Applied Problems of
Nonlinear Oscillation and Wave Theory

Izvestiya Vysshikh Uchebnykh Zavedeniy. Applied Nonlinear Dynamics. 2023;31(1)

Article DOI: 10.18500/0869-6632-003028

Stability thresholds of attractors of the Hopfield network

I. A. Soloviev[1,2], *V. V. Klinshov*[1,2]✉

[1]Institute of Applied Physics of the RAS, Nizhny Novgorod, Russia
[2]National Research Lobachevsky State University of Nizhny Novgorod, Russia
E-mail: solocool46@gmail.com, ✉vladimir.klinshov@ipfran.ru

Received 24.11.2022, accepted 26.12.2022, available online 18.01.2023, published 31.01.2023

Abstract. *Purpose* of the work is the detailed study of the attractors of the Hopfield network and their basins of attraction depending on the parameters of the system, the size of the network and the number of stored images. To characterize the basins of attraction we used the *method* of the so-called stability threshold, i.e., the minimum distance from an attractor to the boundary of its basin of attraction. For useful attractors, this value corresponds to the minimum distortion of the stored image, after which the system is unable to recognize it. In the *result* of the study it is shown that the dependence of the average stability threshold of useful attractors on the number of stored images can be nonmonotonic, due to which the stability of the network can improve when new images are memorized. An analysis of the stability thresholds allowed to estimate the maximum number of images that the network can store without fatal errors in their recognition. In this case, the stability threshold of useful attractors turns out to be close to the minimum possible value, that is, to unity. To *conclude*, calculation of the stability thresholds provides important information about the attraction basins of the network attractors.

Keywords: dynamical networks, collective dynamics, associative memory.

Acknowledgements. This work was supported by Russian Foundation for Basic Research, grant No 20-52-12021.

For citation: Soloviev IA, Klinshov VV. Stability thresholds of attractors of the Hopfield network. Izvestiya VUZ. Applied Nonlinear Dynamics. 2023;31(1):75–85. DOI: 10.18500/0869-6632-003028

Introduction

The Hopfield network [1] is a classical model of associative memory or memory with content addressing. The functioning as a memory system in this network is possible due to its collective dynamic, which ensures the recovery of the information stored in it from the provided fragment. To do this, the connections in the network are selected according to a certain rule, which ensures the formation of attractors in its phase space corresponding to the data stored in it.

The Hopfield model, proposed more than forty years ago, still attracts the close attention of researchers in the field of neuroscience and information systems. The variants of the network were considered both in the classical discrete and continuous versions [2]. Hopfield network implementations have been proposed in the form of physical systems, such as optical systems [3] or phase-locked frequency circuits [4]. It has been shown that in addition to associative memory, the Hopfield network can solve a number of other tasks, [5] such as optimization and linear programming [6]. Much attention in the study of Hopfield networks was paid to the study of memory capacity, that is, the maximum amount of data that the system is able to remember without significant performance degradation [7, 8]. Recent years have been marked by the emergence of so-called new Hopfield networks (modern Hopfield networks, [9]) or dense associative memory networks (dense associative memories, [10]), in which strong nonlinearity leads to to a significant increase in memory capacity.

From the point of view of nonlinear dynamics, the memory capacity of the system is equal to the maximum number of data-encoding attractors that can be created in phase space simultaneously. However, it turns out that in addition to such «useful» attractors, other, «parasitic» attractors may arise. These attractors do not correspond to any stored data. An important characteristic of both useful and parasitic attractors is their pool of attraction — the set of all initial states of the network from which it evolves to this attractor. In the context of associative memory, the attractor attraction pool is a collection of various initial data that the network recognizes as the corresponding image from its memory. In the multidimensional phase space of a network, the pools of attraction of its attractors can be extremely complex sets, and their definition and description is a very non-trivial task.

This work is devoted to the study of both useful and parasitic attractors of the Hopfield network and their pools of attraction. It is shown that when a sufficiently large number of images are stored in the network, numerous parasitic attractors also appear in it, which are located in areas of phase space that are remote from useful attractors. To characterize the pools of attraction of the attractors,a stability threshold was used – a numeric measure equal to the minimum distance between attractor and edge of its basin. [11]. It is shown that the use of this characteristic allows us to obtain important new information about the structure and properties of the basins of attraction.

