Convergence of an iterative neural network learning algorithm for linearly dependent patterns



1990 J. Phys. A: Math. Gen. 23 L223

(http://iopscience.iop.org/0305-4470/23/5/007)

J. Phys. A: Math. Gen. 23 (1990) L223–L228. Printed in the UK

LETTER TO THE EDITOR

# Convergence of an iterative neural network learning algorithm for linearly dependent patterns†

Kenneth W Berryman‡, Mario E Inchiosa, Arthur M Jaffe and Steven A Janowsky§

Department of Physics, Harvard University, Cambridge, MA 02138, USA

Received 26 April 1989, in final form 13 November 1989

**Abstract.** We show that a local iterative learning algorithm of Diederich and Opper converges for any set of patterns, including linearly dependent sets, and that the resulting coupling constants for a given set of patterns are the same as those given by the pseudo-inverse rule for any maximal linearly independent subset of the given set. Thus, the matrix of couplings produced by the algorithm is the desired projection matrix onto the subspace spanned by the training set.

In proposing learning rules for neural network associative memories, researchers have often excluded certain sets of patterns. For example, the Hopfield model with a Hebbian learning rule [4] does not perform well when the patterns to be stored are correlated. When the number of patterns increases linearly with the number of neurons, even the small overlap of randomly generated patterns causes errors in their storage. Also, storage capacity drastically decreases when patterns are 'biased' (unequal numbers of 'on' and 'off' bits) [1].

In [6], Kanter and Sompolinsky deal with correlated patterns by using the non-local pseudoinverse learning rule of Personnaz *et al* [10]. They only implement the pseudo-inverse rule (inspired by the work of Kohonen [9]) for linearly independent patterns, although their formulae extend to the more general case. Diederich and Opper [3] enhanced the usefulness of the pseudoinverse rule by providing a local algorithm for computing it. Locality of learning algorithms is desirable for their efficient hardware implementation in large neural network associative memories. However, Diederich and Opper also only characterised the behaviour of their algorithm for linearly independent sets of training patterns.

Is it necessary to consider storage of a linearly dependent set of training patterns? It could occur that the training patterns, presumably data coming from the real world, are not linearly independent. We would want our system to be able to handle this situation. Also, a rigorous statistical mechanical analysis, such as that of Koch and Piasko [8] or Inchiosa [5] (reported in a preliminary form in Berryman *et al* [2]), demands understanding learning rule behaviour for all possible sets of input patterns. Yet, in the simple case of a set with a pattern appearing twice, the set becomes linearly

† Work supported in part by National Science Foundation grant PHY/DMS 88-16214.

‡ Address after 1 October 1989: Physics Department, Varian Buidling, Stanford University, Stanford, CA 94305, USA

§ Supported in part by National Science Foundation grant PHY 87-06420.

0305-4470/90/050223+06\$03.50 

dependent, prohibiting the use of a learning rule implementation which is restricted to linearly independent sets. These examples clearly illustrate the need to consider storage of linearly dependent patterns.

In the following we will show that linearly dependent patterns are perfectly acceptable inputs for the iterative algorithm of Diederich and Opper. We will show that for a given set of linearly dependent patterns the algorithm converges to the same coupling constants as for any maximal linearly independent subset of the patterns, resulting in a coupling matrix $J$ which is the desired projection matrix. This is the correct set of couplings: learning on a maximal linearly independent subset would cause the entire training set to be memorised because the pseudoinverse rule automatically memorises all valid linear combination states [6]. We will demonstrate the convergence of the iterative learning algorithm by exploiting results on Gauss-Seidel iteration for singular matrices [7]. Finally, our practical experience with this method indicates that, in the examples we have studied, the rate of learning is faster for linearly dependent sets of patterns than for a smaller set obtained by first using a linear independence filter.

Our system has the Hamiltonian

$$\mathscr{H} = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} J_{ij}S_iS_j \tag{1}$$

where $S_i \in \{+1, -1\}$. The coupling constants $\{J_{ij}\}$ are given by learning rule II of [3], with the modification that linearly dependent sets of training patterns are allowed. We are primarily concerned with learning, i.e. adjusting $\{J_{ij}\}$ so that the minima of $\mathscr{H}$ are appropriately placed. The complementary problem of associative recall consists of causing the $\{S_i\}$ to evolve from an initial condition representing partial knowledge of a stored pattern to a local minimum of $\mathscr{H}$ representing full recovery of the pattern. In fact, Kanter and Sompolinsky show that the prototype patterns *and* their (rare) valid linear combinations which are valid Ising spin states are actually global minima of $\mathscr{H}$ [6].

