Statistical physics of interacting neural networks

Wolfgang Kinzel [1], Richard Metzler [1] and Ido Kanter [2]

(1) Institute for Theoretical Physics, University of Würzburg,
Am Hubland, 97074 Würzburg, Germany
(2)Bar Ilan University xxx

June 2001

Abstract

Recent results on the statistical physics of time series generation and prediction are presented. Training a neural network on quasiperiodic and chaotic sequences produces different overlaps to the sequence generator and prediction errors. For each network there exists a sequence for which it fails completely. Two interacting networks show a transition to perfect synchronization. A pool of interacting networks solves the minority game – a model of competition in a closed market. Finally, as a demonstration, a percptron predicts bit sequences produced by human beings.

1 Introduction

Since 1985 the theory of neural networks profitted from the contribution of statistical physics [1, 2, 3, 4, 5]. In the limit of infinitely large networks and for a set of random training examples there exist mathematical tools to calculate the properties of a sytem of interacting neurons and synapses exactly. In particluar one phenomenon has been investigated in great detail: A neural network can earn a rule from a set of examples. A network called "the teacher" produces a set of input/output pairs by which a different network, called the "student" is trained. After traning the student has not only learned the examples but has developed an overlap to the unknown teacher, it has learned to generalize.

Since 1995 the statistical physics of time series prediction has been studied as well [6, 7, 8, 9, 10, 11, 12, 13, 14]. Similar to the static case the series is

generated by a unknown rule - usually a different "teacher"–network - and a "student" network is trained on these data by moving it over the series, as shown in Fig.1.

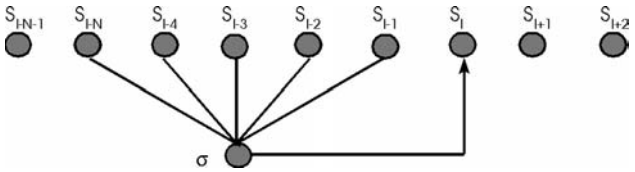

Figure 1: A percptron moves over a time series.

In this talk we present some new results of the statistical physics of time series generation and prediction. We consider only a simple perceptron with one layer of synaptic weights, but we use different transfer functions $f(x)$:

$$\sigma = \text{sign}(\underline{w} \cdot \underline{S}); \tag{1}$$

$$\sigma = \tanh\left(\frac{\beta}{N}\, \underline{w} \cdot \underline{S}\right); \tag{2}$$

$$\sigma = \sin\left(\frac{\beta}{N}\, \underline{w} \cdot \underline{S}\right). \tag{3}$$

β is a parameter giving the slope of the linear part of the transfer function in the continuous cases, $f(x) \simeq \beta x + O(x^3)$.

The aim of our network is to learn a given sequence $S_0, S_1, S_2, \ldots$. This means that the network should find - by some simple algorithms - a weight vector $\underline{w}$ with the property

$$S_t = f\left(\frac{1}{N}\sum_{j=1}^{N} w_j S_{t-j}\right) \tag{4}$$

for all time steps t. For the Boolean function Eq.(1) this set of equations becomes a set of inequalities

$$S_t \sum_{j=1}^{N} w_j S_{t-j} > 0 \tag{5}$$

for all t.

2 Predicting time series

If a neural network cannot generate a given sequence of numbers, it cannot predict it with zero error. But this is not the whole story. Even if the sequence has been generated by an (unknown) neural network (the teacher), a different network (the student) can try to learn and to predict this sequence. In this context we are interested in two questions:

1. When a student network with the identical architecture as the teacher one is trained on the sequence, how does the overlap between student and teacher develop with the number of training examples (= windows of the sequence)?

2. After the student network has been trained on a part of the sequence how well can it predict the sequence several steps ahead?

Recently these questions have been investigated numerically for the simple perceptron [17]. We have to distinguish several scenarios:

1. Boolean versus continuous perceptron

2. On–line versus batch learning

3. Quasiperiodic versus chaotic sequence.

In all cases we consider only the stationary part of a sequence which was generated by a perceptron. The student network is trained on the stationary part only, not on the transient.

