INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **36** (2003) 4543–4550 PII: S0305-4470(03)53799-4

Time series prediction by feedforward neural networks—is it difficult?

Michal Rosen-Zvi[1], **Ido Kanter**[1] **and Wolfgang Kinzel**[2]

[1] Minerva Center and Department of Physics, Bar-Ilan University, Ramat-Gan, 52900 Israel
[2] Institut für Theoretische Physik, Universität Würzburg, Am Hubland 97074 Würzburg, Germany

Received 13 January 2003, in final form 3 March 2003
Published 8 April 2003
Online at stacks.iop.org/JPhysA/36/4543

Abstract

The difficulties that a neural network faces when trying to learn from a quasi-periodic time series are studied analytically using a teacher–student scenario where the random input is divided into two macroscopic regions with different variances, 1 and $1/\gamma^2$ ($\gamma \gg 1$). The generalization error is found to decrease as $\epsilon_g \propto \exp(-\alpha/\gamma^2)$, where α is the number of examples per input dimension. In contradiction to this very slow vanishing generalization error, the next output prediction is found to be almost free of mistakes. This picture is consistent with learning quasi-periodic time series produced by feedforward neural networks, which is dominated by enhanced components of the Fourier spectrum of the input. Simulation results are in good agreement with the analytical results.

PACS numbers: 05.20.−y, 05.45.Tp, 87.18.Sn

1. Introduction

Forecasting future events based on current given data has fascinated people throughout history. In modern research, different methods taken from a variety of fields are employed for this task [1]. Time series that are produced by neural networks have lately been studied in the framework of the statistical physics field [2–6]. One of the novel findings regarding time series produced by neural networks is concerned with series produced by perceptrons with continuous activation function. It was found that creating a sequence using a continuous activation function could result in a quasi-periodic sequence (QPS), in some range of the parameters. Trying to learn a QPS by a similar network, a perceptron with the same activation function, results in poor learning, i.e. a student trained on QPSs obtains very little information about the teacher. However, a host of simulation results covering most of the parameter space show that the student who learned only partially can predict the same sequence over many steps ahead [5, 6].

In this paper we identify the reasons for good and poor teacher–student learning neural-net time series. We suggest an explanation by presenting an analytical solution of a model that

contains a student attempting to learn the teacher's sequence. Both teacher and student are the simplest archetype of feedforward neural networks, the perceptron, and are given the same input which is composed of two regions with different variances (see figure 2). The features of the student, the teacher and the sequence described in the model are very similar to those in the case of learning a time series produced by the teacher, hence the model illuminates the above mysterious partial learning.

Besides the relevance of our model to time series, it is also relevant to the question of what is the space distribution of input that performs poor learning. Poor learning is found in training real data consisting of input with different variances. We present an analytical survey of learning scenarios with inputs that do not cover the whole space, only a subset; an instance is the finding procedure for the native state of a protein which is mapped onto a perceptron-learning problem where the input represents the contact energy and the hydrophobic energy [7].

In the following we first describe the features that are typical of sequences derived by neural networks. In particular, we describe and formulate the quasi-periodic orbits in section 2. In section 3 we present on-line learning results where the input is the time series. We concentrate on the teacher–student scenario [8] in the perceptron for simplicity, where the same behaviour applies to multilayer networks as well. After describing the findings concerning this learning task, we suggest an explanation for the learning phenomenon by introducing a model that imitates the features indicated above. We present an analytical study of the model in section 4 and conclude with a comparison and discussion about the similarities between the learning process in the model and the process of learning time series in section 5.

2. Quasi-periodic orbits

The general form of the perceptron rule generating a sequence is as follows:

$$S^t_{\rm out} = g\left(\beta \frac{1}{\sqrt{N}} \sum_{j=1}^{N} W_j S^t_j\right) \tag{1}$$

where g is a continuous transfer function, $\mathbf{W}$ is an N-dimensional weight vector and β is the gain parameter. The input vector at time $t+1$ is obtained by

$$S^{t+1}_1 = S^t_{\rm out} \qquad S^{t+1}_j = S^t_{j-1} \qquad j = 2, \ldots, N. \tag{2}$$

A well-known fact in contemporary research in such systems is that the gain parameter has a major impact on the stationary solution. For all continuous functions and small gain parameter the stationary sequence is given by the trivial solution, $S^t_{\rm out} = 0$. The first nontrivial solution for larger β is quasi-periodic. In the case of a non-monotonic function, there is another transition to a robust chaos phase at large β [4]. In the discussion below we concentrate on the quasi-periodic solutions.