1. Network model

In this paper, we consider a discrete version of the Hopfield network, in which each element (neuron) is described by a state variable V_i, taking the value 0 or 1. The vector $\mathbf{V} = (V_1, ..., V_N)$, which determines the state of the network, will also sometimes be called «image». The dynamics of the network unfolds in discrete time, and at each time step, the state of all neurons simultaneously or sequentially changes according to the following rule:

$$V_i \to 0, \text{ if } \sum_{j=1}^{N} T_{ij} V_j < U, \tag{1a}$$

$$V_i \to 1, \text{ if } \sum_{j=1}^{N} T_{ij} V_j > U. \tag{1b}$$

Here N is the size of the network, T_{ij} is the coefficients of communication between neurons. The connection matrix is selected in the following form. Let S images $\mathbf{V}^1, ..., \mathbf{V}^S$ be given, then

$$T_{ij} = \sum_{k=1}^{S} \left(2V_i^k - 1\right) \left(2V_j^k - 1\right). \tag{2}$$

It can be shown [1] that with such a connection matrix, each of the images $\mathbf{V}^k$ becomes an attractor of the network, that is, the images are «written» into its memory. Now, when choosing the initial conditions near some of the recorded images, the system will switch to it, that is, it will restore the stored image. Images $\mathbf{V}^k$ will be also called useful attractors of the system.

2. Parasitic attractors

It turns out, however, that along with useful attractors, «parasitic» can arise in the network that do not correspond to any of the remembered images. The choice of the initial state of the network near the parasitic attractor leads to the inability of the network to recognize the image presented to it. Thus, the presence of parasitic attractors has a significant impact on the functioning of the network, so the question of their presence and quantity is important. In addition, it is important to understand whether parasitic attractors are slightly distorted versions of useful ones or represent completely different images.

To study the number and location of parasitic attractors, a large-scale numerical simulation was carried out depending on the size of N and the number of memorized images S. For each set of parameters, 360 different network implementations were considered, with S random generated stored images. For each implementation of the network, 100 initial conditions were selected and the dynamics of the network was modeled was modeled until it evolves to the attractor as a result of which a list of unique parasitic attractors was compiled. It is important to note that for any attractor of the network, the inverted image is also an attractor, which is easy to show based on the rule (1). Therefore, we did not consider the attractors, which are inverted memorized images, to be parasitic. Also, two parasitic attractors, which are inverted versions of each other, were considered as one.

In Fig. 1 and 2 present the results of the study averaged over all network implementations. In Fig. 1, *a* shows the dependence of the number of unique parasitic attractors on the size of the network N, and in Fig. 1, *b* — on the number of saved images S. The dependence of the number of parasitic attractors on the size of the network is monotonically increasing and has a saturating character. Much more interesting is the dependence of the number of parasitic attractors on