First we discuss the Boolean perceptron of size N which has generated a bit cycle with a typical length $L < 2N$. The teacher perceptron has random weights with zero bias, and the cycle is related to one component of the power spectum of the weights. The student network is trained using the perceptron learning rule:

$$\begin{aligned} \Delta w_i &= \tfrac{1}{N} S_t \; S_{t-i} \quad \text{if} \;\; S_t \sum_{j=1}^{N} w_j S_{t-j} < 0; \\ \Delta w_i &= 0 \;\; \text{else.} \end{aligned} \tag{6}$$

For this algorithm there exists a mathematical theorem [1]: If the set of examples can be generated by some perceptron then this algorithm stops, i.e. it finds one out of possibly many solutions. Since we consider examples from a bit sequence generated by a perceptron, this algorithm is guaranteed

to learn the sequence perfectly. On–line and batch training are identical, in this case.

The network is trained on the cycle until the training error is zero. Hence the student network can predict the stationary sequence perfectly. Surprisingly, it turns out that the overlap between student and teacher is small, in fact it is zero for infinitely large networks, $N \to \infty$. The network learns the projection of the teacher's weight vector onto the sequence, but not the complete vector. It behaves like a filter selecting one of the components of the power spectrum of the weights. Although it predicts the sequence perfectly, it does not gain much information on the rule which generates this sequence.

This situation seems to be different in the case of a continuous perceptron. Inverting Eq.(4) for a monotonic transfer function $f(x)$ gives N linear equations for N unknowns w_i. If the stationary part of the sequence is either quasiperiodic or chaotic, all patterns are different and the batch training, using N windows, leads to perfect learning.

This holds true for a chaotic time series. However, for a quasiperiodic one the patterns are almost linearly dependent, yielding an ill–conditioned set of linear equations. Without the $\tanh(x)$ in Eq.(4), one would obtain a two–dimensional space of patterns; with the nonlinearity one obtains small contributions in the other $N-2$ dimensions of the weight space. Nevertheless, depending on the parameter β, even professional computer routines sometimes do not succeed in solving Eq.(4) for quasiperiodic patterns generated by a teacher perceptron.

How does this scenario show up in an on–line training algorithm for a continuous perceptron? If a quasiperiodic sequence is learned step by step without iterating previous steps, using gradient descent to update the weights,

$$\Delta w_i = \frac{\eta}{N}(S_t - f(h)) \cdot f'(h) \cdot S_{t-i} \;\; \text{with} \;\; h = \beta \sum_{j=1}^{N} w_j S_{t-j} \tag{7}$$

then one can distinguish two time scales (time = number of training steps):

1. A fast one increasing the overlap between teacher and student to a value which is still far away from the value one which corresponds to perfect agreement.

2. A slow one increasing the overlap very slowly. Numerical simulations for millions times N training steps yielded an overlap which was still far away from the value one.

Although there is a mathematical theorem on stochastic optimization which seems to guarantee convergence to perfect success [15], our on–line

algorithm cannot gain much information about the teacher network. It would be interesting to know how these two time scales depend on the size of the system. In addition we cannot exclude that there exist on-line algorithms which can learn our ill-conditioned problem in short times.

This is completely different for a chaotic time series generated by a corresponding teacher network with $f(x) = sin(x)$. It turns out that the chaotic series appears like a random one: After a number of training steps of the order of N the overlap relaxes exponentially fast to perfect agreement between teacher and student.

Hence, after training the perceptron with a number of examples of the order of N we obtain the two cases: For a quasiperiodic sequence the student has not obtained much information about the teacher, while for a chaotic sequence the student's weight vector comes close to the one of the teacher. One important question remains: How well can the student predict the time series?

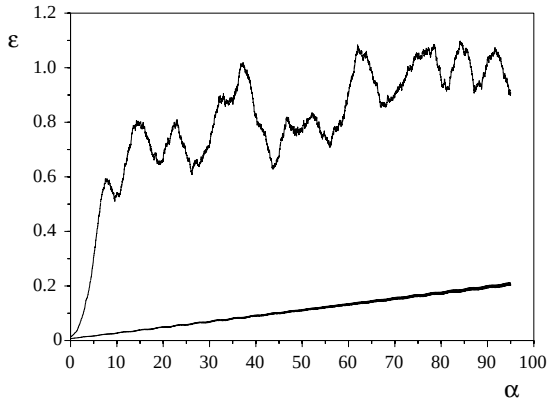

Figure 2: Prediction error as a function of time steps ahead, for a quasiperiodic (lower) and chaotic (upper) series.

Fig.2 shows the prediction error as a function of the time interval over which the student makes the predictions. The student network which has been trained on the quasiperiodic sequence can predict it very well. The error increases linearly with the size of the interval, even predicting $10N$ steps ahead yields an error of about 10% of the total possible range. On the other side, the student trained on the chaotic sequence cannot make predictions. The prediction error increases exponentially with time; already after a few steps the error corresponds to random guessing, $\epsilon \simeq 1$.

