Vol. 40, No.1, March 1996, pp. ***–***. / Journal of Mathematical Psychology

A Widrow-Hoff learning-rule for a generalization of the linear auto-associator

Hervé Abdi*†, Dominique Valentin*, Betty Edelman*, & Alice J. O'Toole*

**The University of Texas at Dallas, and †Université de Bourgogne à Dijon.*

A generalization of the linear auto-associator that allows for differential importance and non-independence of both the stimuli and the units has been described previously by Abdi (1988). This model was shown to implement the general linear model of multivariate statistics. In this note, a proof is given that the Widrow-Hoff learning rule can be similarly generalized and that the weight matrix will converge to a generalized pseudo-inverse when the learning parameter is properly chosen. The value of the learning parameter is shown to be dependent only upon the (generalized) eigenvalues of the weight matrix and not upon the eigenvectors themselves. This proof provides a unified framework to support comparison of neural network models and the general linear model of multivariate statistics.

1. Introduction

The linear auto-associator, first described in the early 1970s by Anderson (1972) and Kohonen (1972), has been applied to many problems in pattern recognition and neural modeling. In brief, an auto-associator is a special case of neural network models in which the association between an input pattern and itself is learned. As such, the linear auto-associator or auto-associative memory constitutes a powerful pattern completion device because it is capable of reconstructing learned patterns when noisy or incomplete versions of the learned patterns are used as "memory keys." Kohonen (1977), for example, showed that an auto-associative memory can be used to store images of faces and reconstruct the original faces when features have been omitted or degraded.

When the linear auto-associator is viewed as a "neural network," elements of the weight matrix are seen as the connection strengths between "cells" or units. In the classical version of the auto-associator, all of the neuron-like units have the same importance and are independent of one another. Similarly, the stimuli stored in the memory are treated as equally important and are not interdependent in any way (*e.g.,* no *a priori* associations exist among stimuli). While this type of model has been shown to be useful in many pattern recognition applications, the assumption of equal importance and independence of both units and stimuli might be quite unrealistic in some cases. Numerous cognitive science or pattern recognition applications require the *a priori* imposition of constraints operating at the level of individual parts of the input code and/or at the level of individual stimuli.

One example of the usefulness of being able to give a differential importance to different units of an auto-associator can be found in the domain of face perception. While it has been shown in the past that a classical auto-associator can be used successfully for recognizing and categorizing faces along visually derived dimensions such as sex or race (cf. Valentin, Abdi & O'Toole, 1994, for a review), this model falls short of exhibiting some properties characteristic of human subjects. For example, it is well known that

Correspondence about this paper should be send to Hervé Abdi, Program in Applied Cognition and Neuroscience, The University of Texas at Dallas, Richardson, TX 75083-0688, U.S.A.; or to Hervé Abdi, Université de Bourgogne, Faculté des Sciences Gabriel, Boulevard Gabriel, 21004 Dijon Cedex, France. E-mail: herve@utdallas.edu; herve@u-bourgogne.fr
Part of this paper has been presented at the 24th Annual Mathematical Psychology Meeting (1991), and the 1992 Mind Meeting on neural networks. We are grateful to Jerry R. Busemeyer, Jean-Claude Falmagne, and an anonymous reviewer for their helpful remarks on a previous version of this paper.

all the features in a face are not equally important to discrimination between faces (Shepherd, Davies, & Ellis, 1981). Likewise, numerous empirical studies indicate that some faces are more distinctive or more memorable than others (Light, Kayra-Stuart, & Hollander, 1979). Hence, to provide a psychologically relevant model of face perception it might be useful to be able to give different importance to certain parts of the code and the stimuli.

Abdi (1988) described a generalization of the linear auto-associator that allows for differential importance and nonindependence of both the stimuli and the neuron-like units. In this model, differential importance and nonindependence are defined as a set of constraints expressed *via* positive definite square matrices. These two constraint matrices operate on the associator; one at the level of stimulus input and the other at the level of individual units. The model then is able to incorporate *a priori* biases in the stimulus set and in different parts of the stimulus code. Abdi (1988), however, does not provide an iterative learning rule for the generalized auto-associator comparable to the standard Widrow-Hoff learning rule used for most neural networks. The main goal of this note is to describe and analyze such a rule.

In previous work (Abdi, Valentin & O'Toole, 1996), we applied this generalized model to the problem of categorizing faces by sex and compared its performance to that of a classical auto-associator. Face images were digitized. A cell or unit of the neural network corresponded to a pixel of the digitized image. Each cell of the memory responded as the inverse of its activation across the learning period. This particular type of constraint emphasized parts of the code (*i.e.,* pixels) that are the most useful to discriminate between different stimuli. Specifically, if a cell is active for all the faces it does not provide information for any discrimination between faces. On the other hand, a cell active for subsets of faces or individual faces is very useful for discriminating between faces or groups of faces. Since for this classification task there was no indication that some faces would be more useful than others, the stimuli were all given the same importance. We showed that, with this specific set of constraints, the generalized model was able to learn the task not only as accurately as a classical auto-associator, but also considerably faster.

Other choices of constraints are possible for both the stimulus set and the stimulus code. For example, to stay in the domain of face perception, empirical data indicate that internal features (*i.e.,* eyes, nose, and mouth) and external features (*i.e.,* outline of the head) are differentially important for familiar and unfamiliar faces. Specifically, while familiar faces are better recognized from their internal features than from their external features, unfamiliar faces are recognized equally well from both types of features (Ellis, Shepherd, & Davies, 1979). The differential importance of internal and external features for recognizing familiar faces can be modeled by assigning a different weight to the pixels corresponding to internal and external features. A way to do that is to assume that the more informative a pixel is, the more weight it should be given in the input representation. Empirical data suggest that the more expressive and hence the more variable a feature is, the more informative it will be (Ellis, 1981). A good way of testing this hypothesis in the framework of a generalized auto-associator would be to weight the pixels as a function of their variability in the set of faces. In this case, the weight matrix would be the diagonal matrix of pixel variances across the set. Similarly, the fact that some individuals are more "distinctive" or more recognizable than others (Light *et al.*, 1979) can be modeled by using diagonal matrices in which the elements correspond, for example, to the distance between the faces and the average face. These two examples show that, in addition to being useful for modeling pattern recognition or human observer behaviors, the generalized auto-associator can be used as a practical tool for testing specific hypotheses.

