Sven Upgang

On the storage capacity of generalized Hopfield models

2019

Mathematik

On the storage capacity of generalized Hopfield models

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften im Fachbereich
Mathematik und Informatik
der Mathematisch-Naturwissenschaftlichen Fakultät
der Westfälischen Wilhelms-Universität Münster

vorgelegt von
Sven Upgang
aus Stadtlohn
- 2019 -

Dekan:	Prof. Dr. Xiaoyi Jiang
Erster Gutachter:	Prof. Dr. Matthias Löwe
Zweiter Gutachter:	Prof. Dr. Gerold Alsmeyer
Tag der mündlichen Prüfung:	11.02.2020
Tag der Promotion:	11.02.2020

Summary

The Hopfield model is an artificial neural network which is able to operate as an auto-associative memory. To achieve a larger storage capacity in this model, Krotov and Hopfield suggest in [KH16] to change the interaction function of its dynamics. We prove that the storage capacity increases from $N/c \log(N)$ to $N^{n-1}/c_n \log(N)$ if the interaction function x^2 of the standard Hopfield model is changed to x^n. This statement considers patterns which are chosen uniformly at random from $\{-1, 1\}^N$. Furthermore, if the patterns are generated by a Curie-Weiss model with $\beta < 1$ and the interaction function is equal to x^n for n odd, we show that the same order for the storage capacity can be obtained. Moreover, an exponential interaction function, which can be seen as a "limit of the polynomial version" if the degree goes to infinity, leads to an even higher storage capacity. For this interaction function we prove that the network is able to store at least $\exp(\alpha N)$ i.i.d. generated patterns with a positive basin of attraction.

In the last chapter we introduce a Hopfield model where the strength of the interaction between each pair of neurons can be influenced through weights. For this model we prove a lower bound for the storage capacity with the same methods as in the standard Hopfield model. Furthermore, we adopt the idea of the hierarchical Hopfield model and choose weights which depend on the graph distance between the nodes. For a regular graph and a lattice graph on a torus we calculate a lower bound for the storage capacity. In both examples we obtain similar results as in the standard Hopfield model if the weights decrease slower with the distance than the neighbourhood of a node grows. The storage capacity is lower if the weights are chosen to decrease faster. If the weights decline too fast, we cannot guarantee that the network is able to store any pattern. At last, we show that the root node in a Galton-Watson tree suggests a similar storage capacity as the root node in a regular tree.

Danksagung

Für die Betreuung meiner Promotion möchte ich mich zunächst bei meinem Doktorvater Prof. Dr. Matthias Löwe bedanken. Du hast dir stets Zeit für meine Anliegen genommen und mich mit deinen hilfreichen Ideen unterstützt. Darüber hinaus möchte ich mich für das entgegengebrachte Vertrauen sowie deine konstruktiven Anmerkungen zu meiner Arbeit bedanken.

Außerdem danke ich Herrn Prof. Dr. Alsmeyer für die hilfsbereite und wissenschaftliche Betreuung als Zweitgutachter.

Des Weiteren bedanke ich mich bei Dr. Franck Vermet, Dr. Judith Heusel und Mete Demircigil für die wissenschaftliche Zusammenarbeit.

Ebenso möchte ich mich beim gesamten Institut für Mathematische Stochastik, für das angenehme und produktive Arbeitsklima, und insbesondere bei den Doktoranden, für eure fachliche und freundschaftliche Unterstützung, bedanken.

Ganz herzlich möchte ich mich auch bei meiner Familie, besonders bei meinen Eltern für ihre Unterstützung und bei meinem Bruder Frank für das Korrekturlesen meiner Arbeit, bei meiner Freundin Hannah für ihren Rückhalt und bei meinen Freunden bedanken.

Contents

1. Introduction to the standard Hopfield model

The standard version of the Hopfield model was first introduced in 1982 by John J. Hopfield [Hop82]. In a similar way as simple electric circuits can be combined to build a computer solving complex tasks, Hopfield asks if there is a natural way how a large number of simple components can build collective computational power. Inspired by biological processes of brain cells and the way they transfer information via electrochemical signals, the author postulates a mathematical model where these collective effects occur. The model is not meant to explain the functionality of the brain, but rather to show how simple dynamics on a microscopic level can lead to more powerful computations on a macroscopic level.

The brain consists of a huge amount of electrically excitable cells called neurons [Pur+13]. These neurons are connected to each other via synapses and communicate with the help of neurotransmitters. The so-called resting potential of each neuron can be affected by incoming signals of other neurons. When the resulting postsynaptic potential reaches a certain level, the neuron is triggered to send signals itself. While the computational power of a single neuron is rather limited, the brain as a collection of neurons is able to solve a variety of complex tasks. It is by far not trivial how this functionality emerges from the interaction of neurons.

The model which Hopfield described in his paper [Hop82] shows how an algorithm defined on the level of neurons generates a specific behaviour of the whole network. The Hopfield model consists of N neurons, where each of them takes a value in a binary state space S. Originally, Hopfield used the set $\{0, 1\}$ as a state space but for a cleaner notation we will continue with $S = \{-1, 1\}$. These values correspond to the neuron "not firing" (-1) and "firing at a maximal rate" (1) [Hop82]. As time goes by, the configuration of the network changes based on its current state. In each time step a neuron is chosen uniformly at random and readjusts its value. The updating procedure works as follows: Assuming the current configuration is given by $\sigma = (\sigma_j)_j \in S^N$, then neuron $i \in \{1, \dots, N\}$ is updating to

$$\sigma_i^{new} = \begin{cases} 1, & \text{if } h_i(\sigma) := \sum_{j \neq i} W_{ij}\sigma_j > \theta_i \\ -1, & \text{if } h_i(\sigma) < \theta_i \end{cases}.$$

Here, $h_i(\sigma)$ is a weighted sum of input signals from other neurons and is called postsynaptic potential. This potential triggers the neuron to fire if it is above a certain level θ_i, and can be seen as a simplified version of the observations from neuroscience. The key factor to create a collective behaviour is the right choice of weights. Alternatively to the sequential/asynchronous updating procedure described above, all neurons can be updated at the same time step. This is called a parallel/synchronous dynamics.

The Hopfield model is an auto-associative memory. This means information can be stored in the net, and even if the input is incomplete or contains corrupted parts, the memory is able to retrieve the correct data. The information we want the net to store is encoded into bit strings of length N, which we denote by $\xi^1, \ldots, \xi^M$. These so-called patterns $\xi^1, \ldots, \xi^M$ are often chosen uniformly at random from the configuration space $S^N = \{-1, 1\}^N$. By generating the information randomly, one hopes for typical or even more difficult data than in a real world application. Alternatively, the patterns can be generated such that different patterns or the bits within a pattern are correlated. This will be of interest in Chapter 3.

The random variables are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. One way to consider a configuration to be stored in the memory is to require the pattern to be a fixed point of the dynamics. In other words, all neurons need to keep their initial value in case they are updated. Additionally, the feature to correct errors and missing data will be induced by a basin of attraction around each pattern. In this case, configurations close to a pattern become successively corrected and end up as one of the patterns.

To achieve this memory functionality, the weight between two neurons is assigned to the correlation of their spins in all patterns. This rule is called Hebb rule [Heb49] and is based on the neurobiological observation that the connection between neurons grows stronger if the cells are often firing simultaneously. For each $i, j \in \{1, \ldots, N\}$ with $i \neq j$ set

$$W_{ij} = \sum_{\mu=1}^{M} \xi_i^\mu \xi_j^\mu \tag{1.1}$$

and $W_{ii} = 0$ for all i. If the neurons i and j have the same sign in a pattern, the weight increases by one. For each pattern where the neurons have opposed signs, the weight decreases. Because of the symmetric setting the thresholds $(\theta_i)_{i \leq N}$ are all chosen to be zero. This leads to the dynamics $T : S^N \to S^N$, whose i-th vector entry T_i maps a configuration σ to the value neuron i is having after its update. The vector entries of $T = (T_i)_{1 \leq i \leq N}$ are defined by

$$T_i(\sigma) = \operatorname{sgn}\left(\sum_{j \neq i} \left(\sum_{\mu=1}^{M} \xi_i^\mu \xi_j^\mu \right) \sigma_j \right), \tag{1.2}$$

where sgn is the signum function. With this definition a neuron has a tendency to align with neurons which often share the same sign in the patterns ($W_{ij} > 0$) and avoid aligning to others ($W_{ij} < 0$).

After proving that the algorithm with asynchronous updating always converges to a stable state, Hopfield stated that the model can store $\alpha_c N$ patterns with $\alpha_c \approx 0.14$ before severe errors occur. His conclusions about the storage capacity were based on Monte-Carlo simulations [Hop82]. Some years later Amit et al. showed in [AGS85a; AGS85b; AGS87] that $\alpha_c \approx 0.138$ with the help of the non-rigorous replica trick. Rigorous results were then obtained by Newman in [New88]. He proved that the model can store αN patterns with $\alpha \leq 0.056$ if you allow a small fraction of errors at the end of the retrieval. This was done by analysing the energy landscape of the model and proving the existence of energy barriers around each pattern. Further improvements to the lower bound of α_c were made by Loukianova ($\alpha \leq 0.071$) in [Lou97] and Talagrand ($\alpha \leq 0.08$) in [Tal98].

A more restrictive concept of storage capacity is considered if you ask the patterns to be fixed points and corrupted patterns to be retrieved perfectly. Results for this perfect retrieval of patterns were obtained by McEliece et al. in [McE+87]. They were able to show that the Hopfield model can store at least $N/c\log(N)$ patterns for an appropriate constant c. We will see more about this in Section 1.1, especially in Theorem 1.1. Results about sharp upper bounds were shown by Bovier in [Bov99]. Komlós and Paturi in [KP88] and Burshtein in [Bur94] extended the results on the radius of attraction and on the number of iterations in a multi-step setting.

1.1. Storage capacity of the standard Hopfield model for a perfect retrieval

We now have a look at the results for a perfect retrieval of patterns. First, let us introduce the Hamming distance as a metric on the configuration space $\{-1,1\}^N$. The distance between $x, y \in \{-1,1\}^N$ is defined by

$$d(x,y) = \sum_{i=1}^{N} \frac{|x_i - y_i|}{2} = \frac{1}{2}\left(N - \sum_{i=1}^{N} x_i y_i\right)$$

and counts the number of vector entries which differ from each other. A sphere around x with radius r is denoted by $\mathcal{S}(x,r)$. In the context of an auto-associative memory, an element of the sphere around a pattern with radius r represents corrupted data with exactly r wrong bits. As long as r is reasonably small, the configuration should evolve from the sphere to the centre driven by the dynamics of the net described in Chapter 1.

In the following theorem a lower bound for the storage capacity is stated in the case, where we require the patterns to be fixed points of the dynamics and having a non-trivial radius of attraction. These statements hold with a probability converging to one while the number of neurons tends to infinity. The results go back to Komlós and Paturi [KP88], McEliece et. al. [McE+87] and Petritis [Pet96]. The theorem is cited from the chapter "On the Storage Capacity of the Hopfield Model" by Matthias Löwe in the book [BP97].

Theorem 1.1
Let $\rho \in [0, \frac{1}{2})$ *and, for each* $\nu = 1, \dots, M(N)$*, let* $\widetilde{\xi}^\nu$ *be an element of* $\mathcal{S}(\xi^\nu, \rho N)$*. Assume that* $M(N) = (1-2\rho)^2 \frac{N}{c \log(N)}$*. Then,*

(1) if $c > 2$*,*

$$\mathbb{P}\left(T(\widetilde{\xi}^\nu) = \xi^\nu\right) = 1 - R_N$$

with $\lim_{N\to\infty} R_N = 0$*.*

(2) if $c \geq 4$*,*

$$\mathbb{P}\Big(\bigcap_{\mu=1}^{M(N)} T(\widetilde{\xi}^\mu) = \xi^\mu\Big) = 1 - R_N$$

with $\lim_{N\to\infty} R_N = 0$*.*

(3) if $c > 6$*,*

$$\mathbb{P}\Big(\liminf_{N\to\infty}\Big(\bigcap_{\mu=1}^{M(N)} T(\widetilde{\xi}^\mu) = \xi^\mu\Big)\Big) = 1$$

that is, the noised patterns are almost surely attracted.

In the proof of Theorem 1.1 we want to work out how the Hopfield model is able to reconstruct the correct patterns and why there are limitations on the amount of patterns which can be stored. Additionally, we set ourselves into the scenario of independent and identical distributed spins, and the proof provides insights which will be important in the upcoming sections. Results of Chapter 2 (see Theorem 2.1) can be seen as a generalization of this theorem and therefore imply statements from above. Since the proof for $\rho > 0$ works analogously, we keep the notations clear and only proof the case $\rho = 0$.

Proof. Suppose we want to store pattern ξ^1 and the net is currently updating neuron i. A close look at the event leads to a necessary condition and therefore an upper bound for the probability of updating neuron i to the wrong value. If the current configuration is

σ, then

$$\begin{aligned}\left\{T_i(\sigma) \neq \xi_i^1\right\} &= \left\{\operatorname{sgn}\left(\sum_{j\neq i}\left(\sum_{\mu=1}^{M}\xi_i^\mu\xi_j^\mu\right)\sigma_j\right) \neq \xi_i^1\right\} \\ &\subseteq \left\{\sum_{\mu=1}^{M}\xi_i^1\xi_i^\mu\sum_{j\neq i}\xi_j^\mu\sigma_j \leq 0\right\}. \end{aligned} \tag{1.3}$$

For each $\mu \in \{1,\dots,M\}$ we call

$$m_i^\mu(\sigma) = \frac{1}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N}\xi_j^\mu\sigma_j \in [-1,1] \tag{1.4}$$

the overlap of pattern ξ^μ with configuration σ (without spin i). Therefore, the net weights each overlap with a sign $(\xi_i^1\xi_i^\mu)$ to account for an alignment with ξ_i^1, and assign neuron i to the value where the sum of overlaps is stronger.

If the initial configuration is ξ^1 (or on a sphere close to ξ^1), then the overlap with this pattern is deterministic and equal (or close) to one. On the other hand, the sum of the remaining terms forms a random distraction. In this case we call $m_i^1(\xi^1)$ the signal term and

$$\sum_{\mu=2}^{M}\xi_i^1\xi_i^\mu(N-1)\, m_i^\mu(\xi^1)$$

the noise term.

From these insights we can conclude that the probability of updating a fixed neuron to the wrong signal is bounded by the probability of a large deviation event of the noise term:

$$\mathbb{P}\left(T_i(\xi^1)\neq\xi_i^1\right) \leq \mathbb{P}\left(-\sum_{\mu=2}^{M}\xi_i^1\xi_i^\mu\sum_{j\neq i}\xi_j^\mu\xi_j^1 \geq N-1\right).$$

A standard way to handle this expression is to use an exponential Chebyshev inequality with parameter $t>0$. Thus,

$$\mathbb{P}\left(T_i(\xi^1)\neq\xi_i^1\right) \leq \exp\left(-t(N-1)\right)\mathbb{E}\left[\exp\left(-t(N-1)\sum_{\mu=2}^{M}\xi_i^1\xi_i^\mu m_i^\mu(\xi^1)\right)\right]. \tag{1.5}$$

The moment generating function of the noise term for the scenario of independent and identically distributed (i.i.d.) patterns is easier to handle with the following observation:

Lemma 1.2
Let $(\xi_j^\mu)_{j\in\{1,\dots,N\}}^{\mu\in\{1,\dots,M\}}$ *be i.i.d. Rademacher-distributed random variables, i.e.*

$$\mathbb{P}\left(\xi_1^1 = 1\right) = \frac{1}{2} = \mathbb{P}\left(\xi_1^1 = -1\right).$$

For $\mu \in \{2,\dots,M\}$ *and* $j \in \{1,\dots,N\}$ *set* $\zeta_j^\mu := \xi_j^1\xi_j^\mu$ *and* $\zeta^\mu = (\zeta_j^\mu)_j$. *Then* $(\zeta^\mu)_{\mu\in\{2,\dots,M\}}$ *is distributed as* $(\xi^\mu)_{\mu\in\{2,\dots,M\}}$. *Hence, for every* $f : \mathbb{R}^{N\times(M-1)} \to \mathbb{R}$

$$\mathbb{E}\left[f(\zeta^2,\dots\zeta^M)\right] = \mathbb{E}\left[f(\xi^2,\dots,\xi^M)\right].$$

Proof of Lemma 1.2. The statement is a simple consequence of the independent and symmetric distributed spins. With these arguments, it is easy to show that for a set $A \subseteq \{-1,1\}^{N\times(M-1)}$

$$\mathbb{P}\left((\zeta^2,\dots,\zeta^M) \in A\right) = \mathbb{P}\left((\xi^2,\dots,\xi^M) \in A\right)$$

is true. The rest of the proof is a simple consequence of the transformation formula. □

If we apply Lemma 1.2 and use the independence of spins in eq. (1.5), it follows that

$$\begin{aligned}\mathbb{P}\left(T_i(\xi^1) \neq \xi_i^1\right) &\leq \exp\left(-t(N-1)\right)\mathbb{E}\left[\exp\left(-t\sum_{\mu=2}^{M}\xi_i^\mu\sum_{j\neq i}\xi_j^\mu\right)\right]\\ &\leq \exp\left(-t(N-1)\right)\cosh(t)^{(N-1)(M-1)}\\ &\leq \exp\left(-t(N-1)\right)\cosh(t)^{NM}.\end{aligned}$$

Here, cosh is the hyperbolic cosine function, which can be characterized as

$$\cosh(x) = \frac{1}{2}\left(e^x + e^{-x}\right).$$

The hyperbolic cosine is bounded from above by $\exp(\frac{x^2}{2})$. Together with $t = \frac{1}{M}$ as well as the identity $M = (1-2\rho)^2\frac{N}{c\log(N)}$, we conclude that

$$\begin{aligned}\mathbb{P}\left(T_i(\xi^1) \neq \xi_i^1\right) &\leq \exp\left(-tN + \frac{t^2}{2}NM + t\right)\\ &= \exp\left(-\frac{1}{2}\frac{N}{M} + t\right) = N^{-\frac{c}{2}}(1+o(1)).\end{aligned}$$

The statements (1) and (2) of Theorem 1.1 follow with a simple union bound argument. An application of the Borel-Cantelli lemma proves Statement (3). □

The proof suggests that if the number of patterns grows too fast in N, then the neural network could be distracted by many small deviations whose probability is not vanishing fast enough. Therefore, M cannot grow arbitrarily fast in N if we want to store and retrieve patterns with a high probability. The techniques we used to prove Theorem 1.1 just allow a statement about the lower bound of the storage capacity. Rigorous results about the upper bound were also done. Bovier showed in [Bov99] that the choice of M in statement (1) is sharp for this definition of storage capacity.

In the proof of Theorem 1.1 we worked with the overlap of configurations. Because of the dynamics, a self-coupling is not possible. Therefore, the spin of the updated neuron is excluded. In general an overlap is defined as follows:

Definition 1.3
For each $\mu \in \{1, \ldots, M\}$ we call

$$m^\mu(\sigma) = \frac{1}{N} \sum_{j=1}^{N} \xi_j^\mu \sigma_j \in [-1, 1] \tag{1.6}$$

the overlap of pattern ξ^μ with configuration σ. Furthermore, we denote by $m_i^\mu(\sigma)$ the overlap without index i, i.e.

$$m_i^\mu(\sigma) = \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \xi_j^\mu \sigma_j \in [-1, 1]. \tag{1.7}$$

These overlaps play a crucial role when the Hopfield model is considered as a spin glass model.

1.2. The standard Hopfield model as a spin glass model

Another approach to study the Hopfield model is to define it as a spin glass model (see [PF77; Tal98], Chapter 5 in [Tal03] or Part 1 in [BP97]). These models are interacting particle systems, which are originally used in physics to describe magnetic spins and their behaviour in different temperature regimes. Famous examples of spin glasses are the Ising model and the Curie-Weiss model (see Chapter IV in [Ell85]). The latter one will be relevant in Chapter 3 to analyse whether and how correlations affect the storage capacity of the Hopfield model. Furthermore, ideas and results about spin glasses will be helpful in Chapter 2 to handle a more general dynamics.

A spin system is defined on a lattice $\Lambda \subseteq \mathbb{Z}^d$ where each site has a spin from the local spin space $\mathcal{S}$. The behaviour of the spin glass is characterized through a Hamiltonian function $H_\Lambda : S^\Lambda \to \mathbb{R}$, which assigns an energy value to each configuration $\sigma \in S^\Lambda$. Then the Hamiltonian function is used to define the finite volume Gibbs measure $\mu_{\beta,\Lambda}$ on $(\mathcal{S}^\Lambda, \mathcal{B}(\mathcal{S}^\Lambda))$ by

$$\mu_{\beta,\Lambda}(\sigma) = \frac{2^{-N}}{Z_{\beta,\Lambda}} \exp\left(-\beta H_\Lambda(\sigma)\right). \tag{1.8}$$

The parameter $\beta \in [0,\infty)$ represents the inverse temperature of the system and the normalizing constant

$$Z_{\beta,\Lambda} = 2^{-N} \sum_{\widetilde{\sigma} \in \{-1,1\}^\Lambda} \exp\left(-\beta H_\Lambda(\widetilde{\sigma})\right)$$

is called partition function. From eq. (1.8) we see that the system favours configurations with low energy values.

In models of statistical physics one often uses order parameters (e.g. the mean magnetization) to analyse the behaviour of the system for $|\Lambda| \to \infty$. These order parameters describe fundamental properties of the system from a macroscopic point of view. For the Hopfield model $\xi^1, \dots, \xi^M$ are chosen uniformly at random from $\{-1,1\}^N$ and are then considered to be fixed for the spin glass. Because of this, the order parameters and the Hamiltonian are random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A natural choice of order parameters for $\omega \in \Omega$ is the vector of overlaps $m[\omega](\sigma)$, which is defined by

$$m[\omega](\sigma) = (m^1[\omega](\sigma), \dots, m^M[\omega](\sigma))$$

(see Definition 1.3) for every $\sigma \in \mathcal{S}^\Lambda$. As Hamiltonian function we define $H_{N,M}[\omega](\sigma)$ to be

$$\begin{aligned} H_{N,M}[\omega](\sigma) &= -\frac{N}{2} \sum_{\mu=1}^{M} m^\mu[\omega](\sigma)^2 - hNm^1[\omega](\sigma) \\ &= -\frac{N}{2} \left\| m[\omega](\sigma) \right\|_2^2 - hNm^1[\omega](\sigma) \\ &= -\frac{1}{2N} \sum_{i,j=1}^{N} \left(\sum_{\mu=1}^{N} \xi_i^\mu[\omega] \xi_j^\mu[\omega] \right) \sigma_i \sigma_j - h \sum_{j=1}^{N} \xi_j^\mu[\omega] \sigma_j \end{aligned} \tag{1.9}$$

for $h \geq 0$ (see eq. (1.2) in [BG97] or eq. (1.2) in [Tal98]). With $\|x\|_2$ we denote the ℓ_2-norm of x.

The spins are interacting with the whole net through the macroscopic parameter rather than with local spins. This is the reason why the lattice structure can be ignored. We are dropping Λ and use $|\Lambda| := N$ instead. If $h > 0$, the last part represents a local field added to the spin glass. This is used to give pattern ξ^1 a special role. Low energy values can be reached if the configuration aligns with the previously generated patterns and especially if pattern ξ^1 is included.

The connection between the Hopfield model defined by the dynamics in eq. (1.2) and the spin glass with the Hamiltonian given in eq. (1.9) can be seen if we consider two configurations σ and $\widetilde{\sigma}$. Let these configurations be equal for all but for one spin, which we denote by σ_k. We think of $\widetilde{\sigma}$ as the starting configuration and σ_k as the spin after updating neuron k. The difference of their energy values (assuming $h = 0$) is given by

$$H_{N,M}[\omega](\sigma) - H_{N,M}[\omega](\widetilde{\sigma}) = -\frac{1}{2N} \sum_{i,j=1}^{N} W_{ij} \left(\sigma_i \sigma_j - \widetilde{\sigma}_i \widetilde{\sigma}_j\right) = -\frac{2}{N}\, \sigma_k \cdot \left(\sum_{j=1}^{N} W_{kj} \widetilde{\sigma}_j \right).$$

Here, we used that the weights (see eq. (1.1)) are chosen to be symmetric ($W_{ij} = W_{ji}$) and without self-coupling ($W_{ii} = 0$) as well as the identity

$$\sigma_i \sigma_j - \widetilde{\sigma}_i \widetilde{\sigma}_j = \begin{cases} 0, & \text{if } i \neq k \text{ and } j \neq k \\ 0, & \text{if } i = j = k \\ 2\sigma_i \sigma_j, & \text{else} \end{cases} .$$

Furthermore, the dynamics T ensures that the new value of neuron k is equal to

$$\sigma_k = T_k(\widetilde{\sigma}) = \operatorname{sgn}\left(\sum_{i=1}^{N} W_{kj} \widetilde{\sigma}_j \right).$$

Therefore, an update according to T never increases the energy value of the configuration. If the configurations σ and $\widetilde{\sigma}$ have the same energy value, it follows that the postsynaptic potential needs to be zero. Since the Hamiltonian is bounded from below and we set $\operatorname{sgn}(0) = 1$, the algorithm will always converge to a local minimum of the energy landscape.

2. Generalized Hopfield model with i.i.d. patterns

We already learned that each pattern influences the dynamics of the net through its overlap with the current configuration. If we want the net to retrieve a specific pattern, the corresponding overlap pushes the net to the right behaviour. But with an increasing number of patterns, the net gets distracted by other patterns if the overall contribution has the same order as the signal term. To reach a larger storage capacity, the order of the overlaps needs to grow faster while two configurations get close to each other. If we can ensure this, the order of the signal and noise term should be distinguishable even for a larger number of patterns. In context of spin glasses, this means the energy value should decrease faster while the network approaches the correct pattern. Based on this observation, Krotov and Hopfield suggested a modification of the dynamics in their work [KH16]:

Define $\widehat{T} = (\widehat{T}_i)_{1\leq i\leq N}$ by

$$\widehat{T}_i(\sigma) = \operatorname{sgn}\left[\sum_{\mu=1}^{M}\Big(F\big(1\cdot\xi_i^\mu + \sum_{j\neq i}\xi_j^\mu\sigma_j\big) - F\big((-1)\cdot\xi_i^\mu + \sum_{j\neq i}\xi_j^\mu\sigma_j\big)\Big)\right], \qquad (2.1)$$

where $F:\mathbb{R}\to\mathbb{R}$ is some smooth function.

The suggested dynamics in eq. (2.1) can be seen as a more general version of the dynamics in the standard Hopfield model. If we choose $F(x) = x^2$, the original dynamics can be regained because neuron i is then updated to the sign of

$$\begin{aligned}
&\sum_{\mu=1}^{M}\left(F\big(1\cdot\xi_i^\mu + \sum_{j\neq i}\xi_j^\mu\sigma_j\big) - F\big((-1)\cdot\xi_i^\mu + \sum_{j\neq i}\xi_j^\mu\sigma_j\big)\right)\\
=\;&\sum_{\mu=1}^{M}\left(1 + 2\sum_{j\neq i}\xi_i^\mu\xi_j^\mu\sigma_j + \Big(\sum_{j\neq i}\xi_j^\mu\sigma_j\Big)^2 - 1 + 2\sum_{j\neq i}\xi_i^\mu\xi_j^\mu\sigma_j - \Big(\sum_{j\neq i}\xi_j^\mu\sigma_j\Big)^2\right)\\
=\;&4\sum_{\mu=1}^{M}\sum_{j\neq i}\xi_i^\mu\xi_j^\mu\sigma_j.
\end{aligned}$$

This again emphasizes the connection between the dynamics of the Hopfield model and the energy landscape of the spin glass model. We can interpret $-\sum_{\mu} F(m^{\mu}(\sigma))$ as an energy value of a configuration σ. Then eq. (2.1) describes exactly the behaviour to switch neuron i to the configuration with a lower energy level. The advantage of $\widehat{T}$ is that we can determine, with the help of F, how fast the energy decreases when the net approaches a specific pattern. A natural choice to enforce a faster decreasing energy function would be $F(x) = x^n$ for $n \geq 2$. By using this function, overlaps of higher orders are used to measure how close a configuration is to one of the patterns:

$$m^{\mu}(\sigma)^{n-1} = \left(\frac{1}{N} \sum_{j=1}^{N} \xi_j^{\mu} \sigma_j \right)^{n-1} = \frac{1}{N^{n-1}} \sum_{j_1, \dots, j_{n-1}} \xi_{j_1}^{\mu} \dots \xi_{j_{n-1}}^{\mu} \sigma_{j_1} \dots \sigma_{j_{n-1}}. \tag{2.2}$$

For larger n, the energy decreases faster if a configuration comes close to the pattern. Therefore, the storage capacity for this kind of dynamics is expected to be higher.

Similar to this approach, in [New88] Newman used overlaps of higher orders but with a different definition of storage capacity. He demanded that there exists an energy barrier around each pattern, which guarantees that a local minimum is in a close neighbourhood. His main theorem states that the probability of

$$H(y) > H(\xi^{\mu}) + \varepsilon N^l \text{ for every } y \text{ in } \mathcal{S}(\xi^{\mu}, \delta)$$

converges to one, where $H(y) = -\sum_{\mu=1}^{M} m^{\mu}(\sigma)^l$. This implies that the net is allowed to end up in a configuration which still has a small amount of errors. In this scenario, he showed that the storage capacity increases to $M \approx \alpha N^{l-1}$ instead of $M \approx \alpha N$. Burshtein proved in [Bur98] that the storage capacity for a perfect retrieval increases if the network is asked to reach the correct pattern after multiple iterations of the synchronous dynamics. His results coincide with our lower bound for the storage capacity. Other results which use higher order overlaps can be found in relation to spin glass models (see [Tal00b; Tal00a; BN01; BKL02]). In the present thesis, we want to show that the storage capacity can be increased if we consider higher order overlaps with a perfect retrieval of patterns. In contrast to [Bur98], in our results the net evolves in one synchronous step to the desired pattern.

The choice of $F(x) = x^n$ causes another interesting change. A closer look at the postsynaptic potential shows that the contribution of each pattern to this potential can be split into two parts:

$$\sum_{\mu=1}^{M} (1 \cdot \xi_i^{\mu} + (N-1) m_i^{\mu}(\sigma))^n - ((-1) \cdot \xi_i^{\mu} + (N-1) m_i^{\mu}(\sigma))^n$$

$$
\begin{aligned}
&= \sum_{\mu=1}^{M} \sum_{k=0}^{n} \binom{n}{k} \left(1-(-1)^k\right) (\xi_i^\mu)^k \left((N-1) m_i^\mu(\sigma)\right)^{n-k} \\
&= 2 \sum_{\mu=1}^{M} \underbrace{\xi_i^\mu \operatorname{sgn}\left(m_i^\mu(\sigma)\right)^{n-1}}_{\text{sign of the data}} \cdot \underbrace{\sum_{\substack{k=0,\\ k \text{ odd}}}^{n} \binom{n}{k} (N-1)^{n-k} \left|m_i^\mu(\sigma)\right|^{n-k}}_{\text{strength of the data}} .
\end{aligned} \tag{2.3}
$$

One part of the postsynaptic potential is called "strength of the data" because it measures how much information the current configuration provides about each pattern. The absolute value of the overlap shows how close the configuration is either to the pattern ξ^μ itself or to its negative counterpart $-\xi^\mu$. From both possibilities, being close to ξ^μ and being close to $-\xi^\mu$, the net receives valuable indications about the correct value of neuron i. Thus, it is reasonable that patterns with a large strength term have a significant impact on the postsynaptic potential and therefore on the next value of neuron i. Whether the patterns hint to a value of 1 or -1 for neuron i is solely determined by the "sign of the data". The strength of the data is always non-negative.

First, let us assume that $n-1$ is an odd number. This case contains the standard Hopfield model ($n=2$). The sign of the data suggests a value of ξ_i^μ if the configuration is close to ξ^μ and it suggests $-\xi_i^\mu$ if the configuration is close to $-\xi^\mu$. This behaviour is well-known from the standard Hopfield model and leads to the fact that together with a pattern the net always stores its negative counterpart.

On the other hand, assume that $n-1$ is an even number. Since the sign function is either 1 or -1 and the exponent is even, the signal of the data is equal to the spin ξ_i^μ. In this case the neural net interprets a configuration close to the negative counterpart of a pattern, namely $-\xi^\mu$, as a hint to ξ_i^μ. Therefore, the net does not store negative counterparts of the patterns but rather takes a step in the direction of ξ^μ.

As eq. (2.3) shows, the strength term in the postsynaptic potential is rather complex and maybe difficult to handle. Since we are interested in results for N large, the most important contribution is coming from the highest order overlap. In our case, this is $m_i^\mu(\sigma)^{n-1}$ and as long as N is large, the difference should be negligible. That is the reason why we replace the strength term in eq. (2.3) by the highest order overlap $m_i^\mu(\sigma)^{n-1}$. We consider the dynamics $\widetilde{T} = (\widetilde{T}_i)_{i \le N}$ on $\{-1,+1\}^N$, where $\widetilde{T}_i$ is defined by

$$
\widetilde{T}_i(\sigma) := \operatorname{sgn}\left(\sum_{j_1,\dots,j_{n-1}} \sigma_{j_1} \cdot \ldots \cdot \sigma_{j_{n-1}} W_{i,j_1,\dots,j_{n-1}} \right) \tag{2.4}
$$

with

$$W_{i_1,\ldots,i_n} = \frac{1}{N^{n-1}} \sum_{\mu=1}^{M} \xi_{i_1}^{\mu} \cdot \ldots \cdot \xi_{i_n}^{\mu}. \tag{2.5}$$

The summation in eq. (2.4) goes from 1 to N and excludes indices where $j_k = i$ for some k.

To see that the dynamics $\widetilde{T}$ in eq. (2.4) really results from replacing the strength term in eq. (2.3), we write:

$$\begin{aligned}
\widetilde{T}_i(\sigma) &= \operatorname{sgn}\left(\sum_{j_1,\ldots,j_{n-1}} \sigma_{j_1} \cdots \sigma_{j_{n-1}} W_{i,j_1\cdots j_{n-1}} \right) \\
&= \operatorname{sgn}\left(\sum_{\mu=1}^{M} \xi_i^{\mu} \sum_{j_1,\ldots,j_{n-1}} \xi_{j_1}^{\mu} \sigma_{j_1} \cdot \ldots \cdot \xi_{j_{n-1}}^{\mu} \sigma_{j_{n-1}} \right) \\
&= \operatorname{sgn}\left(\sum_{\mu=1}^{M} \xi_i^{\mu} (N-1)^{n-1} m_i^{\mu}(\sigma)^{n-1} \right) \\
&= \operatorname{sgn}\left(\sum_{\mu=1}^{M} \underbrace{\xi_i^{\mu} \cdot \operatorname{sgn}\left(m_i^{\mu}(\sigma)\right)^{n-1}}_{\text{sign of the data}} \cdot \underbrace{(N-1)^{n-1} \cdot \left|m_i^{\mu}(\sigma)\right|^{n-1}}_{\text{strength of the data}} \right).
\end{aligned} \tag{2.6}$$

From now on, as long as we are interested in results about the limiting case $N \to \infty$, we will avoid to distinguish between N and $N-1$ or M and $M-1$. These differences are negligible in the limit.

2.1. Generalized dynamics with a polynomial interaction function

In this section the data is given by M bit strings of length N which are chosen uniformly at random from the configuration space $\{-1,1\}^N$. As before the patterns $\xi^1, \ldots, \xi^M$ are defined as $\xi^{\mu} = (\xi_j^{\mu})_{1 \le j \le N}$ where $(\xi_j^{\mu})_{j=1,\ldots,N}^{\mu=1,\ldots,M}$ is a family of i.i.d. random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with

$$\mathbb{P}(\xi_j^{\mu} = 1) = \mathbb{P}(\xi_j^{\mu} = -1) = \frac{1}{2}$$

for all $\mu = 1, \ldots, M$ and all $j = 1, \ldots, N$. The network updates neurons according to the dynamics $(\widetilde{T}_i)_{i \le N}$ (see eq. (2.4)). For a perfect retrieval, we consider a pattern to be stored if the pattern itself is a fixed point of the dynamics. Additionally, we expect a positive basin of attraction. That means a corrupted input string evolves through the net to the desired pattern after each neuron has been updated at least once.

The following theorem shows that our dynamics with higher order overlaps is capable of storing more patterns than the standard Hopfield model. Moreover, there is still a positive basin of attraction. The following result was published in [Dem+17].

Theorem 2.1
Let $M = \frac{N^{n-1}}{c_n \log(N)}$ and let $\xi^1, \ldots, \xi^M$ be M independent patterns chosen uniformly at random from $\{-1, +1\}^N$. The Hopfield model with dynamics $\widetilde{T}$ can store at least M patterns for $c_n > 2(2n-3)!!$ if one wants a fixed pattern to be a fixed point of the dynamics with a probability converging to one.

Moreover, fix $\rho \in [0, \frac{1}{2})$. If $c_n > \frac{2(2n-3)!!}{(1-2\rho)^{2(n-1)}}$, then for any $\widetilde{\xi}^\nu$ taken uniformly at random from $\mathcal{S}(\xi^\nu, \rho N)$, where ρN is assumed to be an integer, it follows that

$$\mathbb{P}\left(\widetilde{T}(\widetilde{\xi}^\nu) = \xi^\nu\right) = 1 - R_N$$

with $\lim_{N\to\infty} R_N = 0$.

Furthermore, if $c_n > \frac{2n(2n-3)!!}{(1-2\rho)^{2(n-1)}}$ then

$$\mathbb{P}\left(\forall \mu \le M : \ \widetilde{T}(\widetilde{\xi}^\mu) = \xi^\mu\right) = 1 - R_N$$

with $\lim_{N\to\infty} R_N = 0$.

Proof. For ease of notation we assume that the network consists of $N+1$ (instead of N) neurons. For fixed $\rho \in [0, \frac{1}{2})$ we choose input data $\widetilde{\xi}^1 \in \mathcal{S}(\xi^1, \rho(N+1))$ and show that pattern ξ^1 can be stored ($\rho = 0$) resp. corrected ($\rho > 0$) with high probability if N tends to infinity. We start with the dynamics $\widetilde{T}$ as stated in eq. (2.6). By multiplying with ξ_i^1, we conclude that a necessary condition for a false update of neuron i, namely $\widetilde{T}_i(\widetilde{\xi}^1) \neq \xi_i^1$, is given by

$$\sum_{\mu=1}^{M} \xi_i^1 \xi_i^\mu N^{n-1} \left(m_i^\mu(\widetilde{\xi}^1)\right)^{n-1} \le 0. \tag{2.7}$$

We split the sum into signal ($\mu = 1$) and noise ($\mu \neq 1$) term. The signal slightly differs whether the bit of neuron i itself is corrupt or not:

$$N^{n-1} m_i^1(\widetilde{\xi}^1)^{n-1} = \Big(\sum_{j \neq i} \xi_j^1 \widetilde{\xi}_j^1 \Big)^{n-1} = \Big((N+1)(1-2\rho) - \xi_i^1 \widetilde{\xi}_i^1 \Big)^{n-1}.$$

A lower bound for the signal is given by

$$N^{n-1} m_i^1(\widetilde{\xi}^1)^{n-1} > (1-2\rho)^{n-1} N^{n-1} u(N),$$

where $u(N) = 1 - 2\rho(1-2\rho)^{-1} N^{-1}$. The term $u(N)$ converges to one for N to infinity and therefore will be negligible in the limit. A necessary condition for the event in eq. (2.7) is

$$-\sum_{\mu=2}^{M} \xi_i^1 \xi_i^\mu N^{n-1} m_i^\mu(\widetilde{\xi}^1)^{n-1} \geq (1-2\rho)^{n-1} N^{n-1} u(N).$$

Hence, the probability for a false update at neuron i can be bounded by

$$\mathbb{P}\left(\widetilde{T}_i(\widetilde{\xi}^1) \neq \xi_i^1 \right) \leq \mathbb{P}\left(-\sum_{\mu=2}^{M} \xi_i^1 \xi_i^\mu \; N^{n-1} m_i^\mu(\widetilde{\xi}^1)^{n-1} \geq (1-2\rho)^{n-1} N^{n-1} u(N) \right).$$

For $t > 0$ we use the exponential Chebyshev inequality and conclude

$$\mathbb{P}\left(\widetilde{T}_i(\widetilde{\xi}^1) \neq \xi_i^1 \right) \leq e^{-t(1-2\rho)^{n-1} N^{n-1} u(N)} \; \mathbb{E}\left[\exp\left(-t \sum_{\mu=2}^{M} \xi_i^1 \xi_i^\mu N^{n-1} m_i^\mu(\widetilde{\xi}^1)^{n-1} \right) \right]. \tag{2.8}$$

The moment generating function in eq. (2.8) can be handled in a similar way as in the proof of the standard Hopfield model (see Theorem 1.1). There we used Lemma 1.2 and with the independence and symmetric distribution of the spins, this lemma is applicable even though we are working with $\widetilde{\xi}^1$. Together with the independence of patterns and after integration with respect to ξ_i^2, the moment generating function equals

$$\begin{aligned} \mathbb{E}\left[\exp\left(-t \sum_{\mu=2}^{M} \xi_i^1 \xi_i^\mu N^{n-1} m_i^\mu(\widetilde{\xi}^1)^{n-1} \right) \right] &= \mathbb{E}\left[\exp\left(-t\, \xi_i^2 N^{n-1} m_i^2(1)^{n-1} \right) \right]^{M-1} \\ &= \mathbb{E}\left[\cosh\left(t N^{\frac{n-1}{2}} \left(\frac{1}{\sqrt{N}} \sum_{j \neq i} \xi_j^2 \right)^{n-1} \right) \right]^{M-1}. \end{aligned}$$

All in all, the probability of updating neuron i to the wrong spin is bounded by

$$\mathbb{P}\left(\widetilde{T}_i(\widetilde{\xi}^1) \neq \xi_i^1\right) \leq e^{-t(1-2\rho)^{n-1}N^{n-1}u(N)}\, \mathbb{E}\left[\cosh\left(tN^{\frac{n-1}{2}}\left(\frac{1}{\sqrt{N}}\sum_{j\neq i}\xi_j^2\right)^{n-1}\right)\right]^M. \quad (2.9)$$

In eq. (2.9) we clearly see the similarity to the standard Hopfield model, but because of $F(x) = x^{n-1}$ the signal term is stronger in this scenario. The noise term is random and has a wider range than before which leads to problems if we would use worst case boundaries. While we have a signal term of order N^{n-1}, the scaled noise term for large N is close to a standard normal distribution. So the noise term has fluctuations of order $N^{\frac{n-1}{2}}$.