In summary we obtain the surprising result:

1. A network trained on a quasiperodic sequence does not obtain much information about the teacher network which generated the sequence. But the network can predict this sequence over many (of the order of N) steps ahead.

2. A network trained on a chaotic sequence obtains almost complete knowledge about the teacher network. But this network cannot make reasonable predictions on the sequence.

It would be interesting to find out whether this result also holds for other prediction algorithms, such as multilayer networks.

3 Predicting with 100% error

Consider some arbitrary prediction algorithm. It may contain all the knowledge of mankind, many experts may have developed it. Now there is a bit sequence $S_1, S_2...$ and the algorithm has been trained on the first t bits $S_1, ..., S_t$. Can it predict the next bit S_{t+1}? Is the prediction error, averaged over a large t interval, less than 50%?

If the bit sequence is random then every algorithm will give a prediction error of 50%. But if there are some correlations in the sequence then a clever algorithm should be able to reduce this error. In fact, for the most powerful algorithm one is tempted to say that for *any* sequence it should perform better than 50% error. However, this is not true [16]. To see this just generate a sequence $S_1, S_2, S_3, ...$ using the following algorithm:

Define S_{t+i} to be the opposite of the prediction of this algorithm which has been trained on $S_1, \ldots, S_t$.

Now, if the same algorithm is trained on this sequence, it will always predict the following bit with 100% error. Hence there is no general prediction machine; to be successful for a class of problems the algorithm needs some preknowledge about it.

The Boolean perceptron is a very simple prediction algorithm for a bit sequence, in particular with the on–line training algorithm (6). How does the bit sequence look like for which the perceptron completely fails?

Following (6) we just have to take the negative value

$$S_t = -\text{sign}\left(\sum_{j=1}^{N} w_j S_{t-j}\right) \tag{8}$$

and then train the network on this new bit

$$\Delta w_j = +\frac{1}{N} S_t \; S_{t-j}. \tag{9}$$

The perceptron is trained on the opposite (= negative) of its own prediction. Starting from (say) random initial states $S_1, \ldots, S_N$ and weights $\underline{w}$, this procedure generates a sequence of bits $S_1, S_2, \ldots S_t, \ldots$ and of vectors $\underline{w}, \underline{w}(1), \underline{w}(2),$
$\ldots \underline{w}(t), \ldots$ as well. Given this sequence and the same initial state, the perceptron which is trained on it yields a prediction error of 100%.

It turns out that this simple algorithm produces a rather complex bit sequence which comes close to a random one. After a transient time the weight vector $\underline{w}(t)$ seems to perform a kind of random walk on a N–dimensional hypersphere. The bit sequence runs to a cycle whose average length L scales exponentially with N,

$$L \simeq 2.2^N. \tag{10}$$

The autocorrelation function of the sequence shows complex properties: It is close to zero up to N, oscillates between N and $3N$ and it is similar to random noise for larger distances. Its entropy is smaller than the one of a random sequence since the frequency of some patterns is suppressed. Of course, it is not random since the prediction error is 100% instead of 50% for a random bit sequence.

When a second perceptron (=student) with different initial state $\underline{w}$ is trained on such a antipredictable sequence generated by Eq.(6) it can perform somewhat better than the teacher: The prediction error goes down to about 78% but it is still larger than 50% for random guessing. However, the student obtains knowledge about the teacher: The angle between the two weight vectors relaxes to about 45 degrees.

4 Learning from each other

In the previous section we have discussed a neural network which learns from itself. But more interesting may be the scenario where several networks are interacting, learning from each other. After all, our living world consists of interacting adaptive systems and recent methods of computer science use interacting agents to solve complex problems. Here we consider a simple system of interacting perceptrons as a first example to develop a theory of cooperative behaviour of adaptive agents.

Consider K Boolean perceptrons, each of which has an N–dimensional weight vector $\underline{w}^\nu, \nu = 1, \ldots, K$. Each perceptron is receiving the same input

vector $S_1, \ldots, S_N$ and produces its own output bit

$$\sigma^\nu = \mathrm{sign}(\underline{w}^\nu \cdot \underline{S}) \tag{11}$$

Now these networks receive information from their neighbours in a ring–like topology: Perceptron $\underline{w}^\nu$ is trained on the output $\sigma^{\nu-1}$ of perceptron $\underline{w}^{\nu-1}$, and $\underline{w}^1$ is trained on σ^K. Training is performed keeping the length of the weight vectors fixed:

$$\underline{w}^\nu(t+1) = \frac{\underline{w}^\nu(t) + (\eta/N)\sigma^{\nu-1}\underline{S}}{|\underline{w}^\nu(t) + (\eta/N)\sigma^{\nu-1}\underline{S}|} \tag{12}$$

The learning rate η is a parameter controlling the speed of learning.