A quasi-periodic solution is the solution of equation (1) above some critical value. A quasi-periodic solution means that the solution at each step changes and one cannot find one point in the return map—the space that is defined by S^t_i versus S^t_{i-1}—that satisfies the update equations. However, there is a certain line in the attractor dimension that the solutions are confined to (see [4–6] for more details). Such a line indicates problems in learning. In particular, for large dimensions N, batch learning does not work well for QPS generated by a teacher perceptron [5].

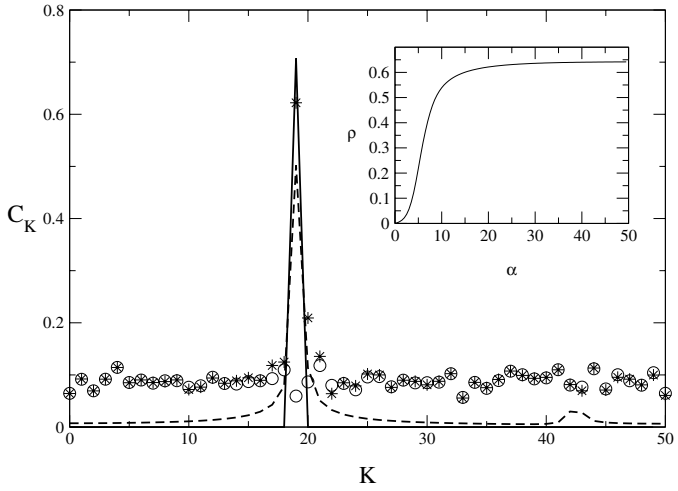

Figure 1. Fourier spectrum of a perceptron weight vector C_K, $N = 100$ (solid line); the averaged Fourier spectrum of the sequence produced with $\beta = 0.19$ and 20 different initial conditions (dashed line); and the averaged Fourier spectrum of the weight vector of a student that tries to learn, at the beginning, $\alpha = 0$ (circles) and at $\alpha = 50$ (stars). Inset: simulation results of averaged overlap between the teacher and the student above, as a function of α.

3. On-line learning

Traditional on-line learning includes an input set that changes with time. Usually the input set is generated at random. In this section we study the case where the time series is taken as an input set and in section 4 we take a specific correlated set as the input set. In each time step the teacher generates an output which is given to the student in addition to the input. The parameter α counts the time steps or equivalently the number of examples given and it is rescaled by the input dimension, $\alpha = t/N$. In the following, to simplify the analytical presentation we use 'sine' as a representative transfer function, and β is chosen to be in the intermediate region, where the solution is quasi-periodic.

We exemplify the phenomenon of partial learning by plotting results of a specific simulation with $\beta = 0.19$ and $N = 100$ (see figure 1). The sequences are initiated at random and the teacher is taken in the special case of enhanced dominant components in the Fourier spectrum where the special learning phenomenon is easy to present. After the transient to the typical attractor occurs we start updating the student, according to the gradient descent method [8]. This procedure is repeated 20 times with different initial input and student weight vector. In figure 1 the Fourier spectrum of the teacher weight vector is plotted (solid line). The average of the Fourier spectrum derived from the stationary part of the sequences (dashed line) and the student weight vectors, at $\alpha = 0$ (circles) and $\alpha = 50$ (stars), are given.

The possibility of learning the weights of the perceptron that produce QPSs as well as the ability to predict its next outputs, the future sequence, are the main issues addressed in this paper. The ability to learn the weights is measured by the correlation between the two vectors, $\rho = \mathbf{W}^S \cdot \mathbf{W}^T / (\|\mathbf{W}^S\| \, \|\mathbf{W}^T\|)$, where $\mathbf{W}^T$ ($\mathbf{W}^S$) is the teacher (student) weight vector and $\rho = 1$ indicates perfect learning. The development of the overlap between student and teacher in the above-mentioned specific example is given in the inset of figure 1 (solid line). As can be seen in the plot, the learning time series is characterized by two different regions: at the beginning, in the small α regime, the rate of learning is fast, whereas in the second

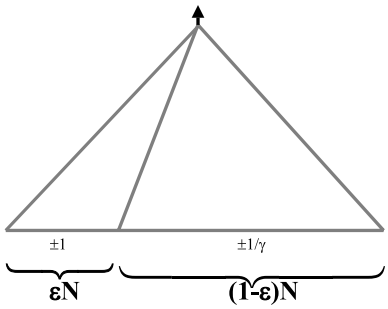

Figure 2. A sketch of a perceptron that receives input which is divided into two parts, εN and $(1-\varepsilon)N$, where the second moments of the random input are 1 and $1/\gamma^2$, respectively.

regime, for large α, the increment of the overlap has slowed down. These two regions are directly connected to the two different regions in the Fourier spectrum, the enhanced range, where the input spectrum is enlarged in the region where the teacher components are learned, as opposed to the other region, where the input spectrum is not enlarged and the student has not succeeded in learning the teacher components.