Define

$$m := \frac{1}{\sqrt{N}}\sum_{j\neq i}\xi_j^2$$

and write the expectation as a sum over all possible values $x \in \{0, \pm\frac{1}{\sqrt{N}}, \dots, \pm\sqrt{N}\}$:

$$\mathbb{E}\left[\cosh\left(tN^{\frac{n-1}{2}}\left(\frac{1}{\sqrt{N}}\sum_{j\neq i}\xi_j^2\right)^{n-1}\right)\right] = \sum_x \cosh\left(tN^{\frac{n-1}{2}}x^{n-1}\right)\cdot \mathbb{P}(m = x).$$

We split the sum into two parts, where the first consists of large outliers of m beyond $\log(N)^\tau$ for a fixed $\tau > \frac{1}{2}$. Observe that x cannot grow faster than $\sqrt{N}$ and together with $\cosh(z) \leq \exp(|z|)$ we conclude that

$$\begin{aligned} &\sum_{x:\log(N)^\tau < |x| \leq \sqrt{N}} \cosh\left(tN^{\frac{n-1}{2}}x^{n-1}\right)\cdot \mathbb{P}(m = x) \\ \leq\; & 2\cosh\left(tN^{n-1}\right)\mathbb{P}(m > \log(N)^\tau) \\ \leq\; & 2\exp\left(tN^{n-1}\right)\exp\left(-\frac{1}{2}\log(N)^{2\tau}\right). \end{aligned}$$

In the last line we used an exponential bound for the probability of i.i.d. Rademacher spins to exceed a certain value. This is stated in Lemma A.5 with $a = \sqrt{N}\log(N)^\tau$ and is a simple application of the exponential Chebyshev inequality.

Now set $t = \frac{a_n}{M}$ for $a_n > 0$ and recall that $M = \frac{(N+1)^{n-1}}{c_n \log(N+1)}$ then

$$t = \frac{a_n c_n \log(N+1)}{(N+1)^{n-1}} \leq \frac{a_n c_n \log(N)}{N^{n-1}},$$

and thus,

$$
\begin{aligned}
&2\exp\left(tN^{n-1}\right)\exp\left(-\frac{1}{2}\log(N)^{2\tau}\right) \\
\leq\ & 2\ \exp\left(\left[a_n c_n - \frac{1}{2}\log(N)^{2\tau-1}\right]\log(N)\right) =: 2\exp\left(h_1(N)\right).
\end{aligned} \tag{2.10}
$$

Since $\tau > \frac{1}{2}$, this part of the expectation converges to zero for $N \to \infty$ because the term in brackets can be bounded from above by a negative value if N is large enough.

The second part of the sum consists of the critical values of m. We use the inequality $\cosh(z) \leq \exp(\frac{z^2}{2})$ and write the exponential function in its Taylor expansion:

$$
\begin{aligned}
& \sum_{x:|x|\leq\log(N)^\tau} \cosh\left(tN^{\frac{n-1}{2}}x^{n-1}\right)\mathbb{P}(m=x) \\
\leq & \sum_{x:|x|\leq\log(N)^\tau} e^{\frac{t^2}{2}N^{n-1}x^{2(n-1)}}\mathbb{P}(m=x) \\
= & \sum_{x:|x|\leq\log(N)^\tau} \left(1+\frac{t^2}{2}N^{n-1}x^{2(n-1)}+\sum_{k=2}^{\infty}\frac{1}{2^k}\frac{(t^2N^{n-1}x^{2(n-1)})^k}{k!}\right)\mathbb{P}(m=x).
\end{aligned}
$$

The distribution of m converges to a standard normal distribution and its moments can be bounded by the moments of the latter. For $l \in \mathbb{N}$ let $\kappa_{2l} = (2l-1)!!$ be the $2l$-th moment of a standard normal distribution. For N large enough, we derive the upper bound

$$
\begin{aligned}
& \sum_{x:|x|\leq\log(N)^\tau} \left(1+\frac{t^2}{2}N^{n-1}x^{2(n-1)}+\sum_{k=2}^{\infty}\frac{1}{2^k}\frac{(t^2N^{n-1}x^{2(n-1)})^k}{k!}\right)\mathbb{P}(m=x) \\
\leq\ & 1+\frac{t^2}{2}N^{n-1}\kappa_{2(n-1)} + \sum_{x:|x|\leq\log(N)^\tau}\sum_{k=2}^{\infty}\frac{1}{2^k}\frac{\left(t^2N^{n-1}x^{2(n-1)}\right)^k}{k!}\cdot\mathbb{P}(m=x).
\end{aligned} \tag{2.11}
$$

In the higher order terms of the Taylor expansion, we bound x by its highest possible value $\log(N)^\tau$ and the probability by one. Hence, the series has an upper bound which is given by

$$
\begin{aligned}
& \sum_{x:|x|\leq\log(N)^\tau}\sum_{k=2}^{\infty}\frac{1}{2^k}\frac{\left(t^2N^{n-1}x^{2(n-1)}\right)^k}{k!}\cdot\mathbb{P}(m=x) \\
\leq\ & \sum_{k=2}^{\infty}\frac{1}{2^k}\frac{\left(t^2N^{n-1}\log(N)^{2\tau(n-1)}\right)^k}{k!}\cdot\mathbb{P}(|m|\leq\log(N)^\tau)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{k=2}^{\infty} \frac{1}{2^k} \frac{(t^2 N^{n-1} \log(N)^{2\tau(n-1)})^k}{k!} \\
&\leq \frac{t^4}{4} N^{2(n-1)} \log(N)^{4\tau(n-1)} \sum_{k=0}^{\infty} \frac{1}{2^k} \frac{(t^2 N^{n-1} \log(N)^{2\tau(n-1)})^k}{k!}. \qquad (2.12)
\end{aligned}
$$

Recall that $t = \frac{a_n}{M}$, where $a_n > 0$ is not yet specified, and $M = \frac{(N+1)^{n-1}}{c_n \log(N+1)}$. With these identities

$$
t \leq \frac{a_n c_n \log(N)}{N^{n-1}}, \qquad (2.13)
$$

and thus, for N large enough,

$$
t^2 N^{n-1} \log(N)^{2\tau(n-1)} \leq \frac{a_n^2 \log(N)^{2\tau(n-1)-2}}{N^{n-1}} < 1.
$$

This guarantees that the series in eq. (2.12) can be bounded by e. Therefore, an upper bound for the series of higher order terms is given by:

$$
\sum_{x:|x|\leq \log(N)^\tau} \sum_{k=2}^{\infty} \frac{1}{2^k} \frac{\left(t^2 N^{n-1} x^{2(n-1)}\right)^k}{k!} \cdot \mathbb{P}(m = x) \leq \frac{t^4}{4} N^{2(n-1)} \log(N)^{4\tau(n-1)} e. \qquad (2.14)
$$

As long as N is large enough, the sum with the critical values of m has the following upper bound (see eqs. (2.11) and (2.14)):

$$
\begin{aligned}
&\sum_{x:|x|\leq \log(N)^\tau} \cosh\left(t N^{\frac{n-1}{2}} x^{n-1}\right) \mathbb{P}(m = x) \\
&\leq 1 + \frac{t^2}{2} N^{n-1} \kappa_{2(n-1)} + \frac{t^4}{4} N^{2n-2} \log(N)^{4\tau(n-1)} e \\
&\leq \exp\left(\frac{t^2}{2} N^{n-1} \kappa_{2(n-1)} + t^4 N^{2(n-1)} \log(N)^{4\tau(n-1)}\right) =: \exp\left(h_2(N)\right). \qquad (2.15)
\end{aligned}
$$

In the last line we applied the inequality $1 + x \leq e^x$ and used that $\frac{e}{4} \leq 1$.

The moment generating function was split into two parts. The first part consists of large outliers of m. With the help of the exponential Chebyshev inequality, we derived the upper bound $2\exp(h_1(N))$ (see eq. (2.10)), which converges to zero. The second part contains the critical values of m and is bounded by $\exp(h_2(N))$ (see eq. (2.15)), which converges to one if N tends to infinity. Now we claim that the moment generating function to the power of M can be bounded by

$$\mathbb{E}\left[\cosh\left(tN^{\frac{n-1}{2}}\left(\frac{1}{\sqrt{N}}\sum_{j\neq i}\xi_j^2\right)^{n-1}\right)\right]^M \leq \exp\left(Mh_2(N)\right)(1+o(1)). \tag{2.16}$$

To prove this, we need to make sure that

$$\Big(1+2\exp\big(h_1(N)-h_2(N)\big)\Big)^M = (1+o(1))$$

or, equivalently, that its logarithm converges to zero. By using $1+x \leq \exp(x)$, the identity for t (see eq. (2.13)) and $\exp(-h_2(N)) \leq 1$, we conclude that

$$\begin{aligned} M\log\left(1+2\exp\left(h_1(N)-h_2(N)\right)\right) &\leq 2\ M\ \exp\left(h_1(N)-h_2(N)\right) \\ &\leq 2\ M\ \exp\left(h_1(N)\right) \\ &= 2\frac{(N+1)^{n-1}}{c_n\log(N+1)}\exp\left(\left[a_nc_n-\frac{1}{2}\log(N)^{2\tau-1}\right]\log(N)\right) \\ &= 2\frac{(N+1)^{n-1}}{c_n\log(N+1)}N^{a_nc_n-\frac{1}{2}\log(N)^{2\tau-1}} = o(1). \end{aligned}$$

The convergence in the last line follows because $\tau > \frac{1}{2}$ and as a consequence

$$a_nc_n-\frac{1}{2}\log(N)^{2\tau-1} \leq -(n-1)$$

if N is large enough.

We apply the statement of eq. (2.16) to eq. (2.9) and derive

$$\mathbb{P}(\widetilde{T}_i(\widetilde{\xi}^1) \neq \xi_i^1) \leq \exp\big(-t(1-2\rho)^{n-1}N^{n-1}u(N)\big)\exp\left(M\,h_2(N)\right)(1+o(1)) \tag{2.17}$$

with $h_2(N) = \frac{t^2}{2}N^{n-1}\kappa_{2(n-1)} + t^4N^{2(n-1)}\log(N)^{4\tau(n-1)}$ and $M = \frac{(N+1)^{n-1}}{c_n\log(N+1)}$. We already mentioned that $u(N)$ goes to one for N tending to infinity. For every $0<\varepsilon<1$ we can choose N large enough such that

$$u(N)\frac{\log(N+1)}{\log(N)}\Big(\frac{N}{N+1}\Big)^{n-1} \geq (1-\varepsilon). \tag{2.18}$$

The parameter introduced by the Chebyshev inequality is set to

$$t = \frac{a_n}{M} \leq \frac{a_nc_n\log(N)}{N^{n-1}}$$

for $a_n > 0$. Hence, the term in eq. (2.17) is bounded from above by

$$\begin{aligned} & \exp\left(-t(1-2\rho)^{n-1}N^{n-1}(1-\varepsilon) + \frac{t^2}{2}N^{n-1}\kappa_{2(n-1)}M + t^4 N^{2(n-1)}\log(N)^{4\tau(n-1)}M\right) \\ = \; & \exp\left(-a_n c_n \log(N)(1-2\rho)^{n-1}(1-\varepsilon) + \frac{1}{2}a_n^2 c_n \log(N)\kappa_{2(n-1)} + o(1)\right) \\ = \; & \exp\left(\left[-a_n c_n\left((1-2\rho)^{n-1}(1-\varepsilon) - \frac{a_n\,\kappa_{2(n-1)}}{2}\right)\right]\log(N)\right)(1+o(1)). \end{aligned}$$

The fourth order term is negligible because

$$t^4 N^{2(n-1)}\log(N)^{4\tau(n-1)}M \leq \frac{a_n^4 c_n^3 \log(N)^{4\tau(n-1)+3}}{N^{n-1}} \to 0$$

if N goes to infinity. Now we choose $a_n = \frac{(1-2\rho)^{n-1}(1-\varepsilon)}{\kappa_{2(n-1)}}$ then

$$-a_n c_n\left((1-2\rho)^{n-1}(1-\varepsilon) - \frac{a_n\,\kappa_{2(n-1)}}{2}\right) = -c_n\frac{(1-2\rho)^{2(n-1)}(1-\varepsilon)^2}{2\kappa_{2(n-1)}}.$$

Thus, for all $0 < \varepsilon < 1$

$$\begin{aligned} \mathbb{P}\left(\exists i \leq N:\; \widetilde{T}_i(\widetilde{\xi}^1) \neq \xi_i^1\right) &\leq N\exp\left(\left[-c_n\frac{(1-2\rho)^{2(n-1)}(1-\varepsilon)^2}{2\kappa_{2(n-1)}}\right]\log(N)\right)(1+o(1)) \\ &= \exp\left(\left[1-c_n\frac{(1-2\rho)^{2(n-1)}(1-\varepsilon)^2}{2\kappa_{2(n-1)}}\right]\log(N)\right)(1+o(1)), \end{aligned}$$

which converges to zero for N to infinity as long as

$$c_n > \frac{2(2n-3)!!}{(1-2\rho)^{2(n-1)}(1-\varepsilon)^2}.$$

For every $c_n > \frac{2(2n-3)!!}{(1-2\rho)^{2(n-1)}}$ there exists a $\varepsilon > 0$ such that

$$c_n > \frac{2(2n-3)!!}{(1-2\rho)^{2(n-1)}(1-\varepsilon)^2} > \frac{2(2n-3)!!}{(1-2\rho)^{2(n-1)}} \tag{2.19}$$

and this proves the first statement.

The second statement of Theorem 2.1 involves all patterns. A union bound argument and the previous bound for the probability lead to

$$\mathbb{P}\left(\exists \mu \leq M:\; \exists i \leq N:\; \widetilde{T}_i(\widetilde{\xi}^1) \neq \xi_i^1\right)$$

$$\leq \; NM \exp\left(\left[-c_n \frac{(1-2\rho)^{2(n-1)}(1-\varepsilon)^2}{2\kappa_{2(n-1)}}\right] \log(N)\right)(1+o(1)) \tag{2.20}$$
$$= \; \exp\left(n\left[1-c_n \frac{(1-2\rho)^{2(n-1)}(1-\varepsilon)^2}{2n\kappa_{2(n-1)}}\right] \log(N) - \log(c_n \log(N))\right)(1+o(1)),$$

which converges to zero for N to infinity as long as

$$c_n > \frac{2n(2n-3)!!}{(1-2\rho)^{2(n-1)}(1-\varepsilon)^2}.$$

Equation (2.20) is true for all $0 < \varepsilon < 1$. This proves the second statement because ε can again be chosen appropriately (similar to eq. (2.19)). □

2.2. Generalized dynamics with an exponential interaction function

In the previous section we saw that a polynomial function F in the dynamics

$$\widehat{T}_i(\sigma) = \operatorname{sgn}\left[\sum_{\mu=1}^{M}\left(F(1\cdot\xi_i^\mu + \sum_{i\neq j}\xi_j^\mu\sigma_j) - F((-1)\cdot\xi_i^\mu + \sum_{j\neq i}\xi_j^\mu\sigma)\right)\right] \tag{2.21}$$

can increase the storage capacity of the model. If we choose $F(x) = x^2$, the dynamics is equivalent to the dynamics of the standard Hopfield model. A result for the more general approach $F(x) = x^n$ was shown in Theorem 2.1. As a next step, we consider the function $F(x) = \exp(x)$, which can be seen as a "limit of the polynomial functions" where n tends to infinity. The following result is cited from [Dem+17].

Theorem 2.2
Consider the generalized Hopfield model with the dynamics described in eq. (2.21) *and interaction function F given by $F(x) = e^x$. For a fixed $0 < \alpha < \frac{\log(2)}{2}$ let $M = \exp(\alpha N)+1$ and let $\xi^1, \dots, \xi^M$ be M patterns chosen uniformly at random from $\{-1,+1\}^N$. Moreover, fix $\rho \in [0, \frac{1}{2})$. For any μ and any $\widetilde{\xi}^\mu$ taken uniformly at random from $\mathcal{S}(\xi^\mu, \rho N)$, where ρN is assumed to be an integer, it holds that*

$$\mathbb{P}\left(\exists\mu\;\exists i : \widehat{T}_i\left(\widetilde{\xi}^\mu\right) \neq \xi_i^\mu\right) \to 0,$$

if α is chosen in dependence of ρ such that

$$\alpha < \frac{1}{2}\min\left\{I(1-2\rho),\; I\big(1-2\rho-\alpha+I(1-2\rho-\alpha)\big)\right\} \tag{2.22}$$

with

$$I(x) = \begin{cases} \frac{1}{2}\Big((1+x)\log(1+x) + (1-x)\log(1-x)\Big), & \textit{if } x \in [-1,1] \\ \infty, & \textit{else} \end{cases}.$$

Note that Theorem 2.2 in particular implies that

$$\mathbb{P}\left(\exists \mu\ \exists i : \widehat{T}_i\left(\xi^\mu\right) \neq \xi_i^\mu\right) \to 0$$

as $N \to \infty$, i.e. with a probability converging to 1, all the patterns are fixed points of the dynamics.

Proof. Starting in one of the patterns (without loss of generality in ξ^1), we want to prove that it is an attractive fixed point of the update dynamics, i.e. we need to show that $\widehat{T}(\xi^1) = \xi^1$ with a probability converging to one. Additionally, we want the model to correct ρN random errors by updating each of the neurons once.

Recall that we can interpret $-\sum_\mu F(m^\mu(\sigma))$ as the energy value of a configuration σ. Equation (2.21) shows that the neural net switches a neuron to the spin which results in a lower energy value. We denote by

$$\Delta_i E(\sigma) := \sum_{\mu=1}^{M} \left(F\left(\sigma_i \xi_i^\mu + \sum_{j \neq i} \xi_j^\mu \sigma_j\right) - F\left(-\sigma_i \xi_i^\mu + \sum_{j \neq i} \xi_j^\mu \sigma_j\right)\right) \tag{2.23}$$

the energy difference between the configuration σ and its counterpart where the sign of neuron i is switched. For now let us exclude the event where $\Delta_i E(\sigma) = 0$ because in the limit $N \to \infty$ this will be negligible. An equivalent formulation to the dynamics in eq. (2.21) is to say that neuron i remains unchanged after an application of the update rule as long as the difference is positive, i.e. $\Delta_i E(\sigma) > 0$. The spin of neuron i will be changed if $\Delta_i E(\sigma)$ is negative.

Since the input is an element of the sphere around ξ^1, the overlap with the desired pattern is given by $m^1(\widetilde{\xi}^1) = (1-2\rho)$. To calculate the energy values of the two configurations, one for each possible value of neuron i, it is important to know the initial value of neuron i. The indices with false signals are chosen uniformly at random. Thus, neuron i could possibly start with a corrupted signal. In the different cases we get

$$-\widetilde{\xi}_i^1 \xi_i^1 + \sum_{j \neq i}^{N} \widetilde{\xi}_j^1 \xi_j^1 = N m^1(\widetilde{\xi}^1) - 2\widetilde{\xi}_i^1 \xi_i^1 = \begin{cases} N(1-2\rho) - 2 & \text{if neuron } i \text{ is correct} \\ N(1-2\rho) + 2 & \text{if neuron } i \text{ is false} \end{cases}. \tag{2.24}$$

The summand in eq. (2.23) where $\mu = 1$ can be interpreted as signal term and together with eq. (2.24) we observe that

$$\begin{aligned} & F\left(Nm^1(\widetilde{\xi}^1)\right) - F\left(Nm^1(\widetilde{\xi}^1) - 2\widetilde{\xi}_i^1\xi_i^1\right) \\ = & \begin{cases} F(N(1-2\rho)) - F(N(1-2\rho)-2) & \text{if neuron } i \text{ is correct} \\ F(N(1-2\rho)) - F(N(1-2\rho)+2) & \text{if neuron } i \text{ is false} \end{cases} . \end{aligned}$$

Therefore, the signal term pushes the net to the right behaviour because

$$\begin{aligned} & F\big(N(1-2\rho)\big) - F\big(N(1-2\rho)-2\big) \geq 0 \text{ on the one hand, and} \\ & F\big(N(1-2\rho)\big) - F\big(N(1-2\rho)+2\big) \leq 0 \text{ on the other hand} \end{aligned}$$

depending on whether neuron i is correct or incorrect. Now we are able to derive a necessary condition for the event that the noise term becomes big enough to distract the neural net. In the case of $\widetilde{\xi}_i^1 = \xi_i^1$ (correct signal), we see that $T_i(\widetilde{\xi}^1) \neq \xi_i^1$ is equivalent to

$$\Delta_i E(\widetilde{\xi}^1) = \sum_{\mu=1}^{M} \Big(F\big(Nm^\mu(\widetilde{\xi}^1)\big) - F\big(Nm^\mu(\widetilde{\xi}^1) - 2\widetilde{\xi}_i^1\xi_i^\mu\big)\Big) < 0$$

which is the same as

$$\begin{aligned} & \sum_{\mu=2}^{M} \Big(F\big(-\xi_i^1\xi_i^\mu + \sum_{j\neq i} \xi_j^\mu\widetilde{\xi}_j^1\big) - F\big(\xi_i^1\xi_i^\mu + \sum_{j\neq i} \xi_j^\mu\widetilde{\xi}_j^1\big)\Big) \\ > \;& F\big(N(1-2\rho)\big) - F\big(N(1-2\rho)-2\big). \end{aligned} \tag{2.25}$$

In the case of $\widetilde{\xi}_i^1 = -\xi_i^1$ (corrupted signal), the event of $T_i(\widetilde{\xi}^1) \neq \xi_i^1$ is true if and only if

$$\Delta_i E(\widetilde{\xi}^1) = \sum_{\mu=1}^{M} \Big(F(Nm^\mu(\widetilde{\xi}^1)) - F(Nm^\mu(\widetilde{\xi}^1) - 2\widetilde{\xi}_i^1\xi_i^\mu)\Big) > 0$$

which is equivalent to

$$\begin{aligned} & \sum_{\mu=2}^{M} \Big(F\big(-\xi_i^1\xi_i^\mu + \sum_{j\neq i} \xi_j^\mu\widetilde{\xi}_j^1\big) - F\big(\xi_i^1\xi_i^\mu + \sum_{j\neq i} \xi_j^\mu\widetilde{\xi}_j^1\big)\Big) \\ > \;& F\big(N(1-2\rho)+2\big) - F\big(N(1-2\rho)\big). \end{aligned} \tag{2.26}$$

For both of these conditions (see eqs. (2.25) and (2.26)) a necessary event is given by

$$\sum_{\mu=2}^{M}\Big(F\big(-\xi_i^1\xi_i^\mu+\sum_{j\neq i}\xi_j^\mu\widetilde{\xi}_j^1\big)-F\big(\xi_i^1\xi_i^\mu+\sum_{j\neq i}\xi_j^\mu\widetilde{\xi}_j^1\big)\Big)\geq F(N(1-2\rho))-F(N(1-2\rho)-2)$$

which equals

$$\sum_{\mu=2}^{M}\left(\exp\left(-\xi_i^1\xi_i^\mu+\sum_{j\neq i}\xi_j^\mu\widetilde{\xi}_j^1\right)-\exp\left(\xi_i^1\xi_i^\mu+\sum_{j\neq i}\xi_j^\mu\widetilde{\xi}_j^1\right)\right)\geq e^{N(1-2\rho)}\left[1-e^{-2}\right] \tag{2.27}$$

for our choice of $F(x)=\exp(x)$. The event in eq. (2.27) also covers the possibility that $\Delta_i E(\sigma)=0$. Thus, it follows that

$$\begin{aligned}&\mathbb{P}\big(\widehat{T}_i(\widetilde{\xi}^1)\neq\xi_i^1\big)\\ \leq\;&\mathbb{P}\left(\sum_{\mu=2}^{M}\left(\exp\left(-\xi_i^1\xi_i^\mu+\sum_{j\neq i}\xi_j^\mu\widetilde{\xi}_j^1\right)-\exp\left(\xi_i^1\xi_i^\mu+\sum_{j\neq i}\xi_j^\mu\widetilde{\xi}_j^1\right)\right)\geq c\,e^{N(1-2\rho)}\right)\end{aligned} \tag{2.28}$$

with $c=(1-e^{-2})$. With the observation that

$$\begin{aligned}&\exp\left(-\xi_i^1\xi_i^\mu+\sum_{j\neq i}\xi_j^\mu\widetilde{\xi}_j^1\right)-\exp\left(\xi_i^1\xi_i^\mu+\sum_{j\neq i}\xi_j^\mu\widetilde{\xi}_j^1\right)\\ =\;&\exp\left(Nm^\mu(\widetilde{\xi}^1)\right)\left[\exp\left(-\xi_i^\mu\xi_i^1-\xi_i^\mu\widetilde{\xi}_i^1\right)-\exp\left(\xi_i^\mu\xi_i^1-\xi_i^\mu\widetilde{\xi}_i^1\right)\right]\leq e^2c\,\exp\left(Nm^\mu(\widetilde{\xi}^1)\right)\end{aligned}$$

the probability in eq. (2.28) can be bounded by

$$\mathbb{P}\big(\widehat{T}_i(\widetilde{\xi}^1)\neq\xi_i^1\big)\leq\mathbb{P}\left(\sum_{\mu=2}^{M}\exp\left(Nm^\mu(\widetilde{\xi}^1)\right)\geq e^{N(1-2\rho)-2}\right). \tag{2.29}$$

The main contribution to the noise term stems from the overlaps between ξ^μ and $\widetilde{\xi}^1$ for $\mu\neq 1$. The patterns are generated as i.i.d. random variables and Lemma 1.2 proves that the overlap behaves in the same way as i.i.d. Rademacher-distributed random variables. In the rest of the proof we want to use a large deviation principle to show that the impact of a small number of patterns cannot be large enough to eliminate the signal. On the other hand, the choice of α guarantees that M does not grow fast enough so that the noise term reaches the order of the signal by adding up a lot of small deviations.

Define $A=\{\mu\in\{2,\dots,M\}:m^\mu(\widetilde{\xi}^1)\geq\beta\}$ for $\beta>0$. The noise term would reach a similar strength as the signal if each pattern would contribute an amount of

$$\frac{1}{M-1}e^{N(1-2\rho)-2}=e^{(1-2\rho-\alpha)N-2} \tag{2.30}$$

to the sum in eq. (2.29). Here, we used $M-1=\exp(\alpha N)$. Set $0<\beta<1-2\rho-\alpha$ then the random set A contains all patterns whose contribution to the noise term is above the average. The random set A provides a partition:

$$\begin{aligned}
&\mathbb{P}\left(\sum_{\mu=2}^{M}\exp(Nm^{\mu}(\widetilde{\xi}^{1}))>e^{N(1-2\rho)-2}\right)\\
=&\sum_{X\subseteq\{2,\dots,M\}}\mathbb{P}\left(\left\{\sum_{\mu=2}^{M}\exp(Nm^{\mu}(\widetilde{\xi}^{1}))>e^{N(1-2\rho)-2}\right\}\cap\{A=X\}\right)\\
=&\sum_{X\subseteq\{2,\dots,M\}}\mathbb{P}\left(A=X\right)\cdot\mathbb{P}\left(\left\{\sum_{\mu=2}^{M}\exp(Nm^{\mu}(\widetilde{\xi}^{1}))>e^{N(1-2\rho)-2}\right\}\;\middle|\;\{A=X\}\right).
\end{aligned}$$

The distribution of A can be simplified by the fact that all patterns are generated as i.i.d. random variables. Similar to the standard Hopfield model, we can get rid of the random distraction through $\widetilde{\xi}^{1}$ with the help of Lemma 1.2. With this insight and the independence of patterns, we set

$$p=\mathbb{P}\left(2\in A\right)=\mathbb{P}\left(m^{2}(\widetilde{\xi}^{1})\geq\beta\right)\tag{2.31}$$

and see that the probability for A to be a fixed set $X\subseteq\{2,\dots,M\}$ equals the probability of a specific coin toss realization. Thus,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{\mu=2}^{M}\exp(Nm^{\mu}(\widetilde{\xi}^{1}))>e^{N(1-2\rho)-2}\right)\\
=&\sum_{k=0}^{M-1}\sum_{X\in P_k(\{2,\dots,M\})}p^{k}(1-p)^{M-1-k}\cdot\mathbb{P}\left(\sum_{\mu=2}^{M}\exp(Nm^{\mu}(\widetilde{\xi}^{1}))>e^{N(1-2\rho)-2}\;\middle|\;A=X\right),
\end{aligned}$$

where P_k denotes all the subsets of size k. Because of the i.i.d. setting the probabilities only vary with the size of A. Without loss of generality we are allowed to use $\bar{X}=\{2,\dots,k+1\}$ if $|A|=k$ for fixed k and conclude

$$\begin{aligned}
&\mathbb{P}\left(\sum_{\mu=2}^{M}\exp(Nm^{\mu}(\widetilde{\xi}^{1}))>e^{N(1-2\rho)-2}\right)\\
=&\sum_{k=0}^{M-1}\binom{M-1}{k}p^{k}(1-p)^{M-1-k}\cdot\mathbb{P}\left(\sum_{\mu=2}^{M}\exp(Nm^{\mu}(\widetilde{\xi}^{1}))>e^{N(1-2\rho)-2}\;\middle|\;A=\bar{X}\right).
\end{aligned}\tag{2.32}$$

The contribution of patterns in A^c to the noise term are bounded by $e^{\beta N}$. As a quick reminder, by definition pattern ξ^1 cannot be part of A or A^c. Therefore, $|A^c| = M-1-|A|$. Thus,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{\mu=2}^{M} \exp(Nm^\mu(\widetilde{\xi}^1)) > e^{N(1-2\rho)-2} \;\Big|\; A = \bar{X}\right) \\
\leq\; &\mathbb{P}\left(\sum_{\mu\in\bar{X}} \exp(Nm^\mu(\widetilde{\xi}^1)) > e^{N(1-2\rho)-2} - (M-1-k)\exp(\beta N) \;\Big|\; A = \bar{X}\right).
\end{aligned}$$

To shorten the equations set

$$\widetilde{c}(k) := e^{N(1-2\rho)-2} - (M-1-k)e^{\beta N}. \tag{2.33}$$

With a union bound argument, the event that the term exceeds $\widetilde{c}(k)$ can be bounded by an event that a large deviation of one pattern occurs:

$$\begin{aligned}
&\mathbb{P}\left(\sum_{\mu\in\bar{X}} \exp(Nm^\mu(\widetilde{\xi}^1)) > \widetilde{c}(k) \;\Big|\; A = \bar{X}\right) \\
\leq\; &\mathbb{P}\left(\max_{\mu\in\bar{X}} \exp\left(Nm^\mu(\widetilde{\xi}^1)\right) > \frac{\widetilde{c}(k)}{k} \;\Big|\; A = \bar{X}\right) \\
\leq\; &|\bar{X}| \cdot \mathbb{P}\left(\exp\left(Nm^2(\widetilde{\xi}^1)\right) > \frac{\widetilde{c}(k)}{k} \;\Big|\; A = \bar{X}\right) \\
=\; &k \cdot \mathbb{P}\left(\exp\left(Nm^2(\widetilde{\xi}^1)\right) > \frac{\widetilde{c}(k)}{k} \;\Big|\; 2 \in A\right).
\end{aligned} \tag{2.34}$$

Now define

$$r(k) := \mathbb{P}\left(\exp\left(Nm^2(\widetilde{\xi}^1)\right) > \frac{\widetilde{c}(k)}{k}\right) \tag{2.35}$$

and together with $p = \mathbb{P}(2 \in A)$ (see eq. (2.31)) it follows that

$$\begin{aligned}
&\mathbb{P}\left(\exp\left(Nm^2(\widetilde{\xi}^1)\right) > \frac{\widetilde{c}(k)}{k} \;\Big|\; 2 \in A\right) \\
\leq\; &\frac{\mathbb{P}\left(\left\{\exp\left(Nm^2(\widetilde{\xi}^1)\right) > \frac{\widetilde{c}(k)}{k}\right\} \cap \{2 \in A\}\right)}{\mathbb{P}(2 \in A)} \leq \frac{r(k)}{p}.
\end{aligned} \tag{2.36}$$

The estimates in eqs. (2.34) and (2.36) change eq. (2.32) into

$$\begin{aligned}\mathbb{P}\big(\widehat{T}_i\big(\widetilde{\xi}^\mu\big) \neq \xi_i^\mu\big) &= \mathbb{P}\left(\sum_{\mu=2}^{M} \exp(Nm^\mu(\widetilde{\xi}^1)) > e^{N(1-2\rho)-2}\right) \\ &\leq \sum_{k=0}^{M-1} \binom{M-1}{k} p^k (1-p)^{M-1-k} k \frac{r(k)}{p} \\ &= \sum_{k=0}^{M-1} k \binom{M-1}{k} p^{k-1} (1-p)^{M-1-k} r(k). \end{aligned} \tag{2.37}$$

In eq. (2.37) we derived an upper, bound which only depends on probabilities involving the overlap of $\widetilde{\xi}^1$ with one single pattern (here without loss of generality ξ^2). These probabilities are p (see eq. (2.31)) and $r(k)$ (see eq. (2.35)).

The random set A contains patterns whose contribution to the noise term is bigger than what on average is needed to exceed the signal term. In eq. (2.37) the size of A is denoted by k. With $M-1 = \exp(\alpha N)$ and the definition of $\widetilde{c}(k)$ in eq. (2.33), we see that

$$\begin{aligned}\frac{\widetilde{c}(k)}{k} = \frac{1}{k}\left(e^{N(1-2\rho)-2} - (M-1-k)e^{\beta N}\right) &= \frac{1}{k}\left(e^{N(1-2\rho)-2} - (M-1)e^{\beta N}\right) + e^{\beta N} \\ &= \frac{1}{k}\left(e^{N(1-2\rho)-2} - e^{(\alpha+\beta)N}\right) + e^{\beta N} \\ &= \frac{1}{k} e^{N(1-2\rho)-2}\left(1 - e^{(\alpha+\beta-(1-2\rho))N+2}\right) + e^{\beta N}, \end{aligned} \tag{2.38}$$

which is decreasing in k because $\beta < 1-2\rho-\alpha$. This means if more patterns are allowed to have a contribution above the average, each of them needs to have a lower impact to cause the net to fail an update. As a consequence, the probability $r(k)$ to reach this barrier $\frac{\widetilde{c}(k)}{k}$ is increasing in k. In other words, if more patterns are contained in A, it is more likely for the overlap of a pattern to exceed the critical value, namely $\frac{1}{N}\log(\frac{\widetilde{c}(k)}{k})$.

Because $|A|$ is binomial distributed with parameters p and $M-1$, the expected size of A is $p(M-1)$. Let us first consider the scenario where the size of A is not unusually big. In context of eq. (2.37), that means we concentrate on the sum over indices k which are less than $2p(M-1)$. Later we will see that sizes of A beyond $2p(M-1)$ can be eliminated with a large deviation argument about the binomial distribution.

The integer part of a number x is denoted by $\lfloor x \rfloor$. With the fact that $r(k)$, and therefore $kr(k)$, is increasing in k, we conclude for the sum with indices up to $\lfloor 2p(M-1) \rfloor$ that

$$\sum_{k=0}^{\lfloor 2p(M-1)\rfloor} \binom{M-1}{k} kr(k) p^{k-1} (1-p)^{M-1-k}$$

$$
\begin{aligned}
&\leq \frac{1}{p} \max_{l \leq \lfloor 2p(M-1) \rfloor} lr(l) \sum_{k=0}^{\lfloor 2p(M-1) \rfloor} \binom{M-1}{k} p^k (1-p)^{M-1-k} \\
&\leq \frac{1}{p} \lfloor 2p(M-1) \rfloor \, r(\lfloor 2p(M-1) \rfloor) \\
&\leq 2(M-1) \, r(2p(M-1)).
\end{aligned}
$$

In the second line, the series is equal to a probability of a binomial distribution and can be bounded by one. Furthermore, we used $\lfloor 2p(M-1) \rfloor \leq 2p(M-1)$ and the fact that $r(k)$ is increasing in k.

Again with the arguments around Lemma 1.2, the overlap $m^2(\widetilde{\xi}^1)$ is distributed like a sum of Rademacher spins. This justifies that we apply Cramér's theorem (see Theorem A.1 and Theorem A.3) to $m^2(\widetilde{\xi}^1)$ and deduce for $y > 0$

$$
\mathbb{P}\left(m^2(\widetilde{\xi}^1) \geq y\right) \leq \exp\left(-NI(y)\right), \tag{2.39}
$$

where I is the rate function of Rademacher spins. The rate function I is equal to

$$
I(x) = \frac{1}{2}\Big((1+x)\log\left((1+x)\right) + (1-x)\log\left(1-x\right)\Big) \tag{2.40}
$$

for $x \in [-1, 1]$ (see Example A.2 b). Cramér's theorem applied to p leads to

$$
\frac{1}{p} = \frac{1}{\mathbb{P}\left(m^2(\widetilde{\xi}^1) \geq \beta\right)} \geq \exp(NI(\beta)). \tag{2.41}
$$

Recall from eq. (2.35) that

$$
r(k) = \mathbb{P}\left(\exp\left(Nm^2(\widetilde{\xi}^1)\right) > \frac{\widetilde{c}(k)}{k}\right)
$$

with $\widetilde{c}(k) = e^{N(1-2\rho)-2} - e^{(\alpha+\beta)N} + ke^{\beta N}$. The identity $M - 1 = \exp(\alpha N)$ and the application of Cramér's theorem on $\frac{1}{p}$ as stated in eq. (2.41) bounds the probability $r(2p(M-1))$ by

$$
\begin{aligned}
r(2p(M-1)) &= \mathbb{P}\left(\exp\left(Nm^2(\widetilde{\xi}^1)\right) > \frac{1}{2p(M-1)}\left(e^{N(1-2\rho)-2} - e^{(\alpha+\beta)N}\right) + e^{\beta N}\right) \\
&= \mathbb{P}\left(\exp\left(Nm^2(\widetilde{\xi}^1)\right) > \frac{1}{2p}\left(e^{N(1-2\rho-\alpha)-2} - e^{\beta N}\right) + e^{\beta N}\right) \\
&\leq \mathbb{P}\left(\exp\left(Nm^2(\widetilde{\xi}^1)\right) > \frac{1}{2}\left(e^{N(1-2\rho-\alpha+I(\beta))-2} - e^{N(\beta+I(\beta))}\right)\right)
\end{aligned}
$$

Since $\beta < 1-2\rho-\alpha$, the dominating term is $e^{N(1-2\rho-\alpha+I(\beta))-2}$ and it follows that

$$\begin{aligned} r(2p(M-1)) &\leq \mathbb{P}\left(\exp\left(Nm^2(\widetilde{\xi}^1)\right) > e^{N(1-2\rho-\alpha+I(\beta))-(2+\log(2))}\left(1+o(1)\right)\right) \\ &= \mathbb{P}\left(Nm^2(\widetilde{\xi}^1) > N\left(1-2\rho-\alpha+I(\beta)-\frac{2+\log(2)}{N}\right)+o(1)\right). \end{aligned} \tag{2.42}$$

The upper bound of $r(2p(M-1))$ corresponds to a large deviation event of $m^2(\widetilde{\xi}^1)$ and as long as we can choose $\beta < 1-2\rho-\alpha$ such that there exists γ with

$$0 < \gamma \leq 1-2\rho-(\alpha-I(\beta))-\frac{2+\log(2)}{N}, \tag{2.43}$$

we can conclude with eq. (2.39) that

$$r(2p(M-1)) \leq e^{-I(\gamma)N}. \tag{2.44}$$

Therefore, the first part of eq. (2.37) is bounded by

$$\sum_{k=0}^{\lfloor 2p(M-1)\rfloor}\binom{M-1}{k}kr(k)p^{k-1}(1-p)^{M-1-k} \leq 2(M-1)\ \exp\left(-NI(\gamma)\right) \tag{2.45}$$

under the assumption that the parameters β and γ can be chosen appropriately (see eq. (2.43)). The right choice of β and γ will be discussed at the end of the proof.

Back to eq. (2.37), for the sum with indices above $\lfloor 2p(M-1)\rfloor$, we are using the identity $k\binom{M-1}{k} = (M-1)\binom{M-2}{k-1}$ and bound the probability $r(k)$ by one. Thus,

$$\begin{aligned} &\sum_{k=\lfloor 2p(M-1)\rfloor+1}^{M-1} k\binom{M-1}{k}p^{k-1}(1-p)^{M-1-k}r(k) \\ \leq\ & (M-1)\sum_{k=\lfloor 2p(M-1)\rfloor}^{M-2}\binom{M-2}{k}p^k(1-p)^{M-2-k} \\ =\ & (M-1)\ \mathbb{P}\left(S_{M-2} \geq \lfloor 2p(M-1)\rfloor\right) \\ \leq\ & (M-1)\ \mathbb{P}\Big(S_{M-2} \geq \frac{3}{2}p(M-2)\Big). \end{aligned} \tag{2.46}$$

Here, S_{M-2} denotes a binomial distributed random variable with parameters $M-2$ and p. The last line follows because $\lfloor 2p(M-1)\rfloor \geq \frac{3}{2}p(M-2)$ if and only if $Mp \geq 2(1-p)$. This is fulfilled because Mp grows to infinity if α is small enough, which we will see at the end of the proof.

The probability in eq. (2.46) can be bounded by Lemma A.4, which provides an exponential bound for a large deviation event of a binomial distribution. The lemma states that for a binomial distributed random variable S_n and $\varepsilon > 0$

$$\mathbb{P}\left(S_n \geq n(p+\varepsilon)\right) \leq \exp\left(-n\frac{\varepsilon^2}{(2p+\varepsilon)}\right).$$

Thus, by Lemma A.4 with $\varepsilon = \frac{p}{2}$, we obtain the following bound for the second part of the sum:

$$\sum_{k=\lfloor 2p(M-1)\rfloor+1}^{M-1} k\binom{M-1}{k}p^{k-1}(1-p)^{M-1-k}r(k) \leq (M-1)\exp\Big(-\frac{p(M-2)}{10}\Big). \quad (2.47)$$

The eqs. (2.45) and (2.47) bound the probability to update neuron i to the wrong spin and together with $M-1 = \exp(\alpha N)$, we conclude that

$$\begin{aligned}
&\mathbb{P}\Big(\exists\mu\ \exists i : \widehat{T}_i\big(\widetilde{\xi}^\mu\big) \neq \xi_i^\mu\Big) \\
\leq\ & N \cdot M \cdot \mathbb{P}\Big(\widehat{T}_i\big(\widetilde{\xi}^\mu\big) \neq \xi_i^\mu\Big) \\
\leq\ & N \cdot M \cdot \left(2(M-1)\exp\big(-NI(\gamma)\big) + (M-1)\exp\Big(-\frac{p(M-2)}{10}\Big)\right) \\
\leq\ & N\frac{M}{M-1}\left[2\exp\left(-N(I(\gamma)-2\alpha)\right) + \exp\Big(2\alpha N - \frac{M-2}{M-1}\cdot\frac{p(M-1)}{10}\Big)\right]. \quad (2.48)
\end{aligned}$$

It remains to prove that the bound in eq. (2.48) converges to zero if N tends to infinity. To achieve this, we show that the parameters β and γ can be chosen in such a way that the previously made assumption in eq. (2.43) is fulfilled, $I(\gamma) - 2\alpha > 0$ and $p(M-1)$ grows faster to infinity than N does.

