On Hopfield Network Attractor Localization and Visualization

Marijan Sorić
Kavli Institute for Systems Neuroscience
Norwegian University of Science and Technology
7491 Trondheim, Norway
marijaso@stud.ntnu.no
https://github.com/soricm/Hopfield-Network-Attractor

Abstract

Hopfield networks are models of associative memory used for pattern recognition and memory retrieval from partial and noisy information. This paper focuses on the localization and visualization of attractors in Hopfield networks, examining their location, number, and network capacity. Using 2D and 3D visualizations, we explore the network's energy landscape and its ability to store and retrieve memories. The study also analyzes the network's dynamics, update rules, and the connection between the energy function and attractor states.

1 Introduction

Artificial neural networks can be used at the same time for data analysis, finding biological representations, features in the data; and for modelling, by generating representation from neural circuitry. By developing neural networks, we want to reproduce the synaptic plasticity from the brain: the ability of synapses to strenghten or weaken over time, in response to increases or decreases in their activity [1]. The paradigm of Hebbian Learning (1949) describes synaptic interactions with the simple rule: *"Neurons that fire together, wire together"* [2].

The Hopfield network is a simple model of associative memory[1], created in 1982 [3]. This model is able to perform association, *i.e.* pattern recognition cued by partial and noisy information. The Hopfield network can store and recall those patterns. A partial cue activates part of the assembly, which is then completed by the dynamics. Interactions between neurons follow the Hebbian's law of association, that depends on the patterns that we want to memorize. Thus, we are able to retrieve dynamically those memories from a partial and noisy information.

The aim of this work is to have a better understanding of Hopfield network's space attractor visualization. Where are they? How many are they? What is the capacity of the network? We will also try to visualize results in 2D and 3D cases.

2 Hopfield network

2.1 Single neuron model background

Neurons can be divided into three different parts: *dendrite* (input), *soma* (central processing unit) and *axon* (output), that are connected by synapses. Action potentials are the neuronal signals, or

[1]The ability to learn and remember the relationship between unrelated items.

electrical pulses, that occur within a small time window and can be recorded. The membrane potential $u(t)$ is defined (at every time) as the electrical charge difference between inside and outside the cell. An input can depolarize the cell if $u(t) > u_{rest}$ or hyperpolarize $u(t) \leq u_{rest}$. Hodgkin and Huxley discovered in 1963 that the shape of the action potential is controlled by ions in the membrane [4]. They also proposed a single neuron model with an electronic circuit, that links the microscopic level of ion channels to the macroscopic level of currents and action potentials. However, this model was mathematically complex because it relied on non-linear differential equations. That is why the Leaky Integrate-and-Fire Model was introduced, a more general model [1].

2.2 Hopfield network definition

We consider a network with N neurons. Each of them can be active or inactive at each time step. There are 2^N different possible states. Then, the network can be fully characterized at any time step t, by the vector: $S(t) \in \{-1, 1\}^N$.

One can notice that Hopfield networks are highly flexible and can handle inputs of various dimensions (1D, 2D, 3D, 1000D...) As a matter of fact, we can visualize a network as a line: vector in $\mathbb{R}^N$, or as a surface, where the input are images in $\mathbb{R}^{\sqrt{N}\times\sqrt{N}}$ or even videos, $\mathbb{R}^{\sqrt{N}\times\sqrt{N}\times\sqrt{N}}$. We can always reshape the input data to suit our visualization needs.

2.3 Store memories

Neurons interactions are described by $W \in \mathbb{R}^{N\times N}$, that is a symmetrical Hollow matrix[2], with $(N-1)(N-2)$ parameters. Since there is no self interaction, we set $W_{ii} = 0$. A pattern, or memory, denoted $\xi^\mu \in \{-1, 1\}^N$, is a set of activity states. Considering asymptotic dynamics, if we store a pattern ξ^μ, and start with neuron firing states similar to the pattern, we hope that the network converges toward the attractor ξ^μ, *i.e.* $\lim_{t\to+\infty} S(t) = \xi^\mu$. We set W, in the Hebbian-style synaptic matrix, to store patterns from ξ:

$$W \triangleq \frac{1}{N}\left[\sum_{\mu=1}^{K} \xi^\mu (\xi^\mu)^T - \operatorname{diag}\left(\left((\xi_i^\mu)^2\right)_i\right)\right]$$

In other words, $W_{ii} = 0$ and $W_{ij} = \frac{1}{N}\sum_{\mu=1}^{K} \xi_i^\mu \xi_j^\mu$. A pattern is stored when it is a fixed point of the dynamics.

