A sequential pseudo-inverse learning rule for networks of formal neurons



1992 J. Phys. A: Math. Gen. 25 4139

(http://iopscience.iop.org/0305-4470/25/15/021)

J. Phys. A: Math. Gen. 25 (1992) 4139–4146. Printed in the UK

# A sequential pseudo-inverse learning rule for networks of formal neurons

A D Linkevich

TDS and Novopolotsk Polytechnical Institute, Parkovaya 12-80, Novopolotsk, 211440, Byelorussia

Received 27 November 1991

**Abstract.** For networks of formal two-state neurons the sequential learning problem is considered: a set of patterns has been memorized by adjusting a matrix of the synaptic efficacies. Then an extra pattern is presented to the network and should be stored in addition to previous patterns in such a way that the latter are not used. The problem has been solved in the framework of the pseudo-inverse learning rules. The resulting synaptic matrix has the same value as for other variants of the pseudo-inverse learning rule if memorizing of the previous patterns was performed by some pseudo-inverse rule.

## 1. Introduction

The learning problem is usually formulated for networks composed of two-state formal neurons (Ising spins) as follows: it is necessary to find a matrix of the synaptic efficacies $J_{ij}$ $(i, j = 1, 2, \ldots, N)$ so that a set of $p$ memorized patterns represented by the $N$-dimensional binary vectors $\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \ldots, \boldsymbol{\xi}^p$ are attractors (fixed points) of the network dynamics [1]. However, if one approaches real biological systems then some additional properties are necessary. So, suppose that the $p$ patterns $\{\boldsymbol{\xi}^\alpha\}$ were presented to the network and after a certain time interval a new set of $q$ patterns $\{\boldsymbol{\zeta}^\beta\}$ should be stored in addition to the previous one. It is obvious that this task can be achieved by memorizing the joint set of $(p+q)$ patterns anew, but in real situations the first set of patterns $\{\boldsymbol{\xi}^\alpha\}$ can already be missing while the second one $\{\boldsymbol{\zeta}^\beta\}$ enters the network. Thus the following problem arises.

Let us assume that a set $M = \{\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \ldots, \boldsymbol{\xi}^p\}$ of $p$ patterns has been memorized by adjusting the synaptic matrix $J$. Then an extra pattern $\boldsymbol{\zeta}$ is presented to the network and should also be stored. It is necessary to find such a matrix $\tilde{J}$ that: (i) all $(p+1)$ patterns will be memorized (i.e. they should be attractors of the network dynamics); and in addition (ii) the matrix $\tilde{J}$ must be determined through the matrix $J$ and vector $\boldsymbol{\zeta}$ only, but not through the set $M$.

This problem will be referred to as the sequential learning problem.

The importance of this problem is caused by the fact that relearning a neural network in order to memorize patterns presented in different time periods can be impossible if information is obtained from the real world. In fact, all biological nervous systems in their everyday life fulfil a sequential learning: new events add new images in the memory. Artificial systems should be learnt sequentially in order to operate in changing situations.

A very simple and elegant solution of the sequential learning problem is given in fact by the Hebbian learning rule [1, 2]:

$$J_{ij} = \frac{1}{N} \sum_{\mu} \xi_i^{\mu} \xi_j^{\mu}. \tag{1}$$

However, as is known, this rule performs well in the case of random uncorrelated patterns only, whereas patterns entering the network from the real world can be correlated.

Correlated patterns are acceptable inputs for the pseudo-inverse (PI) learning rule proposed by Kohonen [3] and Personnaz *et al* [4]:

$$J_{ij} = \frac{1}{N} \sum_{\mu,\nu} C_{\mu\nu}^{-1} \xi_i^{\mu} \xi_j^{\nu} \tag{2}$$

where $C^{-1}$ is the inverse matrix for the pattern correlation matrix

$$C^{\mu\nu} = \frac{1}{N} \sum_{i=1}^{N} \xi_i^{\mu} \xi_i^{\nu} \tag{3}$$

and summing over $\mu$ and $\nu$ is performed from 1 to $p$ if the memorized patterns are all linearly independent. This rule has been formulated in a local iterative way by Diederich and Opper [5]:

$$J_{ij}^{l+1} = J_{ij}^{l} + \frac{1}{N} \sum_{\mu} \left( 1 - \sum_{k} J_{ik}^{l} \xi_i^{\mu} \xi_k^{\mu} \right) \xi_i^{\mu} \xi_j^{\mu} \qquad l = 0, 1, 2, \ldots \tag{4}$$

Berryman *et al* [6] proved that the latter algorithm converges for any set of patterns, including linearly dependent ones, and the resulting synaptic efficacies are the same as for any maximal linearly independent subset of the patterns $L$. The iterative procedure (4) produces the matrix $J$ which can be given by equations (2) and (3) if they are restricted to the subset $L$.