Recent results [5] based on simulations show that despite the difficulties in learning the weight vector of the teacher, the student can infer the *successive output* fairly well. A quantitative measure of the prediction error one step ahead is given by

$$\epsilon_p = \left\langle \left| S_{\text{out}}^{tT} - S_{\text{out}}^{tS} \right| \right\rangle_t \tag{3}$$

where the indices T/S stand for the teacher/student output, respectively, and the average, $\langle\rangle_t$, is taken over the sequence in the stationary part at different times.

We claim that this result pertains to the Fourier spectrum of the perceptrons and the input. Since the Fourier transform of the generated sequence is dominated by certain values of K, which are the largest in the Fourier spectrum of the teacher weight vector, the student is limited to learning only those components. Since the largest components of the student weight vector in the Fourier spectrum, after the learning process is carried out, are those typical of the specific sequence, the student and the teacher will produce very similar sequences. Having input which is largely enhanced in a fraction of the components in the Fourier spectrum (compared to the other components) is the reason for having a learning curve comprising two regions.

Therefore, we present a model that contains such enhancement (see figure 2). Learning in the Fourier spectrum—learning the Fourier components by using the Fourier transformation of the input set—is equivalent to learning in the regular input space. Hence, having a model with enhancement in the regular spectrum is an equivalent representation of the situation described above. This simplification enables analytical study of the unusual learning phenomenon.

4. Toy model

This model contains an random input vector, $\boldsymbol{\xi}$, that has two different regions. The first moments of both regions equal zero whereas the second moments are 1 and $\frac{1}{\gamma^2}$, respectively,

$$\xi_i = \begin{cases} \pm 1 & 1 \leqslant i \leqslant \varepsilon N \\ \pm \frac{1}{\gamma} & \text{otherwise.} \end{cases} \tag{4}$$

The output S is calculated according to the rule

$$S = \sin\left(\frac{\mathbf{W}^T \cdot \boldsymbol{\xi}}{\sqrt{N}} \right). \tag{5}$$

At each time step μ the student receives the pattern $\{\boldsymbol{\xi}_\mu, S_\mu\}$, and tries to learn the teacher's quantities according to the gradient descent method [8]. The learning rule in the case of 'sine' activation function is (see [9])

$$\mathbf{W}^S(\mu) = \mathbf{W}^S(\mu-1) - \frac{\eta}{\sqrt{N}}[\sin y(\mu) - \sin x(\mu)]\cos x(\mu)\boldsymbol{\xi}(\mu). \tag{6}$$

Here $x(\mu)/y(\mu)$ are the student's/teacher's local fields in the μth step (where $x = x_1 + x_2$ and $y = y_1 + y_2$),

$$x_i = \frac{1}{\sqrt{N}}\sum_{i\in V_i} W_i^S\xi_i \qquad y_i = \frac{1}{\sqrt{N}}\sum_{i\in V_i} W_i^T\xi_i \tag{7}$$

and $i = 1, 2$ where $V_1 = 1, \ldots, \varepsilon N$, and $V_2 = \varepsilon N + 1, \ldots, N$.

The macroscopic order parameters are the overlaps between teacher and student and the student's norm in each region

$$R_i = \frac{1}{\|V_i\|}\sum_{i\in V_i} W_i^S W_i^T \qquad Q_i = \frac{1}{\|V_i\|}\sum_{i\in V_i} W_i^S W_i^S \tag{8}$$

where $\|V_1\| = \varepsilon N$ and $\|V_2\| = (1-\varepsilon)N$. For simplicity we assume that the teacher weight vector is normalized, $\|\mathbf{W}^T\| = N$.

