The Hopfield-type Memory Without Catastrophic Forgetting

Iakov Karandashev Yakov.Karandashev@phystech.edu
Boris Kryzhanovsky kryzhanov@mail.ru
Leonid Litinskii litin@mail.ru
Center of Optical Neural Technologies of Scientific Research Institute for System Analysis
Russian Academy of Science
Moscow, Russia

***Abstract*—We analyzed a Hopfield-like model of artificial memory that reproduces some features of the human memory. They are: a) the ability to absorb new information when working; b) the memorized patterns are only a small part of a set of patterns that are written down in connection matrix; c) the more the pattern was shown during the learning process, the better the quality of its recognition. We used the Hebb rule, but each pattern was supplied with its own weight. The weight is proportional to the frequency of the pattern showing during the learning of the network. As a result unlimited number of patterns can be written down in the connection matrix, and that would not lead to the memory destroying (as it has place in the standard Hopfield model). However, only the patterns that were shown rather frequently would be recognized: their weights have to be larger a critical value. For analyzed variants of the weights distribution the storage capacity was estimated as ~0.05*N*-0.06*N*, where *N* is the dimensionality of the problem.**

Keywords - Associative memory; catastrophic forgetting; quasi-Hebbian matrix.

I. INTRODUCTION

In the standard Hopfield model one uses the Hebb matrix constructed with the aid of M random patterns [1], [2]. For definiteness we suppose that *N*-dimensional vector-row $\mathbf{x}^{\mu}=(x_1^{\mu},x_2^{\mu},...,x_N^{\mu})$ with binary coordinates $x_i^{\mu}=\pm 1$ is the μ-th pattern, the number of patterns is equal to M, and the Hebb connection matrix has the form

$$J_{ij}=(1-\delta_{ij})\sum_{\mu=1}^{M}x_i^{\mu}x_j^{\mu}\ .$$

The estimate for number of random patterns that can be memorized in the Hopfield model is well known: $M_c\approx 0.138\cdot N$. If the number of patterns written down into connection matrix is larger than M_c, the catastrophe takes place: the network ceases to recognize patterns at all.

This catastrophic forgetting is a troublesome defect of the Hopfield model. Indeed, let us imagine a robot whose memory is based on the Hopfield model. It is natural to think that his memory is steadily filled up. When the robot sees a new image, it is written additionally to its memory. Catastrophic forgetting means that when the number of memorized patterns exceeds M_c, the memory will be completely destroyed. Everything that was accumulated in the memory would be forgotten. This behavior is contrary to common sense.

Earlier some modifications of the Hebb matrix were proposed to eliminate the memory destruction [3]-[7]. As the result of such modifications an unlimited number of random patterns can be fearlessly written down into matrix elements one by one. However, the memory of the network is restricted. If as previously the maximum number of recognized patterns is denoted by M_c, for the models discussed in [3]-[7] $M_c\approx 0.05\cdot N$. All these models have the same weak point: only patterns that are the last written down in the connection matrix constitute the memory of the network. Let us explain the last statement. Let us number the patterns in the course of their appearance during the learning process: the later the pattern was shown to the network, the larger its number. Then it turns out that the real network memory is formed of patterns whose numbers belong to the interval $\mu\in[M-M_c,M]$. Patterns with order numbers less than $M-M_c$ are excluded from the memory irretrievably. That is the common property of the models [3]-[7].

In our work, we succeeded in eliminating of the catastrophic forgetting of the Hopfield model. In our approach every pattern is supplied by an individual weight. Then in place of the Hebbian matrix we obtain a quasi-Hebbian connection of the form

$$J_{ij}=(1-\delta_{ij})\sum_{\mu=1}^{M}r_{\mu}x_i^{\mu}x_j^{\mu}\ . \qquad (1)$$

Here the weights r_{μ} are positive and put in decreasing order: $r_1\ge r_2\ge ...\ge ...\ge 0$. It is natural to treat the weight r_{μ} as the frequency of the μ-th pattern showing during the learning process.

