{\partial f(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \beta\, \boldsymbol{X}\left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)\boldsymbol{X}^T = \boldsymbol{X}\mathrm{J}_s\boldsymbol{X}^T , \tag{48}$$

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. B1). 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.

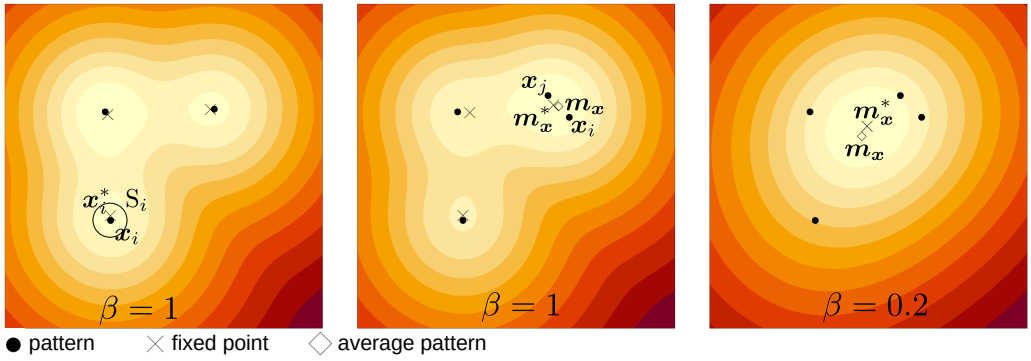

Figure B1: The three cases of fixed points. **a) Stored patterns (fixed point is single pattern)**: patterns are stored if they are well separated. Each pattern $\boldsymbol{x}_i$ has a single fixed point $\boldsymbol{x}_i^*$ close to it. In the sphere S_i, pattern $\boldsymbol{x}_i$ is the only pattern 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.

We begin with a bound on the Jacobian of the iterate, thereby heavily relying on the Jacobian of the softmax from Lemma 24.

Lemma 2. *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:*

$$\|\mathrm{J}\|_2 \leqslant 2 \beta \|\boldsymbol{X}\|_2^2 m \leqslant 2 \beta N M^2 m . \tag{49}$$

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

$$\|\mathrm{J}\|_2 \leqslant 2 \beta N M^2 \epsilon - 2 \epsilon^2 \beta N M^2 < 2 \beta N M^2 \epsilon . \tag{50}$$

Proof. With

$$\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) , \tag{51}$$

the symmetric Jacobian J is

$$\mathrm{J} = \frac{\partial f(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \beta \boldsymbol{X} \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T \right) \boldsymbol{X}^T = \boldsymbol{X} \mathrm{J}_s \boldsymbol{X}^T , \tag{52}$$

where J_s is Jacobian of the softmax.
With $m = \max_i p_i(1 - p_i)$, Eq. (465) from Lemma 24 is

$$\|\mathrm{J}_s\|_2 = \beta \left\| \mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T \right\|_2 \leqslant 2 m \beta . \tag{53}$$

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

$$\|\mathrm{J}\|_2 \leqslant \beta \left\| \boldsymbol{X}^T \right\|_2 \|\mathrm{J}_s\|_2 \|\boldsymbol{X}\|_2 \leqslant 2 m \beta \|\boldsymbol{X}\|_2^2 . \tag{54}$$

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

$$\|\boldsymbol{X}\|_2 \leqslant \|\boldsymbol{X}\|_F = \sqrt{\sum_i \|\boldsymbol{x}_i\|^2} . \tag{55}$$

Therefore, we obtain the first statement of the lemma:

$$\|\mathrm{J}\|_2 \ \leqslant \ 2\,\beta\,\|\boldsymbol{X}\|_2^2\,m \ \leqslant \ 2\,\beta\,N\,M^2\,m\,. \tag{56}$$

With $p_{\max} = \max_i p_i \geq 1 - \epsilon$ Eq. (469) in Lemma 24 is

$$\|\mathrm{J}_s\|_2 \ \leqslant \ 2\,\beta\,\epsilon \ - \ 2\,\epsilon^2\,\beta \ < \ 2\,\beta\,\epsilon\,. \tag{57}$$

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

$$\|\mathrm{J}\|_2 \ \leqslant \ 2\,\beta\,N\,M^2\,\epsilon \ - \ 2\,\epsilon^2\,\beta\,N\,M^2 \ < \ 2\,\beta\,N\,M^2\,\epsilon\,. \tag{58}$$

□

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 B1 (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:*

$$\Delta_i \ = \ \min_{j, j \neq i} \left(\boldsymbol{x}_i^T \boldsymbol{x}_i \ - \ \boldsymbol{x}_i^T \boldsymbol{x}_j \right) \ = \ \boldsymbol{x}_i^T \boldsymbol{x}_i \ - \ \max_{j, j \neq i} \boldsymbol{x}_i^T \boldsymbol{x}_j\,. \tag{59}$$

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

$$\begin{aligned} \Delta_i \ &= \ \min_{j, j \neq i} \frac{1}{2} \left(\|\boldsymbol{x}_i\|^2 \ - \ \|\boldsymbol{x}_j\|^2 \ + \ \|\boldsymbol{x}_i \ - \ \boldsymbol{x}_j\|^2 \right) \\ &= \ \frac{1}{2} \|\boldsymbol{x}_i\|^2 \ - \ \frac{1}{2} \max_{j, j \neq i} \left(\|\boldsymbol{x}_j\|^2 \ - \ \|\boldsymbol{x}_i \ - \ \boldsymbol{x}_j\|^2 \right)\,. \end{aligned} \tag{60}$$

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

$$i \ = \ \arg\max_k \min_{j, j \neq k} \left(\boldsymbol{\xi}^T \boldsymbol{x}_k \ - \ \boldsymbol{\xi}^T \boldsymbol{x}_j \right) \ = \ \arg\max_k \left(\boldsymbol{\xi}^T \boldsymbol{x}_k \ - \ \max_{j, j \neq k} \boldsymbol{\xi}^T \boldsymbol{x}_j \right) \tag{61}$$

$$0 \ \leqslant \ c \ = \ \max_k \min_{j, j \neq k} \left(\boldsymbol{\xi}^T \boldsymbol{x}_k \ - \ \boldsymbol{\xi}^T \boldsymbol{x}_j \right) \ = \ \max_k \left(\boldsymbol{\xi}^T \boldsymbol{x}_k \ - \ \max_{j, j \neq k} \boldsymbol{\xi}^T \boldsymbol{x}_j \right)\,. \tag{62}$$

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

B2.4.2 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

$$\boldsymbol{X} \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T \right) \boldsymbol{X}^T \tag{63}$$

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

$$\boldsymbol{X} \left(\mathrm{diag}(\boldsymbol{p}) \ - \ \boldsymbol{p}\boldsymbol{p}^T \right) \boldsymbol{X}^T \ = \ \boldsymbol{X} \mathrm{diag}(\boldsymbol{p}) \boldsymbol{X}^T \ - \ \boldsymbol{X}\boldsymbol{p}\boldsymbol{p}^T\boldsymbol{X}^T \tag{64}$$

$$= \ \sum_{i=1}^N p_i\,\boldsymbol{x}_i\,\boldsymbol{x}_i^T \ - \ \left(\sum_{i=1}^N p_i\,\boldsymbol{x}_i \right) \left(\sum_{i=1}^N p_i\,\boldsymbol{x}_i \right)^T \tag{65}$$

$$= \ \mathrm{E}_{\boldsymbol{p}}[\boldsymbol{x}\,\boldsymbol{x}^T] \ - \ \mathrm{E}_{\boldsymbol{p}}[\boldsymbol{x}]\,\mathrm{E}_{\boldsymbol{p}}[\boldsymbol{x}]^T \ = \ \mathrm{Var}_{\boldsymbol{p}}[\boldsymbol{x}]\,, \tag{66}$$

therefore we have

$$\mathrm{J} \ = \ \beta\,\mathrm{Var}_{\boldsymbol{p}}[\boldsymbol{x}]\,. \tag{67}$$

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:

$$\boldsymbol{m}_{\boldsymbol{x}} \;=\; \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i \;, \tag{68}$$

$$m_{\max} \;=\; \max_{1 \leqslant i \leqslant N} \|\boldsymbol{x}_i \;-\; \boldsymbol{m}_{\boldsymbol{x}}\|_2 \;. \tag{69}$$

$\boldsymbol{m}_{\boldsymbol{x}}$ is the arithmetic mean (the center) of the patterns. $m_{\max}$ is the maximal distance of the patterns to the center $\boldsymbol{m}_{\boldsymbol{x}}$.
The variance of the patterns is

$$\begin{aligned} \mathrm{Var}_{\boldsymbol{p}}[\boldsymbol{x}] \;&=\; \sum_{i=1}^{N} p_i \; \boldsymbol{x}_i \; \boldsymbol{x}_i^T \;-\; \left(\sum_{i=1}^{N} p_i \; \boldsymbol{x}_i \right) \; \left(\sum_{i=1}^{N} p_i \; \boldsymbol{x}_i \right)^T \\ &=\; \sum_{i=1}^{N} p_i \; \left(\boldsymbol{x}_i \;-\; \sum_{i=1}^{N} p_i \boldsymbol{x}_i \right) \; \left(\boldsymbol{x}_i \;-\; \sum_{i=1}^{N} p_i \boldsymbol{x}_i \right)^T \;. \end{aligned} \tag{70}$$

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

$$\|\mathrm{J}\|_2 \;\leqslant\; \beta \; m_{\max}^2 \;. \tag{71}$$

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

$$f(\boldsymbol{a}) \;=\; \sum_{i=1}^{N} p_i \|\boldsymbol{x}_i \;-\; \boldsymbol{a}\|^2 \;=\; \sum_{i=1}^{N} p_i (\boldsymbol{x}_i \;-\; \boldsymbol{a})^T (\boldsymbol{x}_i \;-\; \boldsymbol{a}) \;. \tag{72}$$

The solution to

$$\frac{\partial f(\boldsymbol{a})}{\partial \boldsymbol{a}} \;=\; 2 \; \sum_{i=1}^{N} p_i (\boldsymbol{a} \;-\; \boldsymbol{x}_i) \;=\; 0 \tag{73}$$

is

$$\boldsymbol{a} \;=\; \sum_{i=1}^{N} p_i \boldsymbol{x}_i \;. \tag{74}$$

The Hessian of f is positive definite since

$$\frac{\partial^2 f(\boldsymbol{a})}{\partial \boldsymbol{a}^2} \;=\; 2 \; \sum_{i=1}^{N} p_i \; \boldsymbol{I} \;=\; 2 \; \boldsymbol{I} \tag{75}$$

and f is a convex function. Hence, the mean

$$\bar{\boldsymbol{x}} \;:=\; \sum_{i=1}^{N} p_i \; \boldsymbol{x}_i \tag{76}$$

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

$$\sum_{i=1}^{N} p_i \|\boldsymbol{x}_i \;-\; \bar{\boldsymbol{x}}\|^2 \;\leqslant\; \sum_{i=1}^{N} p_i \|\boldsymbol{x}_i \;-\; \boldsymbol{m}_{\boldsymbol{x}}\|^2 \;\leqslant\; m_{\max}^2 \;. \tag{77}$$

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:

$$\left\|\boldsymbol{a}\boldsymbol{b}^T\right\|_2 \;=\; \sqrt{\lambda_{\max}(\boldsymbol{b}\boldsymbol{a}^T\boldsymbol{a}\boldsymbol{b}^T)} \;=\; \|\boldsymbol{a}\| \; \sqrt{\lambda_{\max}(\boldsymbol{b}\boldsymbol{b}^T)} \;=\; \|\boldsymbol{a}\| \; \|\boldsymbol{b}\| \;, \tag{78}$$

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:

$$\begin{aligned} \|\mathrm{Var}_{\boldsymbol{p}}[\boldsymbol{x}]\|_2 &\leqslant \sum_{i=1}^{N} p_i \left\| (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \, (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \right\|_2 \\ &= \sum_{i=1}^{N} p_i \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^2 \leqslant \sum_{i=1}^{N} p_i \|\boldsymbol{x}_i - \boldsymbol{m}_{\boldsymbol{x}}\|^2 \leqslant m_{\max}^2 \, . \end{aligned} \tag{79}$$

The bound of the lemma on $\|\mathrm{J}\|_2$ follows from Eq. (67).
For $\|\mathrm{J}\|_2 \leqslant \beta \, m_{\max}^2 < 1$ 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 $\|\mathrm{Var}_{\boldsymbol{p}}[\boldsymbol{x}]\|_2$ via $\|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^2$: 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:

$$\begin{aligned} \|\mathrm{Var}_{\boldsymbol{p}}[\boldsymbol{x}]\|_2 &= \mathrm{Tr}\left(\sum_{i=1}^{N} p_i \, (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \, (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \right) - \sum_{k=2}^{d} e_k \\ &= \sum_{i=1}^{N} p_i \mathrm{Tr}\left((\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \, (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \right) - \sum_{k=2}^{d} e_k \\ &= \sum_{i=1}^{N} p_i \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^2 - \sum_{k=2}^{d} e_k \, . \end{aligned} \tag{80}$$

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 B2. For N patterns $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, we consider the iteration

$$\boldsymbol{\xi}^{\mathrm{new}} = f(\boldsymbol{\xi}) = \boldsymbol{X}\boldsymbol{p} = \boldsymbol{X}\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) \tag{81}$$

using

$$\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) \, . \tag{82}$$

$\boldsymbol{\xi}^{\mathrm{new}}$ 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

$$\boldsymbol{p} = 1/N \, \boldsymbol{1} \tag{83}$$

and

$$\boldsymbol{\xi}^{\mathrm{new}} = 1/N \, \boldsymbol{X} \, \boldsymbol{1} = \boldsymbol{m}_{\boldsymbol{x}} = \boldsymbol{\xi} \, . \tag{84}$$

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

$$\begin{aligned} \left| \boldsymbol{\xi}^T (\boldsymbol{x}_i - \boldsymbol{x}_j) \right| &\leqslant \|\boldsymbol{\xi}\| \, \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \leqslant \|\boldsymbol{\xi}\| \, (\|\boldsymbol{x}_i - \boldsymbol{m}_{\boldsymbol{x}}\| + \|\boldsymbol{x}_j - \boldsymbol{m}_{\boldsymbol{x}}\|) \\ &\leqslant 2 \, m_{\max} \, \|\boldsymbol{\xi}\| \, . \end{aligned} \tag{85}$$

This inequality gives:

$$\begin{aligned} \left|\boldsymbol{\xi}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\right| &\leqslant 2\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|)\,, \\ \left|\boldsymbol{\xi}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\right| &\leqslant 2\, m_{\max}\, M\,, \end{aligned} \tag{86}$$

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

$$\beta\, \left|\boldsymbol{m}_{\boldsymbol{x}}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\right| \leqslant 2\, \beta\, m_{\max}\, \|\boldsymbol{m}_{\boldsymbol{x}}\|\,, \tag{87}$$

$$\beta\, \left|(\boldsymbol{m}_{\boldsymbol{x}}^*)^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\right| \leqslant 2\, \beta\, m_{\max}\, \|\boldsymbol{m}_{\boldsymbol{x}}^*\| \leqslant 2\, \beta\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|)\,, \tag{88}$$

$$\beta\, \left|\boldsymbol{x}_i^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\right| \leqslant 2\, \beta\, m_{\max}\, \|\boldsymbol{x}_i\| \leqslant 2\, \beta\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|)\,. \tag{89}$$

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:

$$\begin{aligned} [\mathrm{softmax}(\beta\, \boldsymbol{X}^T \boldsymbol{\xi})]_i &= \frac{1}{1 + \sum_{j \neq i} \exp(-\beta\, (\boldsymbol{\xi}^T \boldsymbol{x}_i - \boldsymbol{\xi}^T \boldsymbol{x}_j))} \\ &\leqslant \frac{1}{1 + \sum_{j \neq i} \exp(-2\, \beta\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|))} \\ &= \frac{1}{1 + (N-1) \exp(-2\, \beta\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|))} \\ &= \frac{\exp(2\, \beta\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|))}{\exp(2\, \beta\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|)) + (N-1)} \\ &\leqslant 1/N\, \exp(2\, \beta\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|))\,. \end{aligned} \tag{90}$$

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:

$$[\mathrm{softmax}(\beta\, \boldsymbol{X}^T \boldsymbol{m}_{\boldsymbol{x}})]_i \leqslant 1/N\, \exp(2\, \beta\, m_{\max}\, \|\boldsymbol{m}_{\boldsymbol{x}}\|)\,. \tag{91}$$

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:

$$\begin{aligned} [\mathrm{softmax}(\beta\, \boldsymbol{X}^T \boldsymbol{m}_{\boldsymbol{x}}^*)]_i &\leqslant 1/N\, \exp(2\, \beta\, m_{\max}\, \|\boldsymbol{m}_{\boldsymbol{x}}^*\|) \\ &\leqslant 1/N\, \exp(2\, \beta\, m_{\max}\, (m_{\max} + \|\boldsymbol{m}_{\boldsymbol{x}}\|))\,. \end{aligned} \tag{92}$$

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 2:

$$\|\mathrm{J}\|_2 < 1\,. \tag{93}$$

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. (46) in the variable $\boldsymbol{p}$:

$$\boldsymbol{p}^{\mathrm{new}} = g(\boldsymbol{p}) = \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{p})\,. \tag{94}$$

The Jacobian is

$$\mathrm{J}(\boldsymbol{p}) = \frac{\partial g(\boldsymbol{p})}{\partial \boldsymbol{p}} = \boldsymbol{X}^T \boldsymbol{X}\, \mathrm{J}_s \tag{95}$$

with

$$\mathrm{J}_s(\boldsymbol{p}^{\mathrm{new}}) = \beta \left(\mathrm{diag}(\boldsymbol{p}^{\mathrm{new}}) - \boldsymbol{p}^{\mathrm{new}} (\boldsymbol{p}^{\mathrm{new}})^T\right)\,. \tag{96}$$

The mean value theorem 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 $\mathrm{J}_s^m = \int_0^1 \mathrm{J}_s(\lambda \boldsymbol{p})\, \mathrm{d}\lambda$:

$$\boldsymbol{p}^{\mathrm{new}} \;=\; g(\boldsymbol{p}) \;=\; g(\boldsymbol{0}) \;+\; (\mathrm{J}^m)^T \boldsymbol{p} \;=\; g(\boldsymbol{0}) \;+\; \mathrm{J}_s^m\, \boldsymbol{X}^T \boldsymbol{X}\, \boldsymbol{p} \;=\; 1/N\, \boldsymbol{1} \;+\; \mathrm{J}_s^m\, \boldsymbol{X}^T \boldsymbol{X}\, \boldsymbol{p}\,. \quad (97)$$

With $m = \max_i p_i(1-p_i)$, Eq. (465) from Lemma 24 is

$$\|\mathrm{J}_s(\boldsymbol{p})\|_2 \;=\; \beta \left\|\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right\|_2 \;\leqslant\; 2\, m\, \beta\,. \quad (98)$$

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:

$$\|\mathrm{J}_s^m\|_2 \;\leqslant\; 2\, m\, \beta\,. \quad (99)$$

Eq. (468) in Lemma 24 states that every J_s is bounded by $1/2\beta$, therefore also the mean:

$$\|\mathrm{J}_s^m\|_2 \;\leqslant\; 0.5\, \beta\,. \quad (100)$$

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

$$\|\mathrm{J}_s^m\|_2 \;\leqslant\; 2\, \min\{0.25, p_{\max}\}\, \beta\,. \quad (101)$$

Consequently,

$$\|\mathrm{J}^m\|_2 \;\leqslant\; N\, M^2\, 2\, \min\{0.25, p_{\max}\}\, \beta\,, \quad (102)$$

where we used Eq. (159). $\left\|\boldsymbol{X}^T \boldsymbol{X}\right\|_2 = \left\|\boldsymbol{X}\boldsymbol{X}^T\right\|_2$, 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

$$N\, M^2\, 2\, \min\{0.25, p_{\max}\}\, \beta \;<\; 1\,. \quad (103)$$

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

$$\boldsymbol{p}^{\mathrm{new}} \;=\; g(\boldsymbol{p}) \;=\; 1/N\, \boldsymbol{1} \;+\; \mathrm{J}^m\, \boldsymbol{p}\,, \quad (104)$$

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

$$\|\mathrm{J}^m\|_2 \;\leqslant\; N\, M^2\, 2\, \min\{0.25, p_{\max}\}\, \beta \;=\; c \;<\; 1\,. \quad (105)$$

According to Banach fixed point theorem g has a fixed point in the sphere S.
Hölder's inequality gives:

$$\|\boldsymbol{p}\|^2 \;=\; \boldsymbol{p}^T \boldsymbol{p} \;\leqslant\; \|\boldsymbol{p}\|_1 \|\boldsymbol{p}\|_\infty \;=\; \|\boldsymbol{p}\|_\infty \;=\; p_{\max}\,. \quad (106)$$

Alternatively:

$$\|\boldsymbol{p}\|^2 \;=\; \sum_i p_i^2 \;=\; p_{\max} \sum_i \frac{p_i}{p_{\max}}\, p_i \;\leqslant\; p_{\max} \sum_i p_i \;=\; p_{\max}\,. \quad (107)$$

Let now S be the sphere around the origin $\boldsymbol{0}$ with radius $1/\sqrt{N} + \sqrt{p_{\max}}$ and let $\|\mathrm{J}^m(\boldsymbol{p})\|_2 \leqslant c < 1$ 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

$$\|\boldsymbol{p}^{\mathrm{new}}\| \;\leqslant\; 1/\sqrt{N} \;+\; \|\mathrm{J}^m\|_2\, \|\boldsymbol{p}\| \;\leqslant\; 1/\sqrt{N} \;+\; \sqrt{p_{\max}}\,. \quad (108)$$

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 24 and $\|\boldsymbol{p}\|_1 = 1$ to obtain from Eq. (104):

$$\|\boldsymbol{p}^{\mathrm{new}} \;-\; 1/N\, \boldsymbol{1}\|_1 \;\leqslant\; \|\mathrm{J}^m\|_1 \;\leqslant\; 2\, \beta\, m\, \|\boldsymbol{X}\|_\infty\, M_1\,, \quad (109)$$

$$\|\boldsymbol{p}^{\mathrm{new}} \;-\; 1/N\, \boldsymbol{1}\|_1 \;\leqslant\; \|\mathrm{J}^m\|_1 \;\leqslant\; 2\, \beta\, m\, N\, M_\infty\, M_1\,, \quad (110)$$

$$\|\boldsymbol{p}^{\mathrm{new}} \;-\; 1/N\, \boldsymbol{1}\|_1 \;\leqslant\; \|\mathrm{J}^m\|_1 \;\leqslant\; 2\, \beta\, m\, N\, M^2\,, \quad (111)$$

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 $\boldsymbol{X}^T\boldsymbol{X}$:

$$\begin{aligned}\left\|\boldsymbol{X}^T\boldsymbol{X}\right\|_1 &= \max_i \sum_{j=1}^N \left|\boldsymbol{x}_i^T\boldsymbol{x}_j\right| \leqslant \max_i \sum_{j=1}^N \|\boldsymbol{x}_i\|_\infty \|\boldsymbol{x}_j\|_1 \\ &\leqslant M_\infty \sum_{j=1}^N M_1 = N\, M_\infty\, M_1\,,\end{aligned} \tag{112}$$

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

$$\begin{aligned}\left\|\boldsymbol{X}^T\boldsymbol{X}\right\|_1 &= \max_i \sum_{j=1}^N \left|\boldsymbol{x}_i^T\boldsymbol{x}_j\right| \leqslant \max_i \sum_{j=1}^N \|\boldsymbol{x}_i\| \|\boldsymbol{x}_j\| \\ &\leqslant M \sum_{j=1}^N M = N\, M^2\,,\end{aligned} \tag{113}$$

where the first inequality is from Hölder's inequality (here the same as the Cauchy-Schwarz inequality). See proof of Lemma 24 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 $\|\boldsymbol{p}\|$.

$$\begin{aligned}&\min_{\boldsymbol{p}} \ \|\boldsymbol{p}\|^2 \\ &\text{s.t.} \ \sum_i p_i = 1 \\ &\qquad \forall_i : \ p_i \geq 0\,.\end{aligned} \tag{114}$$

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. (108) we obtain

$$1/\sqrt{N} \leqslant \|\boldsymbol{p}^{\text{new}}\| \leqslant 1/\sqrt{N} + \sqrt{p_{\max}}\,. \tag{115}$$

