BINARY OPTIMIZATION: A RELATION BETWEEN THE DEPTH OF A LOCAL MINIMUM AND THE PROBABILITY OF ITS DETECTION

B. V. Kryzhanovsky, V. M. Kryzhanovsky and A. L. Mikaelian
Center of Optical Neural Technologies, SR Institute of System Analisys RAS
44/2 Vavilov Str, Moscow 119333, Russia

Keywords: Binary optimization, neural networks, local minimum.

Abstract: The standard method in optimization problems consists in a random search of the global minimum: a neuron network relaxes in the nearest local minimum from some randomly chosen initial configuration. This procedure is to be repeated many times in order to find as deep an energy minimum as possible. However the question about the reasonable number of such random starts and whether the result of the search can be treated as successful remains always open. In this paper by analyzing the generalized Hopfield model we obtain expressions describing the relationship between the depth of a local minimum and the size of the basin of attraction. Based on this, we present the probability of finding a local minimum as a function of the depth of the minimum. Such a relation can be used in optimization applications: it allows one, basing on a series of already found minima, to estimate the probability of finding a deeper minimum, and to decide in favor of or against further running the program. The theory is in a good agreement with experimental results.

1 INTRODUCTION

Usually a neural system of associative memory is considered as a system performing a recognition or retrieval task. However it can also be considered as a system that solves an optimization problem: the network is expected to find a configuration minimizes an energy function (Hopfield,1982). This property of a neural network can be used to solve different *NP*-complete problems. A conventional approach consists in finding such an architecture and parameters of a neural network, at which the objective function or cost function represents the neural network energy. Successful application of neural networks to the traveling salesman problem (Hopfield and Tank, 1985) had initiated extensive investigations of neural network approaches for the graph bipartition problem (Fu and Anderson, 1986), neural network optimization of the image processing (Poggio and Girosi, 1990) and many other applications. This subfield of the neural network theory is developing rapidly at the moment (Smith, 1999), (Hartmann and Rieger, 2004), (Huajin Tang et al, 2004), (Kwok and Smith, 2004), (Salcedo-Sanz et al, 2004), (Wang et al, 2004, 2006).

The aforementioned investigations have the same common feature: the overwhelming majority of neural network optimization algorithms contain the Hopfield model in their core, and the optimization process is reduced to finding the global minimum of some quadratic functional (the energy) constructed on a given $N \times N$ matrix in an N-dimensional configuration space (Joya, 2002), (Kryzhanovsky et al, 2005). The standard neural network approach to such a problem consists in a random search of an optimal solution. The procedure consists of two stages. During the first stage the neural network is initialized at random, and during the second stage the neural network relaxes into one of the possible stable states, i.e. it optimizes the energy value. Since the sought result is unknown and the search is done at random, the neural network is to be initialized many times in order to find as deep an energy minimum as possible. But the question about the reasonable number of such random starts and whether the result of the search can be regarded as successful always remains open.

In this paper we have obtained expressions that have demonstrated the relationship between the depth of a local minimum of energy and the size of

V. Kryzhanovsky B., M. Kryzhanovsky V. and L. Mikaelian A. (2007).
BINARY OPTIMIZATION: A RELATION BETWEEN THE DEPTH OF A LOCAL MINIMUM AND THE PROBABILITY OF ITS DETECTION.
In *Proceedings of the Fourth International Conference on Informatics in Control, Automation and Robotics*, pages 5-10

the basin of attraction (Kryzhanovsky et al, 2006). Based on this expressions, we presented the probability of finding a local minimum as a function of the depth of the minimum. Such a relation can be used in optimization applications: it allows one, based on a series of already found minima, to estimate the probability of finding a deeper minimum, and to decide in favor of or against further running of the program. Our expressions are obtained from the analysis of generalized Hopfield model, namely, of a neural network with Hebbian matrix. They are however valid for any matrices, because any kind of matrix can be represented as a Hebbian one, constructed on arbitrary number of patterns. A good agreement between our theory and experiment is obtained.