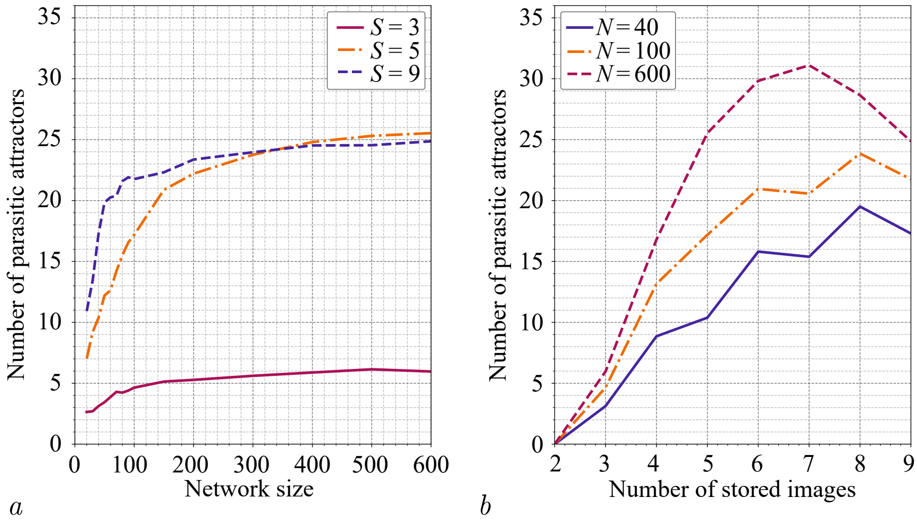
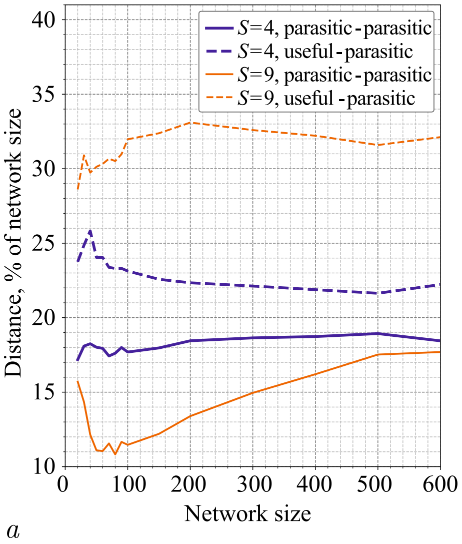


Fig. 1. Average number of parasitic attractors in the hopfield network depending on the network size (*a*) and the number of stored images (*b*)



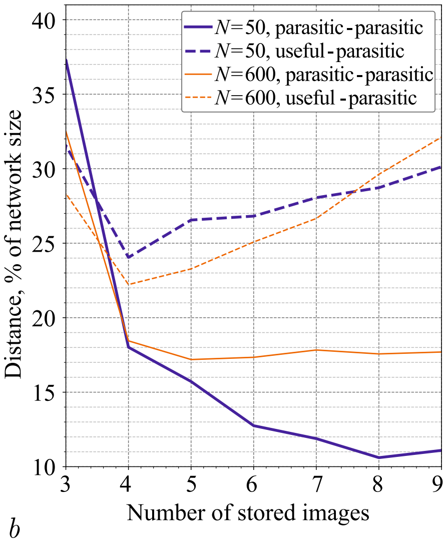


Fig. 2. Average distance between parasitic attractors and the nearest useful attractor, average distance between parasitic attractors and the nearest parasitic attractor depending on the network size (a) and the number of stored images (b) (color online)

the number of memorized images of S. This dependence is nonmonotonic with a maximum at $S = 7$–8, and this value practically does not change when N varies widely from 20 to 600.

In Fig. 2 provides data on the average distance from a parasitic attractor to the nearest useful attractor, as well as to the nearest other parasitic attractor. The distance between the two images was calculated as the Hamming distance, that is, the number of mismatched elements was determined:

$$L(\mathbf{V}^1, \mathbf{V}^2) = \sum_{i=1}^{N} \left| V_i^1 - V_i^2 \right|. \tag{3}$$

The calculated distance is related to the correlation coefficient ρ between the two images as $L = N(1 - \rho)/2$. Note that in the graphs the distance was normalized by the size of the network N. It can be seen that the dependence of the average distance between a parasitic attractor and nearest other attractor is close to linear, especially for large N. Thus, parasitic attractors differ significantly from useful ones and from each other. However, it should be noted that parasitic attractors are characterized by a significant correlation with some of the memorized images, since the distance to the nearest one is always significantly less than $N/2$.

The dependence on the number of memorized images is significantly different for the distance to the nearest parasitic and the nearest useful attractor. In the first case, the dependence is monotonically decreasing, and in the second it shows a minimum at $S = 4$, and this value is almost the same for all N ranging from 20 and up to 600.