To govern the dynamics of recall, we use Kanter and Sompolinsky's approach of eliminating self-interactions, i.e. we use the synaptic matrix $J_{ij}(1-\delta_{ij})$, which preserves the locations of the minima of $\mathscr{H}$ [6]. Thus, at zero temperature the $S_i$ evolve as follows: $S_i \to -S_i$ if and only if $S_i \sum_{j\neq i} J_{ij}S_j < 0$. If self-interaction terms are not eliminated, they can reduce the size of the basins of attraction for the stored patterns. Such terms arise in storing correlated patterns, whether linearly dependent or not.

The input to the learning rule consists of $p$ $N$-bit patterns $\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \ldots, \boldsymbol{\xi}^p$ with $\xi_i^\mu \in \{+1, -1\}$. Roman letters index the $N$ neurons and Greek letters index the $p$ patterns. One applies patterns one at a time to the network and updates the couplings. The change in $J_{ij}$ in learning cycle $l$, $l = 0, 1, 2, \ldots$, due to presentation of pattern $\mu$ is

$$\delta J_{ij}(l, \mu) \equiv N^{-1}\left[1 - \sum_k J_{ik}(l, \mu)\xi_i^\mu \xi_k^\mu\right]\xi_i^\mu \xi_j^\mu \tag{2}$$

where $J_{ij}(l, \mu)$ is the value of $J_{ij}$ before the presentation of the pattern, $J_{ij} + \delta J_{ij}$ is the value afterwards, and $J_{ij}(0, 1) \equiv 0$.

Since individual rows of the coupling matrix $J$ evolve independently of each other, we can consider just one row, say row $i$. Writing $\mathscr{J}_j \equiv J_{ij}$ and $\sigma_j^\mu \equiv \xi_i^\mu \xi_j^\mu$ we have

$$\delta\mathscr{J}_j(l, \mu) = N^{-1}\left[1 - \sum_k \mathscr{J}_k(l, \mu)\sigma_k^\mu\right]\sigma_j^\mu. \tag{3}$$

We can rewrite this equation as

$$\delta \mathscr{J}_j(l, \mu) = N^{-1} \delta x^\mu(l) \sigma_j^\mu \tag{4}$$

where we are introducing a new set of variables $\{x^\mu\}$, $\mu = 1, 2, \ldots, p$, called 'embedding strengths' which evolve according to the equation

$$x^\mu(l+1) - x^\mu(l) \equiv \delta x^\mu(l) = 1 - \sum_k \mathscr{J}_k(l, \mu) \sigma_k^\mu \qquad x^\mu(0) \equiv 0. \tag{5}$$

Since the $\boldsymbol{\sigma}^\mu$ may be linearly dependent, the expression of $\boldsymbol{\mathscr{J}}$ as a linear combination of $\boldsymbol{\sigma}^\mu$ may not be unique. Equation (5) defines a particular choice of a (possibly non-unique) decomposition of $\boldsymbol{\mathscr{J}}$ in terms of $\{\boldsymbol{\sigma}^\mu\}$:

$$\mathscr{J}_k(l, \mu) = N^{-1} \sum_{\nu < \mu} x^\nu(l+1) \sigma_k^\nu + N^{-1} \sum_{\nu \geq \mu} x^\nu(l) \sigma_k^\nu \qquad k = 1, 2, \ldots, N. \tag{6}$$

Substituting (6) into (5) gives the iteration equation

$$x^\mu(l+1) - x^\mu(l) = 1 - \sum_{\nu < \mu} B^{\mu\nu} x^\nu(l+1) - \sum_{\nu \geq \mu} B^{\mu\nu} x^\nu(l) \tag{7}$$

where $B^{\mu\nu} \equiv N^{-1} \Sigma_k \sigma_k^\mu \sigma_k^\nu$. Provided the $\{x^\mu(l)\}$ converge, we obtain from (6) the asymptotic expression

$$\mathscr{J}_k(\infty) \sim N^{-1} \sum_\nu x^\nu(\infty) \sigma_k^\nu. \tag{8}$$

We will show that (7) converges to a solution of

$$\boldsymbol{B}\boldsymbol{x} = \mathbf{1} \tag{9}$$

for any set of patterns $\{\boldsymbol{\xi}^\mu\}$. Our first step will be to show that a solution of (9) always exists. Equation (7) is the Gauss–Seidel iterative method for finding a solution to (9). While the Gauss–Seidel method is well known for non-singular systems, we use a classic, but less widely known, result of Keller [7] for singular systems.