The rate function I on $[0,1]$ is a strictly increasing function. Define $\beta_0 := I^{-1}(\alpha) > 0$. If $\beta < \beta_0$, then $\eta := \frac{1}{2}(\alpha - I(\beta)) > 0$. Cramér's theorem (see Theorem A.1) provides a lower bound for p: For N large enough we deduce that

$$\begin{aligned}
p = \mathbb{P}\left(m^2(\widetilde{\xi}^1) \geq \beta\right) &\geq \exp\left(-N\left(I(\beta)+\eta\right)\right) \\
&= \exp\left(-\frac{N}{2}(I(\beta)+\alpha)\right).
\end{aligned}$$

The lower bound for p together with the identity $M-1 = \exp(\alpha N)$ leads to

$$p(M-1) \geq \exp\left(N\left(\alpha - \frac{1}{2}(I(\beta)+\alpha)\right)\right) = \exp\left(\frac{N}{2}(\alpha - I(\beta))\right)$$

for N large enough. The lower bound goes exponentially fast to infinity if $\alpha - I(\beta) > 0$.

If we ensure that $\beta < \beta_0 = I^{-1}(\alpha)$, then the second term of eq. (2.48) is bounded from above by

$$\begin{aligned} & N \exp\Big(2\alpha N - \frac{M-2}{M-1} \cdot \frac{p(M-1)}{10}\Big) \\ \leq \ & \exp\left(\log(N) + 2\alpha N - \frac{1}{10}(1+o(1)) \exp\left(\frac{\alpha - I(\beta)}{2} N\right)\right) \overset{N\to\infty}{\longrightarrow} 0. \end{aligned} \tag{2.49}$$

Now it remains to show that $I(\gamma) - 2\alpha > 0$ while γ is restricted by eq. (2.43). For N large enough eq. (2.43) is fulfilled if

$$0 < \gamma < 1 - 2\rho - \alpha + I(\beta). \tag{2.50}$$

Since I is an increasing function on $[0,1]$, we want to choose β as large as possible, but the two constraints, namely $\beta < 1 - 2\rho - \alpha$ and $\beta < \beta_0 = I^{-1}(\alpha)$, are limiting this. The first constraint was used in eqs. (2.30), (2.38) and (2.42). The second one was important in eq. (2.49). Altogether, the choice of β is limited by $\min\{1 - 2\rho - \alpha, \beta_0\}$. The value of the minimum depends on ρ and α, which are fixed by the theorem. By assumptions $\rho \in [0, \frac{1}{2})$ and

$$\alpha < \frac{1}{2} \min\big\{I(1-2\rho),\ I\big(1 - 2\rho - \alpha + I(1 - 2\rho - \alpha)\big)\big\}. \tag{2.51}$$

We distinguish between the two cases whether $\min\{1 - 2\rho - \alpha, \beta_0\}$ is equal to $1 - 2\rho - \alpha$ or equal to β_0.

(1) Let $1 - 2\rho - \alpha$ be smaller or equal to β_0. Then $\min\{1 - 2\rho - \alpha, \beta_0\} = 1 - 2\rho - \alpha$ and we need to ensure that $\beta < 1 - 2\rho - \alpha$. Since $I(x) \leq x$, we know from $\alpha < \frac{I(1-2\rho)}{2}$ that $\alpha < 1 - 2\rho$. According to eq. (2.51) we know that

$$\alpha < \frac{I(1 - 2\rho - \alpha + I(1 - 2\rho - \alpha))}{2}.$$

Due to these conditions and because of the continuity of I, there exists $0 < \beta < 1 - 2\rho - \alpha$ with

$$2\alpha < I(1 - 2\rho - \alpha + I(\beta)) < I(1 - 2\rho - \alpha + I(1 - 2\rho - \alpha)).$$

Here, it is important that β can be chosen to be arbitrarily close to $1 - 2\rho - \alpha$. In the same way there exists $0 < \gamma < 1 - 2\rho - \alpha + I(\beta)$ such that

$$2\alpha < I(\gamma) < I(1 - 2\rho - \alpha + I(\beta)).$$

Thus, $I(\gamma) - 2\alpha > 0$.

We need to be careful if $1 - 2\rho - \alpha + I(1 - 2\rho - \alpha) > 1$ because in this case $I(1 - 2\rho - \alpha + I(1 - 2\rho - \alpha)) = \infty$. But since this condition is equivalent to the first inequality of $2\rho + \alpha < I(1 - 2\rho - \alpha) < I(1)$, we can choose $\beta < I^{-1}(2\rho + \alpha)$ such that

$$2\alpha < I(1 - 2\rho - \alpha + I(\beta)) < I(1)$$

because $2\alpha < I(1 - 2\rho - \alpha) < I(1)$. Then $\beta < 1 - 2\rho - \alpha$ is fulfilled. Additionally, we can choose $0 < \gamma < 1 - 2\rho - \alpha + I(\beta)$ such that

$$2\alpha < I(\gamma) < I(1 - 2\rho - \alpha + I(\beta))$$

because I is continuous and $1 - 2\rho - \alpha > 0$.

(2) Let α be smaller than $I(1-2\rho-\alpha)$. It follows that $\beta_0 = I^{-1}(\alpha) < 1-2\rho-\alpha$, which is equivalent to $\min\{1 - 2\rho - \alpha, \beta_0\} = \beta_0$. In this case we only need to guarantee that $\beta < \beta_0$. Thus, β can be chosen such that $\alpha - I(\beta)$ is positive but arbitrarily small. By assumptions (see eq. (2.51))

$$\alpha < \frac{I(1 - 2\rho)}{2}.$$

Because of the continuity of I there exists β such that

$$2\alpha < I(1 - 2\rho - \alpha + I(\beta)) < I(1 - 2\rho). \tag{2.52}$$

To achieve the inequality in (2.52), it is important that β is allowed to be close to β_0. Then β can be chosen such that $1 - 2\rho - \alpha + I(\beta)$ is arbitrarily close to $1 - 2\rho$. In the same way there exists γ such that

$$2\alpha < I(\gamma) < I(1 - 2\rho - \alpha + I(\beta)).$$

Thus, $I(\gamma) - 2\alpha > 0$ and all conditions are met.

This proves that the first term in eq. (2.48) converges to zero even if multiplied by N. All in all, we showed that

$$\mathbb{P}\left(\exists\mu\ \exists i : T_i\left(\widetilde{\xi}^\mu\right) \neq \xi_i^\mu\right) \to 0.$$

□

Figure 2.1 shows a plot of the function

$$f_\rho(\alpha) = \frac{1}{2}\min\{I(1 - 2\rho),\ I(1 - 2\rho - \alpha + I(1 - 2\rho - \alpha))\} - \alpha$$

for different values of ρ. For fixed ρ Theorem 2.2 states that α can be chosen such that the function $f_\rho(\alpha)$ is positive. Therefore, the root of $f_\rho(\alpha)$ determines the upper bound for α to obtain a storage capacity of $M = \exp(\alpha N)$.

All in all, Theorem 2.2 showed that by changing the dynamics such that the energy decreases exponentially fast while approaching a pattern, the net is able to store at least $M = \exp(\alpha N)$ patterns. This includes the stability of patterns and a positive basin of attraction.

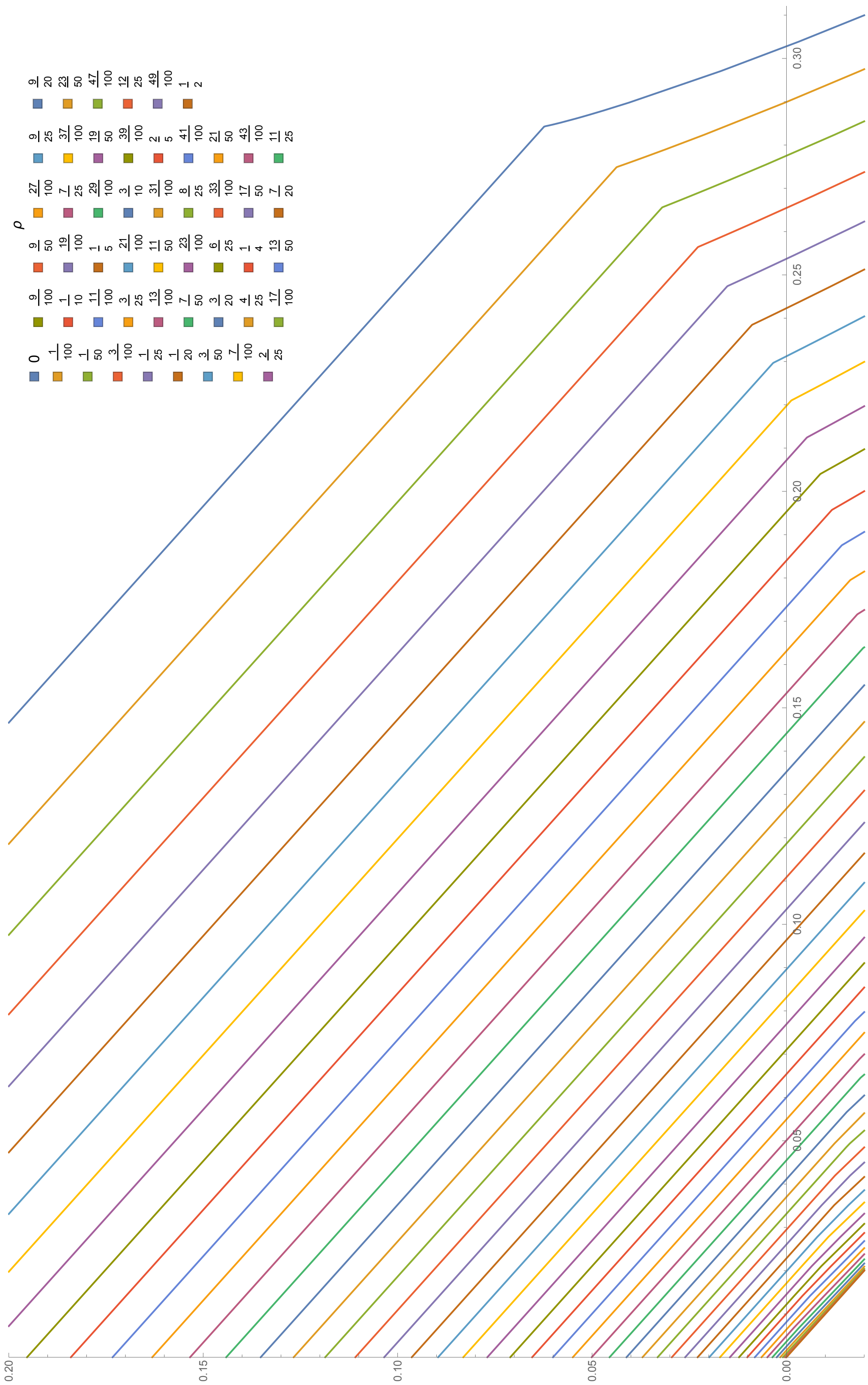

Figure 2.1.: Plots of the function $f_\rho(\alpha)$ for $\rho = 0, \frac{1}{100}, \dots, \frac{1}{2}$

3. Generalized Hopfield model with Curie-Weiss patterns

In Chapter 2 we derived lower bounds for the storage capacity of the Hopfield model with a polynomial and an exponential dynamics. In the next step, we investigate how the dynamics treat spatially dependent patterns. This means that different patterns are still independent but distinct spins of the same pattern are allowed to be dependent. Results with spatially dependent patterns for the standard Hopfield model were shown in [Löw98; LV05]. Löwe and Vermet showed in [LV05] that a moderate deviations principle is sufficient to prove a storage capacity of similar order as in the case of i.i.d. generated patterns. This was demonstrated with patterns created by a Curie-Weiss model and by an Ising model. In the following section, we give an introduction to the Curie-Weiss model and derive some results needed to calculate a storage capacity for a Hopfield model with a polynomial dynamics and Curie-Weiss patterns.

3.1. Introduction to the Curie-Weiss model

Spin glass models were introduced in Section 1.2. The Curie-Weiss model is a mean field model where all spins interact with the mean magnetization. Thus, we do not need to consider any lattice structure. The Hamiltonian of the Curie-Weiss model is given by

$$H_N^{CW}(\sigma) = -\frac{1}{2N}\sum_{i,j=1}^{N}\sigma_i\sigma_j = -\frac{N}{2}\left(\frac{1}{N}\sum_{j=1}^{N}\sigma_j\right)^2 \tag{3.1}$$

with $\sigma \in \{-1,1\}^N$. Therefore, the finite volume Gibbs measure equals

$$\mu_{\beta,N}(\sigma) = \frac{2^{-N}}{Z_{\beta,N}}\exp\left(-\beta H_N^{CW}(\sigma)\right) = \frac{2^{-N}}{Z_{\beta,N}}\exp\left(\frac{\beta N}{2}\left(\frac{1}{N}\sum_{j=1}^{N}\sigma_j\right)^2\right), \tag{3.2}$$

where $Z_{\beta,N}$ is the partition function with

$$Z_{\beta,N} = 2^{-N}\sum_{\widetilde{\sigma}\in\{-1,1\}^N}\exp\left(-\beta H_N^{CW}(\widetilde{\sigma})\right). \tag{3.3}$$

In this chapter we want the patterns to be generated according to a Curie-Weiss model. In the previous part of the work, we introduced $\xi := \xi^1 = (\xi_j)_{j\le N}$ as a vector of i.i.d. spins on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with

$$\mathbb{P}(\xi_1 = 1) = \frac{1}{2} = \mathbb{P}(\xi_1 = -1).$$

Now we denote by $\mathbb{P}_\beta^{CW}$ the probability measure on $(\Omega, \mathcal{F})$ such that the push-forward measure of ξ is equal to the Gibbs measure of the Curie-Weiss model, i.e.

$$\mathbb{P}_\beta^{CW}(\xi \in \,\mathrm{d}\sigma\,) = \mu_{\beta,N}(\,\mathrm{d}\sigma\,). \tag{3.4}$$

The expectation value with respect to $\mathbb{P}_\beta^{CW}$ is labelled $\mathbb{E}^{CW}[\cdot]$. A pattern with i.i.d. spins from the previous section is connected to a pattern according to a Curie-Weiss model by a simple change of measure:

$$\mathbb{P}_\beta^{CW}(\xi \in \,\mathrm{d}\sigma\,) = \frac{1}{Z_{\beta,N}} \exp\left(-\beta H_N^{CW}(\sigma)\right) \mathbb{P}(\xi \in \,\mathrm{d}\sigma\,), \tag{3.5}$$

where

$$Z_{\beta,N} = \mathbb{E}\left[\exp\left(-\beta H_N^{CW}(\xi)\right)\right]. \tag{3.6}$$

In the same way as before, let $\xi^1, \dots, \xi^M$ be independent copies of ξ. To avoid that the probability space depends on N one can define $(\xi_j)_{j\in\mathbb{N}}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ and work with the projection to the first N coordinates.

3.2. Results about the Curie-Weiss model

The Curie-Weiss model is a good first approach for spatially dependent patterns because the spins are exchangeable and the model parameter β can be used to adjust the correlation of spins. A deeper analysis of the model shows that for $0 < \beta < 1$ the mean magnetization converges to zero and the correlation between spins vanishes if N is going to infinity. In this case the spins are not independent but in the limit their behaviour is very similar to independent spins. For $\beta > 1$ the correlation between spins is strong enough such that non-trivial solutions for the mean magnetization appear. A result about the correlation of spins will be derived in Theorem 3.15. For results about the mean magnetization the reader is referred to chapter IV in [Ell85].

3.2.1. Central Limit Theorem for the magnetization

For i.i.d. patterns a useful tool to derive a lower bound for the storage capacity was Lemma 1.2, which shows that the overlap is distributed like a sum of single spins. Standard tools of probability theory, like the Central Limit Theorem, were then used to approximate the limiting behaviour. In the case of Curie-Weiss patterns for $0 < \beta < 1$, we are able to show a Central Limit Theorem for the sum of spins. Since the Curie-Weiss model was introduced to mimic a ferromagnet, the sum of spin is called magnetization.

Definition 3.1
Let $\xi = (\xi_j)_{j \leq N}$ be generated by a Curie-Weiss model. We call $S_N = \sum_{j=1}^N \xi_j$ the magnetization of the model.

Results about the magnetization in a Curie-Weiss model and its fluctuations were derived by Ellis, Newman and Rosen in [EN78a; EN78b; ENR80]. They proved a Law of Large Numbers and a Central Limit Theorem for S_N. In [Ell85] Ellis applied the theory of large deviations to models of statistical mechanics including the Curie-Weiss model. These results were extended by Eichelsbacher and Löwe in [EL04], who proved a moderate deviations principle for S_N. Rates of convergence were achieved with the help of Stein's method in [EL10; CS11]. For our purposes, it is important that the magnetization obeys a Central Limit Theorem if $\beta < 1$:

Theorem 3.2 (see Theorem V.9.4 in [Ell85])
Let $\xi = (\xi_j)_{j \leq N}$ be generated by a Curie-Weiss model with $0 < \beta < 1$. The scaled magnetization $\frac{1}{\sqrt{N}} S_N$ converges in distribution to a normal distribution with expectation value 0 and variance $\sigma_{CW}^2 := (1-\beta)^{-1}$.

Proof. The key observation is that a Central Limit Theorem is valid for $(\xi_j)_{j \leq N}$ under $\mathbb{P}$ because these are i.i.d. random variables. We need to show that

$$\mathbb{P}_\beta^{CW}\left(\frac{S_N}{\sqrt{N}} \in \cdot\right) \Rightarrow \mathcal{N}\left(0, \sigma_{CW}^2\right),$$

which means that for every bounded and continuous function f (short: $f \in \mathcal{C}_b(\mathbb{R})$)

$$\int_\Omega f\left(\frac{S_N}{\sqrt{N}}\right) \mathrm{d}\mathbb{P}_\beta^{CW} \overset{n\to\infty}{\longrightarrow} \mathbb{E}[f(Y)] \tag{3.7}$$

needs to hold. Here, Y is a $(0, \sigma_{CW}^2)$-normal distributed random variable. The observation in eq. (3.5) leads to the following representation of the integral

$$\int_\Omega f\left(\frac{S_N}{\sqrt{N}}\right) \mathrm{d}\mathbb{P}_\beta^{CW} = \frac{1}{Z_{N,\beta}} \int_\Omega f\left(\frac{S_N}{\sqrt{N}}\right) \exp\left(\frac{\beta}{2}\left(\frac{S_N}{\sqrt{N}}\right)^2\right) \mathrm{d}\mathbb{P}\,.$$

The convergence in eq. (3.7) follows if we can verify the statements

$$\int_\Omega f\left(\frac{S_N}{\sqrt{N}}\right)\exp\left(\frac{\beta}{2}\left(\frac{S_N}{\sqrt{N}}\right)^2\right)\,\mathrm{d}\mathbb{P} \overset{N\to\infty}{\longrightarrow} \int_\mathbb{R} f(y)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(1-\beta)y^2}{2}\right)\,\mathrm{d}y$$

and

$$Z_{N,\beta} = \int_\Omega \exp\left(\frac{\beta}{2}\left(\frac{S_N}{\sqrt{N}}\right)^2\right)\,\mathrm{d}\mathbb{P} \overset{N\to\infty}{\longrightarrow} \int_\mathbb{R} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(1-\beta)y^2}{2}\right)\,\mathrm{d}y \;= \sigma_{CW}.$$

Both statements would be an immediate consequence of the weak convergence of $\frac{1}{\sqrt{N}}\sum_{j=1}^N \xi_j$ under $\mathbb{P}$ if the function $g(x) = f(x)\exp(\frac{\beta}{2}x^2)$ would be bounded. Instead, we utilize that g is a continuous function and claim that $(g(\sqrt{N}^{-1}S_N))_{N\in\mathbb{N}}$ is uniformly integrable (see Definition A.7). The weak convergence and uniform integrability is enough to deduce the convergence which is left to show (see Theorem A.9). Thus, Theorem A.9 together with Lemma 3.3, which proves the uniform integrability of $(g(\sqrt{N}^{-1}S_N))_{N\in\mathbb{N}}$, completes the proof. □

Lemma 3.3 (see proof of Theorem 9.4 in Chapter V in [Ell85])
Let $0 < \beta < 1$. Define $(W_N)_{N\in\mathbb{N}}$ by

$$W_N = \exp\left(\frac{\beta}{2}\left(\frac{S_N}{\sqrt{N}}\right)^2\right) = \exp\left(\frac{\beta}{2}\left(\frac{1}{\sqrt{N}}\sum_{j=1}^N \xi_j\right)^2\right).$$

Then $(W_N)_{N\in\mathbb{N}}$ is uniformly integrable under $\mathbb{P}$.

Proof. We need to show that

$$\sup_{N\geq 1}\int_{\{W_N\geq\alpha\}} W_N\,\mathrm{d}\mathbb{P} \overset{\alpha\to\infty}{\longrightarrow} 0.$$

By calculating the expected value of $W_N\,\mathbb{1}_{\{W_N\geq\alpha\}}$ with a standard integral formula, it follows that

$$\begin{aligned}\int_{\{W_N\geq\alpha\}} W_N\,\mathrm{d}\mathbb{P} &= \int_0^\alpha \mathbb{P}(W_N\,\mathbb{1}_{\{W_N\geq\alpha\}} \geq t)\,\mathrm{d}t + \int_\alpha^\infty \mathbb{P}(W_N\,\mathbb{1}_{\{W_N\geq\alpha\}} \geq t)\,\mathrm{d}t \\ &= \alpha\,\mathbb{P}(W_N\geq\alpha) + \int_\alpha^\infty \mathbb{P}(W_N\,\mathbb{1}_{\{W_N\geq\alpha\}} \geq t)\,\mathrm{d}t \\ &= \alpha\,\mathbb{P}(W_N\geq\alpha) + \int_\alpha^\infty \mathbb{P}(W_N\geq t)\,\mathrm{d}t\,. \end{aligned} \tag{3.8}$$

For the tail probability of W_N we derive the following upper bound

$$\begin{aligned}
\mathbb{P}(W_N \geq \alpha) &= \mathbb{P}\left(\exp\left(\frac{\beta}{2}\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\xi_j\right)^2\right) \geq \alpha\right) \\
&= \mathbb{P}\left(\Big|\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\xi_j\Big| \geq \left(\frac{2\log(\alpha)}{\beta}\right)^{\frac{1}{2}}\right) \\
&= \mathbb{P}\left(\Big|\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\xi_j\Big| \geq x\right) \\
&= 2\,\mathbb{P}\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\xi_j \geq x\right) \\
&\leq 2\,\exp\left(-x^2\right)\mathbb{E}\left[\exp\left(x\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\xi_j\right)\right],
\end{aligned}$$

where $x = \left(\frac{2\log(\alpha)}{\beta}\right)^{\frac{1}{2}} > 0$. In the last line we used an exponential Chebyshev inequality with $t = x$. Under $\mathbb{P}$ the spins $\xi_1, \ldots, \xi_N$ are independent and identically Rademacher-distributed. Thus,

$$\mathbb{E}\left[\exp\left(x\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\xi_j\right)\right] = \cosh\left(\frac{x}{\sqrt{N}}\right)^N \leq \exp\left(\frac{x^2}{2}\right)$$

because $\cosh(y) \leq \exp(\frac{y^2}{2})$ for all $y \in \mathbb{R}$. Together, the tail event of W_N is bounded by

$$\mathbb{P}(W_N \geq \alpha) \leq 2\exp\left(-\frac{x^2}{2}\right) = 2\alpha^{-\frac{1}{\beta}}, \tag{3.9}$$

which does not depend on N and follows with the identity for x. By using the bound from eq. (3.9) in eq. (3.8), we conclude for $\beta < 1$ that

$$\begin{aligned}
\int_{\{W_N \geq \alpha\}} W_N \,\mathrm{d}\mathbb{P} &= \alpha\,\mathbb{P}(W_N \geq \alpha) + \int_{\alpha}^{\infty} \mathbb{P}(W_N \geq t)\,\mathrm{d}t \\
&\leq \alpha\, 2\alpha^{-\frac{1}{\beta}} + \int_{\alpha}^{\infty} 2t^{-\frac{1}{\beta}}\,\mathrm{d}t \\
&= 2\alpha^{1-\frac{1}{\beta}} + 2\frac{1}{\frac{1}{\beta}-1}\alpha^{1-\frac{1}{\beta}} = 2\frac{1}{1-\beta}\alpha^{1-\frac{1}{\beta}} = \mathcal{O}(\alpha^{1-\frac{1}{\beta}}).
\end{aligned}$$

Therefore,

$$\sup_{N\geq 1}\int_{\{W_N\geq\alpha\}} W_N \,\mathrm{d}\mathbb{P} \;\leq \mathcal{O}(\alpha^{1-\frac{1}{\beta}}),$$

which converges to 0 for $\alpha\to\infty$ because of $\beta<1$. Thus, $(W_N)_N$ is uniformly integrable. □

A simple consequence of Lemma 3.3 is the uniform integrability of the following family of random variables:

Corollary 3.4
For every $k\in\mathbb{N}$ the family $(Y_N^k)_{N\in\mathbb{N}}$, where Y_n^k is defined by

$$Y_N^k := \left(\frac{S_N}{\sqrt{N}}\right)^k \exp\left(\frac{\beta}{2}\left(\frac{S_N}{\sqrt{N}}\right)^2\right),$$

is uniformly integrable under $\mathbb{P}$.

Proof. Lemma 3.3 showed that for every $0<\beta<1$ the exponential part in Y_n^k is uniformly integrable. But the polynomial part of Y_N^k is negligible. Let $0<\varepsilon<1-\beta$ then there exists $\gamma_k>0$ such that $A_k := \{x\in\mathbb{R} : k\log(|x|) > \frac{1-\varepsilon-\beta}{2}x^2\} \subseteq \{x\in\mathbb{R} : |x|\leq\gamma_k\}$. On the event $\left\{\frac{S_N}{\sqrt{N}}\in A_k^c\right\} = \left\{\left|\frac{S_N}{\sqrt{N}}\right| > \gamma_k\right\}$ we know that

$$\left|\frac{S_N}{\sqrt{N}}\right|^k \leq \exp\left(\frac{1-\varepsilon-\beta}{2}\;\left(\frac{S_N}{\sqrt{N}}\right)^2\right)$$

and therefore

$$\begin{aligned}&\left\{\left|\frac{S_N}{\sqrt{N}}\right|^k \exp\left(\frac{\beta}{2}\left(\frac{S_N}{\sqrt{N}}\right)^2\right)\geq\alpha,\ \frac{S_N}{\sqrt{N}}\in A_k\right\}\\ =\ &\left\{\left|\frac{S_N}{\sqrt{N}}\right|^k \exp\left(\frac{\beta}{2}\left(\frac{S_N}{\sqrt{N}}\right)^2\right)\geq\alpha,\ \left|\frac{S_N}{\sqrt{N}}\right|\leq\gamma_k\right\} = \varnothing\end{aligned}$$

if α is large enough. Thus,

$$\begin{aligned}&\lim_{\alpha\to\infty}\sup_{N\in\mathbb{N}}\int_{\{|Y_N^k|\geq\alpha\}} |Y_N^k|\;\mathrm{d}\mathbb{P}\\ =\ &\lim_{\alpha\to\infty}\sup_{N\in\mathbb{N}}\int |Y_N^k|\cdot\mathbb{1}_{\left\{\left|\frac{S_N}{\sqrt{N}}\right|^k\exp\left(\frac{\beta}{2}\left(\frac{S_N}{\sqrt{N}}\right)^2\right)\geq\alpha,\ \frac{S_N}{\sqrt{N}}\in A_k^c\right\}}\;\mathrm{d}\mathbb{P}\end{aligned}$$

$$\leq \limsup_{\alpha\to\infty} \sup_{N\in\mathbb{N}} \int_{\left\{\exp\left(\frac{1-\varepsilon}{2}\left(\frac{S_N}{\sqrt{N}}\right)^2\right)\geq\alpha\right\}} \exp\left(\frac{1-\varepsilon}{2}\left(\frac{S_N}{\sqrt{N}}\right)^2\right) \mathrm{d}\mathbb{P} = 0.$$

The last term is equal to zero because of Lemma 3.3 with $\widetilde{\beta} = 1-\varepsilon$. □

Corollary 3.4 is helpful to show the convergence of moments in a Curie-Weiss model with $0 < \beta < 1$:

Lemma 3.5

Let $\xi = (\xi_j)_{j\leq N}$ be a Curie-Weiss pattern for $0<\beta<1$ and let $\sigma_{CW}^2 = (1-\beta)^{-1}$. Then the k-th moment of $\sqrt{N}^{-1} S_N$ under $\mathbb{P}_\beta^{CW}$ converges to the k-th moment of a $\mathcal{N}(0, \sigma_{CW}^2)$ distribution, i.e.

$$\mathbb{E}^{CW}\left[\left(\frac{S_N}{\sqrt{N}}\right)^k\right] = \mathbb{E}^{CW}\left[\left(\frac{1}{\sqrt{N}}\sum_{j=1}^N \xi_j\right)^k\right] \to \begin{cases} \sigma_{CW}^k (k-1)!!, & \text{if } k \text{ is even} \\ 0, & \text{if } k \text{ is odd} \end{cases}.$$

Proof. The Central Limit Theorem states that $\sqrt{N}^{-1} S_N \Rightarrow \mathcal{N}(0,1)$ under $\mathbb{P}$ and from Corollary 3.4 we know that Y_N^k is uniformly integrable under $\mathbb{P}$. Theorem A.9 then states the following convergence:

$$\begin{aligned}\mathbb{E}^{CW}\left[\left(\frac{S_N}{\sqrt{N}}\right)^k\right] &= \frac{1}{Z_{N,\beta}}\mathbb{E}\left[\left(\frac{S_N}{\sqrt{N}}\right)^k \exp\left(\frac{\beta}{2}\left(\frac{S_N}{\sqrt{N}}\right)^2\right)\right] \\ &= \frac{1}{Z_{N,\beta}}\mathbb{E}\left[Y_N^k\right] \overset{N\to\infty}{\longrightarrow} \frac{1}{\sigma_{CW}}\mathbb{E}\left[Z^k \exp\left(\frac{\beta}{2}Z^2\right)\right]\end{aligned}$$

with $Z \sim \mathcal{N}(0,1)$. The partition function was handled in the proof of Theorem 3.2, where we saw that $Z_{N,\beta} \overset{N\to\infty}{\longrightarrow} \sigma_{CW}$. A simple calculation shows that

$$\mathbb{E}\left[Z^k \exp\left(\frac{\beta}{2}Z^2\right)\right] = \begin{cases} \sigma_{CW}^{k+1}\cdot(k-1)!! & k \text{ even} \\ 0 & k \text{ uneven} \end{cases},$$

where $(k-1)!! = (k-1)\cdot(k-3)\cdot\ldots\cdot 3\cdot 1$. Thus,

$$\mathbb{E}^{CW}\left[\left(\frac{S_N}{\sqrt{N}}\right)^k\right] \overset{N\to\infty}{\longrightarrow} \begin{cases} \sigma_{CW}^k\cdot(k-1)!! & k \text{ even} \\ 0 & k \text{ uneven} \end{cases}.$$

□

3.2.2. The Gibbs measure of a Curie-Weiss model as a de Finetti type measure

In the last section we proved a Central Limit Theorem for the magnetization. Now we want to calculate the correlation between spins and then prove a Central Limit Theorem for the overlap of two independent Curie-Weiss patterns. For this task it is helpful to use another representation of the Gibbs measure. We show that the Gibbs measure is a de Finetti type measure, which is by de Finetti's theorem (see [Ald85]) connected to exchangeable sequences of random variables. Since the density in a Curie-Weiss model (see eq. (3.2)) only depends on the magnetization, the distribution of a configuration is invariant under a permutation of its entries. Random variables with this characteristic are called exchangeable. We denote by $\mathcal{L}(X)$ the law of a random variable X.

Definition 3.6 (see [Ald85])
A finite sequence $(X_1, \dots, X_n)$ of random variables is called exchangeable if

$$\mathcal{L}(X_1, \dots, X_n) = \mathcal{L}(X_{\pi(1)}, \dots, X_{\pi(n)})$$

for each permutation π of $\{1, \dots, n\}$.
An infinite sequence $(X_1, X_2, \dots)$ of random variables is called exchangeable if

$$\mathcal{L}(X_1, X_2, \dots) = \mathcal{L}(X_{\pi(1)}, X_{\pi(2)}, \dots)$$

for each finite permutation π of $\mathbb{N}$. A permutation of $\mathbb{N}$ is finite if $|\{i : \pi(i) \neq i\}| < \infty$.

For an infinite sequence which is exchangeable de Finetti's theorem states that this sequence can be written as a mixture of i.i.d. random variables (see [Ald85]). In general this is not true for finite sequences, but we will show that the vector of Curie-Weiss spins can still be written as such a mixture of i.i.d. random variables.

Definition 3.7 (see Definition 17 in [HKW15])
Let μ be a probability measure on $[-1, 1]$. We say $\{-1, 1\}$-valued random variables $X_1, \dots, X_n$ are of de Finetti type if

$$\mathbb{P}(X_1 = a_1, \dots, X_n = a_n) = \int_{-1}^{1} P_t((a_1, \dots, a_n)) d\mu(t),$$

where

$$P_t((a_1, \dots, a_n)) = \frac{1}{2^n}(1+t)^{n_+(a_1,\dots,a_n)}(1-t)^{n_-(a_1,\dots,a_n)} \tag{3.10}$$

with $n_+(a_1, \dots, a_n)$ and $n_-(a_1, \dots, a_n)$ counting the occurrences of $+1$ resp. -1. μ is called the de Finetti measure.

The Gibbs measure of a Curie-Weiss model is a measure of de Finetti type.

Theorem 3.8 (see Theorem 20 in [HKW15])
If $\mathbb{P}_\beta^{CW}$ *denotes the Gibbs measure on* $\{-1,1\}^N$ *of a Curie-Weiss model for* $\beta > 0$*, then*

$$\mathbb{P}_\beta^{CW}((\xi_j)_{j\leq N} = \sigma) = Z^{-1} \int_{-1}^{1} P_t(\sigma_1, \ldots, \sigma_N)\, \mathbb{Q}(dt),$$

where

$$\mathbb{Q}(dt) = \frac{\exp(-\frac{N}{2} F_\beta(t))}{1-t^2} dt$$

with

$$F_\beta(t) = \frac{1}{\beta}\left(\frac{1}{2}\log\left(\frac{1+t}{1-t}\right)\right)^2 + \log\left(1-t^2\right) \tag{3.11}$$

and

$$Z = \int_{-1}^{1} \frac{\exp(-\frac{N}{2} F_\beta(t))}{1-t^2} dt.$$

P_t *is defined as in eq.* (3.10)*.*

A helpful tool to prove Theorem 3.8 is the Hubbard-Stratonovich transformation. This transformation can be used to convert the density function to a simpler form such that we can easily use the independence of spins under $\mathbb{P}$.

Theorem 3.9 (Hubbard-Stratonovich transformation)
For $a > 0$ *and* $b \in \mathbb{R}$

$$\sqrt{\frac{\pi}{a}}\exp\left(\frac{b^2}{4a}\right) = \int_{-\infty}^{\infty} \exp\left(-as^2 + bs\right)\, ds\,. \tag{3.12}$$

Proof. Based on the Gaussian density function we know that

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(s-\mu)^2}{2\sigma^2}\right)\, \mathrm{d}s\, = 1$$

for $\mu \in \mathbb{R}$ and $\sigma > 0$. The identity

$$\begin{aligned}\left(\sqrt{a}s - \frac{b}{2\sqrt{a}}\right)^2 &= as^2 - 2\sqrt{a}s\frac{b}{2\sqrt{a}} + \frac{b^2}{4a} \\ &= as^2 - bs + \frac{b^2}{4a}\end{aligned}$$

transforms the right side of eq. (3.12) into

$$\begin{aligned}\int_{-\infty}^{\infty} \exp\left(-as^2 + bs\right) \mathrm{d}s &= \int_{-\infty}^{\infty} \exp\left(-\left(\sqrt{a}s - \frac{b}{2\sqrt{a}}\right)^2\right) \exp\left(\frac{b^2}{4a}\right) \mathrm{d}s \\ &= \exp\left(\frac{b^2}{4a}\right) \int_{-\infty}^{\infty} \exp\left(-a\left(s - \frac{b}{2a}\right)^2\right) \mathrm{d}s \\ &= \exp\left(\frac{b^2}{4a}\right) \sqrt{\frac{\pi}{a}} \int_{-\infty}^{\infty} \sqrt{\frac{a}{\pi}} \exp\left(-\frac{(s - \frac{b}{2a})^2}{2\frac{1}{2a}}\right) \mathrm{d}s \\ &= \sqrt{\frac{\pi}{a}} \exp\left(\frac{b^2}{4a}\right).\end{aligned}$$

In the last step we integrate the density function of a normal distribution with $\mu = \frac{b}{2a}$ and $\sigma^2 = \frac{1}{2a}$. Thus, the integral is equal to one. □

Proof of Theorem 3.8. Let σ be in $\{-1, 1\}^N$. We need to convert the probability

$$\mathbb{P}_\beta^{CW}(\xi = \sigma) = 2^{-N} Z_{\beta,N}^{-1} \exp\left(\frac{\beta}{2}\left(\frac{1}{\sqrt{N}}\sum_{j=1}^N \sigma_j\right)^2\right)$$

into the desired form needed for a de Finetti type measure. The Hubbard-Stratonovich transformation (see Theorem 3.9) applied to the density function of $\mathbb{P}_\beta^{CW}$, i.e.

$$a = \frac{1}{2} \qquad \text{and} \qquad b = \sqrt{\frac{\beta}{N}} \sum_{j=1}^N \sigma_j,$$

leads to the identity

$$\exp\left(\frac{\beta}{2}\left(\frac{1}{\sqrt{N}}\sum_{j=1}^N \sigma_j\right)^2\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{s^2}{2} + s\sqrt{\frac{\beta}{N}}\sum_{j=1}^N \sigma_j\right) \mathrm{d}s\,.$$

Now we are able to transform the exponential of the sum into a product of exponentials. Together with a change of variables $y = s\sqrt{\frac{\beta}{N}}$, we get

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{s^2}{2} + s\sqrt{\frac{\beta}{N}} \sum_{j=1}^{N} \sigma_j\right) \mathrm{d}s \\
&= \sqrt{\frac{N}{2\pi\beta}} \int_{-\infty}^{\infty} \exp\left(-\frac{N}{2\beta} y^2\right) \prod_{j=1}^{N} \exp\left(\sigma_j y\right) \mathrm{d}y \\
&= \sqrt{\frac{N}{2\pi\beta}} \int_{-\infty}^{\infty} \exp\left(-N\left(\frac{1}{2\beta} y^2 - \log(\cosh(y))\right)\right) \prod_{j=1}^{N} \frac{\exp\left(\sigma_j y\right)}{\cosh(y)} \mathrm{d}y\,. \qquad (3.13)
\end{aligned}$$

Additionally, we corrected the exponential functions by a cosh-term. With the symmetry of cosh, we conclude that for every $\sigma_j \in \{-1, 1\}$

$$\begin{aligned}
\frac{e^{\sigma_j y}}{\cosh(y)} = \frac{e^{\sigma_j y}}{\cosh(\sigma_j y)} &= \frac{1}{2} \frac{\left(e^{\sigma_j y} + e^{-\sigma_j y}\right) + \left(e^{\sigma_j y} - e^{\sigma_j y}\right)}{\cosh(\sigma_j y)} \\
&= \frac{1}{2} \frac{\cosh(\sigma_j y) + \sinh(\sigma_j y)}{\cosh(\sigma_j y)} \\
&= \frac{1}{2}\left(1 + \tanh(\sigma_j y)\right) = \frac{1}{2}\left(1 + \sigma_j \tanh(y)\right).
\end{aligned}$$

After a change of variables $t = \tanh(y)$, the product in eq. (3.13) already has the form of P_t:

$$\begin{aligned}
\prod_{j=1}^{N} \frac{\exp\left(\sigma_j y\right)}{\cosh(y)} &= \prod_{j=1}^{N} \frac{1}{2}\left(1 + \sigma_j \tanh(y)\right) \\
&= \frac{1}{2^N} \prod_{j=1}^{N} (1 + \sigma_j t) = P_t(\sigma_1, \ldots, \sigma_N). \qquad (3.14)
\end{aligned}$$

It remains to show that the de Finetti measure $\mathbb{Q}$ and the function F_β is as stated in the theorem. For this we use that

$$y = \tanh^{-1}(t) = \frac{1}{2} \log\left(\frac{1+t}{1-t}\right)$$

as well as

$$\log(\cosh(\tanh(t))) = -\frac{1}{2}\log(1 - \tanh(y)^2) = -\frac{1}{2}\log(1 - t^2)$$

to deduce

$$\begin{aligned}\frac{1}{2\beta}y^2 - \log(\cosh(y)) &= \frac{1}{2\beta}\tanh^{-1}(t)^2 - \log(\cosh(\tanh^{-1}(t))) \\ &= \frac{1}{2\beta}\left(\frac{1}{2}\log\left(\frac{1+t}{1-t}\right)\right)^2 + \frac{1}{2}\log(1-t^2) \\ &= \frac{1}{2}F_\beta(t). \end{aligned} \tag{3.15}$$

Here, $\tanh^{-1}$ is the inverse hyperbolic tangent. Equation (3.15) applied to eq. (3.13) leads to

$$\begin{aligned}\exp\left(\frac{\beta}{2}\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\sigma_j\right)^2\right) &= \sqrt{\frac{N}{2\pi\beta}}\int_{-1}^{1}\frac{\exp\left(-\frac{N}{2}F_\beta(t)\right)}{1-t^2}P_t(\sigma)\,\mathrm{d}t \\ &= \sqrt{\frac{N}{2\pi\beta}}\int_{-1}^{1}P_t(\sigma)\,\mathbb{Q}(\,\mathrm{d}t\,). \end{aligned} \tag{3.16}$$

The term

$$\frac{\mathrm{d}t}{\mathrm{d}y} = (1-\tanh(y)^2) = (1-t^2)$$

occurs because of the change of variables. The identity in eq. (3.16) applied to $Z_{N,\beta}$ shows that

$$\begin{aligned}Z_{N,\beta} &= \sum_{(\sigma_j)_j\in\{-1,1\}^N} 2^{-N}\exp\left(\frac{\beta}{2}\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\sigma_j\right)^2\right) \\ &= 2^{-N}\sqrt{\frac{N}{2\pi\beta}}\int_{-1}^{1}\sum_{(\sigma_j)_j\in\{-1,1\}^N}P_t(\sigma)\,\mathbb{Q}(\mathrm{d}\,t). \end{aligned} \tag{3.17}$$

With the identity for $P_t(\sigma)$ stated in eq. (3.14), we see that the sum over $P_t(\sigma)$ can be expressed as an expectation value of i.i.d. Rademacher spins:

$$\sum_{(\sigma_j)_j\in\{-1,1\}^N} P_t(\sigma) = \Big(\mathbb{E}[(1+\xi_1 t)]\Big)^N = 1. \tag{3.18}$$

All in all, this proves

$$Z_{N,\beta} = 2^{-N} \sqrt{\frac{N}{2\pi\beta}}\, Z,$$

and together with eq. (3.16), we proved

$$\mathbb{P}_{\beta}^{CW}((\xi_j)_{j\leq N} = \sigma) = \frac{1}{Z} \int_{-1}^{1} \frac{\exp\left(-\frac{N}{2} F_\beta(t)\right)}{(1-t^2)} P_t(\sigma)\, \mathrm{d}t\,.$$

□

3.2.3. Laplace's method and the correlation of spins

The expectation of a product of spins is easy to calculate if we work with independent spins. In a Curie-Weiss model the spins are not independent, but the representation as a de Finetti measure can be interpreted as a mixture of independent variables. This can be used to show the following formula to calculate the correlation of spins.