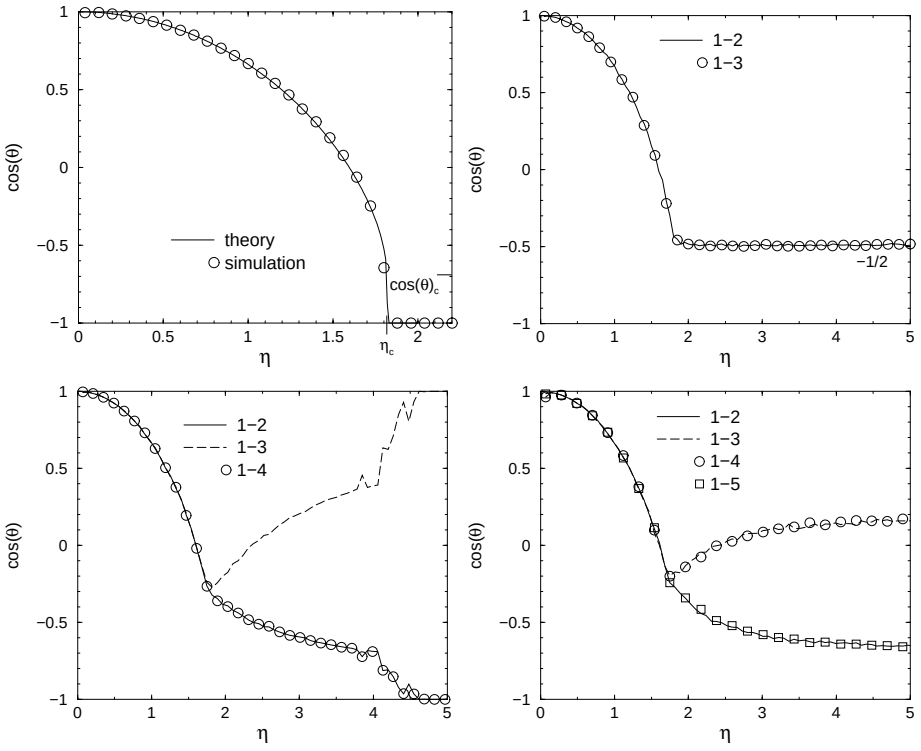

Figure 3: Angles between different networks of a ring of K perceptrons as a function of the learning rate. K takes the values 2,3,5 and 4 clockwise.

This problem has been solved analytically in the limit $N \to \infty$ [14] for random inputs. The system relaxes to a stationary state, where the angles $\theta_{\nu\mu}$

(or overlaps) between different agents take a fixed value. For small learning rate η all of these angles are small, i.e. there is good agreement between the agents. But more surprising: The state of the system is completely symmetric, there is only one common angle $\theta = \theta_{\nu\mu}$ between all pairs of networks. The agents do not recognize the clockwise flow of information.

Increasing the learning rate η the common angle θ increases, too. With larger learning steps the agents tend to have opposite opinion of all of their members. But, due to the symmetry, there is maximal possible angle given by

$$\cos\theta = -\frac{1}{K-1}. \tag{13}$$

In fact, increasing η the system arrives at this maximal angle at some critical value η_c. For larger value of $\eta > \eta_c$ the system undergoes a phase transition: The complete symmetry is broken, but the symmetry of the ring is still conserved:

$$\theta_1 = \theta_{\nu+1,\nu} \quad , \quad \theta_2 = \theta_{\nu+2,\nu}, ...$$

For K agents there are possible $(K-1)/2$ values of θ_i if K is odd, and $K/2-1$ values for even K. These results are demonstrated in Fig.3

This is a simple - but analytically solvable - example of a system of interacting neural networks. We observe a symmetry breaking transition when increasing the learning rate. However, this system does not solve any problem. In the following section we will extend this scenario to a case where indeed neural networks interact to solve a special problem, the minority game.

