Biol. Cybernetics 19, 201—209 (1975)

On Optimal Nonlinear Associative Recall

T. Poggio

Max-Planck-Institut für biologische Kybernetik, Tübingen, FRG

Received: December 10, 1974

Abstract

The problem of determining the nonlinear function ("black-box") which optimally associates (on given criteria) two sets of data is considered. The data are given as discrete, finite column vectors, forming two matrices X ("input") and Y ("output") with the same numbers of columns and an arbitrary numbers of rows. An iteration method based on the concept of the generalized inverse of a matrix provides the polynomial mapping of degree k on X by which Y is retrieved in an optimal way in the least squares sense. The results can be applied to a wide class of problems since such polynomial mappings may approximate any continuous real function from the "input" space to the "output" space to any required degree of accuracy. Conditions under which the optimal estimate is linear are given. Linear transformations on the input key-vectors and analogies with the "whitening" approach are also discussed. Conditions of "stationarity" on the processes of which X and Y are assumed to represent a set of sample sequences can be easily introduced. The optimal linear estimate is given by a discrete counterpart of the Wiener-Hopf equation and, if the key-signals are noise-like, the holographic-like scheme of associative memory is obtained, as the optimal nonlinear estimator. The theory can be applied to the system identification problem. It is finally suggested that the results outlined here may be relevant to the construction of models of associative, distributed memory.

1. Introduction

The problem of determining the operator, which optimally associates (on given criteria) a set of items with another set of items, is of considerable interest in a variety of areas.

First, the problem meets the most natural definition of associative recall as given in terms of stimuli and responses. A black-box can be said to associate two signals if when presented with one as input, it "recalls" the other as output. In the past few years a number of models have dealt with specific networks of this kind; both linear and not-linear systems have been examined (van Heerden, 1963; Anderson, 1968, 1972; Gabor, 1969; Longuet-Higgins, 1968, 1970; Kohonen, 1972; Marr, 1969, 1970; Willshaw, 1971, 1972; Cooper, 1974). At the present time it seems as if such models may be relevant to the problem of how the brain stores and retrieves information.

Secondly, the problem is closely connected with the system identification problem (Balakrishnan, 1967); that is, the task of characterizing a system from known input-output data, of finding the underlying "laws". A general setting for this problem is estimation theory. I will consider here a discrete, "deterministic" case: n m-components column vectors x^j represent the key-signals with which a suitable mapping has to associate optimally n r-components column vectors y^j which are the signals to be retrieved. The purpose of the present paper is to show how to obtain the nonlinear mapping 0 which represents the optimal solution, in the least square sense, of the equation

$$Y = 0(X) \quad Y \in \mathcal{M}(m, n), X \in \mathcal{M}(r, n) \tag{1.1}$$

for given Y and X, Y and X being arbitrary (possibly rectangular) real valued matrices. The class of mappings 0 considered here is restricted (for each component of y^j) to the class of polynomial mappings of degree k in the vector space V (over the real field) to which the vectors x^j belong. The basic Eq. (1.1) can be therefore specialized to

$$Y = L_0 + L_1(X) + L_2(X, X) + \cdots + L_k(X, \ldots, X), \tag{1.2}$$

where Y and X are arbitrary real valued matrices (the set of column vectors y^j and x^j respectively) and L_k is a k-linear (symmetric) mapping (Dieudonné, 1969) $V \times V \times \cdots \times V \to W$. W is a vector space defined, as V, over the real field; the vectors y^j are elements of W. With $(L_k)_{i,\alpha_1 \ldots \alpha_k}$ defined as the k-way real matrix associated to the mapping L_k, Eq. (1.2) can be explicitly rewritten as

$$\begin{aligned} Y_{ij} = (L_0)_{ij} &+ \sum_{\alpha_1} (L_1)_{i,\alpha_1} X_{\alpha_1 j} \\ &+ \sum_{\alpha_1 \alpha_2} (L_2)_{i,\alpha_1 \alpha_2} X_{\alpha_1 j} X_{\alpha_2 j} + \cdots \\ &+ \sum_{\alpha_1 \ldots \alpha_k} (L_k)_{i,\alpha_1 \ldots \alpha_k} X_{\alpha_1 j} \ldots X_{\alpha_k j} . \end{aligned} \tag{1.3}$$

The restriction to real-valued quantities is introduced here for simplicity; for a more general approach see

Poggio (1975b). The paper deals only with estimators (the mappings 0 from X to Y) which are optimal on the mean-square criterion. A discussion of which other criteria yield the same estimators will be given elsewhere.