Proposition 3.10 (see Proposition 18 in [HKW15])
For Curie-Weiss spins $(\xi_j)_{j\leq N}$ the correlation can be calculated as

$$\mathbb{E}^{CW}[\xi_1 \cdot \ldots \cdot \xi_k] = Z^{-1} \int_{-1}^{1} t^k \frac{\exp(-\frac{N}{2} F_\beta(t))}{1-t^2}\, dt\,,$$

where F_β is given by eq. (3.11).

Proof. We use Theorem 3.8 and conclude that for every $\beta > 0$

$$\begin{aligned}
\mathbb{E}^{CW}[\xi_1 \cdot \ldots \cdot \xi_k] &= \sum_{(\sigma_j)_j \in \{-1,1\}^N} \left(\prod_{i=1}^{k} \sigma_i\right) \mathbb{P}_{\beta}^{CW}(\sigma_1, \ldots, \sigma_N) \\
&= \sum_{(\sigma_j)_j \in \{-1,1\}^N} \left(\prod_{i=1}^{k} \sigma_i\right) Z^{-1} \int_{-1}^{1} P_t(\sigma_1, \ldots, \sigma_n)\, \mathbb{Q}(\mathrm{d}\, t) \\
&= Z^{-1} \int_{-1}^{1} \sum_{(\sigma_j)_j \in \{-1,1\}^N} \left(\prod_{i=1}^{k} \sigma_i\right) P_t(\sigma_1, \ldots, \sigma_n)\, \mathbb{Q}(\mathrm{d}\, t).
\end{aligned}$$

Similar to eq. (3.18), the sum can be expressed as an expectation value of independent Rademacher spins:

$$\sum_{(\sigma_j)_j\in\{-1,1\}^N}\left(\prod_{i=1}^{k}\sigma_i\right)P_t(\sigma_1,\ldots,\sigma_n)=\Big(E[\xi_1(1+\xi_1 t)]\Big)^k\Big(\mathbb{E}[(1+\xi_1 t)]\Big)^{N-k}=t^k.$$

This proves

$$\mathbb{E}^{CW}[\xi_1\cdot\ldots\cdot\xi_k]=Z^{-1}\int_{-1}^{1}t^k\,\frac{\exp(-\frac{N}{2}F_\beta(t))}{1-t^2}\,\mathrm{d}t\,.$$

□

As a simple consequence of the formula in Proposition 3.10, we get a first result about the correlation of spins:

Corollary 3.11
Let $\beta>0$. For every $k\in N$

$$\mathbb{E}^{CW}\left[\xi_1\cdot\ldots\cdot\xi_k\right]\geq 0$$

and for k odd

$$\mathbb{E}^{CW}\left[\xi_1\cdot\ldots\cdot\xi_k\right]=0.$$

Proof. Proposition 3.10 and the identity for F (see eq. (3.11)), especially the fact that F is symmetric, i.e. $F(t)=F(-t)$ for all $t>0$, shows that for k odd

$$\mathbb{E}^{CW}[\xi_1\cdot\ldots\cdot\xi_k]=Z^{-1}\int_{-1}^{1}t^k\frac{\exp(-\frac{N}{2}F_\beta(t))}{1-t^2}\,\mathrm{d}t\;=0.$$

If k is even, the integrand is non-negative. Thus, the expectation is non-negative. □

To calculate the order of the correlation with an even number of spins, we introduce Laplace's method. With this method we can determine the behaviour of integrals of the form

$$\int_c^d e^{-NF(x)}\phi_N(x)\,\mathrm{d}x$$

through the minimal points of F. In our case this means we need to calculate the minimum of F_β to get an approximation. For a general version of Laplace's method one needs several assumptions to guarantee that the dominating contribution comes from the exponential part. But as long as F is continuously differentiable and an integrability condition is fulfilled, Laplace's method provides a statement strong enough for our purposes. A more general version can be found in [Olv74; Won01; Kir15]. In the following we prove a version of Laplace's method together with its application to the Curie-Weiss model.

Definition 3.12
The Gamma function is defined by

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \, dt$$

for $x > 0$.

The following theorem is adapted from Proposition 24 in [HKW15] and Theorem 5.10 resp. Corollary 5.12 in [Kir15]. For more details on Laplace's method the reader is referred to [Olv74].

Theorem 3.13 (Laplace's method (see Theorem 5.10 and Corollary 5.12 in [Kir15]))
Suppose $F : (c, d) \to \mathbb{R}$ *(with* $c \in \mathbb{R} \cup \{-\infty\}$ *and* $d \in \mathbb{R} \cup \{\infty\}$*) is* $(k+1)$*-times continuously differentiable for an even* k *and assume that for some* $a \in (c, d)$*:*

$$F(a) = F'(a) = \ldots = F^{(k-1)}(a) = 0 \quad \text{and} \quad F^{(k)}(a) > 0$$

and $F'(x) \neq 0$ *for* $x \neq a$*. Suppose furthermore that* $\phi : (c, d) \to \mathbb{R}$ *is continuous at* a *with* $\phi(a) \neq 0$ *and such that*

$$\int_c^d e^{-F(x)} \left|\phi(x)\right| \left|x - a\right|^l \, dy \, < \infty. \tag{3.19}$$

Then

(1) For l *even*

$$\lim_{N\to\infty} N^{\frac{l+1}{k}} \int_c^d e^{-\frac{N}{2}F(x)} (x-a)^l \phi(x) \, dx \, = 2 \, \phi(a) \left(\frac{k!}{F^{(k)}(a)}\right)^{\frac{l+1}{k}} \frac{2^{\frac{l+1}{k}}}{k} \, \Gamma\left(\frac{l+1}{k}\right).$$

(2) For l *odd*

$$\lim_{N\to\infty} N^{\frac{l+1}{k}} \int_c^d e^{-\frac{N}{2}F(x)} (x-a)^l \phi(x) \, dx \, = 0.$$

Proof. With the assumptions on F, especially that $F^{(k)}(a) > 0$ for k even, it follows that F has a unique minimum in a, i.e.

$$\inf_{|x-a|\geq\delta} F(x) > F(a) = 0$$

for every $\delta > 0$. For a fixed $\delta > 0$ there is a $B > 0$ such that $F(x) \geq B > 0$ for all $x > \delta$.

Without loss of generality $a = 0$ because we are always able to do a change of variables in the integral. Let us first concentrate on the integral over the positive line. For $\delta > 0$ we split the integral into a part close to the minimal point at $a = 0$ and a part bounded away from the critical point:

$$\int_0^d e^{-\frac{N}{2}F(x)} x^l \phi(x)\,\mathrm{d}x = \int_0^\delta e^{-\frac{N}{2}F(x)} x^l \phi(x)\,\mathrm{d}x + \int_\delta^d e^{-\frac{N}{2}F(x)} x^l \phi(x)\,\mathrm{d}x\,.$$

In the second integral F can be bounded from below by B and with the integrability from eq. (3.19) we conclude that

$$\begin{aligned}\int_\delta^d e^{-\frac{N}{2}F(x)} x^l \phi(x)\,\mathrm{d}x &\leq e^{-\frac{N-2}{2}B} \int_\delta^d e^{-F(x)} x^l \phi(x)\,\mathrm{d}x \\ &\leq e^{-\frac{N-2}{2}B} \int_c^d e^{-F(x)} |x|^l |\phi(x)|\,\mathrm{d}x \\ &\leq Ce^{-\frac{N-2}{2}B},\end{aligned}$$

which converges to zero for $N \to \infty$ even if multiplied by $N^{\frac{l+1}{k}}$ for any $l, k \in \mathbb{N}$.

For the integral considering values close to the minimum of F, we use Taylor's theorem (see Theorem A.10). Close to $a = 0$ the function F can be approximated by a polynomial of order k

$$F(x) = \frac{F^{(k)}(0)}{k!} x^k + r(x),$$

where r is the remainder term with $|r(x)| \leq C|x|^{k+1}$ if $x \leq \delta$ and δ small enough. Here, it is important that the first $k-1$ derivatives vanish at a. Thus, with δ small enough

$$\int_0^\delta e^{-\frac{N}{2}F(x)} x^l \phi(x)\,\mathrm{d}x = \int_0^\delta e^{-\frac{N}{2}\frac{F^{(k)}(0)}{k!}x^k + r(x)} x^l \phi(x)\,\mathrm{d}x\,.$$

Set $A = \frac{f^{(k)}(0)}{k!}$ then a change of variables $y = (NA)^{\frac{1}{k}} x$ leads to

$$\begin{aligned}
&\int_0^{\delta} e^{-\frac{N}{2}\frac{F^{(k)}(0)}{k!}x^k + r(x)} x^l \phi(x)\,\mathrm{d}x \\
&= (NA)^{-\frac{l+1}{k}} \int_0^{\delta(NA)^{1/k}} e^{-\frac{1}{2}y^k} y^l e^{-\frac{N}{2}r\left(\frac{y}{(NA)^{1/k}}\right)} \phi\left(\frac{y}{(NA)^{1/k}}\right)\mathrm{d}y \\
&= (NA)^{-\frac{l+1}{k}} \int_0^{\infty} e^{-\frac{1}{2}y^k} y^l\, \mathbb{1}_{[0,\delta(NA)^{1/k}]}(y)\, e^{-\frac{N}{2}r\left(\frac{y}{(NA)^{1/k}}\right)} \phi\left(\frac{y}{(NA)^{1/k}}\right)\mathrm{d}y\,.
\end{aligned}$$

The function ϕ is continuous in $x = 0$ and therefore

$$\phi\left(\frac{y}{(NA)^{\frac{1}{k}}}\right) \to \phi(0)$$

for $N \to \infty$. Taylor's theorem states that $|r(x)| \le C|x|^{k+1}$ for $x \le \delta$. Thus,

$$\left| r\left(\frac{y}{(NA)^{\frac{1}{k}}}\right)\right| \le C\left|\frac{y^{k+1}}{(NA)^{1+\frac{1}{k}}}\right| = \widetilde{C}|y|^{k+1}N^{-(1+\frac{1}{k})} \tag{3.20}$$

and because this is decreasing faster than N, we conclude that

$$\exp\left(-\frac{N}{2}\, r\left(\frac{y}{(NA)^{\frac{1}{k}}}\right)\right) \to 1$$

for $N \to \infty$. Using dominated convergence and assuming that the requirements are fulfilled, it follows that

$$\begin{aligned}
&\lim_{N\to\infty} N^{\frac{l+1}{k}} \int_0^{d} e^{-\frac{N}{2}F(x)} x^l \phi(x)\,\mathrm{d}x \\
&= \lim_{N\to\infty} A^{-\frac{l+1}{k}} \int_0^{\infty} e^{-\frac{1}{2}y^k} y^l\, \mathbb{1}_{[0,\delta(NA)^{1/k}]}(y)\, e^{-\frac{N}{2}r\left(\frac{y}{(NA)^{1/k}}\right)} \phi\left(\frac{y}{(NA)^{1/k}}\right)\mathrm{d}y \\
&= \phi(0)\left(\frac{k!}{f^{(k)}(0)}\right)^{\frac{l+1}{k}} \int_0^{\infty} e^{-\frac{y^k}{2}} y^l\,\mathrm{d}y\,.
\end{aligned}$$

Additionally, with a change of variables, it can easily be shown that

$$\int_0^\infty e^{-\frac{y^k}{2}} y^l \,\mathrm{d}y = \frac{2^{\frac{l+1}{k}}}{k} \Gamma\left(\frac{l+1}{k}\right).$$

If we consider the integral from c to 0 and change variables $y = -x$, we see that with the same arguments as above in the case of l even

$$\lim_{N\to\infty} N^{\frac{l+1}{k}} \int_c^0 e^{-\frac{N}{2}F(x)} x^l \phi(x) \,\mathrm{d}x = \phi(a) \left(\frac{k!}{F^{(k)}(a)}\right)^{\frac{l+1}{k}} \frac{2^{\frac{l+1}{k}}}{k} \Gamma\left(\frac{l+1}{k}\right)$$

and for l odd

$$\lim_{N\to\infty} N^{\frac{l+1}{k}} \int_c^0 e^{-\frac{N}{2}F(x)} x^l \phi(x) \,\mathrm{d}x = -\phi(a) \left(\frac{k!}{F^{(k)}(a)}\right)^{\frac{l+1}{k}} \frac{2^{\frac{l+1}{k}}}{k} \Gamma\left(\frac{l+1}{k}\right).$$

Here, it is important to mention that the function F was approximated by a polynomial of even order. Therefore, the transformation does not affect the exponential function. This proves the statement of the theorem as long as we can justify the application of the dominated convergence.

The dominated convergence is applicable because eq. (3.20) and $y \leq \delta(NA)^{\frac{1}{k}}$ show that a dominating and integrable function is given by

$$g(y) = D y^l e^{-\frac{y^k}{4}}.$$

□

Some analytical calculations provide the minima of F_β in different regimes of β.

Lemma 3.14 (see Proposition 5.16 in [Kir15])
Let F_β be the function in eq. (3.11):

$$F_\beta(t) = \frac{1}{\beta}\left(\frac{1}{2}\log\left(\frac{1+t}{1-t}\right)\right)^2 + \log\left(1-t^2\right).$$

(1) If $\beta < 1$, then F_β has a unique minimum at $t = 0$. Furthermore, $F'(0) = 0$ and $F''(0) = 2(\frac{1}{\beta} - 1) > 0$.

(2) If $\beta = 1$, then F_β has a unique minimum at $t = 0$ with $F'(0) = F''(0) = F'''(0) = 0$ and $F^{(iv)}(0) = 4 > 0$.

(3) For $\beta > 0$ the function F_β has a unique minimum in $[0,1)$ at $t_0 > 0$ and a unique minimum in $(-1,0]$ at $-t_0 < 0$. $F'_\beta(t_0) = F'_\beta(-t_0) = 0$ and $F''_\beta(t_0) = F''_\beta(-t_0) > 0$. t_0 is the unique strictly positive solution of $t = \tanh(\beta t)$. F_β has a local maximum at $t = 0$ with $F'_\beta(0) = 0$ and $F''_\beta(0) = 2\left(\frac{1}{\beta} - 1\right) < 0$.

The proof can be found in the appendix, see Lemma A.12.

We introduced the tools to derive the asymptotic behaviour of the correlation between Curie-Weiss spins. Let a_N, b_N be sequences of real numbers. We write $a_N \approx b_N$ if $\frac{a_N}{b_N} \overset{N\to\infty}{\longrightarrow} 1$. With the help of Laplace's method (see Theorem 3.13) and the result about F_β (see Lemma 3.14), we are able to prove the following theorem:

Theorem 3.15 (see Corollary 28 in [HKW15])
Let $\xi = (\xi_1, \ldots, \xi_N)$ be generated by a Curie-Weiss model with $\beta > 0$. Let l be an even integer. Then,

(1) if $\beta < 1$:

$$\mathbb{E}^{CW}[\xi_1 \cdot \ldots \cdot \xi_l] \approx (l-1)!! \left(\frac{\beta}{1-\beta}\right)^{\frac{l}{2}} N^{-\frac{l}{2}}.$$

(2) if $\beta = 1$:

$$\mathbb{E}^{CW}\left[\xi_1 \cdot \ldots \cdot \xi_l\right] \approx c_l N^{-\frac{l}{4}}.$$

(3) if $\beta > 1$:

$$\mathbb{E}^{CW}\left[\xi_1 \cdot \ldots \cdot \xi_l\right] \approx m(\beta)^l,$$

where $t = m(\beta)$ is the strictly positive solution of $\tanh(\beta t) = t$.

Proof. We define

$$Z_N(l) := \int_{-1}^{1} e^{-\frac{N}{2}F_\beta(x)} x^l \frac{1}{1-x^2}\,\mathrm{d}x\,.$$

The partition function $Z_{N,\beta}$ of the Curie-Weiss model equals $Z_N(0)$ and Proposition 3.10 states that the correlation of spins can be expressed through $Z_N(l)$ as

$$\mathbb{E}^{CW}[\xi_1 \cdot \ldots \cdot \xi_l] = \frac{Z_N(l)}{Z_N(0)}. \tag{3.21}$$

We apply Laplace's method to $Z_N(l)$ and use the insights about F_β from Lemma 3.14. For $\beta \leq 1$, $a = 0$ and $\phi(x) = \frac{1}{1-x^2}$ the requirements of Theorem 3.13 are fulfilled because

$$\int_{-1}^{1} \exp\left(-F_\beta(x)\right) \frac{1}{1-x^2}\,\mathrm{d}x = \int_{-1}^{1} \exp\left(-\frac{1}{\beta}\tanh^{-1}(x)^2\right) \frac{1}{(1-x^2)^2}\,\mathrm{d}x < \infty.$$

If $\beta < 1$, we set $k = 2$ and with Theorem 3.13 and $F''(0) = 2\frac{1-\beta}{\beta}$, we conclude

$$\begin{aligned} N^{\frac{l+1}{2}} Z_N(l) &\approx 2\,\phi(0) \left(\frac{2!}{F''(0)}\right)^{\frac{l+1}{2}} \frac{2^{\frac{l+1}{2}}}{2}\,\Gamma\left(\frac{l+2}{2}\right) \\ &= \left(\frac{\beta}{1-\beta}\right)^{\frac{l+1}{2}} \Gamma\left(\frac{l+1}{2}\right) 2^{\frac{l+1}{2}}. \end{aligned}$$

This leads to

$$Z_N(l) \approx \left(\frac{\beta}{1-\beta}\right)^{\frac{l+1}{2}} \Gamma\left(\frac{l+1}{2}\right) 2^{\frac{l+1}{2}} N^{-\frac{l+1}{2}}.$$

Together with eq. (3.21) we showed that

$$\mathbb{E}^{CW}[\xi_1 \cdot \ldots \cdot \xi_l] = \frac{Z_N(l)}{Z_N(0)} \approx \left(\frac{\beta}{1-\beta}\right)^{\frac{l}{2}} 2^{\frac{l}{2}} \frac{\Gamma\left(\frac{l+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)} N^{-\frac{l}{2}}.$$

With integration by parts it can be verified that $\Gamma(x+1) = x\Gamma(x)$ and by induction it follows that for l even

$$2^{\frac{l}{2}} \frac{\Gamma\left(\frac{l+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)} = (l-1)!!.$$

This proves the first part.

Now assume that $\beta = 1$ and set $k = 4$ then with $F_\beta^{(iv)}(0) = 4$ Laplace's method states

$$\begin{aligned} Z_N(l) &\approx 2\left(\frac{4!}{F_\beta^{(iv)}(0)}\right)^{\frac{l+1}{4}} \frac{2^{\frac{l+1}{4}}}{4}\Gamma\left(\frac{l+1}{4}\right) N^{-\frac{l+1}{4}} \\ &= \frac{1}{2}\, 12^{\frac{l+1}{4}}\, \Gamma\left(\frac{l+1}{4}\right) N^{-\frac{l+1}{4}} \end{aligned}$$

and in the same way as before

$$\mathbb{E}^{CW}[\xi_1 \cdot \ldots \cdot \xi_l] = \frac{Z_N(l)}{Z_N(0)} = c_l N^{-\frac{l}{4}}$$

with $c_l = 12^{\frac{l}{4}} \Gamma\left(\frac{l+1}{4}\right) \Gamma\left(\frac{1}{4}\right)^{-1}$.

For $\beta > 1$ Lemma 3.14 states that F_β has its minima at t_0 and $-t_0$. The binomial identity leads to

$$\begin{aligned} \int_0^1 e^{-\frac{N}{2}F_\beta(x)} x^l \phi(x)\,\mathrm{d}x &= \int_0^1 e^{-\frac{N}{2}F_\beta(x)} (x - t_0 + t_0)^l \phi(x)\,\mathrm{d}x \\ &= \sum_{j=0}^{l} \binom{l}{j} t_0^{l-j} \int_0^1 e^{-\frac{N}{2}F_\beta(x)} (x - t_0)^j \phi(x)\,\mathrm{d}x \end{aligned} \tag{3.22}$$

and

$$\int_{-1}^0 e^{-\frac{N}{2}F_\beta(x)} x^l \phi(x)\,\mathrm{d}x = \sum_{j=0}^{l} \binom{l}{j} (-t_0)^{l-j} \int_{-1}^0 e^{-\frac{N}{2}F_\beta(x)} (x - (-t_0))^j \phi(x)\,\mathrm{d}x\,. \tag{3.23}$$

Because of the symmetry of F_β and ϕ, a change of variables shows that

$$\int_{-1}^0 e^{-\frac{N}{2}F_\beta(x)} (x + t_0)^j \phi(x)\,\mathrm{d}x = (-1)^j \int_0^1 e^{-\frac{N}{2}F_\beta(x)} (x - t_0)^j \phi(x)\,\mathrm{d}x\,. \tag{3.24}$$

Since l is even, the identities in eqs. (3.22) and (3.23) together with eq. (3.24) prove that

$$Z_N(l) = \int_{-1}^1 e^{-\frac{N}{2}F_\beta(x)} x^l \phi(x)\,\mathrm{d}x = 2\sum_{j=0}^{l} \binom{l}{j} t_0^{l-j} \int_0^1 e^{-\frac{N}{2}F_\beta(x)} (x - t_0)^j \phi(x)\,\mathrm{d}x\,.$$

In this form we apply Laplace's method (see Theorem 3.13) with $k = 2$ and $a = t_0$. The requirements are fulfilled because

$$\int_0^1 \exp\Big(-\frac{1}{\beta}\tanh^{-1}(t)^2\Big)\frac{|x - t_0|}{(1 - x^2)^2}\,\mathrm{d}x < \infty.$$

Thus, the integrals where j is an odd number converge to zero even if multiplied by $N^{\frac{j+1}{2}}$. If j is even, the integrals multiplied by $N^{\frac{j+1}{2}}$ converge to different constants. In the limit

$N \to \infty$ the integral with $j = 0$ is the leading term and we conclude that

$$Z_N(l) \approx 2t_0^l \int_0^1 e^{-\frac{N}{2}F_\beta(x)}\phi(x)\,\mathrm{d}x\,. \tag{3.25}$$

Equation (3.25) in combination with Equation (3.21) proves the statement

$$\mathbb{E}^{CW}[\xi_1 \cdot \ldots \cdot \xi_l] = \frac{Z_N(l)}{Z_N(0)} \approx t_0^l.$$

□

3.2.4. Central Limit Theorem for the overlap of spins

From Theorem 3.15 we learn that the correlation vanishes if $\beta \leq 1$ and N is going to infinity. This is an important insight and possibly explains why the standard Hopfield model is able to store Curie-Weiss patterns for $\beta \leq 1$. In a large network the correlation of spins is not strong enough to fundamentally change the functionality of the network. Furthermore, we see that higher order correlations vanish even faster than lower order correlations. We want to use Theorem 3.15 to prove a Central Limit Theorem for the overlap of two independent copies of Curies-Weiss patterns.

Theorem 3.16 (Central Limit Theorem for the overlap of Curie-Weiss patterns)
Let ξ^1, ξ^2 be two independent copies of Curie-Weiss patterns for $0 < \beta < 1$. The scaled overlap $m^1(\xi^2)$ obeys a Central Limit Theorem that means

$$\frac{1}{\sqrt{N}} m^1(\xi^2) = \frac{1}{\sqrt{N}} \sum_{j=1}^N \xi_j^1 \xi_j^2 \Rightarrow \mathcal{N}(0,1)$$

Proof. We want to use the method of moments (see Theorem 30.2 in [Bil95]). For this, we need to verify that all moments of the scaled overlap converge to the moments of a standard normal distribution. Let $k \in \mathbb{N}$ be an odd number. The moment of the overlap can be written as

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{\sqrt{N}}\sum_{j=1}^N \xi_j^1\xi_j^2\right)^k\right] &= N^{-\frac{k}{2}} \sum_{i_1,\ldots,i_k} \mathbb{E}[\xi_{i_1}^1 \cdot \ldots \cdot \xi_{i_k}^1 \cdot \xi_{i_1}^2 \cdot \ldots \cdot \xi_{i_k}^2] \\
&= N^{-\frac{k}{2}} \sum_{i_1,\ldots,i_k} \left(\mathbb{E}[\xi_{i_1}^1 \cdot \ldots \cdot \xi_{i_k}^1]\right)^2.
\end{aligned}$$

Here, we sum over all $i_1, \dots, i_k \in \{1, \dots, N\}$. It is possible that the same index appears multiple times. Because of $\xi_j^\mu \in \{-1, 1\}$, every time a particular spin appears twice or an even number of times it is equal to one. Nevertheless, there always remain an odd number of spins because k is odd. For every combination of indices $i_1, \dots, i_k \in \{1, \dots, N\}$ there exists an odd integer $l(i_1, \dots, i_k)$ such that

$$\left(\mathbb{E}[\xi_{i_1}^1 \cdot \ldots \cdot \xi_{i_k}^1]\right)^2 = \left(\mathbb{E}[\xi_1^1 \cdot \ldots \cdot \xi_l^1]\right)^2 = 0.$$

The last equality follows with Theorem 3.15 because l is odd. With this we get

$$\mathbb{E}\left[\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\xi_j^1\xi_j^2\right)^k\right] = N^{-\frac{k}{2}} \sum_{i_1,\dots,i_k} \left(\mathbb{E}[\xi_{i_1}^1 \cdot \ldots \cdot \xi_{i_k}^1]\right)^2 = 0$$

and showed that all odd moments are equal to zero.

It remains to show that for $k \in \mathbb{N}$ the moment

$$\mathbb{E}\left[\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\xi_j^1\xi_j^2\right)^{2k}\right] = N^{-k} \sum_{i_1,\dots,i_{2k}} \left(\mathbb{E}[\xi_{i_1}^1 \cdot \ldots \cdot \xi_{i_{2k}}^1]\right)^2$$

converges to $(2k-1)!!$ if N tends to infinity.

There are three aspects which determine the behaviour of the moment for N going to infinity:

(1) The first one is the normalizing factor N^{-k}.

(2) The second part refers to the expectation values of the form $\mathbb{E}[\xi_{i_1}^1 \cdot \ldots \cdot \xi_{i_{2k}}^1]$. Here, the tricky part is that the order of $\mathbb{E}[\xi_{i_1}^1 \cdot \ldots \cdot \xi_{i_{2k}}^1]$ depends on the amount of distinct indices in $i_1, \dots, i_{2k}$ which appear odd number of times. Again, spins which occur in pairs and therefore have an even exponent, get cancelled out. In this case with $2k$ indices there always remains an even number (including zero) of spins. In contrast to the odd moments, the order of $\mathbb{E}[\xi_1^1 \cdot \ldots \cdot \xi_l^1]$ (with l even) now changes with β as seen in Theorem 3.15.

(3) For the last one we have to consider the number of combinations to draw indices $i_1, \dots, i_{2k}$ out of $\{1, \dots, N\}$ with the constraint that a specific number of distinct indices have an odd exponent.

To analyse $\sum \left(\mathbb{E}[\xi_{i_1}^1 \cdot \ldots \cdot \xi_{i_{2k}}^1]\right)^2$, we split the sum in parts where the number of distinct indices occurring in $i_1, \dots, i_{2k}$ is fixed. Because of the identical distribution of spin vectors with the same length, we can write:

$$\sum_{i_1,\dots,i_{2k}} \left(\mathbb{E}[\xi^1_{i_1} \cdot \dots \cdot \xi^1_{i_{2k}}]\right)^2 = \sum_{m=1}^{2k} \sum_{\substack{i_1,\dots,i_{2k} \in \{1,\dots,N\} \\ |\{i_1,\dots,i_{2k}\}| = m}} \left(\mathbb{E}[\xi^1_{i_1} \cdot \dots \cdot \xi^1_{i_{2k}}]\right)^2$$
$$= \sum_{m=1}^{2k} \binom{N}{m} \sum_{\substack{i_1,\dots,i_{2k} \in \{1,\dots,m\} \\ |\{i_1,\dots,i_{2k}\}| = m}} \left(\mathbb{E}[\xi^1_{i_1} \cdot \dots \cdot \xi^1_{i_{2k}}]\right)^2.$$

The factor $\binom{N}{m}$ can be split into $\frac{N!}{(N-m)!}$, which has the order N^m and is competing with the factor N^{-k}, and $\frac{1}{m!}$, which will be important to get a convergence to $(2k-1)!!$. An even moment has the form

$$\mathbb{E}\left[\left(\frac{1}{\sqrt{N}} \sum_{j=1}^{N} \xi^1_j \xi^2_j\right)^{2k}\right] = N^{-k} \sum_{m=1}^{2k} \frac{N!}{(N-m)!} \frac{1}{m!} \sum_{\substack{i_1,\dots,i_{2k} \in \{1,\dots,m\} \\ |\{i_1,\dots,i_{2k}\}| = m}} \left(\mathbb{E}[\xi^1_{i_1} \cdot \dots \cdot \xi^1_{i_{2k}}]\right)^2$$
$$= \sum_{m=1}^{2k} N^{m-k} \left(1 + o(1)\right) \frac{1}{m!} \sum_{\substack{i_1,\dots,i_{2k} \in \{1,\dots,m\} \\ |\{i_1,\dots,i_{2k}\}| = m}} \left(\mathbb{E}[\xi^1_{i_1} \cdot \dots \cdot \xi^1_{i_{2k}}]\right)^2.$$

There are three regimes, 1.) $m < k$, 2.) $m > k$ and 3.) $m = k$, with different sorts of behaviour for N going to infinity:

(1) **The number of distinct indices m in $i_1, \dots, i_{2k}$ is less than k:**

In the easiest case m is less than k and with this N^{m-k} goes to zero. The expected value is bounded by one. Therefore, the inner sum can be bounded by a constant $C(k,m)$ counting the possible permutations to draw indices $i_1, \dots, i_{2k}$ from $\{1, \dots m\}$ where all m indices occur at least once. Thus,

$$0 \leq \sum_{m=1}^{k-1} N^{m-k} (1 + o(1)) \frac{1}{m!} \sum_{\substack{i_1,\dots,i_{2k} \in \{1,\dots,m\} \\ |\{i_1,\dots,i_{2k}\}| = m}} \left(\mathbb{E}[\xi^1_{i_1} \cdot \dots \cdot \xi^1_{i_{2k}}]\right)^2$$
$$\leq N^{m-k} \, (k-1) \frac{C(k,m)}{m!} (1 + o(1)) \longrightarrow 0$$

These arguments are valid for all $\beta > 0$.

(2) **The number of distinct indices m appearing in $i_1, \dots, i_{2k}$ is greater than k:**

The problem in this case is, that the number of combinations to choose m indices out of $\{1, \dots, N\}$ is of order N^m beating the normalizing factor N^{-k}. On the other side, with the constraint that all of these m indices need to appear, there are less indices left to cancel out spins in the expectation value. This leads the expectation values to decrease faster to zero while N is going to infinity.

In the inner sum we draw exactly $2k$ times and need to have m distinct indices. Thus, at most $2k - m \geq 0$ indices can be drawn at least twice. Therefore, at least $2k - 2(2k - m) = 2(m - k) \geq 2 > 0$ ($2k$ at all and $2k - m$ twice) indices have an odd exponent. From Theorem 3.15 it follows that $\mathbb{E}[\xi_{i_1}^1 \cdot \ldots \cdot \xi_{i_{2k}}^1]^2$ is at most of order $\mathcal{O}(N^{-2(m-k)})$ (meaning that it could have an order $\mathcal{O}(N^{-l})$ with $l > 2(m-k)$) because at least $2(m - k)$ spins remain with an odd exponent. So as long as $\beta < 1$, we get

$$\begin{aligned}
0 \;\leq\; & \sum_{m=k+1}^{2k} N^{m-k}(1+o(1))\frac{1}{m!} \sum_{\substack{i_1,\dots,i_{2k}\in\{1,\dots,m\} \\ |\{i_1,\dots,i_{2k}\}| = m}} \Big(\mathbb{E}[\xi_{i_1}^1 \cdot \ldots \cdot \xi_{i_{2k}}^1]\Big)^2 \\
\leq\; & \sum_{m=k+1}^{2k} N^{m-k}\, \mathcal{O}(N^{-2(m-k)})\, \frac{C(k,m)}{m!}(1+o(1)) \\
=\; & \sum_{m=k+1}^{2k} \mathcal{O}(N^{-(m-k)})\, \frac{C(k,m)}{m!}(1+o(1)) \longrightarrow 0.
\end{aligned}$$

Therefore, this part of the sum is of order $\mathcal{O}(N^{-(m-k)}) = o(1)$ with $m - k > 0$.

(3) **The number of distinct indices m in $i_1, \dots, i_{2k}$ is equal to k:**

For $m = k$ the normalizing factor and the number of combinations to draw m distinct indices cancel each other out. The expectation values determine the behaviour for large N. In all combinations where at least one spin remains (and because $2k$ is even at least two spins remain) with an odd exponent, the expectation value is of order $\mathcal{O}(N^{-2}) = o(N^{-1})$ for $\beta < 1$ (and $\mathcal{O}(N^{-\frac{1}{2}}) = o(1)$ for $\beta = 1$). The number of permutations for indices $i_1, \dots, i_{2k}$ out of $\{1, \dots, m\}$ with at least one index with an odd exponent is independent of N. Thus, for $\beta \leq 1$

$$\begin{aligned}
& (1+o(1))\,\frac{1}{m!} \sum_{\substack{i_1,\dots,i_{2k}\in\{1,\dots,m\} \\ |\{i_1,\dots,i_{2k}\}| = m}} \Big(\mathbb{E}[\xi_{i_1}^1 \cdot \ldots \cdot \xi_{i_{2k}}^1]\Big)^2 \\
=\; & o(1) + (1+o(1))\frac{1}{m!} \sum_{\substack{i_1,\dots,i_{k}\in\{1,\dots,m\} \\ |\{i_1,\dots,i_{k}\}| = m}} \Big(\mathbb{E}\Big[\big(\xi_{i_1}^1\big)^2 \cdot \ldots \cdot \big(\xi_{i_k}^1\big)^2\Big]\Big)^2 \\
=\; & o(1) + (1+o(1))\,\frac{1}{m!}\, C(m,m).
\end{aligned}$$

The summand with $m = k$ converges to a constant which is determined by the number of choices to draw $2k$ indices out of $\{1, \dots, m\}$ such that each of k distinct indices appear exactly twice. To quantify $C(m, m)$, we change the perspective and draw k times the two spots out of $\{1, \dots, 2k\}$ for each index. Therefore,

$$C(k,k) = \binom{2k}{2}\binom{2k-2}{2}\cdots\binom{2}{2} = \frac{(2k)!}{2^k}.$$

With this, we showed that the summand for $m = k$ is equal to

$$\begin{aligned} o(1) + (1 + o(1))\,\frac{1}{m!}\,C(k,k) &= \frac{(2k)!}{2^k k!}(1 + o(1)) \\ &= (2k-1)!!(1 + o(1)). \end{aligned}$$

Altogether, we showed that

$$\mathbb{E}\left[\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\xi_j^1\xi_j^2\right)^{2k}\right] = (2k-1)!!\;(1 + o(1))$$

for $\beta < 1$ and for all $k \in \mathbb{N}$. With the convergence of all moments to the moments of a standard normal distributed random variable, we proved a Central Limit Theorem for the overlap of two independent Curie-Weiss patterns if $\beta < 1$. □

3.3. Main Result for Curie-Weiss patterns and the polynomial dynamics

Let $\xi^1, \dots, \xi^M$ be independent copies of ξ, which is generated according to a Curie-Weiss model for $\beta < 1$. An important aspect of the proof in Theorem 2.1 was the large deviation principle of the overlap. Since we proved a Central Limit Theorem (see Theorem 3.16), we expect that the overlap of Curie-Weiss patterns in a large network behaves in a similar way as in the independent case. Therefore, we hope that an exponential bound for a tail event can be found in this setting and helps to prove a lower bound for the storage capacity with Curie-Weiss patterns. The next theorem states such an exponential bound for a tail event of an overlap.

Theorem 3.17
Let ξ^1 be a Curie-Weiss pattern for $\beta < 1$. Then for $\gamma > 0$

$$\mathbb{P}_\beta^{CW}\left(\sqrt{N}m^1(\tau) > \gamma\right) \le \frac{1}{Z_{N,\beta}\sqrt{1-\beta}}\exp\left(-\frac{(1-\beta)\gamma^2}{2}\right).$$

for every $\tau \in \{-1,1\}^N$.