2.4 Update rules

The network state can be updated either synchronously or asynchronously at each time step Δt. We will prefer the latter because it might be more biologically plausible, using a sequential order. The evolution of the network is deterministic, based on the weighted sum of the activities of other neurons[3].

$$S_i(t+\Delta t) \triangleq \operatorname{sgn}\left(\sum_{j=1}^{N} W_{ij} S_j(t)\right) = \operatorname{sgn}(\langle W_i^T, S(t)\rangle)$$

Where $\langle ., . \rangle$ denotes the scalar product. Note that we did not write this formula with matrix $S(t+\Delta t)$ but only for $S_i(t+\Delta t)$ because we consider the asynchronous update rule.

2.5 Energy landscape

Hopfield's work introduced an energy function, also called Hamiltonian, that decreases during random asynchronous updates [3].

[2] A is a Hallow matrix whenever $\forall i,\ A_{ii} = 0$.

[3] Note that sometimes, we get $\langle W_i^T, S(t)\rangle = 0$. We will consider that $\operatorname{sgn}(0) = -1$. The bias is set to 0.

$$E(S) \triangleq -\frac{1}{2}\langle WS, S\rangle$$

3 Attractors analysis

Now, let's focus on the definition and characterization of discrete attractors from Hopfield network.

3.1 Localization

Given an initial state $S(0)$, we can asynchronously update each neuron S_i. For large t, the system can have one of these different asymptotic behaviors:

- Fixed point (attractors). The state has converged, and doesn't change anymore. If the initial state is similar to a pattern stored, we hope that the network converges toward the pattern: the network would have successfully retrieved the memory.
- Limit cycles. The state hasn't converged toward a single state, but there are trapped in a cycle defined by at least 2 states.
- Chaos attractors. When neither of the above cases apply. There is no single attractors or "pattern" in the sequence of "limit states" (chaotic or random).

The energy function E decreases at each state flip of a neuron, during random asynchronous updates of the network. Therefore, the dynamics of the network are attracted to these local minima ($E(S(t))$ will tend to the local minima). Thus, the dynamics tend to attractors, that are fixed points of the dynamics and local minima of E. Now, let's try to characterize and locate these local minima, *i.e.* what are the states S that locally minimizes E.

Since W is a symmetric and real matrix, the Spectral theorem holds. We can have a new vector basis: $\mathbb{R}^N = \bigoplus_{\lambda\in \mathrm{Sp}(W)} E_\lambda(W) = \bigoplus_{i=1}^N \mathrm{Span}(V_i)$, where V_i are eigenvectors of W associated with their eigenvalue λ_i. Each state S, can be expressed in this new basis $\mathrm{Span}((V_i)_i) = \mathbb{R}^N$. Let $S = \sum_{i=1}^N \alpha_i V_i$. Then[4],

$$\begin{aligned}
E(S) &= -\frac{1}{2}\langle WS, S\rangle \geq -\frac{N}{2}\rho(W) \\
&= -\frac{1}{2}\left\langle \sum_{i=1}^N \alpha_i W V_i, \sum_{j=1}^N \alpha_j V_j \right\rangle \\
&= -\frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N \alpha_i\alpha_j\langle \lambda_i V_i, V_j\rangle \\
&= -\frac{1}{2}\sum_{i=1}^N \alpha_i^2\lambda_i
\end{aligned}$$