Moreover

$$\begin{aligned}\|\boldsymbol{p}^{\text{new}}\|^2 &= (\boldsymbol{p}^{\text{new}})^T\boldsymbol{p}^{\text{new}} = 1/N + (\boldsymbol{p}^{\text{new}})^T \mathrm{J}^m\, \boldsymbol{p} \leqslant 1/N + \|\mathrm{J}^m\|_2 \|\boldsymbol{p}\| \\ &\leqslant 1/N + \|\mathrm{J}^m\|_2\,,\end{aligned} \tag{116}$$

since $\boldsymbol{p}^{\text{new}} \in \mathrm{S}$ and $\boldsymbol{p} \in \mathrm{S}$.
For the fixed point, we have

$$\|\boldsymbol{p}^*\|^2 = (\boldsymbol{p}^*)^T\boldsymbol{p}^* = 1/N + (\boldsymbol{p}^*)^T \mathrm{J}^m\, \boldsymbol{p}^* \leqslant 1/N + \|\mathrm{J}^m\|_2 \|\boldsymbol{p}^*\|^2\,, \tag{117}$$

and hence

$$1/N \leqslant \|\boldsymbol{p}^*\|^2 \leqslant 1/N \frac{1}{1 - \|\mathrm{J}^m\|_2} = 1/N\,(1 + \frac{\|\mathrm{J}^m\|_2}{1 - \|\mathrm{J}^m\|_2})\,. \tag{118}$$

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

B2.4.3 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 4. *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 $j \neq i$:*

$$i \ = \ \arg\max_k \min_{j,j\neq k} \left(\boldsymbol{\xi}^T \boldsymbol{x}_k \ - \ \boldsymbol{\xi}^T \boldsymbol{x}_j\right) \ = \ \arg\max_k \left(\boldsymbol{\xi}^T \boldsymbol{x}_k \ - \ \max_{j,j\neq k} \boldsymbol{\xi}^T \boldsymbol{x}_j\right) \tag{119}$$

$$0 \ \leqslant \ c \ = \ \max_k \min_{j,j\neq k} \left(\boldsymbol{\xi}^T \boldsymbol{x}_k \ - \ \boldsymbol{\xi}^T \boldsymbol{x}_j\right) \ = \ \max_k \left(\boldsymbol{\xi}^T \boldsymbol{x}_k \ - \ \max_{j,j\neq k} \boldsymbol{\xi}^T \boldsymbol{x}_j\right) . \tag{120}$$

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

$$\|\boldsymbol{x}_i \ - \ f(\boldsymbol{\xi})\| \ \leqslant \ 2 \, \epsilon \, M \, , \tag{121}$$

where

$$M \ = \ \max_i \|\boldsymbol{x}_i\| \, , \tag{122}$$

$$\epsilon \ = \ (N-1) \ \exp(-\beta \, c) \, . \tag{123}$$

Proof. For the softmax component i we have:

$$\begin{aligned} [\text{softmax}(\beta \, \boldsymbol{X}^T \boldsymbol{\xi})]_i \ &= \ \frac{1}{1 \ + \ \sum_{j\neq i} \exp(\beta \, (\boldsymbol{\xi}^T \boldsymbol{x}_j \ - \ \boldsymbol{\xi}^T \boldsymbol{x}_i))} \ \geq \ \frac{1}{1 \ + \ \sum_{j\neq i} \exp(-\beta \, c)} \\ &= \ \frac{1}{1 \ + \ (N-1) \exp(-\beta \, c)} \ = \ 1 \ - \ \frac{(N-1) \exp(-\beta \, c)}{1 \ + \ (N-1) \exp(-\beta \, c)} \\ &\geq \ 1 \ - \ (N-1) \exp(-\beta \, c) \ = \ 1 \ - \ \epsilon \end{aligned} \tag{124}$$

For softmax components $k \neq i$ we have

$$[\text{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})]_k \ = \ \frac{\exp(\beta \, (\boldsymbol{\xi}^T \boldsymbol{x}_k \ - \ \boldsymbol{\xi}^T \boldsymbol{x}_i))}{1 \ + \ \sum_{j\neq i} \exp(\beta \, (\boldsymbol{\xi}^T \boldsymbol{x}_j \ - \ \boldsymbol{\xi}^T \boldsymbol{x}_i))} \ \leqslant \ \exp(-\beta \, c) \ = \ \frac{\epsilon}{N-1} \, . \tag{125}$$

The iteration f can be written as

$$f(\boldsymbol{\xi}) \ = \ \boldsymbol{X} \text{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) \ = \ \sum_{j=1}^N \boldsymbol{x}_j \, [\text{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})]_j \, . \tag{126}$$

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

$$\begin{aligned} \|\boldsymbol{x}_i \ - \ f(\boldsymbol{\xi})\| \ &= \ \left\| \boldsymbol{x}_i \ - \ \sum_{j=1}^N [\text{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})]_j \, \boldsymbol{x}_j \right\| \\ &= \ \left\| (1 - [\text{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})]_i) \, \boldsymbol{x}_i \ - \ \sum_{j=1, j\neq i}^N [\text{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})]_j \, \boldsymbol{x}_j \right\| \\ &\leqslant \ \epsilon \, \|\boldsymbol{x}_i\| \ + \ \frac{\epsilon}{N-1} \ \sum_{j=1, j\neq i}^N \|\boldsymbol{x}_j\| \\ &\leqslant \ \epsilon \, M \ + \ \frac{\epsilon}{N-1} \ \sum_{j=1, j\neq i}^N M \ = \ 2 \, \epsilon \, M \, . \end{aligned} \tag{127}$$

□

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:

$$\Delta_i \ = \ \min_{j,j\neq i} \left(\boldsymbol{x}_i^T \boldsymbol{x}_i \ - \ \boldsymbol{x}_i^T \boldsymbol{x}_j \right) \ = \ \boldsymbol{x}_i^T \boldsymbol{x}_i \ - \ \max_{j,j\neq i} \boldsymbol{x}_i^T \boldsymbol{x}_j \ . \tag{128}$$

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

$$\begin{aligned} \Delta_i \ &= \ \min_{j,j\neq i} \frac{1}{2} \ \left(\|\boldsymbol{x}_i\|^2 \ - \ \|\boldsymbol{x}_j\|^2 \ + \ \|\boldsymbol{x}_i \ - \ \boldsymbol{x}_j\|^2 \right) \\ &= \ \frac{1}{2} \|\boldsymbol{x}_i\|^2 \ - \ \frac{1}{2} \ \max_{j,j\neq i} \left(\|\boldsymbol{x}_j\|^2 \ - \ \|\boldsymbol{x}_i \ - \ \boldsymbol{x}_j\|^2 \right) \ . \end{aligned} \tag{129}$$

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 B2 (Sphere S_i)**.** *The sphere S_i is defined as*

$$\mathrm{S}_i \ := \ \left\{ \boldsymbol{\xi} \mid \|\boldsymbol{\xi} \ - \ \boldsymbol{x}_i\| \ \leqslant \ \frac{1}{\beta \ N \ M} \right\} \ . \tag{130}$$

Lemma 5. *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:

$$\Delta_i \ \geq \ \frac{2}{\beta \ N} \ + \ \frac{1}{\beta} \ \ln \left(2 \ (N-1) \ N \ \beta \ M^2 \right) \tag{131}$$

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.

$$\tilde{\Delta}_i \ = \ \min_{j,j\neq i} \left(\boldsymbol{\xi}^T \boldsymbol{x}_i \ - \ \boldsymbol{\xi}^T \boldsymbol{x}_j \right) \ . \tag{132}$$

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

$$\left| \boldsymbol{\xi}^T \boldsymbol{x}_j \ - \ \boldsymbol{x}_i^T \boldsymbol{x}_j \right| \leqslant \ \|\boldsymbol{\xi} \ - \ \boldsymbol{x}_i\| \ \|\boldsymbol{x}_j\| \ \leqslant \ \|\boldsymbol{\xi} \ - \ \boldsymbol{x}_i\| \ M \ . \tag{133}$$

We have the lower bound

$$\begin{aligned} \tilde{\Delta}_i \ &\geq \ \min_{j,j\neq i} \left(\left(\boldsymbol{x}_i^T \boldsymbol{x}_i \ - \ \|\boldsymbol{\xi} \ - \ \boldsymbol{x}_i\| \ M \right) \ - \ \left(\boldsymbol{x}_i^T \boldsymbol{x}_j \ + \ \|\boldsymbol{\xi} \ - \ \boldsymbol{x}_i\| \ M \right) \right) \\ &= \ - \ 2 \ \|\boldsymbol{\xi} \ - \ \boldsymbol{x}_i\| \ M \ + \ \min_{j,j\neq i} \left(\boldsymbol{x}_i^T \boldsymbol{x}_i \ - \ \boldsymbol{x}_i^T \boldsymbol{x}_j \right) \ = \ \Delta_i \ - \ 2 \ \|\boldsymbol{\xi} \ - \ \boldsymbol{x}_i\| \ M \\ &\geq \ \Delta_i \ - \ \frac{2}{\beta \ N} \ , \end{aligned} \tag{134}$$

where we used the assumption (A1) of the lemma.
From the proof in Lemma 4 we have

$$p_{\max} \ = \ [\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})]_i \ \geq \ 1 \ - \ (N-1) \ \exp(- \ \beta \ \tilde{\Delta}_i) \ = \ 1 \ - \ \tilde{\epsilon} \ . \tag{135}$$

Lemma 4 states that

$$\begin{aligned} \|\boldsymbol{x}_i \ - \ f(\boldsymbol{\xi})\| \ &\leqslant \ 2 \ \tilde{\epsilon} \ M \ = \ 2 \ (N-1) \ \exp(- \ \beta \ \tilde{\Delta}_i) \ M \\ &\leqslant \ 2 \ (N-1) \ \exp(- \ \beta \ (\Delta_i \ - \ \frac{2}{\beta \ N})) \ M \ . \end{aligned} \tag{136}$$

We have

$$\begin{aligned} &\|\boldsymbol{x}_i \ - \ f(\boldsymbol{\xi})\| \\ &\leqslant \ 2 \ (N-1) \ \exp(- \ \beta \ (\frac{2}{\beta \ N} \ + \ \frac{1}{\beta} \ \ln \left(2 \ (N-1) \ N \ \beta \ M^2 \right) \ - \ \frac{2}{\beta \ N})) \ M \\ &= \ 2 \ (N-1) \ \exp(- \ \ln \left(2 \ (N-1) \ N \ \beta \ M^2 \right)) \ M \\ &= \ \frac{1}{N \ \beta \ M} \ , \end{aligned} \tag{137}$$

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 6. *Assume that*

A1:

$$\Delta_i \;\geq\; \frac{2}{\beta\, N} \;+\; \frac{1}{\beta}\, \ln\left(2\,(N-1)\,N\,\beta\,M^2\right)\,, \tag{138}$$

then f is a contraction mapping in S_i.

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

$$f(\boldsymbol{\xi}) \;=\; f(\boldsymbol{x}_i) \;+\; \mathrm{J}^m\,(\boldsymbol{\xi} \;-\; \boldsymbol{x}_i)\,. \tag{139}$$

Therefore

$$\|f(\boldsymbol{\xi}) \;-\; f(\boldsymbol{x}_i)\| \;\leqslant\; \|\mathrm{J}^m\|_2\,\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|\,. \tag{140}$$

We define $\tilde{\boldsymbol{\xi}} = \lambda\boldsymbol{\xi} + (1-\lambda)\boldsymbol{x}_i$ for some $\lambda \in [0,1]$. From the proof in Lemma 4 we have

$$p_{\max}(\tilde{\boldsymbol{\xi}}) \;=\; [\mathrm{softmax}(\beta\,\boldsymbol{X}^T\,\tilde{\boldsymbol{\xi}})]_i \;\geq\; 1 \;-\; (N-1)\,\exp(-\,\beta\,\tilde{\Delta}_i) \;=\; 1 \;-\; \tilde{\epsilon}\,, \tag{141}$$

$$\tilde{\epsilon} \;=\; (N-1)\,\exp(-\,\beta\,\tilde{\Delta}_i)\,, \tag{142}$$

$$\tilde{\Delta}_i \;=\; \min_{j,j\neq i}\left(\tilde{\boldsymbol{\xi}}^T\boldsymbol{x}_i \;-\; \tilde{\boldsymbol{\xi}}^T\boldsymbol{x}_j\right)\,. \tag{143}$$

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 $1 \leqslant j \leqslant N$:

$$\left|\tilde{\boldsymbol{\xi}}^T\boldsymbol{x}_j \;-\; \boldsymbol{x}_i^T\boldsymbol{x}_j\right| \leqslant \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{x}_i\right\|\,\|\boldsymbol{x}_j\| \;\leqslant\; \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{x}_i\right\|\,M\,. \tag{144}$$

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

$$\begin{aligned}
\tilde{\Delta}_i \;&\geq\; \min_{j,j\neq i}\left(\left(\boldsymbol{x}_i^T\boldsymbol{x}_i \;-\; \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{x}_i\right\|\,M\right) \;-\; \left(\boldsymbol{x}_i^T\boldsymbol{x}_j \;+\; \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{x}_i\right\|\,M\right)\right) \\
&=\; -\,2\left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{x}_i\right\|\,M \;+\; \min_{j,j\neq i}\left(\boldsymbol{x}_i^T\boldsymbol{x}_i \;-\; \boldsymbol{x}_i^T\boldsymbol{x}_j\right) \;=\; \Delta_i \;-\; 2\left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{x}_i\right\|\,M \\
&\geq\; \Delta_i \;-\; 2\,\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|\,M\,,
\end{aligned} \tag{145}$$

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. (141) we have

$$\begin{aligned}
\tilde{\epsilon} \;&=\; (N-1)\,\exp(-\,\beta\,\tilde{\Delta}_i) \\
&\leqslant\; (N-1)\,\exp\left(-\,\beta\,\left(\Delta_i \;-\; 2\,\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|\,M\right)\right) \\
&\leqslant\; (N-1)\,\exp\left(-\,\beta\,\left(\Delta_i \;-\; \frac{2}{\beta\,N}\right)\right)\,,
\end{aligned} \tag{146}$$

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 $\tilde{\Delta}_i$:

$$\begin{aligned}
\tilde{\Delta}_i \;&\leqslant\; \min_{j,j\neq i}\left(\left(\boldsymbol{x}_i^T\boldsymbol{x}_i \;+\; \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{x}_i\right\|\,M\right) \;-\; \left(\boldsymbol{x}_i^T\boldsymbol{x}_j \;-\; \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{x}_i\right\|\,M\right)\right) \\
&=\; 2\left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{x}_i\right\|\,M \;+\; \min_{j,j\neq i}\left(\boldsymbol{x}_i^T\boldsymbol{x}_i \;-\; \boldsymbol{x}_i^T\boldsymbol{x}_j\right) \;=\; \Delta_i \;+\; 2\left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{x}_i\right\|\,M \\
&\leqslant\; \Delta_i \;+\; 2\,\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|\,M\,,
\end{aligned} \tag{147}$$

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. (141) we have

$$\begin{aligned}
\tilde{\epsilon} \;&=\; (N-1)\,\exp(-\,\beta\,\tilde{\Delta}_i) \\
&\geq\; (N-1)\,\exp\left(-\,\beta\,\left(\Delta_i \;+\; 2\,\|\boldsymbol{\xi} \;-\; \boldsymbol{x}_i\|\,M\right)\right) \\
&\geq\; (N-1)\,\exp\left(-\,\beta\,\left(\Delta_i \;+\; \frac{2}{\beta\,N}\right)\right)\,,
\end{aligned} \tag{148}$$

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 24 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. (50) in Lemma 2:

$$\begin{aligned}\left\| \mathrm{J}(\tilde{\boldsymbol{\xi}}) \right\|_2 &\leqslant 2\,\beta\,N\,M^2\,\tilde{\epsilon} - 2\,\tilde{\epsilon}^2\,\beta\,N\,M^2 \\ &\leqslant 2\,\beta\,N\,M^2\,(N-1)\,\exp\left(-\beta\,\left(\Delta_i - \frac{2}{\beta\,N}\right)\right) - \\ &\quad 2\,(N-1)^2\,\exp\left(-2\,\beta\,\left(\Delta_i + \frac{2}{\beta\,N}\right)\right)\,\beta\,N\,M^2\,.\end{aligned} \tag{149}$$

The bound Eq. (149) holds for the mean J^m, too, since it averages over $\mathrm{J}(\tilde{\boldsymbol{\xi}})$:

$$\begin{aligned}\|\mathrm{J}^m\|_2 &\leqslant 2\,\beta\,N\,M^2\,(N-1)\,\exp\left(-\beta\,\left(\Delta_i - \frac{2}{\beta\,N}\right)\right) - \\ &\quad 2\,(N-1)^2\,\exp\left(-2\,\beta\,\left(\Delta_i + \frac{2}{\beta\,N}\right)\right)\,\beta\,N\,M^2\,.\end{aligned} \tag{150}$$

The assumption of the lemma is

$$\Delta_i \geq \frac{2}{\beta\,N} + \frac{1}{\beta}\,\ln\left(2\,(N-1)\,N\,\beta\,M^2\right)\,, \tag{151}$$

This is

$$\Delta_i - \frac{2}{\beta\,N} \geq \frac{1}{\beta}\,\ln\left(2\,(N-1)\,N\,\beta\,M^2\right)\,, \tag{152}$$

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

$$\begin{aligned}&\|\mathrm{J}^m\|_2 \leqslant 2\,\beta\,(N-1)\,\exp\left(-\beta\,\frac{1}{\beta}\,\ln\left(2\,(N-1)\,N\,\beta\,M^2\right)\right)\,N\,M^2 - \\ &2\,(N-1)^2\,\exp\left(-2\,\beta\,\left(\Delta_i + \frac{2}{\beta\,N}\right)\right)\,\beta\,N\,M^2 \\ &= 2\,\beta\,(N-1)\,\frac{1}{2\,(N-1)\,N\,\beta\,M^2}\,N\,M^2 - \\ &2\,(N-1)^2\,\exp\left(-2\,\beta\,\left(\Delta_i + \frac{2}{\beta\,N}\right)\right)\,\beta\,N\,M^2 \\ &= 1 - 2\,(N-1)^2\,\exp\left(-2\,\beta\,\left(\Delta_i + \frac{2}{\beta\,N}\right)\right)\,\beta\,N\,M^2 < 1\,.\end{aligned} \tag{153}$$

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 7. *Assume that*

A1:

$$\Delta_i \geq \frac{2}{\beta\,N} + \frac{1}{\beta}\,\ln\left(2\,(N-1)\,N\,\beta\,M^2\right)\,, \tag{154}$$

then f has a fixed point in S_i.

Proof. We use Banach fixed point theorem: Lemma 5 says that f maps from S_i into S_i. Lemma 6 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, we have with $\mathrm{J}^m = \int_0^1 \mathrm{J}(\lambda \boldsymbol{\xi} + (1-\lambda)\boldsymbol{x}_i^*)\, \mathrm{d}\lambda$:

$$\|f(\boldsymbol{\xi}) \, - \, \boldsymbol{x}_i^*\| \, = \, \|f(\boldsymbol{\xi}) \, - \, f(\boldsymbol{x}_i^*)\| \, \leqslant \, \|\mathrm{J}^m\|_2 \, \|\boldsymbol{\xi} \, - \, \boldsymbol{x}_i^*\| \tag{155}$$

According to Lemma 24, 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

$$\|\mathrm{J}_s(\tilde{\boldsymbol{x}})\|_2 \, < \, 2 \, \epsilon \, \beta \, . \tag{156}$$

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

$$\|\mathrm{J}(\tilde{\boldsymbol{x}})\|_2 \, \leqslant \, 2 \, \beta \, \|\boldsymbol{X}\|_2^2 \, \epsilon \, \leqslant \, 2 \, \beta \, NM^2 \, \epsilon \, . \tag{157}$$

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:

$$\|\boldsymbol{X}\|_2 \, \leqslant \, \|\boldsymbol{X}\|_F \, = \, \sqrt{\sum_i \|\boldsymbol{x}_i\|^2} \, . \tag{158}$$

Therefore

$$\|\boldsymbol{X}\|_2^2 \, \leqslant \, N \, M^2 \, . \tag{159}$$

The norm of Jacobian of the fixed point iteration is bounded

$$\|\mathrm{J}^m\|_2 \, \leqslant \, 2 \, \beta \, \|\boldsymbol{X}\|_2^2 \, \epsilon \, \leqslant \, 2 \, \beta \, NM^2 \, \epsilon \, . \tag{160}$$

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

$$\Delta_i \, = \, \min_{j, j \neq i} \left(\boldsymbol{x}_i^T \boldsymbol{x}_i \, - \, \boldsymbol{x}_i^T \boldsymbol{x}_j \right) \, = \, \boldsymbol{x}_i^T \boldsymbol{x}_i \, - \, \max_{j, j \neq i} \boldsymbol{x}_i^T \boldsymbol{x}_j \, . \tag{161}$$

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

$$\tilde{\Delta}_i \, = \, \min_{j, j \neq i} \left(\tilde{\boldsymbol{x}}^T \boldsymbol{x}_i \, - \, \tilde{\boldsymbol{x}}^T \boldsymbol{x}_j \right) \, . \tag{162}$$

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

$$\left| \tilde{\boldsymbol{x}}^T \boldsymbol{x}_j \, - \, \boldsymbol{x}_i^T \boldsymbol{x}_j \right| \leqslant \, \|\tilde{\boldsymbol{x}} \, - \, \boldsymbol{x}_i\| \, \|\boldsymbol{x}_j\| \, \leqslant \, \|\tilde{\boldsymbol{x}} \, - \, \boldsymbol{x}_i\| \, M \, . \tag{163}$$

We have the lower bound

$$\begin{aligned} \tilde{\Delta}_i \, &\geq \, \min_{j, j \neq i} \left(\left(\boldsymbol{x}_i^T \boldsymbol{x}_i \, - \, \|\tilde{\boldsymbol{x}} \, - \, \boldsymbol{x}_i\| \, M \right) \, - \, \left(\boldsymbol{x}_i^T \boldsymbol{x}_j \, + \, \|\tilde{\boldsymbol{x}} \, - \, \boldsymbol{x}_i\| \, M \right) \right) \\ &= \, - \, 2 \, \|\tilde{\boldsymbol{x}} \, - \, \boldsymbol{x}_i\| \, M \, + \, \min_{j, j \neq i} \left(\boldsymbol{x}_i^T \boldsymbol{x}_i \, - \, \boldsymbol{x}_i^T \boldsymbol{x}_j \right) \, = \, \Delta_i \, - \, 2 \, \|\tilde{\boldsymbol{x}} \, - \, \boldsymbol{x}_i\| \, M \, . \end{aligned} \tag{164}$$

Since

$$\begin{aligned} \|\tilde{\boldsymbol{x}} \, - \, \boldsymbol{x}_i\| \, &= \, \|\lambda \boldsymbol{\xi} + (1-\lambda)\boldsymbol{x}_i^* \, - \, \boldsymbol{x}_i\| \\ &\leqslant \, \lambda \, \|\boldsymbol{\xi} \, - \, \boldsymbol{x}_i\| \, + \, (1-\lambda) \, \|\boldsymbol{x}_i^* \, - \, \boldsymbol{x}_i\| \\ &\leqslant \, \max\{\|\boldsymbol{\xi} \, - \, \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \, - \, \boldsymbol{x}_i\|\} \, , \end{aligned} \tag{165}$$

we have

$$\tilde{\Delta}_i \, \geq \, \Delta_i \, - \, 2 \, \max\{\|\boldsymbol{\xi} \, - \, \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \, - \, \boldsymbol{x}_i\|\} \, M \, . \tag{166}$$