Another interesting reason for generalizing the linear auto-associator comes from the fact that using a linear auto-associator as a content addressable memory is equivalent to creating a cross-product matrix of the input patterns (*i.e.,* the weight matrix) and computing its eigen-decomposition (Abdi, 1987, 1994a; Anderson, Silverstein, Ritz, & Jones, 1977; Baldi & Hornik, 1989; Boulard & Kamp, 1988; Knapp & Anderson, 1984; Kohonen, 1977; Krogh & Hertz, 1990; Linsker, 1988, 1989; Oja, 1982, 1989; Rubner & Tavan, 1989). This amounts to performing principal components analysis, also known in the engineering literature as Karhunen-Loève decomposition, on the cross-product matrix.

One advantage of this type of analysis is that it makes clear that classical auto-associators implement least-squares approximations. The interest of the generalized auto-associator is that it implements a generalized least-squares approximation or a least-squares approximation under (linear) constraints. In other words, while the classical auto-associator relies on the notion of euclidean distance, the generalized auto-associator implements the family of generalized Euclidean distances. Some well-known examples of these are the Mahalanobis distance (used in discriminant analysis) and the chi-square distance (used in correspondence analysis).

This concept of distance is important in understanding a variety of current models of psychological processes. For example, the notions of distance and similarity between stimuli or between percepts constitute a key feature of most recent models of categorization (cf. Ashby, 1992; Nosofsky, 1992). For example, Nosofsky (1992), in his generalized context model (GCM), represents stimuli with a parameter standing for the strength of a stimulus and with features weighted by an attentional parameter. Nosofsky's model can be seen as equivalent to representing the strength of a stimulus by a diagonal matrix applied to the stimulus and to representing the attentional weights by the diagonal terms of a matrix applied to the units of the memory. Categorization can then be considered to be a function of the generalized distance to the centers of the categories.

To conclude, a second useful property of the generalized model is that it provides a unified framework within which frequently encountered psychological models of categorization (such as those cited) can be compared systematically to their neural-network-implemented cousins. Comparison among these models is especially complicated, though no less important, when the network models involve iteratively applied learning algorithms such as the Widrow-Hoff learning rule (cf. Duda & Hart, 1973), also known as the "Delta" rule (McClelland & Rumelhart, 1986). These learning rules are very commonly applied in both the computational and psychological modeling literatures.

The purpose of the present paper is to show that the Widrow-Hoff learning rule can be adapted to the generalized auto-associator previously shown to implement a generalized Euclidean distance metric (Abdi, 1988; Abdi *et al.*, 1996). This exercise involves the notions of generalized eigenvectors and eigenvalues. We show that convergence, and the choice of the learning constant to achieve it, depend only upon the (generalized) eigenvalues of the connection matrix. With this proof in hand, several distance metrics common in the categorization literature and in the neural network literature can be compared and analyzed within a unified framework.

This paper is organized as follows. First, we define the notation used in the model presentation and proof. Second, we review the main features of the generalized auto-associator. Third, we present the generalized Widrow-Hoff learning rule in the context of the generalized model. Finally, we prove the convergence of the Widrow-Hoff rule in terms of the generalized eigenvectors of the matrix of synaptic connections. In order to do that, we begin by showing that Widrow-Hoff learning affects only the eigenvalues of the weight matrix. Then, we provide the specific expression for the eigenvalues of the weight matrix at a given time. We include also an Appendix reviewing the computation of the generalized eigenvectors.

2. Notation

In what follows, boldface lowercase letters denote (column) vectors (*e.g.,* $\mathbf{x}$), and boldface uppercase letters denote matrices (*e.g.,* $\mathbf{W}$). The superscript T indicates the transposition operation (*e.g.,* $\mathbf{x}^T$ is a row vector). The square matrix $\mathbf{I}$ is the identity matrix, and $\mathbf{0}$ is a null matrix (*i.e.,* filled with zeros). The operator $\mathrm{diag}(\mathbf{x})$ creates a square matrix with values of $\mathbf{x}$ on the diagonal.

The set of K stimuli to be stored in an auto-associative memory made of I units or cells is represented by an $I \times K$ matrix $\mathbf{X} = [x_{i,k}]$, with $x_{i,k}$ being a real number denoting the level of activation of the ith unit for the kth pattern ($k = \{1, \ldots, K\}$, $\quad i = \{1, \ldots, I\}$). The total of the rows, columns, and grand total are denoted, respectively, as:

$$x_{i+} = \sum_k x_{i,k} \qquad x_{+k} = \sum_i x_{i,k} \qquad x_{++} = \sum_{i,k} x_{i,k} \ .$$

Each unit is connected to all the other units in the memory. The intensity (or weight) of the connections is given by the $I \times I$ matrix $\mathbf{W}$.

The set of constraints imposed on the units is represented by a positive definite matrix of order $I \times I$ noted $\mathbf{B}$. For example, if the cells of the memory are supposed to fire with a rate inversely proportional to their activation across the entire set of stimuli, $\mathbf{B}$ will be a diagonal matrix with $b_{i,i} = x_{i+}^{-1}$ (in this case, we implicitly suppose that $x_{i+} \neq 0, \forall i$). When the matrix $\mathbf{B}$ is equal to the identity matrix, all the $\mathbf{B}$-generalized notions defined below reduce to their standard equivalent.