The PI learning rule also possesses other good properties [7], but in general this rule is a solution of the learning problem but not of the sequential one. In fact the latter problem has been solved in the framework of the PI rule for orthogonal patterns by Personnaz *et al* [8], but this solution is equivalent to the Hebbian rule (1) (see equations (16) and (18) below). Moreover, as is pointed out above, the patterns to be stored can be from the real world and therefore they can be non-orthogonal.

Here the restriction of orthogonality of the stored patterns will be removed. In other words, the aim of this paper is to find a solution of the sequential learning problem in the framework of the PI approach for arbitrary patterns.

The rule defined by equations (2) and (3) or, equivalently, by equation (4) will be referred to as the standard PI learning rule as opposed to the sequential PI rule given below. The synaptic matrices produced will be called standard and sequential PI synaptic matrices, respectively.

The paper is organized as follows. In section 2 a sequential PI learning rule is constructed, section 3 is devoted to comparison between the sequential and standard PI synaptic matrices and section 4 contains some conclusions.

## 2. Finding a sequential pseudo-inverse learning rule

The PI learning rule defines matrix $J$ as a solution of the equations [3, 4]

$$\sum_{j=1}^{N} J_{ij}\xi_j^\mu = \xi_i^\mu \qquad \mu = 1, 2, \ldots, p \qquad i = 1, 2, \ldots, N \tag{5}$$

or in the matrix form

$$JK = K \tag{6}$$

where matrix $K$ consists of the components of the memorized vectors: $K_{i\mu} = \xi_i^\mu$. In the framework of the PI approach the sequential learning problem can be formulated as follows. Let us assume that for a given matrix $K$ matrix $J$ has been adjusted so that equation (6) is satisfied. Then for a given vector $\boldsymbol{\zeta}$ it is necessary to find a matrix $\tilde{J}$ so that: (i) the equations

$$\tilde{J}K = K \tag{7}$$

$$\tilde{J}\boldsymbol{\zeta} = \boldsymbol{\zeta} \tag{8}$$

are satisfied; and (ii) matrix $\tilde{J}$ is expressed through matrix $J$ and vector $\boldsymbol{\zeta}$ only, but not through matrix $K$.

We start with equation (8). Its general solution can be represented in the form [8]

$$\tilde{J} = \eta + J(I-\eta) + D(I-\eta) \tag{9}$$

$$\eta = \frac{1}{N}\boldsymbol{\zeta} \otimes \boldsymbol{\zeta}. \tag{10}$$

Here the sign $\otimes$ denotes the direct product, i.e.

$$[\boldsymbol{A} \otimes \boldsymbol{B}]_{ij} = A_i B_j$$

for arbitrary vectors $\boldsymbol{A}$, $\boldsymbol{B}$; $I$ is the unit matrix and $D$ is an arbitrary matrix. Substitution of equation (9) into equation (7) yields the following equation for the matrix $D$:

$$D(I-\eta)K = (J-I)\eta K. \tag{11}$$

Let $Q$ be the PI matrix [9] for the matrix $(I-\eta)K$. Then the solution of equation (11) is

$$D = (J-I)\eta K Q$$

and equation (9) takes the form

$$\tilde{J} = \eta + J(I-\eta) + (J-I)\eta K Q(I-\eta). \tag{12}$$

(Note that for calculating matrix $Q$ one can formulate an iterative procedure on the analogy of [5, 10].) Thus equation (12) determines a learning rule for the network. However, this is a solution of the learning problem only, but not of the sequential learning problem because the third term in equation (12) depends on matrix $K$ (directly and through matrix $Q$).