2 DESCRIPTION OF THE MODEL

Let us consider Hopfield model, i.e a system of N Ising spins-neurons $s_i = \pm 1$, $i = \overline{1,N}$. A state of such a neural network can be characterized by a configuration $\boldsymbol{S} = (s_1, s_2, ..., s_N)$. Here we consider a generalized model, in which the connection matrix:

$$T_{ij} = \sum_{m=1}^{M} r_m s_i^{(m)} s_j^{(m)}, \qquad \sum r_m^2 = 1 \tag{1}$$

is constructed following Hebbian rule on M binary N-dimensional patterns $\boldsymbol{S}_m = (s_1^{(m)}, s_2^{(m)}, ..., s_N^{(m)})$, $m = \overline{1,M}$. The diagonal matrix elements are equal to zero ($T_{ii} = 0$). The generalization consists in the fact, that each pattern $\boldsymbol{S}_m$ is added to the matrix T_{ij} with its statistical weight r_m. We normalize the statistical weights to simplify the expressions without loss of generality. Such a slight modification of the model turns out to be essential, since in contrast to the conventional model it allows one to describe a neural network with a non-degenerate spectrum of minima.

The energy of the neural network is given by the expression:

$$E = -\tfrac{1}{2} \sum_{i,j=1}^{N} s_i T_{ij} s_j \tag{2}$$

and its (asynchronous) dynamics consist in the following. Let $\boldsymbol{S}$ be an initial state of the network. Then the local field $h_i = -\partial E / \partial s_i$, which acts on a randomly chosen i-th spin, can be calculated, and the energy of the spin in this field $\varepsilon_i = -s_i h_i$ can be determined. If the direction of the spin coincides with the direction of the local field ($\varepsilon_i < 0$), then its state is stable, and in the subsequent moment ($t+1$) its state will undergo no changes. In the opposite case ($\varepsilon_i > 0$) the state of the spin is unstable and it flips along the direction of the local field, so that $s_i(t+1) = -s_i(t)$ with the energy $\varepsilon_i(t+1) < 0$. Such a procedure is to be sequentially applied to all the spins of the neural network. Each spin flip is accompanied by a lowering of the neural network energy. It means that after a finite number of steps the network will relax to a stable state, which corresponds to a local energy minimum.

3 BASIN OF ATTRACTION

Let us examine at which conditions the pattern $\boldsymbol{S}_m$ embeded in the matrix (1) will be a stable point, at which the energy E of the system reaches its (local) minimum E_m. In order to obtain correct estimates we consider the asymptotic limit $N \to \infty$. We determine the basin of attraction of a pattern $\boldsymbol{S}_m$ as a set of the points of N-dimensional space, from which the neural network relaxes into the configuration $\boldsymbol{S}_m$. Let us try to estimate the size of this basin. Let the initial state of the network $\boldsymbol{S}$ be located in a vicinity of the pattern $\boldsymbol{S}_m$. Then the probability of the network convergation into the point $\boldsymbol{S}_m$ is given by the expression:

$$\mathbf{Pr} = \left(\frac{1 + \mathbf{erf}\, \gamma}{2} \right)^N \tag{3}$$

where $\mathbf{erf}\, \gamma$ is the error function of the variable γ:

$$\gamma = \frac{r_m \sqrt{N}}{\sqrt{2(1 - r_m^2)}} \left(1 - \frac{2n}{N} \right) \tag{4}$$

and n is Hemming distance between $\boldsymbol{S}_\mathrm{m}$ and $\boldsymbol{S}$. The expression (3) can be obtained with the help of the methods of probability theory, repeating the well-known calculation (Perez-Vincente, 1989) for conventional Hopfield model.

It follows from (3) that the basin of attraction is determined as the set of the points of the configuration space close to $\boldsymbol{S}_m$, for which $n \le n_m$:

$$n_m = \frac{N}{2}\left(1 - \frac{r_0\sqrt{1-r_m^2}}{r_m\sqrt{1-r_0^2}}\right) \qquad (5)$$

where

$$r_0 = \sqrt{2\ln N / N} \qquad (6)$$

Indeed, if $n \le n_m$ we have $\mathbf{Pr} \to 1$ for $N \to \infty$, i.e. the probability of the convergence to the point $\boldsymbol{S}_m$ asymptotically tends to 1. In the opposite case ($n > n_m$) we have $\mathbf{Pr} \to 0$. It means that the quantity n_m can be considered as the radius of the basin of attraction of the local minimum E_m.

It follows from (5) that the radius of basin of attraction tends to zero when $r_m \to r_0$ (Fig.1). It means that the patterns added to the matrix (1), whose statistical weight is smaller than r_0, simply do not form local minima. Local minima exist only in those points $\boldsymbol{S}_m$, whose statistical weight is relatively large: $r_m > r_0$.