3. Pools of attraction and their characteristics

An important characteristic of the attractor is its pool of attraction — the set of initial states of the network, which due to its dynamics converge to this attractor. In the context of associative memory, the pool of attraction of a useful attractor is a set of images recognized by the system. The pools of attraction of parasitic attractors, in turn, determine a set of images, the presentation of which leads to a malfunction in the system. For the memory system to work well, the pools of attraction of useful attractors should be in some sense large, and the pools of

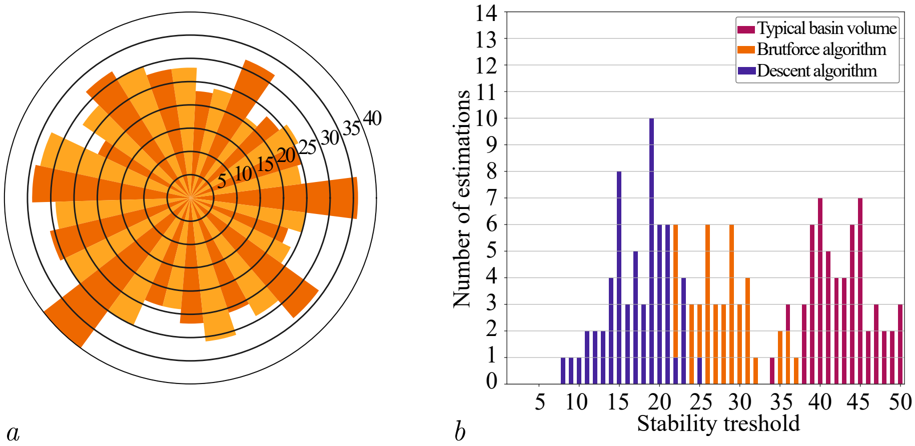


Fig. 3. *a* — Stability treshold estimations of one attractor for several runs of the algorithm with different initial conditions. *b* — Distributions of stability treshold estimation for descent and brutforce algorithms compared to the typical basin volume (color online)

parasitic attractors — small. To formalize this intuitive consideration, it is necessary to introduce some ways of quantifying the basins of attraction.

In a multidimensional phase space, the pools of attraction of attractors are extremely complex sets, and their description is a non-trivial task. Several different numerical measures have been proposed to characterize the size of basins of complex dynamical systems. For example, basin stability [12] characterizes the volume of the basin in phase space. The stability threshold [11] is the shortest distance from the attractor to the boundary of its attraction basin. For an associative memory system, the stability threshold of useful attractors seems to be the most informative characteristic, since it sets the maximum degree of distortion of the image, which is guaranteed to be recognized by the network. At the same time, the disturbance corresponding to the stability threshold is a minimal distortion of the image, which leads to a malfunction of the system. Finding such a disturbance allows you to determine the «weak point» of the system, vulnerable to possible attacks [13].

To determine the stability threshold of the Hopfield network, we used a variant of gradient descent proposed in [11], modified for discrete systems. The algorithm starts with a random image of $\mathbf{U}$ located outside the attraction pool of the attractor of interest to us $\mathbf{V}$, and tries to approach the attractor by taking steps in its direction. The step is done as follows: from the set of elements of the image $\mathbf{U}$ that differ from the corresponding elements of the vector $\mathbf{V}$, one element is randomly selected and its value is inverted, which leads to a decrease in the distance between $\mathbf{U}$ and $\mathbf{V}$. If the image $\mathbf{U}$ falls into the pool of attraction of the attractor $\mathbf{V}$, the step taken is rejected, in the opposite case it is accepted. The described steps are repeated until a situation arises in which no step is possible. In this case, the image $\mathbf{U}$ is located on the boundary of the attraction pool of the attractor $\mathbf{V}$, and the local minimum of the distance between the attractor and the boundary of its attraction pool is reached. Starting from various initial images $\mathbf{U}$, the algorithm finds a set of local minima, and the minimum of them is taken as an estimate of the stability threshold.