5 Competing in the minority game

Recently a mathematical model of economy receives a lot of attention in the community of statistical physics [20]. It is a simple model of a closed market: There are K agents who have to make a binary decision $\sigma^\nu \in \{+1, -1\}$ at each time step. All of the agents who belong to the minority gain one point, the majority has to pay one point (to a cashier which always wins). The global loss is given by

$$G = \left| \sum_{\nu=1}^{K} \sigma^\nu \right| \tag{14}$$

If the agents come to an agreement before they make a new decision, it is easy to minimize G : $(K-1)/2$ agents have to choose $+1$, then $G = 1$. However, this is not the rule of the game, the agents are not allowed to cooperate. Each agent knows only the history of the minority decision, $S_1, S_2, S_3, \ldots$,

but otherwise he/she has no information. Can the agent find an algorithm to maximize his/her profit?

If each agent makes a random decision, then G is of the order of $\sqrt{K}$. It is not easy to find algorithms which perform better than random [18, 19].

Here we use a perceptron for each agent to make a decision based on the past N steps $\underline{S} = (S_{t-N}, ..., S_{t-1})$ of the minority decision. The decision of agent $\underline{w}^\nu$ is given by

$$\sigma^\nu = \mathrm{sign}(\underline{w}^\nu \ \underline{S}). \tag{15}$$

After the bit S_t of the minority has been determined, each perceptron is trained on this new example $(\underline{S}, S_t)$,

$$\Delta\underline{w}^\nu = \frac{\eta}{N} \ \ S_t \ \underline{S}. \tag{16}$$

This problem could be solved analytically [14]. The average global loss for $\eta \to 0$ is given by

$$\langle G^2 \rangle = (1 - 2/\pi)K \simeq 0.363 \ K. \tag{17}$$

Hence, for small enough learning rates the system of interacting neural networks performs better than random decisions. A pool of adaptive perceptrons can organize itself to yield a successful cooperation.

6 Predicting human beings

As a final example of a perceptron predicting a bit sequence we discuss a real application. Assume that the bit sequence $S_0, S_1, S_2, ...$ is produced by a human being. Now a simple perceptron (1) with on–line learning (6) takes the last N bits and makes a prediction for the next bit. Then the network is trained on the new true bit, which afterwards appears as part of the input for the following prediction.

Eq.(6) is a simple deterministic equation describing the change of weights according to the new bit and the past N bits. Can such a simple equation foresee the reaction of a human being? On the other side, if a person can calculate or estimate the outcome of Eq. (6), then he/she can just do the opposite, and the network completely fails to predict.

To answer these questions we have written a little C program which receives the two bits 0 and 1 from the keyboard [21]. The program needs two fields `neuron` and `weight` which contain the variable S_i and w_i, respectively. Here are the main steps:

1. Repeat:
   `while (1) {`

2. Calculate the vector product $\underline{w}\,\underline{S}$:
   `for (h=0; i=0; i<N; i++) h+=weight[i]*neuron[i];`

3. Read a new bit:
   `if(getchar()=='1')   input=1; else input =-1;`

4. Calculate the prediction error:
   `if(h*input<0) {error ++;`

5. Train:
   `for(i=0; i<N; i++) weight[i]+=input* neuron[i]/(double)N;}`

6. Shift the input window:
   `for(i=N-1; i>0; i--) neuron[i]=neuron[i-1]; neuron[0] =input;`
   `}`

A graphical version of this program can be called over the internet:

http://theorie.physik.uni-wuerzburg.de/~kinzel

Now we ask a person to generate a bit sequence for which the prediction error of the network is high. We already know from section 2 what happens if the candidate produces a rhythm: if its length is smaller than $1.7N$ the perceptron can learn it perfectly, without errors. Hence the candidate should either produce random numbers which give 50% errors or he/she should try to calculate the prediction of the perceptron, in this case an error higher than 50% is possible.

We have tested this program on students of our class. Each student had to send a file with one thousand bits, generated by hand. It turns out that on average the network predicts with an error of about 35%. The distribution of errors is broad with a range between 20% and 50%. Hence a human being is not a good random number generator. The simple perceptron (1) and (6) succeeds in predicting human behaviour!

Some students submitted sequences with 50% error. It was obvious - and later confessed - that they used random number generators, digits of π, the logistic map, etc. instead of their own fingers. One student submitted a sequence with 100% error. He was the supervisor of our computer system, knew the program and submitted the sequence described in section 5.

7 Summary

The theory of time series generation and prediction is a new field of statistical physics. The properties of perceptrons, simple one–layer neural networks being trained on sequences which were produced by other perceptrons, have been studied.