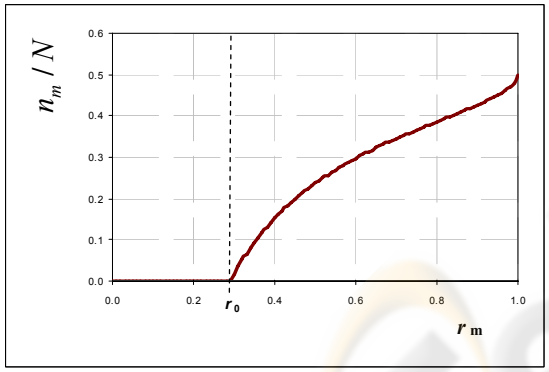

Figure 1: A typical dependence of the width of the basin of attraction n_m on the statistical weight of the pattern r_m. A local minimum exists only for those patterns, whose statistical weight is greater than r_0. For $r_m \to r_0$ the size of the basin of attraction tends to zero, i.e. the patters whose statistical weight $r_m \le r_0$ do not form local minima.

4 DEPTH OF LOCAL MINIMUM

From analysis of Eq. (2) it follows that the energy of a local minimum E_m can be represented in the form:

$$E_m = -r_m N^2 \qquad (7)$$

with the accuracy up to an insignificant fluctuation of the order of

$$\sigma_m = N\sqrt{1-r_m^2} \qquad (8)$$

Then, taking into account Eqs. (5) and (7), one can easily obtain the following expression:

$$E_m = E_{\min}\frac{1}{\sqrt{(1-2n_m/N)^2 + E_{\min}^2/E_{\max}^2}} \qquad (9)$$

where

$$E_{\min} = -N\sqrt{2N\ln N}\ , \quad E_{\max} = -\left(\sum_{m=1}^{M} E_m^2\right)^{1/2} \qquad (10)$$

which yield a relationship between the depth of the local minimum and the size of its basin of attraction. One can see that the wider the basin of attraction, the deeper the local minimum and vice versa: the deeper the minimum, the wider its basin of attraction (see Fig.2).

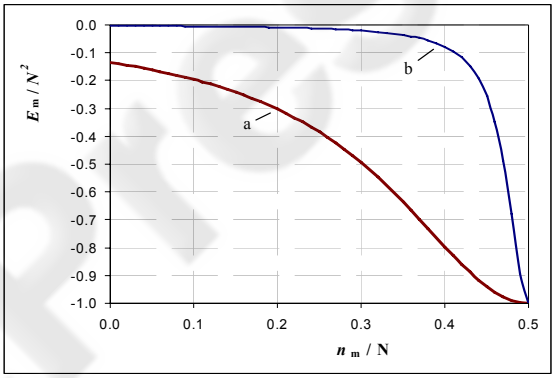

Figure 2: The dependence of the energy of a local minimum on the size of the basin of attraction: a) N=50; b) N=5000.

We have introduced here also a constant E_{max}, which we make use of in what follows. It denotes the maximal possible depth of a local minimum. In the adopted normalization, there is no special need to introduce this new notation, since it follows from (7)-(9) that $E_{\max} = -N^2$. However for other normalizations some other dependencies of $E_{\max}$ on N are possible, which can lead to a misunderstanding.

The quantity E_{min} introduced in (10) characterizes simultaneously two parameters of the neural network. First, it determines the half-width of the Lorentzian distribution (9). Second, it follows from (9) that:

$$E_{\max} \le E_m \le E_{\min} \qquad (11)$$

i.e. $E_{\min}$ is the upper boundary of the local minimum spectrum and characterizes the minimal possible depth of the local minimum. These results are in a good agreement with the results of computer experiments aimed to check whether there is a local minimum at the point $\boldsymbol{S}_{\mathrm{m}}$ or not. The results of one of these experiments (N=500, M=25) are shown in Fig.3. One can see a good linear dependence of the energy of the local minimum on the value of the statistical weight of the pattern. Note that the overwhelming number of the experimental points corresponding to the local minima are situated in the right lower quadrant, where $r_m > r_0$ and $E_{\mathrm{m}} < E_{\min}$. One can also see from Fig.3 that, in accordance with (8), the dispersion of the energies of the minima decreases with the increase of the statistical weight.