Nevertheless, equation (12) gives the possibility of formulating an algorithm for sequential learning. To this end let us represent the vector $\boldsymbol{\zeta}$ as

$$\boldsymbol{\zeta} = \sum_{\sigma=1}^{p} a^\sigma \boldsymbol{\xi}^\sigma + b\boldsymbol{\psi} \tag{13}$$

where the first term is a projection of the vector $\zeta$ on to the subspace $H_\xi$ spanned by the memorized vectors $\xi^1, \xi^2, \ldots, \xi^p$ and the second term is an orthogonal complement of the vector $\zeta$, i.e. the binary vector $\psi$ is orthogonal to the subspace $H_\xi$:

$$(\psi\xi^\sigma) = 0 \qquad \sigma = 1, 2, \ldots, p. \tag{14}$$

The coefficients $a^1, a^2, \ldots, a^p$ and $b$ in equation (13) are real constants. (Note that two $N$-dimensional binary vectors $\boldsymbol{A}$, $\boldsymbol{B}$ could in principle be orthogonal, i.e. $(\boldsymbol{AB}) \equiv \Sigma A_i B_i = 0$, if the number $N$ is even. Therefore, our method is directly applied for an even $N$ only. If the number of the neurons $N$ is odd then one auxiliary neuron can, in addition, be introduced.)

First let us consider two particular cases.

(i) Let the vector $\zeta$ have no orthogonal part, i.e. $b\psi = 0$ and $\zeta = \Sigma a^\sigma \xi^\sigma$. Then $(J - I)\eta = 0$ and equation (12) gives $\tilde{J} = J$, i.e. no changes appear in the matrix of the synaptic efficacies. The reason is that the PI learning rule automatically memorizes all valid linear combinations of the embedded vectors [4, 7] or, in other words, the PI rule matrix $J$ is a matrix projecting onto the subspace $H_\xi$ spanned by the memorized patterns, and adding linearly dependent patterns does not change this subspace [4, 6].

(ii) Let the vector $\zeta$ have no projection part, i.e. $\Sigma a^\sigma \xi^\sigma = 0$ or $\zeta = b\psi$. Since $\zeta$ and $\psi$ are both binary vectors then $b = \pm 1$ and

$$\zeta = \pm\psi. \tag{15}$$

As a result $\eta K = 0$ and equation (12) takes the form

$$\tilde{J} = \eta + J(I - \eta)$$

or

$$\tilde{J} = J + (I - J)\frac{1}{N}\zeta \otimes \zeta. \tag{16}$$

This result coincides, as it should, with the learning rule [8] proposed by Personnaz *et al* for orthogonal patterns. Note that from equations (2) and (14) the condition follows that

$$J\psi = 0 \tag{17}$$

which is due to the fact that the PI matrix $J$ is a matrix projecting on to the subspace $H_\xi$, but $\psi$ is orthogonal to $H_\xi$. Taking into account equations (15) and (17), from equation (16) one obtains

$$\tilde{J} = J + \frac{1}{N}\zeta \otimes \zeta. \tag{18}$$

This is merely the Hebbian learning rule (1).

Consider now the general case (13). As is mentioned above, the PI learning rule automatically memorizes all valid linear combinations of the memorized vectors. Therefore it is sufficient to store the orthogonal component $\psi$ of the vector $\zeta$ only. Thus on the analogy of equation (18) the following expression for the matrix $\tilde{J}$ is obtained:

$$\tilde{J} = J + \frac{1}{N}\psi \otimes \psi. \tag{19}$$

It may be checked that the matrix $\tilde{J}$ defined by this equation satisfies equations (7) and (8). Thus the sequential learning problem is reduced to finding the orthogonal component $\boldsymbol{\psi}$ for an arbitrary vector $\boldsymbol{\zeta}$.

The latter problem can be treated as follows. First let us multiply equation (13) by the vector $\boldsymbol{\psi}$. Then taking into account equation (14) one has

$$b=\frac{1}{N}(\boldsymbol{\psi}\boldsymbol{\zeta}). \tag{20}$$

Further, from equation (5) it follows that

$$(I-J)\sum_{\sigma=1}^{p} a^{\alpha}\boldsymbol{\xi}^{\sigma}=0. \tag{21}$$

Since according to equation (13)

$$\sum_{\sigma=1}^{p} a^{\sigma}\boldsymbol{\xi}^{\sigma}=\boldsymbol{\zeta}-b\boldsymbol{\psi}$$

then equation (21) takes the form

$$b(I-J)\boldsymbol{\psi}=(I-J)\boldsymbol{\zeta}$$

or

$$b\boldsymbol{\psi}=\boldsymbol{\zeta}-J\boldsymbol{\zeta} \tag{22}$$

where condition (17) was used. The last equation clearly means that since $J\boldsymbol{\zeta}$ is a projection of the vector $\boldsymbol{\zeta}$ on to the subspace $H_{\xi}$, hence $(\boldsymbol{\zeta}-J\boldsymbol{\zeta})$ is that part of the vector $\boldsymbol{\zeta}$ that is orthogonal to $H_{\xi}$.