Fig. 3 illustrates typical results of the search for the stability threshold of one of the attractors of the network. The algorithm started 100 times with various random images, and the pie chart in Fig. 3, *a* displays the distances to the found local minima. Let's note a few important points. Firstly, there are a lot of local minima: almost every time the algorithm finds a new minimum. Secondly, the distance from the attractor to the found minima is significantly

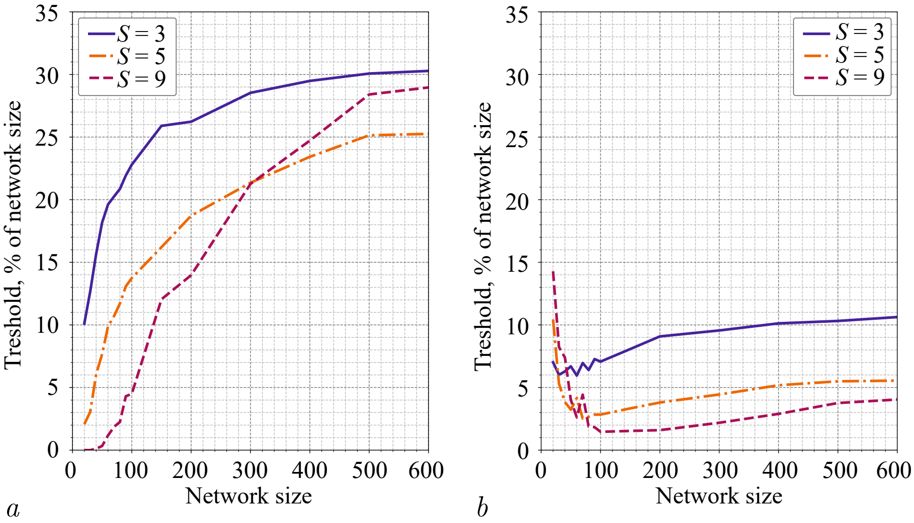


Fig. 4. Average stability treshold depending on the network size N with different numbers of stored images S for useful (a) and parasitic (b) attractors

less than the characteristic size of the attractor pool.The characteristic size of the attractor pool was calculated as the radius of a hyperchar with a volume equal to the volume of this pool., which is also marked in the figure. These features indicate an extremely complex, rugged shape of the attraction basin, which is characterized by the presence of many deep and narrow «depressions» (cf. [14]). Determining the stability threshold for a basin of this shape is extremely important, since it allows you to determine very specific weak disturbances that lead to errors in the system.

The results obtained also indicate the effectiveness of our algorithm for obtaining an estimate of the stability threshold. In Fig. 3, b the distribution of stability thresholds obtained using the proposed descent algorithm for attractors of 60 different networks with the same parameters is given. For comparison, the same figure shows the estimates obtained using the <brute force> method, that is, by randomly iterating through various initial conditions and checking their belonging to the pool. The parameters of both algorithms were chosen in such a way as to ensure the same computing time. At the same time, the descent algorithm gives significantly better results.

4. Investigation of system stability thresholds

With the help of the proposed algorithm, a systematic study of the attraction pools of both useful and parasitic attractors of the Hopfield network was carried out, depending on the size of the network N and the number of images S memorized. For each set of parameters N and S, 120 different network implementations were considered, in each of them the stored images were randomly selected. For each implementation, all parasitic attractors were first found, as described above, and then the stability thresholds of useful and parasitic attractors were determined. The results obtained were averaged over all attractors and over all network implementations.

The results of the study are presented in Fig. 4, a, b, which show the dependence of the stability thresholds on the size of the network and the number of stored images. Note that in the graphs, the threshold is normalized by the size of the network N. The dependence of the normalized stability threshold on the size of the network is saturated at large N for both useful and parasitic attractors. Thus, attractors in large networks of $N \gg S$ are characterized by sufficiently strong stability: to remove the system from their pool of attraction, it is necessary to apply a perturbation of the order of 0.15-0.25, that is, to change the state of 15-25% of the elements.

Of great interest is the dependence of the stability thresholds on the number of images stored in the network. Let us first stop at the thresholds of stability of useful attractors. From Fig. 5, *a* it can be seen that the dependence is different for small ($N \leqslant 100$) and large ($N > 100$) networks. If for small networks the stability threshold decreases monotonically with the increase in the number of stored images, then for large networks this dependence is non-monotonic: the stability threshold first decreases, reaches a local minimum (at $S = 4$ for $N = 600$), then increases, and begins to decrease again at sufficiently large S. The presence of an increasing area on this dependence contradicts intuition, since when new attractors are added to the system, the average size of their attraction pools decreases, and it is natural to expect a decrease in stability thresholds as well. However, it turns out that the shape of the basins is «smoothed», and the deepest «depressions» become smaller, which leads to an increase in the stability threshold. This observation leads to unexpected recommendations to increase the system's resistance to external attacks: sometimes it is enough to add new images to the system, which will lead to an increase in the stability thresholds.