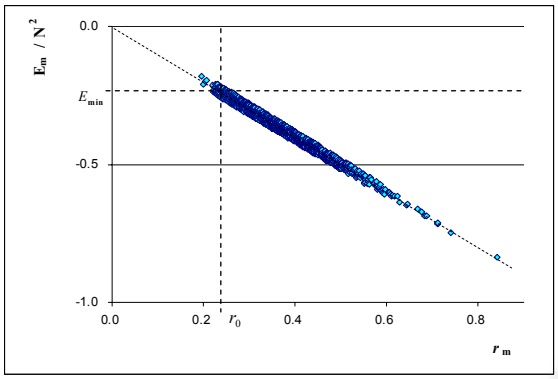

Figure 3: The dependence of the energy E_m of a local minimum on the statistical weight r_m of the pattern.

5 THE PROBABILITY OF FINDING THE MINIMUM

Let us find the probability W of finding a local minimum E_m at a random search. By definition, this probability coincides with the probability for a randomly chosen initial configuration to get to the basin of attraction of the pattern $\boldsymbol{S}_m$. Consequently, the quantity $W = W(n_m)$ is the number of points in a sphere of a radius n_m, reduced to the total number of the points in the N-dimensional space:

$$W = 2^{-N} \sum_{n=1}^{n_m} C_N^n \qquad (12)$$

Equations (5) and (12) define implicitly a connection between the depth of the local minimum and the probability of its finding. Applying asymptotical Stirling expansion to the binomial coefficients and passing from summation to integration one can represent (12) as

$$W = W_0 \, e^{-Nh} \qquad (13)$$

where h is generalized Shannon function

$$h = \frac{n_m}{N} \ln \frac{n_m}{N} + \left(1 - \frac{n_m}{N}\right) \ln\left(1 - \frac{n_m}{N}\right) + \ln 2 \qquad (14)$$

Here W_0 is an insignificant for the further analysis slow function of E_m. It can be obtained from the asymptotic estimate (13) under the condition $n_m >> 1$, and the dependence $W = W(n_m)$ is determined completely by the fast exponent.

It follows from (14) that the probability of finding a local minimum of a small depth ($E_m \sim E_{\min}$) is small and decreases as $W \sim 2^{-N}$. The probability W becomes visibly non-zero only for deep enough minima $|E_m| >> |E_{\min}|$, whose basin of attraction sizes are comparable with $N/2$. Taking into account (9), the expression (14) can be transformed in this case to a dependence $W = W(E_m)$ given by

$$W = W_0 \exp\left[- N E_{\min}^2 \left(\frac{1}{E_m^2} - \frac{1}{E_{\max}^2} \right) \right] \qquad (15)$$

It follows from (14) that the probability to find a minimum increases with the increase of its depth. This dependence"the deeper minimum $\rightarrow$ the larger the basin of attraction $\rightarrow$ the larger the probability to get to this minimum" is confirmed by the results of numerous experiments. In Fig.4 the solid line is computed from Eq. (13), and the points correspond to the experiment (Hebbian matrix with a small loading parameter $M / N \le 0.1$). One can see that a good agreement is achieved first of all for the deepest minima, which correspond to the patterns $\boldsymbol{S}_m$ (the energy interval $E_m \le -0.49 N^2$ in Fig.4). The experimentally found minima of small depth (the points in the region $E_m > -0.44 N^2$) are the so-called "chimeras". In standard Hopfield model ($r_m \equiv 1/\sqrt{M}$) they appear at relatively large loading parameter $M / N > 0.05$. In the more general case, which we consider here, they can appear also earlier. The reasons leading to their appearance are well examined with the help of the methods of statistical physics in (Amit et al, 1985), where it was shown that the chimeras appear as a consequence of

interference of the minima of $\boldsymbol{S}_m$. At a small loading parameter the chimeras are separated from the minima of $\boldsymbol{S}_m$ by an energy gap clearly seen in Fig.4.

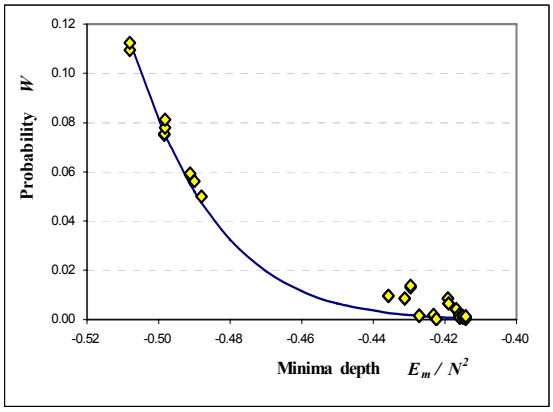

Figure 4: The dependence of the probability W to find a local minimum on its depth E_m: theory - solid line, experiment – points.