Suppose we have a given set of $p$ patterns, with $q$ being the maximal size of any linearly independent subset of the patterns. For convenience we assume the first $q$ patterns are linearly independent. We can make this assumption without loss of generality because writing out (9) in terms of the $\boldsymbol{\xi}^\mu$ shows that re-ordering the patterns simply corresponds to re-ordering the $x^\mu$.

Let $B^\#$ denote the $q \times q$ matrix consisting of the first $q$ rows of the first $q$ columns of $B$. Consider the equation obtained by restricting (9) to the first $q$ patterns ($\boldsymbol{x}^\#$ and $\mathbf{1}^\#$ are $q$-vectors):

$$\boldsymbol{B}^\# \boldsymbol{x}^\# = \mathbf{1}^\#. \tag{10}$$

There exists a unique solution for $\boldsymbol{x}^\#$ if $B^\#$ is invertible. Now since

$$(\boldsymbol{v}, \boldsymbol{B}^\# \boldsymbol{v}) = \sum_{\mu=1}^{q} \sum_{\nu=1}^{q} B^{\#\mu\nu} v^\mu v^\nu = N^{-1} \sum_k \left( \sum_{\mu=1}^{q} v^\mu \xi_i^\mu \xi_k^\mu \right)^2 \tag{11}$$

and since the linear independence of the first $q$ patterns implies that $\Sigma_\mu\, y^\mu \xi_k^\mu = 0$ for all $k$ if and only if $\boldsymbol{y} = \mathbf{0}$, we see that $(\boldsymbol{v}, \boldsymbol{B}^\# \boldsymbol{v}) > 0$ for all $\boldsymbol{v} \neq \mathbf{0}$; i.e. $B^\#$ is positive definite and therefore invertible.

Let $\tilde{x} \equiv (B^{\#})^{-1}1^{\#}$, the solution of (10), and let $\bar{x} \equiv (\tilde{x}^1, \tilde{x}^2, \ldots, \tilde{x}^q, 0, \ldots, 0)$ be a $p$-vector. We claim that $\bar{x}$ is a solution of (9). First notice that since the last $p-q$ patterns are linearly dependent on the first $q$, there exists a set of real numbers $\{c^{\mu\alpha}\}$, $\mu = q+1, \ldots, p$, $\alpha = 1, \ldots, q$, such that

$$\xi_k^\mu = \sum_{\alpha=1}^{q} c^{\mu\alpha}\xi_k^\alpha \tag{12}$$

for $\mu > q$. Now for $\mu \leqslant q$,

$$\sum_{\nu=1}^{p} B^{\mu\nu}\bar{x}^\nu = \sum_{\nu=1}^{q} B^{\#\mu\nu}\tilde{x}^\nu = 1 \tag{13}$$

while for $\mu > q$,

$$\begin{aligned}\sum_{\nu=1}^{p} B^{\mu\nu}\bar{x}^\nu &= \xi_i^\mu\xi_i^\mu \sum_{\nu=1}^{p} B^{\mu\nu}\bar{x}^\nu \\ &= \xi_i^\mu \sum_{\nu=1}^{p} N^{-1}\sum_k \xi_k^\mu\sigma_k^\nu\bar{x}^\nu \\ &= \xi_i^\mu \sum_{\nu=1}^{p} N^{-1}\sum_k\sum_{\alpha=1}^{q} c^{\mu\alpha}\xi_k^\alpha\sigma_k^\nu\bar{x}^\nu \end{aligned}\tag{14}$$

using the linear dependence (12). Then using $(\xi_i^\alpha)^2 = 1$ (again),

$$\begin{aligned}(14) &= \xi_i^\mu \sum_{\alpha=1}^{q} c^{\mu\alpha}\xi_i^\alpha \sum_{\nu=1}^{p} N^{-1}\sum_k \xi_i^\alpha\xi_k^\alpha\sigma_k^\nu\bar{x}^\nu \\ &= \xi_i^\mu \sum_{\alpha=1}^{q} c^{\mu\alpha}\xi_i^\alpha \sum_{\nu=1}^{p} B^{\alpha\nu}\bar{x}^\nu \\ &= \xi_i^\mu \sum_{\alpha=1}^{q} c^{\mu\alpha}\xi_i^\alpha = \xi_i^\mu\xi_i^\mu = 1.\end{aligned}\tag{15}$$

So we have proved our claim.