Not all "black-boxes" operating on discrete matrices of data can have a representation of the form (1.2). However, the polynomial representation (1.2) is fairly general as shown by the classical Weierstrass-Stone theorem (see Dieudonné, 1969). This provides a simple proof that the real valued polynomials in n-variables, defined on a compact subset E of $\mathbb{R}^n$, are dense in the family of real-valued continuous functions on E in the uniform norm topology. The theorem can be applied vector-wise and component-wise to Eq. (1.2), implying that the "black-boxes" considered here may include all real continuous m-components functions in r variables. This topic will be further discussed in a forthcoming paper (Poggio, 1975b) together with some extensions of (1.1) and (1.2).

When the optimum mapping 0 is restricted to the linear term L_1 of (1.2) the solution to this problem is in terms of a generalized inverse for matrices, as given by Penrose (1956) (see Appendix). The significance of this result in the context of associative models of memory has been pointed out by Kohonen and Ruohonen (1973). In this paper the solution to the general nonlinear problem is in the form of an iterative method capable of approximating the optimal sequence $\{L_0, \ldots L_k\}$ to an arbitrary degree of accuracy. The plan of the paper is as follows. Section 2 gives the optimal nonlinear correction of degree k as the best approximate k-linear mapping which solves the equation $E = L_k(X, \ldots, X)$. In Section 3 the conditions under which the optimal estimate of the form (1.2) is linear are discussed. An iteration method which provides the optimal polynomial estimation of a chosen degree is developed in Section 4. At first the optimal solution of zero degree $\hat{L}_0^1$ for the given X and Y is determined; the first order optimal correction $\hat{L}_1^1$ when X, Y, and $\hat{L}_0^1$ are given is then computed and so on until the k order correction. At that point the calculation are iterated, starting again with the optimal zero order optimal correction when X, Y, $\hat{L}_0^1 \ldots \hat{L}_k^1$ are given. Linear "codings" of the input set are discussed in Section 5. Section 6 deals with linear optimal estimation under restrictions of "stationarity" on the processes of which the vectors $\{x^j\}$ and $\{y^j\}$ are assumed to be sample sequences. Finally the system identification problem is briefly discussed in Section 7.

The following notation will be used: L_k is a k-linear mapping and $(L_k)_{i,\alpha_1 \ldots \alpha_k}$ the associated k-way matrix with real elements. $\hat{L}_k$ indicates an estimation of L_k. A is a rectangular matrix, with transpose A^* and generalized inverse $A^\dagger$. The sum of the squares of the elements of A is written $\|A\|$ and the commutator $AB - BA$, if it exists, is written as $[A, B]$. AB, without indices, indicates the product of two matrices, in the usual sense.

2. Optimal N-Order Correction

The optimal zero-order solution of (1.2) with $Y \in \mathcal{M}(m, n)$, $X \in \mathcal{M}(r, n)$ is given by the associated normal equation as

$$(L_0)_{ij} = \frac{1}{m} \Sigma_q Y_{qj} \qquad i = 1, \ldots, m \qquad j = 1, \ldots, n . \tag{2.1}$$

Unless otherwise stated I will assume, for simplicity, that the zero-order optimal solution is zero. In other words the signal vectors are assumed to have zero mean.

The first order optimal solution of (1.2) is the best approximate solution of

$$Y = L_1(X) \tag{2.2}$$

and it is given as

$$(L_1)_{ij} = (YX^\dagger)_{ij}, \tag{2.3}$$

where $X^\dagger$ is the generalized inverse of X. This result is due to Penrose (1956). $X^\dagger$ always exists for any arbitrary matrix X; it is unique and can be calculated in a number of ways; its main properties are given in the Appendix. Of course if the matrix X is square and non-singular $X^{-1} = X^\dagger$ exists and then (2.3) is an exact solution of (1.2).