Proof. First, we derive an upper bound for the moment generating function

$$\mathbb{E}^{CW}\left[\exp\left(t\sqrt{N}m^1(\tau)\right)\right] = \mathbb{E}^{CW}\left[\exp\left(\frac{t}{\sqrt{N}}\sum_{j=1}^{N}\xi_j^1\tau_j\right)\right],$$

where $t > 0$ and $\tau = (\tau_j)_{j\le N} \in \{-1,1\}^N$ is a fixed configuration. Denote by k the amount of positive spins in τ. Without loss of generality we set $\tau_j = 1$ for $j \le k$ and $\tau_j = -1$ for $k < j \le N$. For a configuration $\sigma = (\sigma_j)_{j\le N} \in \{-1,1\}^N$ define

$$y_1(\sigma) := \frac{1}{\sqrt{N}}\sum_{j=1}^{k}\sigma_j \quad \text{and} \quad y_2(\sigma) := \frac{1}{\sqrt{N}}\sum_{j=k}^{N}\sigma_j.$$

as well as $y(\sigma) := y_1(\sigma) + y_2(\sigma)$. The probability of ξ^1 to be equal to a specific spin configuration σ in the Curie-Weiss model is given by

$$\mathbb{P}_\beta^{CW}\left(\xi^1 = \sigma\right) = \frac{2^{-N}}{Z_{N,\beta}}\exp\left(\frac{\beta}{2}\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\sigma_j\right)^2\right) = \frac{2^{-N}}{Z_{N,\beta}}\exp\left(\frac{\beta}{2}y(\sigma)^2\right).$$

(see eq. (3.5)). With this we get

$$\begin{aligned}
&\mathbb{E}^{CW}\left[\exp\left(\frac{t}{\sqrt{N}}\sum_{j=1}^{N}\xi_j^2\tau_j\right)\right]\\
=\;&\mathbb{E}^{CW}\left[\exp\left(\frac{t}{\sqrt{N}}\left(\sum_{j=1}^{k}\xi_j^2 - \sum_{j=k+1}^{N}\xi_j^2\right)\right)\right]\\
=\;&\sum_{\sigma\in\{-1,1\}^N}\exp\left(t\big(y_1(\sigma) - y_2(\sigma)\big)\right)\frac{2^{-N}}{Z_{N,\beta}}\exp\left(\frac{\beta}{2}y(\sigma)^2\right).
\end{aligned}$$

The Hubbard-Stratonovich transformation (see Theorem 3.9) for $a = \frac{1}{2}$ and $b = \sqrt{\beta}y$ justifies the equation

$$\begin{aligned}
&\mathbb{E}^{CW}\left[\exp\left(\frac{t}{\sqrt{N}}\sum_{j=1}^{N}\xi_j^1\tau_j\right)\right]\\
=\;&\frac{2^{-N}}{Z_{N,\beta}}\sum_{\sigma\in\{-1,1\}^N}\exp\left(t\big(y_1(\sigma) - y_2(\sigma)\big)\right)\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}s^2 + y(\sigma)s\sqrt{\beta}\right)\,\mathrm{d}s\,.
\end{aligned}$$

We sum over a finite number of indices and the linearity of the integral leads to

$$
\begin{aligned}
&\frac{1}{\sqrt{2\pi}}\frac{2^{-N}}{Z_{N,\beta}}\int\limits_{-\infty}^{\infty}\sum_{\sigma\in\{-1,1\}^N}\exp\Big(t\big(y_1(\sigma)-y_2(\sigma)\big)\Big)\exp\left(-\frac{1}{2}s^2+y(\sigma)s\sqrt{\beta}\right)\,\mathrm{d}s\\
=\;&\frac{1}{\sqrt{2\pi}}\frac{2^{-N}}{Z_{N,\beta}}\int\limits_{-\infty}^{\infty}\sum_{\sigma\in\{-1,1\}^N}\exp\Big((t+s\sqrt{\beta})y_1(\sigma)\Big)\exp\Big((-t+s\sqrt{\beta})y_2(\sigma)\Big)\,e^{-\frac{s^2}{2}}\,\mathrm{d}s\\
=\;&\frac{1}{\sqrt{2\pi}}\frac{2^{-N}}{Z_{N,\beta}}\int\limits_{-\infty}^{\infty}\sum_{(\sigma_j)_{j\le k}}\exp\Big((t+s\sqrt{\beta})y_1(\sigma)\Big)\sum_{(\sigma_j)_{j>k}}\exp\Big((-t+s\sqrt{\beta})y_2(\sigma)\Big)e^{-\frac{s^2}{2}}\,\mathrm{d}s\,.
\end{aligned}
$$

For every spin we sum over the two possible states -1 and 1 therefore the exponential becomes a disturbed version of the cosh-term, which appeared in the case of independent patterns (see proof of Theorem 2.1) before:

$$
\sum_{\sigma_1\in\{-1,1\}}\exp\left((t+s\sqrt{\beta})\frac{\sigma_j}{\sqrt{N}}\right)=2\cosh\left(\frac{(t+s\sqrt{\beta})}{\sqrt{N}}\right).
$$

This leads to

$$
\begin{aligned}
&\frac{1}{\sqrt{2\pi}}\frac{2^{-N}}{Z_{N,\beta}}\int\limits_{-\infty}^{\infty}\left(2\cosh\left(\frac{t+s\sqrt{\beta}}{\sqrt{N}}\right)\right)^k\left(2\cosh\left(\frac{-t+s\sqrt{\beta}}{\sqrt{N}}\right)\right)^{N-k}e^{-\frac{s^2}{2}}\,\mathrm{d}s\\
=\;&\frac{1}{\sqrt{2\pi}}\frac{1}{Z_{N,\beta}}\int\limits_{-\infty}^{\infty}\cosh\left(\frac{t+s\sqrt{\beta}}{\sqrt{N}}\right)^k\cosh\left(\frac{-t+s\sqrt{\beta}}{\sqrt{N}}\right)^{N-k}e^{-\frac{s^2}{2}}\,\mathrm{d}s\,.
\end{aligned}
$$

For every $t\in\mathbb{R}$ we can bound $\cosh(t)\le\exp\left(\frac{t^2}{2}\right)$ and conclude that the moment generating function is bounded by

$$
\begin{aligned}
&\mathbb{E}^{CW}\left[\exp\left(\frac{t}{\sqrt{N}}\sum_{j=1}^N\xi_j^1\tau_j\right)\right]\\
\le\;&\frac{1}{\sqrt{2\pi}}\frac{1}{Z_{N,\beta}}\int\limits_{-\infty}^{\infty}\exp\left(\frac{k}{2N}\left(t+s\sqrt{\beta}\right)^2+\frac{N-k}{2N}\left(-t+s\sqrt{\beta}\right)^2\right)e^{-\frac{s^2}{2}}\,\mathrm{d}s\,.
\end{aligned}
$$

If we expand the quadratic terms, we get

$$
\frac{1}{\sqrt{2\pi}}\frac{1}{Z_{N,\beta}}\int\limits_{-\infty}^{\infty}\exp\left(\frac{k}{2N}\left(t^2+2ts\sqrt{\beta}+s^2\beta\right)+\frac{N-k}{2N}\left(t^2-2ts\sqrt{\beta}+s^2\beta\right)\right)e^{-\frac{s^2}{2}}\,\mathrm{d}s
$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{Z_{N,\beta}} \int_{-\infty}^{\infty} \exp\left(\frac{1}{2}\left(t^2 + s^2\beta\right) - \frac{N-2k}{N} ts\sqrt{\beta} - \frac{1}{2}s^2 \right) \mathrm{d}s\,.$$

In the next step we use the Hubbard-Stratonovich transformation (see Theorem 3.9 with $a = \frac{1-\beta}{2}$ and $b = -\frac{N-2k}{N}t\sqrt{\beta}$) and form the integral back to an exponential term:

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}} \frac{1}{Z_{N,\beta}} \exp\left(\frac{t^2}{2}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}s^2(1-\beta) - \frac{N-2k}{N} ts\sqrt{\beta} \right) \mathrm{d}s \\
=\ & \frac{1}{\sqrt{2\pi}} \frac{1}{Z_{N,\beta}} \exp\left(\frac{t^2}{2}\right) \sqrt{\frac{2\pi}{1-\beta}} \exp\left(\left(\frac{N-2k}{N}\right)^2 \frac{\beta t^2}{2(1-\beta)} \right) \\
=\ & \frac{1}{\sqrt{1-\beta}Z_{N,\beta}} \exp\left(\frac{t^2}{2}\left[1 + \left(\frac{N-2k}{N}\right)^2 \frac{\beta}{1-\beta} \right] \right).
\end{aligned}$$

The number of negative spins in τ was denoted by $k \in \{0, \ldots, N\}$. Thus, $N - 2k \in \{-N, \ldots, N\}$ as well as $\left(\frac{N-2k}{N}\right)^2 \in [0,1]$. An upper bound is therefore

$$\begin{aligned}
&\frac{1}{\sqrt{1-\beta}Z_{N,\beta}} \exp\left(\frac{t^2}{2}\left[1 + \frac{\beta}{1-\beta} \right] \right) \\
=\ & \frac{1}{\sqrt{1-\beta}Z_{N,\beta}} \exp\left(\frac{1}{1-\beta}\frac{t^2}{2} \right)
\end{aligned}$$

because $0 < \beta < 1$. The proof of the theorem is completed after using an exponential Chebyshev inequality. With $c_N = (\sqrt{1-\beta}Z_{N,\beta})^{-1}$ and $t = (1-\beta)\gamma$, we conclude that

$$\begin{aligned}
\mathbb{P}_\beta^{CW}\left(\sqrt{N}m^1(\tau) > \gamma \right) &\leq \exp\left(-\gamma t\right) \mathbb{E}^{CW}\left[\exp\left(\frac{t}{\sqrt{N}} \sum_{j=1}^{N} \xi_j^1 \tau_j \right) \right] \\
&\leq c_N \exp\left(-\gamma t + \frac{1}{1-\beta}\frac{t^2}{2} \right) \\
&= c_N \exp\left(-\frac{(1-\beta)\gamma^2}{2} \right).
\end{aligned}$$

In Theorem 3.2 we proved that $Z_{N,\beta} \to \sigma_{CW} = \sqrt{1-\beta}^{-1}$. Thus, c_N converges to one for $N \to \infty$. □

With the help of Theorem 3.17, we are able to calculate a lower bound for the storage capacity of the Hopfield model with polynomial dynamics $\widetilde{T}$ (see eq. (2.4)) and patterns which are generated according to a Curie-Weiss model with $0 < \beta < 1$ as long as $n - 1$ is chosen to be even.

Theorem 3.18
Let $M = \frac{N^{n-1}}{c_n \log(N)}$ and let $\xi^1, \dots, \xi^M$ be M independent patterns chosen according to Curie-Weiss model with $\beta < 1$. Let $n-1$ be an even integer. The Hopfield model with dynamics $\widetilde{T}$ can store at least M patterns for $c_n > 2(2n-3)!!\sigma_{CW}^{2(n-1)}$ if one wants a fixed pattern to be a fixed point of the dynamics with a probability converging to one.

Moreover fix $\rho \in [0, \frac{1}{2})$. If $c_n > \frac{2(2n-3)!!\sigma_{CW}^{2(n-1)}}{(1-2\rho)^{2(n-1)}}$, then for any $\widetilde{\xi}^\nu$ taken uniformly at random from $\mathcal{S}(\xi^\nu, \rho N)$, where ρN is assumed to be an integer, it follows that

$$\mathbb{P}_\beta^{CW}\left(\widetilde{T}(\widetilde{\xi}^\nu) = \xi^\nu\right) = 1 - R_N,$$

where $R_N \to 0$ for $N \to \infty$.

Furthermore, if $c_n > \frac{2n(2n-3)!!\sigma_{CW}^{2(n-1)}}{(1-2\rho)^{2(n-1)}}$, then

$$\mathbb{P}_\beta^{CW}\left(\forall \mu \leq M:\ \widetilde{T}(\widetilde{\xi}^\mu) = \xi^\mu\right) = 1 - R_N,$$

where $R_N \to 0$ for $N \to \infty$.

Proof. Similarly to the proof of Theorem 2.1, we assume that the network consists of $N+1$ (instead of N) neurons. Without loss of generality we focus on pattern ξ^1 resp. $\widetilde{\xi}^1 \in \mathcal{S}(\xi^1, \rho N)$. Now with the same arguments which we used to derive an upper bound in the case of i.i.d. patterns (see eq. (2.8)), we can conclude that

$$\mathbb{P}_\beta^{CW}\left(T_i(\widetilde{\xi}^1) \neq \xi_i^1\right) \leq e^{-t(1-2\rho)^{n-1}N^{n-1}u(N)}\, \mathbb{E}^{CW}\left[\exp\left(-t\sum_{\mu=2}^{M} \xi_i^1 \xi_i^\mu N^{n-1} m_i^\mu(\widetilde{\xi}^1)^{n-1}\right)\right], \tag{3.26}$$

where $u(N) \overset{N\to\infty}{\longrightarrow} 1$. In contrast to the proof of Theorem 2.1, the spins in one pattern are not independent. Thus, we cannot use a statement like Lemma 1.2. But for the moment generating function conditioned on ξ^1 and $\widetilde{\xi}^1$, we can use the independence of patterns $\xi^2, \dots, \xi^M$:

$$\begin{aligned} &\mathbb{E}^{CW}\left[\exp\left(-t\sum_{\mu=2}^{M} \xi_i^1 \xi_i^\mu N^{n-1} m_i^\mu(\widetilde{\xi}^1)^{n-1}\right)\right] \\ =\ &\mathbb{E}^{CW}\left[\mathbb{E}^{CW}\left[\exp\left(-t\sum_{\mu=2}^{M} \xi_i^1 \xi_i^\mu N^{n-1} m_i^\mu(\widetilde{\xi}^1)^{n-1}\right) \,\middle|\, \xi^1, \widetilde{\xi}^1\right]\right] \\ =\ &\mathbb{E}^{CW}\left[\mathbb{E}^{CW}\left[\exp\left(-t\xi_i^1 \xi_i^2 N^{n-1} m_i^2(\widetilde{\xi}^1)^{n-1}\right) \,\middle|\, \xi^1, \widetilde{\xi}^1\right]^{M-1}\right]. \end{aligned}$$

The strategy is to show that the conditional moment generating function can still be bounded similar to the i.i.d. case. By Theorem 3.17 the probability of large values for the overlap decays exponentially fast for every realization of ξ^1 resp. $\widetilde{\xi}^1$. Furthermore, higher order terms of the Taylor expansion can be bounded with the help of a truncation argument and for the critical values we use that the sum of spins obey a Central Limit Theorem (see Theorem 3.2).

We want to work with an arbitrary realization of $\widetilde{\xi}^1$ and separate large values of the overlap through the random set

$$A = A_{\widetilde{\xi}^1} = \left\{ (\sigma_j)_{j \leq N} \,:\, \Big| \frac{1}{\sqrt{N}} \sum_{j \neq i} \widetilde{\xi}^1_j \sigma_j \Big| \leq \gamma \right\}$$

for $\gamma > 0$. The moment generation function conditioned on ξ^1 and $\widetilde{\xi}^1$ can be split into parts where $\{\xi^2 \in A\}$ respectively $\{\xi^2 \in A^c\}$ holds:

$$\begin{aligned} &\mathbb{E}^{CW}\left[\exp\left(-t\xi_i^1\xi_i^\mu N^{n-1} m_i^2(\widetilde{\xi}^1)^{n-1}\right) \,\Big|\, \xi^1, \widetilde{\xi}^1\right] \\ =\ &\mathbb{E}^{CW}\left[\exp\left(-t\xi_i^1\xi_i^\mu N^{n-1} m_i^2(\widetilde{\xi}^1)^{n-1}\right)\left(\mathbb{1}_{\{\xi^2\in A^c\}} + \mathbb{1}_{\{\xi^2\in A\}}\right) \,\Big|\, \xi^1, \widetilde{\xi}^1\right]. \end{aligned} \tag{3.27}$$

In the first part of eq. (3.27) we bound the exponential function by its maximal value:

$$\begin{aligned} &\mathbb{E}^{CW}\left[\exp\left(-t\xi_i^1\xi_i^2 N^{n-1} m_i^2(\widetilde{\xi}^1)^{n-1}\right) \mathbb{1}_{\{\xi^2\in A^c\}} \,\Big|\, \xi^1, \widetilde{\xi}^1\right] \\ \leq\ &\exp\left(tN^{n-1}\right) \mathbb{P}_\beta^{CW}\left(\xi^2 \in A^c \,\Big|\, \xi^1, \widetilde{\xi}^1\right). \end{aligned} \tag{3.28}$$

The conditional probability can be handled with the help of Theorem 3.17, which provides a large deviation estimate for the overlap with an arbitrary deterministic configuration. Whether the summation includes i or not makes no difference. Thus, applying Theorem 3.17 leads to

$$\begin{aligned} &\mathbb{P}_\beta^{CW}\left(\xi^2 \in A^c \,\Big|\, \widetilde{\xi}^1\right) \\ =\ &\mathbb{P}_\beta^{CW}\left(\frac{1}{\sqrt{N}}\sum_{j\neq i} \widetilde{\xi}_j^1 \xi_j^2 > \gamma \,\Big|\, \widetilde{\xi}^1\right) + \mathbb{P}_\beta^{CW}\left(\frac{1}{\sqrt{N}}\sum_{j\neq i} \left(-\widetilde{\xi}_j^1\right)\xi_j^2 > \gamma \,\Big|\, \widetilde{\xi}^1\right) \\ \leq\ &2\widetilde{c}_N \exp\left(-\frac{(1-\beta)\gamma^2}{2}\right), \end{aligned} \tag{3.29}$$

where $\widetilde{c}_N = (Z_{N,\beta}\sqrt{1-\beta})^{-1}$. The estimates in eqs. (3.28) and (3.29) show that

$$\mathbb{E}^{CW}\left[\exp\left(-t\xi_i^1\xi_i^2 N^{n-1} m_i^2(\widetilde{\xi}^1)^{n-1}\right) \mathbb{1}_{\{\xi^2\in A^c\}} \,\Big|\, \xi^1, \widetilde{\xi}^1\right]$$

$$\leq\ 2\widetilde{c}_N \exp\left(tN^{n-1} - \frac{(1-\beta)\gamma^2}{2}\right). \tag{3.30}$$

Choose $t = \frac{a_n}{M}$ for $a_n > 0$ and $\gamma = \log(N)^\tau$ for a fixed $\tau > \frac{1}{2}$. Then

$$t = \frac{a_n c_n \log(N+1)}{(N+1)^{n-1}} \leq \frac{a_n c_n \log(N)}{N^{n-1}}$$

and the bound in eq. (3.30) is smaller than

$$2\widetilde{c}_N \exp\left(tN^{n-1} - \frac{(1-\beta)\gamma^2}{2}\right) = 2\widetilde{c}_N \exp\left(\log(N)\left[a_n c_n - \frac{1-\beta}{2}\log(N)^{2\tau-1}\right]\right).$$

Because of $0 < \beta < 1$ and $\tau > \frac{1}{2}$ we know that $2\tau - 1 > 0$. Thus, for large N the term in brackets can be bounded from above by a constant less than zero. With this we showed that

$$\begin{aligned}
&\mathbb{E}^{CW}\left[\exp\left(-t\xi_i^1\xi_i^2 N^{n-1} m_i^2(\widetilde{\xi}^1)^{n-1}\right)\mathbb{1}_{\{\xi^2\in A^c\}}\ \Big|\ \xi^1,\widetilde{\xi}^1\right]\\
\leq\ & 2\widetilde{c}_N \exp\left(\left[a_n c_n - \frac{1-\beta}{2}\log(N)^{2\tau-1}\right]\log(N)\right) \qquad (3.31)\\
=\ & 2\widetilde{c}_N \exp\left(h_1(N)\right) = o(1),
\end{aligned}$$

where $h_1(N) = (a_n c_n - \frac{1-\beta}{2}\log(N)^{2\tau-1})\log(N)$.

For the second part of eq. (3.27) we use the Taylor expansion of the exponential function:

$$\begin{aligned}
&\mathbb{E}^{CW}\left[\exp\left(-t\,\xi_i^1\xi_i^2 N^{n-1} m_i^2(\widetilde{\xi}^1)^{n-1}\right)\mathbb{1}_{\{\xi^2\in A\}}\ \Big|\ \xi^1,\widetilde{\xi}^1\right]\\
=\ & \mathbb{E}^{CW}\left[\left(1+\sum_{k=1}^{\infty}\frac{(-t)^k}{k!}\left(\xi_i^1\right)^k\left(\xi_i^2\right)^k N^{k(n-1)} m_i^2(\widetilde{\xi}^1)^{k(n-1)}\right)\mathbb{1}_{\{\xi^2\in A\}}\ \Big|\ \xi^1,\widetilde{\xi}^1\right]\\
\leq\ & 1+\sum_{k=1}^{\infty}\frac{(-t)^k}{k!}\left(\xi_i^1\right)^k\cdot\mathbb{E}^{CW}\left[\left(\xi_i^2\right)^k N^{k(n-1)} m_i^2(\widetilde{\xi}^1)^{k(n-1)}\mathbb{1}_{\{\xi^2\in A\}}\ \Big|\ \xi^1,\widetilde{\xi}^1\right].
\end{aligned}$$

Large outliers of the overlap are eliminated by the indicator function for the event $\{\xi^2 \in A\}$ and this truncation guarantees that

$$|m_i^2(\widetilde{\xi}^1)| \leq N^{-\frac{1}{2}}\gamma = N^{-\frac{1}{2}}\log(N)^\tau.$$

Together with $\xi_i^1, \xi_i^2 \leq 1$ and $\mathbb{P}_\beta^{CW}\left(\xi^2 \in A\ \Big|\ \xi^1, \widetilde{\xi}^1\right) \leq 1$, we can deduce for the higher order terms that

$$\begin{aligned}
&\sum_{k=3}^{\infty} \frac{(-t)^k}{k!} \left(\xi_i^1\right)^k \cdot \mathbb{E}^{CW} \left[\left(\xi_i^2\right)^k N^{k(n-1)} m_i^2(\widetilde{\xi}^1)^{k(n-1)} \mathbb{1}_{\{\xi^2 \in A\}} \;\middle|\; \xi^1, \widetilde{\xi}^1 \right] \\
\leq\ & \sum_{k=3}^{\infty} \frac{t^k}{k!} \cdot N^{\frac{k(n-1)}{2}} \gamma^{k(n-1)} \\
\leq\ & \frac{\left(tN^{\frac{n-1}{2}}\gamma^{n-1}\right)^3}{6} \sum_{k=0}^{\infty} \frac{\left(tN^{\frac{n-1}{2}}\gamma^{n-1}\right)^k}{k!} \\
\leq\ & t^3 N^{3\frac{n-1}{2}} \gamma^{3(n-1)}.
\end{aligned}$$

For N large the series is bounded by e because with $t \leq \frac{a_n c_n \log(N)}{N^{n-1}}$ and $\gamma = \log(N)^\tau$ we have that

$$tN^{\frac{n-1}{2}}\gamma^{n-1} \leq \frac{a_n c_n \log(N)}{N^{n-1}} N^{\frac{n-1}{2}} \log(N)^{\tau(n-1)} = a_n c_n N^{-\frac{n-1}{2}} \log(N)^{\tau(n-1)+1} \to 0$$

as N tends to infinity. Thus, $tN^{\frac{n-1}{2}}\gamma^{n-1} \leq 1$ for N large enough.

For the first order term it is important that $n-1$ is an even number. Corollary 3.11 states that the expectation value of an odd number of spins is equal to zero because of their symmetric distribution. We know that $k(n-1)+1$ is an odd number. Thus, for all k odd, especially for $k=1$, it follows with Corollary 3.11 that

$$\begin{aligned}
&-t \left(\xi_i^1\right)^k \mathbb{E}^{CW} \left[\left(\xi_i^2\right)^k N^{k(n-1)} m_i^2(\widetilde{\xi}^1)^{k(n-1)} \;\middle|\; \xi^1, \widetilde{\xi}^1 \right] \\
=\ & -t \sum_{j_1,\dots,j_{k(n-1)}} \xi_i^1 \widetilde{\xi}_{j_1}^1 \cdot \ldots \cdot \widetilde{\xi}_{j_{k(n-1)}}^1 \cdot \mathbb{E}^{CW} \left[\xi_i^2 \xi_{j_1}^2 \cdot \ldots \cdot \xi_{j_{k(n-1)}}^2 \;\middle|\; \xi^1, \widetilde{\xi}^1 \right] \\
=\ & 0.
\end{aligned}$$

The second order term of the Taylor expansion is bounded by

$$\frac{t^2}{2} \mathbb{E}^{CW} \left[N^{2(n-1)} m_i^2(\widetilde{\xi}^1)^{2(n-1)} \mathbb{1}_{\{\xi^2 \in A\}} \;\middle|\; \xi^1, \widetilde{\xi}^1 \right] \leq \frac{t^2}{2} N^{n-1} \, (2n-3)!! \, \sigma_{CW}^{2(n-1)} (1+\varepsilon),$$

which follows with the fact that for every $\varepsilon > 0$ and N large enough

$$\begin{aligned}
&\mathbb{E}^{CW} \left[N^{(n-1)} m_i^2(\widetilde{\xi}^1)^{2(n-1)} \mathbb{1}_{\{\xi^2 \in A\}} \;\middle|\; \xi^1, \widetilde{\xi}^1 \right] \\
=\ & N^{-(n-1)} \sum_{j_1,\dots,j_{2(n-1)}} \widetilde{\xi}_{j_1}^1 \cdot \ldots \cdot \widetilde{\xi}_{j_{2(n-1)}}^1 \cdot \mathbb{E}^{CW} \left[\xi_{j_1}^2 \cdot \ldots \cdot \xi_{j_{2(n-1)}}^2 \mathbb{1}_{\{\xi^2 \in A\}} \;\middle|\; \xi^1, \widetilde{\xi}^1 \right]
\end{aligned}$$

$$\leq \mathbb{E}^{CW}\left[\left(\frac{1}{\sqrt{N}}\sum_{j\neq i}\xi_j^2\right)^{2(n-1)}\right] \leq \kappa_{2(n-1)}\,\sigma_{CW}^{2(n-1)}(1+\varepsilon).$$

Here, we bounded the spins by one and used Lemma 3.5. Furthermore, $\sigma_{CW}^2 = (1-\beta)^{-1}$ and $\kappa_{2l} = (2l-1)!!$.

Altogether, we showed for the second part of eq. (3.27) that

$$\begin{aligned}
&\mathbb{E}^{CW}\left[\exp\left(-t\,\xi_i^1\xi_i^2 N^{n-1} m_i^2(\widetilde{\xi}^1)^{n-1}\right)\,\mathbb{1}_{\{\xi^2\in A\}}\;\middle|\;\xi^1,\widetilde{\xi}^1\right]\\
\leq\;& 1+\frac{t^2}{2}N^{n-1}\kappa_{2(n-1)}\sigma_{CW}^{2(n-1)}(1+\varepsilon)+\left(tN^{\frac{n-1}{2}}\gamma^{n-1}\right)^3\\
\leq\;& \exp\left(\frac{t^2}{2}N^{n-1}\kappa_{2(n-1)}\sigma_{CW}^{2(n-1)}(1+\varepsilon)+\left(tN^{\frac{n-1}{2}}\gamma^{n-1}\right)^3\right)=\exp\left(h_2(N)\right), \qquad (3.32)
\end{aligned}$$

where $h_2(N) = \frac{t^2}{2}N^{n-1}\kappa_{2(n-1)}\sigma_{CW}^{2(n-1)}(1+\varepsilon)+\left(tN^{\frac{n-1}{2}}\gamma^{n-1}\right)^3$.

We know that

$$M\exp\left(h_1(N)\right) = \frac{N^{n-1}}{c_n\log(N)}\exp\left(\left[a_nc_n-\frac{1-\beta}{2}\log(N)^{2\tau-1}\right]\log(N)\right) = o(1) \qquad (3.33)$$

because with $\tau > \frac{1}{2}$ and for N large enough

$$a_nc_n-\frac{1-\beta}{2}\log(N)^{2\tau-1} \leq -(n-1)$$

is true. Similar to the case of i.i.d. patterns (see eq. (2.16)), the statement in eq. (3.33) is sufficient to conclude that the term in eq. (3.32) determines the order of the upper bound for the moment generating function:

$$\mathbb{E}^{CW}\left[\mathbb{E}^{CW}\left[\exp\left(-t\xi_i^1\xi_i^2N^{n-1}m_i^2(\widetilde{\xi}^1)^{n-1}\right)\;\middle|\;\xi^1,\widetilde{\xi}^1\right]^M\right] \leq \exp\left(Mh_2(N)\right)(1+o(1)).$$

Going back to eq. (3.26), we showed with the previous calculations that

$$\begin{aligned}
&\mathbb{P}_\beta^{CW}\left(T_i(\xi^1)\neq\xi_i^1\right)\\
\leq\;& \exp\left(-t(1-2\rho)^{n-1}N^{n-1}u(N)\right)\cdot\exp\left(Mh_2(N)\right)\cdot(1+o(1))\\
=\;& \exp\Big(-t(1-2\rho)^{n-1}N^{n-1}u(N)+\frac{t^2}{2}MN^{n-1}\kappa_{2(n-1)}\sigma_{CW}^{2(n-1)}(1+\varepsilon)+M\left(tN^{\frac{n-1}{2}}\gamma^{n-1}\right)^3+o(1)\Big).
\end{aligned}$$

Analogously to the proof of Theorem 2.1 (see eq. (2.18)), we know that for every $0<\widetilde{\varepsilon}<1$

we can choose N large enough such that

$$u(N)\frac{\log(N+1)}{\log(N)}\Big(\frac{N}{N+1}\Big)^{n-1} \geq (1-\widetilde{\varepsilon}).$$

Hence, we deduce that for all $\varepsilon, \widetilde{\varepsilon} > 0$ and N large enough

$$\begin{aligned}
&\mathbb{P}_\beta^{CW}\left(T_i(\xi^1) \neq \xi_i^1\right) \\
\leq\ & \exp\left(-a_n c_n (1-2\rho)^{n-1}\log(N)(1-\widetilde{\varepsilon}) + \frac{a_n^2 c_n}{2}(2n-3)!!\sigma_{CW}^{2(n-1)}(1+\varepsilon)\log(N) + o(1)\right) \\
=\ & \exp\left(\left[-a_n c_n\left((1-2\rho)^{n-1}(1-\widetilde{\varepsilon}) - \frac{a_n(2n-3)!!\sigma_{CW}^{2(n-1)}(1+\varepsilon)}{2}\right)\right]\log(N) + o(1)\right).
\end{aligned}$$

We use a similar argumentation as in the independent case. First choose

$$a_n = \frac{(1-2\rho)^{n-1}}{(2n-3)!!\sigma_{CW}^{2(n-1)}}\frac{(1-\widetilde{\varepsilon})}{(1+\varepsilon)}.$$

Then, the term in brackets equals

$$-a_n c_n\left((1-2\rho)^{n-1}(1-\widetilde{\varepsilon}) - \frac{a_n(2n-3)!!\sigma_{CW}^{2(n-1)}(1+\varepsilon)}{2}\right) = -\frac{c_n(1-2\rho)^{2(n-1)}}{2(2n-3)!!\sigma_{CW}^{2(n-1)}}\frac{(1-\widetilde{\varepsilon})^2}{(1+\varepsilon)}.$$

These observations are true for all $\varepsilon, \widetilde{\varepsilon} > 0$ and N large enough. Therefore, we conclude that

$$\mathbb{P}_\beta^{CW}\left(\exists i \leq N:\ T_i(\xi^1) \neq \xi_i^1\right) \leq \exp\left(\log(N)\left[1 - \frac{c_n(1-2\rho)^{2(n-1)}}{2(2n-3)!!\sigma_{CW}^{2(n-1)}}\right] + o(1)\right),$$

which converges to zero due to the assumption that

$$c_n > \frac{2\,(2n-3)!!\,\sigma_{CW}^{2(n-1)}}{(1-2\rho)^{2(n-1)}}.$$

The last statement of the theorem follows with

$$\begin{aligned}
&\mathbb{P}_\beta^{CW}\left(\exists \mu \leq M, \exists i \leq N:\ T_i(\xi^1) \neq \xi_i^1\right) \\
\leq\ & \exp\left(\log(N)\left[1 - \frac{c_n(1-2\rho)^{2(n-1)}}{2(2n-3)!!\sigma_{CW}^{2(n-1)}}\right] + (n-1)\log(N) - \log(c_n\log(N)) + o(1)\right) \\
\leq\ & \exp\left(\log(N)\left[n - \frac{c_n(1-2\rho)^{2(n-1)}}{2(2n-3)!!\sigma_{CW}^{2(n-1)}}\right] + o(1)\right),
\end{aligned}$$

which converges to zero as long as

$$c_n > \frac{2n\,(2n-3)!!\,\sigma_{CW}^{2(n-1)}}{(1-2\rho)^{2(n-1)}}.$$

□

3.4. Conclusions about the polynomial dynamics

Theorem 2.1 and Theorem 3.18 proved that the Hopfield model with a polynomial dynamics $\widetilde{T}$ is able to store at least $\frac{N^{n-1}}{c_n \log(N)}$ patterns. This fits to the results of Newman in [New88], where the storage capacity increased from αN to αN^{l-1} by using higher order overlaps. In contrast to Newman, our results consider a perfect retrieval of patterns. Additionally to the stability of the patterns, the net is able to correct an error rate of $\rho \in [0, \frac{1}{2})$.

Furthermore, with Theorem 3.18 we showed that the network is able to store patterns with correlated spins. A result for the standard Hopfield model ($n = 2$) and a more general setting of correlated patterns were achieved in [LV05]. In contrast to these results, we considered a more general version of the dynamics and proved results for a polynomial interaction function x^n with n odd but only for Curie-Weiss patterns with $\beta < 1$. The assumption that n is an odd integer was an important argument to achieve a symmetric distribution in the first order term of the Taylor expansion. While using higher order overlaps, a distinction between an odd and an even exponent and a different behaviour of the net resp. spin glass in these two cases were also observed in [Tal00b; BN01].

4. Generalized Hopfield model with weights

The standard Hopfield model allows all neurons to communicate with each other and there are no further restrictions on the interaction between them. The underlying structure of the model is a complete graph. Furthermore, the influence between two neurons is symmetric and only depends on the patterns $\xi^1, \dots, \xi^M$ (see eq. (1.1)). In a more general approach one can define the Hopfield model on an arbitrary graph $G = (V, E)$. Here, V is a set containing all neurons and E is a set of undirected edges, which determine whether two neurons interchange signals or not. Results about Hopfield models on random graphs can be found for example in [BG92; BG93; KP93; LV11; LV15]. These models allow neurons to interact with each other only if they are directly connected in the underlying graph. A slightly different idea is used in the hierarchical Hopfield model in which they only have to be connected through a path of any length. In this approach neurons influence each other although they are separated by several edges. Direct neighbours interact in the same way as before, but neurons which are more than one edge apart have a weaker signal strength. For this purpose a new parameter φ is introduced, which determines the decrease of the signal strength at greater distances. Agliari et. al. introduce the hierarchical Hopfield model in their publications [Agl+15a; Agl+15b] through recursively defined Hamiltonians.

4.1. The hierarchical Hopfield model

Set $H_0(\sigma) = 0$ for $\sigma \in \{-1, 1\}$. For $k \geq 1$ and a configuration $\sigma = (\sigma_j)_{1 \leq j \leq 2^{k+1}}$ define the Hamiltonian recursively through

$$H_{k+1}(\sigma) = H_k(\sigma_1) + H_k(\sigma_2) - \frac{1}{2^{2\varphi(k+1)}} \sum_{i<j=1}^{2^{k+1}} \left(\sum_{\mu=1}^{M} \xi_i^\mu \xi_j^\mu \right) \sigma_i \sigma_j, \tag{4.1}$$

where $\sigma_1 = (\sigma_j)_{1 \leq j \leq 2^k}$ and $\sigma_2 = (\sigma_j)_{2^k+1 \leq j \leq 2^{k+1}}$. As mentioned before, $\varphi \in (\frac{1}{2}, 1)$ is the parameter to weaken the influence of signals at greater distances. The patterns $\xi^1, \dots, \xi^M$ are assumed to be independent with independent spins where each spin is chosen uniformly at random from $\{-1, 1\}$.

The basic idea is that in each iteration we combine two copies of the previous network (represented by the configurations σ_1 and σ_2), including their existing connections, and build a new network by linking these two copies together (see fig. 4.1). All paths between the nodes have weights assigned to them. These weights consist of two parts: a term representing the influence through the patterns (Hebb rule, see eq. (1.1)) and a discount factor. The discount factor depends on the iteration at which the path is firstly built. Thus, the signal they share is weaker if they are firstly connected to each other at a later step of the iteration. In fig. 4.1 the black connections are built in the first step, the red connections are built in the second step and the blue connections are built in the third step of the iteration. The form of the discount factor and the similarity of

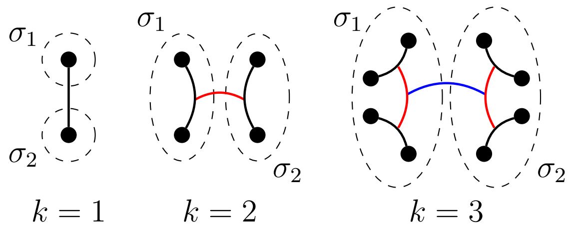

Figure 4.1.: Recursion of the Hamiltonian of the hierarchical Hopfield model (adapted from Figure 1 in [Agl+15b])

the hierarchical Hopfield model to the standard Hopfield model is easier to see if the Hamiltonian is written down explicitly. For this, we denote the first iteration where the nodes i and j are connected by d_{ij}.

Theorem 4.1
Let H_k be recursively defined through eq. (4.1). For every $k \in \mathbb{N}$ and $\sigma \in \{-1, 1\}^{2^k}$ the Hamiltonian after k iterations is equal to

$$H_k(\sigma) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j \tag{4.2}$$

with

$$J_{ij} = J(d_{ij}, k, \varphi) = \frac{4^{-\varphi(d_{ij}-1)} - 4^{-\varphi k}}{4^{\varphi} - 1} \sum_{\mu=1}^{M} \xi_i^{\mu}\xi_j^{\mu}. \tag{4.3}$$

Proof. The statement can be proven by induction. With the initial value $H_0(\sigma) = 0$ and $k = 1$, the recursive formula in eq. (4.1) leads to

$$H_1(\sigma) = H_0(\sigma_1) + H_0(\sigma_2) - \frac{1}{2^{2\varphi}} \sum_{\substack{i,j=1 \\ i<j}}^{2} W_{ij}\,\sigma_i\sigma_j = -\frac{1}{4^{\varphi}} \sum_{\substack{i,j=1 \\ i<j}}^{2} W_{ij}\,\sigma_i\sigma_j.$$

W_{ij} is defined as in eq. (1.1). The two available nodes are connected in the first step of the iteration, and thus, $d_{ij} = 1$. The fact that

$$\frac{4^{-\varphi(d_{ij}-1)} - 4^{-\varphi k}}{4^{\varphi} - 1} = \frac{4^{-\varphi(1-1)} - 4^{-\varphi}}{4^{\varphi} - 1} = \frac{1 - 4^{-\varphi}}{4^{\varphi} - 1} = \frac{1}{4^{\varphi}}$$

proves the statement for $k = 1$.

Assume that eq. (4.2) is true for an arbitrary but fixed k. With the recursion rule it follows that

$$H_{k+1}(\sigma) = H_k(\sigma_1) + H_k(\sigma_2) - \frac{1}{2^{2\varphi(k+1)}} \sum_{\substack{i,j=1 \\ i<j}}^{2^{k+1}} W_{ij}\sigma_i\sigma_j, \tag{4.4}$$

where $H_k(\sigma_1)$ resp. $H_k(\sigma_2)$ each depend on one half of the nodes. In the same way we split the last sum of eq. (4.4) into

$$\sum_{\substack{i,j=1 \\ i<j}}^{2^{k+1}} W_{ij}\sigma_i\sigma_j = \underbrace{\sum_{\substack{i,j=1 \\ i<j}}^{2^{k}} W_{ij}\sigma_i\sigma_j}_{\text{spins of } \sigma_1} + \underbrace{\sum_{\substack{i,j=2^k+1 \\ i<j}}^{2^{k+1}} W_{ij}\sigma_i\sigma_j}_{\text{spins of } \sigma_2} + \underbrace{\sum_{\substack{1\le i\le 2^k \\ 2^k+1\le j\le 2^{k+1}}} W_{ij}\sigma_i\sigma_j}_{\text{connections between } \sigma_1,\sigma_2} .$$

Together with the induction hypothesis for $H_k(\sigma_1)$, we conclude

$$\begin{aligned} H_k(\sigma_1) - \frac{1}{2^{2\varphi(k+1)}} \sum_{\substack{i,j=1 \\ i<j}}^{2^{k}} W_{ij}\sigma_i\sigma_j &= -\sum_{\substack{i,j=1 \\ i<j}}^{2^{k}} \left(\frac{4^{-\varphi(d_{ij}-1)} - 4^{-\varphi k}}{4^{\varphi} - 1} + \frac{1}{2^{2\varphi(k+1)}} \right) W_{ij}\sigma_i\sigma_j \\ &= -\sum_{\substack{i,j=1 \\ i<j}}^{2^{k}} \frac{4^{-\varphi(d_{ij}-1)} - 4^{-\varphi(k+1)}}{4^{\varphi} - 1} W_{ij}\sigma_i\sigma_j. \end{aligned}$$

The same equality can be shown for $H_k(\sigma_2)$ and the spins of σ_2. All connections between σ_1 and σ_2 appear for the first time at iteration $k+1$. For these i, j's we have $d_{ij} = k+1$ and in this case

$$\frac{4^{-\varphi(d_{ij}-1)} - 4^{-\varphi(k+1)}}{4^{\varphi} - 1} = \frac{1}{2^{2\varphi(k+1)}}.$$

Altogether this proves

$$
\begin{aligned}
H_{k+1}(\sigma) = H_k(\sigma_1) - \frac{1}{2^{2\varphi(k+1)}} \sum_{\substack{i,j=1\\ i<j}}^{2^k} W_{ij}\sigma_i\sigma_j + H_k(\sigma_2) - \frac{1}{2^{2\varphi(k+1)}} \sum_{\substack{i,j=2^k+1\\ i<j}}^{2^{k+1}} W_{ij}\sigma_i\sigma_j \\
- \frac{1}{2^{2\varphi(k+1)}} \sum_{\substack{1\le i\le 2^k\\ 2^k+1\le j\le 2^{k+1}}} W_{ij}\sigma_i\sigma_j \\
= -\sum_{i<j} \frac{4^{-\varphi(d_{ij}-1)} - 4^{-\varphi(k+1)}}{4^{\varphi}-1} W_{ij}\sigma_i\sigma_j = -\sum_{i<j} J(d_{ij}, k+1, \varphi)\, \sigma_i\sigma_j.
\end{aligned}
$$

□

The explicit formula in eq. (4.2) for the Hamiltonian of the hierarchical Hopfield model compared to the Hamiltonian of the standard Hopfield model (see eq. (1.9) with $h = 0$) shows the additional discount factor which comes into play:

$$
\frac{4^{-\varphi(d_{ij}-1)} - 4^{-\varphi k}}{4^{\varphi}-1}. \tag{4.5}
$$

The $(J_{ij})_{i,j}$ from eq. (4.3), which are equal to the discount factor multiplied by the weight according to the Hebb rule, determine the influence of one neuron to another. In eq. (4.5) we see that for fixed k the value of the discount factor decreases if the distance (measured through d_{ij}) increases. In the next chapter we define a Hopfield model with inhomogeneous weights to prove a lower bound for the storage capacity of this model. Then we adopt the idea of the hierarchical Hopfield model and analyse the storage capacity on different graphs, where the weights depend on the graph distance between the nodes.

Agliari et. al. focused in [Agl+15a; Agl+15b] on another aspect of the model. Since the last connection which is added in a step of the recursion always has the smallest discount factor and wires two strongly interconnected parts of the network together, the network can work as a serial and parallel processor. This means, additionally to the retrieval of one pattern, the model can for example retrieve two patterns, one on each subnetwork, at the same time. For detailed results the reader is referred to [Agl+15a].

4.2. The inhomogeneous Hopfield model

Let $G = (V, E)$ be a simple graph, which means there are neither multi-edges nor loops. V is a set of vertices and E represents a set of undirected edges. We assume that $N :=$

$|V| < \infty$. On this graph we want to define a Hopfield model with weights decreasing with the distance between two nodes.

The patterns we want the net to store are denoted by $\xi^1, \dots, \xi^M$. We assume that these patterns are independent and each pattern is chosen uniformly at random from the configuration space $\{-1, 1\}^N$. Furthermore, let $w = (w_{ij})_{i,j \in V}$ be weights for which we assume that $w_{ij} \geq 0$ and $w_{ij} = w_{ji}$ for all $i, j \in V$. For each $i \in V$ we set $w_i = (w_{ij})_{j \in V}$. We define the dynamics $T^w = (T_i^w)_{i \in V}$ through

$$T_i^w(\sigma) = \operatorname{sgn}\left(\sum_{j \in V} w_{ij} \left(\sum_{\mu=1}^{M} \xi_i^\mu \xi_j^\mu\right) \sigma_j\right). \tag{4.6}$$

This corresponds to the standard Hopfield model with additional weights for each connection. Indeed, with $w_{ij} = 1$ for all i, j we retrieve the dynamics of the standard Hopfield model.

Theorem 4.2
For weights $(w_{ij})_{i,j \in V}$ as defined above, set

$$M(N) = \frac{\min_{i \in V} W_i}{c \log(N)}$$

with

$$W_i = \frac{\left(\sum_{j \in V} \omega_{ij}\right)^2}{\left(\sum_{j \in V} \omega_{ij}^2\right)}.$$

(1) If $c > 2$, then for any $\nu = 1, \dots, M$

$$\mathbb{P}\left(T^w(\xi^\nu) = \xi^\nu\right) = 1 - R_N$$

with $\lim_{N \to \infty} R_N = 0$.

(2) If $c \geq 4$, then

$$\mathbb{P}\left(\forall \mu \leq M : \, T^w(\xi^\mu) = \xi^\mu\right) = 1 - R_N$$

with $\lim_{N \to \infty} R_N = 0$.