For the softmax component i we have:

$$\begin{aligned} &[\mathrm{softmax}(\beta \, \boldsymbol{X}^T \tilde{\boldsymbol{\xi}})]_i \, = \, \frac{1}{1 \, + \, \sum_{j \neq i} \exp(\beta \, (\tilde{\boldsymbol{\xi}}^T \boldsymbol{x}_j \, - \, \tilde{\boldsymbol{\xi}}^T \boldsymbol{x}_i))} \\ &\geq \, \frac{1}{1 \, + \, \sum_{j \neq i} \exp(- \, \beta \, (\Delta_i \, - \, 2 \, \max\{\|\boldsymbol{\xi} \, - \, \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \, - \, \boldsymbol{x}_i\|\} \, M))} \\ &= \, \frac{1}{1 \, + \, (N-1) \exp(- \, \beta \, (\Delta_i \, - \, 2 \, \max\{\|\boldsymbol{\xi} \, - \, \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \, - \, \boldsymbol{x}_i\|\} \, M))} \\ &= \, 1 \, - \, \frac{(N-1) \exp(- \, \beta \, (\Delta_i \, - \, 2 \, \max\{\|\boldsymbol{\xi} \, - \, \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \, - \, \boldsymbol{x}_i\|\} \, M))}{1 \, + \, (N-1) \exp(- \, \beta \, (\Delta_i \, - \, 2 \, \max\{\|\boldsymbol{\xi} \, - \, \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \, - \, \boldsymbol{x}_i\|\} \, M))} \\ &\geq \, 1 \, - \, (N-1) \exp(- \, \beta \, (\Delta_i \, - \, 2 \, \max\{\|\boldsymbol{\xi} \, - \, \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \, - \, \boldsymbol{x}_i\|\} \, M)) \\ &= \, 1 \, - \, \epsilon \, . \end{aligned} \tag{167}$$

Therefore

$$\epsilon \;=\; (N-1)\exp(-\,\beta\,(\Delta_i \;-\; 2\;\max\{\|\boldsymbol{\xi}\;-\;\boldsymbol{x}_i\|, \|\boldsymbol{x}_i^*\;-\;\boldsymbol{x}_i\|\}\;M))\,. \tag{168}$$

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

$$\|\mathrm{J}^m\|_2 \;\leqslant\; 2\,\beta\,N\,M^2\,(N-1)\exp(-\,\beta\,(\Delta_i \;-\; 2\;\max\{\|\boldsymbol{\xi}\;-\;\boldsymbol{x}_i\|, \|\boldsymbol{x}_i^*\;-\;\boldsymbol{x}_i\|\}\;M))\,. \tag{169}$$

For a contraction mapping we require

$$\|\mathrm{J}^m\|_2 \;<\; 1\,, \tag{170}$$

which can be ensured by

$$2\,\beta\,NM^2\,(N-1)\exp(-\,\beta\,(\Delta_i \;-\; 2\;\max\{\|\boldsymbol{\xi}\;-\;\boldsymbol{x}_i\|, \|\boldsymbol{x}_i^*\;-\;\boldsymbol{x}_i\|\}\;M)) \;<\; 1\,. \tag{171}$$

Solving this inequality for Δ_i gives

$$\Delta_i \;>\; 2\;\max\{\|\boldsymbol{\xi}\;-\;\boldsymbol{x}_i\|, \|\boldsymbol{x}_i^*\;-\;\boldsymbol{x}_i\|\}\;M \;+\; \frac{1}{\beta}\;\ln\left(2\,(N-1)\,N\,\beta\,M^2\right)\,. \tag{172}$$

In an environment around $\boldsymbol{x}_i^*$ in which Eq. (172) 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 $\boldsymbol{x}_i$:

$$\|f(\boldsymbol{\xi})\;-\;\boldsymbol{x}_i^*\| \;\leqslant\; \|\mathrm{J}^m\|_2\,\|\boldsymbol{\xi}\;-\;\boldsymbol{x}_i^*\| \;<\; \|\boldsymbol{\xi}\;-\;\boldsymbol{x}_i^*\|\,. \tag{173}$$

Using

$$\|f(\boldsymbol{\xi})\;-\;\boldsymbol{x}_i^*\| \;\leqslant\; \|\mathrm{J}^m\|_2\,\|\boldsymbol{\xi}\;-\;\boldsymbol{x}_i^*\| \;\leqslant\; \|\mathrm{J}^m\|_2\,\|\boldsymbol{\xi}\;-\;f(\boldsymbol{\xi})\| \;+\; \|\mathrm{J}^m\|_2\,\|f(\boldsymbol{\xi})\;-\;\boldsymbol{x}_i^*\|\,, \tag{174}$$

we obtain

$$\|f(\boldsymbol{\xi})\;-\;\boldsymbol{x}_i^*\| \;\leqslant\; \frac{\|\mathrm{J}^m\|_2}{1\;-\;\|\mathrm{J}^m\|_2}\,\|\boldsymbol{\xi}\;-\;f(\boldsymbol{\xi})\|\,. \tag{175}$$

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

B2.4.4 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

$$\mathrm{J} \;=\; \beta\,\boldsymbol{X}\left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)\boldsymbol{X}^T \;=\; \boldsymbol{X}\mathrm{J}_s\boldsymbol{X}^T\,. \tag{176}$$

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

$$\boldsymbol{X}\left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)\boldsymbol{X}^T \tag{177}$$

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

$$\boldsymbol{X}\left(\mathrm{diag}(\boldsymbol{p}) \;-\; \boldsymbol{p}\boldsymbol{p}^T\right)\boldsymbol{X}^T \;=\; \boldsymbol{X}\mathrm{diag}(\boldsymbol{p})\boldsymbol{X}^T \;-\; \boldsymbol{X}\boldsymbol{p}\boldsymbol{p}^T\boldsymbol{X}^T \tag{178}$$

$$=\; \sum_{i=1}^N p_i\,\boldsymbol{x}_i\,\boldsymbol{x}_i^T \;-\; \left(\sum_{i=1}^N p_i\,\boldsymbol{x}_i\right)\left(\sum_{i=1}^N p_i\,\boldsymbol{x}_i\right)^T \tag{179}$$

$$=\; \mathrm{E}_{\boldsymbol{p}}[\boldsymbol{x}\,\boldsymbol{x}^T] \;-\; \mathrm{E}_{\boldsymbol{p}}[\boldsymbol{x}]\,\mathrm{E}_{\boldsymbol{p}}[\boldsymbol{x}]^T \;=\; \mathrm{Var}_{\boldsymbol{p}}[\boldsymbol{x}]\,, \tag{180}$$

therefore we have

$$\mathrm{J} \;=\; \beta\,\mathrm{Var}_{\boldsymbol{p}}[\boldsymbol{x}]\,. \tag{181}$$

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. (46). 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:

$$M := \max_i \|\boldsymbol{x}_i\| \,, \tag{182}$$

$$\gamma = \sum_{i=l+1}^{N} p_i \leqslant \epsilon \,, \tag{183}$$

$$1-\gamma = \sum_{i=1}^{l} p_i \geq 1 - \epsilon \,, \tag{184}$$

$$\tilde{p}_i := \frac{p_i}{1-\gamma} \leqslant p_i/(1-\epsilon) \,, \tag{185}$$

$$\sum_{i=1}^{l} \tilde{p}_i = 1 \,, \tag{186}$$

$$\boldsymbol{m}_{\boldsymbol{x}} = \frac{1}{l} \sum_{i=1}^{l} \boldsymbol{x}_i \,, \tag{187}$$

$$m_{\max} = \max_{1 \leqslant i \leqslant l} \|\boldsymbol{x}_i - \boldsymbol{m}_{\boldsymbol{x}}\| \,. \tag{188}$$

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

$$\begin{aligned} \mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}] &= \sum_{i=1}^{l} \tilde{p}_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T - \left(\sum_{i=1}^{l} \tilde{p}_i \, \boldsymbol{x}_i \right) \left(\sum_{i=1}^{l} \tilde{p}_i \, \boldsymbol{x}_i \right)^T \\ &= \sum_{i=1}^{l} \tilde{p}_i \left(\boldsymbol{x}_i - \sum_{i=1}^{l} \tilde{p}_i \boldsymbol{x}_i \right) \left(\boldsymbol{x}_i - \sum_{i=1}^{l} \tilde{p}_i \boldsymbol{x}_i \right)^T . \end{aligned} \tag{189}$$

Lemma 8. *With the definitions in Eq. (182) to Eq. (189), 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*

$$\|\mathrm{J}\|_2 \leqslant \beta \left((1-\gamma) \, m_{\max}^2 + \gamma \, 2 \, (2 - \gamma) \, M^2 \right) \tag{190}$$

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

$$\|\mathrm{J}\|_2 \leqslant \beta \left(m_{\max}^2 + \epsilon \, 2 \, (2 - \epsilon) \, M^2 \right) . \tag{191}$$

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

$$
\begin{aligned}
(1-\gamma)\,\mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}] \;&=\; \sum_{i=1}^{l} p_i \,\left(\boldsymbol{x}_i \;-\; \frac{1}{1-\gamma}\,\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)\,\left(\boldsymbol{x}_i \;-\; \frac{1}{1-\gamma}\,\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)^T \qquad (192)\\
&=\; \sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\,\boldsymbol{x}_i^T \;-\; \left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)\,\frac{1}{1-\gamma}\,\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)^T\\
&-\; \frac{1}{1-\gamma}\,\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)\,\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)^T\\
&+\; \frac{\sum_{i=1}^{l} p_i}{(1-\gamma)^2}\,\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)\,\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)^T\\
&=\; \sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\,\boldsymbol{x}_i^T \;-\; \frac{1}{1-\gamma}\,\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)\,\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)^T\\
&=\; \sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\,\boldsymbol{x}_i^T \;-\; \left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)\,\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)^T \;+\; \left(1 \;-\; \frac{1}{1-\gamma}\right)\,\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)\,\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)^T\\
&=\; \sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\,\boldsymbol{x}_i^T \;-\; \left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)\,\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)^T \;-\; \frac{\gamma}{1-\gamma}\,\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)\,\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)^T\,.
\end{aligned}
$$

Therefore we have

$$
\begin{aligned}
&\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\,\boldsymbol{x}_i^T \;-\; \left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)\,\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)^T \qquad (193)\\
&=\; (1-\gamma)\,\mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}] \;+\; \frac{\gamma}{1-\gamma}\,\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)\,\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)^T\,.
\end{aligned}
$$

We now can reformulate the Jacobian J:

$$\begin{aligned}
\mathrm{J} \;=\; & \beta \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T \;+\; \sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T \right. \\
& \left. - \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \;+\; \sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \;+\; \sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right)^T \right) \\
= \; & \beta \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T \;-\; \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right)^T \right. \\
& + \; \sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T \;-\; \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right)^T \\
& \left. - \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right)^T \;-\; \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right)^T \right) \\
= \; & \beta \left((1-\gamma) \, \mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}] \;+\; \frac{\gamma}{1-\gamma} \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right)^T \right. \\
& + \; \sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T \;-\; \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right)^T \\
& \left. - \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right)^T \;-\; \left(\sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right) \left(\sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right)^T \right) \; .
\end{aligned} \tag{194}$$

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

$$\left\| \boldsymbol{a} \boldsymbol{b}^T \right\|_2 \;=\; \sqrt{\lambda_{\max}(\boldsymbol{b} \boldsymbol{a}^T \boldsymbol{a} \boldsymbol{b}^T)} \;=\; \|\boldsymbol{a}\| \, \sqrt{\lambda_{\max}(\boldsymbol{b} \boldsymbol{b}^T)} \;=\; \|\boldsymbol{a}\| \, \|\boldsymbol{b}\| \; , \tag{195}$$

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:

$$\left\| \sum_{i=1}^{l} p_i \, \boldsymbol{x}_i \right\| \;\leqslant\; \sum_{i=1}^{l} p_i \, \|\boldsymbol{x}_i\| \;\leqslant\; (1-\gamma) \, M \; , \tag{196}$$

$$\left\| \sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \right\| \;\leqslant\; \sum_{i=l+1}^{N} p_i \, \|\boldsymbol{x}_i\| \;\leqslant\; \gamma \, M \; , \tag{197}$$

$$\left\| \sum_{i=l+1}^{N} p_i \, \boldsymbol{x}_i \, \boldsymbol{x}_i^T \right\|_2 \;\leqslant\; \sum_{i=l+1}^{N} p_i \left\| \boldsymbol{x}_i \, \boldsymbol{x}_i^T \right\|_2 \;=\; \sum_{i=l+1}^{N} p_i \, \|\boldsymbol{x}_i\|^2 \;\leqslant\; \sum_{i=l+1}^{N} p_i \, M^2 \;=\; \gamma \, M^2 \; . \tag{198}$$

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

$$f(\boldsymbol{a}) \;=\; \sum_{i=1}^{l} p_i \|\boldsymbol{x}_i \;-\; \boldsymbol{a}\|^2 \;=\; \sum_{i=1}^{l} p_i (\boldsymbol{x}_i \;-\; \boldsymbol{a})^T (\boldsymbol{x}_i \;-\; \boldsymbol{a}) \; . \tag{199}$$

The solution to

$$\frac{\partial f(\boldsymbol{a})}{\partial \boldsymbol{a}} \;=\; 2 \sum_{i=1}^{N} p_i (\boldsymbol{a} \;-\; \boldsymbol{x}_i) \;=\; 0 \tag{200}$$

is

$$\boldsymbol{a} \;=\; \sum_{i=1}^{N} p_i \boldsymbol{x}_i \; . \tag{201}$$

The Hessian of f is positive definite since

$$\frac{\partial^2 f(\boldsymbol{a})}{\partial \boldsymbol{a}^2} \;=\; 2\,\sum_{i=1}^{N} p_i\,\boldsymbol{I} \;=\; 2\,\boldsymbol{I} \tag{202}$$

and f is a convex function. Hence, the mean

$$\bar{\boldsymbol{x}} \;:=\; \sum_{i=1}^{N} p_i\,\boldsymbol{x}_i \tag{203}$$

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

$$\sum_{i=1}^{l} p_i \|\boldsymbol{x}_i \;-\; \bar{\boldsymbol{x}}\|^2 \;\leqslant\; \sum_{i=1}^{l} p_i \|\boldsymbol{x}_i \;-\; \boldsymbol{m}_{\boldsymbol{x}}\|^2 \;\leqslant\; (1 \;-\; \gamma)\, m_{\max}^2\,. \tag{204}$$

We now bound the variance on the first l patterns:

$$\begin{aligned}(1-\gamma)\,\|\mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}]\|_2 \;&\leqslant\; \sum_{i=1}^{l} p_i \Big\| (\boldsymbol{x}_i \;-\; \bar{\boldsymbol{x}})\;(\boldsymbol{x}_i \;-\; \bar{\boldsymbol{x}})^T \Big\|_2 \\ &=\; \sum_{i=1}^{l} p_i \|\boldsymbol{x}_i \;-\; \bar{\boldsymbol{x}}\|^2 \;\leqslant\; \sum_{i=1}^{l} p_i \|\boldsymbol{x}_i \;-\; \boldsymbol{m}_{\boldsymbol{x}}\|^2 \;\leqslant\; (1 \;-\; \gamma)\, m_{\max}^2\,.\end{aligned} \tag{205}$$

We obtain for the spectral norm of J:

$$\begin{aligned}\|\mathrm{J}\|_2 \;&\leqslant\; \beta\,\big((1-\gamma)\,\|\mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}]\|_2 \\ &+\; \frac{\gamma}{1-\gamma}\left\| \left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)\;\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)^T \right\|_2 \\ &+\; \left\| \sum_{i=l+1}^{N} p_i\,\boldsymbol{x}_i\,\boldsymbol{x}_i^T \right\|_2 \;+\; \left\| \left(\sum_{i=l+1}^{N} p_i\,\boldsymbol{x}_i\right)\;\left(\sum_{i=l+1}^{N} p_i\,\boldsymbol{x}_i\right)^T \right\|_2 \\ &+\; \left\| \left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)\;\left(\sum_{i=l+1}^{N} p_i\,\boldsymbol{x}_i\right)^T \right\|_2 \;+\; \left\| \left(\sum_{i=l+1}^{N} p_i\,\boldsymbol{x}_i\right)\left(\sum_{i=1}^{l} p_i\,\boldsymbol{x}_i\right)^T \right\|_2 \Big) \\ &\leqslant\; \beta\,\big((1-\gamma)\,\|\mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}]\|_2 \;+\; \gamma\,(1-\gamma)\,M^2 \;+\; \gamma\,M^2 \;+\; \gamma^2\,M^2 \;+ \\ &\gamma\,(1-\gamma)\,M^2 \;+\; \gamma\,(1-\gamma)\,M^2\big) \\ &=\; \beta\,\big((1-\gamma)\;\|\mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}]\|_2 \;+\; \gamma\,2\,(2 \;-\; \gamma)\,M^2\big)\;.\end{aligned} \tag{206}$$

Combining the previous two estimates immediately leads to Eq. (190).
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 $\|\mathrm{J}\|_2$:

$$\|\mathrm{J}\|_2 \;\leqslant\; \beta\,\big(\,m_{\max}^2 \;+\; \epsilon\,2\,(2 \;-\; \epsilon)\,M^2\big)\;. \tag{207}$$

□

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:

$$\begin{aligned}\|\mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}]\|_2 \;&=\; \mathrm{Tr}\left(\sum_{i=1}^{l} p_i\,(\boldsymbol{x}_i \;-\; \bar{\boldsymbol{x}})\;(\boldsymbol{x}_i \;-\; \bar{\boldsymbol{x}})^T\right) \;-\; \sum_{k=2}^{d} e_k \\ &=\; \sum_{i=1}^{l} p_i \mathrm{Tr}\left((\boldsymbol{x}_i \;-\; \bar{\boldsymbol{x}})\;(\boldsymbol{x}_i \;-\; \bar{\boldsymbol{x}})^T\right) \;-\; \sum_{k=2}^{d} e_k \\ &=\; \sum_{i=1}^{l} p_i \|\boldsymbol{x}_i \;-\; \bar{\boldsymbol{x}}\|^2 \;-\; \sum_{k=2}^{d} e_k\,.\end{aligned} \tag{208}$$

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:

$$M \ := \ \max_i \|\boldsymbol{x}_i\| \ , \tag{209}$$

$$\gamma \ = \ \sum_{i=l+1}^{N} p_i \ \leqslant \ \epsilon \ , \tag{210}$$

$$1-\gamma \ = \ \sum_{i=1}^{l} p_i \ \geq \ 1 \ - \ \epsilon \ , \tag{211}$$

$$\tilde{p}_i \ := \ \frac{p_i}{1-\gamma} \ \leqslant \ p_i/(1-\epsilon) \ , \tag{212}$$

$$\sum_{i=1}^{l} \tilde{p}_i \ = \ 1 \ , \tag{213}$$

$$\boldsymbol{m}_{\boldsymbol{x}} \ = \ \frac{1}{l} \ \sum_{i=1}^{l} \boldsymbol{x}_i \ , \tag{214}$$

$$m_{\max} \ = \ \max_{1 \leqslant i \leqslant l} \|\boldsymbol{x}_i \ - \ \boldsymbol{m}_{\boldsymbol{x}}\| \ . \tag{215}$$

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

$$0 \ \leqslant \ c \ = \ \min_{j,l<j} \left(\boldsymbol{\xi}^T \boldsymbol{m}_{\boldsymbol{x}} \ - \ \boldsymbol{\xi}^T \boldsymbol{x}_j \right) \ = \ \boldsymbol{\xi}^T \boldsymbol{m}_{\boldsymbol{x}} \ - \ \max_{j,l<j} \boldsymbol{\xi}^T \boldsymbol{x}_j \ . \tag{216}$$

The following holds:

$$\|\boldsymbol{m}_{\boldsymbol{x}} \ - \ f(\boldsymbol{\xi})\| \ \leqslant \ m_{\max} \ + \ 2 \, \gamma \, M \ \leqslant \ m_{\max} \ + \ 2 \, \epsilon \, M \ , \tag{217}$$

where

$$M \ = \ \max_i \|\boldsymbol{x}_i\| \ , \tag{218}$$

$$\epsilon \ = \ (N-l) \ \exp(-\beta \, c) \ . \tag{219}$$

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

$$[\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})]_j \ = \ \frac{\exp(\beta \ (\boldsymbol{\xi}^T \boldsymbol{x}_j \ - \ \boldsymbol{\xi}^T \boldsymbol{x}_s))}{1 \ + \ \sum_{k,k \neq s} \exp(\beta \ (\boldsymbol{\xi}^T \boldsymbol{x}_k \ - \ \boldsymbol{\xi}^T \boldsymbol{x}_s))} \ \leqslant \ \exp(-\beta \, c) \ = \ \frac{\epsilon}{N-l} \ , \tag{220}$$

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

$$f(\boldsymbol{\xi}) \ = \ \boldsymbol{X} \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) \ = \ \sum_{j=1}^{N} \boldsymbol{x}_j \ [\mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi})]_j \ . \tag{221}$$

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

$$\left\| \boldsymbol{m}_{\boldsymbol{x}} - \sum_{j=1}^{l} \frac{p_j}{1-\gamma} \boldsymbol{x}_j \right\|^2 = \left\| \sum_{j=1}^{l} \frac{p_j}{1-\gamma} (\boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{x}_j) \right\|^2 \tag{222}$$

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

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

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

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

It follows that

$$\left\| \boldsymbol{m}_{\boldsymbol{x}} - \sum_{j=1}^{l} \frac{p_j}{1-\gamma} \boldsymbol{x}_j \right\| \leqslant m_{\max} \tag{223}$$

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

$$\begin{aligned}
\|\boldsymbol{m}_{\boldsymbol{x}} - f(\boldsymbol{\xi})\| &= \left\| \boldsymbol{m}_{\boldsymbol{x}} - \sum_{j=1}^{N} p_j \boldsymbol{x}_j \right\| \\
&= \left\| \boldsymbol{m}_{\boldsymbol{x}} - \sum_{j=1}^{l} p_j \boldsymbol{x}_j - \sum_{j=l+1}^{N} p_j \boldsymbol{x}_j \right\| \\
&= \left\| \boldsymbol{m}_{\boldsymbol{x}} - \sum_{j=1}^{l} \frac{p_j}{1-\gamma} \boldsymbol{x}_j + \frac{\gamma}{1-\gamma} \sum_{j=1}^{l} p_j \boldsymbol{x}_j - \sum_{j=l+1}^{N} p_j \boldsymbol{x}_j \right\| \\
&\leqslant \left\| \boldsymbol{m}_{\boldsymbol{x}} - \sum_{j=1}^{l} \frac{p_j}{1-\gamma} \boldsymbol{x}_j \right\| + \frac{\gamma}{1-\gamma} \left\| \sum_{j=1}^{l} p_j \boldsymbol{x}_j \right\| + \left\| \sum_{j=l+1}^{N} p_j \boldsymbol{x}_j \right\| \\
&\leqslant \left\| \boldsymbol{m}_{\boldsymbol{x}} - \sum_{j=1}^{l} \frac{p_j}{1-\gamma} \boldsymbol{x}_j \right\| + \frac{\gamma}{1-\gamma} \sum_{j=1}^{l} p_j M + \sum_{j=l+1}^{N} p_j M \\
&\leqslant \left\| \boldsymbol{m}_{\boldsymbol{x}} - \sum_{j=1}^{l} \frac{p_j}{1-\gamma} \boldsymbol{x}_j \right\| + 2 \gamma M \\
&\leqslant m_{\max} + 2 \gamma M \leqslant m_{\max} + 2 \epsilon M ,
\end{aligned} \tag{224}$$

where we applied Eq. (222) 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

$$\Delta_m = \min_{j,l<j} \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \right) = \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} - \max_{j,l<j} \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j . \tag{225}$$

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. (129), Δ_m can also be expressed as

$$\begin{aligned} \Delta_m &= \min_{j,l<j} \frac{1}{2} \left(\|\boldsymbol{m}_{\boldsymbol{x}}\|^2 - \|\boldsymbol{x}_j\|^2 + \|\boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{x}_j\|^2 \right) \\ &= \frac{1}{2} \|\boldsymbol{m}_{\boldsymbol{x}}\|^2 - \frac{1}{2} \max_{j,l<j} \left(\|\boldsymbol{x}_j\|^2 - \|\boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{x}_j\|^2 \right) . \end{aligned} \tag{226}$$

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 B3 (Sphere S_m)**.** *The sphere S_m is defined as*

$$\mathrm{S}_m := \left\{ \boldsymbol{\xi} \mid \|\boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}}\| \leqslant \frac{1}{\beta\, m_{\max}} \right\} . \tag{227}$$

Lemma 10. *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 $l < j$:

$$\Delta_m \geq \frac{2\, M}{\beta\, m_{\max}} - \frac{1}{\beta} \ln \left(\frac{1 - \beta\, m_{\max}^2}{2\, \beta\, (N-l)\, M\, \max\{m_{\max},\, 2\, M\}} \right) , \tag{228}$$