In many cases the estimation (2.3) does not provide an exact "retrieval". I define the first order error matrix to be

$$E_1 = Y - \hat{L}_1(X) = Y(I - X^\dagger X) \qquad E_1 \in \mathcal{M}(m, n) . \tag{2.4}$$

Every linear correction ΔL_1 to the optimum linear estimator is identically zero. The verification is trivial since the optimum linear solution of $E_1 = \Delta L_1 X$ is $\hat{\Delta} L_1 = E_1 X^\dagger = Y(I - X^\dagger X) X^\dagger \equiv 0$ because of property (A.2). I therefore ask for the optimal second order correction which is equivalent to finding the best approximate solution $\hat{L}_2$ of

$$E_1 = L_2(X, X) \tag{2.5}$$

or, written explicitly,

$$(E_1)_{ij} = \sum_{\alpha_1 \alpha_2} (L_2)_{i,\alpha_1 \alpha_2} X_{\alpha_1 j} X_{\alpha_2 j} . \tag{2.6}$$

It is here convenient to define the matrix

$$(C_2)_{a,j}=(C_2)_{\alpha_1\alpha_2,j}=X_{\alpha_1 j}X_{\alpha_2 j} \quad C_2\in\mathcal{M}(r^2,n), \quad (2.7)$$

where the map $\Pi^2:\alpha_1\alpha_2\to a$ uses collective indices

$$\begin{array}{rccccccc}\alpha_1\alpha_2= & 11 & 12 & \ldots & 21 & \ldots & rr\\ a= & 1 & 2 & \ldots & r+1 & \ldots & r^2\,.\end{array}$$

The transpose of (2.7) is defined as

$$(C_2)^*_{i,b}=(C_2)^*_{i,\beta_1\beta_2}=X^*_{i,\beta_1}X^*_{i,\beta_2}. \quad (2.8)$$

For instance the matrix C_2 associated with a 2×3 matrix X

$$X=\begin{pmatrix}x_{11} & x_{12} & x_{13}\\ x_{21} & x_{22} & x_{23}\end{pmatrix}$$

is given by (2.7) as a 4×3 matrix

$$C_2=\begin{pmatrix}x_{11}\cdot x_{11} & x_{12}\cdot x_{12} & x_{13}\cdot x_{13}\\ x_{11}\cdot x_{21} & x_{12}\cdot x_{22} & x_{13}\cdot x_{23}\\ x_{21}\cdot x_{11} & x_{22}\cdot x_{12} & x_{23}\cdot x_{13}\\ x_{21}\cdot x_{21} & x_{22}\cdot x_{22} & x_{23}\cdot x_{23}\end{pmatrix}.$$

It is clear how C_2 can be explicitly constructed for arbitrary X.

The extension to higher order is obvious: (2.4) and (2.7) become

$$E_k=Y-\hat{L}_k(X,\ldots,X), \quad (2.9)$$

$$(C_k)_{a,j}=(C_k)_{\alpha_1\ldots\alpha k,j}=X_{\alpha_1 j}\ldots X_{\alpha_k j}, \quad C_k\in\mathcal{M}(r^k,n) \quad (2.10)$$

under the map Π^k. Of course $C_1=X$ and Π^1 is the associated identity map.

Theorem 2.1

The best approximate solution $\hat{L}_k(k\geqq 1)$ *of the equation* $E_{k-1}=\hat{L}_k(X,\ldots,X)$ *is*

$$(\hat{L}_k)_{ij}=(E_{k-1}C_k^\dagger)_{ij}. \quad (2.11)$$

Proof: The equation

$$(E_{k-1})_{ij}=\sum_{\alpha_1,\ldots,\alpha_k}(L_k)_{i,\alpha_1\ldots\alpha_k}X_{\alpha_1 j}\ldots X_{\alpha_k j}$$

is rewritten under the Π^k map as the linear equation

$$(E_{k-1})_{ij}=\sum_a(L_k)_{i,a}(C_k)_{a,j}, \quad (L_k)_{i,a}\in\mathcal{M}(m,k)$$

$$C_k\in\mathcal{M}(r^k,n)$$

to which Theorem A-3 applies. Of course when $k=1$ the theorem is identical with Theorem A-3. One can easily check that the normal equation associated with a minimum of

$$\|E_k\|=\mathrm{tr}\{[E_{k-1}-L_k(X,\ldots,X)][E_{k-1}-L_k(X,\ldots,X)]^*\}$$

is satisfied by (2.11) and that (2.11) does give a minimum of $\|E_k\|$.

The following lemma provides an easy way to calculate $C_k^\dagger$.

Lemma 2.1

$$C_k^*C_k=(X^*X)^{⑤}, \quad (2.12)$$

where ⑤ indicates the operation of raising the elements of a matrix to the k-th power.