Proof. The proof goes along the lines as the one of Theorem 1.1. The initial configuration is ξ^1 which corresponds to the case of $\rho = 0$ in the setting of Theorem 1.1. With the exponential Chebyshev inequality, an upper bound for the probability of a false update is given by

$$
\begin{aligned}
\mathbb{P}(T_i^w(\xi^1) \neq \xi_i^1) &\leq \mathbb{P}\Bigg(-\sum_{\mu=2}^{M} \xi_i^1 \xi_i^\mu \sum_{j \in V} \omega_{ij} \xi_j^1 \xi_j^\mu \geq \sum_{j \in V} \omega_{ij} \Bigg) \\
&\leq \exp\Bigg(-t \sum_{j \in V} w_{ij} \Bigg) \mathbb{E}\Bigg[\exp\Bigg(-t \sum_{\mu=1}^{M} \xi_i^1 \xi_i^\mu \sum_{j \in V} w_{ij} \xi_j^1 \xi_j^\mu \Bigg) \Bigg] \\
&\leq \exp\Bigg(-t \sum_{j \in V} w_{ij} \Bigg) \mathbb{E}\Bigg[\prod_{j \in V} \cosh\Bigg(-t\, w_{ij}\, \xi_j^1 \xi_j^\mu \Bigg) \Bigg]^M \\
&\leq \exp\Bigg(-t \sum_{j \in V} \omega_{ij} \Bigg) \exp\Bigg(\frac{t^2}{2} \sum_{j \in V} \omega_{ij}^2 \Bigg)^M .
\end{aligned}
$$

The arguments are known from the previous proof: We used Lemma 1.2 and the inequality $\cosh(x) \leq \exp(\frac{x^2}{2})$. If we choose

$$
t = \frac{1}{M} \frac{\left(\sum_{j \in V} \omega_{ij} \right)}{\left(\sum_{j \in V} \omega_{ij}^2 \right)},
$$

the bound is equal to

$$
\mathbb{P}(T_i^w(\xi^1) \neq \xi_i^1) \leq \exp\left(-\frac{1}{2} \frac{\left(\sum_{j \in V} \omega_{ij} \right)^2}{\left(\sum_{j \in V} \omega_{ij}^2 \right)} \frac{1}{M} \right) = \exp\left(-\frac{1}{2} \frac{W_i}{M} \right).
$$

Because of the weights, we do not have a uniform bound for the probability anymore. We see that the term W_i determines how fast the probability for a wrong update of neuron i is decreasing. Thus, W_i can be seen as a measurement of the susceptibility to errors for neuron i. Boole's inequality together with the minimal value of W_i as an upper bound leads to

$$
\begin{aligned}
\mathbb{P}\left(\exists i \in V :\ T_i^w(\xi^1) \neq \xi_i^1 \right) &\leq N \max_{i \in V} \exp\left(-\frac{1}{2} \frac{W_i}{M} \right) \\
&\leq \exp\left(\log(N) - \frac{1}{2} \frac{\min_{i \in V} W_i}{M} \right) \\
&\leq \exp\left(-\log(N) \left[\frac{1}{2} \frac{\min_{i \in V} W_i}{M \log(N)} - 1 \right] \right).
\end{aligned}
\tag{4.7}
$$

Since

$$M = \frac{\min_{i\in V} W_i}{c\log(N)}$$

for $c > 2$, we conclude that

$$\mathbb{P}(\exists i \in V:\ T_i^w(\xi^1) \neq \xi_i^1) \leq N^{1-\frac{c}{2}} \overset{N\to\infty}{\longrightarrow} 0.$$

This proves the first statement. For the second statement, we use Boole's inequality again and conclude that

$$\mathbb{P}(\exists \mu \leq M:\ T^w(\xi^\mu) \neq \xi^\mu) \leq M\, N^{1-\frac{c}{2}}. \tag{4.8}$$

Now let $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the ℓ_1-norm resp. ℓ_2-norm on $\mathbb{R}^n$, i.e. for $x \in \mathbb{R}^n$

$$\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2} \qquad \text{and} \qquad \|x\|_1 = \sqrt{\sum_{i=1}^n |x_i|}. \tag{4.9}$$

and write

$$W_i = \frac{\left(\sum_{j\in V} w_{ij}\right)^2}{\sum_{j\in V} w_{ij}^2} = \left(\frac{(\|w_i\|_1)^2}{\|w_i\|_2}\right)^2.$$

Since W_i, and therefore $\min_i W_i$, is invariant under scaling the weights by any factor, we can assume that $0 \leq w_{ij} \leq 1$ for all i, j. Now with the Cauchy-Schwarz inequality (see proof of Lemma A.11) as well as $0 \leq w_{ij} \leq 1$, we know that

$$\sum_{j\in V} \omega_{ij} \leq W_i \leq N.$$

Therefore, the bound for the storage capacity of the standard Hopfield model is an upper bound for the number of patterns in this setting, i.e.

$$M \leq \frac{N}{c\log(N)}.$$

Thus, the probability in eq. (4.8) can be bounded by

$$\mathbb{P}(\exists \mu:\ T^w(\xi^\mu) \neq \xi^\mu) \leq M\, N^{1-\frac{c}{2}} \leq \frac{N^{2-\frac{c}{2}}}{c\log(N)}.$$

If $c \geq 4$, this converges to zero and proves the second statement. $\square$

First, we notice that the storage capacity of the Hopfield model with weights can be bottlenecked by a single neuron. The value of M in Theorem 4.2 depends on the smallest of all W_i, $i \in V$. This is reasonable and can be compared to results about the Hopfield model on random graphs. To get a decent storage capacity, the random graph needs to have a uniform connectivity. This requirement is met by making assumptions on the minimal degree, the average degree and the spectral gap of the graph (see [KP93; LV15]). Graphs with bottlenecks cause problems and can prevent the net from storing any pattern. The same problem can occur for the Hopfield model with inhomogeneous weights.

Additionally, if we set $w_{ij} = 1$ for all i, j, we retrieve the dynamics of the standard Hopfield model. It follows that $W_i = N$ for all $i \in V$ and that the lower bound for the storage capacity from Theorem 4.2 coincides with the bound from Theorem 1.1. As we have seen in the proof of Theorem 4.2 the storage capacity of the standard Hopfield model is an upper bound for the storage capacity with weights.

4.3. The Hopfield model with weights on a deterministic graph

With Theorem 4.2 we derived a lower bound for the storage capacity in a Hopfield model with weights. The new terms which appear in the result, namely W_i for $i \in V$, are difficult to calculate for arbitrary weights. Based on the idea of a hierarchical Hopfield model, in this section we choose weights which decrease with the distance between two nodes and calculate the asymptotic behaviour of W_i for certain graphs. As a metric we use the graph distance $d_g(i, j)$ between two nodes $i \neq j$, which is equal to the number of edges in the shortest path connecting them. On a disconnected graph the model can be analysed separately on each component. Because of this, we assume that the graph G is connected. Thus, $d_g(i, j) < \infty$ for all pairs of nodes.

4.3.1. Lattice structure on a torus

The Hopfield model with weights is defined on a finite graph. For the first application we define $G_k^{(1)} = (V_k^{(1)}, E_k^{(1)})$ to be the graph $\mathbb{Z}^2 \cap [-k, k]$ with periodic boundary conditions. This means $G_k^{(1)}$ is a subgraph of the lattice $\mathbb{Z}^2$ and for every $m \in \mathbb{Z} \cap [-k, k]$ the vertices (m, k) and (k, m) are connected to the vertices $(m, -k)$ resp. $(-k, m)$. The total amount of vertices of $G_k^{(1)}$ is equal to $N := |V_k^{(1)}| = (2k + 1)^2$.

This graph is chosen to be the first example because its structure is very homogeneous. If we denote by $Z_l(i)$ the number of vertices at distance l from the vertex $i \in V$, then

$$Z_0(i) = 1, \tag{4.10}$$

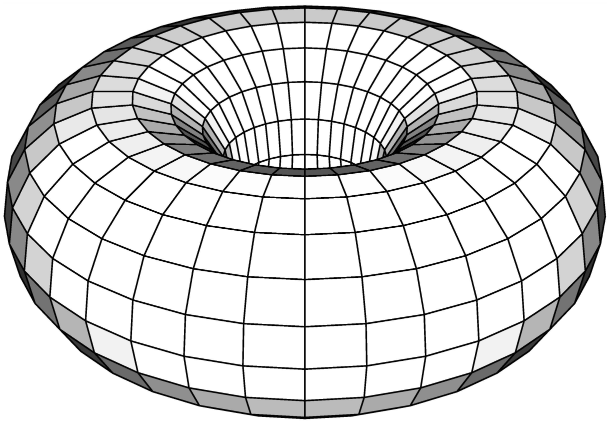

Figure 4.2.: Lattice on a torus

$$
\begin{aligned}
Z_l(i) &= 4l && \text{if } 1 \le l \le k,\\
Z_l(i) &= 4(l-1) = 4k && \text{if } l = k+1,\\
\text{and } Z_l(i) &= 4k - 4(l-k-1) = 4(2k-l+1) && \text{if } k+2 \le l \le 2k.
\end{aligned}
$$

This does not depend on the choice of i and since we want the weights to depend only on the distance between nodes, the factor W_i has the same value for all $i \in V_k^{(1)}$. It is important to mention that each vertex is only counted once because just the shortest connection determines the distance between two nodes. This fits to the Hopfield model where we want to consider each signal only once. Furthermore, the decreasing number of nodes for large l is a technical phenomenon. Since we need a finite graph, the growth of the graph gets cut off at some point. On the infinite lattice graph $\mathbb{Z}^2$ the number of nodes at distance l grows like $4l$ for all $l \in \mathbb{N}$.

For $\varphi > 0$ and nodes i, j with $d_g(i,j) = l > 0$ we set the weight w_{ij} equal to

$$w_l := w_{ij} = l^{-\varphi} \tag{4.11}$$

and $w_{ii} = 0$ for all $i \in V_k^{(1)}$. Compared to the discount factor in the hierarchical Hopfield model (see eq. (4.3)), in this approach the discount factor is of a simpler form. In both models, larger values of φ lead to a faster decrease of signals at greater distances. Additionally, the parameter φ allows us to compensate for the increasing number of nodes.

As mentioned before, since $Z_l(i)$ does not depend on i, the value of W_i is constant in i. Thus, to determine $\min_i W_i$, we need to calculate

$$W_i = \frac{\left(\sum_{j\in V} w_{ij}\right)^2}{\sum_{j\in V} w_{ij}^2} = \frac{\left(\sum_{l=0}^{2k} Z_l\, l^{-\varphi}\right)^2}{\sum_{l=0}^{2k} Z_l\, l^{-2\varphi}}. \tag{4.12}$$

An important observation is that the numerator and the denominator in eq. (4.12) are of the same structure. To determine the asymptotic behaviour of W_i, it is sufficient to calculate the asymptotics of $\sum_{l=0}^{2k} Z_l \; l^{-\varphi}$ for an arbitrary $\varphi > 0$. Together with the insights about Z_l from eq. (4.10), we need to analyse the asymptotic order of

$$\begin{aligned}\sum_{l=0}^{2k} Z_l l^{-\varphi} &= \sum_{l=1}^{k} 4l^{1-\varphi} + \sum_{l=k+1}^{k+1} 4(l-1)l^{-\varphi} + \sum_{l=k+2}^{2k} 4(2k-l+1)\, l^{-\varphi} \\ &= 4\left(\sum_{l=1}^{k} l^{1-\varphi} + k^{1-\varphi}\left(1+\frac{1}{k}\right)^{-\varphi} + (2k+1)\sum_{l=k+2}^{2k} l^{-\varphi} - \sum_{l=k+2}^{2k} l^{1-\varphi}\right)\end{aligned} \tag{4.13}$$

for $\varphi > 0$. The sums appearing in eq. (4.13) can be bounded from below and from above by an appropriate integral. To do so, we use the function x^τ on the positive half-line, which is increasing if $\tau \geq 0$ and decreasing otherwise. For every $a, b \in \mathbb{N}$ with $2 \leq a < b$ the sum for $\tau \geq 0$ is bounded from above by

$$\sum_{l=a}^{b} l^\tau \leq \int_a^{b+1} x^\tau \,\mathrm{d}x = \frac{1}{1+\tau}\left((b+1)^{1+\tau} - a^{1+\tau}\right) \tag{4.14}$$

and from below by

$$\sum_{l=a}^{b} l^\tau \geq \int_{a-1}^{b} x^\tau \,\mathrm{d}x = \frac{1}{1+\tau}\left(b^{1+\tau} - (a-1)^{1+\tau}\right). \tag{4.15}$$

In the case of $\tau < 0$ an upper bound for the sum is

$$\sum_{l=a}^{b} l^\tau \leq \int_{a-1}^{b} x^\tau \,\mathrm{d}x = \begin{cases} \dfrac{1}{1+\tau}\left(b^{1+\tau} - (a-1)^{1+\tau}\right) & \text{if } \tau \neq -1 \\ \log(b) - \log(a-1) & \text{if } \tau = -1 \end{cases} \tag{4.16}$$

and as a lower bound we have

$$\sum_{l=a}^{b} l^\tau \geq \int_{a}^{b+1} x^\tau \,\mathrm{d}x = \begin{cases} \dfrac{1}{1+\tau}\left((b+1)^{1+\tau} - a^{1+\tau}\right) & \text{if } \tau \neq -1 \\ \log(b+1) - \log(a) & \text{if } \tau = -1 \end{cases}. \tag{4.17}$$

These four inequalities (eqs. (4.14) to (4.17)) can be used to analyse the sums involved in eq. (4.13). We write $a_k \approx b_k$ if $\frac{a_k}{b_k} \overset{k\to\infty}{\longrightarrow} 1$.

(1) In the first sum we exclude the index $l = 1$ to avoid problems with the integrability

of the upper bound. Applying the statements in eqs. (4.16) and (4.17) leads to

$$
\begin{aligned}
1+\sum_{l=2}^{k} l^{1-\varphi} &\leq \begin{cases} 1+\dfrac{1}{2-\varphi}\left((k+1)^{2-\varphi}-2^{2-\varphi}\right), & \text{if } 0<\varphi\leq 1 \\ 1+\dfrac{1}{2-\varphi}\left(k^{2-\varphi}-1^{2-\varphi}\right), & \text{if } 1<\varphi \text{ and } \varphi\neq 2 \\ 1+\log(k), & \text{if } \varphi=2 \end{cases} \\
&\approx \begin{cases} \dfrac{1}{2-\varphi}k^{2-\varphi}, & \text{if } 0<\varphi<2 \\ \log(k), & \text{if } \varphi=2 \\ 1, & \text{if } \varphi>2 \end{cases}
\end{aligned}
$$

and

$$
\begin{aligned}
1+\sum_{l=2}^{k} l^{1-\varphi} &\geq \begin{cases} 1+\dfrac{1}{2-\varphi}\left(k^{2-\varphi}-1^{2-\varphi}\right), & \text{if } 0<\varphi\leq 1 \\ 1+\dfrac{1}{2-\varphi}\left((k+1)^{2-\varphi}-2^{2-\varphi}\right), & \text{if } 1<\varphi \text{ and } \varphi\neq 2 \\ 1+\left(\log(k+1)-\log(2)\right), & \text{if } \varphi=2 \end{cases} \\
&\approx \begin{cases} \dfrac{1}{2-\varphi}k^{2-\varphi}, & \text{if } 0<\varphi<2 \\ \log(k), & \text{if } \varphi=2 \\ 1, & \text{if } \varphi>2 \end{cases}
\end{aligned}
$$

Since the asymptotic behaviour of both bounds coincide, the asymptotics of the first sum is equal to

$$
\sum_{l=1}^{k} l^{1-\varphi} \approx \begin{cases} \dfrac{1}{2-\varphi}k^{2-\varphi}, & \text{if } 0<\varphi<2 \\ \log(k), & \text{if } \varphi=2 \\ 1, & \text{if } \varphi>2 \end{cases}.
$$

(2) With the same arguments applied to the second sum, we get an upper bound

$$
\sum_{l=k+2}^{2k} l^{-\varphi} \leq \begin{cases} \dfrac{1}{1-\varphi}\left((2k)^{1-\varphi}-(k+1)^{1-\varphi}\right), & \text{if } \varphi>0 \text{ and } \varphi\neq 1 \\ \left(\log(2k)-\log(k+1)\right), & \text{if } \varphi=1 \end{cases}
$$

and a lower bound

$$\sum_{l=k+2}^{2k} l^{-\varphi} \geq \begin{cases} \dfrac{1}{1-\varphi}\left((2k+1)^{1-\varphi}-(k+2)^{1-\varphi}\right), & \text{if } \varphi > 0 \text{ and } \varphi \neq 1 \\ (\log(2k+1)-\log(k+2)), & \text{if } \varphi = 1 \end{cases}.$$

Asymptotically both bounds have the same behaviour. Thus,

$$\sum_{l=k+2}^{2k} l^{-\varphi} \approx \begin{cases} \dfrac{(2^{1-\varphi}-1)}{1-\varphi}k^{1-\varphi}, & \text{if } 0 < \varphi,\ \varphi \neq 1 \\ \log(2), & \text{if } \varphi = 1 \end{cases}.$$

Additionally, we need to consider the factor $(2k+1)$ in front of the sum. Taking this into account, we conclude that

$$(2k+1)\sum_{l=k+2}^{2k} l^{-\varphi} \approx \begin{cases} \dfrac{2(2^{1-\varphi}-1)}{1-\varphi}k^{2-\varphi}, & \text{if } 0 < \varphi,\ \varphi \neq 1 \\ 2\log(2)k, & \text{if } \varphi = 1 \end{cases}.$$

Here, we need to mention that $k^{2-\varphi}$ goes to zero if $\varphi > 2$ and k tends to infinity.

(3) The third sum is similar to the first one but it has a different range of indices. By using eqs. (4.14) to (4.17) we derive the following inequalities:

$$\sum_{l=k+2}^{2k} l^{1-\varphi} \leq \begin{cases} \dfrac{1}{2-\varphi}\left((2k+1)^{2-\varphi}-(k+2)^{2-\varphi}\right), & \text{if } 0 < \varphi \leq 1 \\ \dfrac{1}{2-\varphi}\left((2k)^{2-\varphi}-(k+1)^{2-\varphi}\right), & \text{if } 1 < \varphi \text{ and } \varphi \neq 2 \\ \log(2k)-\log(k+1), & \text{if } \varphi = 2 \end{cases}$$

and

$$\sum_{l=k+2}^{2k} l^{1-\varphi} \geq \begin{cases} \dfrac{1}{2-\varphi}\left((2k)^{2-\varphi}-(k+1)^{2-\varphi}\right), & \text{if } 0 < \varphi \leq 1 \\ \dfrac{1}{2-\varphi}\left((2k+1)^{2-\varphi}-(k+2)^{2-\varphi}\right), & \text{if } 1 < \varphi \text{ and } \varphi \neq 2 \\ \log(2k+1)-\log(k+2), & \text{if } \varphi = 2 \end{cases}.$$

For $\varphi \neq 2$ we see that

$$(2k+1)^{2-\varphi}-(k+2)^{2-\varphi} \approx (2^{2-\varphi}-1)k^{2-\varphi}$$

and the same asymptotics is true for $(2k)^{2-\varphi}-(k+1)^{2-\varphi}$. If $\varphi = 2$, both bounds converge to $\log(2)$ for $k \to \infty$. Therefore, the third sum in eq. (4.13) is asymptotically equivalent to

$$\sum_{l=k+2}^{2k} l^{1-\varphi} \approx \begin{cases} \frac{(2^{2-\varphi}-1)}{2-\varphi} k^{2-\varphi}, & \text{if } 0<\varphi,\, \varphi \neq 2 \\ \log(2), & \text{if } \varphi = 2 \end{cases}.$$

These insights are enough to calculate the asymptotic behaviour of the term in eq. (4.13) for $k \to \infty$. In the case where $\varphi \notin \{1,2\}$, we observe that the term grows to infinity if $0 < \varphi < 2$ and converges to a constant in the case of $\varphi > 2$:

$$\begin{aligned} & 4\left(\sum_{l=1}^{k} l^{1-\varphi} + k^{1-\varphi}\left(1+\frac{1}{k}\right)^{-\varphi} + (2k+1)\sum_{l=k+2}^{2k} l^{-\varphi} - \sum_{l=k+2}^{2k} l^{1-\varphi}\right) \\ \approx\ & 4\left(1+\frac{1}{2-\varphi}k^{2-\varphi} + k^{1-\varphi}\left(1+\frac{1}{k}\right)^{-\varphi} + \frac{2(2^{1-\varphi}-1)}{1-\varphi}k^{2-\varphi} - \frac{2^{2-\varphi}-1}{2-\varphi}k^{2-\varphi}\right) \\ =\ & 4+\frac{4}{2-\varphi}k^{2-\varphi}\left(1+\frac{1}{k}\left(1+\frac{1}{k}\right)^{-\varphi} + 2(2^{1-\varphi}-1)\frac{2-\varphi}{1-\varphi} - \left(2^{2-\varphi}+1\right)\right) \\ =\ & 4+\frac{4}{2-\varphi}k^{2-\varphi}\left(2(2^{1-\varphi}-1)\frac{1}{1-\varphi} + \frac{1}{k}\left(1+\frac{1}{k}\right)^{-\varphi}\right) \\ \approx\ & 4+\frac{8}{2-\varphi}\frac{2^{1-\varphi}-1}{1-\varphi}k^{2-\varphi} \end{aligned} \tag{4.18}$$

The special case where $\varphi = 1$ is just a technical difficulty and shows the same behaviour

$$\begin{aligned} & 4\left(\sum_{l=1}^{k} l^{1-\varphi} + k^{1-\varphi}\left(1+\frac{1}{k}\right)^{-\varphi} + (2k+1)\sum_{l=k+2}^{2k} l^{-\varphi} - \sum_{l=k+2}^{2k} l^{1-\varphi}\right) \\ \approx\ & 4\left(\frac{1}{2-\varphi}k^{2-\varphi} + k^{1-\varphi}\left(1+\frac{1}{k}\right)^{-\varphi} + 2\log(2)k - \frac{2^{2-\varphi}-1}{2-\varphi}k^{2-\varphi}\right) \\ =\ & \frac{4}{2-\varphi}k^{2-\varphi}\left(\frac{1}{k}\left(1+\frac{1}{k}\right)^{-\varphi} + 2\log(2) - \left(2^{2-\varphi}+2\right)\right) \\ =\ & 8\log(2)k\left(1+\frac{1}{8\log(2)}\frac{1}{k}\left(1+\frac{1}{k}\right)^{-1}\right) \\ \approx\ & 8\log(2)k. \end{aligned} \tag{4.19}$$

Even the constant fits perfectly to the results from before because

$$\lim_{\varphi\to 1}\frac{2^{1-\varphi}-1}{1-\varphi} = \log(2).$$

In between the diverging case $\varphi < 2$ and the converging case $\varphi > 2$, we have $\varphi = 2$ where the term also goes to infinity but only in logarithmic speed:

$$\begin{aligned}
&4\left(\sum_{l=1}^{k} l^{1-\varphi} + k^{1-\varphi}\left(1+\frac{1}{k}\right)^{-\varphi} + (2k+1)\sum_{l=k+2}^{2k} l^{-\varphi} - \sum_{l=k+2}^{2k} l^{1-\varphi}\right) \\
\approx\ & 4\left(\log(k) + k^{1-\varphi}\left(1+\frac{1}{k}\right)^{-\varphi} + \frac{2(2^{1-\varphi}-1)}{1-\varphi}k^{2-\varphi} - \log(2)\right) \\
=\ & 4\log(k)\left(1 + \frac{1}{k\log(k)}\left(1+\frac{1}{k}\right)^{-2} + \frac{1}{\log(k)} - \frac{\log(2)}{\log(k)}\right) \\
\approx\ & 4\log(k). \qquad (4.20)
\end{aligned}$$

The asymptotic equivalences we showed in eqs. (4.18) to (4.20) can be used to derive the asymptotic behaviour of W_i by applying them to the numerator and denominator (using 2φ for the latter one).

(1) If $0 < \varphi < 1$ and $\varphi \neq \frac{1}{2}$, the asymptotics of the numerator and the denominator are given in eq. (4.18). Thus,

$$\begin{aligned}
\min_i W_i \approx \frac{\left(\frac{8}{2-\varphi}\frac{2^{1-\varphi}-1}{1-\varphi}k^{2-\varphi}\right)^2}{\frac{8}{2-2\varphi}\frac{2^{1-2\varphi}-1}{1-2\varphi}k^{2-2\varphi}} &= \frac{16(2^{1-\varphi}-1)^2}{(2^{1-2\varphi}-1)}\frac{(1-2\varphi)}{(2-\varphi)^2(1-\varphi)}k^2 \\
&\approx \frac{4(2^{1-\varphi}-1)^2}{(2^{1-2\varphi}-1)}\frac{(1-2\varphi)}{(2-\varphi)^2(1-\varphi)}N.
\end{aligned}$$

In the last line we used that the lattice graph on a torus has a total number of nodes of $N := |V| = (2k+1) \approx 4k^2$. This makes the results comparable to former results and fits to the notation of the standard Hopfield model.

(2) If $\varphi = \frac{1}{2}$, the asymptotics of the numerator is given in eq. (4.18) and the asymptotics of the denominator is given in eq. (4.19). Thus,

$$\begin{aligned}
\min_i W_i \approx \frac{\left(\frac{8}{2-\varphi}\frac{2^{1-\varphi}-1}{1-\varphi}k^{2-\varphi}\right)^2}{8\log(2)k} &= \frac{128(\sqrt{2}-1)^2}{9\log(2)}k^2 \\
&\approx \frac{32(\sqrt{2}-1)^2}{9\log(2)}N.
\end{aligned}$$

(3) If $\varphi = 1$, the asymptotics of the numerator is given in eq. (4.19) and the asymptotics

of the denominator is given in eq. (4.20). Thus,

$$\min_i W_i \approx \frac{(8\log(2)k)^2}{4\log(k)} = 4\log(2)^2 \frac{4k^2}{\log(k)}$$
$$\approx 8\log(2)^2 \frac{N}{\log(N)}.$$

Here, we used that $\log(k) \approx \frac{1}{2}\log(N)$.

(4) If $1 < \varphi < 2$, the asymptotics of both, numerator and denominator, are given in eq. (4.18). In the denominator the term $k^{2-2\varphi}$ converges to zero because $2\varphi > 2$. Therefore, the denominator converges to a constant. Thus,

$$\min_i W_i \approx \frac{\left(\frac{8}{2-\varphi}\frac{2^{1-\varphi}-1}{1-\varphi}k^{2-\varphi}\right)^2}{4 + \frac{8}{2-2\varphi}\frac{2^{1-2\varphi}-1}{1-2\varphi}k^{2-2\varphi}} \approx \frac{\left(\frac{8}{2-\varphi}\frac{2^{1-\varphi}-1}{1-\varphi}k^{2-\varphi}\right)^2}{4}$$
$$= \frac{16(2^{1-\varphi}-1)^2}{(2-\varphi)^2(1-\varphi)^2}k^{2(2-\varphi)}$$
$$\approx \frac{4^{\varphi}(2^{1-\varphi}-1)^2}{(2-\varphi)^2(1-\varphi)^2}N^{2-\varphi}.$$

(5) If $\varphi = 2$, the asymptotics of the numerator is given in eq. (4.20) and the asymptotics of the denominator is given in eq. (4.18). Thus,

$$\min_i W_i \approx \frac{(4\log(k))^2}{4 + \frac{8}{2-2\varphi}\frac{2^{1-2\varphi}-1}{1-2\varphi}k^{2-2\varphi}} = \frac{(4\log(k))^2}{4} = 4\log(k)^2 \approx \log(N)^2$$

(6) If $\varphi > 2$, the asymptotics of both, numerator and denominator, are given in eq. (4.18) and for $\varphi > 2$ both terms converge to a constant. Thus,

$$\min_i W_i \approx \frac{\left(4 + \frac{8}{2-\varphi}\frac{2^{1-\varphi}-1}{1-\varphi}k^{2-\varphi}\right)^2}{4 + \frac{8}{2-2\varphi}\frac{2^{1-2\varphi}-1}{1-2\varphi}k^{2-2\varphi}} = \frac{4^2}{4} = 4.$$

In Theorem 4.2 the amount of patterns were chosen to be equal to

$$M = \frac{\min_i W_i}{c\log(N)}.$$

Since we calculated the asymptotic behaviour of $\min_i W_i$, we proved the following theorem:

Theorem 4.3

Consider a Hopfield model on a torus $G = (V, E)$ with dynamics $T^w = (T_i^w)_{i \in V}$ defined as in eq. (4.6) and weights according to eq. (4.11). Then there exist constants $c_0(\varphi), \dots, c_4(\varphi)$ such that

$$\min_i W_i \approx \left\{ \begin{array}{ll} c_0(\varphi) N, & \text{if } 0 < \varphi < 1 \\ c_1(\varphi) \dfrac{N}{\log(N)}, & \text{if } \varphi = 1 \\ c_2(\varphi) N^{2-\varphi}, & \text{if } 1 < \varphi < 2 \\ c_3(\varphi) \log(N)^2, & \text{if } \varphi = 2 \\ c_4(\varphi), & \text{if } \varphi > 2 \end{array} \right\} =: \widetilde{W}(\varphi).$$

In particular, for $M = \frac{\widetilde{W}(\varphi)}{c \log(N)}$ and $\xi^1, \dots, \xi^M$ chosen uniformly at random from $\{-1, 1\}^N$, it follows that:

(1) If $c > 2$ and $\varphi \le 2$, then for any $\nu = 1, \dots, M$

$$\mathbb{P}\left(T^w(\xi^\nu) = \xi^\nu\right) = 1 - R_N$$

with $\lim_{N \to \infty} R_N = 0$.

(2) If $c \ge 4$ and $\varphi \le 2$, then

$$\mathbb{P}\left(\forall \mu \le M : \ T^w(\xi^\mu) = \xi^\mu\right) = 1 - R_N$$

with $\lim_{N \to \infty} R_N = 0$.

The lower bounds for the storage capacity can be summarized as

	$0 < \varphi < 1$	$\varphi = 1$	$1 < \varphi < 2$	$\varphi = 2$	$\varphi > 2$
M	$\frac{N}{c_0 \log(N)}$	$\frac{N}{c_1 \log(N)^2}$	$\frac{N^{(2-\varphi)}}{c_2 \log(N)}$	$c_3 \log(N)$	0

where $c_0, \dots, c_4$ are positive constants.

The lattice graph is a homogeneous example and works as a basis for the Hopfield model with weights. The parameter φ can be used to control the decay of signals from far away. If φ is too large, the incoming signals are not strong enough to preserve the storage capacity. For $\varphi > 2$ we cannot guarantee that the net is able to store any pattern. A special feature of the lattice graph is that the observable neighbourhood is the same for each node. This made the calculations easier. In the next example this is will not be true and we need to analyse how W_i behaves for different i.

4.3.2. Regular tree graph

For the next example, we take a look at a μ-regular tree graph. Starting with the root, in this graph each vertex has exactly $\mu \in \mathbb{N}$ descendants. Let $G_k^{(2)} = (V_k^{(2)}, E_k^{(2)})$ be a μ-regular tree graph up to generation k. In contrast to the previous example, the amount of nodes with a specific distance to a fixed vertex now varies with its generation. At generation zero we only have the root node. By construction, the root has μ direct neighbours and with each step one node arises μ new vertices. Thus, for all $0 \leq l \leq k$, the amount of nodes at distance l to generation 0 is equal to $Z_l(0) = \mu^l$. Therefore, the total number of vertices in $G_k^{(2)}$ is given by

$$N := |V_k^{(2)}| = \sum_{l=0}^{k} \mu^l = \frac{\mu^{k+1} - 1}{\mu - 1} \approx \mu^k \frac{\mu}{\mu - 1}. \tag{4.21}$$

For a node at an arbitrary generation n, $0 \leq n \leq k$, the number of vertices at distance l, $0 \leq l \leq 2k$, can be expressed as

$$Z_l(n) = \underbrace{\mu^l \, \mathbb{1}_{\{0 \leq l \leq k-n\}}}_{\color{green}\bullet} + \underbrace{\mathbb{1}_{\{1 \leq l \leq n\}}}_{\color{red}\bullet} + \underbrace{\sum_{j=2}^{n+1} \mu^{l-j} \, \mathbb{1}_{\{j \leq l \leq k-n+2j-2\}}}_{\color{blue}\bullet} . \tag{4.22}$$

With the help of fig. 4.3, we see that the first term (green nodes) is counting the vertices of the μ-regular subtree starting at generation n. The red nodes are included in the second term and to each of these red nodes there is another μ-regular subtree connected (blue nodes). The blue subtrees have different heights because they are starting at each generation from 1 to n and all are ending at generation k. The blue nodes are covered by the third term, where each summand corresponds to one of the subtrees. For example, the smallest subtree is represented by the term with $j = 2$. Each subtree contains nodes whose distance from the fixed vertex at generation n is between j and $k - n + 2j - 2$.

In the previous example (Torus, see Section 4.3.1), Z_l grew linearly in l and the weights were set to $w_{ij} = l^{-\varphi}$. In the case of a tree graph, we observe an exponential growth of $Z_l(n)$ in l. For this reason, we define the weight as follows: For $\varphi > 0$ and nodes i, j with $d_g(i, j) = l \geq 0$ let w_{ij} be equal to

$$w_l := w_{ij} = \mu^{-\varphi l}. \tag{4.23}$$

The weights only depend on the distance between nodes, and since all nodes within the same generation are exposed to the same neighbourhood structure, we are now interested

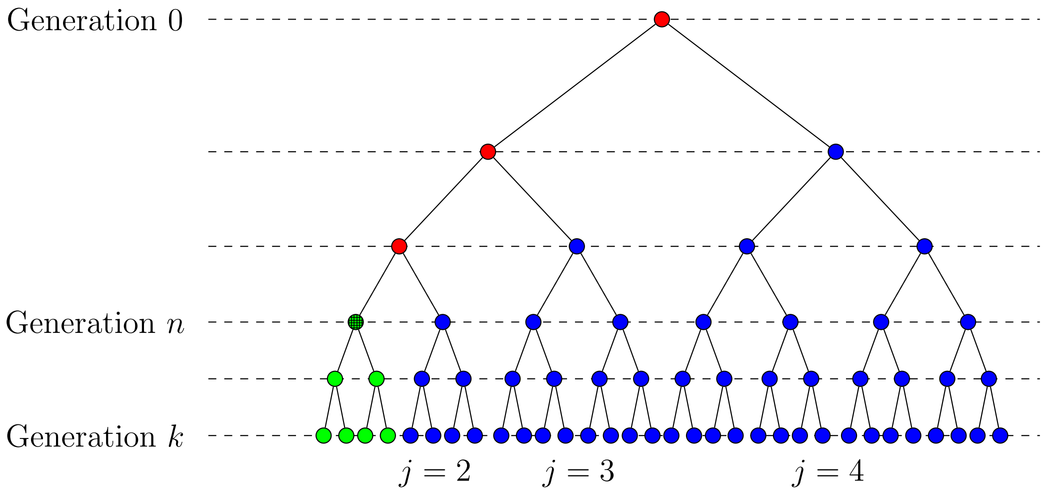

Figure 4.3.: A 2-regular tree up to generation $k = 5$.

in the asymptotic behaviour of

$$W_n = \frac{\left(\sum_{l=0}^{2k} Z_l(n)\,\mu^{-\varphi l}\right)^2}{\sum_{l=0}^{2k} Z_l(n)\,\mu^{-2\varphi l}}$$

as $k \to \infty$. Here, n denotes the generation of the fixed neuron which is updated and can be a function of k. For example, for a leaf node n would be equal to k.

A good strategy in the previous example was to first derive the asymptotic behaviour of the numerator of W_n. Let $0 \le n \le k$, $\varphi > 0$ and $\mu \in \mathbb{N}$, then with the help of eq. (4.22) we conclude that

$$\begin{aligned}\sum_{l=0}^{2k} Z_l(n)\mu^{-\varphi l} &= \sum_{l=0}^{2k}\left(\mu^l\,\mathbb{1}_{\{0\le l\le k-n\}} + \mathbb{1}_{\{1\le l\le n\}} + \sum_{j=2}^{n+1}\mu^{l-j}\,\mathbb{1}_{\{j\le l\le k-n+2j-2\}}\right)\mu^{-\varphi l}\\ &= \sum_{l=0}^{k-n}\mu^{(1-\varphi)l} + \sum_{l=1}^{n}\mu^{-\varphi l} + \sum_{l=2}^{2k}\sum_{j=2}^{n+1}\mu^{(1-\varphi)l}\mu^{-j}\,\mathbb{1}_{\{j\le l\le k-n+2j-2\}}\,. \qquad (4.24)\end{aligned}$$

The value of a geometric series can be calculated explicitly. Thus, we are not forced to find upper and lower bounds as before. For $\tau \ne 0$ and $0 \le a < b$ we use the identity

$$\sum_{l=a}^{b}\mu^{l\tau} = \mu^{a\tau}\frac{\mu^{(b-a+1)\tau}-1}{\mu^\tau - 1} = \frac{\mu^{(b+1)\tau}-\mu^{a\tau}}{\mu^\tau - 1} \qquad (4.25)$$

and every time the exponent appears to be zero, the corresponding value can be calculated with the help of continuity arguments. If $\varphi \neq 1$, for the first sum in eq. (4.24) we conclude that

$$\sum_{l=0}^{k-n} \mu^{(1-\varphi)l} = \frac{\mu^{(1-\varphi)(k-n+1)} - 1}{\mu^{(1-\varphi)} - 1}, \tag{4.26}$$

and for all $\varphi > 0$ the second sum in eq. (4.24) equals

$$\sum_{l=1}^{n} \mu^{-\varphi l} = \mu^{-\varphi} \frac{\mu^{-\varphi n} - 1}{\mu^{-\varphi} - 1}. \tag{4.27}$$

The third term is easier to handle if we swap the two sums. This means, instead of collecting all nodes with the same distance to the fixed generation, we focus on the (blue) subtrees (see fig. 4.3) first and then sum over all subtrees. This leads to

$$\begin{aligned}
\sum_{l=2}^{2k} \sum_{j=2}^{n+1} \mu^{(1-\varphi)l} \mu^{-j} \, \mathbb{1}_{\{j \le l \le k-n+2j-2\}} &= \sum_{l=2}^{2k} \sum_{j=2}^{n+1} \mu^{l-j} \mu^{-\varphi l} \, \mathbb{1}_{\{j \le l \le k-n+2j-2\}} \\
&= \sum_{j=2}^{n+1} \sum_{l=2}^{2k} \mu^{l-j} \mu^{-\varphi l} \, \mathbb{1}_{\{j \le l \le k-n+2j-2\}} \\
&= \sum_{j=2}^{n+1} \sum_{l=j}^{k-n+2j-2} \mu^{l-j} \mu^{-\varphi l} \\
&= \sum_{j=2}^{n+1} \sum_{l=0}^{k-n+j-2} \mu^{l} \mu^{-\varphi(l+j)} \\
&= \sum_{j=2}^{n+1} \mu^{-\varphi j} \sum_{l=0}^{k-n+j-2} \mu^{(1-\varphi)l} \\
&= \mu^{-2\varphi} \sum_{j=0}^{n-1} \mu^{-\varphi j} \sum_{l=0}^{k-n+j} \mu^{(1-\varphi)l}.
\end{aligned}$$

The index j is still referring to the different subtrees and each of them has a contribution in form of a geometric sum. Because of the different heights of the subtrees their values depend on j. If we use eq. (4.25), it follows that

$$\sum_{l=0}^{k-n+j} \mu^{(1-\varphi)l} = \frac{1}{\mu^{(1-\varphi)} - 1} \left(\mu^{(1-\varphi)(k-(n-1))} \mu^{(1-\varphi)j} - 1 \right),$$

and thus,

$$\mu^{-2\varphi}\sum_{j=0}^{n-1}\mu^{-\varphi j}\sum_{l=0}^{k-n+j}\mu^{(1-\varphi)l} = \frac{\mu^{-2\varphi}}{\mu^{(1-\varphi)}-1}\sum_{j=0}^{n-1}\mu^{-\varphi j}\left(\mu^{(1-\varphi)(k-n+1)}\mu^{(1-\varphi)j}-1\right)$$
$$= \frac{\mu^{-2\varphi}}{\mu^{(1-\varphi)}-1}\left(\mu^{(1-\varphi)(k-n+1)}\sum_{j=0}^{n-1}\mu^{(1-2\varphi)j}-\sum_{j=0}^{n-1}\mu^{-\varphi j}\right).$$

As long as we exclude that $\varphi = \frac{1}{2}$, we can use eq. (4.25) again to derive

$$\sum_{j=0}^{n-1}\mu^{(1-2\varphi)j} = \frac{1}{\mu^{(1-2\varphi)}-1}\left(\mu^{(1-2\varphi)n}-1\right) \tag{4.28}$$

and

$$\sum_{j=0}^{n-1}\mu^{-\varphi j} = \frac{1}{\mu^{-\varphi}-1}\left(\mu^{-\varphi n}-1\right). \tag{4.29}$$