In previous papers [8], [9] with the aid of statistical physics methods the main equation for the Hopfield model with the quasi-Hebbian matrix was obtained. This equation generalizes the classical equation of the standard Hopfield model [1]. For a special distribution of the weights we succeeded in solution of the main equation: the case when only one coefficient differed from all other that were identically equal: $r_1=\tau$, $r_2=r_3=...=r_M=1$ was examined. Theoretical results were confirmed by computer simulations. (In what follows we use the notations introduced in [8], [9].)

In this paper, we present the results of solving of the main equation in the general case. The main obtained result is as follows. For every weights distribution $\{r_{\mu}\}$ there is such a critical value r_c that only patterns whose weights are greater than r_c will be recognized by the network. Other patterns are not recognized. Now we know only the algorithm of calculation of the critical value r_c . It is not clear, if it is possible to obtain an analytical expression for r_c .

The case of the weights, which are the terms of a decreasing geometric series $r_\mu = q^\mu$ ($q < 1$) we discuss in details. For this weights distribution the number of the recognized patterns is $\sim 0.05 \cdot N$. The results are confirmed by computer simulations.

Note, for the first time the quasi-Hebbian connection matrix (1) was discussed many years ago. For this matrix the implicit form of the main equation (2) was obtained in [6]. However, the authors of [6] examined the case of the standard Hopfield model only ($r_\mu \equiv 1$). Our contribution is the solution of the main equation in the general form.

II. SOLUTION OF MAIN EQUATION

The main equation for the k -th pattern, which we obtained in [8], [9], has the form:

$$\frac{\gamma^2}{\alpha} = \frac{1}{M-1}\sum_{\mu\neq k}^{M}\left(\frac{r_\mu}{r_k\varphi - r_\mu}\right)^2 . \qquad (2)$$

It is supposed that M and N are very large: $M, N >> 1$, α is the load parameter: $\alpha = M / N$, $\gamma = \gamma(y)$ and $\varphi = \varphi(y)$ are functions of an auxiliary argument $y \geq 0$:

$$\gamma(y) = \sqrt{\frac{2}{\pi}}e^{-y^2} , \quad \varphi(y) = \frac{\sqrt{\pi} erf(y)}{2y}e^{y^2} .$$

The function $\gamma(y)$ decreases monotonically, and $\varphi(y)$ increases monotonically from the minimal value $\varphi(0) = 1$. For what follows it is important that $\varphi(y) \geq 1$. When all the weights are equal to each other, equation (2) turns into the equation for the standard Hopfield model [1], [2], [6].

Let us transform Eq.(2) dividing the left hand and right hand sides by M . Moreover we tend the upper limit of the sum in r.h.s. of Eq. (2) to infinity. In fact, we pass to the model with infinitely number of patterns. The main equation takes form:

$$N = \sum_{\mu\neq k}^{\infty} f_\mu^{(k)}(y) , \qquad (3)$$

where $f_\mu^{(k)}$ are the functions of γ , φ and r_μ :

$$f_\mu^{(k)}(y) = \left(\frac{t_\mu^{(k)}}{\gamma(\varphi - t_\mu^{(k)})}\right)^2 , \quad t_\mu^{(k)} = \frac{r_\mu}{r_k} , \quad \mu \neq k . \qquad (4)$$

Eq. (3) connects the dimensionality N of the problem with number k of the pattern. When we find the solution y_0 of this equation, we calculate the overlap of the k -th pattern with the nearest fixed point:

$$m_k = erf\left(y_0\right) . \qquad (5)$$

When $m_k \approx 1$, in the vicinity of the k -th pattern there is a fixed point of the network. This situation is interpreted as recognition of the k -th pattern by the network. If $m_k \approx 0$, the k -th pattern is not recognized by the network.