The *generalized norm (or* $\mathbf{B}$*-norm)* of vector $\mathbf{x}_k$ is defined as

$$\|\mathbf{x}_k\|_\mathbf{B} = \sqrt{\mathbf{x}_k^T \mathbf{B} \mathbf{x}_k} \ . \tag{1}$$

Vectors $\mathbf{x}_k$ and $\mathbf{x}_{k'}$ will be called $\mathbf{B}$-*orthogonal* if

$$\mathbf{x}_k^T \mathbf{B} \mathbf{x}_{k'} = 0 \ . \tag{2}$$

The generalized cosine (or $\mathbf{B}$-cosine) is defined as

$$\cos_\mathbf{B}(\mathbf{x}_k, \mathbf{x}_{k'}) = \frac{\mathbf{x}_k^T \mathbf{B} \mathbf{x}_{k'}}{\|\mathbf{x}_k\|_\mathbf{B} \ \|\mathbf{x}_{k'}\|_\mathbf{B}} \ . \tag{3}$$

In general, one can assume for simplicity that the stimuli are $\mathbf{B}$-normalized (*i.e.,* $\mathbf{x}_k^T \mathbf{B} \mathbf{x}_k = 1$).

The (squared) generalized Euclidean distance between two stimuli, k and k', is defined as

$$d_\mathbf{B}^2(k, k') = (\mathbf{x}_k - \mathbf{x}_{k'})^T \mathbf{B} (\mathbf{x}_k - \mathbf{x}_{k'}) \ . \tag{4}$$

For example, if the $\mathbf{B}$ matrix is defined as a diagonal matrix with $b_{i,i} = x_{i+}^{-1} \times x_{++}$, then the distance implemnted is the chi-square" distance (Benzécri, 1977). If $\mathbf{B}$ is equal to the inverse of the stimulus covariance matrix, then $d_\mathbf{B}^2(k, k')$ is the Mahalanobis distance used in discriminant analysis and in several recent models of categorization (*e.g.,* Ashby, 1992; Nosofsky, 1992).

The set of constraints imposed on the stimuli is represented by a positive definite matrix of order $K \times K$ noted $\mathbf{M}$. For example, if the stimuli should be given an importance proportional to the general intensity of the stimulation, $\mathbf{M}$ will be a diagonal matrix with $m_{k,k} = x_{+k}$. It is worth noting, however, that choices other than a diagonal matrix are possible for $\mathbf{B}$ and $\mathbf{M}$. For example, temporal interference between stimuli can be modeled by defining $\mathbf{M}$ as a band diagonal matrix, with values decreasing proportionally to their distance from the diagonal. This would implement an interdependence between stimuli as an inverse function of the relative position of the stimuli in the learning sequence.

3. STORAGE AND RECALL FROM THE MEMORY

In this section the features of the generalized model are briefly reviewed. When the matrices $\mathbf{B}$ and $\mathbf{M}$ are replaced by identity matrices, the generalized auto-associator behaves like the classical linear auto-associator (cf. Anderson, 1972; Kohonen, 1972).

The stimuli are stored in the memory by setting the weights of the connections between cells. These values are stored in a matrix $\mathbf{W}$ using Hebbian learning:

$$\mathbf{W} = \mathbf{X}\mathbf{M}\mathbf{X}^T \ . \tag{5}$$

In the particular case for which $\mathbf{M}$ is diagonal Eq. 5 becomes

$$\mathbf{W} = \mathbf{X}\mathbf{M}\mathbf{X}^T = \sum_k m_k \mathbf{x}_k \mathbf{x}_k^T \ . \tag{6}$$

Recall of a pattern is achieved by presenting the pattern to the memory after multiplication by the matrix $\mathbf{B}$. Specifically, recall of pattern $\mathbf{x}_\ell$ is achieved as

$$\widehat{\mathbf{x}}_\ell = \mathbf{W}\mathbf{B}\mathbf{x}_\ell = \mathbf{X}\mathbf{M}\mathbf{X}^T \mathbf{B}\mathbf{x}_\ell \ . \tag{7}$$

Again the specific case of $\mathbf{M}$ diagonal leads to some simplifications. It is clear, in this case, that recall from the memory will involve some cross talk between the different stimuli stored:

$$\begin{aligned} \widehat{\mathbf{x}}_\ell &= \mathbf{W}\mathbf{B}\mathbf{x}_\ell \\ &= \sum_k m_k \mathbf{x}_k \mathbf{x}_k^T \mathbf{B}\mathbf{x}_\ell \\ &= m_\ell \mathbf{x}_\ell \mathbf{x}_\ell^T \mathbf{B}\mathbf{x}_\ell + \sum_{\ell \neq k} m_k \mathbf{x}_k \mathbf{x}_k^T \mathbf{B}\mathbf{x}_\ell \ . \end{aligned} \tag{8}$$

Since stimuli are assumed to be $\mathbf{B}$-normalized, Eq. 8 reduces to

$$\widehat{\mathbf{x}}_\ell = m_\ell \mathbf{x}_\ell \mathbf{x}_\ell^T \mathbf{B}\mathbf{x}_\ell + \sum_{\ell \neq k} m_k \cos_\mathbf{B}(\mathbf{x}_k, \mathbf{x}_\ell)\mathbf{x}_k \ . \tag{9}$$

The term $\mathbf{x}_\ell^T \mathbf{B}\mathbf{x}_\ell$ is a scalar, which is denoted by γ_ℓ; then Eq. 9 can be rewritten as

$$\widehat{\mathbf{x}}_\ell = m_\ell \gamma_\ell \mathbf{x}_\ell + \sum_{\ell \neq k} m_k \cos_\mathbf{B}(\mathbf{x}_k, \mathbf{x}_\ell)\mathbf{x}_k \ , \tag{10}$$

which shows that the stimulus recalled is composed of two terms, the first one being the original stimulus and the second one reflecting the interference between the original stimulus and the other stimuli in the set. In general, the cosine between $\mathbf{x}_\ell$ and $\widehat{\mathbf{x}}_\ell$ is used as a measure of quality of recall.