Note, however, that an anomalous increasing dependence of the stability threshold on the number of memorized images can be observed only with a small (compared to N) number of images. Further growth of S always leads to a decrease in the stability threshold. For parasitic attractors, the stability threshold also decreases with the growth of S, as shown in 5, *b*. However, with large S, the stability threshold of parasitic attractors decreases more slowly than the stability threshold of useful attractors, and with a certain number of stored images, their values become equal, as shown in Fig. 6. The number of images at which such equality is achieved is an important characteristic of the system, which determines the maximum number of images that the system can remember and correctly recognize. When trying to memorize a larger number of images, the stability threshold of useful attractors quickly drops to zero, which means that they lose stability and the system fails.

The critical number of images S^* can be defined as the number of stored images under which average stability thresholds of parasitic and useful attractors are equal. On fig. 7, *a* the dependence of the critical number of images on the size of the system is presented, and this

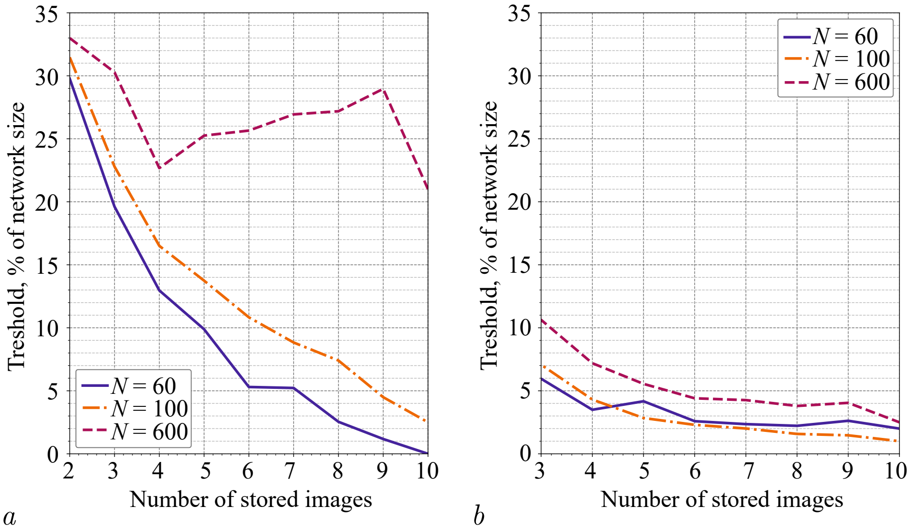


Fig. 5. Average stability treshold depending on the number of stored images S with different network sizes N for useful (a) and parasitic (b) attractors

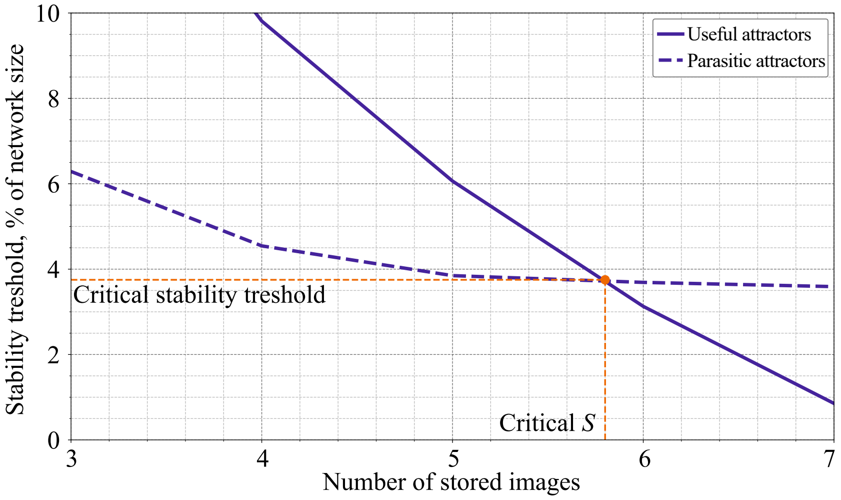


Fig. 6. Average stability treshold depending on the number of stored images S with network size $N = 40$ for useful and parasitic attractors. The point of intersection determines the critical number of stored images and the critical stability treshold