Since each region of the input is taken to be extensive in N, it is possible to derive analytical equations that describe the development of the above-mentioned order parameters [8]. The equations over the order parameters are derived by multiplying equation (6) by the teacher weight vector and the student weight vector, respectively, and taking the summation in each region separately. The result is coupled update equations of the form $R_i(\mu+1) = R_i(\mu) + F_i^R(x, y)/N$ and $Q_i(\mu+1) = Q_i(\mu) + F_i^Q(x, y)/N$, where F is some function over the local fields and $i = 1, 2$ (see [8] for more details about the derivation of these standard equations). The next step is to take the average over the local fields. This toy model does not have the regular joint probability distribution. In this case the joint probability distribution is composed of two independent distributions

$$P(x, y) = P(x_1, y_1|R_1, Q_1, \varepsilon)P(x_2, y_2|R_2, Q_2, \varepsilon, \gamma). \tag{9}$$

Each one of the joint probability distributions is calculated according to its correlation matrix

$$\bar{\bar{C}}_1 = \begin{pmatrix} \varepsilon & \varepsilon R_1 \\ \varepsilon R_1 & \varepsilon Q_1 \end{pmatrix} \qquad \bar{\bar{C}}_2 = \begin{pmatrix} \frac{1-\varepsilon}{\gamma^2} & \frac{1-\varepsilon}{\gamma^2}R_2 \\ \frac{1-\varepsilon}{\gamma^2}R_2 & \frac{1-\varepsilon}{\gamma^2}Q_2 \end{pmatrix}. \tag{10}$$

We derived the equations of motion over the four order parameters:

$$\begin{aligned}
\frac{\mathrm{d}R_1}{\mathrm{d}\alpha} &= \frac{\eta}{2}[(R_1+1)A_+ - (R_1-1)A_- - 2R_1C] \\
\frac{\mathrm{d}Q_1}{\mathrm{d}\alpha} &= \eta[(R_1+Q_1)A_+ - (Q_1-R_1)A_- - 2Q_1C] + \frac{\hat{\eta}}{8} \\
\frac{\mathrm{d}R_2}{\mathrm{d}\alpha} &= \frac{\eta}{2\gamma^2}[(R_2+1)A_+ - (R_2-1)A_- - 2R_2C] \\
\frac{\mathrm{d}Q_2}{\mathrm{d}\alpha} &= \frac{\eta}{\gamma^2}[(R_2+Q_2)A_+ - (Q_2-R_2)A_- - 2Q_2C] + \frac{\hat{\eta}}{8\gamma^2}
\end{aligned} \tag{11}$$

where

$$\begin{aligned}
A_\pm &= \mathrm{e}^{-\frac{1}{2}[\varepsilon(Q_1\pm 2R_1+1)+\frac{1-\varepsilon}{\gamma^2}(Q_2\pm 2R_2+1)]} \\
B_\pm &= \mathrm{e}^{-\frac{1}{2}[\varepsilon(1+9Q_1\pm 6R_1)-\frac{1-\varepsilon}{\gamma^2}(1+9Q_2\pm 6R_2)]} \\
C &= \mathrm{e}^{-2[\varepsilon Q_1+\frac{(1-\varepsilon)}{\gamma^2}Q_2]} \qquad D = \mathrm{e}^{-2[\varepsilon+\frac{(1-\varepsilon)}{\gamma^2}]} \\
\hat{\eta} &= \eta^2\big(3 - 2A_- + 2A_+ - A_-^4 - A_+^4 - 2B_- + 2B_+ + 2C - C^4 - 2D\big).
\end{aligned} \tag{12}$$

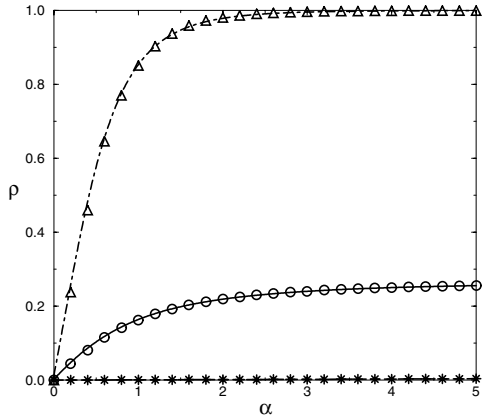

Figure 3. Analytical results of the overlaps ρ_1 (dot-dashed line), ρ_2 (dashed line) and ρ (solid line), as a function of α for $\gamma = 40$ and $\varepsilon = 0.2$. Simulation results (symbols) are averaged over ten different runs with $N = 1000$. Error bars are smaller than symbols.

The transformation to two order parameters that represent the overall overlaps is straightforward.

