Learning of time series through neuron-to-neuron instruction

2003 J. Phys. A: Math. Gen. 36 1315

(http://iopscience.iop.org/0305-4470/36/5/309)

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

J. Phys. A: Math. Gen. **36** (2003) 1315–1322 PII: S0305-4470(03)55674-8

Learning of time series through neuron-to-neuron instruction

Y Miyazaki[1]**, W Kinzel**[2] **and S Shinomoto**[1]

[1] Department of Physics, Kyoto University, Kyoto 606-8502, Japan
[2] Institut für Theoretische Physik, Universität Wurzburg, 97074 Wurzburg, Germany

E-mail: yoichi@scphys.kyoto-u.ac.jp and shinomoto@scphys.kyoto-u.ac.jp

Received 29 October 2002, in final form 10 December 2002
Published 22 January 2003
Online at stacks.iop.org/JPhysA/36/1315

Abstract
A model neuron with delayline feedback connections can learn a time series generated by another model neuron. It has been known that some student neurons that have completed such learning under the instruction of a teacher's quasi-periodic sequence mimic the teacher's time series over a long interval, even after instruction has ceased. We found that in addition to such faithful students, there are unfaithful students whose time series eventually diverge exponentially from that of the teacher. In order to understand the circumstances that allow for such a variety of students, the orbit dimension was estimated numerically. The quasi-periodic orbits in question were found to be confined in spaces with dimensions significantly smaller than that of the full phase space.

PACS numbers: 05.45.−a, 07.05.Mh, 84.35.+i

1. Introduction

A single neuron or a network of neurons with delayline feedback connections generates time series autonomously. Recent studies have revealed various interesting characteristics of sequence generators of this kind [1, 2]. A model neuron with a monotonic transfer function can generate a sequence that is stationary (fixed), periodic or quasi-periodic. In addition, there are tiny parameter regions in which low-dimensional chaotic sequences are generated (so-called fragile chaos) [3]. In contrast, a model neuron with a nonmonotonic transfer function can generate a high-dimensional chaotic sequence (robust chaos) [4], in addition to the fixed, periodic or quasi-periodic sequences.

A model neuron (student) can also learn a time series generated by another model neuron (teacher). Recently Freking *et al* [5, 6] performed a numerical study and reached the following conclusions: (A) a student neuron trained with a quasi-periodic sequence does not obtain much information about the teacher neuron, but it can mimic the teacher's time series over a long interval; (B) a student neuron trained with a high-dimensional chaotic sequence obtains almost complete knowledge about the teacher, but it cannot mimic the teacher's time series over a long interval.

We performed a detailed numerical study for the neuron-to-neuron instruction. We have two main results. First, we obtain an understanding of the mechanism with which we can explain (A) and (B). In the case that the generated sequence is quasi-periodic, the neuron's internal states are confined to a lower dimensional sub-space of the full phase space, and this is the reason why the student does not acquire much information about the teacher. In the case that the generated sequence is chaotic, the neuron's internal states occupy the full dimension of the phase space, and therefore, in the course of the instruction, the student acquires information about all aspects of the teacher.

Our second main result is that we have found phenomena that are inconsistent with the above conjecture (A). With this finding, we propose the following revised version:

(A′) A student neuron trained with a quasi-periodic sequence does not obtain much information about the teacher. Students can be classified into two types: (1) faithful students that are able to mimic the teacher's time series over a long interval even after instruction has ceased, and (2) unfaithful students whose time series diverge exponentially from that of the teacher.

2. The model neuron

We consider sequence generators that are represented by recurrence equations of the form

$$s_t = f\left(\sum_{i=1}^{N} w_i s_{t-i}\right) \tag{1}$$

where $f(x)$ represents the neuron's nonlinear transfer function, and $\mathbf{w} = (w_1, w_2, \ldots, w_N)$ is a weight vector characterizing the weights of delayline connections from its output. We consider a single value of the order of recurrence (or the dimension) $N = 20$ throughout the present paper, but the features of the results do not depend on the choice of the dimension.