6 DISCUSSION

Our analysis shows that the properties of the generalized model are described by three parameters r_0, $E_{\min}$ and $E_{\max}$. The first determines the minimal value of the statistical weight at which the pattern forms a local minimum. The second and third parameters are accordingly the minimal and the maximal depth of the local minima. It is important that these parameters are independent from the number of embeded patterns M.

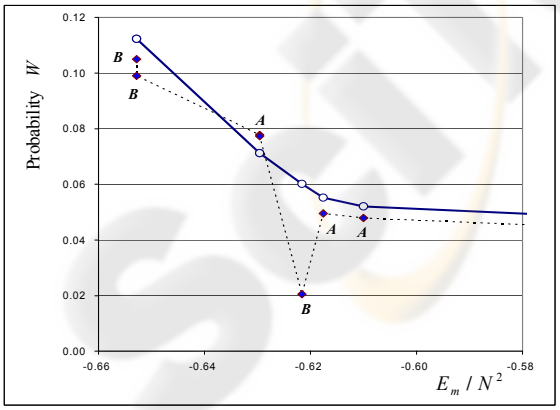

Figure 5: The comparison of the predicted probabilities (solid line) and the experimentally found values (points connected with the dashed line).

Now we are able to formulate a heuristic approach of finding the global minimum of the functional (2) for any given matrix (not necessarily Hebbian one). The idea is to use the expression (15) with unknown parameters W_0, $E_{\min}$ and $E_{\max}$. To do this one starts the procedure of the random search and finds some minima. Using the obtained data, one determines typical values of $E_{\min}$ and $E_{\max}$ and the fitting parameter W_0 for the given matrix. Substituting these values into (15) one can estimate the probability of finding an unknown deeper minimum E_m (if it exists) and decide in favor or against (if the estimate is a pessimistic one) the further running of the program.

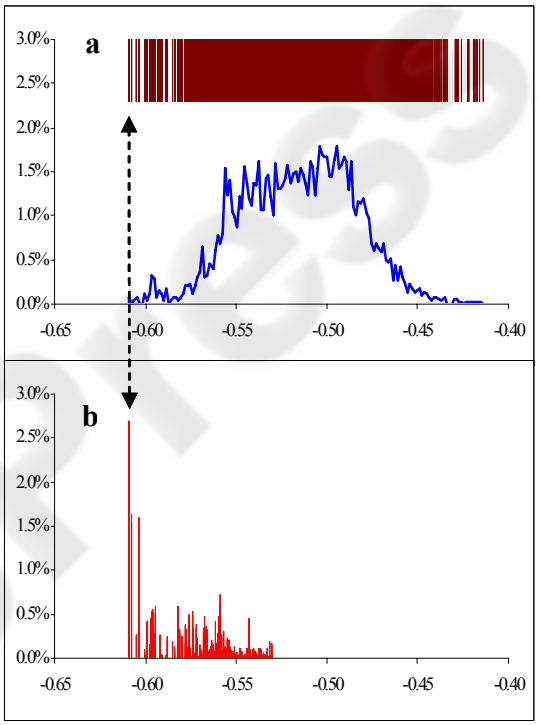

Figure 6: The case of matrix with a quasi-continuous type of spectrum. a) The upper part of the figure shows the spectrum of minima distribution – each vertical line corresponds to a particular minimum. The solid line denotes the spectral density of minima (the number of minima at length ΔE). The Y-axis presents spectral density and the X-axis is the normalized values of energy minima E / N^2. b) Probability of finding a minimum with energy E. The Y-axis is the probability of finding a particular minimum (%) and the X-axis is the normalized values of energy minima.

This approach was tested with Hebbian matrices at relatively large values of the loading parameter ($M / N \geq 0.2 \div 10$). The result of one of the experiments is shown in Fig.5. In this experiment

with the aid of the found minima (the points A) the parameters W_0, $E_{\min}$ and $E_{\max}$ were calculated, and the dependence $W = W(E_m)$ (solid line) was found. After repeating the procedure of the random search over and over again ($\sim 10^5$ random starts) other minima (points B) and the precise probabilities of getting into them were found. One can see that although some dispersion is present, the predicted values in the order of magnitude are in a good agreement with the precise probabilities.