The normalized overlaps, $\rho_i = R_i/\sqrt{Q_i}$, and the overall overlap, $\rho = \epsilon\rho_1 + (1-\epsilon)\rho_2$, as a function of α for the case $\gamma = 40$, $\varepsilon = 0.2$ and $\eta = 1$ derived from equations (11), are plotted in figure 3. The initial conditions are $Q_1 = Q_2 = 0.5, R_1 = R_2 = 0$. The simulation results (symbols) are averaged over ten different runs with $N = 1000$. One can see good agreement with the analytical results derived from the equations above. Indeed the performance of the second regime, ρ_2, is very slow and therefore the overall overlap, ρ, evolves slowly too.

The generalization error, ϵ_g, which is calculated by averaging over *random inputs* is exactly the same as for the simple perceptron, [9]. The prediction error of one step ahead, equation (3), is equivalent to calculating the error the student will produce having the *special input*, equation (4). Employing equation (9) for the average over the inputs yields

$$\epsilon_p = \tfrac{1}{2}\big[1 - A_- + A_+ - \tfrac{1}{2}(C+D)\big]. \tag{13}$$

It is simple to verify that within the limit of large γ, when $R_1 \to 1$ and $Q_1 \to 1$, equation (13) depends merely on the asymptotic of the first region quantities, $\epsilon_p \sim \varepsilon[(1-R_1)(1+\mathrm{e}^{-2\varepsilon}) - (1-Q_1)(1+3\,\mathrm{e}^{-2\varepsilon})/2]/2$. Note that this model is characterized by the same discrepancy as in the case of time series prediction; the prediction error does vanish even though the generalization error is not small.

In figure 4 the generalization error (solid line) [9] and the prediction error (dashed line) equation (13), are plotted in the case of $\gamma = 40$ and $\varepsilon = 0.2$. One can see that although the student learned poorly (for a large number of examples $\rho \cong 0.2$, $\epsilon_g \cong 0.3$), the prediction error is almost zero since the learning in the enhanced region is almost perfect. Simulation results (symbols) for $N = 1000$ and averaged over ten samples are in good agreement with the analytical curves (error bars are smaller than the symbols).

The generalization error decays to zero exponentially. The rate of the decay is given by linearizing equations (11) and is governed by the smallest absolute eigenvalue, $\ln \epsilon_g \propto -\min|\lambda|\alpha$, [8]. The smaller the absolute eigenvalues the slower the decay to perfect generalization. The eigenvalues that determined the rate of convergence in the above specific example are $\{-1.5227, -0.5670, -0.0010, -0.0005\}$. One can see that one of the eigenvalues

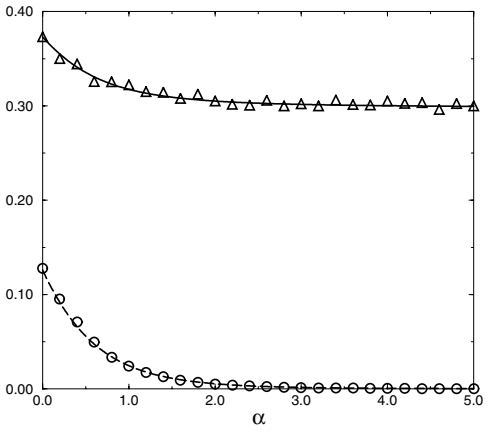

Figure 4. Analytical results of the prediction error, ϵ_p (dashed line), as obtained from equation (13) and the generalization error, ϵ_g (solid line) (taken from [9]), as a function of α for $\gamma = 40$ and $\varepsilon = 0.2$. Simulation results of ϵ_p (circles) and ϵ_g (triangles) are averaged over ten different runs with $N = 1000$. Error bars are smaller than the symbols.