In the particular case of a set of **B**-orthogonal stimuli, and when **M** is diagonal, the previous equation reduces to

$$\begin{aligned}\widehat{\mathbf{x}}_\ell &= m_\ell\gamma_\ell\mathbf{x}_\ell + \sum_{\ell\neq k} m_k \cos_{\mathbf{B}}(\mathbf{x}_k, \mathbf{x}_\ell)\mathbf{x}_k \\ &= m_\ell\gamma_\ell\mathbf{x}_\ell \end{aligned} \tag{11}$$

which shows that in this case the stimulus recalled is proportional to the stimulus stored in the matrix.

In the general case where the stimuli stored in the memory do not form a **B**-orthogonal set, the memory will not reconstruct them perfectly. However, some patterns will be perfectly reconstructed. If the $\mathbf{x}_k$'s are **B**-normalized, the maximal values for the cosine between the inputs and the responses are obtained when the response of the matrix is equal or proportional to the input, namely when

$$\begin{aligned}\widehat{\mathbf{x}}_\ell &= \mathbf{W}\mathbf{B}\mathbf{x}_\ell \\ &= \lambda_\ell\mathbf{x}_\ell \end{aligned} \tag{12}$$

(as $\mathbf{x}_\ell^T\mathbf{B}\mathbf{x}_\ell = 1$). This is equivalent to a (generalized) eigen-decomposition problem (cf. Wilkinson, 1965), where the problem is to find the generalized eigenvectors $\mathbf{u}_k$ and generalized eigenvalues λ_k such that

$$\mathbf{W}\mathbf{u}_k = \lambda_k\mathbf{u}_k \quad \text{with } \mathbf{u}_k^T\mathbf{B}\mathbf{u}_k = 1 \ . \tag{13}$$

The equivalence between these two formulations, and the computation of the generalized eigenvectors are detailed in the Appendix.

As previously noted, Anderson *et al.* (1977) have pointed out that for the classical auto-associator, finding the eigenvectors of **W** is equivalent to performing a principal components analysis on the set of units because **W** is a cross-product matrix. For the generalized version of this model, Abdi (1988) has noted that with an appropriate choice of the matrices **M** and **B**, finding the (generalized) eigenvectors of **W** is equivalent to implementing the general linear model (see also Greenacre, 1984, pp. 347–349). For example, when

$$\mathbf{B} = \operatorname{diag}(\mathbf{b}) = [b_{i,i}] = \frac{1}{x_{i+}/x_{++}} \tag{14}$$

and

$$\mathbf{M} = \operatorname{diag}(\mathbf{m}) = [m_{k,k}] = \frac{x_{+k}}{x_{++}} \tag{15}$$

the generalized Euclidean distance used by the model is the Chi-square distance (cf. Benzécri, 1977; Greenacre, 1984; Weller & Romney, 1990), and the computation of the generalized eigenvectors of **W** is equivalent to the statistical technique of "correspondence analysis" (Abdi, 1988; Benzécri, 1977; Greenacre, 1984; Weller & Romney, 1990). An example of this approach, together with applications to the recognition and categorization of faces is given in Abdi (1988) and Abdi *et al.* (1996).

4. Generalized Widrow-Hoff Learning Rule

Most applications of the linear auto-associator use the Widrow-Hoff learning rule to improve the storage capacity of the memory. This is equivalent to using a gradient descent method to adjust the weights of the connections so as to reduce the squared error between stimuli and their reconstructions.

In this section, the Widrow-Hoff learning rule is adapted to the generalized linear auto-associator.

Learning is incremental (*i.e.,* occurs in discrete steps) and proceeds by comparison of the response of the system with the target response. Specifically, the weight matrix at time $t+1$ is obtained as

$$\begin{aligned}\mathbf{W}_{[t+1]} &= \mathbf{W}_{[t]} + \eta(\mathbf{X} - \mathbf{W}_{[t]}\mathbf{B}\mathbf{X})\mathbf{M}\mathbf{X}^T \\ &= \mathbf{W}_{[t]} + \Delta_{[t+1]} \ , \end{aligned} \tag{16}$$

with η being a (small) positive real number called the *learning constant.*[1] When **M** and **B** are identity

[1] The matrix notation used here seems to imply that for learning to occur, all the stimuli need to be present at the same time, which makes it psychologically unrealistic. It is, however, possible to rewrite Eq. 16 so that only one stimulus is present at a given time. This implies that matrix **M** is diagonal. Denote by m_k its kth diagonal element. If we assume that the error signal is computed for each stimulus and applied at the end of the learning period, then the term $\Delta_{[t+1]}$ of Eq. 16 can be rewritten as $\Delta_{[t+1]} = \eta\sum_k(\mathbf{x}_k - \mathbf{W}_{[t]}\mathbf{B}\mathbf{x}_k)m_k\mathbf{x}_k^T$, which shows that the error signal can be computed with only one stimulus present at a given time. Using the complete matrix **X** for computing the error signal is equivalent to

matrices, Eq. 16 describes the standard Widrow-Hoff learning rule (Anderson *et al.*, 1977).

It is shown in the next section that (when η is appropriately chosen), the learning rule will converge toward

$$\mathbf{W}_{[\infty]} = \mathbf{U}\mathbf{U}^T , \tag{17}$$

where $\mathbf{U}$ is the matrix of the generalized eigenvectors of $\mathbf{W}$ (*i.e.,* $\mathbf{U}^T\mathbf{B}\mathbf{U} = \mathbf{I}$).