Proof: I write explicitly

$$\begin{aligned}(C_k^*C_k)_{ij}&=\sum_{\alpha_1,\ldots,\alpha_k}C^*_{i,\alpha_1\ldots\alpha_k}C_{\alpha_1\ldots\alpha_k,j}\\ &=\sum_{\alpha_1,\ldots,\alpha_k}X^*_{i\alpha_1}\ldots X^*_{i\alpha_k}X_{\alpha_1 j}\ldots X_{\alpha_k j}\\ &=\left(\sum_\alpha X^*_{i\alpha}X_{\alpha j}\right)^k=[(X^*X)_{ij}]^k\\ &=(X^*X)^{⑤}.\end{aligned}$$

Using (2.12) and (A.11) $C_k^\dagger$ can be obtained from

$$C_k^\dagger=[(X^*X)^{⑤}]^\dagger C_k^* \quad (2.13)$$

and since $(X^*X)^{⑤}$ is square and symmetric one may, for instance, use the method outlined in the Appendix to compute its generalized inverse. Moreover the following theorem holds

Theorem 2.2

If $(X^*X)^{⑤}$ *is non-singular the* k-th *order optimal correction is exact.*

Proof: If $(X^*X)^{⑤}$ is non-singular it has an inverse. Because of Lemma 2.1 the k-th order optimal correction gives

$$L_k(X,\ldots,X)=E_{k-1}C_k^\dagger C_k=E_{k-1}[(X^*X)^{⑤}]^\dagger[(X^*X)^{⑤}],$$

which is E_{k-1}, if the inverse of $(X^*X)^{⑤}$ exists.

In particular if $(X^*X)^{-1}$ exists the optimal linear approximation is exact; if $[(X^*X)^{②}]^{-1}$ exists the second order correction is exact, that is it provides an estimation with an associated zero mean square error.

3. Linearity of the Optimal Estimation

It is natural to ask for which class of matrices X is the linear optimal estimation also the optimal estimation of the type (1.2). If $(X^*X)^{-1}$ exists the optimal estimation is obviously linear. The following theorem attempts a more general characterization.

Theorem 3.1

The optimal linear solution of $Y=0(X)$ *cannot be improved, for arbitrary* Y, *by a* k-th *order correction*

if and only if

$$X^\dagger X = C_k^\dagger C_k + Z(I - C_k^\dagger C_k), \tag{3.1}$$

where $Z \in \mathcal{M}(n, n)$ *is arbitrary.*

Proof: The condition for the k-th correction to be zero is

$$\sum_q \sum_{\alpha_1, \ldots, \alpha_k} (E_1)_{iq} (C_k^\dagger)_{q, \alpha_1 \ldots \alpha_k} C_{\alpha_1 \ldots \alpha_k, j} = 0,$$

that is

$$Y(I - X^\dagger X)\, C_k^\dagger C_k = 0. \tag{3.2}$$

Theorem A.2 guarantees that a solution $X^\dagger X$ of (3.2) always exists; moreover it provides the general solution of (3.2) in the form

$$X^\dagger X = Y^\dagger Y(I - Z)\, C_k^\dagger C_k + Z, \tag{3.3}$$

where Z is arbitrary. If (3.2) has to be valid for arbitrary Y, it reduces to

$$C_k^\dagger C_k = X^\dagger X\, C_k^\dagger C_k. \tag{3.4}$$

Again Theorem A.2 gives the general solution $X^\dagger X$ of (3.4) as (3.1).

A particular class of matrices X for which the optimal linear estimation cannot be improved by any k-th order correction is further characterized by means of the following lemma

Lemma 3.1

If $X^* X$ is a square blocks diagonal matrix

$$X^* X = \begin{pmatrix} A & \ldots & 0 \\ \ldots & B & \ldots \\ 0 & \ldots & \ldots \end{pmatrix} \tag{3.5}$$

with $A, B \ldots$ square submatrices with all elements equal (in particular $A, B \ldots$ can be simply numbers), then

$$X^\dagger X = C_k^\dagger C_k \quad \text{for all} \quad k. \tag{3.6}$$

Proof: If $(X^* X)$ has the structure (3.5), $C_k^* C_k$ has the same structure because of Lemma 2.1. Furthermore,

$$C_k^* C_k = \begin{pmatrix} A^{\circledast} & \ldots & 0 \\ \ldots & B^{\circledast} & \ldots \\ 0 & \ldots & \end{pmatrix}. \tag{3.7}$$