After substitution of equation (22) into equation (20) the following expression is given:

$$b^2=N^{-1}(N-\boldsymbol{\zeta}^T J\boldsymbol{\zeta})=N^{-1}\left(N-\sum_{i,j}\zeta_i J_{ij}\zeta_j\right). \tag{23}$$

From equations (19), (22) and (23) one obtains the final expression for the matrix $\tilde{J}$:

$$\tilde{J}=J+\frac{(\boldsymbol{\zeta}-J\boldsymbol{\zeta})\otimes(\boldsymbol{\zeta}-J\boldsymbol{\zeta})}{N-\boldsymbol{\zeta}^T J\boldsymbol{\zeta}} \tag{24}$$

or

$$\tilde{J}_{ij}=J_{ij}+\left(N-\sum_{k,l=1}^{N}\zeta_k J_{kl}\zeta_l\right)^{-1}\left(\zeta_i-\sum_{k=1}^{N}J_{ik}\zeta_k\right)\left(\zeta_j-\sum_{k=1}^{N}J_{jk}\zeta_k\right).$$

It is easy to check that this matrix indeed gives a solution of equations (7) and (8) and therefore equation (24) is a sequential PI synaptic matrix.

It is obvious that all patterns presented to the network can be memorized by this procedure so that the synaptic matrix for patters $\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \ldots, \boldsymbol{\xi}^p$ is given by the following recurrent expression:

$$J^{\mu}=J^{\mu-1}+(N-(\boldsymbol{\xi}^{\mu})^T J^{\mu-1}\boldsymbol{\xi}^{\mu})^{-1}(\boldsymbol{\xi}^{\mu}-J^{\mu-1}\boldsymbol{\xi}^{\mu})\otimes(\boldsymbol{\xi}^{\mu}-J^{\mu-1}\boldsymbol{\xi}^{\mu}) \tag{25}$$

where $\mu=1,2,\ldots,p$.

## 3. A connection between sequential and standard pseudo-inverse synaptic matrices

Let us consider how matrix $\tilde{J}$, given by the sequential rule (24), is connected with the synaptic matrix produced by the standard PI learning rule (2), (3) or, equivalently, by the iterative algorithm (4). This question has no obvious answer. Indeed the general solution of the equation $JK = K$ has the form [8]

$$J = KK^I + B(I - KK^I)$$

where $K^I$ is the PI matrix for matrix $K$ and $B$ is an arbitrary matrix. Therefore two PI synaptic matrices can be different due to different matrices $B$. Plainly, matrix $\tilde{J}$ depends on what matrix $J$ is used in equation (24). First consider the case when $J$ is given by the standard PI rule (2), (3) or (4).

To be more exact, note that we consider the following two sets of patterns: $M = \{\xi^1, \xi^2, \ldots, \xi^p\}$ and $M_1 = \{\xi^1, \xi^2, \ldots, \xi^p, \xi^{p+1}\} = \{M, \zeta\}$ with one additional pattern $\zeta \equiv \xi^{p+1}$. Matrix $J$ is the standard PI synaptic matrix for the set $M$. For the set $M_1$ one can construct two synaptic matrices: the sequential matrix $\tilde{J}$ by means of rule (24) and a standard PI matrix $J_1$ using rules (2), (3) or (4). Our aim in this section is to compare the matrices $\tilde{J}$ and $J_1$.

To do this let us consider how the matrices $J$ and $J_1$ can be calculated. Let $L$ be a maximal linearly independent subset of the patterns (MLISP) for the set $M$. Then matrix $J$ can be defined by equations (2) and (3) where only the patterns from the subset $L$ contribute.

In the computation of matrix $J_1$ it is reasonable to begin with two particular cases, as in the previous section.

(i) Let vector $\zeta$ be a linear combination of vectors $\xi^1, \xi^2, \ldots, \xi^p$. Then the subset $L$ can be used as an MLISP for the set $M_1$. Therefore in this case $J_1 = J$ and $\tilde{J} = J_1$.