It should be noted, in passing, that in the particular case where $\mathbf{W}$ is full rank, $\mathbf{W}_{[\infty]}$ will converge toward the "trivial" solution $\mathbf{W}_{[\infty]} = \mathbf{B}^{-1}$. In the classical case, for $\mathbf{W}$ full rank, the trivial convergence gives $\mathbf{W}_{[\infty]} = \mathbf{I}$ (Anderson *et al.*, 1977; Kohonen, 1977).

Moreover, it is shown that $\mathbf{W}_{[t]}$ can be expressed as a function of only η, t, and the eigenvectors and eigenvalues of $\mathbf{W}$. Specifically

$$\mathbf{W}_{[t]} = \mathbf{U}\left[\mathbf{I} - (\mathbf{I} - \eta\mathbf{\Lambda})^t\right]\mathbf{U}^T . \tag{18}$$

As a consequence, for convergence to be reached, the expression

$$\left[\mathbf{I} - (\mathbf{I} - \eta\mathbf{\Lambda})^t\right]$$

in Eq. 18 should be equal to the identity matrix (*i.e.,* Eq. 18 is then equivalent to Eq. 17). In other words, convergence will be reached if and only if

$$\lim_{t\to\infty} (\mathbf{I} - \eta\mathbf{\Lambda})^t = \mathbf{0} , \tag{19}$$

equivalently if and only if

$$\lim_{t\to\infty} (1 - \eta\lambda_i)^t = 0 \quad \forall i , \tag{20}$$

which is equivalent to having

$$0 < \eta < 2\lambda_i^{-1} \quad \forall i . \tag{21}$$

If $\lambda_{\max}$ denotes the largest eigenvalue, then

$$\lambda_{\max}^{-1} \leq \lambda_i^{-1} \quad \forall i . \tag{22}$$

Since $\mathbf{W} = \mathbf{X}\mathbf{M}\mathbf{X}^T$ (Eq. 5) and because $\mathbf{M}$ is a positive definite matrix, $\mathbf{W}$ is positive semi-definite, and hence its eigenvalues are always positive or zero. Therefore, convergence will be reached when

$$0 < \eta < 2\lambda_{\max}^{-1} . \tag{23}$$

This generalizes the classical condition for convergence in the standard case (cf. Abdi, 1994b; Widrow & Stearns, 1985).

---

the "batch-mode" learning of back-propagation networks (cf. Haykin, 1994).

5. Lemmas

Before going into the details of the proof, we provide the following two lemmas:

Let $\mathbf{W}$ be an $I \times I$ matrix and

$$\mathbf{W} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \qquad \text{with: } \mathbf{U}^T\mathbf{B}\mathbf{U} = \mathbf{I} \tag{24}$$

be the generalized eigenvalue decomposition of $\mathbf{W}$ (with $\mathbf{\Lambda}$ including the zero-eigenvalues if any), then the following properties can be derived:

Lemma 1: *If* $\mathbf{\Phi}$ *is an* $I \times I$ *diagonal matrix, then*

$$\mathbf{U}\mathbf{\Phi}\mathbf{U}^T + \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \mathbf{U}(\mathbf{\Phi} + \mathbf{\Lambda})\mathbf{U}^T . \tag{25}$$

Proof: It suffices to factor to obtain the result.

Lemma 2: *If* $\mathbf{\Phi}$ *is a diagonal matrix then*

$$\mathbf{\Phi}^0 + \mathbf{\Phi}^1 + \cdots + \mathbf{\Phi}^t = (\mathbf{I} - \mathbf{\Phi}^{t+1})(\mathbf{I} - \mathbf{\Phi})^{-1} .$$

Proof: The result follows from the power series expression

$$1 + x + x^2 + \cdots + x^t = \frac{(1 - x^{t+1})}{(1 - x)} \tag{26}$$

applied to the diagonal elements of $\mathbf{\Phi}$.

6. Proof

The following proof consists of two parts: in the first part, we prove that the Widrow-Hoff learning for the generalized auto-associator affects only the non-zero eigenvalues of $\mathbf{W}$. In the second part, we derive the specific formula for the eigenvalues at time t. Specifically, we show (using the notation defined in Eq. 24) that the matrix $\mathbf{W}$ at time t can be expressed as

$$\mathbf{W}_{[t]} = \mathbf{U}\left[\mathbf{I} - (\mathbf{I} - \eta\mathbf{\Lambda})^t\right]\mathbf{U}^T . \tag{27}$$

6.1. Part 1: Learning and the Eigenvalues of the Auto-associative Matrix.

Denoting by $\mathbf{\Phi}_{[t]}$ the eigenvalue matrix of $\mathbf{W}_{[t]}$, we prove that $\mathbf{W}_{[t+1]}$ can be expressed as a function of $\mathbf{\Phi}_{[t]}$ and the eigenvectors of $\mathbf{W}$, specifically:

$$\mathbf{W}_{[t+1]} = \mathbf{U}\left[\mathbf{\Phi}_{[t]} + \left(\mathbf{I} - \mathbf{\Phi}_{[t]}\right)\eta\mathbf{\Lambda}\right]\mathbf{U}^T \tag{28}$$

(with notation from Eq. 24). This can be shown by induction.

Proof: For convenience, assume that the weight matrix is initialized with zero values:

$$\mathbf{W}_{[0]} = \mathbf{0} . \tag{29}$$

The first iteration of the generalized Widrow-Hoff learning rule (cf. Eq. 16) corresponds (*i.e.,* is proportional) to Hebbian learning as defined by Eq. 5:

$$\begin{aligned}\mathbf{W}_{[1]} &= \mathbf{W}_{[0]} + \eta\left(\mathbf{X} - \mathbf{W}_{[0]}\mathbf{B}\mathbf{X}\right)\mathbf{M}\mathbf{X}^T \\ &= \eta\mathbf{X}\mathbf{M}\mathbf{X}^T \\ &= \eta\mathbf{W} \ . \end{aligned} \tag{30}$$

Equation 28 holds for $t = 1$ as:

$$\begin{aligned}\mathbf{W}_{[1]} &= \eta\mathbf{W} \\ &= \eta\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T \\ &= \mathbf{U}\eta\boldsymbol{\Lambda}\mathbf{U}^T \\ &= \mathbf{U}\left[\mathbf{0} + (\mathbf{I} - \mathbf{0})\,\eta\boldsymbol{\Lambda}\right]\mathbf{U}^T \end{aligned} \tag{31}$$

(because the eigenvalues of $\mathbf{W}_{[0]} = \mathbf{0}$ are all zeros, and from Eqs. 24 and 30).