We have shown that for any set of input patterns there exists a solution to the system given by equation (9). We now wish to show that the existence of this solution is enough to guarantee convergence of the iterative procedure. As a consequence of theorems 1 and 2 of Keller [7] we obtain the following.

*Theorem.* Let $B$ be a $p$th-order Hermitian matrix and $L$ be a non-singular matrix of order $p$ for which $P \equiv L + L^\dagger - B$ is positive definite. If $Bx = f$ has a solution, then the following two equivalent statements hold if and only if $B$ is positive semidefinite.

($a$) For every $x(0)$ the sequence $\{x(l)\}$ of

$$Lx(l+1) = (L-B)x(l) + f \qquad l = 0, 1, 2, \ldots \tag{16}$$

converges to a solution of $Bx = f$.

($b$) Define the iteration matrix $T \equiv I - L^{-1}B$. For every $e(0)$ the sequence $\{e(l)\}$ of

$$e(l+1) = Te(l) \qquad l = 0, 1, 2, \ldots \tag{17}$$

converges to a vector in the null space of $B$.

Equation (16) corresponds to our Gauss-Seidel iteration procedure (7) if we let $L$ be the lower triangular part of $B$ (including the diagonal). Since $L^{\mu\mu} = B^{\mu\mu} = 1$, $\det L = 1$ and $L$ is invertible. Also, $P = L + L^{\dagger} - B = I$, and $I$ is trivially positive definite. Our matrix $B$ is positive semidefinite, since

$$(\boldsymbol{v}, B\boldsymbol{v}) = N^{-1} \sum_k \left( \sum_{\mu=1}^{p} \sigma_k^\mu v^\mu \right)^2 \geqslant 0. \tag{18}$$

Since we have already produced a solution to (9), the conclusions of the theorem apply.

Using (8), we note that

$$\mathscr{J}_k[\tilde{\boldsymbol{x}}] = N^{-1} \sum_{\mu=1}^{q} \tilde{x}^\mu \sigma_k^\mu = N^{-1} \sum_{\mu=1}^{p} \bar{x}^\mu \sigma_k^\mu = \mathscr{J}_k[\bar{\boldsymbol{x}}]. \tag{19}$$

Thus $\{J_{ij}\}$ for the constructed solution $\bar{\boldsymbol{x}}$ equals $\{J_{ij}\}$ for the solution in the case of linearly independent patterns, $\tilde{\boldsymbol{x}}$.

The iterative procedure gives us a solution of the form $\hat{\boldsymbol{x}} = \bar{\boldsymbol{x}} + \boldsymbol{w}$, where $\boldsymbol{w}$ is some vector satisfying $B\boldsymbol{w} = \boldsymbol{0}$. But in that case,

$$0 = (\boldsymbol{w}, B\boldsymbol{w}) = N^{-1} \sum_k \left( \sum_{\mu=1}^{p} \sigma_k^\mu w^\mu \right)^2. \tag{20}$$

Therefore $\Sigma_\mu \sigma_k^\mu w^\mu = 0$, and

$$\mathscr{J}_k[\hat{\boldsymbol{x}}] = N^{-1} \sum_{\mu=1}^{p} (\bar{x}^\mu + w^\mu) \sigma_k^\mu = \mathscr{J}_k[\bar{\boldsymbol{x}}]. \tag{21}$$

Equations (19) and (21) imply the couplings $\{J_{ij}\}$ produced by the iterative procedure are the same as the couplings obtained by considering only the $q$ linearly independent patterns.