(ii) Let the vector $\zeta$ be orthogonal to the set $M$. Then let us take the set $\{L, \zeta\}$ as an MLISP for the set $M_1$. According to equation (3) one finds the correlation matrix $C_1$ for this subset:

$$C_1 = \begin{vmatrix} C & \mathbf{0}_q \\ \mathbf{0}_q^T & 1 \end{vmatrix}$$

where $C$ is the correlation $q \times q$-matrix for the patterns from the subset $L$ and $\mathbf{0}_q$ is the $q$-dimensional zero vector ($q$ is the number of patterns in the subset $L$). It is obvious that

$$C_1^{-1} = \begin{vmatrix} C^{-1} & \mathbf{0}_q \\ \mathbf{0}_q^T & 1 \end{vmatrix}$$

and in turn equation (2) gives

$$J_1 = J + \frac{1}{N} \zeta \otimes \zeta \tag{26}$$

i.e. $\tilde{J} = J_1$ again (cf equations (18) and (26)).

Now turn to the general case (13). Matrix $J_1$ can be given by equations (2) and (3) where $\{L, \zeta\}$ is used as an MLISP. Let us also consider, however, the two following auxiliary sets: $M_2 = \{M, \psi\}$ and $M_3 = \{M, \psi, \zeta\}$, and denote by $J_2$ and $J_3$ the synaptic matrices produced by the rule (2) and (3) for these sets. For set $M_2$ we take the subset

$\{L, \psi\}$ as an MLISP and analogously to equation (26) we have

$$J_2 = J + \frac{1}{N} \psi \otimes \psi \tag{27}$$

i.e. $J_2 = \tilde{J}$ (cf equations (19) and (27)).

For set $M_3$ we can use the subset $\{L, \psi\}$ as an MLISP and therefore we have $J_3 = J_2$. On the other hand, we can take the subset $\{L, \zeta\}$ as an MLISP and in this case we obtain $J_3 = J_1$. But the standard PI synaptic matrix is unique and as a result we arrive at the chain $J_1 = J_3 = J_2 = \tilde{J}$. Thus for an arbitrary pattern $\zeta$ one has $\tilde{J} = J_1$.

Now consider the case when all patterns presented to the network are stored by means of the sequential rule, i.e. prescription (25) is used. Let us assume that $J^0 = 0$ (*tabula rasa*). Then equation (25) yields for $p = 1$:

$$J^1 = \frac{1}{N} \xi^1 \otimes \xi^1.$$

But the same expression is given by the standard PI rule (2), (3). Taking into account the result proved above (coincidence of $\tilde{J}$ and $J_1$), we conclude by induction that the sequential and standard PI learning rules produce the same synaptic matrices.

## 4. Conclusions

Projective properties of the synaptic matrices produced by PI learning rules made it possible to construct the sequential learning rule (24). This rule allows a new pattern entering the network to be memorized in a simple manner, in addition to the patterns previously stored. An essential feature of this rule is that patterns can be correlated (and, in particular, can be linearly dependent), as can occur for patterns arriving from the real world. Another attractive property is that the resulting synaptic matrix is the same as in other variants of the PI learning rule (see equations (2), (3) or (4)) if (i) previous storing of a set of patterns $\xi^1, \xi^2, \ldots, \xi^p$ was performed by rules (2), (3) or (4), or (ii) all patterns presented to the network are memorized by the sequential rule (25). Hence for these cases the storage capacity and basins of attraction have same values for both standard and sequential PI learning rules.

Note that, at least, a rough analogy can be made between the sequential rule and learning in real biological systems: first an incoming signal ($\zeta$) is perceived and treated on the basis of previous experience ($J$) and, further, only new information ($\psi$) is memorized.

## Acknowledgments

The author wishes to thank A V Kartynnick for useful discussions and the referees for stimulating comments.

## References

[1] Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554
[2] Hebb D 1949 *The Organization of Behavior* (New York: Wiley)

[3] Kohonen T 1988 *Self-organization and Associative Memory* (Berlin: Springer)
[4] Personnaz L, Guyon I and Dreyfus G 1985 *J. Physique Lett.* **46** L359; 1986 *Phys. Rev.* A **34** 4217
[5] Diederich S and Opper M 1987 *Phys. Rev. Lett.* **58** 949
[6] Berryman K W, Inchiosa M E, Jaffe A M and Janowsky S A 1990 *J. Phys. A: Math. Gen.* **23** L223
[7] Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380
[8] Personnaz L, Guyon I, Dreyfus G and Toulouse G 1986 *J. Stat. Phys.* **43** 411
[9] Albert A 1972 *Regression and Moore–Penrose Pseudo-inverse* (New York: Academic)
Strang G 1976 *Linear Algebra and its Applications* (New York: Academic)
[10] Opper M 1987 *Europhys. Lett.* 4 979