With eqs. (4.28) and (4.29) and for $\varphi \notin \left\{\frac{1}{2}, 1\right\}$ the third sum can be written as

$$\begin{aligned}
&\frac{\mu^{-2\varphi}}{\mu^{(1-\varphi)}-1}\left(\mu^{(1-\varphi)(k-n+1)}\sum_{j=0}^{n-1}\mu^{(1-2\varphi)j}-\sum_{j=0}^{n-1}\mu^{-\varphi j}\right)\\
=\ &\frac{\mu^{-2\varphi}}{\mu^{(1-\varphi)}-1}\left(\mu^{(1-\varphi)(k-n+1)}\frac{1}{\mu^{(1-2\varphi)}-1}\left(\mu^{(1-2\varphi)n}-1\right)-\frac{1}{\mu^{-\varphi}-1}\left(\mu^{-\varphi n}-1\right)\right)\\
=\ &\frac{\mu^{-2\varphi}}{\mu^{(1-\varphi)}-1}\left[\frac{1}{\mu^{(1-2\varphi)}-1}\left(\mu^{(1-\varphi)(k-n+1)+(1-2\varphi)n}-\mu^{(1-\varphi)(k-n+1)}\right)-\frac{1}{\mu^{-\varphi}-1}(\mu^{-\varphi n}-1)\right]\\
=\ &\frac{\mu^{(1-3\varphi)}}{(\mu^{(1-\varphi)}-1)(\mu^{(1-2\varphi)}-1)}\mu^{(1-\varphi)k}\left(\mu^{-\varphi n}-\mu^{-(1-\varphi)n}\right)-\frac{\mu^{-2\varphi}}{(\mu^{(1-\varphi)}-1)(\mu^{-\varphi}-1)}(\mu^{-\varphi n}-1),
\end{aligned} \tag{4.30}$$

where we used that

$$(1-\varphi)(k-n+1)+(1-2\varphi)n = (1-\varphi)k+(1-\varphi)-\varphi n.$$

The numerator of W_n in the case of a μ-regular tree graph can be calculated with the help of eqs. (4.26), (4.27) and (4.30). For $\varphi \notin \left\{\frac{1}{2}, 1\right\}$ we conclude that

$$\sum_{l=0}^{2k} Z_l(n)\mu^{-\varphi l}$$

$$
\begin{aligned}
&= \sum_{l=0}^{k-n} \mu^{(1-\varphi)l} + \sum_{l=1}^{n} \mu^{-\varphi l} + \sum_{l=2}^{2k}\sum_{j=2}^{n+1} \mu^{(1-\varphi)l}\mu^{-j}\, \mathbb{1}_{\{j\le l\le k-n+2j-2\}} \\
&= \frac{\mu^{(1-\varphi)(k-n+1)}-1}{\mu^{(1-\varphi)}-1} + \mu^{-\varphi}\frac{\mu^{-\varphi n}-1}{\mu^{-\varphi}-1} + \frac{\mu^{(1-3\varphi)}\mu^{(1-\varphi)k}}{(\mu^{(1-\varphi)}-1)(\mu^{(1-2\varphi)}-1)}\left(\mu^{-\varphi n}-\mu^{-(1-\varphi)n}\right) \\
&\quad - \frac{\mu^{-2\varphi}}{(\mu^{(1-\varphi)}-1)(\mu^{-\varphi}-1)}(\mu^{-\varphi n}-1) \\
&= C_1(\varphi,\mu) + \mu^{(1-\varphi)(k-n)}C_2(\varphi,\mu) + \mu^{-\varphi n}\mu^{(1-\varphi)k}C_3(\varphi,\mu) + \mu^{-\varphi n}C_4(\varphi,\mu),
\end{aligned}
$$

where the constants are given by

$$
\begin{aligned}
C_1 = C_1(\varphi,\mu) &= -\frac{1-\mu+\mu^{2\varphi}}{(\mu^{\varphi}-1)(\mu-\mu^{\varphi})} \qquad & C_2 = C_2(\varphi,\mu) &= -\frac{\mu(1-\mu+\mu^{2\varphi})}{(\mu-\mu^{\varphi})(\mu-\mu^{2\varphi})} \\
C_3 = C_3(\varphi,\mu) &= \frac{\mu}{(\mu-\mu^{\varphi})(\mu-\mu^{2\varphi})} \qquad & C_4 = C_4(\varphi,\mu) &= -\frac{-1+\mu-\mu^{-\varphi}}{(\mu^{\varphi}-1)(\mu-\mu^{\varphi})}.
\end{aligned}
$$

To verify this identity, we need to rearrange the parts according to their dependency on n and k and simplify them. The constant terms lead to the value of $C_1(\varphi,\mu)$ through

$$
\begin{aligned}
&-\frac{1}{\mu^{(1-\varphi)}-1} - \frac{\mu^{-\varphi}}{\mu^{-\varphi}-1} + \frac{\mu^{-2\varphi}}{(\mu^{(1-\varphi)}-1)(\mu^{-\varphi}-1)} \\
&= -\frac{\mu^{\varphi}}{\mu-\mu^{\varphi}} + \frac{1}{\mu^{\varphi}-1} - \frac{1}{(\mu-\mu^{\varphi})(\mu^{\varphi}-1)} \\
&= -\frac{(\mu^{\varphi}(\mu^{\varphi}-1)-(\mu-\mu^{\varphi})+1)}{(\mu^{\varphi}-1)(\mu-\mu^{\varphi})} \\
&= -\frac{1-\mu+\mu^{2\varphi}}{(\mu^{\varphi}-1)(\mu-\mu^{\varphi})} = C_1(\varphi,\mu).
\end{aligned}
$$

$C_2(\varphi,\mu)$ collects all terms which depend on $\mu^{-(1-\varphi)(k-n)}$:

$$
\begin{aligned}
\frac{\mu^{(1-\varphi)}}{\mu^{(1-\varphi)}-1} - \frac{\mu^{(1-3\varphi)}}{(\mu^{(1-\varphi)}-1)(\mu^{(1-2\varphi)}-1)} &= \frac{\mu}{(\mu-\mu^{\varphi})} - \frac{\mu}{(\mu-\mu^{\varphi})(\mu-\mu^{2\varphi})} \\
&= -\frac{-\mu(\mu-\mu^{2\varphi})+\mu}{(\mu-\mu^{\varphi})(\mu-\mu^{2\varphi})} \\
&= -\frac{\mu(1-\mu+\mu^{2\varphi})}{(\mu-\mu^{\varphi})(\mu-\mu^{2\varphi})} = C_2(\varphi,\mu).
\end{aligned}
$$

In the same manner, $C_3(\varphi,\mu)$ is the prefactor of $\mu^{(1-\varphi)k-\varphi n}$ and is equal to

$$
\frac{\mu^{(1-3\varphi)}}{(\mu^{(1-\varphi)}-1)(\mu^{(1-2\varphi)}-1)} = \frac{\mu}{(\mu-\mu^{\varphi})(\mu-\mu^{2\varphi})} = C_3(\varphi,\mu).
$$

$C_4(\varphi,\mu)$ collects all terms with $\mu^{-\varphi n}$:

$$\begin{aligned}\frac{\mu^{-\varphi}}{\mu^{-\varphi}-1}-\frac{\mu^{-2\varphi}}{(\mu^{(1-\varphi)}-1)(\mu^{-\varphi}-1)} &= -\frac{1}{\mu^{\varphi}-1}+\frac{1}{(\mu-\mu^{\varphi})(\mu^{\varphi}-1)}\\ &= -\frac{(\mu-\mu^{\varphi})-1}{(\mu^{\varphi}-1)(\mu-\mu^{\varphi})}\\ &= -\frac{-1+\mu-\mu^{-\varphi}}{(\mu^{\varphi}-1)(\mu-\mu^{\varphi})}=C_4(\varphi,\mu).\end{aligned}$$

From the proof of Theorem 4.2 we know that W_n measures how fast the probability to update a neuron at generation n to the wrong bit decreases (see eq. (4.7)). In the following, we analyse the asymptotic behaviour of the terms in W_n for different regimes of n. We distinguish three cases:

(a) n is bounded in k.

(b) n and $k-n$ grow to infinity with k and $\frac{n}{k}\to\alpha\in[0,1]$.

(c) n grows like k meaning that $\frac{n}{k}\to 1$ but $n-k$ is bounded in k.

The aim is to determine how W_n grows with k, and for this we will write C as a representative for all terms which do not depend on k or do not grow with k. The explicit value of C depends on the regime and the values of φ,μ and n. For *b.*) and *c.*) we assume that n is a sequence in k for which $\frac{n}{k}$ converges. Since $0\le n\le k$, the sequence $\frac{n}{k}$ has values in $[0,1]$ and due to the compactness of $[0,1]$ we can always choose a suitable subsequence.

Regime (a):
This case includes the root node and the generations close to the root. For the different values of φ the numerator of W_n behaves asymptotically as follows:

(1) If $0<\varphi<1$, the only term which grows with k is $\mu^{(1-\varphi)k}$. Therefore,

$$C_1(\varphi)+\mu^{(1-\varphi)(k-n)}C_2(\varphi)+\mu^{-\varphi n}\mu^{(1-\varphi)k}C_3(\varphi)+\mu^{-\varphi n}C_4(\varphi)\approx\mu^{(1-\varphi)k}C \tag{4.31}$$

with $C=C_2(\varphi)\mu^{(1-\varphi)n}+C_3(\varphi)\mu^{-\varphi n}$. Since n is bounded in k, the constant C includes terms depending on n.

(2) If $\varphi>1$, the convergence of $\mu^{(1-\varphi)k}$ to zero leads to

$$C_1(\varphi)+\mu^{(1-\varphi)(k-n)}C_2(\varphi)+\mu^{-\varphi n}\mu^{(1-\varphi)k}C_3(\varphi)+\mu^{-\varphi n}C_4(\varphi)\approx C \tag{4.32}$$

with $C=C_1(\varphi)+C_4(\varphi)\mu^{-\varphi n}$.

The results from eqs. (4.31) and (4.32) can be used to derive the asymptotic behaviour of W_n:

(1) If $0 < \varphi < \frac{1}{2}$, both, numerator and denominator, can be handled with eq. (4.31):

$$W_n \approx \frac{\mu^{2(1-\varphi)k}}{\mu^{(1-2\varphi)k}} C = \mu^k C$$

with $C = \frac{\left(C_2(\varphi)+\mu^{(1-\varphi)n}+C_3(\varphi)\mu^{-\varphi n}\right)^2}{C_2(2\varphi)+\mu^{(1-2\varphi)n}+C_3(2\varphi)\mu^{-2\varphi n}}$.

(2) If $\frac{1}{2} < \varphi < 1$, the denominator is converging to a constant (see eq. (4.32)). Thus,

$$W_n \approx \mu^{2(1-\varphi)k} C$$

with $C = \frac{\left(C_2(\varphi)+\mu^{(1-\varphi)n}+C_3(\varphi)\mu^{-\varphi n}\right)^2}{C_1(2\varphi)+C_4(2\varphi)\mu^{-2\varphi n}}$. Because of $\varphi > \frac{1}{2}$ we have $2(1-\varphi) < 1$. Thus, W_n grows with a lower rate than in the case of $0 < \varphi < \frac{1}{2}$.

(3) If $1 < \varphi$, we apply the result of eq. (4.32) to both and get

$$W_n \approx C$$

with $C = \frac{\left(C_1(\varphi)+C_4(\varphi)\mu^{-\varphi n}\right)^2}{C_1(2\varphi)+C_4(2\varphi)\mu^{-2\varphi n}}$.

Regime (b):

This case contains all nodes with an appropriate distance to the root and to the leaf nodes. Since n is growing with k, we assume that $\frac{n}{k}$ converges to $\alpha \in [0, 1]$. Even if $\alpha = 1$, the constraint that $n - k$ tends to infinity states that the distance to the leafs is getting arbitrarily large. This is relevant for the asymptotic behaviour of the numerator and denominator as well as for W_n itself.

(1) If $0 < \varphi < \frac{1}{2}$, the term $\mu^{-\varphi n}$ converges slower to zero than $\mu^{-(1-\varphi)n}$ does, i.e.

$$\frac{\mu^{-(1-\varphi)n}}{\mu^{-\varphi n}} = \mu^{-(1-2\varphi)n} \overset{k\to\infty}{\longrightarrow} 0.$$

Therefore,

$$\begin{aligned} & C_1(\varphi) + \mu^{(1-\varphi)(k-n)} C_2(\varphi) + \mu^{-\varphi n}\mu^{(1-\varphi)k} C_3(\varphi) + \mu^{-\varphi n} C_4(\varphi) \\ = \; & \mu^{(1-\varphi)k-\varphi n} \left(C_1(\varphi)\mu^{-k+\varphi(k+n)} + \mu^{-(1-2\varphi)n} C_2(\varphi) + C_3(\varphi) + \mu^{-(1-\varphi)k} C_4(\varphi) \right). \end{aligned}$$

Since n grows at most like k and $\varphi < \frac{1}{2}$, we know that

$$\mu^{-k+\varphi(k+n)} \overset{k\to\infty}{\longrightarrow} 0.$$

Hence,

$$C_1(\varphi) + \mu^{(1-\varphi)(k-n)} C_2(\varphi) + \mu^{-\varphi n} \mu^{(1-\varphi)k} C_3(\varphi) + \mu^{-\varphi n} C_4(\varphi) \approx \mu^{(1-\varphi)k - \varphi n} C \quad (4.33)$$

with $C = C_3(\varphi)$.

(2) If $\frac{1}{2} < \varphi < 1$, the term $\mu^{-(1-\varphi)n}$ converges slower to zero than $\mu^{-\varphi n}$ does, i.e.

$$\frac{\mu^{-\varphi n}}{\mu^{-(1-\varphi)n}} = \mu^{-(2\varphi-1)n} \overset{k\to\infty}{\longrightarrow} 0.$$

As mentioned before, in this regime $k - n$ grows to infinity. Thus,

$$\begin{aligned} & C_1(\varphi) + \mu^{(1-\varphi)(k-n)} C_2(\varphi) + \mu^{-\varphi n} \mu^{(1-\varphi)k} C_3(\varphi) + \mu^{-\varphi n} C_4(\varphi) \\ = \ & \mu^{(1-\varphi)(k-n)} \left(C_1(\varphi) \mu^{-(1-\varphi)(k-n)} + C_2(\varphi) + \mu^{-(2\varphi-1)n} C_3(\varphi) + \mu^{-(1-\varphi)k-(2\varphi-1)n} C_4(\varphi) \right) \\ \approx \ & \mu^{(1-\varphi)(k-n)} C \end{aligned} \quad (4.34)$$

with $C = C_2(\varphi)$ because

$$\mu^{-(1-\varphi)(k-n)} \overset{k\to\infty}{\longrightarrow} 0$$

and

$$\mu^{-(1-\varphi)k+(1-2\varphi)n} = \mu^{-(1-\varphi)k-(2\varphi-1)n} \overset{k\to\infty}{\longrightarrow} 0.$$

(3) If $1 < \varphi$, the term converges to a constant:

$$C_1 + \mu^{-(1-\varphi)n} \mu^{(1-\varphi)k} C_2 + \mu^{-\varphi n} \mu^{(1-\varphi)k} C_3 + \mu^{-\varphi n} C_4 \approx C \quad (4.35)$$

with $C = C_1(\varphi)$ because n and $k - n$ grow to infinity.

These results lead to the following asymptotic behaviour of W_n:

(1) If $0 < \varphi < \frac{1}{4}$, with eq. (4.33) we conclude that

$$W_n \approx \frac{\mu^{2(1-\varphi)k-2\varphi n}}{\mu^{(1-2\varphi)k-2\varphi n}} C = \mu^k C$$

with $C = \frac{C_3(\varphi)^2}{C_3(2\varphi)}$.

(2) If $\frac{1}{4} < \varphi < \frac{1}{2}$, we use eq. (4.34) to handle the denominator and eq. (4.33) to handle the numerator. Thus,

$$W_n \approx \frac{\mu^{2(1-\varphi)k-2\varphi n}}{\mu^{(1-2\varphi)(k-n)}} C \approx \mu^{k(1-(4\varphi-1)\alpha)} C$$

with $C = \frac{C_3(\varphi)^2}{C_2(2\varphi)}$.

(3) If $\frac{1}{2} < \varphi < 1$, the denominator converges to a constant (see eq. (4.35)) and the numerator can be treated with eq. (4.34):

$$W_n \approx \mu^{2(1-\varphi)k\left(1-\frac{n}{k}\right)} C \approx \mu^{2(1-\varphi)k(1-\alpha)} C$$

with $C = \frac{C_2(\varphi)^2}{C_1(2\varphi)}$.

(4) If $1 < \varphi$, both, denominator and numerator, converge to a constant (see eq. (4.35)). Therefore,

$$W_n \approx C$$

with $C = \frac{C_1(\varphi)^2}{C_1(2\varphi)}$.

Regime (c):
In this case we do not allow the distance between the fixed neuron and the last generation to grow arbitrary large. This regime contains all nodes which are close to the leafs. By the geometric growth of the graph, the majority of nodes can be found within a small number of generations away from the leafs. The results are similar to the results of regime (b), but terms depending on $n-k$ can be considered to be constant or at least to be bounded. We assume $n-k$ to be constant to avoid a distinction of further cases.

(1) If $0 < \varphi < \frac{1}{2}$ and n grows with the same speed as k, then

$$\mu^{(1-\varphi)k-\varphi n} = \mu^{k\left(1-\varphi\left(1+\frac{n}{k}\right)\right)} \overset{k\to\infty}{\longrightarrow} \infty.$$

The asymptotics are the same as in regime (b) and with $\frac{n}{k} \to 1$ we get

$$C_1(\varphi) + \mu^{(1-\varphi)(k-n)} C_2(\varphi) + \mu^{-\varphi n}\mu^{(1-\varphi)k} C_3(\varphi) + \mu^{-\varphi n} C_4(\varphi) \approx \mu^{k(1-2\varphi)} C \tag{4.36}$$

with $C = C_3(\varphi)$.

(2) If $\frac{1}{2} < \varphi$, we see that

$$\mu^{(1-\varphi)k-\varphi n} = \mu^{k\left(1-\varphi\left(1+\frac{n}{k}\right)\right)} \overset{k\to\infty}{\longrightarrow} 0$$

and in contrast to regime (b), $\varphi > \frac{1}{2}$ is enough for this term to converge to a constant:

$$C_1(\varphi) + \mu^{(1-\varphi)(k-n)} C_2(\varphi) + \mu^{-\varphi n}\mu^{(1-\varphi)k} C_3(\varphi) + \mu^{-\varphi n} C_4(\varphi) \approx C \tag{4.37}$$

with $C = C_1(\varphi) + \mu^{(1-\varphi)(k-n)} C_2(\varphi)$. This C does not grow with k because we assumed $n-k$ to be constant.

The results lead to the following asymptotic behaviour of W_n:

(1) If $0 < \varphi < \frac{1}{4}$, we get the same result for W_n as in regime (b):

$$W_n \approx \frac{\mu^{2(1-2\varphi)k}}{\mu^{(1-4\varphi)k}} C = \mu^k C$$

with $C = \frac{C_3(\varphi)^2}{C_3(2\varphi)}$.

(2) If $\frac{1}{4} < \varphi < \frac{1}{2}$, the denominator converges to a constant (see eq. (4.37)). We still have the same growth rate for W_n as in regime (b) but with a different constant:

$$W_n \approx \mu^{2(1-2\varphi)k} C$$

where $C = \frac{C_3(\varphi)^2}{C_1(2\varphi)+\mu^{(1-\varphi)(k-n)}C_2(2\varphi)}$.

(3) If $\frac{1}{2} < \varphi$, the term W_n converges to a constant

$$W_n \approx C$$

with $C = \frac{C_1(\varphi)+\mu^{(1-\varphi)(k-n)}C_2(\varphi)^2}{C_1(2\varphi)+\mu^{(1-2\varphi)(k-n)}C_2(2\varphi)}$.

The lower bound for the storage capacity in the sense of Theorem 4.2 is equal to the minimal value of W_n divided by $c\log(N)$. Here, N is the total number of nodes in the finite graph G_k. The value of $\min_n W_n$ and the generation where the minimum is attained are hard to calculate analytically for all values of φ and μ. But since we are interested in results for large networks, we see that regime (c) leads to the smallest growth in k. Thus, if k is large enough, the storage capacity is determined by the nodes which are close to the leafs.

Since we calculated the asymptotic behaviour of $\min_i W_i$, we proved the following theorem:

Theorem 4.4
Consider a Hopfield model on a μ-regular tree $G = (V, E)$ with dynamics $T^w = (T_i^w)_{i \in V}$ defined as in eq. (4.6) and weights according to eq. (4.23). Then there exist constants $c_0(\varphi), c_1(\varphi), c_2(\varphi)$ such that

$$\min_i W_i \approx \begin{Bmatrix} c_0(\varphi)N, & \text{if } 0 < \varphi < \frac{1}{4} \\ c_1(\varphi)N^{2(1-2\varphi)}, & \text{if } \frac{1}{4} < \varphi < \frac{1}{2} \\ c_2(\varphi), & \text{if } \varphi > \frac{1}{2} \end{Bmatrix} =: \widetilde{W}(\varphi).$$

In particular, for $M = \frac{\widetilde{W}(\varphi)}{c\log(N)}$ *and* $\xi^1, \ldots, \xi^M$ *chosen uniformly at random from* $\{-1,1\}^N$*, it follows that:*

(1) If $c > 2$ *and* $\varphi \leq \frac{1}{2}$*, then for any* $\nu = 1, \ldots, M$

$$\mathbb{P}\left(T^w(\xi^\nu) = \xi^\nu\right) = 1 - R_N$$

with $\lim_{N\to\infty} R_N = 0$.

(2) If $c \geq 4$ *and* $\varphi \leq \frac{1}{2}$*, then*

$$\mathbb{P}\left(\forall \mu \leq M : \; T^w(\xi^\mu) = \xi^\mu\right) = 1 - R_N$$

with $\lim_{N\to\infty} R_N = 0$.

The lower bounds for the storage capacity can be summarized as

Regime (c)	$0 < \varphi < \frac{1}{4}$	$\frac{1}{4} < \varphi < \frac{1}{2}$	$\varphi > \frac{1}{2}$
M	$\frac{N}{c_0 \log(N)}$	$\frac{N^{2(1-2\varphi)}}{c_1 \log(N)}$	0

where c_0, c_1, c_2 are positive constants.

The storage capacity of the net is determined by the minimum of all W_n but we can still do the same calculations for the other two regimes. For regime (a) the results are

Regime (a)	$0 < \varphi < \frac{1}{2}$	$\frac{1}{2} < \varphi < 1$	$\varphi > 1$
$\frac{W_n}{c\log(N)}$	$\frac{N}{\widetilde{c}_0 \log(N)}$	$\frac{N^{2(1-\varphi)}}{\widetilde{c}_1 \log(N)}$	0

where $\widetilde{c}_0, \widetilde{c}_1, \widetilde{c}_2$ are positive constants. The signal strength for nodes in the first part of the graph is stronger because regime (a) would lead to a higher storage capacity than regime (c). The functionality of the neural network in (c) decreases faster resp. collapses earlier than in (a). This is due to the growth rate of the neighbourhood structure. The number of vertices at distance l to the root grows like μ^l whereas the number of vertices at distance l to a leaf only grows with a rate of $\mu^{\frac{l}{2}}$ (see eq. (4.22)).

To complete the picture, the regime (b) leads to

Regime (b)	$0 < \varphi < \frac{1}{4}$	$\frac{1}{4} < \varphi < \frac{1}{2}$	$\frac{1}{2} < \varphi < 1$	$\varphi > 1$
$\frac{W_n}{c\log(N)}$	$\frac{N}{\widehat{c}_0 \log(N)}$	$\frac{N^{1-(4\varphi-1)\alpha}}{\widehat{c}_1 \log(N)}$	$\frac{N^{2(1-\varphi)(1-\alpha)}}{\widehat{c}_2 \log(N)}$	0

where $\widehat{c}_0, \widehat{c}_1, \widehat{c}_2$ are positive constants and α is the limit of $\frac{n}{k}$.

4.4. Conclusions about the inhomogeneous Hopfield model on a deterministic graph

The examples above showed that the most decisive topological quantity that influences the storage capacity is the overall growth rate of the neighbourhood structure. As long as the decrease of the signal strength is sufficiently lower than this growth rate, the model behaves very similar to the standard Hopfield model and reaches a comparable storage capacity. Raising the parameter leads to a loss of storage capacity without immediately destroying the functionality of the net. If the parameter φ exceeds a certain bound, the lower bound for the storage capacity is equal to zero.

Furthermore, in this model each node has a different probability to get updated to the wrong bit because, depending on the graph, they can be exposed to different neighbourhood structures. A good example for this is the regular tree graph. A closer look to eq. (4.22) shows that the growth rate for a leaf is of order $\mu^{\frac{l}{2}}$ and the root has a growth rate of order μ^l. This explains why the storage capacity in regime (a) is higher than in regime (c).

4.5. The Hopfield model with weights on a random graph

In the previous section we successfully derived a lower bound for the storage capacity of two deterministic graphs: the lattice on a torus and the regular tree graph. In these examples the most decisive quantity was the difference between the growth rate of neighbours and the rate at which the weights decline at greater distances. This raises the question how important the regularity of a graph is and if fluctuations in the neighbourhood structure have an impact on the results. To answer this question, we generate a graph with the help of a Galton-Watson process and analyse the Hopfield model on a realization of this random graph. Our aim is to compare the storage capacity which is suggested by the root node of a Galton-Watson tree to the storage capacity in regime (a) of a regular tree graph.

In the following section we introduce the Galton-Watson process and state some results about its asymptotic growth rate. For these results we refer the reader to [AN72] and [Als].

4.5.1. General results about the Galton-Watson process

Definition 4.5 (see Chapter I in [AN72])
Let $(\Lambda, \mathcal{A}, \mathbb{Q})$ be a probability space. A Galton-Watson process with offspring distribution

$(p_j)_{j\geq 0}$ is a Markov-Chain $(Z_n)_{n\in\mathbb{N}}$ on the non-negative integers such that

$$\mathbb{Q}\left(Z_{n+1} = j \mid Z_n = i\right) = p_j^{\star(i)},$$

where $p_j^{\star(i)}$ is the i-fold convolution of $(p_j)_{j\geq 0}$.

In Definition 4.5 the i-fold convolution of $(p_j)_{j\geq 0}$ in the case of $i = 0$ is equal to the Dirac measure in 0.

Let $(Z_n)_{n\in\mathbb{N}}$ be a Galton-Watson process with offspring distribution $(p_j)_{j\geq 0}$ and $Z_0 = 1$. Furthermore, we assume that the expectation value and the variance of $(p_j)_{j\geq 0}$ is finite. Each realization of a Galton-Watson process can be represented as a tree graph (see [Ott49; Jan12]). This graph is infinite if and only if the Galton-Watson process survives, i.e. $Z_n > 0$ for all $n \in \mathbb{N}$. The Hopfield model can be defined on a finite graph. We denote by $G_k^{(3)} = (V_k^{(3)}, E_k^{(3)})$ the graph which is generated by a Galton-Watson process $(Z_n)_{n\in\mathbb{N}}$ up to generation k. Since we are interested in the asymptotic behaviour in the case where the number of nodes grows to infinity, we want the Galton-Watson process to generate an infinite graph. We assume that $p_0 = 0$ to assure this. Additionally, we assume that $p_1 < 1$ because otherwise the graph is just a 1-regular tree graph. The expectation value of the offspring distribution with this assumptions is greater than one:

$$\mu := \mathbb{E}[Z_1] = \sum_{j=1}^{\infty} j\, p_j > \sum_{j=1}^{\infty} p_j = 1.$$

The Galton-Watson process is said to be in the supercritical case.

Lemma 4.6
Let $(Z_n)_{n\in\mathbb{N}}$ be a Galton-Watson process with $\mu < \infty$. Then the process $(Y_n)_{n\in\mathbb{N}}$ defined by

$$Y_n = \frac{Z_n}{\mu^n}$$

is a non-negative martingale and converges almost surely to a random variable Y.

Proof. The process is clearly non-negative. The Markov property leads to

$$\mathbb{E}[Z_{n+m} \mid Z_n = i_n, \ldots, Z_0 = i_0] = \mathbb{E}[Z_{n+m} \mid Z_n = i_n] = i_n\, \mathbb{E}[Z_m \mid Z_0 = 1] = i_n \mu^m.$$

Therefore,

$$\mathbb{E}[Y_{n+m} \mid Y_0, \ldots Y_n] = Y_n \quad \text{a.s.}$$

The existence of Y and the convergence $Y_n \to Y$ a.s. is a consequence of the martingale convergence theorem. □

From Lemma 4.6 we see that $Z_n(\omega)$ grows like $\mu^n Y(\omega)$. This is very helpful for our purposes as long as $Y(\omega) > 0$. Surely, if $(Z_n)_{n\in\mathbb{N}}$ gets extinct, the limit Y needs to be zero. But the opposite implication is not true since Y can be zero just because the scaling factor μ^n is too strong. A result by Kesten and Stigum [KS66] provides a condition which ensures that the events $\{Y = 0\}$ and $\{Z_n \to \infty\}$ almost surely coincide. Denote by q the extinction probability of a Galton-Watson process, i.e. $q = \mathbb{Q}(\lim_n Z_n = 0)$.

Theorem 4.7 (Kesten, Stigum)
Let $(Z_n)_{n\in\mathbb{N}}$ be a supercritical Galton-Watson process with $Z_0 = 1$ and $\mu \in (1,\infty)$. Then the following statements are equivalent

(1) $\mathbb{Q}(Y = 0) = q$

(2) $\mathbb{E}[Y] = 1$

(3) $\mathbb{E}[Z_1 \log(Z_1)] = \sum_{l\geq 1} p_k k \log(k) < \infty.$

The last statement is often called $Z\log(Z)$ condition.

For the proof we refer to [KS66] and [Als]. A probabilistic proof can be found in [LPP95].

Our assumptions eliminate the possibility for the Galton-Watson process to die out, i.e. $q = 0$. Thus, with Theorem 4.7 and the $Z\log(Z)$ condition, it is true that $Y > 0$ a.s. Even if the $Z\log(Z)$ condition is not fulfilled there exists a normalizing sequence such that the limit is zero with probability q. This result goes back to Heyde and Seneta [Sen68; Hey70b; Hey70a]:

Theorem 4.8 (Heyde, Seneta, Part I Theorem 10.3 in [AN72])
Let $(Z_n)_{n\in\mathbb{N}}$ be a supercritical Galton-Watson process with $Z_0 = 1$ and $\mu \in (1,\infty)$. Then, there always exists a sequence of constants $(C_n)_{n\in\mathbb{N}}$ with $C_n \to \infty$ and $C_n^{-1} C_{n-1} \to \mu$ as $n \to \infty$, such that the random variable $Y_n^* := C_n^{-1} Z_n$ converges a.s. to a random variable Y^* with $Q(Y^* > 0) = 1 - q$.

For the proof we refer to [AN72] and [Sen68; Hey70b; Hey70a].

In our context, Theorem 4.7 and Theorem 4.8 show the existence of a scaling sequence such that the limit is a.s. positive. If the $Z\log(Z)$ condition is fulfilled, μ^n is an appropriate choice. If the condition (3) in Theorem 4.7 is not true, there exists a sequence which at least fulfils $C_n^{-1} C_{n-1} \to \mu$ as $n \to \infty$. These results allow us to show the following lemma, which is a modified version of Theorem 10.2(a) in [Als]:

Lemma 4.9
Let Y and Y^* be the limits of $(Y_n)_{n\in\mathbb{N}}$ resp. $(Y_n^*)_{n\in\mathbb{N}}$. For every $\varphi \in [0,1)$

$$\mu^{-(1-\varphi)k} \sum_{l=0}^{k} Z_l \mu^{-\varphi l} \to Y \frac{\mu^{(1-\varphi)}}{\mu^{(1-\varphi)}-1} \quad a.s.$$

and

$$C_k^{-(1-\varphi)} \sum_{l=0}^{k} Z_l C_l^{-\varphi} \to Y^* \frac{\mu^{(1-\varphi)}}{\mu^{(1-\varphi)}-1} \quad a.s.$$

as $k \to \infty$.

Proof. We only prove the second statement because the first one is a consequence of the same and partially easier arguments. Let f be the generating function of Z_1 and f_n the n-th fold composition of f. Let g_n be the inverse function of f_n on the interval $[q,1]$. A proof of Theorem 4.8 (see proof of Theorem 6.1 in [Als]) shows that we can choose $C_n = (1-g_n(s))^{-1}$ for $s \in (q,1)$, $n \geq 0$ as a norming sequence. Additionally, the sequence satisfies

$$\mu_0^{n-j} \leq \frac{1-g_j(s)}{1-g_n(s)} = \frac{C_n}{C_j} \leq \mu^{n-j} \tag{4.38}$$

for all $\mu_0 \in (1,\mu)$ and all $n \geq j \geq J(s,\mu_0)$, where $J(s,\mu_0) \in \mathbb{N}_0$ is chosen appropriately. For $\varepsilon \in (0,\mu-1)$ let $J = J(s,\mu-\varepsilon)$ and define $\tau = \sup\{l \in \mathbb{N} : Y_l^* \geq (1+\varepsilon)Y^*\} \vee J$. The idea is that the first summands until τ are negligible compared to $\mu^{-(1-\varphi)k}$. But τ is chosen in a way such that Y_l for $l \geq \tau$ is close to its limit and $C_k^{-1}C_l$ is close to μ^{j-k}. Therefore the summands above τ are close to a geometric sum. To prove the statement, we split the sum at τ

$$C_k^{-(1-\varphi)} \sum_{l=0}^{k} Z_l C_l^{-\varphi} = \sum_{l=0}^{\tau} Y_l^* \frac{C_l^{(1-\varphi)}}{C_k^{(1-\varphi)}} + \sum_{l=\tau+1}^{k} Y_l^* \frac{C_l^{(1-\varphi)}}{C_k^{(1-\varphi)}}.$$

The first term converges a.s. to zero with $k \to \infty$ because $C_k^{(1-\varphi)} \overset{k\to\infty}{\longrightarrow} \infty$. The inequality in (4.38) leads to an upper bound of the second term, which is given by

$$\begin{aligned}\sum_{l=\tau+1}^{k} Y_l^* \frac{C_l^{(1-\varphi)}}{C_k^{(1-\varphi)}} &\leq (1+\varepsilon)Y^* \sum_{l=\tau+1}^{k} \frac{C_l^{(1-\varphi)}}{C_k^{(1-\varphi)}} \\ &\leq (1+\varepsilon)Y^* \sum_{l=\tau+1}^{k} (\mu-\varepsilon)^{l-k}\end{aligned}$$

$$= (1+\varepsilon)Y^* \frac{(\mu-\varepsilon)^{(1-\varphi)}}{(\mu-\varepsilon)^{(1-\varphi)}-1}\left(1-(\mu-\varepsilon)^{-(1-\varphi)(k-\tau)}\right).$$

For k large enough we showed that for every $\varepsilon \in (0, \mu - 1)$

$$C_k^{-(1-\varphi)} \sum_{l=0}^{k} Z_l C_l^{-\varphi} \leq Y^* \frac{(1+\varepsilon)(\mu-\varepsilon)^{(1-\varphi)}}{(\mu-\varepsilon)^{(1-\varphi)}-1}$$

and conclude that

$$\limsup_{k\to\infty} C_k^{-(1-\varphi)} \sum_{l=0}^{k} Z_l C_l^{-\varphi} \leq Y^* \frac{\mu^{(1-\varphi)}}{\mu^{(1-\varphi)}-1}.$$

Since all summands are positive, we know that for $m \geq 0$

$$\liminf_{k\to\infty} C_k^{-(1-\varphi)} \sum_{l=0}^{k} Z_l C_l^{-\varphi} \geq \liminf_{k\to\infty} \sum_{l=k-m}^{k} Y_l^* \frac{C_l^{(1-\varphi)}}{C_k^{(1-\varphi)}} = \sum_{l=0}^{m} \liminf_{k\to\infty} Y_{k-l}^* \frac{C_{k-l}^{(1-\varphi)}}{C_k^{(1-\varphi)}}.$$

From eq. (4.38) we know that

$$\frac{C_{k-l}}{C_k} = \frac{C_{k-l}}{C_{k-l+1}} \cdot \ldots \cdot \frac{C_{k-1}}{C_k} \geq \mu^{-l}$$

for k large enough and together with $Y_{k-l}^* \overset{k\to\infty}{\longrightarrow} Y^*$ it follows that

$$\liminf_{k\to\infty} C_k^{-(1-\varphi)} \sum_{l=0}^{k} Z_l C_l^{-\varphi} \geq Y^* \sum_{l=0}^{m} \mu^{-l(1-\varphi)} \tag{4.39}$$

for all $m \geq 0$. Because of

$$\sum_{l=0}^{\infty} \mu^{-l(1-\varphi)} = \frac{1}{1-\mu^{-(1-\varphi)}} = \frac{\mu^{(1-\varphi)}}{\mu^{(1-\varphi)}-1}$$

and eq. (4.39) for all $m \geq 0$, we showed that

$$\liminf_{k\to\infty} C_k^{-(1-\varphi)} \sum_{l=0}^{k} Z_l C_l^{-\varphi} \geq Y^* \frac{\mu^{(1-\varphi)}}{\mu^{(1-\varphi)}-1}$$

which completes the proof. □

For our purpose Lemma 4.9 has two applications. If we set $\varphi = 0$, the lemma gives a statement about the asymptotic behaviour of the total number of nodes in a Galton-

Watson tree:

$$N \approx Y^* C_k \frac{\mu}{\mu - 1} \quad \text{a.s.} \tag{4.40}$$

Furthermore, for an arbitrary $\varphi \in [0, 1)$ the lemma can be used to calculate the asymptotic behaviour of W_i.

4.5.2. The inhomogeneous Hopfield model on a Galton-Watson tree

Let $(Z_n)_{n\in\mathbb{N}}$ be a Galton-Watson process with the assumptions made in Section 4.5.1 and $G_k^{(3)} = (V_k^{(3)}, E_k^{(3)})$ be the corresponding random graph up to generation k. The Galton-Watson process represents the neighbourhood structure from the perspective of the root. The previous example (see Section 4.3.2) showed that the number of neighbours with a specific distance, and therefore the value of W_n for an arbitrary generation n, can be very difficult to calculate. In a Galton-Watson tree it is not true that every neuron in one generation leads to the same value for W_i. In this section we determine the asymptotic behaviour of the storage capacity from the perspective of the root node similar to the regime (a) of the previous example. In general, the storage capacity in the sense of Theorem 4.2 for an arbitrary generation is expected to be worse. The same behaviour was observable in the case of a regular tree graph.

The random graph does not have a fixed number of descendants for each generation. From Theorem 4.7 and Theorem 4.8 we learned that the best approximation for the growth rate is given by μ^k resp. C_k depending on whether the $Z \log(Z)$ condition is fulfilled or not. If we keep in mind that in the first case C_k can be set to μ^k, we define the weights for the Hopfield model as follows: For $\varphi > 0$ and nodes i, j with $d_g(i, j) = l \geq 0$ we set w_{ij} equal to

$$w_l := w_{ij} = C_l^{-\varphi}, \tag{4.41}$$

where $(C_n)_{n\in\mathbb{N}}$ is the sequence from the proof of Lemma 4.9. This definition coincides with eq. (4.23) but instead of the deterministic number of descendants we are using the corresponding asymptotic growth rate C_l. All in all, the setting is quite similar to the regular tree with the difference that the amount of nodes in a generation is a random variable. With this, we observe random fluctuations around the expected number of descendants in every generation. This difficulty can be handled with the results from Section 4.5.1.

We are interested in the asymptotic behaviour of

$$W_0 = \frac{\left(\sum\limits_{l=0}^{k} Z_l\, C_l^{-\varphi}\right)^2}{\sum\limits_{l=0}^{k} Z_l\, C_l^{-2\varphi}} = \frac{\left(\sum\limits_{l=0}^{k} Y_l^*\, C_l^{(1-\varphi)l}\right)^2}{\sum\limits_{l=0}^{k} Y_l^*\, C_l^{(1-2\varphi)l}}$$

for $k \to \infty$. If $0 < \varphi < 1$, the convergence from $(Y_l^*)_{l\in\mathbb{N}}$ to Y^*, combined with the fact that the number of nodes in the last generations dominates all former generations, can be used to derive the asymptotics of the numerator. This was done in Lemma 4.9, which states that

$$\sum_{l=0}^{k} Z_l\, C_l^{-\varphi} \approx Y^* C_k^{(1-\varphi)} \frac{\mu^{(1-\varphi)}}{\mu^{(1-\varphi)} - 1} \quad \text{a.s.} \tag{4.42}$$

Therefore, if $0 < \varphi < \frac{1}{2}$, we get

$$\begin{aligned} W_0 = \frac{\left(\sum\limits_{l=0}^{k} Z_l\, C_l^{-\varphi}\right)^2}{\sum\limits_{l=0}^{k} Z_l\, C_l^{-2\varphi}} &\approx \frac{\left(Y^* C_k^{(1-\varphi)}\right)^2}{Y^* C_k^{(1-2\varphi)}} \frac{\mu^{2(1-\varphi)}\left(\mu^{(1-2\varphi)} - 1\right)}{\mu^{(1-2\varphi)}\left(\mu^{(1-\varphi)} - 1\right)^2} \\ &= Y^* C_k \frac{\mu^{2(1-\varphi)}\left(\mu^{(1-2\varphi)} - 1\right)}{\mu^{(1-2\varphi)}\left(\mu^{(1-\varphi)} - 1\right)^2} = Y^*\, C_k\, C(\varphi, \mu) \quad \text{a.s.} \end{aligned}$$

with $C(\varphi,\mu) = \frac{\mu^{2(1-\varphi)}\left(\mu^{(1-2\varphi)}-1\right)}{\mu^{(1-2\varphi)}\left(\mu^{(1-\varphi)}-1\right)^2}$. Together with the representation of the total number of nodes from eq. (4.40) we can write

$$W_0 \approx N\widetilde{C}(\varphi, \mu)$$

with $\widetilde{C}(\varphi,\mu) = \frac{\mu-1}{\mu} C(\varphi,\mu)$.