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

$$\beta\, m_{\max}^2 \leqslant 1 \tag{229}$$

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 $\boldsymbol{X} = (\boldsymbol{x}_{l+1}, \ldots, \boldsymbol{x}_N)$.

$$\tilde{\Delta}_m = \min_{j,l<j} \left(\boldsymbol{\xi}^T \boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{\xi}^T \boldsymbol{x}_j \right) . \tag{230}$$

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

$$\left| \boldsymbol{\xi}^T \boldsymbol{x}_j - \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \right| \leqslant \|\boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}}\|\, \|\boldsymbol{x}_j\| \leqslant \|\boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}}\|\, M . \tag{231}$$

We have the lower bound

$$\begin{aligned} \tilde{\Delta}_m &\geq \min_{j,l<j} \left(\left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} - \|\boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}}\|\, M \right) - \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j + \|\boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}}\|\, M \right) \right) \\ &= -2\, \|\boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}}\|\, M + \min_{j,l<j} \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \right) = \Delta_m - 2\, \|\boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}}\|\, M \\ &\geq \Delta_m - 2\, \frac{M}{\beta\, m_{\max}} , \end{aligned} \tag{232}$$

where we used the assumption (A1) of the lemma.
From the proof in Lemma 9 we have

$$\sum_{i=1}^{l} p_i \geq 1 - (N-l)\, \exp(-\beta\, \tilde{\Delta}_m) = 1 - \tilde{\epsilon} , \tag{233}$$

$$\sum_{i=l+1}^{N} p_i \leqslant (N-l)\, \exp(-\beta\, \tilde{\Delta}_m) = \tilde{\epsilon} . \tag{234}$$

Lemma 9 states that

$$\begin{aligned} \|\boldsymbol{m}_{\boldsymbol{x}} - f(\boldsymbol{\xi})\| &\leqslant m_{\max} + 2\, \tilde{\epsilon}\, M \\ &\leqslant m_{\max} + 2\, (N-l)\, \exp(-\beta\, \tilde{\Delta}_m)\, M . \\ &\leqslant m_{\max} + 2\, (N-l)\, \exp(-\beta\, (\Delta_m - 2\, \frac{M}{\beta\, m_{\max}}))\, M . \end{aligned} \tag{235}$$

Therefore we have

$$\|\boldsymbol{m}_{\boldsymbol{x}} \,-\, f(\boldsymbol{\xi})\| \,\leqslant\, m_{\max} \,+\, 2\,(N-l)\,\exp\left(-\,\beta\,(\Delta_m \,-\, 2\,\frac{M}{\beta\,m_{\max}})\right)\,M \tag{236}$$

$$\leqslant\, m_{\max} \,+\, 2\,(N-l)\,\exp\left(-\,\beta\,\left(\frac{2\,M}{\beta\,m_{\max}} \,-\right.\right.$$

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

$$=\, m_{\max} \,+\, 2\,(N-l)\,\frac{1\,-\,\beta\,m_{\max}^2}{2\,\beta\,(N-l)\,M\,\max\{m_{\max}\,,\,2\,M\}}\,M$$

$$\leqslant\, m_{\max} \,+\, \frac{1\,-\,\beta\,m_{\max}^2}{\beta\,m_{\max}} \,=\, \frac{1}{\beta\,m_{\max}}\,,$$

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.

$$m_{\max} = \max_{1\leqslant i\leqslant l} \|\boldsymbol{x}_i - \boldsymbol{m}_{\boldsymbol{x}}\| \tag{237}$$

$$= \max_{1\leqslant i\leqslant l} \left\| \boldsymbol{x}_i - 1/l \sum_{j=1}^{l} \boldsymbol{x}_j \right\| \tag{238}$$

$$= \max_{1\leqslant i\leqslant l} \left\| 1/l \sum_{j=1}^{l} (\boldsymbol{x}_i - \boldsymbol{x}_j) \right\| \tag{239}$$

$$\leqslant \max_{1\leqslant i,j\leqslant l} \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \tag{240}$$

$$\leqslant \max_{1\leqslant i\leqslant l} \|\boldsymbol{x}_i\| + \max_{1\leqslant j\leqslant l} \|\boldsymbol{x}_i\| \tag{241}$$

$$\leqslant 2M \tag{242}$$

□

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

Lemma 11. *Assume that*

A1:

$$\Delta_m \,\geq\, \frac{2\,M}{\beta\,m_{\max}} \,-\, \frac{1}{\beta}\,\ln\left(\frac{1\,-\,\beta\,m_{\max}^2}{2\,\beta\,(N-l)\,M\,\max\{m_{\max}\,,\,2\,M\}}\right)\,, \tag{243}$$

and

A2:

$$\beta\,m_{\max}^2 \,\leqslant\, 1\,, \tag{244}$$

then f is a contraction mapping in S_m.

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

$$f(\boldsymbol{\xi}) \,=\, f(\boldsymbol{m}_{\boldsymbol{x}}) \,+\, \mathrm{J}^m\,(\boldsymbol{\xi} \,-\, \boldsymbol{m}_{\boldsymbol{x}})\,. \tag{245}$$

In complete analogy to Lemma 6, we get:

$$\|f(\boldsymbol{\xi}) \,-\, f(\boldsymbol{m}_{\boldsymbol{x}})\| \,\leqslant\, \|\mathrm{J}^m\|_2\,\|\boldsymbol{\xi} \,-\, \boldsymbol{m}_{\boldsymbol{x}}\|\,. \tag{246}$$

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 $\boldsymbol{X} = (\boldsymbol{x}_{l+1}, \ldots, \boldsymbol{x}_N)$.

$$\tilde{\Delta}_m \,=\, \min_{j,l<j} \left(\tilde{\boldsymbol{\xi}}^T\boldsymbol{m}_{\boldsymbol{x}} \,-\, \tilde{\boldsymbol{\xi}}^T\boldsymbol{x}_j\right)\,. \tag{247}$$

From the proof in Lemma 9 we have

$$\tilde{\epsilon} \;=\; (N-l)\,\exp(-\,\beta\,\tilde{\Delta}_m)\,, \tag{248}$$

$$\sum_{i=1}^{l} p_i(\tilde{\boldsymbol{\xi}}) \;\geq\; 1 \;-\; (N-l)\,\exp(-\,\beta\,\tilde{\Delta}_m) \;=\; 1 \;-\; \tilde{\epsilon}\,, \tag{249}$$

$$\sum_{i=l+1}^{N} p_i(\tilde{\boldsymbol{\xi}}) \;\leqslant\; (N-l)\,\exp(-\,\beta\,\tilde{\Delta}_m) \;=\; \tilde{\epsilon}\,. \tag{250}$$

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

$$\left|\tilde{\boldsymbol{\xi}}^T \boldsymbol{x}_j \;-\; \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j\right| \leqslant \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{m}_{\boldsymbol{x}}\right\| \|\boldsymbol{x}_j\| \;\leqslant\; \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{m}_{\boldsymbol{x}}\right\| M\,. \tag{251}$$

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

$$\begin{aligned}
\tilde{\Delta}_m \;&\geq\; \min_{j,l<j} \left(\left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} \;-\; \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{m}_{\boldsymbol{x}}\right\| M\right) \;-\; \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \;+\; \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{m}_{\boldsymbol{x}}\right\| M\right)\right) \\
&=\; -\,2\left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{m}_{\boldsymbol{x}}\right\| M \;+\; \min_{j,l<j} \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} \;-\; \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j\right) \;=\; \Delta_m \;-\; 2\left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{m}_{\boldsymbol{x}}\right\| M \\
&\geq\; \Delta_m \;-\; 2\,\|\boldsymbol{\xi} \;-\; \boldsymbol{m}_{\boldsymbol{x}}\|\,M\,.
\end{aligned} \tag{252}$$

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 $\tilde{\epsilon}$:

$$\begin{aligned}
\tilde{\epsilon} \;&\leqslant\; (N-l)\,\exp\left(-\,\beta\,\left(\Delta_m \;-\; 2\,\|\boldsymbol{\xi} \;-\; \boldsymbol{m}_{\boldsymbol{x}}\|\,M\right)\right) \\
&\leqslant\; (N-l)\,\exp\left(-\,\beta\,\left(\Delta_m \;-\; \frac{2\,M}{\beta\,m_{\max}}\right)\right)\,.
\end{aligned} \tag{253}$$

where we used that in the sphere S_i holds:

$$\|\boldsymbol{\xi} \;-\; \boldsymbol{m}_{\boldsymbol{x}}\| \;\leqslant\; \frac{1}{\beta\,m_{\max}}\,, \tag{254}$$

therefore

$$2\,\|\boldsymbol{\xi} \;-\; \boldsymbol{m}_{\boldsymbol{x}}\|\,M \;\leqslant\; \frac{2\,M}{\beta\,m_{\max}}\,. \tag{255}$$

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. (147) in combination with Eq. (255).

$$\begin{aligned}
\tilde{\Delta}_m \;&\geq\; \min_{j,l<j} \left(\left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} \;+\; \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{m}_{\boldsymbol{x}}\right\| M\right) \;-\; \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \;-\; \left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{m}_{\boldsymbol{x}}\right\| M\right)\right) \\
&=\; 2\left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{m}_{\boldsymbol{x}}\right\| M \;+\; \min_{j,l<j} \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} \;-\; \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j\right) \;=\; \Delta_m \;+\; 2\left\|\tilde{\boldsymbol{\xi}} \;-\; \boldsymbol{m}_{\boldsymbol{x}}\right\| M \\
&\geq\; \Delta_m \;+\; 2\,\|\boldsymbol{\xi} \;-\; \boldsymbol{m}_{\boldsymbol{x}}\|\,M\,.
\end{aligned} \tag{256}$$

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 $\tilde{\epsilon}$:

$$\tilde{\epsilon} \;\geq\; (N-l)\,\exp\left(-\,\beta\,\left(\Delta_m \;+\; \frac{2\,M}{\beta\,m_{\max}}\right)\right)\,, \tag{257}$$

where we used that in the sphere S_i holds:

$$\|\boldsymbol{\xi} \;-\; \boldsymbol{m}_{\boldsymbol{x}}\| \;\leqslant\; \frac{1}{\beta\,m_{\max}}\,, \tag{258}$$

therefore

$$2\,\|\boldsymbol{\xi} \;-\; \boldsymbol{m}_{\boldsymbol{x}}\|\,M \;\leqslant\; \frac{2\,M}{\beta\,m_{\max}}\,. \tag{259}$$

From Lemma 8 we have

$$\begin{aligned}\left\|\mathrm{J}(\tilde{\boldsymbol{\xi}})\right\|_2 &\leqslant \beta\left(m_{\max}^2 + \tilde{\epsilon}\,2\,(2 - \tilde{\epsilon})\,M^2\right) \\ &= \beta\left(m_{\max}^2 + \tilde{\epsilon}4\,M^2 - 2\,\tilde{\epsilon}^2\,M^2\right) \\ &\leqslant \beta\left(m_{\max}^2 + (N-l)\,\exp\left(-\beta\,\left(\Delta_m - \frac{2\,M}{\beta\,m_{\max}}\right)\right)4\,M^2 - \right. \\ &\left. 2\,(N-l)^2\,\exp\left(-2\,\beta\,\left(\Delta_m + \frac{2\,M}{\beta\,m_{\max}}\right)\right)M^2\right)\,.\end{aligned} \tag{260}$$

The bound Eq. (260) holds for the mean J^m, too, since it averages over $\mathrm{J}(\tilde{\boldsymbol{\xi}})$:

$$\begin{aligned}\|\mathrm{J}^m\|_2 \leqslant \beta\left(m_{\max}^2 + (N-l)\,\exp\left(-\beta\,\left(\Delta_m - \frac{2\,M}{\beta\,m_{\max}}\right)\right)4\,M^2 - \right. \\ \left. 2\,(N-l)^2\,\exp\left(-2\,\beta\,\left(\Delta_m + \frac{2\,M}{\beta\,m_{\max}}\right)\right)M^2\right)\,.\end{aligned} \tag{261}$$

The assumption of the lemma is

$$\Delta_m \geq \frac{2\,M}{\beta\,m_{\max}} - \frac{1}{\beta}\,\ln\left(\frac{1 - \beta\,m_{\max}^2}{2\,\beta\,(N-l)\,M\,\max\{m_{\max}\,,\,2\,M\}}\right)\,, \tag{262}$$

Therefore we have

$$\Delta_m - \frac{2\,M}{\beta\,m_{\max}} \geq -\frac{1}{\beta}\,\ln\left(\frac{1 - \beta\,m_{\max}^2}{2\,\beta\,(N-l)\,M\,\max\{m_{\max}\,,\,2\,M\}}\right)\,. \tag{263}$$

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

$$\begin{aligned}&\|\mathrm{J}^m\|_2 \leqslant \\ &\beta\left(m_{\max}^2 + (N-l)\,\exp\left(-\beta\,\left(-\frac{1}{\beta}\,\ln\left(\frac{1 - \beta\,m_{\max}^2}{2\,\beta\,(N-l)\,M\,\max\{m_{\max}\,,\,2\,M\}}\right)\right)\right)\right. \\ &\left. 4\,M^2 - 2\,(N-l)^2\,\exp\left(-2\,\beta\,\left(\Delta_m + \frac{2\,M}{\beta\,m_{\max}}\right)\right)M^2\right) \\ &= \beta\left(m_{\max}^2 + (N-l)\,\exp\left(\ln\left(\frac{1 - \beta\,m_{\max}^2}{2\,\beta\,(N-l)\,M\,\max\{m_{\max}\,,\,2\,M\}}\right)\right)\right. \\ &\left. 4\,M^2 - 2\,(N-l)^2\,\exp\left(-2\,\beta\,\left(\Delta_m + \frac{2\,M}{\beta\,m_{\max}}\right)\right)M^2\right) \\ &= \beta\left(m_{\max}^2 + (N-l)\,\frac{1 - \beta\,m_{\max}^2}{2\,\beta\,(N-l)\,M\,\max\{m_{\max}\,,\,2\,M\}}\,4\,M^2 - \right. \\ &\left. 2\,(N-l)^2\,\exp\left(-2\,\beta\,\left(\Delta_m + \frac{2\,M}{\beta\,m_{\max}}\right)\right)M^2\right) \\ &= \beta m_{\max}^2 + \frac{1 - \beta\,m_{\max}^2}{\max\{m_{\max}\,,\,2\,M\}}\,2\,M - \\ &\beta\,2\,(N-l)^2\,\exp\left(-2\,\beta\,\left(\Delta_m + \frac{2\,M}{\beta\,m_{\max}}\right)\right)M^2 \\ &\leqslant \beta m_{\max}^2 + 1 - \beta\,m_{\max}^2 - \beta\,2\,(N-l)^2\,\exp\left(-2\,\beta\,\left(\Delta_m + \frac{2\,M}{\beta\,m_{\max}}\right)\right)M^2 \\ &= 1 - \beta\,2\,(N-l)^2\,\exp\left(-2\,\beta\,\left(\Delta_m + \frac{2\,M}{\beta\,m_{\max}}\right)\right)M^2 < 1\,.\end{aligned} \tag{264}$$

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 12. *Assume that*

A1:

$$\Delta_m \ \geq \ \frac{2\,M}{\beta\,m_{\max}} \ - \ \frac{1}{\beta}\,\ln\left(\frac{1\ -\ \beta\,m_{\max}^2}{2\,\beta\,(N-l)\,M\ \max\{m_{\max}\,,\,2\,M\}}\right)\,, \tag{265}$$

and

A2:

$$\beta\,m_{\max}^2 \ \leqslant \ 1\,, \tag{266}$$

then f has a fixed point in S_m.

Proof. We use Banach fixed point theorem: Lemma 10 says that f maps from the compact set S_m into the same compact set S_m. Lemma 11 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:

$$M \ := \ \max_i \|\boldsymbol{x}_i\|\,, \tag{267}$$

$$\gamma \ = \ \sum_{i=l+1}^{N} p_i \ \leqslant \ \epsilon\,, \tag{268}$$

$$1-\gamma \ = \ \sum_{i=1}^{l} p_i \ \geq \ 1 \ - \ \epsilon\,, \tag{269}$$

$$\tilde{p}_i \ := \ \frac{p_i}{1-\gamma} \ \leqslant \ p_i/(1-\epsilon)\,, \tag{270}$$

$$\sum_{i=1}^{l} \tilde{p}_i \ = \ 1\,, \tag{271}$$

$$\boldsymbol{m}_{\boldsymbol{x}} \ = \ \frac{1}{l}\,\sum_{i=1}^{l} \boldsymbol{x}_i\,, \tag{272}$$

$$m_{\max} \ = \ \max_{1\leqslant i\leqslant l} \|\boldsymbol{x}_i \ - \ \boldsymbol{m}_{\boldsymbol{x}}\|\,. \tag{273}$$

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

$$\begin{aligned}\mathrm{Var}_{\tilde{p}}[\boldsymbol{x}_{1:l}] \ &= \ \sum_{i=1}^{l} \tilde{p}_i\,\boldsymbol{x}_i\,\boldsymbol{x}_i^T \ - \ \left(\sum_{i=1}^{l} \tilde{p}_i\,\boldsymbol{x}_i\right)\,\left(\sum_{i=1}^{l} \tilde{p}_i\,\boldsymbol{x}_i\right)^T \\ &= \ \sum_{i=1}^{l} \tilde{p}_i\,\left(\boldsymbol{x}_i \ - \ \sum_{i=1}^{l} \tilde{p}_i \boldsymbol{x}_i\right)\,\left(\boldsymbol{x}_i \ - \ \sum_{i=1}^{l} \tilde{p}_i \boldsymbol{x}_i\right)^T\,.\end{aligned} \tag{274}$$

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, we have with $\mathrm{J}^m = \int_0^1 \mathrm{J}(\lambda\boldsymbol{\xi} + (1-\lambda)\boldsymbol{m}_{\boldsymbol{x}}^*)\,\mathrm{d}\lambda$:

$$\|f(\boldsymbol{\xi}) \ - \ \boldsymbol{m}_{\boldsymbol{x}}^*\| \ = \ \|f(\boldsymbol{\xi}) \ - \ f(\boldsymbol{m}_{\boldsymbol{x}}^*)\| \ \leqslant \ \|\mathrm{J}^m\|_2\,\|\boldsymbol{\xi} \ - \ \boldsymbol{m}_{\boldsymbol{x}}^*\| \tag{275}$$

According to Lemma 8 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

$$\|\mathrm{J}\|_2 \ \leqslant \ \beta\left((1-\gamma)\,m_{\max}^2 \ + \ \gamma\,2\,(2 \ - \ \gamma)\,M^2\right)\,, \tag{276}$$

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

$$\|\mathrm{J}\|_2 \leqslant \beta \left(m_{\max}^2 + \epsilon \, 2 \, (2 - \epsilon) \, M^2 \right) . \tag{277}$$

From the last condition we require for a contraction mapping:

$$\beta \; m_{\max}^2 < 1 . \tag{278}$$

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

$$\Delta_m = \min_{j,l<j} \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \right) = \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} - \max_{j,l<j} \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j . \tag{279}$$

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

$$\tilde{\Delta}_m = \min_{j,l<j} \left(\tilde{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} - \tilde{\boldsymbol{x}}^T \boldsymbol{x}_j \right) . \tag{280}$$

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

$$\left| \tilde{\boldsymbol{x}}^T \boldsymbol{x}_j - \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \right| \leqslant \| \tilde{\boldsymbol{x}} - \boldsymbol{m}_{\boldsymbol{x}} \| \, \| \boldsymbol{x}_j \| \leqslant \| \tilde{\boldsymbol{x}} - \boldsymbol{m}_{\boldsymbol{x}} \| \, M . \tag{281}$$

We have the lower bound

$$\begin{aligned} \tilde{\Delta}_m &\geq \min_{j,l<j} \left(\left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} - \| \tilde{\boldsymbol{x}} - \boldsymbol{m}_{\boldsymbol{x}} \| \, M \right) - \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j + \| \tilde{\boldsymbol{x}} - \boldsymbol{m}_{\boldsymbol{x}} \| \, M \right) \right) \\ &= -2 \, \| \tilde{\boldsymbol{x}} - \boldsymbol{m}_{\boldsymbol{x}} \| \, M + \min_{j,l<j} \left(\boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{m}_{\boldsymbol{x}} - \boldsymbol{m}_{\boldsymbol{x}}^T \boldsymbol{x}_j \right) = \Delta_m - 2 \, \| \tilde{\boldsymbol{x}} - \boldsymbol{m}_{\boldsymbol{x}} \| \, M . \end{aligned} \tag{282}$$

Since

$$\begin{aligned} \| \tilde{\boldsymbol{x}} - \boldsymbol{m}_{\boldsymbol{x}} \| &= \| \lambda \boldsymbol{\xi} + (1 - \lambda) \boldsymbol{m}_{\boldsymbol{x}}^* - \boldsymbol{m}_{\boldsymbol{x}} \| \\ &\leqslant \lambda \, \| \boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}} \| + (1 - \lambda) \, \| \boldsymbol{m}_{\boldsymbol{x}}^* - \boldsymbol{m}_{\boldsymbol{x}} \| \\ &\leqslant \max\{ \| \boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}} \|, \| \boldsymbol{m}_{\boldsymbol{x}}^* - \boldsymbol{m}_{\boldsymbol{x}} \| \} , \end{aligned} \tag{283}$$

we have

$$\tilde{\Delta}_m \geq \Delta_m - 2 \, \max\{ \| \boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}} \|, \| \boldsymbol{m}_{\boldsymbol{x}}^* - \boldsymbol{m}_{\boldsymbol{x}} \| \} \, M . \tag{284}$$

$$\epsilon = (N - l) \exp(- \beta \, (\Delta_m - 2 \, \max\{ \| \boldsymbol{\xi} - \boldsymbol{m}_{\boldsymbol{x}} \|, \| \boldsymbol{m}_{\boldsymbol{x}}^* - \boldsymbol{m}_{\boldsymbol{x}} \| \} \, M)) . \tag{285}$$

B2.5 Properties of Fixed Points Near Stored Pattern

In Subsection B2.4.3 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.