The values $t_\mu^{(k)}$ are arranged in decreasing order. For what follows it is important that the first $k-1$ of these values are larger than 1, and the other ones are less than 1:

$$t_1^{(k)} > t_2^{(k)} > ... > t_{k-1}^{(k)} > 1 > t_{k+1}^{(k)} > t_{k+2}^{(k)} > .... \qquad (6)$$

The r.h.s. of Eq. (3) is the result of summarizing over the set of functions $f_\mu^{(k)}(y)$ (4). It is easy to see that when $y \to \infty$ the denominator $\gamma(\varphi - t_\mu^{(k)})$ of any function $f_\mu^{(k)}(y)$ tends to 0. In other words, at the infinity each function $f_\mu^{(k)}(y)$ increases unrestrictedly. As a result, at the infinity the r.h.s. of Eq.(3) increases unrestrictedly too.

The behavior of the function $f_\mu^{(k)}(y)$ for finite values of argument depends on the constant $t_\mu^{(k)}$ in its denominator. If $t_\mu^{(k)} < 1$, the function $f_\mu^{(k)}(y)$ is everywhere continuous and limited. If $t_\mu^{(k)} > 1$, the function $f_\mu^{(k)}(y)$ has a singular point. In this case the denominator of the function $f_\mu^{(k)}(y)$ is equal to zero for some value $y_\mu^{(k)}$ of its argument:

$$\varphi\left(y_\mu^{(k)}\right) = t_\mu^{(k)} \quad \Leftrightarrow \quad y_\mu^{(k)} = \varphi^{-1}\left(t_\mu^{(k)}\right),$$

where φ^{-1} is the inverse function with regard to φ . We see that for every $t_\mu^{(k)} > 1$ the function $f_\mu^{(k)}(y)$ has the discontinuity of the second kind in the point $y_\mu^{(k)}$. Since in the series (6) the first $k-1$ values of $t_\mu^{(k)}$ are greater than 1, it is easy to understand that the r.h.s. of Eq.(3) has the discontinuities of the second kind in $k-1$ points $y_{k-1}^{(k)} < y_{k-2}^{(k)} < ... < y_1^{(k)}$. At the infinity the r.h.s. of Eq.(3) increases unrestrictedly.

For simplicity let us go in Eq.(3) to inverse quantities:

$$\frac{1}{N} = F_k(y) \text{, where } F_k(y) = \left(\sum_{\mu\neq k}^{\infty} f_\mu^{(k)}(y)\right)^{-1} . \qquad (7)$$

It is evident that nonnegative function $F_k(y)$ in the r.h.s. of Eq.(7) is equal to zero in the points $y_{k-1}^{(k)} < y_{k-2}^{(k)} <\ldots< y_1^{(k)}$. At the infinity $F_k(y)$ tends to zero. The typical behavior of the function $F_k(y)$ is shown in Fig.1. To the right of the rightmost zero $y_1^{(k)}$, where the inequality $\varphi(y) > t_1^{(k)}$ holds, the function $F_k(y)$ at first increases, and then after reaching its maximum the function $F_k(y)$ decreases monotonically. Let $y_c^{(k)}$ be the coordinate of the rightmost maximum of the function $F_k(y)$. The value of $F_k(y)$ in the point $y_c^{(k)}$ determines the critical characteristics related to the recognition of the k -th pattern. Let us explain what it means.

Generally speaking, Eq.(7) has several solutions. Their number is equal to the number of intersections of the function $F_k(y)$ with the straight line that is parallel to abscissa axis at the height $1/N$ (see Fig. 1). These solutions correspond to stationary points of the solution of the saddle-point equation [1], [2]. However, only one of these intersections is important. Its coordinate is to the right of the rightmost maximum $y_c^{(k)}$. This intersection corresponds to the minimum of the solution of the saddle-point equation. Other solutions of Eq.(7) can be omitted.