It is easily verified (from A.1 to A.4) that

$$\begin{pmatrix} S & \ldots & 0 \\ \ldots & T & \ldots \\ 0 & \ldots & \end{pmatrix}^\dagger \begin{pmatrix} S & \ldots & 0 \\ \ldots & T & \ldots \\ 0 & \ldots & \end{pmatrix} = \begin{pmatrix} S^\dagger S & \ldots & 0 \\ \ldots & T^\dagger T & \ldots \\ 0 & \ldots & \end{pmatrix}. \tag{3.8}$$

Therefore consideration of $(A^{\circledast\dagger} A^{\circledast})$ is enough. The square matrix $A \in \mathcal{M}(h, h)$ with $h \leqq n$ can be expressed as

$$A = a \mathbb{1}, \tag{3.9}$$

where $\mathbb{1}$ is the matrix $h \times h$ with all elements equal to 1. Note that $\mathbb{1}^\dagger = \frac{1}{h^2} \mathbb{1}$. Because of (A.7)

$$A^\dagger A = a^\dagger a \mathbb{1}^\dagger \mathbb{1} \tag{3.10}$$

and since $(a^n)^\dagger a^n = a^\dagger a$, one obtains

$$(A^{\circledast})^\dagger A^{\circledast} = A^\dagger A. \tag{3.11}$$

From (A.11), (3.8), and (3.11) the following relation is satisfied

$$C_k^\dagger C_k = [(X^* X)^{\circledast}]^\dagger [(X^* X)^{\circledast}] = (X^* X)^\dagger (X^* X) \tag{3.12}$$

and finally,

$$C_k^\dagger C_k = X^\dagger X. \tag{3.13}$$

Matrices with other structures also satisfy property (3.13): for instance the matrices X for which $(X^* X)^{-1}$ exists, or the matrices X such that $X^* X$ contains only 0 and 1. This brings us to Theorem 3.2.

Theorem 3.2

For a matrix X *which is such that either* $(X^* X)^{-1}$ *exists or* $(X^* X)$ *has elements equal either to* 0 *or to* 1 *or* $(X^* X)$ *has the diagonal square blocks structure of Lemma 3.1, the optimal linear solution of* (1.1) *is the optimal nonlinear solution of the form* (1.2).

Proof: In all cases $X^\dagger X = C_k^\dagger C_k$, which is a particular solution of (3.1). Theorem 3.1 says that any nonlinear correction cannot improve the linear optimal estimation which is therefore the optimal one.

It is interesting that the key-vectors characterized by Theorem 3.2 have some analogies with the "noiselike" signals usually considered in holographic-like memory schemes. The analogy will become clearer later on. It is also possible to generalize the results of this section to characterize the conditions under which an optimal k-order estimation is optimal.

4. The Iteration Method

There is no reason to assume *a priori* that the second order correction to the best linear approximation is equal to the optimal second order approximation of the class (1.2). Although in many practical cases, either $(X^* X)^{-1}$ or $(C_2^* C_2)^{-1}$ exist, which automatically solves the problem, in general this is not the case, as the next theorem indicates.

Theorem 4.1

The second order correction to a first order optimal solution of (1.2) *gives the optimal second order solution if and only if* X *satisfies to*

$$X^\dagger X\, C_2^\dagger C_2\, X^\dagger = C_2^\dagger C_2\, X^\dagger. \tag{4.1}$$

Proof: The optimal quadratic correction to the optimal linear solution of $Y = 0(X)$ is given by

$$E_1\, C_2^\dagger C_2 \qquad E_1 = Y(I - X^\dagger X). \tag{4.2}$$

The optimal linear correction to

$$\hat{Y} = Y X^\dagger X + E_1 C_2^\dagger C_2 \tag{4.3}$$

is

$$E_2 X^\dagger X \quad \text{with} \quad E_2 = Y - \hat{Y}. \tag{4.4}$$

Equation (4.3) is the optimal second order solution if and only if the zero order correction is zero (and this is automatically satisfied if $\Sigma_k Y_{k\alpha} = 0$ as assumed previously) and

$$E_2 X^\dagger \equiv 0\,. \tag{4.5}$$

Provided that for arbitrary Y (4.5) holds, Eq. (4.1) follows easily.