A perceptron which is trained on a quasiperiodic sequence can predict it very well, but it does not obtain much information on the rule generating the sequence. On the other side, for a chaotic sequence the overlap between student and teacher is almost perfect, but prediction of the sequence is not possible.

For any prediction algorithm there is a sequence for which it completely fails. For a simple perceptron such a sequence is rather complex, with huge cycles and low autocorrelations. Another perceptron which is trained on such a sequence reduces the prediction error from 100% to 78% and obtains overlap to the generating network.

When perceptrons learn from each other, the system relaxes to a symmetric state. Above a critical learning rate there is a phase transition to a state with lower symmetry.

Two perceptrons can perfectly synchronize by mutual learning their output bits.

A system of interacting neural network can develop algorithms for the minority game, a model of a closed economy of competing agents.

Finally it has been demonstrated that human beings are not good random number generators. Even a simple perceptron can predict the bits typed by hand with an error of less than 50%.

Acknowledgement:

The authors thank the Minerva Center and the German-Isreal Science Foundation for support.

References

[1] Hertz, J. and Krogh, A., and Palmer, R.G.: *Introduction to the Theory of Neural Computation*, (Addison Wesley, Redwood City, 1991)

[2] Engel, A. and Van den Broek, C.: *Statistical Mechanics Learning*, Cambridge University Press, 2001)

[3] M. Biehl and N. Caticha: Statistical Mechanics of On-line Learning and Generalization, *The Handbook of Brain Theory and Neural Networks*, ed. by M. A. Arbib (MIT Press, Berlin 2001)

[4] W. Kinzel: *Statistical physics of neural networks*, Comp. Phys. Comm. **121**, 86-93 (1999)

[5] M. Opper and W.Kinzel: Statistical Mechanics of Generalization, *Models of Neural Networks III*, ed. by E. Domany and J.L. van Hemmen and K. Schulten, 151-209 (Springer Verlag, Heidelberg 1995)

[6] E. Eisenstein and I. Kanter and D.A. Kessler and W. Kinzel: *Generation and Prediction of Time Series by a Neural Network*, Phys. Rev. Letters **74** 1, 6-9 (1995)

[7] M. Schröder and W. Kinzel: *Limit cycles of a perceptron*, J. Phys. A **31**, 9131-9147 (1998)

[8] A. Priel and I. Kanter: *Long-term properties of time series generated by a perceptron with various transfer functions*, Phys. Rev. E, **59** 3, 3368-3375 (1999)

[9] M. Schröder and W. Kinzel and I. Kanter: *Training a perceptron by a bit sequence: storage capacity*, J. Phys. A **29**, 7965 (1996)

[10] A. Priel and I. Kanter: *Learning and generation of long-range correlated sequences*, Phys. Rev. E, in press

[11] A. Priel and I. Kanter: *Robust chaos generation by a perceptron*, Europhys. Lett. **51**, 244-250 (2000)

[12] I. Kanter and D.A. Kessler and A. Priel and E. Eisenstein: *Analytical Study of Time Series Generation by Feed-Forward Networks*, Phys. Rev. Lett. **75** 13, 2614-2617 (1995)

[13] L. Ein-Dor and I. Kanter: *Time Series Generation by Multi-layer networks*, Phys. Rev. E **57**, 6564 (1998)

[14] W. Kinzel, R. Metzler, I. Kanter, *Dynamics of interacting neural networks*, J. Phys. A **33** L141-L147 (2000);
R. Metzler and W. Kinzel and I. Kanter: *Interacting Neural Networks*, Phys. Rev. E **62** 2, 2555 (2000)

[15] C. M. Bishop: *Neural Networks for Pattern Recognition* (Oxford University Press, New York 1995)

[16] H. Zhu and W. Kinzel: *Anti-Predictable Sequences: Harder to Predict Than A Random Sequence*, Neural Computation **10**, 2219-2230 (1998)

[17] A. Freking and W. Kinzel and I. Kanter: *unpublished*

[18] D. Challet and M. Marsili and R. Zecchina: *Statistical Mechanics of Systems with Heterogeneous Agents: Minority Games*, Phys. Rev. Lett. **84** 8, 1824-1827 (2000)

[19] G. Reents and R. Metzler and W. Kinzel: *A New Stochastic Strategy for the Minority Game*, cond-mat/0007351 (2000)

[20] Econophysics homepage: http://www.unifr.ch/econophysics/

[21] Kinzel,W. and Reents,G.: *Physics by computer*, (Springer-Verlag, Heidelberg 1998)