In conclusion we stress once again that any given matrix can be performed in the form of Hebbian matrix (1) constructed on an arbitrary number of patterns (for instance, $M \to \infty$) with arbitrary statistical weights. It means that the dependence "the deeper minimum $\leftrightarrow$ the larger the basin of attraction $\leftrightarrow$ the larger the probability to get to this minimum" as well as all other results obtained in this paper are valid for all kinds of matrices. To prove this dependence, we have generated random matrices, with uniformly distributed elements on [-1,1] segment. The results of a local minima search on one of such matrices are shown in Fig. 6. The value of normalized energy is shown on the X-scale and on the Y-scale the spectral density is noted. As we can see, there are a lot of local minima, and most of them concentrated in central part of spectrum (Fig 6.a). Despite of such a complex view of the spectrum of minima, the deepest minimum is found with maximum probability (Fig 6.b). The same perfect accordance of the theory and the experimental results are also obtained in the case of random matrices, the elements of which are subjected to the Gaussian distribution with a zero mean.

The work supported by RFBR grant # 06-01-00109.

REFERENCES

Amit, D.J., Gutfreund, H., Sompolinsky, H., 1985. Spin-glass models of neural networks. Physical Review A, v.32, pp.1007-1018.

Fu, Y., Anderson, P.W., 1986. Application of statistical mechanics to NP-complete problems in combinatorial optimization. Journal of Physics A. , v.19, pp.1605-1620.

Hartmann, A.K., Rieger, H., 2004. New Optimization Algorithms in Physics., Wiley-VCH, Berlin.

Hopfield, J.J. 1982. Neural Networks and physical systems with emergent collective computational abilities. Proc. Nat. Acad. Sci.USA. v.79, pp.2554-2558 .

Hopfield, J.J., Tank, D.W., 1985. Neural computation of decisions in optimization problems. Biological Cybernetics, v.52, pp.141-152.

Huajin Tang; Tan, K.C.; Zhang Yi, 2004. A columnar competitive model for solving combinatorial optimization problems. IEEE Trans. Neural Networks v.15, pp.1568 – 1574.

Joya, G., Atencia, M., Sandoval, F., 2002. Hopfield Neural Networks for Optimization: Study of the Different Dynamics. Neurocomputing, v.43, pp. 219-237.

Kryzhanovsky, B., Magomedov, B., 2005. Application of domain neural network to optimization tasks. Proc. of ICANN'2005. Warsaw. LNCS 3697, Part II, pp.397-403.

Kryzhanovsky, B., Magomedov, B., Fonarev, A., 2006. On the Probability of Finding Local Minima in Optimization Problems. Proc. of International Joint Conf. on Neural Networks IJCNN-2006 Vancouver, pp.5882-5887.

Kwok, T., Smith, K.A., 2004. A noisy self-organizing neural network with bifurcation dynamics for combinatorial optimization. IEEE Trans. Neural Networks v.15, pp.84 – 98.

Perez-Vincente, C.J., 1989. Finite capacity of sparce-coding model. Europhys. Lett., v.10, pp.627-631.

Poggio, T., Girosi, F., 1990. Regularization algorithms for learning that are equivalent to multilayer networks. Science 247, pp.978-982.

Salcedo-Sanz, S.; Santiago-Mozos, R.; Bousono-Calzon, C., 2004. A hybrid Hopfield network-simulated annealing approach for frequency assignment in satellite communications systems. IEEE Trans. Systems, Man and Cybernetics, v. 34, 1108 – 1116

Smith, K.A. 1999. Neural Networks for Combinatorial Optimization: A Review of More Than a Decade of Research. INFORMS Journal on Computing v.11 (1), pp.15-34.

Wang, L.P., Li, S., Tian F.Y, Fu, X.J., 2004. A noisy chaotic neural network for solving combinatorial optimization problems: Stochastic chaotic simulated annealing. IEEE Trans. System, Man, Cybern, Part B - Cybernetics v.34, pp. 2119-2125.

Wang, L.P., Shi, H., 2006: A gradual noisy chaotic neural network for solving the broadcast scheduling problem in packet radio networks. IEEE Trans. Neural Networks, vol.17, pp.989 - 1000.