Assuming that the expression is true at time t, then the learning iteration at step $t+1$ is

$$\begin{aligned}\mathbf{W}_{[t+1]} &= \mathbf{W}_{[t]} + \eta\left(\mathbf{X} - \mathbf{W}_{[t]}\mathbf{B}\mathbf{X}\right)\mathbf{M}\mathbf{X}^T \\ &= \mathbf{U}\boldsymbol{\Phi}_{[t]}\mathbf{U}^T + \eta\left(\mathbf{X} - \mathbf{U}\boldsymbol{\Phi}_{[t]}\mathbf{U}^T\mathbf{B}\mathbf{X}\right)\mathbf{M}\mathbf{X}^T \\ &= \mathbf{U}\boldsymbol{\Phi}_{[t]}\mathbf{U}^T + \eta\left(\mathbf{I} - \mathbf{U}\boldsymbol{\Phi}_{[t]}\mathbf{U}^T\mathbf{B}\right)\mathbf{X}\mathbf{M}\mathbf{X}^T \\ &= \mathbf{U}\boldsymbol{\Phi}_{[t]}\mathbf{U}^T + \eta\left(\mathbf{I} - \mathbf{U}\boldsymbol{\Phi}_{[t]}\mathbf{U}^T\mathbf{B}\right)\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T \\ &= \mathbf{U}\boldsymbol{\Phi}_{[t]}\mathbf{U}^T + \eta\left(\mathbf{U} - \mathbf{U}\boldsymbol{\Phi}_{[t]}\mathbf{U}^T\mathbf{B}\mathbf{U}\right)\boldsymbol{\Lambda}\mathbf{U}^T \\ &= \mathbf{U}\boldsymbol{\Phi}_{[t]}\mathbf{U}^T + \eta\left(\mathbf{U} - \mathbf{U}\boldsymbol{\Phi}_{[t]}\right)\boldsymbol{\Lambda}\mathbf{U}^T \\ &= \mathbf{U}\boldsymbol{\Phi}_{[t]}\mathbf{U}^T + \eta\mathbf{U}\left(\mathbf{I} - \boldsymbol{\Phi}_{[t]}\right)\boldsymbol{\Lambda}\mathbf{U}^T \\ &= \mathbf{U}\boldsymbol{\Phi}_{[t]}\mathbf{U}^T + \mathbf{U}\left(\mathbf{I} - \boldsymbol{\Phi}_{[t]}\right)\eta\boldsymbol{\Lambda}\mathbf{U}^T \\ &\qquad\qquad \text{(because } \eta \text{ is a scalar)} \\ &= \mathbf{U}\left[\boldsymbol{\Phi}_{[t]} + \left(\mathbf{I} - \boldsymbol{\Phi}_{[t]}\right)\eta\boldsymbol{\Lambda}\right]\mathbf{U}^T \\ &\qquad\qquad \text{(from Lemma 1)} \ , \end{aligned} \tag{32}$$

which shows that learning affects only the eigenvalues of $\mathbf{W}$ and completes the first part of the proof.

6.2. **Part 2: Specific Formula for the Eigenvalues of the Weight Matrix.** We will now derive the specific expression for the eigenvalues of $\mathbf{W}_{[t]}$. Specifically we will show that the eigenvalues of $\mathbf{W}_{[t]}$ can be expressed as:

$$\boldsymbol{\Phi}_{[t]} = \mathbf{I} - (\mathbf{I} - \eta\boldsymbol{\Lambda})^t \ . \tag{33}$$

The process starts with (cf. Eq. 29)

$$\boldsymbol{\Phi}_{[0]} = \mathbf{0} \ . \tag{34}$$

At step 1

$$\begin{aligned}\boldsymbol{\Phi}_{[1]} &= \eta\boldsymbol{\Lambda} \\ &= \eta\boldsymbol{\Lambda}(\mathbf{I} - \eta\boldsymbol{\Lambda})^0 \ . \end{aligned} \tag{35}$$

(from Eqs. 24 and 30).

At step 2

$$\begin{aligned}\boldsymbol{\Phi}_{[2]} &= \boldsymbol{\Phi}_{[1]} + \left(\mathbf{I} - \boldsymbol{\Phi}_{[1]}\right)\eta\boldsymbol{\Lambda} \qquad \text{(from Eq. 32)} \\ &= \eta\boldsymbol{\Lambda}(\mathbf{I} - \eta\boldsymbol{\Lambda})^0 + \left[\mathbf{I} - \eta\boldsymbol{\Lambda}(\mathbf{I} - \eta\boldsymbol{\Lambda})^0\right]\eta\boldsymbol{\Lambda} \\ &= \eta\boldsymbol{\Lambda}(\mathbf{I} - \eta\boldsymbol{\Lambda})^0 + \eta\boldsymbol{\Lambda}(\mathbf{I} - \eta\boldsymbol{\Lambda}) \ . \end{aligned} \tag{36}$$

Using induction, the matrix of eigenvalues at time t can be expressed as

$$\boldsymbol{\Phi}_{[t]} = \eta\boldsymbol{\Lambda}\sum_{i=0}^{t-1}(\mathbf{I} - \eta\boldsymbol{\Lambda})^i \ . \tag{37}$$