B2.5.1 Exponentially Many Patterns can be Stored

From Subsection B2.4.3 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

$$\Delta_i = \min_{j,j \neq i} \left(\boldsymbol{x}_i^T \boldsymbol{x}_i - \boldsymbol{x}_i^T \boldsymbol{x}_j \right) = \boldsymbol{x}_i^T \boldsymbol{x}_i - \max_{j,j \neq i} \boldsymbol{x}_i^T \boldsymbol{x}_j . \tag{286}$$

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

$$\begin{aligned} \Delta_i &= \min_{j,j \neq i} \frac{1}{2} \left(\| \boldsymbol{x}_i \|^2 - \| \boldsymbol{x}_j \|^2 + \| \boldsymbol{x}_i - \boldsymbol{x}_j \|^2 \right) \\ &= \frac{1}{2} \| \boldsymbol{x}_i \|^2 - \frac{1}{2} \max_{j,j \neq i} \left(\| \boldsymbol{x}_j \|^2 - \| \boldsymbol{x}_i - \boldsymbol{x}_j \|^2 \right) . \end{aligned} \tag{287}$$

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

$$\mathrm{S}_i = \left\{ \boldsymbol{\xi} \mid \|\boldsymbol{\xi} - \boldsymbol{x}_i\| \leqslant \frac{1}{\beta\, N\, M} \right\} . \tag{288}$$

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 B4 (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. (81) 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:

$$\Delta_i \geq \frac{2}{\beta\, N} + \frac{1}{\beta} \ln\left(2\,(N-1)\,N\,\beta\,M^2\right) . \tag{289}$$

This is the assumption in Lemma 7 to ensure a fixed point in sphere S_i. Since replacing $(N-1)N$ by N^2 gives

$$\frac{2}{\beta\, N} + \frac{1}{\beta} \ln\left(2\,N^2\,\beta\,M^2\right) > \frac{2}{\beta\, N} + \frac{1}{\beta} \ln\left(2\,(N-1)\,N\,\beta\,M^2\right) , \tag{290}$$

the inequality follows from following master inequality

$$\Delta_i \geq \frac{2}{\beta\, N} + \frac{1}{\beta} \ln\left(2\,N^2\,\beta\,M^2\right) , \tag{291}$$

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

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \leqslant \frac{2}{\beta\, N\, M} . \tag{292}$$

Therefore we have using the Cauchy-Schwarz inequality:

$$\Delta_i \leqslant \boldsymbol{x}_i^T (\boldsymbol{x}_i - \boldsymbol{x}_j) \leqslant \|\boldsymbol{x}_i\|\, \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \leqslant M \frac{2}{\beta\, N\, M} = \frac{2}{\beta\, N} . \tag{293}$$

The last inequality is a contraction to Eq. (291) if we assume that

$$1 < 2\,(N-1)\,N\,\beta\,M^2 . \tag{294}$$

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

$$\Delta_{\min} = \min_{1 \leqslant i \leqslant N} \Delta_i \tag{295}$$

to obtain the master inequality

$$\Delta_{\min} \geq \frac{2}{\beta\, N} + \frac{1}{\beta} \ln\left(2\,N^2\,\beta\,M^2\right) . \tag{296}$$

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:

$$\forall_i : \|\boldsymbol{x}_i\| = M . \tag{297}$$

Under assumption Eq. (294) we have only to show that the master inequality Eq. (296) 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

$$\alpha_{\min} = \min_{1 \leqslant i < j \leqslant N} \alpha_{ij} . \tag{298}$$

On the sphere with radius M we have

$$\Delta_{\min} = \min_{1 \leqslant i < j \leqslant N} M^2(1 - \cos(\alpha_{ij})) = M^2(1 - \cos(\alpha_{\min})) , \tag{299}$$

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

$$M^2(1 - \cos(\alpha_{\min})) \geq \frac{2}{\beta\, N} + \frac{1}{\beta} \ln\left(2\,N^2\,\beta\,M^2\right) . \tag{300}$$

Under assumption Eq. (294) we have only to show that the master inequality Eq. (296) is fulfilled for $\Delta_{\min}$. We consider patterns on the sphere, therefore the master inequality Eq. (296) becomes Eq. (300). 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 B3 (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*

$$N \ = \ 2^{2(d-1)} \ . \tag{301}$$

Proof. For random patterns on the sphere, we have to show that the master inequality Eq. (300) holds:

$$M^2(1 \ - \ \cos(\alpha_{\min})) \ \geq \ \frac{2}{\beta \, N} \ + \ \frac{1}{\beta} \, \ln\left(2 \, N^2 \, \beta \, M^2\right) \ . \tag{302}$$

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

$$\forall_i : \ \min_{j, j \neq i} \alpha_{ij} \ = \ \alpha_{\min} \ , \tag{303}$$

In a d-dimensional space we can place

$$N \ = \ \left(\frac{2\pi}{\alpha_{\min}}\right)^{d-1} \tag{304}$$

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

$$\alpha_{\min} \ = \ \frac{2\pi}{N^{1/(d-1)}} \ . \tag{305}$$

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

$$M^2 \left(1 \ - \ \cos\left(\frac{2\pi}{N^{1/(d-1)}}\right)\right) \ \geq \ \frac{2}{\beta \, N} \ + \ \frac{1}{\beta} \, \ln\left(2 \, N^2 \, \beta \, M^2\right) \ . \tag{306}$$

We set $N = 2^{2(d-1)}$ and obtain for Eq. (306):

$$M^2 \left(1 \ - \ \cos\left(\frac{\pi}{2}\right)\right) \ \geq \ \frac{2}{\beta \, 2^{3(d-1)}} \ + \ \frac{1}{\beta} \, \ln\left(2 \, \beta \, M^2\right) \ + \ \frac{1}{\beta} \, 4 \, (d-1) \ln 2 \, . \tag{307}$$

This inequality is equivalent to

$$\beta \, M^2 \ \geq \ \frac{1}{2^{2(d-1)-1}} \ + \ \ln\left(2 \, \beta \, M^2\right) \ + \ 4 \, (d-1) \ln 2 \, . \tag{308}$$

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 B4 (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*

$$N \ = \ 2^{3(d-1)} \ . \tag{309}$$

Proof. We set $N = 2^{3(d-1)}$ and obtain for Eq. (306):

$$M^2 \left(1 - \cos\left(\frac{\pi}{4}\right)\right) \geq \frac{2}{\beta\, 2^{3(d-1)}} + \frac{1}{\beta} \ln\left(2\,\beta\, M^2\right) + \frac{1}{\beta}\, 6\,(d-1) \ln 2\,. \tag{310}$$

This inequality is equivalent to

$$\beta\, M^2 \left(1 - \frac{\sqrt{2}}{2}\right) \geq \frac{1}{2^{3(d-1)-1}} + \ln\left(2\,\beta\, M^2\right) + 6\,(d-1) \ln 2\,. \tag{311}$$

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. (294) we have to show that the master inequality Eq. (300) 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 [12] and [10]. Theorem 2 in [12] gives the distribution of the minimal angle for random patterns on the unit sphere. Proposition 3.5 in [10] 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[10] 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 13 ([10])**.** *Let d be the dimension of the pattern space,*

$$\kappa_d := \frac{1}{d\,\sqrt{\pi}} \frac{\Gamma((d+1)/2)}{\Gamma(d/2)}\,. \tag{312}$$

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

$$\Pr(N^{\frac{2}{d-1}} \alpha_{\min} \geq \delta) \geq 1 - \frac{\kappa_{d-1}}{2}\, \delta^{d-1}\,. \tag{313}$$

Proof. The statement of the lemma is Eq. (3-6) from Proposition 3.5 in [10]. □

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

Lemma 14. *For κ_d defined in Eq. (312) we have the following bounds for every $d \geq 1$:*

$$\frac{1}{\exp(1/6)\,\sqrt{e\,\pi\,d}} \leqslant \kappa_d \leqslant \frac{\exp(1/12)}{\sqrt{2\,\pi\,d}} < 1\,. \tag{314}$$

Proof. We use for $x > 0$ the following bound related to Stirling's approximation formula for the gamma function, c.f. [35, (5.6.1)]:

$$1 < \Gamma(x)\,(2\,\pi)^{-\frac{1}{2}} x^{\frac{1}{2}-x} \exp(x) < \exp\left(\frac{1}{12\,x}\right)\,. \tag{315}$$

Using Stirling's formula Eq. (315), we upper bound κ_d:

$$\begin{aligned} \kappa_d &= \frac{1}{d\,\sqrt{\pi}} \frac{\Gamma((d+1)/2)}{\Gamma(d/2)} < \frac{1}{d\,\sqrt{\pi}} \frac{\exp\left(\frac{1}{6(d+1)}\right) \exp\left(-\frac{d+1}{2}\right) \left(\frac{d+1}{2}\right)^{\frac{d}{2}}}{\exp\left(-\frac{d}{2}\right) \left(\frac{d}{2}\right)^{\frac{d}{2}-\frac{1}{2}}} \\ &= \frac{1}{d\,\sqrt{\pi\,e}} \exp\left(\frac{1}{6(d+1)}\right) \left(1 + \frac{1}{d}\right)^{\frac{d}{2}} \sqrt{\frac{d}{2}} \leqslant \frac{\exp\left(\frac{1}{12}\right)}{\sqrt{2\,\pi}\,\sqrt{d}}\,. \end{aligned} \tag{316}$$

For the first inequality, we applied Eq. (315), 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. (315):

$$\begin{aligned} \kappa_d &= \frac{1}{d\,\sqrt{\pi}}\,\frac{\Gamma((d+1)/2)}{\Gamma(d/2)} > \frac{1}{d\,\sqrt{\pi}}\,\frac{\exp\left(-\frac{d+1}{2}\right)\,\left(\frac{d+1}{2}\right)^{\frac{d}{2}}}{\exp\left(\frac{1}{6\,d}\right)\,\exp\left(-\frac{d}{2}\right)\,\left(\frac{d}{2}\right)^{\frac{d}{2}-\frac{1}{2}}} \\ &= \frac{1}{d\,\sqrt{\pi\,e}\,\exp\left(\frac{1}{6\,d}\right)}\,\left(1+\frac{1}{d}\right)^{\frac{d}{2}}\,\sqrt{\frac{d}{2}} \geq \frac{1}{\exp\left(\frac{1}{6}\right)\,\sqrt{e\,\pi\,d}}\,, \end{aligned} \tag{317}$$

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. (300).

Lemma 15. *For $0 \leqslant x \leqslant \pi$ the function* cos *can be upper bounded by:*

$$\cos(x) \;=\; 1 \;-\; \frac{x^2}{5}\,. \tag{318}$$

Proof. We use the infinite product representation of cos from [35, (4.22.2)]:

$$\cos(x) \;=\; \prod_{n=1}^{\infty}\left(1-\frac{4\,x^2}{(2n-1)^2\,\pi^2}\right)\,. \tag{319}$$

It holds

$$1 \;-\; \frac{4\,x^2}{(2n-1)^2\,\pi^2} \;\leqslant\; 1 \tag{320}$$

for $|x| \leqslant \pi$ and $n \geq 2$, we can get the following upper bound on Eq. (319):

$$\begin{aligned} \cos(x) \;&\leqslant\; \prod_{n=1}^{2}\left(1-\frac{4\,x^2}{(2n-1)^2\pi^2}\right) \;=\; \left(1\;-\;\frac{4\,x^2}{\pi^2}\right)\,\left(1\;-\;\frac{4\,x^2}{9\,\pi^2}\right) \\ &=\; 1\;-\;\frac{40\,x^2}{9\,\pi^2}\;+\;\frac{16\,x^4}{9\,\pi^4} \;\leqslant\; 1\;-\;\frac{40\,x^2}{9\,\pi^2}\;+\;\frac{16\,x^2}{9\,\pi^2} \\ &=\; 1\;-\;\frac{24\,x^2}{9\,\pi^2} \;\leqslant\; 1\;-\;\frac{x^2}{5}\,. \end{aligned} \tag{321}$$

The last but one inequality uses $x \leqslant \pi$, which implies $x/\pi \leqslant 1$. Thus Eq. (318) 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 B5 (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*

$$\begin{aligned} a \;&:=\; \frac{2}{d-1}\,(1\;+\;\ln(2\,\beta\,K^2\,p\,(d-1)))\,, \quad b \;:=\; \frac{2\,K^2\,\beta}{5}\,, \\ c \;&=\; \frac{b}{W_0(\exp(a\;+\;\ln(b))}\,, \end{aligned} \tag{322}$$

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

$$c \;\geq\; \left(\frac{2}{\sqrt{p}}\right)^{\frac{4}{d-1}}\,. \tag{323}$$

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

$$N \;\geq\; \sqrt{p}\,c^{\frac{d-1}{4}}\,. \tag{324}$$

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. (300) is fulfilled:

$$\Pr\left(M^2(1\,-\cos(\alpha_{\min}))\;\geq\;\frac{2}{\beta\,N}\;+\;\frac{1}{\beta}\,\ln\left(2\,N^2\,\beta\,M^2\right)\right)\;\geq\;1\;-\;p\,. \tag{325}$$

Using Eq. (318), we have:

$$1\;-\;\cos(\alpha_{\min})\;\geq\;\frac{1}{5}\,\alpha_{\min}^2\,. \tag{326}$$

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

$$\Pr\left(M^2\frac{\alpha_{min}^2}{5}\;\geq\;\frac{2}{\beta\,N}\;+\;\frac{1}{\beta}\,\ln\left(2\,N^2\,\beta\,M^2\right)\right)\;\geq\;1\;-\;p\,. \tag{327}$$

This inequality is equivalent to

$$\Pr\left(N^{\frac{2}{d-1}}\,\alpha_{min}\;\geq\;\frac{\sqrt{5}\,N^{\frac{2}{d-1}}}{M}\,\left(\frac{2}{\beta\,N}\;+\;\frac{1}{\beta}\,\ln\left(2\,N^2\,\beta\,M^2\right)\right)^{\frac{1}{2}}\right)\;\geq\;1\;-\;p\,. \tag{328}$$

We use Eq. (313) to obtain:

$$\begin{aligned}&\Pr\left(N^{\frac{2}{d-1}}\,\alpha_{min}\;\geq\;\frac{\sqrt{5}\,N^{\frac{2}{d-1}}}{M}\,\left(\frac{2}{\beta\,N}\;+\;\frac{1}{\beta}\,\ln\left(2\,N^2\,\beta\,M^2\right)\right)^{\frac{1}{2}}\right)\\ &\geq\;1\;-\;\frac{\kappa_{d-1}}{2}\,5^{\frac{d-1}{2}}\,N^2\,M^{-(d-1)}\left(\frac{2}{\beta\,N}\;+\;\frac{1}{\beta}\,\ln\left(2\,N^2\,\beta\,M^2\right)\right)^{\frac{d-1}{2}}\,.\end{aligned} \tag{329}$$

For Eq. (328) to be fulfilled, it is sufficient that

$$\frac{\kappa_{d-1}}{2}\,5^{\frac{d-1}{2}}\,N^2\,M^{-(d-1)}\left(\frac{2}{\beta\,N}\;+\;\frac{1}{\beta}\,\ln\left(2\,N^2\,\beta M^2\right)\right)^{\frac{d-1}{2}}\;-\;p\;\leqslant\;0\,. \tag{330}$$

If we insert the assumption Eq. (323) of the theorem into Eq. (324), then we obtain $N\geq 2$. We now apply the upper bound $\kappa_{d-1}/2<\kappa_{d-1}<1$ from Eq. (314) and the upper bound $\frac{2}{\beta N}\leqslant\frac{1}{\beta}$ from $N\geq 2$ to inequality Eq. (330). 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:

$$5^{\frac{d-1}{2}}\,p\,c^{\frac{d-1}{2}}\,M^{-(d-1)}\left(\frac{1}{\beta}\;+\;\frac{1}{\beta}\,\ln\left(2\,p\,c^{\frac{d-1}{2}}\,\beta M^2\right)\right)^{\frac{d-1}{2}}\;\leqslant\;p\,. \tag{331}$$

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

$$\frac{5\,c}{K^2\,(d-1)}\left(\frac{1}{\beta}\;+\;\frac{1}{\beta}\,\ln\left(2\,\beta\,c^{\frac{d-1}{2}}\,p\,K^2\,(d-1)\right)\right)\;-\;1\;\leqslant\;0\,. \tag{332}$$

After some algebraic manipulation, this inequality can be written as

$$a\,c\;+\;c\,\ln(c)\;-\;b\;\leqslant\;0\,, \tag{333}$$

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. (333) equal to zero. We solve

$$a\,\hat{c}\;+\;\hat{c}\,\ln(\hat{c})\;-\;b\;=\;0 \tag{334}$$

for $\hat{c}$:

$$\begin{aligned}&a\,\hat{c}\;+\;\hat{c}\,\ln(\hat{c})\;-\;b\;=\;0\\ \Leftrightarrow\;&a\;+\;\ln(\hat{c})\;=\;b/\hat{c}\\ \Leftrightarrow\;&a\;+\;\ln(b)\;+\;\ln(\hat{c}/b)\;=\;b/\hat{c}\\ \Leftrightarrow\;&b/\hat{c}\;+\;\ln(b/\hat{c})\;=\;a\;+\;\ln(b)\\ \Leftrightarrow\;&b/\hat{c}\,\exp(b/\hat{c})\;=\;\exp(a\;+\;\ln(b))\\ \Leftrightarrow\;&b/\hat{c}\;=\;W_0(\exp(a\;+\;\ln(b)))\\ \Leftrightarrow\;&\hat{c}\;=\;\frac{b}{W_0(\exp(a\;+\;\ln(b))}\,,\end{aligned} \tag{335}$$

where W_0 is the upper branch of the Lambert W function (see Def. B10). Hence, the solution is

$$\hat{c} \;=\; \frac{b}{W_0(\exp(a \;+\; \ln(b))} \;. \tag{336}$$

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. (333) and therefore also Eq. (331), we have a lower bound on the storage capacity N:

$$N \;\geq\; \sqrt{p}\, \hat{c}^{\,\frac{d-1}{4}} \;. \tag{337}$$

□

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 1. *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

$$c \;=\; \begin{cases} b\,\ln\left(\frac{\Omega\,\exp(a\,+\,\ln(b))\,+\,1}{\Omega\,(1\,+\,\Omega)}\right)^{-1} & \text{for } a \;+\; \ln(b) \;\leqslant\; 0\;, \\ b\,(a \;+\; \ln(b))^{-\frac{a\,+\,\ln(b)}{a\,+\,\ln(b)\,+\,1}} & \text{for } a \;+\; \ln(b) \;>\; 0 \end{cases} \tag{338}$$

and ensure

$$c \;\geq\; \left(\frac{2}{\sqrt{p}}\right)^{\frac{4}{d-1}} \;. \tag{339}$$

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

$$N \;\geq\; \sqrt{p}\, c^{\,\frac{d-1}{4}} \;. \tag{340}$$

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 B5. According to [26, Theorem 2.3] we have for any real u and $y > \frac{1}{e}$:

$$W_0(\exp(u)) \;\leqslant\; \ln\left(\frac{\exp(u) \;+\; y}{1 \;+\; \ln(y)}\right) \;. \tag{341}$$

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

$$y \;=\; 1/W_0(1) \;=\; 1/\Omega \;=\; \exp\Omega \;=\; -\,1/\ln\Omega \;\approx\; 1.76322\;, \tag{342}$$

where the Omega constant Ω is

$$\Omega \;=\; \left(\int_{-\infty}^{\infty} \frac{\mathrm{d}t}{(e^t \;-\; t)^2 \;+\; \pi^2}\right)^{-1} \;-\; 1 \;\approx\; 0.56714329\;. \tag{343}$$

See for these equations the special values of the Lambert W function in Lemma 31. We have the upper bound on W_0:

$$W_0(\exp(u)) \;\leqslant\; \ln\left(\frac{\exp(u) \;+\; 1/\Omega}{1 \;+\; \ln(1/\Omega)}\right) \;=\; \ln\left(\frac{\Omega\,\exp(u) \;+\; 1}{\Omega(1 \;+\; \Omega)}\right) \;. \tag{344}$$

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

$$\ln\left(\frac{\Omega\,1\,+\,1}{\Omega(1\,+\,\Omega)}\right) \;=\; \ln\left(\frac{1}{\Omega}\right) \;=\; -\,\ln\left(\Omega\right) \;=\; \Omega \;=\; W_0(1)\,. \tag{345}$$

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 [26, Corollary 2.6] the following bound on $W_0(\exp(u))$ for $1 < \exp(u)$, or, equivalently $0 < u$:

$$W_0(\exp(u)) \;\leqslant\; u^{\frac{u}{1\,+\,u}}\,. \tag{346}$$

A lower bound on $\hat{c}$ is obtained via the upper bounds Eq. (346) and Eq. (344) on W_0 as $W_0 > 0$. We set $u = a + \ln(b)$ and obtain

$$W_0(\exp(a\,+\,\ln(b))) \;\leqslant\; \begin{cases} \ln\left(\frac{\Omega\,\exp(a\,+\,\ln(b))\,+\,1}{\Omega\,(1\,+\,\Omega)}\right)^{-1} & \text{for } a\,+\,\ln(b) \;\leqslant\; 0\,, \\ (a\,+\,\ln(b))^{-\frac{a\,+\,\ln(b)}{a\,+\,\ln(b)\,+\,1}} & \text{for } a\,+\,\ln(b) \;>\; 0 \end{cases} \tag{347}$$

We insert this bound into Eq. (336), 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 B6 (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*

$$\begin{aligned} a \;&:=\; \frac{\ln(c)}{2} \;-\; \frac{K^2\,\beta}{5\,c}\,, \quad b \;:=\; 1 \;+\; \ln\left(2\,p\,\beta\,K^2\right)\,, \\ d \;&=\; \begin{cases} 1 \;+\; \frac{1}{a}\,W(a\,\exp(-b)) & \textit{for } a \neq 0\,, \\ 1 \;+\; \exp(-b) & \textit{for } a = 0\,, \end{cases} \end{aligned} \tag{348}$$

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

$$c \;\geq\; \left(\frac{2}{\sqrt{p}}\right)^{\frac{4}{d-1}}\,, \quad -\frac{1}{e} \;\leqslant\; a\,\exp(-b)\,, \tag{349}$$

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

$$N \;\geq\; \sqrt{p}\,c^{\frac{d-1}{4}}\,. \tag{350}$$

Proof. We consider the probability that the master inequality Eq. (300) is fulfilled:

$$\Pr\left(M^2(1\,-\cos(\alpha_{\min}))) \;\geq\; \frac{2}{\beta\,N} \;+\; \frac{1}{\beta}\,\ln\left(2\,N^2\,\beta\,M^2\right)\right) \;\geq\; 1\;-\;p\,. \tag{351}$$

Using Eq. (318), we have:

$$1\;-\;\cos(\alpha_{\min}) \;\geq\; \frac{1}{5}\,\alpha_{\min}^2\,. \tag{352}$$

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

$$\Pr\left(M^2\frac{\alpha_{min}^2}{5} \;\geq\; \frac{2}{\beta\,N} \;+\; \frac{1}{\beta}\,\ln\left(2\,N^2\,\beta\,M^2\right)\right) \;\geq\; 1\;-\;p\,. \tag{353}$$

This inequality is equivalent to

$$\Pr\left(N^{\frac{2}{d-1}}\,\alpha_{min} \;\geq\; \frac{\sqrt{5}\,N^{\frac{2}{d-1}}}{M}\,\left(\frac{2}{\beta\,N} \;+\; \frac{1}{\beta}\,\ln\left(2\,N^2\,\beta\,M^2\right)\right)^{\frac{1}{2}}\right) \;\geq\; 1\;-\;p\,. \tag{354}$$

We use Eq. (313) to obtain:

$$\Pr\left(N^{\frac{2}{d-1}}\,\alpha_{min} \;\geq\; \frac{\sqrt{5}\,N^{\frac{2}{d-1}}}{M}\,\left(\frac{2}{\beta\,N} \;+\; \frac{1}{\beta}\,\ln\left(2\,N^2\,\beta\,M^2\right)\right)^{\frac{1}{2}}\right) \tag{355}$$
$$\geq\; 1 \;-\; \frac{\kappa_{d-1}}{2}\,5^{\frac{d-1}{2}}\,N^2\,M^{-(d-1)}\left(\frac{2}{\beta\,N} \;+\; \frac{1}{\beta}\,\ln\left(2\,N^2\,\beta\,M^2\right)\right)^{\frac{d-1}{2}}\,.$$

For Eq. (354) to be fulfilled, it is sufficient that

$$\frac{\kappa_{d-1}}{2}\,5^{\frac{d-1}{2}}\,N^2\,M^{-(d-1)}\left(\frac{2}{\beta\,N} \;+\; \frac{1}{\beta}\,\ln\left(2\,N^2\,\beta M^2\right)\right)^{\frac{d-1}{2}} \;-\; p \;\leqslant\; 0\,. \tag{356}$$

If we insert the assumption Eq. (349) of the theorem into Eq. (350), then we obtain $N \geq 2$. We now apply the upper bound $\kappa_{d-1}/2 < \kappa_{d-1} < 1$ from Eq. (314) and the upper bound $\frac{2}{\beta N} \leqslant \frac{1}{\beta}$ from $N \geq 2$ to inequality Eq. (356). 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:

$$5^{\frac{d-1}{2}}\,p\,c^{\frac{d-1}{2}}\,M^{-(d-1)}\left(\frac{1}{\beta} \;+\; \frac{1}{\beta}\,\ln\left(2\,p\,c^{\frac{d-1}{2}}\,\beta M^2\right)\right)^{\frac{d-1}{2}} \;\leqslant\; p\,. \tag{357}$$

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

$$\frac{5\,c}{K^2\,(d-1)}\left(\frac{1}{\beta} \;+\; \frac{1}{\beta}\,\ln\left(2\,\beta\,c^{\frac{d-1}{2}}\,p\,K^2\,(d-1)\right)\right) \;-\; 1 \;\leqslant\; 0\,. \tag{358}$$

This inequality Eq. (358) can be reformulated as:

$$1 \;+\; \ln\left(2\,p\,\beta\,c^{\frac{d-1}{2}}\,K^2\,(d-1)\right) \;-\; \frac{(d-1)\,K^2\,\beta}{5\,c} \;\leqslant\; 0\,. \tag{359}$$