Thus,

$$\min_S E = \min_{S=\sum \alpha_i V_i} -\frac{1}{2}\sum_{i=1}^N \alpha_i^2\lambda_i \geq -\frac{1}{2}\left(\sum_{i=1}^N \alpha_i^2\right)\max \mathrm{Sp}(W) = -\frac{N}{2}\max \mathrm{Sp}(W)$$

We obtained a lower bound for E, because here $\langle S, S\rangle = N$. E is minimal when S is collinear with the eigenvector associated with the highest eigenvalue. However, S might not belong to $\{-1, 1\}^N$. Note that as $\mathrm{Tr}(W) = 0$, and if $W \neq 0_{N,N}$, then there exist at least on eigenvalue strictly positive

[4]Where we defined the spectral radius, $\rho(W) = \{|\lambda| : \lambda \in \mathrm{Sp}(W)\}$.

and one strictly negative. So W can't be positive semi definite (case where $E = -\frac{1}{2}S^T W S \geq 0$). The local minima of E depends on the singular value decomposition of W that catches the effect of the update rule $W_i^T S(t)$.

3.2 Associative memory capacity

Considering a Hopfield network with N neurons, the goal is to store patterns, and be able to retrieve them. However, the maximum number of patterns that can be stored depends on the number of neurons, but also on the designing of the energy function [5]. Since attractors are located in local minima of E, this function must accurately describe interactions between state and patterns. The storage capacity with small retrieved error in the vanilla Hopfield network is in $\mathcal{O}(N)$. More precisely $K^{\max} \approx 0.14N$ [3]. As a matter of fact, we can write the Hamiltonian for the standard Hopfield network, with K patterns[5]:

$$\begin{aligned} E(S) &= -\frac{1}{2}\langle WS, S\rangle \\ &= -\frac{1}{2}\left[\frac{1}{N}\sum_{\mu=1}^{K}\langle \xi^\mu (\xi^\mu)^T S, S\rangle - \left\langle \operatorname{diag}\left(\left((\xi_i^\mu)^2\right)_i\right) S, S\right\rangle\right] \\ &= -\frac{1}{2N}\left[\sum_{\mu=1}^{K}\langle \xi^\mu, S\rangle^2 - \langle \xi^\mu \odot \xi^\mu \odot S, S\rangle\right] \in \mathbb{R} \end{aligned}$$

Note that in case, $W_{ii} = 0$, E is not necessarly negative. Here the function underlying is $F : x \mapsto x^2$, that characterizes the energy function. In some cases, it is possible to obtain $K^{\max} \approx N$ by modifying it [6]. J. Hopfield explains in *Dense associative memory for pattern recognition*, that the vanilla Hopfield gets confused when many stored memories have the same contribution to E. If ξ^μ $\xi^{\mu'}$ are close to each other (high overlap), instead of having two distinct local minima, we might have a single local minimum in between. By using F a monomial function $F : x \mapsto x^n$ [7] or rectified polynomial energy function, the storage capacity of the network, increases drastically: $K^{\max} \approx \alpha_n N^{n-1}$ [5].

3.3 Visualization

We ran some experimentation [1], with various Hopfield network sizes, in order to observe where are attractors located for some specific examples. For each size of neural network N, we plot the energy function in the space $E : [-1, 1]^N \to \mathbb{R}$. We also visualized the convergence of initial states with colors. We map the space with $n = 7$ points between -1 and +1 in each dimension, and run asynchronous sequential update, with $T \gg N$ in order to reach the asymptotic state.

Case $N = 2$

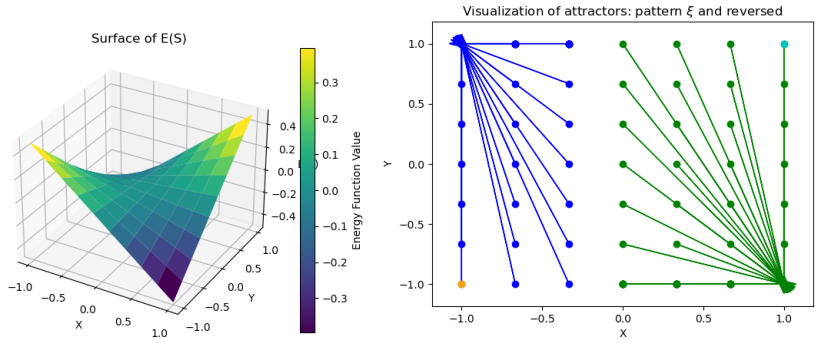

Figure 1: Energy surface and attractor visualization for $\xi = (-1, 1)$.