dependence is close to linear. Interestingly, in this way a certain critical number of images is close to the estimate of the maximum allowable number of images $S \approx 0.1N$ given in the original paper [1]. Thus, a comparison of the stability thresholds of useful and parasitic attractors allows us to obtain an independent estimate of the memory capacity of the Hopfield network, which is in good agreement with the traditional one. It is interesting to analyze the value of the stability threshold at a critical number of images. The dependence of this critical threshold on the size of the network is shown in Fig. 7, *b*, and it should be noted that, unlike the other graphs, the absolute value of the threshold is presented here, not normalized to the size of the network. It can be seen from the graph that the critical threshold of stability in absolute terms is close to unity, that is, the disturbance of just one element of the memorized image leads to the impossibility of its recognition, that is, system failure.

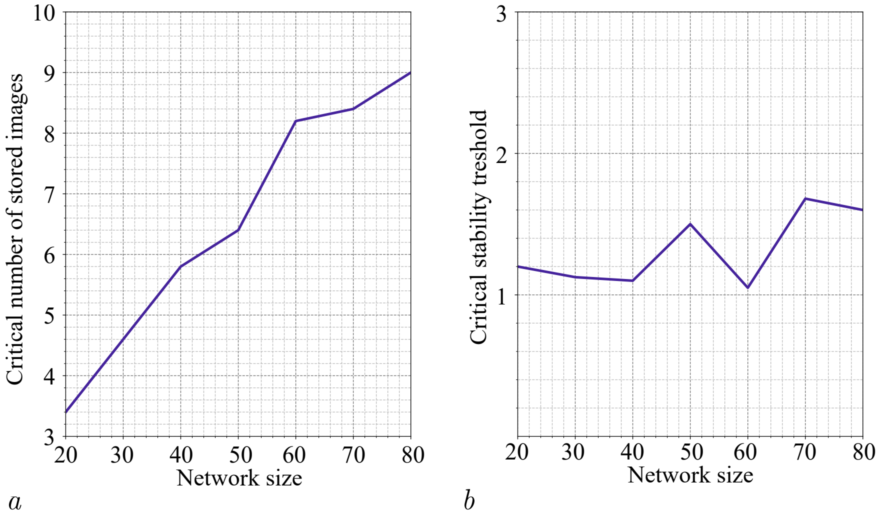


Fig. 7. Critical number of stored images (*a*) and critical stability treshold (*b*) depending on the network size

Conclusion

Attractors of the Hopfield network and their basins of attraction were investigated in this paper. It is shown that when more than two images are recorded in the network, in addition to useful attractors corresponding to these images, parasitic attractors also appear in the system. They do not correspond to any of stored images. The number of parasitic attractors demonstrates a saturating dependence on the size of the network and a non-monotonic dependence with a pronounced maximum on the number of memorized images. Parasitic attractors are located far enough away from useful attractors and from each other.

To study the attraction basins of attractors, a method based on the stability threshold proposed in [11] was used. This method allows us to find the minimum amplitude perturbations of the attractor, leading to the exit from its basins of attraction. In the context of associative memory implemented in the Hopfield network, the stability threshold corresponds to minimal distortion of the image, leading to its incorrect recognition. The definition of such disturbances is important from the point of view of vulnerability to possible attacks.

An algorithm for calculating the stability threshold in discrete systems was proposed, and based on it, a detailed study of the basins in the Hopfield network was carried out depending on its parameters. It is shown that a typical attractor attraction basin is characterized by a complex shape, in which there are numerous narrow and deep depressions. The dependences of the stability thresholds of both useful and parasitic attractors of the network on its parameters were studied. The most interesting is the dependence of the average stability threshold on the number of stored images for large network sizes N. This dependence demonstrates a pronounced minimum, which implies an unexpected possibility of increasing the network's resistance to external attacks by adding additional images to its memory.