We first considered the case of a monotonic transfer function, $f(x) = \tanh(\beta x)$, in which β represents the input gain. We chose the value of each weight element w_i randomly from a normal distribution with mean 0 and variance $1/N$. Therefore, the norms of the weight vectors will be distributed around 1. We chose $\beta = 1$, because with a smaller value there is a tendency towards the trivial fixed point $s_t = 0$, and with a larger value there is a tendency for the sequence to become quasi-discrete $s_t \sim \pm 1$. We chose 25 weight vectors of dimension $N = 20$ to observe the characteristics of the sequence generators. We found that in one case the sequence converged to the trivial fixed point $s_t = 0$, in two cases it converged to nonzero fixed points, in two cases it exhibited period-two oscillation and in 20 cases it exhibited a quasi-periodic sequence. Figure 1 displays a sample quasi-periodic sequence. We checked by applying small disturbances that all the asymptotic orbits are locally stable. As discussed in [3, 4], the range of parameter values for which there is chaotic behaviour for this monotonic transfer function is very small. In fact, chaotic sequences were not encountered in any of the many simulations we carried out. In addition, we observed that many of those sequence generators have multiple stable attractors. In any case, for each trial, we used only one sequence generated by the teacher as the training sequence.

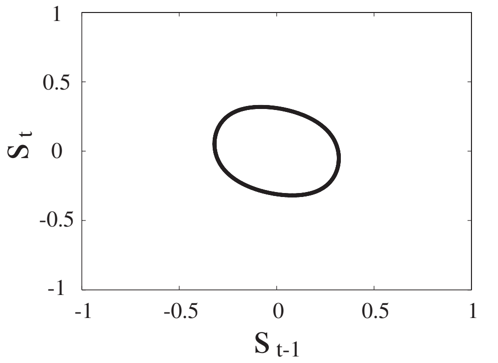

Figure 1. The map of s_t versus s_{t-1} for the asymptotic quasi-periodic sequence generated by a teacher. Sequences generated by faithful students are indistinguishable from this.

3. Learning quasi-periodic sequences

We consider only the situation where the teacher and student have the same transfer function. The learning is carried out with the online gradient descent method, in which the student repeatedly adjusts its weight vector so as to cause its output to be as close as possible to the teacher's output [7]. During the learning stage, s_t is generated for both the teacher and the students by inputting the vector representing the most recent N values of the teacher's time sequence, in f, $(s_{t-1}, s_{t-2}, \ldots, s_{t-N})$. We monitored the average deviation of the student's output o_t from the teacher's output s_t which is obtained for the common inputs,

$$\epsilon = \frac{1}{T} \sum_{t=t_0+1}^{t_0+T} |s_t - o_t| \tag{2}$$

for every time interval of length $T = 20$. The learning was stopped if the condition $\epsilon < 10^{-3}$ was satisfied for 50 000–100 000 consecutive time steps. The length of the sequence necessary to realize this level of learning varies from case to case, but it was always more than one million, and typically about five million. After the learning, we determined how closely the student's final weight vector $\mathbf{w}^S$ approximated the teacher's weight vector $\mathbf{w}^T$.

In each trial, one teacher taught multiple students. The initial weight vector of each student was chosen randomly according to the same prescription as used for the teacher. All the students eventually succeeded in learning to the level specified above. After the criterion for successful learning was satisfied, the student was allowed to proceed independently, by now using its own output for s_{t-i} in f of equation (1). We then observed how the deviation between the teacher and the student evolved.

It was found that the sequence produced thereafter by many students deviated from the teacher's sequence linearly in time and remained close to it for a long period. Despite this slow divergence of the student's time series from the teacher's time series, it is found that there are in fact significant differences between the final student weight vectors and the teacher weight vector. These results are consistent with conjecture (A).

4. Numerical estimate of orbit dimension

If the teacher's asymptotic internal states are confined to a space of dimension M, smaller than the dimension N of the full phase space, then the components of the weight vector orthogonal to this space are irrelevant, and are free from the *learning pressure* that acts on the students' weights. In order to confirm whether the teacher's asymptotic quasi-periodic orbit is indeed confined to such a smaller space, we attempted to determine the spatial dimension of the internal state distribution of the teacher neuron according to the following prescription.