[5]Where $(A \odot B)_{ij} = A_{ij}B_{ij}$ is the Hadamard product: the element-wise (or pointwise) multiplication.

This is the simplest case. We observe that there are two attractors: ξ the stored pattern and $-\xi$, the reversed state. We can draw a line in the surface $[-1, 1]^2$ that delimits initial states and that will end up in ξ or $-\xi$. Note that the boundary is also influenced by the chosen threshold. Attractors are local minima of E.

Case $N = 3$

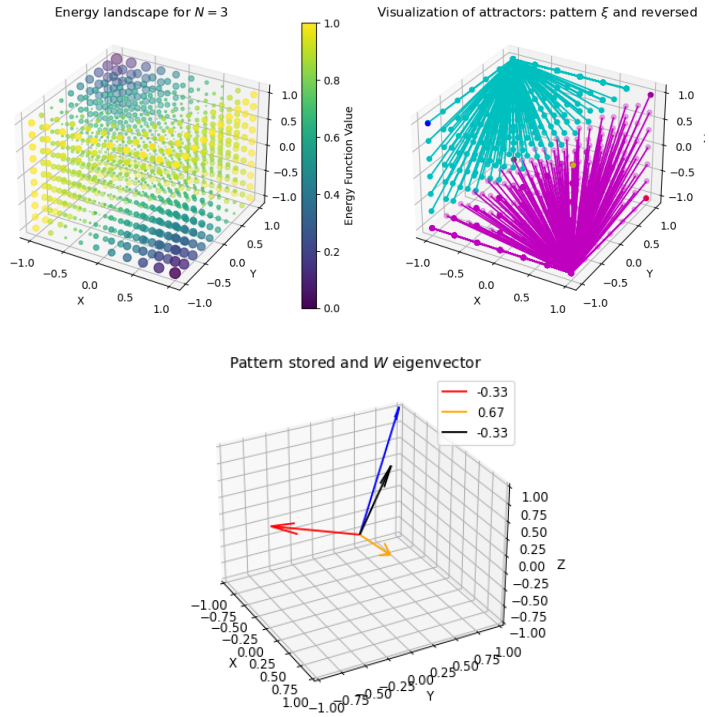

Figure 2: Energy, attractor visualization and eigenvector W for $\xi = (-1, 1, 1)$.

There is a hyperplane, *a plane here*, that draws two subspaces that defined the basin of attraction of each attractor point. In this case, we also ploted the eigenvectors of W and the attractor. We notice that the vector with the smallest eigenvalue is near ξ. Attractors are local minima of E.

Case $N = 4$

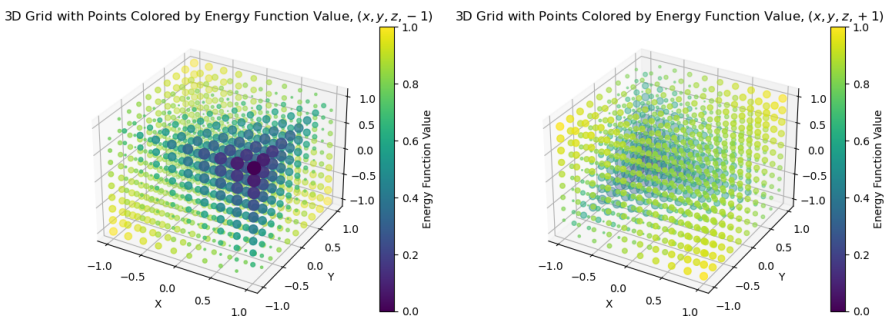

Figure 3: Energy value for ± 1 in the fourth dimension, for $\xi = (1, -1, 1, -1)$.