Based on the study, a new criterion is proposed for determining the maximum number of images that the system is able to store without a significant deterioration in the quality of their recognition: this is the number of images at which the threshold of stability of useful attractors becomes equal (on average) to the threshold of stability of parasitic attractors. It is shown that in this way a certain critical number of images is close to the classical estimate of the capacity of the system $0.1N$, and the critical threshold is close to unity. Thus, the method of stability thresholds allowed us to obtain important new information about the properties of the Hopfield network.

References

1. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. U. S. A. 1982;79(8):2554–2558. DOI: 10.1073/pnas.79.8.2554.
2. Hopfield JJ. Neurons with graded response have collective computational properties like those of two-state neurons. Proc. Natl. Acad. Sci. U. S. A. 1984;81(10):3088–3092. DOI: 10.1073/pnas.81.10.3088.
3. Farhat NH, Psaltis D, Prata A, Paek E. Optical implementation of the Hopfield model. Applied Optics. 1985;24(10):1469–1475. DOI: 10.1364/AO.24.001469.
4. Hoppensteadt FC, Izhikevich EM. Pattern recognition via synchronization in phase-locked loop neural networks. IEEE Transactions on Neural Networks. 2000;11(3):734–738. DOI: 10.1109/72.846744.
5. Joya G, Atencia MA, Sandoval F. Hopfield neural networks for optimization: study of the different dynamics. Neurocomputing. 2002;43(1–4):219–237. DOI: 10.1016/S0925-2312(01)00337-X.
6. Wen UP, Lan KM, Shih HS. A review of Hopfield neural networks for solving mathematical

programming problems. European Journal of Operational Research. 2009;198(3):675–687. DOI: 10.1016/j.ejor.2008.11.002.

7. McEliece R, Posner E, Rodemich E, Venkatesh S. The capacity of the Hopfield associative memory. IEEE Transactions on Information Theory. 1987;33(4):461–482. DOI: 10.1109/TIT.1987.1057328.
8. Storkey A. Increasing the capacity of a hopfield network without sacrificing functionality. In: Gerstner W, Germond A, Hasler M, Nicoud JD. editors. Artificial Neural Networks — ICANN'97. ICANN 1997. Vol. 1327 of Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 1997. P. 451–456. DOI: 10.1007/BFb0020196.
9. Krotov D, Hopfield JJ. Dense associative memory for pattern recognition. In: NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems. 5–10 December 2016, Barcelona, Spain. New York: Curran Associates Inc.; 2016. P. 1180–1188. DOI: 10.5555/3157096.3157228.
10. Ramsauer H, Schäfl B, Lehner J, Seidl P, Widrich M, Adler T, Gruber L, Holzleitner M, Pavlović M, Sandve GK, Greiff V, Kreil D, Kopp M, Klambauer G, Brandstetter J, Hochreiter S. Hopfield networks is all you need [Electronic resource]. arXiv:2008.02217. arXiv Preprint; 2020. 94 p. Available from: https://arxiv.org/abs/2008.02217.
11. Klinshov VV, Nekorkin VI, Kurths J. Stability threshold approach for complex dynamical systems. New Journal of Physics. 2016;18(1):013004. DOI: 10.1088/1367-2630/18/1/013004.
12. Menck PJ, Heitzig J, Marwan N, Kurths J. How basin stability complements the linear-stability paradigm. Nature Physics. 2013;9(2):89–92. DOI: 10.1038/nphys2516.
13. Chakraborty A, Alam M, Dey V, Chattopadhyay A, Mukhopadhyay D. Adversarial attacks and defences: A survey [Electronic resource]. arXiv:1810.00069. arXiv Preprint; 2018. 31 p. Available from: https://arxiv.org/abs/1810.00069.
14. Amari SI, Maginu K. Statistical neurodynamics of associative memory. Neural Networks. 1988;1(1):63–73. DOI: 10.1016/0893-6080(88)90022-6.