A different behaviour can be observed for larger values of φ. For $\varphi > 1$ the numerator of W_0 converges to finite random variable:

$$\sum_{l=0}^{k} Z_l\, C_l^{-\varphi} \overset{k\to\infty}{\longrightarrow} \sum_{l=0}^{\infty} Z_l\, C_l^{-\varphi} =: S_\infty(\varphi) \quad \text{a.s.} \tag{4.43}$$

This random variable strongly depends on the fluctuations in the first generations. For $\varepsilon \in (0, \mu - 1)$ define $\tau = \sup\{l \in \mathbb{N} : Y_l^* \notin [(1-\varepsilon)Y^*,\ (1+\varepsilon)Y^*]\} \vee J$ with J as in the proof of Lemma 4.9. By the convergence of $(Y_l^*)_{l\in\mathbb{N}}$ (see Theorem 4.8) we know that τ is a.s. finite. If we write

$$S_\infty(\varphi) = \sum_{l=0}^{\infty} Z_l \, C_l^{-\varphi} = \sum_{l=0}^{\tau} Y_l^* C_l^{(1-\varphi)} + \sum_{l=\tau+1}^{\infty} Y_l^* C_l^{(1-\varphi)},$$

the first sum is finite and does not grow with k. In the second sum Y_l^* is close to Y^* and C_l can be bounded as in eq. (4.38) since $l \geq \tau$. An upper bound is given by

$$\begin{aligned}
\sum_{l=\tau+1}^{\infty} Y_l^* C_l^{(1-\varphi)} &\leq Y^*(1+\varepsilon) \sum_{l=\tau+1}^{\infty} C_l^{(1-\varphi)} \\
&= Y^*(1+\varepsilon) C_{\tau+1}^{(1-\varphi)} \sum_{l=\tau+1}^{\infty} \left(\frac{C_{l-1}}{C_l} \cdot \ldots \cdot \frac{C_{\tau+1}}{C_{\tau+2}} \right)^{\varphi-1} \\
&\leq Y^*(1+\varepsilon) C_{\tau+1}^{(1-\varphi)} \sum_{l=0}^{\infty} (\mu-\varepsilon)^{(1-\varphi)l} \\
&\leq Y^*(1+\varepsilon) C_{\tau+1}^{(1-\varphi)} \frac{(\mu-\varepsilon)^{(1-\varphi)}}{(\mu-\varepsilon)^{(1-\varphi)}-1},
\end{aligned}$$

and with the same calculations the term is bounded from below by

$$\sum_{l=\tau+1}^{\infty} Y_l^* C_l^{(1-\varphi)} \geq Y^*(1-\varepsilon) C_{\tau+1}^{(1-\varphi)} \frac{\mu^{(1-\varphi)}}{\mu^{(1-\varphi)}-1}.$$

If $\frac{1}{2} < \varphi < 1$, applying eq. (4.42) to the numerator and eq. (4.43) to the denominator leads to

$$W_0 \approx C_k^{2(1-\varphi)} \frac{\left(Y^* \mu^{(1-\varphi)}\right)^2}{S_\infty(2\varphi) \left(\mu^{(1-\varphi)}-1\right)^2},$$

and by eq. (4.40) the term can represented through N as

$$W_0 \approx N^{2(1-\varphi)} \frac{Y^* \mu^{2(1-\varphi)}(\mu-1)}{S_\infty(2\varphi)\mu \left(\mu^{(1-\varphi)}-1\right)^2} \quad \text{a.s.}$$

In the last case, which is $\varphi > 1$, W_0 converges a.s. to a finite random variable. Thus,

$$W \approx \frac{S_\infty(\varphi)^2}{S_\infty(2\varphi)} \quad \text{a.s.}$$

All in all, the calculations of W_0 would lead to a bound for the storage capacity in sense of Theorem 4.2 as follows

	$0 < \varphi < \frac{1}{2}$	$\frac{1}{2} < \varphi < 1$	$1 < \varphi$
$\frac{W_0}{c \log(N)}$	$\frac{N}{C_0 \log(N)}$	$\frac{N^{2(1-\varphi)}}{C_1 \log(N)}$	0

where C_0, C_1, C_2 are positive random variables not depending on N.

4.6. Conclusions about the inhomogeneous Hopfield model on a Galton-Watson tree

Similar to the deterministic graphs, we identified a regime where the signal strength decreases slow enough to obtain results, which would lead to the storage capacity of the standard Hopfield model. Except from the fact that the total number of nodes is random, and therefore the storage capacity contains randomness, we basically got the same results as in the regular tree graph and regime (a). The fluctuations of the random graph make a difference for larger values of φ because the constants mainly depend on the realization of the first generations. But since values of φ greater than one do not lead to a functioning neural network, this does not play a very import role. In the case of a Galton-Watson tree, we only analysed the value of W_0. Thus, to state a theorem similar to Theorem 4.3 or Theorem 4.4 one needs to evaluate W_i for an arbitrary neuron i. Especially neurons close to the leafs would be interesting, since they determine the storage capacity in a μ-regular tree.

A. Appendix

A.1. Large Deviation Theory

Theorem A.1 (Cramér, see Chapter 2.2 in [DZ98])
Let $X_1, X_2, \dots$ be independent and identical distributed (i.i.d.) real-valued random variables. Denote by μ_n the pushforward measure of $\frac{1}{n}S_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ and define Λ^* as the Legendre transform of the cumulant generating function:

$$\Lambda^*(x) := \sup_{\lambda\in\mathbb{R}} \left\{\lambda x - \log\left(\mathbb{E}\left[e^{\lambda X_1}\right]\right)\right\}.$$

The sequence $\{\mu_n\}_n$ satisfies a Large Deviation Principle (LDP) with the convex rate function $\Lambda^*(\cdot)$, namely:

(1) For any closed set $F \subseteq \mathbb{R}$,

$$\limsup_{n\to\infty} \frac{1}{n} \log\left(\mu_n(F)\right) \leq -\inf_{x\in F} \Lambda^*(x).$$

(2) For any open set $G \subseteq \mathbb{R}$,

$$\liminf_{n\to\infty} \frac{1}{n} \log\left(\mu_n(G)\right) \geq -\inf_{x\in G} \Lambda^*(x).$$

Example A.2
We are interested in an application of Theorem A.1 in the following two settings:

(1) Let $X_1, X_2, \dots$ be i.i.d. $\mathrm{Ber}(p)$-distributed random variables for $p \in [0,1]$. Then the rate function is

$$\Lambda^*(x) = x\log\left(\frac{x}{p}\right) + (1-x)\log\left(\frac{1-x}{1-p}\right).$$

Proof. First we calculate the cumulant generating function:

$$\log\left(\mathbb{E}\left[e^{\lambda X_1}\right]\right) = \log\left(e^{\lambda}p + (1-p)\right).$$

Then simple analysis shows that $\lambda x - \log\left(\mathbb{E}\left[e^{\lambda X_1}\right]\right)$ *is maximal at*

$$\lambda^* = \log\left(\frac{1-p}{p}\frac{x}{1-x}\right)$$

Together,

$$\begin{aligned}\Lambda^*(x) &= x\log\left(\frac{1-p}{p}\frac{x}{1-x}\right) - \log\left((1-p)\frac{x}{1-x} + (1-p)\right)\\ &= x\log\left(\frac{x}{p}\right) + (1-x)\log\left(\frac{1-x}{1-p}\right).\end{aligned}$$

(2) Let $Y_1, Y_2, \dots$ *be i.i.d. random variables with* $\mathbb{P}(Y_1 = 1) = p = 1 - \mathbb{P}(Y_1 = -1)$ *for* $p \in [0,1]$*. Then the rate function is*

$$\Lambda^*(x) = \frac{1}{2}\left((1+x)\log\left(\frac{(1+x)}{2p}\right) + (1-x)\log\left(\frac{1-x}{2(1-p)}\right)\right),$$

especially if $p = \frac{1}{2}$

$$\Lambda^*(x) = \frac{1}{2}\left((1+x)\log(1+x) + (1-x)\log(1-x)\right).$$

This is a simple consequence of the example in (a) and the transformation $X_i = \frac{1}{2}(1+Y_i)$*.*

Lemma A.3
Let S_n *be a sum of i.i.d. random variables* X_i *with* $\mathbb{P}(X_1 = 1) = p = 1 - \mathbb{P}(X_1 = -1)$*. Let* Λ^* *be as in Theorem A.1. Then for* $x > 2p - 1$*, we have:*

$$\mathbb{P}(S_n \geq nx) \leq \exp(-n\Lambda^*(x)), \tag{A.1}$$

where $\Lambda^*(x) = \frac{1}{2}\left((1+x)\log\left(\frac{(1+x)}{2p}\right) + (1-x)\log\left(\frac{1-x}{2(1-p)}\right)\right)$*.*

Proof. The exponential Chebyshev inequality with $t > 0$ leads to

$$\mathbb{P}(S_n \geq nx) \leq \exp(-tnx)\mathbb{E}\left[\exp(tS_n)\right] = \exp\left(-n\left(tx - \log\left(\mathbb{E}\left[e^{\lambda X_1}\right]\right)\right)\right).$$

Since this is true for all $t > 0$, we know that

$$\mathbb{P}(S_n \geq nx) \leq \exp\left(-n\sup_{t>0}\left\{tx - \log\left(\mathbb{E}\left[e^{tX_1}\right]\right)\right\}\right).$$

For $x > 2p - 1$ the eq. (A.1) follows with

$$\exp\big(- n \sup_{t>0} \big\{ tx - \log\big(\mathbb{E}\left[e^{tX_1}\right]\big) \big\} \big) = \exp\big(- n\Lambda^*(x)\big),$$

which is true because the term has its maximum at

$$t = \frac{1}{2} \log\Big(\frac{1+x}{1-x} \frac{1-p}{p} \Big).$$

This value is positive if $x > 2p - 1$. □

Lemma A.4
Let S_n be a binomially distributed random variable with parameters n and p. Then for $\varepsilon > 0$, we have:

$$\mathbb{P}\left(S_n \geq n(p+\varepsilon)\right) \leq \exp\left(-n \frac{\varepsilon^2}{(2p+\varepsilon)} \right)$$

Proof. The exponential Chebyshev inequality provides an upper bound for the probability of a tail event:

$$\mathbb{P}\left(S_n \geq n(p+\varepsilon)\right) \leq \exp(-tn(p+\varepsilon))\, \mathbb{E}[\exp(tS_n)].$$

The moment generating function can be bounded with the help of the inequality $1 + x \leq \exp(x)$ for all $x \in \mathbb{R}$. Thus,

$$\mathbb{E}[\exp(tS_n)] = (e^t p + (1-p))^n = (1 + p(e^t - 1))^n \leq \exp(np(e^t - 1)).$$

Together with $t = \log(1 + \frac{\varepsilon}{p})$, we know that the probability is bounded from above by

$$\begin{aligned} \mathbb{P}\left(S_n \geq n(p+\varepsilon)\right) &\leq \exp(-tn(p+\varepsilon) + np(e^t - 1)) \\ &= \exp\Big(- \log\big(1 + \frac{\varepsilon}{p}\big) n(p+\varepsilon) + n\varepsilon \Big). \end{aligned}$$

The logarithm can be bounded from below by

$$\frac{x}{1 + \frac{x}{2}} \leq \log(1 + x)$$

for all $x \geq 0$. Altogether, this leads to

$$\begin{aligned} \mathbb{P}\left(S_n \geq n(p+\varepsilon)\right) &\leq \exp\Big(- \log\big(1 + \frac{\varepsilon}{p}\big) n(p+\varepsilon) + n\varepsilon \Big) \\ &\leq \exp\Big(- \frac{\varepsilon}{p + \frac{\varepsilon}{2}} n(p+\varepsilon) + n\varepsilon \Big) \end{aligned}$$

$$= \exp\Big(-n\frac{\varepsilon^2}{2p+\varepsilon}\Big).$$

□

Lemma A.5
Let S_n be a sum of i.i.d. random variables X_i with $\mathbb{P}(X_1 = 1) = \frac{1}{2} = 1 - \mathbb{P}(X_1 = -1)$. Then for any $a > 0$

$$\mathbb{P}(S_n \geq a) \leq \exp\left(-\frac{a^2}{2n}\right).$$

Proof. Let $a > 0$ and define $t = \frac{a}{n} > 0$. With the exponential Chebyshev inequality, it follows that

$$\mathbb{P}(S_n \geq a) \leq \exp(-ta)\mathbb{E}\left[\exp(tS_n)\right] = \exp(-ta)\cosh(t)^n.$$

Here, $\cosh(x) = \frac{1}{2}(e^x + e^{-x})$. Since $\cosh(t) \leq \exp(\frac{t^2}{2})$ for $t > 0$, we conclude that

$$\mathbb{P}(S_n \geq a) \leq \exp\left(-ta + \frac{t^2}{2}n\right) = \exp\left(-\frac{a^2}{2n}\right)$$

because of $t = \frac{a}{n}$. □

A.2. Weak Convergence

Theorem A.6 (see Theorem 3.4 in [Bil99])
Let X be a random variable and $(X_n)_{n\in\mathbb{N}}$ be a family of random variables. If $X_n \Rightarrow X$, i.e. X_n converges weakly to X, then $\mathbb{E}|X| \leq \liminf_n \mathbb{E}[|X_n|]$.

Proof. By the continuous mapping theorem we know that $|X_n| \Rightarrow |X|$. Therefore, $\mathbb{P}(|X_n| > t) \overset{n\to\infty}{\longrightarrow} \mathbb{P}(|X| > t)$ for all but countably many $t \geq 0$. With Fatou's Lemma we get

$$\begin{aligned}\mathbb{E}[|X|] = \int_0^\infty \mathbb{P}(|X| > t)\, dt = \int_0^\infty \lim_{n\to\infty} \mathbb{P}(|X_n| > t)\, dt \\ \leq \liminf_{n\to\infty} \int_0^\infty \mathbb{P}(|X_n| > t)\, dt = \liminf_{n\to\infty} \mathbb{E}[|X_n|].\end{aligned}$$

□

Definition A.7
A family of random variables $\{X_i : i \in I\}$ *satisfying*

$$\sup_{i\in I} \int_{\{|X_i|\geq\alpha\}} |X_i| \, d\mathbb{P} \overset{\alpha\to\infty}{\longrightarrow} 0.$$

is called uniformly integrable.

Corollary A.8
A family of uniformly integrable random variables has bounded expectation values.

Proof. If α is large enough such that $\sup_{i\in I} \int_{\{|X_i|\geq\alpha\}} |X_i| \, d\mathbb{P} \leq 1$, then

$$\sup_{i\in I} \mathbb{E}|X_n| \leq 1 + \alpha < \infty$$

□

Theorem A.9 (see Theorem 3.5 in [Bil99], Theorem A.8.6 in [Ell85])
Suppose that $(X_n)_{n\in\mathbb{N}}$ *converges weakly to a random variable* X *and let* $h : \mathbb{R} \to \mathbb{R}$ *be a continuous function such that* $(h(X_n))_{n\in\mathbb{N}}$ *is uniformly integrable then*

$$\mathbb{E}_n[h(X_n)] \overset{n\to\infty}{\longrightarrow} \mathbb{E}[h(X)].$$

Proof. h is a continuous function. Thus, with $X_n \Rightarrow X$, we deduce $h(X_n) \Rightarrow h(X)$ by the continuous mapping theorem. With Theorem A.6 we know that $h(X)$ is integrable because $(h(X_n))_{n\in\mathbb{N}}$ are uniformly integrable. Therefore, $\sup_{n\in\mathbb{N}} \mathbb{E}[|h(X_n)|] < \infty$.

With the continuous mapping theorem, we know that $h(X_n)^+ \Rightarrow h(X)^+$ and $h(X_n)^- \Rightarrow h(X)^-$. Thus, we can assume without loss of generality that $(h(X_n))_{n\in\mathbb{N}}$ and $h(X)$ are non-negative.

We show that for every $\varepsilon > 0$ there exists a $N \in \mathbb{N}$ such that $|\mathbb{E}[h(X_n)] - \mathbb{E}[h(X)]| \leq \varepsilon$ for all $n \geq N$. We know that $(h(X_n))_{n\in\mathbb{N}}$ and $h(X)$ are uniformly integrable so there exists a $\alpha > 0$ such that for all $n \in \mathbb{N}$

$$\mathbb{E}[h(X_n) \, \mathbb{1}_{\{h(X_n)\geq\alpha\}}] \leq \frac{\varepsilon}{3} \quad \text{and} \quad \mathbb{E}[h(X) \, \mathbb{1}_{\{h(X)\geq\alpha\}}] \leq \frac{\varepsilon}{3}.$$

Therefore, we get the following inequality

$$\begin{aligned} \Big|\mathbb{E}[h(X_n)] - \mathbb{E}[h(X)]\Big| &\leq \Big|\mathbb{E}[h(X_n) \, \mathbb{1}_{\{h(X_n)<\alpha\}}] - \mathbb{E}[h(X) \, \mathbb{1}_{\{h(X)<\alpha\}}]\Big| \\ &\quad + \Big|\mathbb{E}[h(X_n) \, \mathbb{1}_{\{h(X_n)\geq\alpha\}}]\Big| + \Big|\mathbb{E}[h(X) \, \mathbb{1}_{\{h(X)\geq\alpha\}}]\Big| \\ &\leq \Big|\mathbb{E}[h(X_n) \, \mathbb{1}_{\{h(X_n)<\alpha\}}] - \mathbb{E}[h(X) \, \mathbb{1}_{\{h(X)<\alpha\}}]\Big| + \frac{2}{3}\varepsilon \end{aligned}$$

With the identities

$$\mathbb{E}[h(X_n)\,\mathbb{1}_{\{h(X_n)<\alpha\}}\,] = \int_0^\infty \mathbb{P}(h(X_n)\,\mathbb{1}_{\{h(X_n)<\alpha\}} > t)\;dt$$
$$= \int_0^\alpha \mathbb{P}(t < h(X_n) < \alpha)\;dt$$

and

$$\mathbb{E}[h(X)\,\mathbb{1}_{\{h(X)<\alpha\}}\,] = \int_0^\alpha \mathbb{P}(t < h(X) < \alpha)\;dt$$

the last part of the proof is completed by using $h(X_n) \Rightarrow h(X)$ together with dominated convergence theorem and the fact that α can be chosen such that $\mathbb{P}(h(X) = \alpha) = 0$. We have $\mathbb{1}_{[0,\alpha]}$ as integrable dominating function. Thus, with dominated convergence

$$\begin{aligned}
&\lim_{n\to\infty} \left|\mathbb{E}[h(X_n)\,\mathbb{1}_{\{h(X_n)<\alpha\}}\,] - \mathbb{E}[h(X)\,\mathbb{1}_{\{h(X)<\alpha\}}\,]\right| \\
=& \left|\int_0^\alpha \lim_{n\to\infty} \left(\mathbb{P}(t < h(X_n) < \alpha) - \mathbb{P}(t < h(X) < \alpha)\right) dt\right| \\
\leq& \int_0^\alpha \lim_{n\to\infty} \left(|\mathbb{P}(h(X_n) \leq \alpha) - \mathbb{P}(h(X) \leq \alpha)| + |\mathbb{P}(h(X_n) \leq t) - \mathbb{P}(h(X) \leq t)|\right) dt = 0
\end{aligned}$$

because the weak convergence is equivalent to the convergence of the distribution function for all continuous points.

□

A.3. Analysis

Theorem A.10 (Taylor series, see Theorem 1 in Chapter 22 in [For11])
Let $f : I \to \mathbb{R}$ be a $(k+1)$-times continuously differentiable function and $a \in I$. Then for all $x \in I$

$$f(x) = \sum_{l=0}^{k} \frac{f^{(l)}(a)}{l!}(x-a)^l + R_{k+1}(x)$$

where

$$R_{k+1} = \frac{1}{k!}\int_a^x (x-t)^k f^{(k+1)}(t)\;dt$$

and $f^{(l)}$ denotes the lth derivative of f.

Lemma A.11
Let $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the ℓ_1-norm resp. ℓ_2-norm on $\mathbb{R}^n$. For $x \in \mathbb{R}^n$

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$$

Proof. Let $x \in \mathbb{R}^n$. Since $|x_i| \cdot |x_j| \geq 0$, we conclude that

$$\|x\|_1^2 = \Big(\sum_{i=1}^n |x_i|\Big)^2 = \sum_{i,j=1}^n |x_i| \cdot |x_j| \geq \sum_{i=1}^n |x_i|^2 = \|x\|_2^2.$$

Taking the square root on both sides proves the first inequality.

By using the Cauchy-Schwarz inequality, we know that

$$\|x\|_1 = \sum_{i=1}^n |x_i| \cdot 1 \leq \sqrt{\sum_{i=1}^n x_i^2}\sqrt{\sum_{i=1}^n 1^2} = \sqrt{n}\|x\|_2.$$

This proves the second inequality. □

A.4. Auxiliary Results

Lemma A.12 (see Proposition 5.16 in [Kir15])
Let F_β be the function in eq. (3.11):

$$F_\beta(t) = \frac{1}{\beta}\left(\frac{1}{2}\log\left(\frac{1+t}{1-t}\right)\right)^2 + \log\left(1-t^2\right).$$

(1) If $\beta < 1$, then F_β has a unique minimum at $t = 0$. Furthermore, $F'(0) = 0$ and $F''(0) = 2(\frac{1}{\beta} - 1) > 0$.

(2) If $\beta = 1$, then F_β has a unique minimum at $t = 0$ with $F'(0) = F''(0) = F'''(0) = 0$ and $F^{(iv)}(0) = 4 > 0$.

(3) For $\beta > 0$ the function F_β has a unique minimum in $[0, 1)$ at $t_0 > 0$ and a unique minimum in $(-1, 0]$ at $-t_0 < 0$. $F_\beta'(t_0) = F_\beta'(-t_0) = 0$ and $F_\beta''(t_0) = F_\beta''(-t_0) > 0$. t_0 is the unique strictly positive solution of $t = \tanh(\beta t)$. F_β has a local maximum at $t = 0$ with $F_\beta'(0) = 0$ and $F_\beta''(0) = 2\left(\frac{1}{\beta} - 1\right) < 0$.

Proof. The derivative of F_β is equal to

$$F_\beta'(t) = \frac{2}{\beta}\Big(\frac{1}{2}\log\Big(\frac{1+t}{1-t}\Big)\Big)\frac{1}{2}\frac{1-t}{1+t}\frac{2}{(1-t)^2} - \frac{2t}{1-t^2}$$

$$= \frac{1}{1-t^2}\Big(\frac{1}{\beta}\log\Big(\frac{1+t}{1-t}\Big) - 2t\Big),$$

where we used that for $g(t) := \frac{1+t}{1-t}$

$$g'(t) = \frac{2}{(1-t)^2}.$$

The second derivative of F_β is equal to

$$\begin{aligned} F''_\beta(t) &= 2\frac{1}{\beta}\frac{1}{(1-t^2)^2}\Big(1 + t\log\Big(\frac{1+t}{1-t}\Big)\Big) - 2\frac{1+t^2}{(1-t^2)^2} \\ &= \frac{2}{(1-t^2)^2}\Big[\frac{1}{\beta} - (1+t^2) + \frac{1}{\beta}t\log\Big(\frac{1+t}{1-t}\Big)\Big]. \end{aligned}$$

With the identity

$$\frac{1}{2}\log\Big(\frac{1+t}{1-t}\Big) = \tanh^{-1}(t),$$

we see that

$$F'_\beta(t) = \frac{2}{1-t^2}\Big(\frac{1}{\beta}\tanh^{-1}(t) - t\Big)$$

and

$$F''_\beta(t) = \frac{2}{(1-t^2)^2}\Big[\frac{1}{\beta} - (1+t^2) + \frac{2t}{\beta}\tanh^{-1}(t)\Big]. \tag{A.2}$$

Thus, it follows that $F'_\beta(t) \gtreqless 0$ is equivalent to

$$t \gtreqless \tanh(\beta t).$$

The critical points ($F'_\beta(t) = 0$) are the same as the solutions to the equation

$$\tanh(\beta t) = t. \tag{A.3}$$

For all β a solution of eq. (A.3) is given by $t = 0$. The second derivative at this point is equal to

$$F''_\beta(0) = 2\Big(\frac{1}{\beta} - 1\Big)\begin{cases} > 0, & \text{if } \beta < 1 \\ = 0, & \text{if } \beta = 1 \\ < 0, & \text{if } \beta > 1 \end{cases}.$$

If $\beta < 1$ then $t > \tanh(\beta t)$ for all $t \in (0,1)$. Thus, the derivative $F'_\beta(t)$ is positive and $F_\beta(t)$ is strictly monotone on $(0,1)$. Together with the symmetry of F_β, i.e. $F_\beta(t) = F_\beta(-t)$, this shows that for $\beta < 1$ F_β has a unique minimum at $t = 0$.

In the case $\beta > 1$ the function F_β has a local maximum at $t = 0$. Let $f(t) = \tanh(\beta t)$ then $f(0) = 0$, $f'(0) = \beta > 1$ and $f(t) \to 1$. Let $g(t) = t$ then $g(0) = 0$, $g'(0) = 1$ and $g(t) \to \infty$. Therefore, there exists a positive solution $t_0 \in (0,1)$ to eq. (A.3). The statement $t_0 < 1$ is a consequence of $\tanh(\beta t) < 1$ for all t. Since

$$f'(t) = \frac{\beta}{\cosh(\beta t)^2}$$

is decreasing the solution t_0 is unique. Because of $\beta t_0 = \tanh(t_0)$, we know from eq. (A.2) that

$$\begin{aligned} F''_\beta(t_0) &= \frac{2}{(1-t_0^2)^2}\left[\frac{1}{\beta} - (1+t_0^2) + \frac{2t_0}{\beta}\beta t_0\right] \\ &= \frac{2}{(1-t_0^2)^2}\left[\frac{1}{\beta} - 1 + t_0^2\right] > 0. \end{aligned}$$

Due to the symmetry of F_β the same statements are true for $-t_0$.

For $\beta = 1$ a straightforward calculation leads to

$$F'''_\beta(t) = \frac{4t^3 - 2\left(3t^2+1\right)\log\left(\frac{1+t}{1-t}\right)}{\left(t^2-1\right)^3}$$

and

$$F^{(iv)}(t) = \frac{4\left(-3t^4 + 6\left(t^3+t\right)\log\left(\frac{1+t}{1-t}\right)+1\right)}{\left(t^2-1\right)^4}.$$

□

References

[Agl+15a] E. Agliari, A. Barra, A. Galluzzi, F. Guerra, D. Tantari, and F. Tavani. "Hierarchical neural networks perform both serial and parallel processing". In: *Neural Networks* 66 (June 2015), pp. 22–35.

[Agl+15b] E. Agliari, A. Barra, A. Galluzzi, F. Guerra, D. Tantari, and F. Tavani. "Retrieval Capabilities of Hierarchical Networks: From Dyson to Hopfield". In: *Physical Review Letters* 114.2 (Jan. 2015).

[Ald85] D. J. Aldous. "Exchangeability and related topics". In: *Lecture Notes in Mathematics*. Vol. 1117. Lecture Notes in Math. Springer Berlin Heidelberg, 1985, pp. 1–198.

[Als] G. Alsmeyer. "Kapitel I - Der einfache Galton-Watson-Prozeß". URL: `https://www.uni-muenster.de/Stochastik/alsmeyer/Skripten/GWP01.pdf`.

[AGS85a] D. J. Amit, H. Gutfreund, and H. Sompolinsky. "Spin-glass models of neural networks". In: *Physical Review A* 32.2 (Aug. 1985), pp. 1007–1018.

[AGS85b] D. J. Amit, H. Gutfreund, and H. Sompolinsky. "Storing Infinite Numbers of Patterns in a Spin-Glass Model of Neural Networks". In: *Physical Review Letters* 55.14 (Sept. 1985), pp. 1530–1533.

[AGS87] D. J. Amit, H. Gutfreund, and H. Sompolinsky. "Statistical mechanics of neural networks near saturation". In: *Annals of Physics* 173.1 (Jan. 1987), pp. 30–67.

[AN72] K. B. Athreya and P. E. Ney. *Branching Processes (Grundlehren der mathematischen Wissenschaften)*. Springer, 1972, pp. xi+287.

[Bil95] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley-Interscience, 1995.

[Bil99] P. Billingsley. *Convergence of Probability Measures*. Ed. by P. Billingsley. Second. Wiley Series in Probability and Statistics: Probability and Statistics. A Wiley-Interscience Publication. John Wiley & Sons, Inc., July 16, 1999, pp. x+277. 298 pp.

[Bov99] A. Bovier. "Sharp upper bounds on perfect retrieval in the Hopfield Model". In: *Journal of Applied Probability* 36.3 (Sept. 1999), pp. 941–950.

[BG92] A. Bovier and V. Gayrard. "Rigorous bounds on the storage capacity of the dilute Hopfield model". In: *Journal of Statistical Physics* 69.3-4 (Nov. 1992), pp. 597–627.

[BG93] A. Bovier and V. Gayrard. "Rigorous results on the thermodynamics of the dilute Hopfield model". In: *Journal of Statistical Physics* 72.1-2 (July 1993), pp. 79–112.

[BG97] A. Bovier and V. Gayrard. "The retrieval phase of the Hopfield model: A rigorous analysis of the overlap distribution". In: *Probability Theory and Related Fields* 107.1 (Jan. 1997), pp. 61–98.

[BKL02] A. Bovier, I. Kurkova, and M. Löwe. "Fluctuations of the free energy in the REM and the p-spin SK models". In: *The Annals of Probability* 30.2 (Apr. 2002), pp. 605–651.

[BN01] A. Bovier and B. Niederhauser. "The spin-glass phase-transition in the Hopfield model with p-spin interactions". In: *Advances in Theoretical and Mathematical Physics* 5.6 (Aug. 14, 2001), pp. 1001–1046.

[BP97] A. Bovier and P. Picco. *Mathematical Aspects of Spin Glasses and Neural Networks*. Birkhäuser Boston, Dec. 18, 1997. 396 pp.

[Bur94] D. Burshtein. "Nondirect convergence radius and number of iterations of the Hopfield associative memory". In: *IEEE Transactions on Information Theory* 40.3 (May 1994), pp. 838–847.

[Bur98] D. Burshtein. "Long-term attraction in higher order neural networks". In: *IEEE Transactions on Neural Networks* 9.1 (1998), pp. 42–50.

[CS11] S. Chatterjee and Q.-M. Shao. "Nonnormal approximation by Stein's method of exchangeable pairs with application to the Curie–Weiss model". In: *The Annals of Applied Probability* 21.2 (Apr. 2011), pp. 464–483.

[DZ98] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications (Stochastic Modelling and Applied Probability)*. Springer, 1998.

[Dem+17] M. Demircigil, J. Heusel, M. Löwe, S. Upgang, and F. Vermet. "On a Model of Associative Memory with Huge Storage Capacity". In: *Journal of Statistical Physics* 168.2 (July 1, 2017), pp. 288–299.

[EL04] P. Eichelsbacher and M. Löwe. "Moderate deviations for a class of mean-field models". In: *Markov Process. Related Fields* 10.2 (2004), pp. 345–366.

[EL10] P. Eichelsbacher and M. Löwe. "Stein's Method for Dependent Random Variables Occuring in Statistical Mechanics". In: *Electronic Journal of Probability* 15.0 (2010), pp. 962–988.

[Ell85] R. S. Ellis. *Entropy, Large Deviations, and Statistical Mechanics*. Preprint: 3-540-29059-1. Springer New York, 1985.

[EN78a] R. S. Ellis and C. M. Newman. "Limit theorems for sums of dependent random variables occurring in statistical mechanics". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 44.2 (1978), pp. 117–139.

[EN78b] R. S. Ellis and C. M. Newman. "The statistics of Curie-Weiss models". In: *Journal of Statistical Physics* 19.2 (Aug. 1978), pp. 149–161.

[ENR80] R. S. Ellis, C. M. Newman, and J. S. Rosen. "Limit theorems for sums of dependent random variables occurring in statistical mechanics". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 51.2 (1980), pp. 153–169.

[For11] O. Forster. *Analysis 1 : Differential- und Integralrechnung einer Veränderlichen*. Grundkurs Mathematik. Vieweg+Teubner Verlag, 2011, pp. x+324.

[Heb49] D. O. Hebb. *The Organization of Behavior*. A Wiley book in clinical psychology. John Wiley & Sons, 1949.

[Hey70a] C. C. Heyde. "A rate of convergence result for the super-critical Galton-Watson process". In: *Journal of Applied Probability* 7.2 (Aug. 1970), pp. 451–454.

[Hey70b] C. C. Heyde. "Extension of a Result of Seneta for the Super-Critical Galton-Watson Process". In: *The Annals of Mathematical Statistics* 41.2 (Apr. 1970), pp. 739–742.

[HKW15] W. Hochstättler, W. Kirsch, and S. Warzel. "Semicircle Law for a Matrix Ensemble with Dependent Entries". In: *Journal of Theoretical Probability* 29.3 (Feb. 2015), pp. 1047–1068.

[Hop82] J. J. Hopfield. "Neural networks and physical systems with emergent collective computational abilities". In: *Proceedings of the National Academy of Sciences* 79.8 (Apr. 1982), pp. 2554–2558.

[Jan12] S. Janson. "Simply generated trees, conditioned Galton–Watson trees, random allocations and condensation". In: *Probability Surveys*. Discrete Math. Theor. Comput. Sci. Proc., AQ 9.0 (2012), pp. 103–252.

[KS66] H. Kesten and B. P. Stigum. "A Limit Theorem for Multidimensional Galton-Watson Processes". In: *The Annals of Mathematical Statistics* 37.5 (Oct. 1966), pp. 1211–1223.

[Kir15] W. Kirsch. "A survey of the moment method". 2015. URL: `https://www.fernuni-hagen.de/stochastik/downloads/momente.pdf`.

[KP88] J. Komlós and R. Paturi. "Convergence results in an associative memory model". In: *Neural Networks* 1.3 (Jan. 1988), pp. 239–250.

[KP93] J. Komlós and R. Paturi. "Effect of connectivity in an associative memory model". In: *Journal of Computer and System Sciences* 47.2 (Oct. 1993), pp. 350–373.

[KH16] D. Krotov and J. J. Hopfield. "Dense Associative Memory for Pattern Recognition". In: *Advances in Neural Information Processing Systems 29 (2016), 1172-1180* (June 3, 2016).

[Lou97] D. Loukianova. "Lower bounds on the restitution error in the Hopfield model". In: *Probability Theory and Related Fields* 107.2 (Feb. 1997), pp. 161–176.

[Löw98] M. Löwe. "On the storage capacity of Hopfield models with correlated patterns". In: *The Annals of Applied Probability* 8.4 (Nov. 1998), pp. 1216–1250.

[LV05] M. Löwe and F. Vermet. "The storage capacity of the Hopfield model and moderate deviations". In: *Statistics & Probability Letters* 75.4 (Dec. 2005), pp. 237–248.

[LV11] M. Löwe and F. Vermet. "The Hopfield model on a sparse Erdös-Renyi graph". In: *Journal of Statistical Physics* 143.1 (Mar. 2011), pp. 205–214.

[LV15] M. Löwe and F. Vermet. "Capacity of an associative memory model on random graph architectures". In: *Bernoulli* 21.3 (Aug. 2015), pp. 1884–1910.

[LPP95] R. Lyons, R. Pemantle, and Y. Peres. "Conceptual Proofs of L Log L Criteria for Mean Behavior of Branching Processes". In: *The Annals of Probability* 23.3 (July 1995), pp. 1125–1138.

[McE+87] R. J. McEliece, E. C. Posner, E. R. Rodemich, and S. S. Venkatesh. "The capacity of the Hopfield associative memory". In: *IEEE Transactions on Information Theory* 33.4 (July 1987), pp. 461–482.

[New88] C. M. Newman. "Memory capacity in neural network models: Rigorous lower bounds". In: *Neural Networks* 1.3 (Jan. 1988), pp. 223–238.

[Olv74] F. Olver. *Asymptotics and special functions*. New York: Academic Press, 1974.

[Ott49] R. Otter. "The Multiplicative Process". In: *The Annals of Mathematical Statistics* 20.2 (June 1949), pp. 206–224.

[PF77] L. A. Pastur and A. Figotin. "Exactly soluble model of a spin glass". In: *Sov. J. Low Temp. Phys.* 3.6 (1977), pp. 193–210.

[Pet96] D. Petritis. "Thermodynamic Formalism of Neural Computing". In: *Dynamics of Complex Interacting Systems*. Vol. 2. Nonlinear Phenomena and Complex Systems. Netherlands: Springer, 1996, pp. 81–146.

[Pur+13] D. Purves, K. S. LaBar, M. L. Platt, R. Cabeza, and S. A. Huettel. *Principles of Cognitive Neuroscience*. 2. Oxford University Press, 2013.

[Sen68] E. Seneta. "On Recent Theorems Concerning the Supercritical Galton-Watson Process". In: *The Annals of Mathematical Statistics* 39.6 (Dec. 1968), pp. 2098–2102.

[Tal98] M. Talagrand. "Rigorous results for the Hopfield model with many patterns". In: *Probability Theory and Related Fields* 110.2 (Feb. 1998), pp. 177–275.

[Tal00a] M. Talagrand. "On the p-spin interaction model at low temperature". In: *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics* 331.11 (Dec. 2000), pp. 939–942.

[Tal00b] M. Talagrand. "Rigorous low-temperature results for the mean field p -spins interaction model". In: *Probability Theory and Related Fields* 117.3 (July 2000), pp. 303–360.

[Tal03] M. Talagrand. *Spin Glasses: A Challenge for Mathematicians*. Springer Berlin Heidelberg, July 11, 2003. 600 pp.

[Won01] R. Wong. *Asymptotic Approximation of Integrals (Classics in Applied Mathematics)*. Classics in Applied Mathematics. SIAM: Society for Industrial and Applied Mathematics, 2001.

List of symbols

$S = \{-1, 1\}, S^N$	Binary state space S and configuration space S^N.
N, M	Number of neurons in a neural network and number of patterns.
$(W_{ij})_{i,j}$	Weights of the neural network.
$W_{i_1,\ldots,i_n}$	Weights for the Hopfield model with polynomial dynamics, see eq. (2.5).
θ_i	Threshold for neuron i.
$h_i(\sigma) = \sum_j W_{ij}\sigma_j$	Postsynaptic potential of neuron i.
$\text{sgn}(x)$	Signum function $\text{sgn}(x) = \mathbb{1}_{\{x\geq 0\}} - \mathbb{1}_{\{x<0\}}$.
$\xi^1, \ldots, \xi^M$	Patterns which we want the neural network to store. $\xi^1, \ldots, \xi^M$ are i.i.d.
$(\Omega, \mathcal{F})$	Probability space of $\xi^1, \ldots \xi^M$.
$\mathcal{L}(X)$	Law of a random variable X.
$\mathbb{P}, \mathbb{E}[\cdot]$	Probability measure and corresponding expectation value. Under $\mathbb{P}$ ξ^1 has i.i.d. Rademacher spins.
$\mathbb{P}_\beta^{CW}, \mathbb{E}^{CW}[\cdot]$	Probability measure and corresponding expectation value. Under $\mathbb{P}_\beta^{CW}$ ξ^1 is generated according to a Curie-Weiss model with parameter β.
$\sigma_{CW}^2 = (1-\beta)^{-1}$	Variance for the Curie-Weiss magnetization.
$T = (T_i)_{i\leq N}$	Transfer function of the standard Hopfield model, see eq. (1.2).
$\widehat{T} = (\widehat{T}_i)_{i\leq N}$	Transfer function of the generalized Hopfield model, see eq. (2.1).
$\widetilde{T} = (\widetilde{T}_i)_{i\leq N}$	Transfer function of the Hopfield model with polynomial dynamics, see eq. (2.4).
$T^w = (T_i^w)_{i\leq N}$	Transfer function of the Hopfield model with weights, see eq. (4.6).
$d(x, y)$	Hamming distance between $x, y \in \{-1, 1\}^N$.
$\mathcal{S}(x, r)$	Sphere around x with radius r with respect to the Hamming distance.
$\|\cdot\|_1, \|\cdot\|_2$	l_1- resp. l_2-norm.
$\widetilde{\xi}^1 \in \mathcal{S}(\xi^1, \rho N)$	Corrupted version of the pattern ξ^1.
$m^\mu(\sigma), m_i^\mu(\sigma)$	Overlap of pattern ξ^μ with configuration σ with and without neuron i, see Definition 1.3.
Λ, β	Lattice $\Lambda \subseteq \mathbb{Z}^d$ and the inverse temperature β.
$\kappa_{2l} = (2l-1)!!$	$2l$-th moment of a standard normal distribution.

$\mathcal{S}^\Lambda, \mathcal{B}(\mathcal{S}^\Lambda)$	Configuration space and the corresponding Borel-σ-algebra.
$H_\Lambda(\sigma), Z_{\beta,\Lambda}$	Hamiltonian function and partition function of a spin glass model.
$H_{N,M}[\omega](\sigma)$	Hamiltonian function of the standard Hopfield model, see eq. (1.9).
$H_N^{CW}(\sigma), Z_{\beta,N}$	Hamiltonian function and partition function of the Curie-Weiss model, see eqs. (3.1) and (3.3).
h	Local field of a spin glass model.
$\mu_{\beta,\Lambda}(\sigma)$	Gibbs measure, see eq. (1.8).
$\mu_{\beta,N}(\sigma)$	Gibbs measure of the Curie-Weiss model, see eq. (3.2).
$S_N = \sum_{j=1}^N \xi_j$	Magnetization in a Curie-Weiss model, see Definition 3.1.
$H^{CW}(\sigma)$	Hamiltonian function of the Curie-Weiss model.
$\mu_{\beta,N}(\sigma)$	Gibbes measure of the Curie-Weiss model.
$\Gamma(x)$	Gamma function, see 3.12.
$\cosh(x)$	Hyperbolic cosine.
$\tanh(x), \tanh^{-1}(x)$	Hyperbolic tangent and inverse hyperbolic tangent.
$I(x)$	Rate function of Rademacher spins.
$X_n \Rightarrow X$	Weak convergence of X_n to X.
$\mathcal{C}_b(\mathbb{R})$	Set of continuous and bounded functions on $\mathbb{R}$.
$a_n \approx b_n$	Asymptotic equivalence, meaning that $\frac{a_n}{b_n} \overset{n\to\infty}{\longrightarrow} 1$.
$f(x) = \mathcal{O}(g(x))$	Notation for $\limsup_x \left\lvert\frac{f(x)}{g(x)}\right\rvert < \infty$
$f(x) = o(g(x))$	Notation for $\lim_x \frac{f(x)}{g(x)} = 0$ if $g(x) \neq 0$.
$G = (V, E)$	Graph with vertex set V and set of edges E.
$G_k^{(1)} = (V_k^{(1)}, E_k^{(1)})$	Lattice on a torus.
$G_k^{(2)} = (V_k^{(1)}, E_k^{(2)})$	μ-regular graph up to generation k.
$G_k^{(3)} = (V_k^{(1)}, E_k^{(3)})$	Galton-Watson tree up to generation k.
$w = (w_{ij})_{i,j\in V}$	Weights for each connection in the inhomogeneous Hopfield model.
$w_i = (w_{i,j})_{i\in V}$	Weights of neuron i in the inhomogeneous Hopfield model.
$(w_l)_{l\in\mathbb{N}}$	Weights which only depend on the distance between nodes.
$W_i = \frac{\left(\sum_{j\in V} w_{ij}\right)^2}{\sum_{j\in V} w_{ij}^2}$	Factor in the storage capacity of the inhomogeneous Hopfield model, see Theorem 4.2.
W_n	Factor W_i for a neuron i at generation n.
$Z_l(i), Z_l(n)$	Number of vertices at a distance l to node i resp. to a node at generation n.
$(Z_n)_{n\in\mathbb{N}}$	Galton-Watson process.
$Y_n = \frac{Z_n}{\mu^n} \to Y$	Martingale of a Galton-Watson process and its limit Y.
$Y_n^* = \frac{Z_n}{C^n} \to Y^*$	Martingale of a Galton-Watson process and its limit Y^*.
f, f_n	Generating function of Z_1, where $(Z_n)_n$ is Galton-Watson process, and the n-th fold composition of f.