Again, here we have that $E(\pm\xi)$ is the lowest value, and $\pm\xi$ are attractors points.

3.3.1 Case $N = 3$, $K = 2$

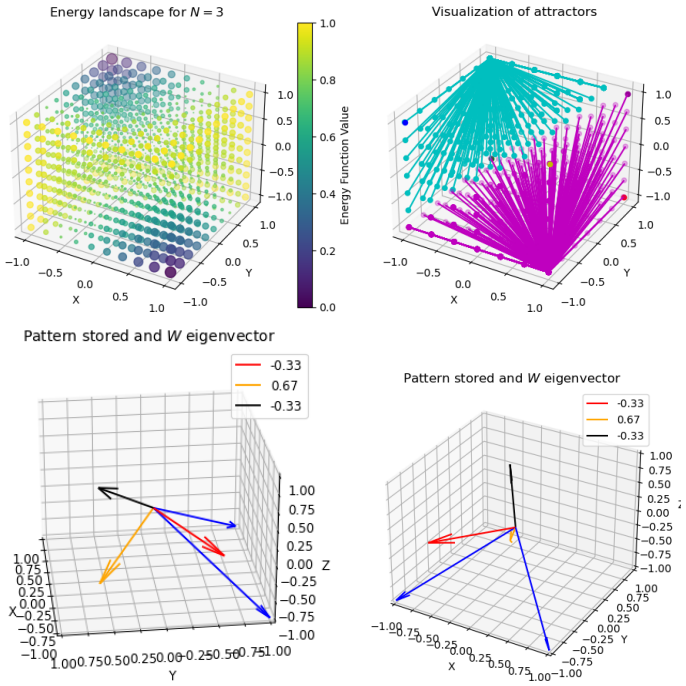

Figure 4: Energy, attractor visualization and W's eigenvector (2 orientations) for two patterns ξ^μ.

Now, we want to store three patterns. Actually, it is not possible, but we will try it to see what is happening in this case. The patterns are $(-1,-1,-1),(1,-1,-1)$. Attractors are located at $(1,-1,-1),(-1,1,1)$, and they are opposed and correspond to the second pattern. The eigenvector in black is in the same direction as these attractors and define a plane that separates the space in two basins of attraction. Attractors are local minima of E.

4 Conclusion

This study aimed to understand the localization and visualization of attractors in Hopfield networks, focusing on their spatial distribution, number, and capacity. We saw that it depends on network size N, but also on the definition of energy function. We saw that in 2D and 3D, attractors correspond to the local minima of the energy landscape. Hebbian learning states a definition for neurons interaction W, however, when there is no self interaction, $\min E$ is not a convex problem and thus has multiple local minima. By changing E, we can also improve the network capacity. Hopfield networks are characterized by discrete attractors points, that seems to be linked with W's eigenvectors, even though the relationship is more complex. If the goal is to approximate biological neurons, certain assumptions should be removed, such as symmetrical weights, binary neuron states, *etc...* While these assumptions simplify calculations, they do not accurately reflect the complexity of biological systems. Future research could explore with bigger networks and observe attractors for different level of overlapping.

References

[1] Gerstner W, Kistler WM, Naud R, Paninski L. Neuronal Dynamics online book: From single neurons to networks and models of cognition. 2014.

[2] ayan P, Abbott L.F. Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. 2001. Chap 8.

[3] Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the national academy of sciences. 1982 Apr;79(8):2554-8.

[4] Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol. 1952;117(4):500-544.

[5] Krotov D, Hopfield JJ. Dense associative memory for pattern recognition. Advances in neural information processing systems. 2016;29.

[6] Kanter I, Sompolinsky H. Associative recall of memory without errors. Physical Review A. 1987 Jan 1;35(1):380.

[7] Baldi P, Venkatesh SS. Number of stable points for spin-glasses and neural networks of higher orders. Physical Review Letters. 1987 Mar 2;58(9):913.