PROOF: The equation is true for $t = 1$ (cf. Eq. 35), suppose it holds for $t-1$, then

$$\begin{aligned}\boldsymbol{\Phi}_{[t]} &= \boldsymbol{\Phi}_{[t-1]} + \left(\mathbf{I} - \boldsymbol{\Phi}_{[t-1]}\right)\eta\boldsymbol{\Lambda} \\ &= \eta\boldsymbol{\Lambda}\sum_{i=0}^{t-2}(\mathbf{I} - \eta\boldsymbol{\Lambda})^i + \left[\mathbf{I} - \eta\boldsymbol{\Lambda}\sum_{i=0}^{t-2}(\mathbf{I} - \eta\boldsymbol{\Lambda})^i\right]\eta\boldsymbol{\Lambda} \\ &= \eta\boldsymbol{\Lambda}\sum_{i=0}^{t-2}(\mathbf{I} - \eta\boldsymbol{\Lambda})^i + \eta\boldsymbol{\Lambda} - \eta^2\boldsymbol{\Lambda}^2\sum_{i=0}^{t-2}(\mathbf{I} - \eta\boldsymbol{\Lambda})^i \\ &= \eta\boldsymbol{\Lambda}\sum_{i=0}^{t-2}(\mathbf{I} - \eta\boldsymbol{\Lambda})^i(\mathbf{I} - \eta\boldsymbol{\Lambda}) + \eta\boldsymbol{\Lambda} \\ &= \eta\boldsymbol{\Lambda}\sum_{i=0}^{t-1}(\mathbf{I} - \eta\boldsymbol{\Lambda})^i \ , \end{aligned} \tag{38}$$

thus proving Eq. 37.

Using Lemma 2, Eq. 37 can also be expressed as

$$\begin{aligned}\mathbf{\Phi}_{[t]} &= \eta\mathbf{\Lambda}\sum_{i=0}^{t-1}(\mathbf{I}-\eta\mathbf{\Lambda})^i \\ &= \eta\mathbf{\Lambda}\left[\mathbf{I}-(\mathbf{I}-\eta\mathbf{\Lambda})^t\right]\left[\mathbf{I}-(\mathbf{I}-\eta\mathbf{\Lambda})\ \right]^{-1} \\ &= \mathbf{I}-(\mathbf{I}-\eta\mathbf{\Lambda})^t \end{aligned} \tag{39}$$

which completes the proof.

7. Conclusion

The proof presented in this note completes and analyzes the inner workings of the generalized linear auto-associator originally presented by Abdi (1988). This model allows for the *a priori* imposition of constraints operating at the level of the input code and at the level of individual stimuli. We showed that the Widrow-Hoff error correction learning rule used by most neural-network practical applications can be adapted to the generalized auto-associator. Hence, in addition to providing a theoretical link between the neural network "learning perspective" and the statistical perspective of the general linear model of multivariate statistics, this proof will allow for a wider and more practical use of the generalized model.

Appendix: Computation of the generalized eigenvectors

The generalized eigen-decomposition of a matrix can be computed from the standard eigen-decomposition. This can be made clearer after recalling the definition of the power of a matrix (or more generally of the function of a matrix, cf. Perlis, 1953, p. 166–168).

Preliminaries: power of a matrix. Suppose that $\mathbf{B}$ is an $I \times I$ positive definite matrix with eigen-decomposition

$$\begin{aligned}\mathbf{B} &= \mathbf{P\Gamma P}^T \\ &= \mathbf{P}\,\mathrm{diag}(\boldsymbol{\gamma})\mathbf{P}^T \\ &= \mathbf{P}\,\mathrm{diag}(\gamma_1,\ldots,\gamma_i,\ldots,\gamma_I)\mathbf{P}^T\ ,\end{aligned} \tag{40}$$

(with γ_i being the ith eigenvalue of $\mathbf{B}$); then, the xth power of $\mathbf{B}$ is defined as

$$\mathbf{B}^x = \mathbf{P\Gamma}^x\mathbf{P}^T = \mathbf{P}\,\mathrm{diag}(\gamma_1^x,\ldots,\gamma_i^x,\ldots,\gamma_I^x)\mathbf{P}^T\ . \tag{41}$$

Generalized eigenvectors. The generalized eigen-decomposition of $\mathbf{W}$ begins by defining a new matrix $\widetilde{\mathbf{W}}$ and computing its standard eigen-decomposition

$$\widetilde{\mathbf{W}} = \mathbf{B}^{\frac{1}{2}}\mathbf{W}\mathbf{B}^{\frac{1}{2}} = \widetilde{\mathbf{U}}\mathbf{\Lambda}\widetilde{\mathbf{U}}^T \qquad \text{with: } \widetilde{\mathbf{U}}^T\widetilde{\mathbf{U}} = \mathbf{I} \tag{42}$$

(with $\mathbf{B}^{\frac{1}{2}}$ obtained from Eq. 41). Then the generalized eigenvectors of $\mathbf{W}$ are

$$\mathbf{U} = \mathbf{B}^{-\frac{1}{2}}\widetilde{\mathbf{U}}\ . \tag{43}$$

Clearly, $\mathbf{U}$ verifies the definition of the generalized eigenvectors as

$$\mathbf{U}^T\mathbf{B}\mathbf{U} = \widetilde{\mathbf{U}}^T\mathbf{B}^{-\frac{1}{2}}\mathbf{B}\mathbf{B}^{-\frac{1}{2}}\widetilde{\mathbf{U}} = \mathbf{I} \tag{44}$$

and

$$\begin{aligned}\mathbf{U\Lambda U}^T &= \mathbf{B}^{-\frac{1}{2}}\widetilde{\mathbf{U}}\mathbf{\Lambda}\widetilde{\mathbf{U}}^T\mathbf{B}^{-\frac{1}{2}} \\ &= \mathbf{B}^{-\frac{1}{2}}\widetilde{\mathbf{W}}\mathbf{B}^{-\frac{1}{2}} \\ &= \mathbf{B}^{-\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\mathbf{W}\mathbf{B}^{\frac{1}{2}}\mathbf{B}^{-\frac{1}{2}} \\ &= \mathbf{W}\ .\end{aligned} \tag{45}$$