Had we chosen a different set of $q$ linearly independent patterns, we would have $\tilde{\boldsymbol{x}}' = (B^{\#\prime})^{-1} \boldsymbol{1}^{\#}$. In analogy with (19) we would have $\mathscr{J}_k[\tilde{\boldsymbol{x}}'] = \mathscr{J}_k[\bar{\boldsymbol{x}}']$. Since $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{x}}'$ are both solutions of $B\boldsymbol{x} = \boldsymbol{1}$, we have $\bar{\boldsymbol{x}}' = \bar{\boldsymbol{x}} + \boldsymbol{w}'$, with $B\boldsymbol{w}' = \boldsymbol{0}$. By analogy with (21) we have $\mathscr{J}_k[\bar{\boldsymbol{x}}'] = \mathscr{J}_k[\bar{\boldsymbol{x}}]$. Thus $\mathscr{J}_k[\tilde{\boldsymbol{x}}'] = \mathscr{J}_k[\bar{\boldsymbol{x}}] = \mathscr{J}_k[\tilde{\boldsymbol{x}}]$, verifying the irrelevance of the choice of maximal linearly independent subset.

Finally, let us analyse the resulting matrix $J$. For the case where the patterns are all linearly independent, it is easy to show [3] that the iterative learning rule converges to the same matrix as that of Personnaz *et al* [10]

$$J_{ij} = N^{-1} \sum_{\mu\nu} (C^{-1})^{\mu\nu} \xi_i^\mu \xi_j^\nu \tag{22}$$

where $C$ is the pattern correlation matrix:

$$C^{\mu\nu} = N^{-1} \sum_k \xi_k^\mu \xi_k^\nu. \tag{23}$$

When the patterns are linearly dependent, we have shown that the iterative procedure produces the $\{J_{ij}\}$ obtained by restricting (22) and (23) to any maximal linearly independent subset of the patterns.

One of the hopes for using neural networks as memory devices is that they behave in a 'fail-soft' manner—when we try to learn too many patterns, or patterns of the wrong type, we hope that most of the knowledge stored within the network remains accessible.

Many learning rules have difficulty storing strongly correlated patterns. In one sense, sets of patterns which are linearly dependent are more strongly correlated than any other sets of patterns. Therefore, understanding what happens when we try to store linearly dependent patterns should give us some insight into the worst case behaviour of such learning rules.

For the Diederich and Opper algorithm we have shown that even for linearly dependent patterns the iterative procedure converges. The result has a nice interpretation in terms of the result for any maximal linearly independent subset of the patterns: the couplings obtained are identical. Thus, *adding linearly dependent patterns to the set of stored patterns does not change the couplings.* This is as it should be, since the pseudoinverse rule coupling matrix is a projection matrix onto the subspace spanned by the training patterns [2, 10], and adding linearly dependent patterns does not change this subspace.

Typical neural network simulations (or applications) do not check prototype patterns for linear independence or other criteria needed to ensure consistent behaviour. Therefore, learning algorithms must be able to handle gracefully any set of training patterns. If neural networks are going to become mainstream computational and storage devices, their behaviour on unexpected input must be considered.

## References

[1] Amit D J, Gutfreund H and Sompolinsky H 1987 Information Storage in Neural Networks with Low Levels of Activity *Phys. Rev.* A **35** 2293-303

[2] Berryman K W, Inchiosa M E, Jaffe A M and Janowsky S A 1990 Extending the Pseudoinverse Rule *Neural Networks and Spin Glasses* ed W K Theumann and R Köberle (Singapore: World Scientific) in press

[3] Diederich S and Opper M 1987 Learning of Correlated Patterns in Spin-Glass Networks by Local Learning Rules *Phys. Rev. Lett.* **58** 949-52

[4] Hopfield J J 1982 Neural Networks and Physical Systems with Emergent Collective Computational Abilities *Proc. Nat. Acad. Sci.* **79** 2554

[5] Inchiosa M Rigorous Results for Associative Memory Using the Pseudoinverse Rule in preparation

[6] Kanter I and Sompolinsky H 1987 Associative Recall of Memory without Errors *Phys. Rev.* A **35** 380-92

[7] Keller H B 1965 On the Solution of Singular and Semidefinite Linear Systems by Iteration *SIAM J. Numer. Anal.* series B, **2** 281-90

[8] Koch H and Piasko J 1989 Some Rigorous Results on the Hopfield Neural Network Model *J. Stat. Phys.* **55** 903-28

[9] Kohonen T 1984 *Self-Organization and Associative Memory* (Berlin:Springer)

[10] Personnaz L, Guyon I and Dreyfus G 1985 Information Storage and Retrieval in Spin-Glass like Neural Networks *J. Physique Lett.* **46** L359-L365