(i) Generate an N-dimensional vector from the sequences as $\mathbf{s}_t = (s_t, s_{t-1}, \ldots, s_{t-N+1})$. Construct an $N \times N$ matrix $\mathbf{S}$ from N consecutive vectors,

$$\mathbf{S} = (\mathbf{s}_t, \mathbf{s}_{t-1}, \ldots, \mathbf{s}_{t-N+1})^t. \tag{3}$$

If these N vectors are linearly dependent, the determinant of $\mathbf{S}$ vanishes. Due to numerical error, it is not feasible to realize a value of exactly zero in a calculation, but a very small absolute value of this determinant can be interpreted as indicating linear dependence.

(ii) For a given value of $M < N$, consider all of the $M \times M$ matrices that can be obtained by removing $N - M$ rows and $N - M$ columns from $\mathbf{S}$. If the N vectors in question span M-dimensional space, there is at least one such $M \times M$ matrix that has a nonzero determinant. This procedure is carried out for $M = N - 1, N - 2, \ldots$, until an $M \times M$ matrix with nonzero determinant is found.

(iii) Rather than carrying out such calculations for all $M \times M$ matrices of the kind described above (which would be much too numerous for the values of $N = 20$ used in actual simulations), for a given value of M, we randomly chose 10 000 matrices. We then judged whether the N vectors span an M-dimensional space by determining if any of those determinants was significantly large to be considered nonzero.

The significance level for judging a determinant to be vanishing or nonvanishing was empirically determined to be 10^{-5} by applying the above procedure to prepared samples whose dimensions were already known. In studying these prepared samples, it was also found that the procedure is effective in determining the actual dimension.

We applied this method to the quasi-periodic sequences that were generated by 20 teachers. The dimensions M of those 20 quasi-periodic sequences were found to range from 4 to 13, significantly smaller than the dimension of the full phase space, $N = 20$. The most common dimension was $M = 6$ (found in six cases). Each training sequence was used to teach 20 students. We also determined the dimension L of the space spanned by those weight vectors, using the same method. Their dimensions were found to range from 5 to 14. In 9 of the 20 cases, the dimension of the training sequence and the dimension of the student weight vectors were approximately complementary, satisfying $M + L = N \pm 2$. In other cases, however, $M + L$ was significantly smaller than N. In the smallest case, $M + L = 13$.

5. Unfaithful students

All the students eventually passed the learning criterion (i.e. realized $\epsilon < 10^{-3}$), which means that they were able to make short-range predictions with sufficient precision. When the students are freed from the teacher's supervision, many of them remain close to the teacher's sequence, and the deviation develops linearly in time, as mentioned above. However, we found that there are students that deviate from the teacher's sequence and asymptotically generate sequences that bear no resemblance to the teacher's sequence. Figure 2 depicts the sequence of such an unfaithful student resulting from the training sequence displayed in figure 1. We found that the deviation from the teacher's sequence grows exponentially in this case (see figure 3). We refer to these two groups of students as 'faithful students' and 'unfaithful students'.

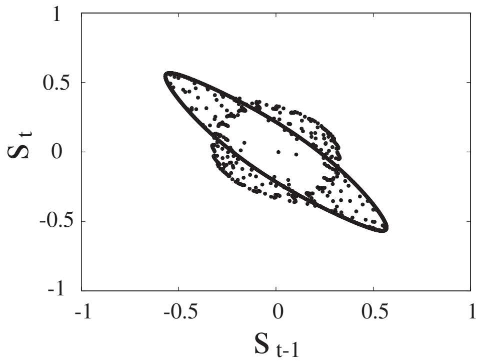

Figure 2. The map of s_t versus s_{t-1} for the sequence generated by an unfaithful student after the learning has ended. After transients, the system enters an orbit that bears no resemblance to the teacher's orbit, shown in figure 1.

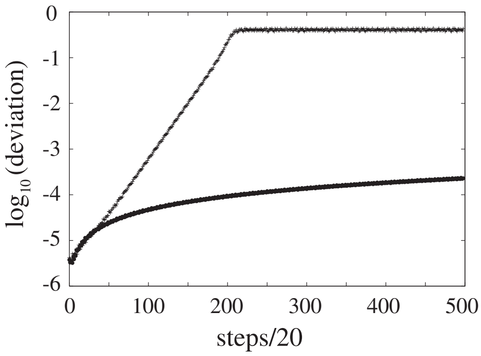

Figure 3. Evolution of the deviation between the student and teacher, after the students are freed from the teacher's supervision. The deviation for the faithful student, depicted in figure 1, exhibits a linear increase and that for the unfaithful student, depicted in figure 2, exhibits an exponential increase. In those cases, the learning was performed until $\epsilon < 10^{-5}$.