As example in Fig.1 the behavior of the r.h.s. of Eq.(7) for the pattern number 5 ($k=5$) is shown for the weights that are the terms of harmonic series $r_\mu = 1/\mu$. Four points $y_4^{(5)} < y_3^{(5)} < y_2^{(5)} < y_1^{(5)}$ are zeros of the function $F_5(y)$. For y that are greater than $y_1^{(5)}$ ($y > y_1^{(5)}$) the function $F_5(y)$ at first increases up to its local maximum in the point $y_c^{(5)}$ and then decreases monotonically. The dashed line that is parallel to the abscissa axis is drawn at the height 0.001. When the l.h.s. of Eq.(7) is equal to 0.001, we obtain $N = 1000$. In other words, for this quasi-Hebbian matrix of the dimensionality $N = 1000$ in the vicinity of the 5-th pattern there is a fixed point certainly. Since the solution of Eq.(7) is large enough, $y_0 \approx 3.5$, the overlap (5) of the pattern and the fixed point is close to 1.

Let us little by little decrease the dimensionality N. The dashed straight line will go up, and the solution $y_0(N)$ of Eq.(7) will shift in the region of smaller values. This will go on till y_0 coincides with the critical value $y_c^{(5)}$. Just this defines the minimal dimensionality $N_{\min}$ for which the fixed point in the vicinity of the 5-th pattern still exists. Since for $N < N_{\min}$ equation (7) has no solutions in the region $y > y_1^{(5)}$, there is no fixed points in the vicinity of the 5-th pattern. From the point of view of the neural network memory this means that when $N < N_{\min}$ the 5-th pattern is not recognized by the network.

Up to now we fixed the number of the pattern k and decreased the dimensionality N. However, it is reasonable to fix the dimensionality N and increase k little by little. We seek its maximal value for which Eq.(7) has a solution. It is easy to show that when k increases the critical point $y_c^{(k)}$ shifts to the right, and the value of the maximum of $F_k(y_c^{(k)})$ decreases steadily. Then it is not difficult to find the maximal value of k for which Eq.(7) has a solution. By $k_m = k_m(N)$ we denote this maximal value.

For given dimensionality N the pattern with the number k_m is the last in whose vicinity there is a fixed point. For $k < k_m$ Eq.(7) has a solution in the region $y > y_c^{(k)}$ too. Consequently, these patterns will also be recognized. On the contrary, for $k > k_m$ Eq.(7) has no solutions in the region $y > y_c^{(k)}$. Consequently, the patterns with such numbers will not be recognized.

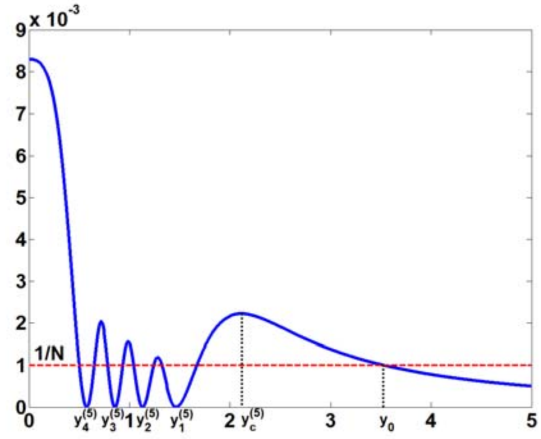

Figure 1. The behavior of the function $F_k(y)$ defined by Eq.(7) when the weights are equal to $r_\mu = 1/\mu$. Here $k=5$, y_0 is the solution of the equation (7) for $1/N = 0.001$ and $y_c^{(5)}$ is the critical value.

Let r_c be the weight corresponding to the pattern with the number k_m: $r_c = r_{k_m}$. Our consideration shows that only the patterns, whose weights are not less than the critical value r_c will be recognized. Patterns whose weights are less than the critical value r_c are not recognized in spite of the fact that these patterns take part in the construction of the quasi-Hebbian matrix. The memory of the network is limited, but the catastrophic forgetting does not occur.