Using

$$a \;:=\; \frac{\ln(c)}{2} \;-\; \frac{K^2\,\beta}{5\,c}\,, \quad b \;:=\; 1 \;+\; \ln\left(2\,p\,\beta\,K^2\right)\,, \tag{360}$$

we write inequality Eq. (359) as

$$\ln(d-1) \;+\; a\,(d-1) \;+\; b \;\leqslant\; 0\,. \tag{361}$$

We determine the value $\hat{d}$ of d which makes the inequality Eq. (361) equal to zero. We solve

$$\ln(\hat{d}-1) \;+\; a\,(\hat{d}-1) \;+\; b \;=\; 0\,. \tag{362}$$

for $\hat{d}$
For $a \neq 0$ we have

$$\begin{aligned}
&\ln(\hat{d}-1) \;+\; a\,(\hat{d}-1) \;+\; b \;=\; 0 \qquad (363)\\
\Leftrightarrow\; &a\,(\hat{d}-1) \;+\; \ln(\hat{d}-1) \;=\; -\,b\\
\Leftrightarrow\; &(\hat{d}-1)\exp(a\,(\hat{d}-1)) \;=\; \exp(-b)\\
\Leftrightarrow\; &a\,(\hat{d}-1)\exp(a\,(\hat{d}-1)) \;=\; a\,\exp(-b)\\
\Leftrightarrow\; &a\,(\hat{d}-1) \;=\; W(a\,\exp(-b))\\
\Leftrightarrow\; &\hat{d} \;-\; 1 \;=\; \frac{1}{a}\,W(a\,\exp(-b))\\
\Leftrightarrow\; &\hat{d} \;=\; 1 \;+\; \frac{1}{a}\,W(a\,\exp(-b))\,,
\end{aligned}$$

where W is the Lambert W function (see Def. B10). 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

$$\hat{d} \;=\; 1 \;+\; \frac{1}{a}\, W(a \exp(-b)) \;. \tag{364}$$

Since $\hat{d}$ fulfills inequality Eq. (358) and therefore also Eq. (357), we have a lower bound on the storage capacity N:

$$N \;\geq\; \sqrt{p}\, \hat{c}^{\,\frac{d-1}{4}} \;. \tag{365}$$

□

Corollary 2. *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*

$$\begin{aligned} a \;&:=\; \frac{\ln(c)}{2} \;-\; \frac{K^2\,\beta}{5\,c} \;, \quad b \;:=\; 1 \;+\; \ln\left(2\,p\,\beta\,K^2\right) \;, \\ d \;&=\; 1 \;+\; \frac{1}{a}\,\left(-\,\ln(-a) \;+\; b\right) \;, \end{aligned} \tag{366}$$

and ensure

$$c \;\geq\; \left(\frac{2}{\sqrt{p}}\right)^{\frac{4}{d-1}} \;, \quad -\frac{1}{e} \;\leqslant\; a\, \exp(-b) \;, \quad a \;<\; 0 \;, \tag{367}$$

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

$$N \;\geq\; \sqrt{p}\, c^{\,\frac{d-1}{4}} \;. \tag{368}$$

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

Proof. For $a < 0$ the Eq. (348) from Theorem (B6) can be written as

$$d \;=\; 1 \;+\; \frac{W_{-1}(a \exp(-b))}{a} \;=\; 1 \;+\; \frac{W_{-1}(-\exp\left(-(-\ln(-a) + b - 1) - 1\right))}{a} \tag{369}$$

From [2, Theorem 3.1] we get the following bound on W_{-1}:

$$-\,\frac{e}{e-1}\,(u+1) \;<\; W_{-1}(-\,\exp(-u-1)) \;<\; -\,(u+1) \;. \tag{370}$$

for $u > 0$. We apply Eq. (370) to Eq. (369) with $u = -\ln(-a) + b - 1$.
Since $a < 0$ we get

$$d \;>\; 1 \;+\; \frac{-\ln(-a) + b}{a} \;. \tag{371}$$

□

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 16 (Proposition 3.6 in [10])**.** *We have the following lower bound on the expectation of $\alpha_{\min}$:*

$$\mathrm{E}\left[N^{\frac{2}{d-1}}\;\alpha_{\min}\right] \;\geq\; \left(\frac{\Gamma(\frac{d}{2})}{2(d-1)\,\sqrt{\pi}\,\Gamma(\frac{d-1}{2})}\right)^{-\frac{1}{d-1}} \Gamma(1+\frac{1}{d-1})\,\frac{d^{-\frac{1}{d-1}}}{\Gamma(2+\frac{1}{d-1})} \;:=\; C_{d-1}. \tag{372}$$

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. (372):

Lemma 17. *The following estimate holds for $d \geq 2$:*

$$C_d \ \geq \ 1 \ - \ \frac{\ln(d+1)}{d} \ . \tag{373}$$

Proof. The recursion formula for the Gamma function is [35, (5.5.1)]:

$$\Gamma(x+1) \ = \ x \, \Gamma(x) \ . \tag{374}$$

We use Eq. (314) and the fact that $d^{\frac{1}{d}} \geq 1$ for $d \geq 1$ to obtain:

$$\begin{aligned} C_d \ &\geq \ (2\sqrt{d})^{\frac{1}{d}} \Gamma(1+\frac{1}{d}) \, \frac{(d+1)^{-\frac{1}{d}}}{\Gamma(2+\frac{1}{d})} \ = \ (2\sqrt{d})^{\frac{1}{d}} \frac{(d+1)^{-\frac{1}{d}}}{1-\frac{1}{d}} \ > \ (d+1)^{\frac{1}{d}} \\ &= \ \exp(-\frac{1}{d} \, \ln(d+1)) \ \geq \ 1 \ - \ \frac{1}{d} \, \ln(d+1) \ , \end{aligned} \tag{375}$$

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 B7 (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*

$$\frac{1}{5} \, (d-1) \, K^2 \, c^{-1} (1 \ - \ \frac{\ln(d-1)}{(d-1)})^2 \ \geq \ \frac{2}{\beta \, c^{\frac{d-1}{4}}} \ + \ \frac{1}{\beta} \, \ln\left(2 \, c^{\frac{d-1}{2}} \, \beta \, (d-1) \, K^2\right) \tag{376}$$

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

$$N \ = \ c^{\frac{d-1}{4}} \ . \tag{377}$$

The inequality Eq. (376) 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. (300) is fulfilled we now consider whether this inequality is fulfilled for the expected minimal distance. We consider the expectation of the minimal distance $\Delta_{\min}$:

$$\mathrm{E}[\Delta_{\min}] \ = \ \mathrm{E}[M^2(1 \ - \ \cos(\alpha_{\min}))] \ = \ M^2(1 \ - \ \mathrm{E}[\cos(\alpha_{\min})]) \ . \tag{378}$$

For this expectation, the master inequality Eq. (300) becomes

$$M^2(1 \ - \ \mathrm{E}[\cos(\alpha_{\min})]) \ \geq \ \frac{2}{\beta \, N} \ + \ \frac{1}{\beta} \, \ln\left(2 \, N^2 \, \beta \, M^2\right) \ . \tag{379}$$

We want to find the largest N that fulfills this inequality.
We apply Eq. (318) and Jensen's inequality to deduce the following lower bound:

$$1 \ - \ \mathrm{E}[\cos(\alpha_{\min})] \ \geq \ \frac{1}{5} \, \mathrm{E}\left[\alpha_{\min}^2\right] \ \geq \ \frac{1}{5} \, \mathrm{E}[\alpha_{\min}]^2 \ . \tag{380}$$

Now we use Eq. (372) and Eq. (373) to arrive at

$$\mathrm{E}[\alpha_{\min}]^2 \ \geq \ N^{-\frac{4}{d-1}} \, \mathrm{E}[N^{\frac{2}{d-1}} \, \alpha_{\min}]^2 \ \geq \ N^{-\frac{4}{d-1}} \, C_{d-1}^2 \ \geq \ N^{-\frac{4}{d-1}} \, (1 - \frac{\ln(d-1)}{(d-1)})^2 \ , \tag{381}$$

for sufficiently large d. Thus in order to fulfill Eq. (379), it is enough to find values that satisfy Eq. (376).

□

B2.5.2 Convergence after One Update and Small Retrieval Error

Theorem B8 (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:*

$$\|f(\boldsymbol{\xi}) \ - \ \boldsymbol{x}_i^*\| \ \leqslant \ \|\mathrm{J}^m\|_2 \, \|\boldsymbol{\xi} \ - \ \boldsymbol{x}_i^*\| \ , \tag{382}$$

$$\|\mathrm{J}^m\|_2 \ \leqslant \ 2 \, \beta \, N \, M^2 \, (N-1) \exp(-\beta \, (\Delta_i \ - \ 2 \, \max\{\|\boldsymbol{\xi} \ - \ \boldsymbol{x}_i\|, \|\boldsymbol{x}_i^* \ - \ \boldsymbol{x}_i\|\} \, M)) \ . \tag{383}$$

Proof. From Eq. (169) we have

$$\|\mathrm{J}^m\|_2 \leqslant 2\,\beta\,N\,M^2\,(N-1)\exp(-\beta\,(\Delta_i - 2\,\max\{\|\boldsymbol{\xi}-\boldsymbol{x}_i\|,\|\boldsymbol{x}_i^*-\boldsymbol{x}_i\|\}\,M))\,. \tag{384}$$

After every iteration the mapped point $f(\boldsymbol{\xi})$ is closer to the fixed point $\boldsymbol{x}_i^*$ than the original point $\boldsymbol{x}_i$:

$$\|f(\boldsymbol{\xi}) - \boldsymbol{x}_i^*\| \leqslant \|\mathrm{J}^m\|_2\,\|\boldsymbol{\xi}-\boldsymbol{x}_i^*\|\,. \tag{385}$$

□

We want to estimate how large Δ_i is. For $\boldsymbol{x}_i$ we have:

$$\Delta_i = \min_{j,j\neq i}\left(\boldsymbol{x}_i^T\boldsymbol{x}_i - \boldsymbol{x}_i^T\boldsymbol{x}_j\right) = \boldsymbol{x}_i^T\boldsymbol{x}_i - \max_{j,j\neq i}\boldsymbol{x}_i^T\boldsymbol{x}_j\,. \tag{386}$$

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

$$\mathrm{E}\left[\boldsymbol{x}^T\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{y}\right] = \sum_{j=1}^d \mathrm{E}\left[x_j^2\right] + \sum_{j=1}^d \mathrm{E}\left[x_j\right]\sum_{j=1}^d \mathrm{E}\left[y_j\right] = d\,. \tag{387}$$

The variance of the separation of two points with normally distributed components is

$$\begin{aligned}
&\mathrm{Var}\left[\boldsymbol{x}^T\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{y}\right] = \mathrm{E}\left[\left(\boldsymbol{x}^T\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{y}\right)^2\right] - d^2 \qquad (388)\\
&= \sum_{j=1}^d \mathrm{E}\left[x_j^4\right] + \sum_{j=1,k=1,k\neq j}^d \mathrm{E}\left[x_j^2\right]\,\mathrm{E}\left[x_k^2\right] - 2\sum_{j=1}^d \mathrm{E}\left[x_j^3\right]\mathrm{E}\left[y_j\right] -\\
&2\sum_{j=1,k=1,k\neq j}^d \mathrm{E}\left[x_j^2\right]\mathrm{E}\left[x_k\right]\mathrm{E}\left[y_k\right] + \sum_{j=1}^d \mathrm{E}\left[x_j^2\right]\,\mathrm{E}\left[y_j^2\right] +\\
&\sum_{j=1,k=1,k\neq j}^d \mathrm{E}\left[x_j\right]\mathrm{E}\left[y_j\right]\mathrm{E}\left[x_k\right]\mathrm{E}\left[y_k\right] - d^2\\
&= 3\,d + d\,(d-1) + d - d^2 = 3\,d\,.
\end{aligned}$$

The expected value for the separation of two random vectors gives:

$$\|\mathrm{J}^m\|_2 \leqslant 2\,\beta\,N\,M^2\,(N-1)\exp(-\beta\,(d - 2\,\max\{\|\boldsymbol{\xi}-\boldsymbol{x}_i\|,\|\boldsymbol{x}_i^*-\boldsymbol{x}_i\|\}\,M))\,. \tag{389}$$

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 B9 (Exponentially Small Retrieval Error)**.** *The retrieval error $\|\boldsymbol{x}_i - \boldsymbol{x}_i^*\|$ of pattern $\boldsymbol{x}_i$ is bounded by*

$$\|\boldsymbol{x}_i - \boldsymbol{x}_i^*\| \leqslant 2\,(N-1)\,\exp(-\beta\,(\Delta_i - 2\,\|\boldsymbol{x}_i^* - \boldsymbol{x}_i\|\,M))\,M \tag{390}$$

and for $\|\boldsymbol{x}_i - \boldsymbol{x}_i^*\| \leqslant \frac{1}{2\,\beta\,M}$ by

$$\|\boldsymbol{x}_i - \boldsymbol{x}_i^*\| \leqslant e\,(N-1)\,M\,\exp(-\beta\,\Delta_i)\,. \tag{391}$$

Proof. We compute the retrieval error which is just $\|\boldsymbol{x}_i - \boldsymbol{x}_i^*\|$. From Lemma 4 we have

$$\|\boldsymbol{x}_i - f(\boldsymbol{\xi})\| \leqslant 2\,\epsilon\,M\,, \tag{392}$$

From Eq. (168) we have

$$\epsilon = (N-1)\exp(-\beta\,(\Delta_i - 2\,\max\{\|\boldsymbol{\xi}-\boldsymbol{x}_i\|,\|\boldsymbol{x}_i^*-\boldsymbol{x}_i\|\}\,M))\,. \tag{393}$$

We use $\boldsymbol{\xi} = \boldsymbol{x}_i^*$ and get

$$\epsilon = (N-1)\exp(-\beta\,(\Delta_i - 2\,\|\boldsymbol{x}_i^* - \boldsymbol{x}_i\|\,M))\,. \tag{394}$$

We obtain

$$\|\boldsymbol{x}_i - \boldsymbol{x}_i^*\| \leqslant 2\,(N-1)\,\exp(-\beta\,(\Delta_i - 2\,\|\boldsymbol{x}_i^* - \boldsymbol{x}_i\|\,M))\,M\,. \tag{395}$$

For $\|\boldsymbol{x}_i - \boldsymbol{x}_i^*\| \leqslant \frac{1}{2\,\beta\,M}$ inequality Eq. (395) gives

$$\|\boldsymbol{x}_i - \boldsymbol{x}_i^*\| \leqslant e\,(N-1)\,M\,\exp(-\beta\,\Delta_i)\,. \tag{396}$$

□

B2.6 Learning Associations

B2.6.1 Initialization: Random Matrix Theory

For the initial matrices and scaling, the random matrix theory is of interest. For matrix entries with variance σ^2 we know from the circular law [43] and the Marchenko-Pastur quarter circular law [31, 50, 8] that $1/(\sigma 2\sqrt{N})\boldsymbol{X}$ has a singular value density concentrated at values smaller than one. The maximal singular value of $\boldsymbol{X} \in \mathbb{R}^{N \times n}$ is $s_{\max}(\boldsymbol{X}) \propto \sqrt{N} + \sqrt{n}$ [38]. Furthermore large singular values have lower density according the quarter circular law. Initialization of mappings to the space, where the modern Hopfield networks works, can be based on the largest singular value. Therefore we can estimate the largest possible norm M of the patterns as we used in the theory.

B2.6.2 Directly Learning Associations

In the first setting, $\boldsymbol{x}$ is mapped by $\boldsymbol{W}\boldsymbol{x}$ to the query space, where the query $\boldsymbol{\xi}$ lives. With and the largest norm of a pattern

$$M_W \ = \ \max_i \left\| \boldsymbol{W}^T \boldsymbol{x}_i \right\| , \tag{397}$$

the energy function E is now

$$\mathrm{E} \ = \ - \ \mathrm{lse}(\beta, \boldsymbol{X}^T \boldsymbol{W}^T \boldsymbol{\xi}) \ + \ \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \ + \ \beta^{-1} \ln N \ + \ \frac{1}{2} M_W^2 \tag{398}$$

$$= \ - \ \beta^{-1} \ln \left(\sum_{i=1}^{N} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{W}^T \boldsymbol{\xi}) \right) \ + \ \frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi} \ + \ \beta^{-1} \ln N \ + \ \frac{1}{2} M_W^2 \ . \tag{399}$$

The derivative of the energy E with respect to $\boldsymbol{\xi}$ is

$$\frac{\partial \mathrm{E}}{\partial \boldsymbol{\xi}} \ = \ - \ \boldsymbol{W} \boldsymbol{X} \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{W}^T \boldsymbol{\xi}) \ + \ \boldsymbol{\xi} \ = \ - \ \boldsymbol{W} \boldsymbol{X} \boldsymbol{p} \ + \ \boldsymbol{\xi} \ , \tag{400}$$

where we used

$$\boldsymbol{p} \ = \ \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{W}^T \boldsymbol{\xi}) \ . \tag{401}$$

The gradient update rule gives

$$\boldsymbol{\xi}^{\mathrm{new}} \ = \ \boldsymbol{W} \boldsymbol{X} \boldsymbol{p} \ = \ \boldsymbol{\xi} \ - \ \frac{\partial \mathrm{E}}{\partial \boldsymbol{\xi}} \ . \tag{402}$$

We consider the query $\boldsymbol{\xi}$ with result $\boldsymbol{y}$:

$$\boldsymbol{y} \ = \ \boldsymbol{W} \boldsymbol{X} \boldsymbol{p} \ = \ \boldsymbol{W} \boldsymbol{X} \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{W}^T \boldsymbol{\xi}) \tag{403}$$

Since the retrieved vector $\boldsymbol{y}$ is mapped by a weight matrix $\boldsymbol{V}$ to another vector, we consider the simplified update rule:

$$\boldsymbol{y} \ = \ \boldsymbol{X} \boldsymbol{p} \ = \ \boldsymbol{X} \mathrm{softmax}(\beta \boldsymbol{X}^T \boldsymbol{W}^T \boldsymbol{\xi}) \tag{404}$$

The derivative with respect to $\boldsymbol{W}$ is

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{y}}{\partial \boldsymbol{W}} \ = \ \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{W}} \frac{\partial \boldsymbol{a}^T \boldsymbol{y}}{\partial \boldsymbol{y}} \ = \ \frac{\partial \boldsymbol{y}}{\partial (\boldsymbol{W}^T \boldsymbol{\xi})} \frac{\partial (\boldsymbol{W}^T \boldsymbol{\xi})}{\partial \boldsymbol{W}} \frac{\partial \boldsymbol{a}^T \boldsymbol{y}}{\partial \boldsymbol{y}} \ . \tag{405}$$

$$\frac{\partial \boldsymbol{y}}{\partial (\boldsymbol{W}^T \boldsymbol{\xi})} \ = \ \beta \ \boldsymbol{X} \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p} \boldsymbol{p}^T \right) \boldsymbol{X}^T \tag{406}$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{y}}{\partial \boldsymbol{y}} \ = \ \boldsymbol{a} \ . \tag{407}$$

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:

$$\frac{\partial (\boldsymbol{W}^T \boldsymbol{\xi})}{\partial \boldsymbol{W}} \frac{\partial \boldsymbol{a}^T \boldsymbol{y}}{\partial \boldsymbol{y}} \ = \ \frac{\partial (\boldsymbol{W}^T \boldsymbol{\xi})}{\partial \boldsymbol{W}} \boldsymbol{a} \ = \ \boldsymbol{\xi}^T \boldsymbol{a} \boldsymbol{I} \ . \tag{408}$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{y}}{\partial \boldsymbol{W}} \ = \ \beta \ \boldsymbol{X} \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p} \boldsymbol{p}^T \right) \boldsymbol{X}^T \ (\boldsymbol{\xi}^T \boldsymbol{a}) \ . \tag{409}$$

B2.6.3 Learning the Mappings to the Association Space

We consider the patterns $\boldsymbol{x}$ that are mapped to $\tilde{\boldsymbol{x}}$ in the association space $\mathbb{R}^d$ by $\tilde{\boldsymbol{x}} = \boldsymbol{W}^K \boldsymbol{x}$. The query $\boldsymbol{\xi}$ is mapped to $\tilde{\boldsymbol{\xi}}$ in the space $\mathbb{R}^d$ by $\tilde{\boldsymbol{\xi}} = \boldsymbol{W}^Q \boldsymbol{\xi}$, too.
With and the largest norm of a pattern

$$M_W \ = \ \max_i \left\| \boldsymbol{W}^K \boldsymbol{x}_i \right\| , \tag{410}$$

the energy function E with mappings $\boldsymbol{W}^K$ and $\boldsymbol{W}^Q$ is

$$\begin{aligned} \mathrm{E} \ &= \ - \ \mathrm{lse}(\beta, \boldsymbol{X}^T (\boldsymbol{W}^K)^T \boldsymbol{W}^Q \boldsymbol{\xi}) \ + \ \frac{1}{2} \boldsymbol{\xi}^T (\boldsymbol{W}^Q)^T \boldsymbol{W}^Q \boldsymbol{\xi} \\ &+ \ \beta^{-1} \ln N \ + \ \frac{1}{2} M_W^2 \\ &= \ - \ \beta^{-1} \ln \left(\sum_{i=1}^N \exp(\beta \boldsymbol{x}_i^T (\boldsymbol{W}^K)^T \boldsymbol{W}^Q \boldsymbol{\xi}) \right) \ + \ \frac{1}{2} \boldsymbol{\xi}^T (\boldsymbol{W}^Q)^T \boldsymbol{W}^Q \boldsymbol{\xi} \\ &+ \ \beta^{-1} \ln N \ + \ \frac{1}{2} M_W^2 \ . \end{aligned} \tag{411}$$

In the association space that is

$$\mathrm{E} \ = \ - \ \mathrm{lse}(\beta, \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{\xi}}) \ + \ \frac{1}{2} \tilde{\boldsymbol{\xi}}^T \tilde{\boldsymbol{\xi}} \ + \ \beta^{-1} \ln N \ + \ \frac{1}{2} \tilde{\boldsymbol{x}}_{\max}^T \tilde{\boldsymbol{x}}_{\max} \tag{412}$$

$$= \ - \ \beta^{-1} \ln \left(\sum_{i=1}^N \exp(\beta \tilde{\boldsymbol{x}}_i^T \tilde{\boldsymbol{\xi}}) \right) \ + \ \frac{1}{2} \tilde{\boldsymbol{\xi}}^T \tilde{\boldsymbol{\xi}} \ + \ \beta^{-1} \ln N \ + \ \frac{1}{2} \tilde{\boldsymbol{x}}_{\max}^T \tilde{\boldsymbol{x}}_{\max} \ . \tag{413}$$

The derivative of the energy E with respect to $\tilde{\boldsymbol{\xi}}$ is

$$\frac{\partial \mathrm{E}}{\partial \tilde{\boldsymbol{\xi}}} \ = \ - \ \tilde{\boldsymbol{X}} \mathrm{softmax}(\beta \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{\xi}}) \ + \ \tilde{\boldsymbol{\xi}} \ = \ - \ \tilde{\boldsymbol{X}} \boldsymbol{p} \ + \ \tilde{\boldsymbol{\xi}} \, , \tag{414}$$

where we used

$$\boldsymbol{p} \ = \ \mathrm{softmax}(\beta \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{\xi}}) \, . \tag{415}$$

The gradient update rule gives

$$\tilde{\boldsymbol{\xi}}^{\mathrm{new}} \ = \ \tilde{\boldsymbol{X}} \boldsymbol{p} \ = \ \tilde{\boldsymbol{\xi}} \ - \ \frac{\partial \mathrm{E}}{\partial \tilde{\boldsymbol{\xi}}} \, . \tag{416}$$

We consider the query $\boldsymbol{\xi}$ that is mapped to $\tilde{\boldsymbol{\xi}}$ to obtain $\tilde{\boldsymbol{\xi}}^{\mathrm{new}}$:

$$\tilde{\boldsymbol{\xi}}^{\mathrm{new}} \ = \ \boldsymbol{W}^Q \boldsymbol{\xi}^{\mathrm{new}} \ = \ \boldsymbol{W}^K \boldsymbol{X} \boldsymbol{p} \ = \ \boldsymbol{W}^K \boldsymbol{X} \mathrm{softmax}(\beta \boldsymbol{X}^T (\boldsymbol{W}^K)^T \boldsymbol{W}^Q \boldsymbol{\xi}) \, . \tag{417}$$