The students' initial weight vectors were generated according to the same prescription as the teacher's weight vectors: each element w_i was chosen randomly from a normal distribution with mean 0 and variance $1/N$. We found that among 20 teachers, nine had at least one unfaithful student out of 20 students. The number of unfaithful students ranged from 0 to 9. The average fraction of unfaithful students was $30/400$.

The maps that have sufficiently large weight vectors would have their own stable and unstable orbits. In the learning process they modify their weight vectors according to the instruction. In some cases they readjust their stable orbits to the teacher's orbit, but in other cases they may readjust their unstable orbits. We interpret unfaithful students as representing the situation in which the student was able to adjust an unstable orbit to the teacher's stable orbit. The student initially keeps close to the teacher's sequence, but the deviation from the teacher's sequence develops exponentially in time. Whether or not such a phenomenon occurs depends on the nature of the training sequence and the initial student weight vector. We have also investigated the case of a student with initial weight vector $\mathbf{w} = (0, 0, \ldots, 0)$, with which a neuron generates only the trivial stationary fixed sequence $s_t = 0$. We found that all 20 teachers succeeded in making this student faithful. In these cases the students, which have

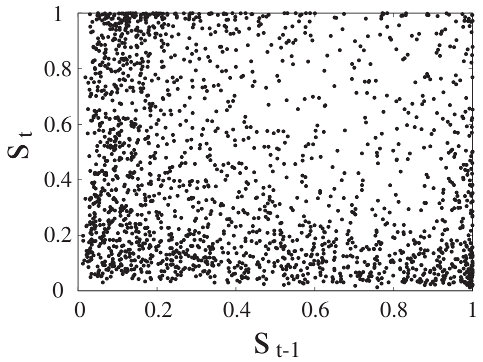

Figure 4. The map of s_t versus s_{t-1} for the asymptotic chaotic sequence generated by the transfer function $1/\cosh(\beta x)$, with $\beta = 4$.

each only one stable fixed point, start to generate stable quasi-periodic orbits according to the instruction and succeed in attaining the desired stable orbit.

It is also worthwhile to see how the learning is perturbed by the noise entered in the transfer process of the teacher's signals. We confirmed that the students could learn the teacher's orbit within an accuracy bounded by the noise level and become either faithful or unfaithful. By increasing the noise level, some of the students that are originally destined to be unfaithful in the noiseless limit, turn out to be faithful, while the opposite case was not observed. It should be noted, however, that the students may remain unfaithful even in the presence of (low-level) noise.

A neuron with a weight vector given by a linear combination of two students' final weight vectors,

$$\mathbf{w} = c\mathbf{w}_1 + (1 - c)\mathbf{w}_2 \tag{4}$$

should satisfy the learning criterion, because both students have satisfied the learning criterion, which means that both $\mathbf{w}_1$ and $\mathbf{w}_2$ must be almost identical to the teacher's weight vector in the sub-space of dimension $M(<N)$ in which the internal states are confined. We tested the ability to remain close to the teacher's time sequence of a hybrid neuron with the weight vector given by equation (4). Interestingly, we found that hybrid neurons generated from two unfaithful students were unfaithful in some cases but not all. Conversely, hybrid neurons generated from two faithful students were faithful in some cases but not all.

6. Learning chaotic and quasi-periodic sequences with nonmonotonic transfer function

We also examined the case of a nonmonotonic transfer function, $f(x) = 1/\cosh(\beta x)$. For small β, the recurrence equation (1) generates a stationary (fixed) sequence. As β increases, the sequence becomes periodic or quasi-periodic, and above a certain critical value of β, it becomes chaotic (see figure 4). Considering a chaotic sequence, we confirmed conjecture (B): after learning is completed, the student's and teacher's weight vectors were almost identical $\mathbf{w}^S \approx \mathbf{w}^T$ or opposite, $\mathbf{w}^S \approx -\mathbf{w}^T$, which happens due to the mirror symmetry of the function $1/\cosh(\beta x)$. However, when the teacher's supervision ceases, the student's sequence rapidly deviates from the teacher's.