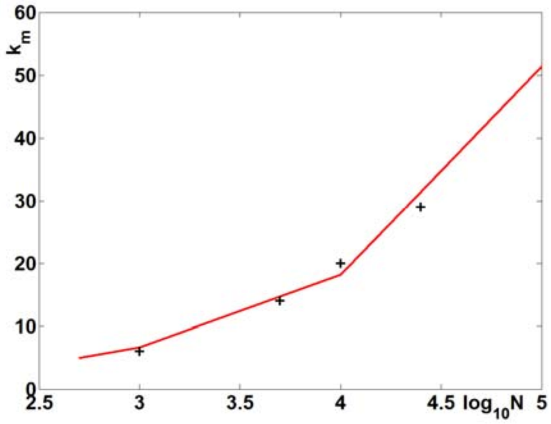

Figure 2. Maximal number of the pattern $k_m(N)$ that can be recognized for the weights $r_\mu = 1/\mu$. The daggers show the results of computer simulations $k_m^{\exp}$.

Let us show how this approach works, choosing as an example matrices whose weights are the terms of harmonic series $r_\mu = 1/\mu$. For the dimensionalities $N = 10^3$, $5\cdot 10^3$, 10^4 , $2.5\cdot 10^4$ and 10^5 we obtained the maximal number $k_m(N)$ of the pattern that could be recognized solving Eq.(7) numerically. In Fig.2 the value of $k_m(N)$ as function of $\log_{10} N$ is shown. The daggers are the results of computer simulation $k_m^{\exp}$. For the given N we generated a random matrix with the weights that are the terms of harmonic series and defined the maximal number of the pattern that was a fixed point. (Patterns with larger numbers differed notably from the nearest fixed points.) The number $k_m^{\exp}$ was obtained as a result of averaging over 10 random matrices. We see that the experimental values are in good agreement with the theoretical results. The numerical solution of Eq.(7) shows that the storage capacity of such a network asymptotically tends to zero:

$$\lim_{N\to\infty} k_{\mathrm{m}}(N)/N \sim \lim_{N\to\infty} 1/\sqrt{N\ln N} = 0 .$$

In the next section we examine in details a particular distribution of the weights.

III. Geometric Progression as Weights

Let us discuss in details the case of the weights in the form of decreasing geometric progression $r_\mu = q^\mu$, where $q \in (0,1)$. For the first the weight of such a type were mentioned in [5] and [6]. It is natural to assume that in Eq.(3) the first value of the summation index is equal to zero and $r_0 = 1$. Now Eq.(3) has the form

$$N = \frac{1}{\gamma^2}\sum_{\mu=0\neq k}^{\infty}\left(\frac{q^\mu}{s_k - q^\mu}\right)^2 \text{, where } s_k = q^k\varphi(y) . \qquad (8)$$

Our interest is the solution of this equation for large values of the argument when the inequality $s_k = q^k\varphi(y) > 1$ is fulfilled. In the r.h.s. of Eq.(8) we replace summation by integration. Then in both sides of the equation we pass to inverse quantities and obtain the analogue of equation (7):

$$\frac{1}{N} = \frac{\gamma^2(\varphi-1)^2}{(\varphi-1)^2\Phi_k(y)-1}, \qquad (9)$$

where

$$\Phi_k(y) = \frac{\ln\left(\frac{s_k-1}{s_k}\right) + \frac{1}{s_k-1}}{|\ln q|} .$$

When solving Eq.(9) numerically we can find the maximal number of the pattern k_m which is recognized by the network yet, $k_m = k_m(N,q)$. Our goal is to find such value of the parameter q that corresponds to the maximum of the storage capacity of the network. In other words, we find the value of the parameter q for which $k_m(N,q)$ is maximal: $\tilde{k}(N) = \max_q k_m(N,q)$. It is evident that this optimal q have to exist; $q_m = q_m(N)$ denotes its value.