Since the retrieved vector is mapped by a weight matrix $\boldsymbol{V}$ to another vector, we consider the simplified update rule. The retrieved vector is now $\boldsymbol{y}$ given by

$$\boldsymbol{y} \ = \ \boldsymbol{X} \boldsymbol{p} \ = \ \boldsymbol{X} \mathrm{softmax}(\beta \boldsymbol{X}^T (\boldsymbol{W}^K)^T \boldsymbol{W}^Q \boldsymbol{\xi}) \, . \tag{418}$$

The vector $\boldsymbol{y}$ does not live in the association space but in the pattern space of $\boldsymbol{x}$. Only $\boldsymbol{W}^K$ would map it to the association space.
The derivative with respect to $\boldsymbol{W}^Q$ is

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{y}}{\partial \boldsymbol{W}^Q} \ = \ \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{W}^Q} \frac{\partial \boldsymbol{a}^T \boldsymbol{y}}{\partial \boldsymbol{y}} \ = \ \frac{\partial \boldsymbol{y}}{\partial (\boldsymbol{W}^Q \boldsymbol{\xi})} \frac{\partial (\boldsymbol{W}^Q \boldsymbol{\xi})}{\partial \boldsymbol{W}^Q} \frac{\partial \boldsymbol{a}^T \boldsymbol{y}}{\partial \boldsymbol{y}} \ . \tag{419}$$

$$\frac{\partial \boldsymbol{y}}{\partial (\boldsymbol{W}^Q \boldsymbol{\xi})} \ = \ \beta \, \boldsymbol{X} \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p} \boldsymbol{p}^T \right) \boldsymbol{X}^T (\boldsymbol{W}^K)^T \tag{420}$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{y}}{\partial \boldsymbol{y}} \ = \ \boldsymbol{a} \, . \tag{421}$$

We have the product of the 3-dimensional tensor $\frac{\partial(\boldsymbol{W}^Q\boldsymbol{\xi})}{\partial \boldsymbol{W}^Q}$ with the vector $\boldsymbol{a}$ which gives a 2-dimensional tensor, i.e. a matrix:

$$\frac{\partial(\boldsymbol{W}^Q\boldsymbol{\xi})}{\partial \boldsymbol{W}^Q}\frac{\partial \boldsymbol{a}^T\boldsymbol{y}}{\partial \boldsymbol{y}} \ = \ \frac{\partial(\boldsymbol{W}^Q\boldsymbol{\xi})}{\partial \boldsymbol{W}^Q}\boldsymbol{a} \ = \ \boldsymbol{\xi}^T\boldsymbol{a}\boldsymbol{I} \ . \tag{422}$$

$$\frac{\partial \boldsymbol{a}^T\boldsymbol{y}}{\partial \boldsymbol{W}} \ = \ \beta \ \boldsymbol{X} \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right) \boldsymbol{X}^T \ (\boldsymbol{W}^K)^T(\boldsymbol{\xi}^T\boldsymbol{a}) \ . \tag{423}$$

The derivative with respect to $\boldsymbol{W}^K$ is

$$\frac{\partial \boldsymbol{a}^T\boldsymbol{y}}{\partial \boldsymbol{W}^K} \ = \ \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{W}^K}\frac{\partial \boldsymbol{a}^T\boldsymbol{y}}{\partial \boldsymbol{y}} \ = \ \frac{\partial \boldsymbol{y}}{\partial((\boldsymbol{W}^K)^T\boldsymbol{W}^Q\boldsymbol{\xi})}\frac{\partial((\boldsymbol{W}^K)^T\boldsymbol{W}^Q\boldsymbol{\xi})}{\partial \boldsymbol{W}^K}\frac{\partial \boldsymbol{a}^T\boldsymbol{y}}{\partial \boldsymbol{y}} \ . \tag{424}$$

$$\frac{\partial \boldsymbol{y}}{\partial((\boldsymbol{W}^K)^T\boldsymbol{W}^Q\boldsymbol{\xi})} \ = \ \beta \ \boldsymbol{X} \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right) \boldsymbol{X}^T \tag{425}$$

$$\frac{\partial \boldsymbol{a}^T\boldsymbol{y}}{\partial \boldsymbol{y}} \ = \ \boldsymbol{a} \ . \tag{426}$$

We have the product of the 3-dimensional tensor $\frac{\partial(\boldsymbol{W}\boldsymbol{\xi})}{\partial \boldsymbol{W}^K}$ with the vector $\boldsymbol{a}$ which gives a 2-dimensional tensor, i.e. a matrix:

$$\frac{\partial((\boldsymbol{W}^K)^T\boldsymbol{W}^Q\boldsymbol{\xi})}{\partial \boldsymbol{W}^K}\frac{\partial \boldsymbol{a}^T\boldsymbol{y}}{\partial \boldsymbol{y}} \ = \ \frac{\partial((\boldsymbol{W}^K)^T\boldsymbol{W}^Q\boldsymbol{\xi})}{\partial \boldsymbol{W}^K}\boldsymbol{a} \ = \ (\boldsymbol{W}^Q)^T\boldsymbol{\xi}^T\boldsymbol{a}\boldsymbol{I} \ . \tag{427}$$

$$\frac{\partial \boldsymbol{a}^T\boldsymbol{y}}{\partial \boldsymbol{W}^K} \ = \ \beta \ \boldsymbol{X} \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right) \boldsymbol{X}^T \ ((\boldsymbol{W}^Q)^T\boldsymbol{\xi}^T\boldsymbol{a}) \ . \tag{428}$$

B2.7 Sequential Softmax Associative Memory

B2.7.1 Infinite Softmax Associative Memory

We have infinite many patterns $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ that are represented by the infinite matrix

$$\boldsymbol{X} \ = \ (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots,) \ . \tag{429}$$

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

$$\boldsymbol{X}_t \ = \ (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t) \ . \tag{430}$$

The query at time t is $\boldsymbol{\xi}_t$.
The energy function at time t is E_t

$$\mathrm{E}_t \ = \ - \ \mathrm{lse}(\beta, \boldsymbol{X}_t^T\boldsymbol{\xi}_t) \ + \ \frac{1}{2}\boldsymbol{\xi}_t^T\boldsymbol{\xi}_t \ + \ \beta^{-1}\ln N \ + \ \frac{1}{2}M^2 \tag{431}$$

$$= \ - \ \beta^{-1}\ln\left(\sum_{i=1}^{T}\exp(\beta \boldsymbol{x}_i^T\boldsymbol{\xi}_t)\right) \ + \ \frac{1}{2}\boldsymbol{\xi}_t^T\boldsymbol{\xi}_t \ + \ \beta^{-1}\ln N \ + \ \frac{1}{2}M^2 \ . \tag{432}$$

The derivative of the energy E_t with respect to $\boldsymbol{\xi}_t$ is

$$\frac{\partial \mathrm{E}_t}{\partial \boldsymbol{\xi}_t} \ = \ - \ \boldsymbol{X}_t\mathrm{softmax}(\beta \boldsymbol{X}_t^T\boldsymbol{\xi}_t) \ + \ \boldsymbol{\xi}_t \ = \ - \ \boldsymbol{X}_t\boldsymbol{p}_t \ + \ \boldsymbol{\xi}_t \ , \tag{433}$$

where we used

$$\boldsymbol{p}_t \ = \ \mathrm{softmax}(\beta \boldsymbol{X}_t^T\boldsymbol{\xi}_t) \ . \tag{434}$$

The fixed point iteration is

$$\boldsymbol{\xi}_t^{\mathrm{new}} \ = \ \boldsymbol{X}_t\boldsymbol{p}_t \ = \ \boldsymbol{\xi}_t \ - \ \frac{\partial \mathrm{E}_t}{\partial \boldsymbol{\xi}_t} \ . \tag{435}$$

$$\boldsymbol{\xi}_t^{\text{new}} \ = \ \boldsymbol{X}_t \boldsymbol{p}_t \ = \ \boldsymbol{X}_t \text{softmax}(\beta \boldsymbol{X}_t^T \boldsymbol{\xi}_t) \ . \tag{436}$$

We can use an infinite pattern matrix with an infinite softmax. The pattern matrix at time t is

$$\boldsymbol{X}_t \ = \ (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t, -\alpha \boldsymbol{\xi}_t, -\alpha \boldsymbol{\xi}_t, \ldots) \ , \tag{437}$$

with the query $\boldsymbol{\xi}_t$ and $\alpha \to \infty$. The energy function at time t is E_t

$$\text{E}_t \ = \ - \ \text{lse}(\beta, \boldsymbol{X}_t^T \boldsymbol{\xi}_t) \ + \ \frac{1}{2} \boldsymbol{\xi}_t^T \boldsymbol{\xi}_t \tag{438}$$

$$= \ - \ \beta^{-1} \ln \left(\sum_{i=1}^{t} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}_t) \ + \ \sum_{i=t+1}^{\lfloor \alpha \rfloor} \exp(-\beta \alpha \| \boldsymbol{\xi}_i \|^2) \right) \ + \ \frac{1}{2} \boldsymbol{\xi}_t^T \boldsymbol{\xi}_t \ . \tag{439}$$

For $\alpha \to \infty$ and $\| \boldsymbol{\xi}_t \| \geq k > \mathbf{0}$ this becomes

$$\text{E}_t \ = \ - \ \text{lse}(\beta, \boldsymbol{X}_t^T \boldsymbol{\xi}_t) \ + \ \frac{1}{2} \boldsymbol{\xi}_t^T \boldsymbol{\xi}_t \tag{440}$$

$$= \ - \ \beta^{-1} \ln \left(\sum_{i=1}^{t} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}_t) \right) \ + \ \frac{1}{2} \boldsymbol{\xi}_t^T \boldsymbol{\xi}_t \ . \tag{441}$$

B2.7.2 Forgetting Softmax Associative Memory

We have infinite many patterns $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ that are represented by the infinite matrix

$$\boldsymbol{X} \ = \ (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots,) \ . \tag{442}$$

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

$$\boldsymbol{X}_t \ = \ (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t) \ . \tag{443}$$

The query at time t is $\boldsymbol{\xi}_t$.
The energy function with forgetting parameter γ at time t is E_t

$$\text{E}_t \ = \ - \ \text{lse}(\beta, \boldsymbol{X}_t^T \boldsymbol{\xi}_t \ - \ \gamma(t-1, t-2, \ldots, 0)^T) \ + \ \frac{1}{2} \boldsymbol{\xi}_t^T \boldsymbol{\xi}_t \ + \ \beta^{-1} \ln N \ + \ \frac{1}{2} M^2 \tag{444}$$

$$= \ - \ \beta^{-1} \ln \left(\sum_{i=1}^{T} \exp(\beta \boldsymbol{x}_i^T \boldsymbol{\xi}_t \ - \ \gamma(t-i)) \right) \ + \ \frac{1}{2} \boldsymbol{\xi}_t^T \boldsymbol{\xi}_t \ + \ \beta^{-1} \ln N \ + \ \frac{1}{2} M^2 \ . \tag{445}$$

The derivative of the energy E_t with respect to $\boldsymbol{\xi}_t$ is

$$\frac{\partial \text{E}_t}{\partial \boldsymbol{\xi}_t} \ = \ - \ \boldsymbol{X}_t \text{softmax}(\beta \boldsymbol{X}_t^T \boldsymbol{\xi}_t \ - \ \gamma(t-1, t-2, \ldots, 0)^T) \ + \ \boldsymbol{\xi}_t \ = \ - \ \boldsymbol{X}_t \boldsymbol{p}_t \ + \ \boldsymbol{\xi}_t \ , \tag{446}$$

where we used

$$\boldsymbol{p}_t \ = \ \text{softmax}(\beta \boldsymbol{X}_t^T \boldsymbol{\xi}_t) \ . \tag{447}$$

The fixed point iteration is

$$\boldsymbol{\xi}_t^{\text{new}} \ = \ \boldsymbol{X}_t \boldsymbol{p}_t \ = \ \boldsymbol{\xi}_t \ - \ \frac{\partial \text{E}_t}{\partial \boldsymbol{\xi}_t} \ . \tag{448}$$

$$\boldsymbol{\xi}_t^{\text{new}} \ = \ \boldsymbol{X}_t \boldsymbol{p}_t \ = \ \boldsymbol{X}_t \text{softmax}(\beta \boldsymbol{X}_t^T \boldsymbol{\xi}_t) \ . \tag{449}$$

B3 Properties of Softmax, Log-Sum-Exponential, Legendre Transform, Lambert W Function

For $\beta > 0$, the *softmax* is defined as

Definition B5 (Softmax)**.**

$$\boldsymbol{p} \ = \ \text{softmax}(\beta \boldsymbol{x}) \tag{450}$$

$$p_i \ = \ [\text{softmax}(\beta \boldsymbol{x})]_i \ = \ \frac{\exp(\beta x_i)}{\sum_k \exp(\beta x_k)} \ . \tag{451}$$

We also need the *log-sum-exp function* (lse), defined as

Definition B6 (Log-Sum-Exp Function)**.**

$$\mathrm{lse}(\beta, \boldsymbol{x}) \;=\; \beta^{-1} \ln \left(\sum_{i=1}^{N} \exp(\beta x_i) \right) \;. \tag{452}$$

Next, we give the relation between the softmax and the lse function.

Lemma 18. *The softmax is the gradient of the* lse*:*

$$\mathrm{softmax}(\beta \boldsymbol{x}) \;=\; \nabla_{\boldsymbol{x}} \mathrm{lse}(\beta, \boldsymbol{x}) \;. \tag{453}$$

In the next lemma we report some important properties of the lse function.

Lemma 19. *We define*

$$\mathrm{L} \;:=\; \boldsymbol{z}^T \boldsymbol{x} \;-\; \beta^{-1} \sum_{i=1}^{N} z_i \ln z_i \tag{454}$$

with $\mathrm{L} \geq \boldsymbol{p}^T \boldsymbol{x}$*. The* lse *is the maximum of* L *on the* N*-dimensional simplex* D *with* $D = \{\boldsymbol{z} \mid \sum_i z_i = 1, 0 \leqslant z_i\}$*:*

$$\mathrm{lse}(\beta, \boldsymbol{x}) \;=\; \max_{\boldsymbol{z} \in D} \boldsymbol{z}^T \boldsymbol{x} \;-\; \beta^{-1} \sum_{i=1}^{N} z_i \ln z_i \;. \tag{455}$$

The softmax $\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{x})$ *is the argument of the maximum of* L *on the* N*-dimensional simplex* D *with* $D = \{\boldsymbol{z} \mid \sum_i z_i = 1, 0 \leqslant z_i\}$*:*

$$\boldsymbol{p} \;=\; \mathrm{softmax}(\beta \boldsymbol{x}) \;=\; \arg\max_{\boldsymbol{z} \in D} \boldsymbol{z}^T \boldsymbol{x} \;-\; \beta^{-1} \sum_{i=1}^{N} z_i \ln z_i \;. \tag{456}$$

Proof. Eq. (455) is obtained from Equation (8) in [22] and Eq. (456) from Equation (11) in [22]. □

From a physical point of view, the lse function represents the "free energy" in statistical thermodynamics [22].
Next we consider the Jacobian of the softmax and its properties.

Lemma 20. *The Jacobian* J_s *of the softmax* $\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{x})$ *is*

$$\mathrm{J}_s \;=\; \frac{\partial \mathrm{softmax}(\beta \boldsymbol{x})}{\partial \boldsymbol{x}} \;=\; \beta \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T \right) \;, \tag{457}$$

which gives the elements

$$[\mathrm{J}_s]_{ij} \;=\; \begin{cases} \beta p_i (1 - p_i) & \textit{for } i = j \\ -\beta p_i p_j & \textit{for } i \neq j \end{cases} \;. \tag{458}$$

Next we show that J_s has eigenvalue 0.

Lemma 21. *The Jacobian* J_s *of the softmax function* $\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{x})$ *has a zero eigenvalue with eigenvector* $\mathbf{1}$*.*

Proof.

$$[\mathrm{J}_s \mathbf{1}]_i \;=\; \beta \left(p_i (1 - p_i) \;-\; \sum_{j, j \neq i} p_i p_j \right) \;=\; \beta \, p_i (1 \;-\; \sum_j p_j) \;= 0 \;. \tag{459}$$

□

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 22. *The Jacobian* J_s *of the softmax* $\boldsymbol{p} = \mathrm{softmax}(\beta \boldsymbol{\xi})$ *is symmetric and positive semi-definite.*

Proof. For an arbitrary $\boldsymbol{y}$, we have

$$\begin{aligned} \boldsymbol{y}^T \left(\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right) \boldsymbol{y} \;&=\; \sum_i p_i y_i^2 - \left(\sum_i p_i y_i\right)^2 \\ &=\; \left(\sum_i p_i y_i^2\right) \left(\sum_i p_i\right) - \left(\sum_i p_i y_i\right)^2 \;\geq\; 0\,. \end{aligned} \tag{460}$$

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 $a_i = y_i\sqrt{p_i}$ and $b_i = \sqrt{p_i}$. Consequently $\left(\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)$ is positive semi-definite.
Alternatively $\sum_i p_i y_i^2 - \left(\sum_i p_i y_i\right)^2$ 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 23. *The softmax $\boldsymbol{p} = \operatorname{softmax}(\beta\boldsymbol{x})$ is monotone, that is,*

$$\left(\operatorname{softmax}(\beta\boldsymbol{x}) \;-\; \operatorname{softmax}(\beta\boldsymbol{x}')\right)^T (\boldsymbol{x} \;-\; \boldsymbol{x}') \;\geq\; 0\,. \tag{461}$$

Proof. We use the mean value theorem with the symmetric matrix $\mathrm{J}_s^m = \int_0^1 \mathrm{J}_s(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{x}')\,\mathrm{d}\lambda$:

$$\operatorname{softmax}(\boldsymbol{x}) \;-\; \operatorname{softmax}(\boldsymbol{x}') \;=\; \mathrm{J}_s^m \;(\boldsymbol{x} \;-\; \boldsymbol{x}')\;. \tag{462}$$

Therefore

$$\left(\operatorname{softmax}(\boldsymbol{x}) \;-\; \operatorname{softmax}(\boldsymbol{x}')\right)^T (\boldsymbol{x} \;-\; \boldsymbol{x}') \;=\; (\boldsymbol{x} \;-\; \boldsymbol{x}')^T\, \mathrm{J}_s^m \;(\boldsymbol{x} \;-\; \boldsymbol{x}') \;\geq\; 0\,, \tag{463}$$

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 22. Since

$$\boldsymbol{x}^T \mathrm{J}_s^m \boldsymbol{x} \;=\; \int_0^1 \boldsymbol{x}^T \mathrm{J}_s(\lambda\boldsymbol{x} \;+\; (1-\lambda)\boldsymbol{x}')\,\boldsymbol{x}\,\mathrm{d}\lambda \;\geq\; 0 \tag{464}$$

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 24. *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:*

$$\|\mathrm{J}_s\|_2 \;\leqslant\; 2\,m\,\beta\,, \tag{465}$$
$$\|\mathrm{J}_s\|_1 \;\leqslant\; 2\,m\,\beta\,, \tag{466}$$
$$\|\mathrm{J}_s\|_\infty \;\leqslant\; 2\,m\,\beta\,. \tag{467}$$

In particular everywhere holds

$$\|\mathrm{J}_s\|_2 \;\leqslant\; \frac{1}{2}\,\beta\,. \tag{468}$$

If $p_{\max} = \max_i p_i \geq 1 - \epsilon \geq 0.5$, then for the spectral norm of the Jacobian holds

$$\|\mathrm{J}_s\|_2 \;\leqslant\; 2\,\epsilon\,\beta \;-\; 2\,\epsilon^2\,\beta \;<\; 2\,\epsilon\,\beta\,. \tag{469}$$

Proof. We consider the maximum absolute column sum norm

$$\|\boldsymbol{A}\|_1 \;=\; \max_j \sum_i |a_{ij}| \tag{470}$$

and the maximum absolute row sum norm

$$\|\boldsymbol{A}\|_\infty \;=\; \max_i \sum_j |a_{ij}|\,. \tag{471}$$

We have for $\boldsymbol{A} = \mathrm{J}_s = \beta \left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)$

$$\begin{aligned}\sum_j |a_{ij}| &= \beta \left(p_i(1-p_i) + \sum_{j,j\neq i} p_i p_j \right) = \beta\, p_i\, (1 - 2p_i + \sum_j p_j) \\ &= 2\,\beta\, p_i\, (1-p_i) \leqslant 2\, m\, \beta\,, \end{aligned} \tag{472}$$

$$\begin{aligned}\sum_i |a_{ij}| &= \beta \left(p_j\, (1-p_j) + \sum_{i,i\neq j} p_j p_i \right) = \beta\, p_j\, (1 - 2p_j + \sum_i p_i) \\ &= 2\,\beta\, p_j\, (1-p_j) \leqslant 2\, m\, \beta\,. \end{aligned} \tag{473}$$

Therefore we have

$$\|\mathrm{J}_s\|_1 \leqslant 2\, m\, \beta\,, \tag{474}$$

$$\|\mathrm{J}_s\|_\infty \leqslant 2\, m\, \beta\,, \tag{475}$$

$$\|\mathrm{J}_s\|_2 \leqslant \sqrt{\|\mathrm{J}_s\|_1 \|\mathrm{J}_s\|_\infty} \leqslant 2\, m\, \beta\,. \tag{476}$$

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 25. *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 mean value theorem states for the symmetric matrix $\mathrm{J}_s^m = \int_0^1 \mathrm{J}(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{x}')\, \mathrm{d}\lambda$:

$$\mathrm{softmax}(\boldsymbol{x}) - \mathrm{softmax}(\boldsymbol{x}') = \mathrm{J}_s^m\, (\boldsymbol{x} - \boldsymbol{x}')\,. \tag{477}$$

According to Lemma 24 for all $\tilde{\boldsymbol{x}} = \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{x}'$

$$\|\mathrm{J}_s(\tilde{\boldsymbol{x}})\|_2 \leqslant 2\, \tilde{m}\, \beta\,, \tag{478}$$

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

$$\|\mathrm{J}_s(\tilde{\boldsymbol{x}})\|_2 \leqslant 2\, m\, \beta \tag{479}$$

which also holds for the mean:

$$\|\mathrm{J}_s^m\|_2 \leqslant 2\, m\, \beta\,. \tag{480}$$

Therefore

$$\|\mathrm{softmax}(\boldsymbol{x}) - \mathrm{softmax}(\boldsymbol{x}')\| \leqslant \|\mathrm{J}_s^m\|_2\, \|\boldsymbol{x} - \boldsymbol{x}'\| \leqslant 2\, m\, \beta\, \|\boldsymbol{x} - \boldsymbol{x}'\|\,. \tag{481}$$

From Lemma 24 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 24: $m \leqslant \epsilon$. □

For completeness we present a result about cocoercivity of the softmax:

Lemma 26. *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,*

$$(\mathrm{softmax}(\boldsymbol{x}) - \mathrm{softmax}(\boldsymbol{x}'))^T\, (\boldsymbol{x} - \boldsymbol{x}') \geq \frac{1}{2\, m\, \beta} \|\mathrm{softmax}(\boldsymbol{x}) - \mathrm{softmax}(\boldsymbol{x}')\|. \tag{482}$$

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 [22]) together with Lemma 25. □

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 B7 (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*

$$f^*(\boldsymbol{x}^*) \;=\; \sup_{\boldsymbol{x}\in X}(\boldsymbol{x}^T\boldsymbol{x}^* \;-\; f(\boldsymbol{x}))\,, \quad \boldsymbol{x}^* \in X \tag{483}$$

See page 219 Def. 13.1 in [7] and page 134 in [23]. Next we define the Legendre transform, which is a more restrictive version of the convex conjugate.

Definition B8 (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*

$$f^*(\boldsymbol{x}^*) \;=\; \sup_{\boldsymbol{x}\in X}(\boldsymbol{x}^T\boldsymbol{x}^* \;-\; f(\boldsymbol{x}))\,, \quad \boldsymbol{x}^* \in X^*\,, \tag{484}$$

$$X^* \;=\; \left\{\boldsymbol{x}^* \in \mathbb{R}^n \mid \sup_{\boldsymbol{x}\in X}(\boldsymbol{x}^T\boldsymbol{x}^* \;-\; f(\boldsymbol{x})) < \infty\right\}\,. \tag{485}$$

See page 91 in [9].