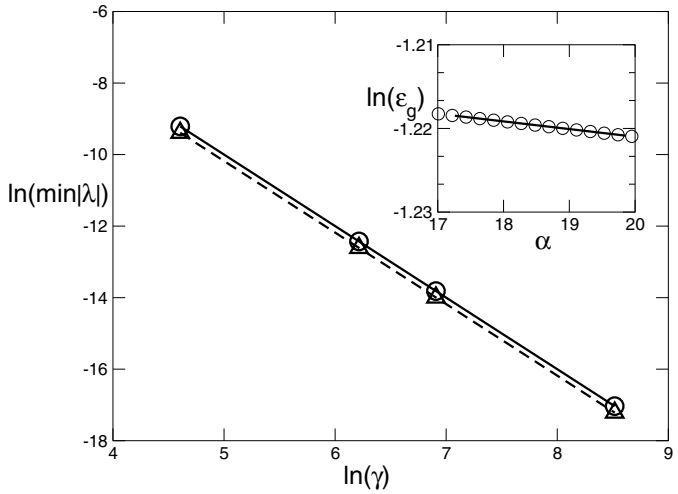

Figure 5. Analytical results for λ for several values of γ having $\eta = 1$ and $\varepsilon = 0.2/0.0001$ (triangles/circles). Lines are linear curves fitted to the logarithm of min $|\lambda|$ as a function of $\ln(\gamma)$; their slope is $\sim$2. Inset: analytical results (circles) of $\ln \epsilon_g$ as a function of α for $\gamma = 40$, $\varepsilon = 0.2$ and $\eta = 1$. The slope of the linear curve fitted is $\sim$0.0009.

is relatively small, i.e., it will take a long time—many examples—to get perfect generalization. A semi-log plot of the generalization error as a function of α in the specific example above is plotted in the inset of figure 5. A linear curve fitted to the analytical result has a slope of $\cong 0.0009$ which is of the order of the smallest absolute eigenvalue, as expected. The values of the smallest absolute eigenvalue for $\varepsilon = 0.2$ and $\varepsilon = 0.001$ and various γ values, as found from the asymptotic expansion of equations (11) with $\eta = 1$, are given by the log–log plot in figure 5. It is apparent that as γ is increased it becomes practically impossible to achieve perfect learning. Fitted linear curves for $\varepsilon = 0.2$ (dashed line) and in the case of $\varepsilon = 0.001$

(solid line) indicate that the slope is approximately 2. Their smallest absolute eigenvalue obeys a power law, $\min|\lambda| \propto \gamma^{-2}$.

5. Discussion

In summary, we found that in a manner similar to the learning procedure of time series, when a student tries to learn from a teacher but is restricted to a specific input spectrum in which at a certain region the inputs are enhanced, the learning does not end in perfect learning within a reasonable time. However, that same enhanced region can be learned perfectly. The parallel drawn between these two cases is clear when comparing the inset of figure 1 and figure 3. One can see in both cases poor performance of the overlap between teacher and student where in both cases the student seems to be stuck. However, the student can learn a certain region as indicated by ρ_1, the dashed line in figure 3, and by the stars around $k = 20$ that show the performance of the student in the Fourier regime (figure 1).

In the case of time series QPSs, the two regions are of the order of 1 and $1/N$. The adaptation of this result to our toy model results in $\gamma \sim N$. Therefore, by extrapolating the analytical model results to infinitely large γ of the order of N, we can estimate that the exponential decay of the generalization error scales with $\alpha \sim N^2$ and the number of required examples to learn the sequence is $O(N^3)$. This estimation might serve as an explanation for the observation in [5], regarding batch learning. In the case of a monotonic activation function such as 'tanh', one may think of the alternative approach of inverting the matrix as a suitable way of finding the teacher's weight vector. Note that the complexity of inverting the matrix also scales with $O(N^3)$ as was already mentioned in [5]. It turns out that even professional computer routines often fail to perform the required matrix inversion: the patterns are almost linearly dependent. The toy model presented in this paper might explain this partial dependence between inputs.

Acknowledgments

We thank Ansgar Freking for fruitful discussions and critical reading of the manuscript. We thank the anonymous reviewers for helpful comments and suggestions. IK acknowledges partial support by the Israel Academy of Science. This paper is part of the PhD Thesis of MR.

References

[1] Weigend A and Gershenfeld N A 1994 *Time Series Prediction: Forecasting the Future and Understanding the Past (Santa Fe)* (Reading, MA: Addison-Wesley)
[2] Kanter I, Kessler D A, Priel A and Eisenstein E 1995 *Phys. Rev. Lett.* **75** 2614
[3] Shröder M and Kinzel W 1998 *J. Phys. A: Math. Gen.* **31** 9131
[4] Kinzel W 2001 *Computational Statistical Physics, From Billiards to Monte Carlo* ed K H Hoffmann and M Schreiber (Berlin: Springer)
[5] Freking A, Kinzel W and Kanter I 2002 *Phys. Rev.* E **65** 050903(R)
[6] Priel A and Kanter I 1999 *Phys. Rev.* E **59** 3668
[7] Vendruscolo M and Domany E 2000 *Vitam. Horm.* **58** 171
[8] Engel A and Van den Broeck C 2001 *Statistical Mechanics of Learning* (Cambridge: Cambridge University Press)
[9] Rosen-Zvi M, Biehl M and Kanter I 1998 *Phys. Rev.* E **58** 3606