On the left panel of Fig.3 the dependence of the ratio $k_m(N,q)/N$ on q for three dimensionalities N is shown. We see that curves have distinct points of maximum, but for all cases the value of the all maximums are the same:

$$\lim_{N\to\infty} \tilde{k}(N)/N \approx 0.05 . \qquad (10)$$

In other words, the maximal number of patterns that can be memorized by the network is expressed as $M_c \approx 0.05\cdot N$. It is more than two times less than the storage capacity for the standard Hopfield model ($0.138\cdot N$), but the catastrophic destruction of the memory does not occur. Let us note that the optimal values $q_m(N)$ are rather close to 1: $q_m = 0.992, 0.9992$ and 0.99992 for N =1000, 10000 and 100000, respectively.

When the value of q becomes larger than q_m , the number of recognized patterns decreases. It is clear that as far as the parameter q tends to 1, our model more and more resemble the standard Hopfield model for which the dimensionality N is finite and the number of the patterns M is infinitely large. It is clear that when $q = 1$ the network memory is destroyed: patterns are not recognized by the network. It turns out that the destruction of the memory occurs even before q becomes equal to 1.

Let q_c denotes the critical value of q under which only the first pattern (with the maximal weight) is recognized; it is clear, that $q_c > q_m$. It is possible to obtain an analytic estimate for q_c : $q_c = 1-\delta$, where $\delta \approx 1/0.329N$. Up to now we do not succeeded in analytic estimation of q_m . From the fitting of the results of the numerical solution of Eq.(9) we obtain: $q_m \approx 1 - 2.75\cdot\delta$.

For the last pattern with the number k_m that can be recognized by the network, on the right panel of Fig.3 the dependence of the overlap on q is shown. In the point of the solution "breakdown" q_c all overlaps have approximately the same values $m_c \approx 0.933$.

The results of this section were obtained by means of numerical solution of eq. (9). At the present time we try to obtain these results analytically. Moreover, it is necessary to compare our predictions with the computer simulations in the case when the weights are terms of geometrical progression. First experiments show good agreement with theoretical predictions. It is interesting to analyze other distributions of the weights r_μ differ from a decreasing geometrical progression. Now these investigations are in progress.

IV. Conclusions

The common experience indicates that the learning process is a continuous one. In the brain there are mechanisms allowing one to accumulate steadily the obtained information. The Hopfield model is an artificial model of the human ability of learning. For this reason the catastrophic destruction of the memory due to too much number of patterns is absolutely inadmissible. An artificial memory of a robot also has to work even if it obtains new information continuously written down.

The method of eliminating of catastrophic destruction of the memory, proposed in our paper, seems to be very attractive. It turns out that it is possible to learn the network uninterruptedly, even during the process of the patterns recognition. Indeed, each pattern that finds itself in the field of vision of a robot can be automatically added to the connection matrix. Each pattern modifies matrix elements according to the standard Hebbian rule. If the pattern is the same as the one written down previously, its weight increases by 1. If the pattern is new, it is written down into the connection matrix with the initial weight equals to 1. According to the statistics of the patterns appearance the weights distribution is online modified. There is no catastrophic destruction of the memory since every moment the real memory of the network consists of patterns, whose weights are larger than the current critical value r_c .

Let the weight of a pattern be less than r_c but we need this pattern to be recognized by the network. It is sufficient to increase the weight of this pattern doing it to be larger than the critical value, and this pattern will be recognized. It is possible that at the same time some other patterns cease to be recognized. (They are those whose weights are only slightly exceeds the critical value r_c). Such replacement of patterns by other ones does not contradict to the common sense. It corresponds to the general conception of the human memory.

ACKNOWLEDGMENT

The work was supported by the program of Russian Academy of Sciences “Information technologies and analysis of complex systems” (project 1.7) and in part by Russian Basic Research Foundation (grant 09-07-00159).

REFERENCES

[1] D. Amit, H. Gutfreund, and H. Sompolinsky, “Statistical mechanics of neural networks near saturation,” Annals of Physics, vol. 173, pp. 30-67, 1987.

[2] J. Hertz, A. Krogh, and R. Palmer, Introduction to the Theory of Neural Computation. Massachusetts: Addison-Wesley, 1991.