The generalized eigenvectors of $\mathbf{W}$ are also eigenvectors of $\mathbf{WB}$. To show this, it suffices to note that the eigen-equation of $\mathbf{W}$ can be written as

$$\mathbf{U\Lambda} = \mathbf{WU}\ . \tag{46}$$

transposing, right-multiplying by $\mathbf{BU}$, and simplifying gives

$$\begin{aligned}\mathbf{\Lambda U}^T &= \mathbf{U}^T\mathbf{W} \\ \mathbf{\Lambda U}^T\mathbf{BU} &= \mathbf{U}^T\mathbf{WBU} \\ \mathbf{\Lambda} &= \mathbf{U}^T\mathbf{WBU}\end{aligned} \tag{47}$$

which shows that $\mathbf{U}$ diagonalizes $\mathbf{WB}$.

References

[1] Abdi, H. (1987). Do we really need a contingency model for concept formation? *British Journal of Psychology*, **78**, 113–125.

[2] Abdi, H. (1988). A generalized approach for connectionist auto-associative memories: Interpretation, implications and illustration for face processing. In J. Demongeot (Ed.), *Artificial intelligence and cognitive sciences.* Manchester: Manchester University Press.

[3] Abdi, H. (1994a). *Les réseaux de neurones.* Grenoble: Presses Universitaires de Grenoble.

[4] Abdi, H. (1994b). A neural network primer. *Journal of Biological Systems*, **2**, 247–282.

[5] Abdi, H., Valentin, D., & O'Toole, A. J. (1996). A generalized auto-associator model for face processing and sex categorization: From principal component analysis to multivariate analysis. In D. Levine (Ed.), *Optimality in biological and artificial networks.* Hillsdale: Erlbaum.

[6] Anderson, J. A. (1972). A simple neural network generating an interactive memory. *Mathematical Biosciences*, **14**, 197–220.

[7] Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model, *Psychological Review*, **84**, 413–451.

[8] Ashby, F.G. (1992). Multidimensional models of categorization. In F.G. Ashby (Ed.), *Multidimensional models of perception and cognition.* Hillsdale: Erlbaum.

[9] Baldi, P., & Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima, *Neural Networks*, **2**, 53–58.

[10] Benzécri, J. P. (1977). *L'analyse des données.* Paris: Dunod.

[11] Boulard, H., & Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition, *Biological Cybernetics*, **59**, 291–294.

[12] Duda, R. O., & Hart P. E. (1973). *Pattern classification and scene analysis.* New York: Wiley.

[13] Ellis, H.D. (1981). Theoretical aspects of face recognition. In G.M. Davies, H.D. Ellis & J.W. Shepherd (Eds.) *Perceiving and remembering faces.* London: Academic Press.

[14] Ellis, H.D., Sheperd, J.W., Davis, G.M. (1979). Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition. *Perception*, **8**, 431–439.

[15] Greenacre, M. J. (1984). *Theory and applications of correspondence analysis.* London: Academic Press.

[16] Haykin, S. (1994). *Neural networks.* New York: MacMillan.

[17] Knapp, A. G., & Anderson, J. A. (1984). Theory of categorization based on distributed memory storage, *Journal of Experimental Psychology: Learning Memory and Cognition*, **10**, 616–637.

[18] Kohonen, T. (1972). Correlation matrix memories. *I.E.E.E. Transactions on Computers*, **C-21**, 353–359.

[19] Kohonen, T. (1977). *Associative memory: A system theoretical approach.* Berlin: Springer Verlag.

[20] Krogh, A., & Hertz, J. A. (1990). Hebbian learning of principal components. In R. Eckmiller, G. Hartmann, & G. Hauske (Eds.), *Parallel processing in neural systems and computers.* Amsterdam: Elsevier.

[21] Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory*, **5**, 212–228.

[22] Linsker, R. (1988). Self organization in a perceptual network, *Computer*, **21**, 105–117.

[23] Linsker, R. (1989). *Designing a sensory processing system: What can be learned from components analysis?* IBM-Research report # 668916. (7 pages).

[24] McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel distributed processing.* Cambridge MA: MIT Press.

[25] Nosofsky, R. M. (1992). Exemplar-based approach to relating categorization, identification, and recognition. In F.G. Ashby (Ed.), *Multidimensional models of perception and cognition.* Hillsdale: Erlbaum.

[26] Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, **15**, 267–273.

[27] Oja, E. (1989). Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, **1**, 61–68.

[28] Perlis, S. (1953). *Theory of matrices.* Reading, MA: Addison-Wesley.

[29] Rubner, J. & Tavan, P. (1989). A self-organizing network for principal component analysis. *Europhysics Letters*, **10**, 693–698.

[30] Sheperd, J. W., Davies, G. M., & Ellis, H. D. (1981). Studies of cue saliency. In G. Davies, H. Ellis, & J. Shepherd (Eds.), *Perceiving and remembering faces.* London: Academic Press.

[31] Valentin, D., Abdi, H., O'Toole, A.J. (1994). Categorization and identification of human face images by neural networks: A review of the linear auto-associative and principal component approaches. *Journal of Biological Systems*, **2**, 413–429

[32] Weller S. C., & A. K. Romney (1990). *Metric scaling: Correspondence analysis.* Newsbury Park (CA): Sage.

[33] Widrow, B., & Stearns, S. D. (1985). *Adaptive signal processing.* Englewood Cliffs NJ: Prentice–Hall.

[34] Wilkinson, J. H. (1965). *The algebraic eigenvalue problem.* Oxford: Clarenton Press.