Definition B9 (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*

$$f\square g : X \to [-\infty, \infty]\,,\; \boldsymbol{x} \mapsto \inf_{\boldsymbol{y}\in X}\left(f(\boldsymbol{y}) + g(\boldsymbol{x}-\boldsymbol{y})\right) \tag{486}$$

See Def. 12.1 in [7].

Lemma 27. *Let f and g be functions from X to $(-\infty, \infty]$. Then the following hold:*

1. *Convex Conjugate of norm squared*

$$\left(\frac{1}{2}\|.\|^2\right)^* \;=\; \frac{1}{2}\|.\|^2\,. \tag{487}$$

2. *Convex Conjugate of a function multiplied by scalar $0 < \alpha \in \mathbb{R}$*

$$(\alpha\, f)^* \;=\; \alpha\, f^*(./\alpha)\,. \tag{488}$$

3. *Convex Conjugate of the sum of a function and a scalar $\beta \in \mathbb{R}$*

$$(f \;+\; \beta)^* \;=\; f^* \;-\; \beta\,. \tag{489}$$

4. *Convex Conjugate of affine transformation of the arguments. Let $\boldsymbol{A}$ be a non-singular matrix and $\boldsymbol{b}$ a vector*

$$(f\,(\boldsymbol{A}\boldsymbol{x} \;+\; \boldsymbol{b}))^* \;=\; f^*\left(\boldsymbol{A}^{-T}\boldsymbol{x}^*\right) \;-\; \boldsymbol{b}^T\boldsymbol{A}^{-T}\boldsymbol{x}^*\,. \tag{490}$$

5. *Convex Conjugate of epi-sums*

$$(f\square g)^* \;=\; f^* + g^*\,. \tag{491}$$

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 [23] 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 [7].
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 [7] and Proposition 11.4.2 in [23] 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 28. *The Legendre transform of the* lse *is the negative entropy function, restricted to the probability simplex and vice versa. For the log-sum exponential*

$$f(\boldsymbol{x}) \;=\; \ln\left(\sum_{i=1}^{n}\exp(x_i)\right)\;, \tag{492}$$

the Legendre transform is the negative entropy function, restricted to the probability simplex:

$$f^{*}(\boldsymbol{x}^{*}) \;=\; \begin{cases}\sum_{i=1}^{n}x_i^{*}\ln(x_i^{*}) & \text{for } 0\leqslant x_i^{*} \text{ and } \sum_{i=1}^{n}x_i^{*}=1\\ \infty & \text{otherwise}\end{cases}\;. \tag{493}$$

For the negative entropy function, restricted to the probability simplex:

$$f(\boldsymbol{x}) \;=\; \begin{cases}\sum_{i=1}^{n}x_i\ln(x_i) & \text{for } 0\leqslant x_i \text{ and } \sum_{i=1}^{n}x_i=1\\ \infty & \text{otherwise}\end{cases}\;. \tag{494}$$

the Legendre transform is the log-sum exponential

$$f^{*}(\boldsymbol{x}^{*}) \;=\; \ln\left(\sum_{i=1}^{n}\exp(x_i^{*})\right)\;, \tag{495}$$

Proof. See page 93 Example 3.25 in [9] and [22]. If f is a regular convex function (lower semi-continuous convex function), then $f^{**}=f$ according to page 135 Exercise 11.2.3 in [23]. If f is lower semi-continuous and convex, then $f^{**}=f$ according to Theorem 13.37 (Fenchel-Moreau) in [7]. The log-sum-exponential is continuous and convex. □

Lemma 29. *Let $\boldsymbol{X}\boldsymbol{X}^{T}$ be non-singular and X a Hilbert space. We define*

$$X^{*} \;=\; \left\{\boldsymbol{a}\mid 0\;\leqslant\;\boldsymbol{X}^{T}\left(\boldsymbol{X}\boldsymbol{X}^{T}\right)^{-1}\boldsymbol{a}\;,\;\;\mathbf{1}^{T}\boldsymbol{X}^{T}\left(\boldsymbol{X}\boldsymbol{X}^{T}\right)^{-1}\boldsymbol{a}\;=\;1\right\}\;. \tag{496}$$

and

$$X^{v} \;=\; \left\{\boldsymbol{a}\mid\boldsymbol{a}=\boldsymbol{X}^{T}\boldsymbol{\xi}\;,\;\;\boldsymbol{\xi}\in X\right\}\;. \tag{497}$$

The Legendre transform of $\operatorname{lse}(\beta,\boldsymbol{X}^{T}\boldsymbol{\xi})$ *with* $\boldsymbol{\xi}\in X$ *is*

$$\left(\operatorname{lse}(\beta,\boldsymbol{X}^{T}\boldsymbol{\xi})\right)^{*}\left(\boldsymbol{\xi}^{*}\right) \;=\; \left(\operatorname{lse}(\beta,\boldsymbol{v})\right)^{*}\left(\boldsymbol{X}^{T}\left(\boldsymbol{X}\boldsymbol{X}^{T}\right)^{-1}\boldsymbol{\xi}^{*}\right)\;, \tag{498}$$

with $\boldsymbol{\xi}^{*}\in X^{*}$ and $\boldsymbol{v}\in X^{v}$. The domain of $\left(\operatorname{lse}(\beta,\boldsymbol{X}^{T}\boldsymbol{\xi})\right)^{*}$ is X^{*}.
Furthermore we have

$$\left(\operatorname{lse}(\beta,\boldsymbol{X}^{T}\boldsymbol{\xi})\right)^{**} \;=\; \operatorname{lse}(\beta,\boldsymbol{X}^{T}\boldsymbol{\xi})\;. \tag{499}$$

Proof. We use the definition of the Legendre transform:

$$\begin{aligned}
\left(\text{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi})\right)^* (\boldsymbol{\xi}^*) &= \sup_{\boldsymbol{\xi} \in X} \boldsymbol{\xi}^T \boldsymbol{\xi}^* - \text{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) \\
&= \sup_{\boldsymbol{\xi} \in X} \left(\boldsymbol{X}^T \boldsymbol{\xi}\right)^T \boldsymbol{X}^T \left(\boldsymbol{X} \boldsymbol{X}^T\right)^{-1} \boldsymbol{\xi}^* - \text{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) \\
&= \sup_{\boldsymbol{v} \in X^v} \boldsymbol{v}^T \boldsymbol{X}^T \left(\boldsymbol{X} \boldsymbol{X}^T\right)^{-1} \boldsymbol{\xi}^* - \text{lse}(\beta, \boldsymbol{v}) \\
&= \sup_{\boldsymbol{v} \in X^v} \boldsymbol{v}^T \boldsymbol{v}^* - \text{lse}(\beta, \boldsymbol{v}) \\
&= (\text{lse}(\beta, \boldsymbol{v}))^* (\boldsymbol{v}^*) = (\text{lse}(\beta, \boldsymbol{v}))^* \left(\boldsymbol{X}^T \left(\boldsymbol{X} \boldsymbol{X}^T\right)^{-1} \boldsymbol{\xi}^*\right) ,
\end{aligned} \tag{500}$$

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 [9], the equations for the maximum $\max_{\boldsymbol{v} \in X^v} \boldsymbol{v}^T \boldsymbol{v}^* - \text{lse}(\beta, \boldsymbol{v})$ 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 $\left(\text{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi})\right)^*$ is X^*, since on page 93 Example 3.25 in [9] it was shown that outside X^* the $\sup_{\boldsymbol{v} \in X^v} \boldsymbol{v}^T \boldsymbol{v}^* - \text{lse}(\beta, \boldsymbol{v})$ is not bounded.
Using

$$\boldsymbol{p} = \text{softmax}(\beta \boldsymbol{X}^T \boldsymbol{\xi}) , \tag{501}$$

the Hessian of $\text{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi})$

$$\frac{\partial^2 \text{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi})}{\partial \boldsymbol{\xi}^2} = \beta \boldsymbol{X} \left(\text{diag}(\boldsymbol{p}) - \boldsymbol{p} \boldsymbol{p}^T\right) \boldsymbol{X}^T \tag{502}$$

is positive semi-definite since $\text{diag}(\boldsymbol{p}) - \boldsymbol{p} \boldsymbol{p}^T$ is positive semi-definite according to Lemma 22. Therefore $\text{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi})$ 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 [23]. If f is lower semi-continuous and convex, then $f^{**} = f$ according to Theorem 13.37 (Fenchel-Moreau) in [7]. Consequently we have

$$\left(\text{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi})\right)^{**} = \text{lse}(\beta, \boldsymbol{X}^T \boldsymbol{\xi}) . \tag{503}$$

□

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 B10 (Lambert Function)**.** *The* Lambert W function *is the inverse function of*

$$f(y) = y e^y . \tag{504}$$

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

$$W(x) = y \Rightarrow y e^y = x . \tag{505}$$

We present some identities for the Lambert W function:

Lemma 30. *Identities for the Lambert W function are*

$$W(x)\, e^{W(x)} \;=\; x\,, \tag{506}$$
$$W(xe^{x}) \;=\; x\,, \tag{507}$$
$$e^{W(x)} \;=\; \frac{x}{W(x)}\,, \tag{508}$$
$$e^{-W(x)} \;=\; \frac{W(x)}{x}\,, \tag{509}$$
$$e^{nW(x)} \;=\; \left(\frac{x}{W(x)}\right)^{n}\,, \tag{510}$$
$$W_0\left(x\, \ln x\right) \;=\; \ln x \quad \textit{for } x \;\geq\; \frac{1}{e}\,, \tag{511}$$
$$W_{-1}\left(x\, \ln x\right) \;=\; \ln x \quad \textit{for } x \;\leqslant\; \frac{1}{e}\,, \tag{512}$$
$$W(x) \;=\; \ln \frac{x}{W(x)} \quad \textit{for } x \;\geq\; -\,\frac{1}{e}\,, \tag{513}$$
$$W\left(\frac{n\, x^{n}}{W\left(x\right)^{n-1}}\right) \;=\; n\, W(x) \quad \textit{for } n, x \;>\; 0\,, \tag{514}$$
$$W(x) \;+\; W(y) \;=\; W\left(x\, y\, \left(\frac{1}{W(x)} \;+\; \frac{1}{W(y)}\right)\right) \quad \textit{for } x, y \;>\; 0\,, \tag{515}$$
$$W_0\left(-\,\frac{\ln x}{x}\right) \;=\; -\,\ln x \quad \textit{for } 0 \;<\; x \;\leqslant\; e\,, \tag{516}$$
$$W_{-1}\left(-\,\frac{\ln x}{x}\right) \;=\; -\,\ln x \quad \textit{for } x \;>\; e\,, \tag{517}$$
$$e^{-\,W(-\,\ln x)} \;=\; \frac{W(-\,\ln x)}{-\,\ln x} \quad \textit{for } x \;\neq\; 1\,. \tag{518}$$

We also present some special values for the Lambert W function:

Lemma 31.

$$W(0) \;=\; 0\,, \tag{519}$$
$$W(e) \;=\; 1\,, \tag{520}$$
$$W\left(-\frac{1}{e}\right) \;=\; -1\,, \tag{521}$$
$$W\left(e^{1+e}\right) \;=\; e\,, \tag{522}$$
$$W\left(2 \ln 2\right) \;=\; \ln 2\,, \tag{523}$$
$$W(1) \;=\; \Omega\,, \tag{524}$$
$$W(1) \;=\; e^{-W(1)} \;=\; \ln\left(\frac{1}{W(1)}\right) \;=\; -\,\ln W(1)\,, \tag{525}$$
$$W\left(-\frac{\pi}{2}\right) \;=\; \frac{i\pi}{2}\,, \tag{526}$$
$$W(-1) \;\approx\; -0.31813 + 1.33723i\,, \tag{527}$$

where the Omega constant Ω is

$$\Omega \;=\; \left(\int_{-\infty}^{\infty} \frac{\mathrm{d}t}{\left(e^{t} \;-\; t\right)^{2} \;+\; \pi^{2}}\right)^{-1} \;-\; 1 \;\approx\; 0.56714329\,. \tag{528}$$

B4 Modern Hopfield Networks: Binary States (Krotov and Hopfield)

B4.1 Modern Hopfield Networks: Introduction

B4.1.1 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 [24]. The NTM can write to the memory and can read from it. A memory network [48] 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 $\mathrm{softmax}$ [40, 41]. Linear Memory Network use a linear autoencoder for sequences as a memory [13].
To enhance RNNs with additional associative memory like Hopfield networks have been proposed [3, 4]. 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 [54]. LSTMs with associative memory via Holographic Reduced Representations have been proposed [15].
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 [24]. End to end memory networks (EMN) make the attention scheme of memory networks [48] differentiable by replacing $\arg\max$ through a $\mathrm{softmax}$ [40, 41]. EMN with dot products became very popular and implement a key-value attention [16] for self-attention. An enhancement of EMN is the transformer [45, 46] and its extensions [17]. The transformer had great impact on the natural language processing (NLP) community as new records in NLP benchmarks have been achieved [45, 46]. MEMO uses the transformer attention mechanism for reasoning over longer distances [5]. Current state-of-the-art for language processing is a transformer architecture called "the Bidirectional Encoder Representations from Transformers" (BERT) [19, 20].

B4.1.2 Modern Hopfield networks: Overview

The storage capacity of classical binary Hopfield networks [27] 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 [33]. The same bound holds for nonlinear learning rules [32]. Using tricks-of-trade and allowing small retrieval errors, the storage capacity is about $0.138d$ [14, 25, 44]. If the learning rule is not related to the Hebb rule then up to d patterns can be stored [1]. Using a Hopfield networks with non-zero diagonal matrices, the storage can be increased to $Cd\ln(d)$ [21]. In contrast to the storage capacity, the number of energy minima (spurious states, stable states) of Hopfield networks is exponentially in d [42, 11, 47].
Recent advances in the field of binary Hopfield networks [27] 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 [6]. 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 [28, 29, 18] (see [29] 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 $c_n > 2(2n-3)!!$ [18].
Interaction functions of the form $F(x) = \exp(x)$ lead to *exponential* storage capacity of $2^{d/2}$ where all stored pattern are fixed points but the radius of attraction vanishes [18]. It has been shown that the network converges even after one update [18].

B4.2 Energy and Update Rule for Binary Modern Hopfield Networks

We follow [18] where the goal is to store a set of input data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ that are represented by the matrix

$$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \ . \tag{529}$$

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 [28] defined the energy function E with the

interaction function F that evaluates the dot product between patterns $\boldsymbol{x}_i$ and the actual state $\boldsymbol{\xi}$:

$$\mathrm{E} \ = \ - \sum_{i=1}^{N} F\left(\boldsymbol{\xi}^T \boldsymbol{x}_i\right) \tag{530}$$

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 [28]. Krotov and Hopfield [28] suggested for minimizing this energy an asynchronous updating dynamics $T = (T_j)$ for component ξ_j:

$$T_j(\boldsymbol{\xi}) \ := \ \mathrm{sgn}\left[\sum_{i=1}^{N}\left(F\left(x_{ij} \ + \ \sum_{l \neq j} x_{il} \ \xi_l\right) \ - \ F\left(- \ x_{ij} \ + \ \sum_{l \neq j} x_{il} \ \xi_l\right)\right)\right] \tag{531}$$

While Krotov and Hopfield used $F(a) = a^n$, Demircigil et al. [18] 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 B10 (Storage Capacity for Binary Modern Hopfield Nets (Demircigil et al. 2017))**.** *Consider the generalized Hopfield model with the dynamics described in Eq.* (531) *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*

$$\Pr\left(\exists i \ \exists j : \ T_j\left(\widetilde{\boldsymbol{x}}_i\right) \ \neq \ x_{ij}\right) \ \rightarrow \ 0 \ ,$$

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 [18]. □

The number of patterns $N = \exp(\alpha d) + 1$ is exponential in the number d of components. The result

$$\Pr\left(\exists i \ \exists j : \ T_j\left(\widetilde{\boldsymbol{x}}_i\right) \ \neq \ x_{ij}\right) \ \rightarrow \ 0$$

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 ϱN and the number $N = \exp(\alpha d) + 1$ of pattern that can be stored.

Theorem B10 in particular implies that

$$\Pr\left(\exists i \ \exists j : \ T_j\left(\boldsymbol{x}_i\right) \ \neq \ x_{ij}\right) \ \rightarrow \ 0$$

as $d \rightarrow \infty$, i.e. with a probability converging to 1, all the patterns are fixed points of the dynamics. In this case we can have $\alpha \rightarrow \frac{I(1)}{2} = \ln(2)/2$.

Krotov and Hopfield define the update dynamics $T_j(\boldsymbol{\xi})$ in Eq. (531) via energy differences of the energy in Eq. (530). First we express the energy in Eq. (530) with $F(a) = \exp(a)$ [18] by the lse function. Then we use the mean value theorem to express the update dynamics $T_j(\boldsymbol{\xi})$ in Eq. (531) by the softmax function. For simplicity, we set $\beta = 1$ in the following. There exists a $v \in [-1, 1]$ with

$$T_j(\boldsymbol{\xi}) \ = \ \mathrm{sgn}\left[\mathrm{E}(\xi_j = 1) \ - \ \mathrm{E}(\xi_j = -1)\right] \ = \ \mathrm{sgn}\left[- \ \exp(\mathrm{lse}(\xi_j = 1)) \ + \ \exp(\mathrm{lse}(\xi_j = -1))\right] \tag{532}$$

$$= \ \mathrm{sgn}\left[(2\boldsymbol{e}_j)^T \nabla_{\boldsymbol{\xi}} \mathrm{E}(\xi_j = v)\right] \ = \ \mathrm{sgn}\left[\exp(\mathrm{lse}(\xi_j = v)) \ (2\boldsymbol{e}_j)^T \frac{\mathrm{lse}(\xi_j = v)}{\partial \boldsymbol{\xi}}\right]$$

$$= \ \mathrm{sgn}\left[\exp(\mathrm{lse}(\xi_j = 1)) \ (2\boldsymbol{e}_j)^T \boldsymbol{X} \mathrm{softmax}(\boldsymbol{X}^T \boldsymbol{\xi}(\xi_j = v))\right]$$

$$= \ \mathrm{sgn}\left[[\boldsymbol{X} \mathrm{softmax}(\boldsymbol{X}^T \boldsymbol{\xi}(\xi_j = v))]_j\right] \ = \ \mathrm{sgn}\left[[\boldsymbol{X}\boldsymbol{p}(\xi_j = v)]_j\right] \ ,$$

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

$$\boldsymbol{p} \ = \ \mathrm{softmax}(\boldsymbol{X}^T \boldsymbol{\xi}) \ . \tag{533}$$

B5 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. B2). 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 $\mathrm{softmax} \in \mathbb{R}^N$ changed to a row vector, we obtain for the update rule Eq. (17) multiplied by $\boldsymbol{W}_V$:

$$\mathrm{softmax}\left(1/\sqrt{d_k}\, \boldsymbol{Q}\, \boldsymbol{K}^T\right)\, \boldsymbol{V}\,. \tag{534}$$

This formula is the transformer attention.

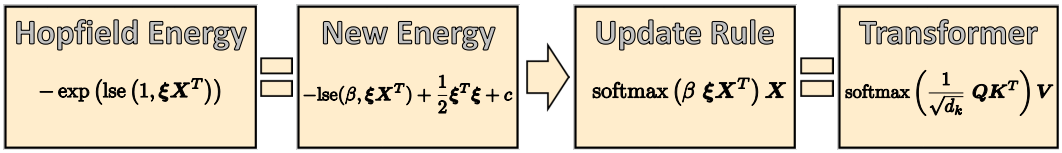

Figure B2: We generalized the energy of binary modern Hopfield networks for allowing continuous states while keeping convergence and storage capacity properties. We defined for the new energy also a new update rule that minimizes the energy. The new update rule is the attention mechanism of the transformer. Formulae are modified to express $\mathrm{softmax}$ as row vector as for transformers. "="-sign means "keeps the properties".

References

[1] Y. Abu-Mostafa and J.-M-StJacques. Information capacity of the Hopfield model. *IEEE Transactions on Information Theory*, 31, 1985.

[2] F. Alzahrani and A. Salem. Sharp bounds for the lambert w function. *Integral Transforms and Special Functions*, 29(12):971–978, 2018.

[3] 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.

[4] J. Ba, G. E. Hinton, V. Mnih, J. Z. Leibo, and C. Ionescu. Using fast weights to attend to the recent past. *ArXiv*, 2016.

[5] 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*, 2020.

[6] A. Barra, M. Beccaria, and A. Fachechi. A new mechanical approach to handle generalized Hopfield neural networks. *Neural Networks*, 106:205–222, 2018.

[7] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Cham: Springer International Publishing, 2nd edition, 2017.

[8] C. Bordenave and D. Chafaï. Around the circular law. *Probab. Surveys*, 9:1–89, 2012.

[9] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 7th edition, 2009.

[10] 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.

[11] 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.

[12] 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.

[13] A. Carta, A. Sperduti, and D. Bacciu. Encoding-based memory modules for recurrent neural networks. *ArXiv*, 2020.

[14] 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.

[15] 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.

[16] M. Daniluk, T. Rocktäschel, J. Welbl, and S. Riedel. Frustratingly short attention spans in neural language modeling. *ArXiv*, 2017. appeared in ICRL 2017.

[17] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and L. Kaiser. Universal transformers. *ArXiv*, 2018. Published at ICLR 2019.

[18] 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.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *ArXiv*, 2018.

[20] 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.

[21] V. Folli, M. Leonetti, and G. Ruocco. On the maximum storage capacity of the Hopfield model. *Frontiers in Computational Neuroscience*, 10(144), 2017.

[22] B. Gao and L. Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *ArXiv*, 2017.

[23] D. J. H. Garling. *Analysis on Polish Spaces and an Introduction to Optimal Transportation*. London Mathematical Society Student Texts. Cambridge University Press, 2017.

[24] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *ArXiv*, 2014.

[25] 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.

[26] 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.

[27] 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.

[28] 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 29*, pages 1172–1180. Curran Associates, Inc., 2016.

[29] D. Krotov and J. J. Hopfield. Dense associative memory is robust to adversarial inputs. *Neural Computation*, 30(12):3151–3167, 2018.

[30] T. Lipp and S. Boyd. Variations and extension of the convex–concave procedure. *Optimization and Engineering*, 17(2):263–287, 2016.

[31] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues or some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[32] C. Mazza. On the storage capacity of nonlinear neural networks. *Neural Networks*, 10(4):593–597, 1997.

[33] 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.

[34] R. R. Meyer. Sufficient conditions for the convergence of monotonic mathematical programming algorithms. *Journal of Computer and System Sciences*, 12(1):108–121, 1976.

[35] 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.

[36] A. Rangarajan, S. Gold, and E. Mjolsness. A novel optimizing network architecture with applications. *Neural Computation*, 8(5):1041–1060, 1996.

[37] A. Rangarajan, A. Yuille, and Eric E. Mjolsness. Convergence properties of the softassign quadratic assignment algorithm. *Neural Computation*, 11(6):1455–1474, 1999.

[38] A. Soshnikov. A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *J. Statist. Phys.*, 108(5-6):1033–1056, 2002.

[39] 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.

[40] 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.

[41] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. End-to-end memory networks. *ArXiv*, 2015.

[42] 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.

[43] T. Tao and V. Vu. Random matrices: Universality of ESDs and the circular law. *Ann. Probab.*, 38:2023–2065, 2010. With an appendix by M. Krishnapur.

[44] J. J. Torres, L. Pantic, and Hilbert H. J. Kappen. Storage capacity of attractor neural networks with depressing synapses. *Phys. Rev. E*, 66:061910, 2002.

[45] 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.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *ArXiv*, 2017.

[47] G. Wainrib and J. Touboul. Topological and dynamical complexity of random neural networks. *Phys. Rev. Lett.*, 110:118101, 2013.

[48] J. Weston, S. Chopra, and A. Bordes. Memory networks. *ArXiv*, 2014.

[49] J. C. F. Wu. On the convergence properties of the em algorithm. *Ann. Statist.*, 11(1):95–103, 1983.

[50] Y. Q. Yin. Limiting spectral distribution for a class of random matrices. *Journal of Multivariate Analysis*, 20(1):50–68, 1986.

[51] 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.

[52] A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.

[53] W. I. Zangwill. *Nonlinear programming: a unified approach*. Prentice-Hall international series in management. Englewood Cliffs, N.J., 1969.

[54] W. Zhang and B. Zhou. Learning to update auto-associative memory in recurrent neural networks for improving sequence memorization. *ArXiv*, 2017.