Again in this case we determined the dimension of the training sequence using the method described above. We used a number of orbits and found in each case that $M \approx 20$ the full dimension. This means that during the learning process, all components of the student weight

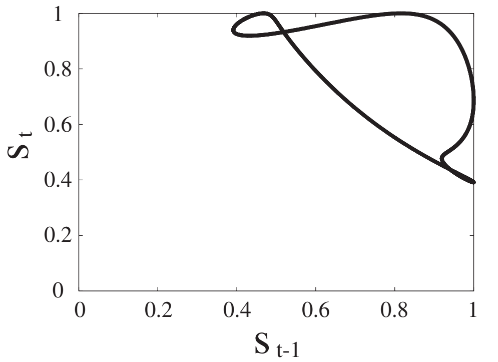

Figure 5. The map of s_t versus s_{t-1} of the asymptotic quasi-periodic sequence generated by the transfer function $1/\cosh(\beta x)$, with $\beta = 1.5$.

vector are subject to the learning pressure, and only the overall sign remains arbitrary. The student time series eventually becomes quite close to the teacher time series. However, the resulting orbit is unstable. For this reason, after the learning process has been terminated, because there will inevitably be a finite difference between the teacher's and student's weight vectors there will also be a finite difference between their time series, and the difference will grow exponentially in time, reflecting the chaotic nature of the orbits.

The transfer function $f(x) = 1/\cosh(\beta x)$ can also generate quasi-periodic sequences if a moderate value of β is used. Figure 5 depicts a quasi-periodic sequence that bears no symmetry. We also carried out a numerical study of this case to determine if the behaviour described by (A$'$) is observed for this kind of nonmonotonic transfer function. We found that the conclusion is unchanged. In particular, a teacher can produce both faithful and unfaithful students. One difference between the nonmonotonic case and the monotonic case studied above is that in the present case the final orbits of unfaithful students can be either stable quasi-periodic sequences or unstable chaotic sequences.

7. Discussion

In the present paper we have studied numerically the learning of time series through neuron-to-neuron instruction. First, we have confirmed the knowledge reported by Freking *et al* [5, 6] and performed the numerical analysis to determine the dimension of the quasi-periodic orbit in question. The orbit dimension was found to be significantly smaller than that of the full phase space. This is consistent with the fact that the learning of a quasi-periodic sequence does not fully determine the weight vector.

Second, we found that in addition to the faithful students, which mimic the teacher's time series over a long interval, there are unfaithful students whose time series diverge exponentially from that of the teacher. Both faithful and unfaithful students can reproduce the teacher's time series during the learning period. When the students are freed from the teacher's supervision, the faithful students stay in orbits that are almost identical to the teacher's orbit, while the unfaithful students eventually enter into individual orbits that are completely different from the teacher's orbit. This evidence is also consistent with the above-mentioned fact that the orbit is confined to a lower dimensional sub-space of the full phase space: the learned orbit could have been made unstable to the orthogonal direction to the space of the learned orbit.

In addition, we have tested to see whether or not unfaithful students are fragile to the noises. The unfaithful students may turn faithful in the presence of high-level noise, but they stay unfaithful in the presence of low-level noise. We also have observed unfaithful students in the learning with a nonmonotonic transfer function. These results imply that the emergence of unfaithful students could be seen in the more general context of time-series learning.

Acknowledgments

We thank Yasuhiro Tsubo and Keiji Miura for stimulating discussions, and Haim Sompolinsky for advising to add noise. This study is supported by Grants-in-Aid for Scientific Research (no 12680382) to SS by the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

[1] Eisenstein E, Kanter I, Kessler D A and Kinzel W 1995 *Phys. Rev. Lett.* **74** 6
[2] Kanter I, Kessler D A, Priel A and Eisenstein E 1995 *Phys. Rev. Lett.* **75** 2614
[3] Priel A and Kanter I 1999 *Phys. Rev.* E **59** 3368
[4] Priel A and Kanter I 2000 *Europhys. Lett.* **51** 230
[5] Freking A, Kinzel W and Kanter I 2002 *Phys. Rev.* E **65** 050903(R)
[6] Kinzel W 2001 *Computational Statistical Physics, From Billiards to Monte Carlo* ed K H Hoffmann and M Schreiber (Berlin: Springer)
[7] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Redwood City, CA: Addison-Wesley)