[3] G. Parisi, “A memory which forgets,” Journal of Physics A vol. 19, pp. L617-L620, 1986.

[4] J.P. Nadal, G. Toulouse, J.P. Changeux, and S. Dehaene, “Networks of Formal Neurons and Memory Palimpsest,” Europhysics Letters vol.1(10), pp. 535-542, 1986.

[5] J.L. van Hemmen, G. Keller, and R. Kuhn, “Forgetful Memories,” Europhysics Letters, vol. 5(7), pp. 663-668, 1988.

[6] J.L. van Hemmen and R. Kuhn, “Collective Phenomena in Neural Networks,”. in Models of Neural Networks, E. Domany, J.L van Hemmen and K. Shulten, Eds. Berlin: Springer, 1992, pp. 1-105.

[7] A. Sandberg, O. Ekeberg, and A. Lansner, “An incremental Bayesian learning rule for palimpsest memory,” preprint.

[8] Ja. Karandashev, B. Kryzhanovsky, and L. Litinskii, “Local Minima of a Quadratic Binary Functional with a Quasi-Hebbian Connection Matrix,” Proc. 20th International Conference on Artificial Neural Networks (ICANN’10) Springer, 2010, Part 3, pp. 41-50, LNCS 6352.

[9] Ya. Karandashev, B. Kryzhanovsky, and L. Litinskii, “Local Minima of a Quadratic Binary Functional with a Quasi-Hebbian Connection Matrix,” arXiv:1012.4981v1.

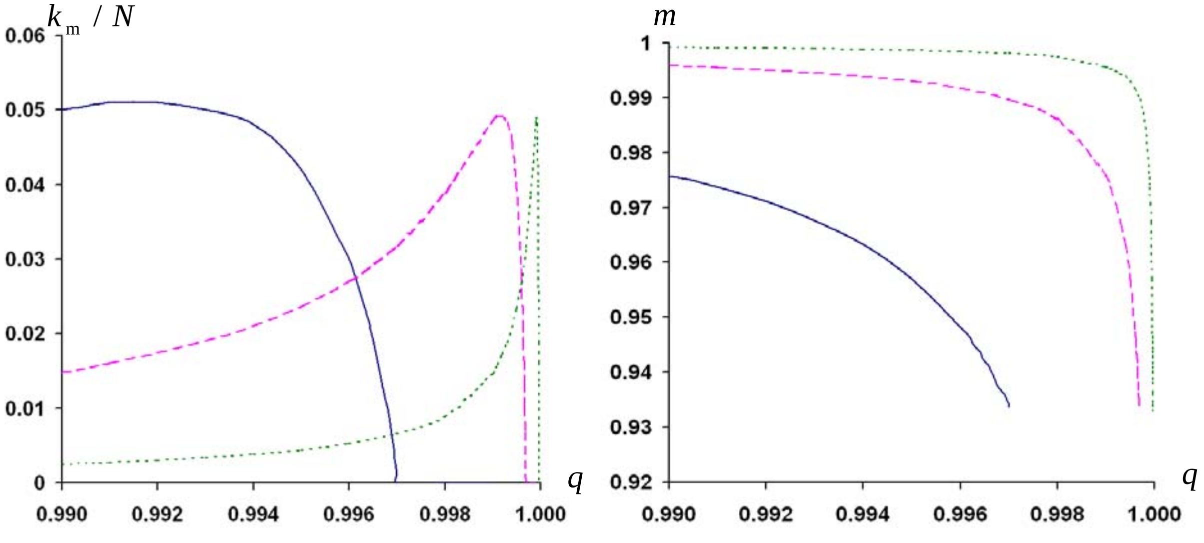

Figure 3. For three dimensionalities N we show: on the left panel the dependence k_m / N on q; on the right panel synchronous values of the pattern overlaps with the nearest fixed point. On both panels the solid line corresponds to the dimensionality N=1000, the dashed line corresponds to N=10000, the point line corresponds to N=100000.