For instance, condition (4.1) will be satisfied if either $[C_2^\dagger C_2, X^\dagger] = 0$ or $[C_2^\dagger C_2, X^\dagger X] = 0$. Those X matrices for which the optimal linear solution cannot be improved by a second order correction [they satisfy (3.1) for $k = 2$] consistently satisfy (4.1). Counterexamples to (4.1) can be constructed. Therefore to find the general optimal approximation of degree k it seems necessary to resort to an iterative scheme. The steps are:

I) The optimum approximation of zero degree (which is no longer restricted to zero) and the sequences of optimal corrections up to the degree k are calculated, as outlined before.

II) The optimum corrections to the result of step I) are computed for all degrees, starting again from the zero order degree.

III) ....

The iteration results in a series of i-ways matrices ($i = 1, 2, \ldots, k$) and in the associated mean square errors

$$\begin{aligned} &\hat{L}_0^{\mathrm{I}}, \hat{L}_1^{\mathrm{I}}, \ldots, \hat{L}_k^{\mathrm{I}}; \hat{L}_0^{\mathrm{II}}, \hat{L}_1^{\mathrm{II}}, \ldots, \hat{L}_k^{\mathrm{II}}; \ldots \\ &\|E_0^{\mathrm{I}}\|, \|E_1^{\mathrm{I}}\|, \ldots, \|E_k^{\mathrm{I}}\|; \|E_0^{\mathrm{II}}\|, \|E_1^{\mathrm{II}}\|, \ldots, \|E_k^{\mathrm{II}}\|; \ldots . \end{aligned} \tag{4.6}$$

The iteration algorithm outlined here (adapted from Katzenelson *et al.*, 1964) gives at each step a meaningful result. Convergence of the iteration algorithm, as well as the uniqueness of the optimum estimator $\{\hat{L}_0, \ldots, \hat{L}_k\}$ up to an estimator which does not affect the mean square error, are proved in the next theorem.

Theorem 4.2

The iterative algorithm

a) *has a limit* k-th *order estimator;*

b) *the limit estimator is the optimal* k-th *order estimator in the least square sense;*

c) *the limit estimator is unique up to an estimator of the same degree which does not affect the mean square error.*

Proof: a) The sequence (4.6) has the property

$$\|E_m^N\| \geqq \|E_h^k\| \quad \text{if} \quad h > m, k \geqq N\,. \tag{4.7}$$

Thus the series $\|E_k^N\|$ $(N = 1, 2, \ldots)$ is monotonically decreasing. Since $\|E_r^N\| \geqq 0$ for every N and r, the series is bounded from below. Therefore the limit

$$\lim_{N \to \infty} \|E_k^N\| = \|E\| \tag{4.8}$$

exists. Corresponding to $\|E\|$ one obtains the limit estimator $\{\hat{L}_0, \hat{L}_1, \ldots, \hat{L}_k\}$ as

$$\hat{L}_i = \lim_{N \to \infty} \hat{L}_i^N \qquad i = i, \ldots k\,. \tag{4.9}$$

b) At each correction the mean square error decreases or remains the same. In fact (4.7) holds and moreover it is easy to prove [from (2.1)] that

$$\|E_k^M\| \geqq \|E_0^{M+1}\|\,. \tag{4.10}$$

Assume now that $\{\hat{L}_0, \hat{L}_1, \ldots \hat{L}_k\}$, the limit of the iterative algorithm, is not a solution of the normal equations

$$\begin{aligned} &\Sigma_i Y_{i\alpha} = \Sigma_i [(L_0)_\alpha + (L_1 X)_{i\alpha} + \cdots + (L_k C_k)_{i\alpha}] \\ &Y X^* = (L_0 + L_1 X + \cdots + L_k C_k) X^* \\ &\quad \vdots \\ &Y C_k^* = (L_0 + L_1 X + \cdots + L_k C_k) C_k^*\,. \end{aligned} \tag{4.11}$$

In this hypothesis $\hat{L}_1 \ldots \hat{L}_k$ can be assumed and the optimal zero-th order correction can be found as $\Delta \hat{L}_0$. If $\Delta \hat{L}_0 \neq 0$ the associated mean square error is now smaller than $\|E\|$ which is in contradiction with (a).

c) Suppose that the two sequences

$$\{\hat{L}_0, \hat{L}_1; \ldots \hat{L}_k\}\ \{\hat{L}'_0, \hat{L}'_1, \ldots \hat{L}'_k\}$$

are both limits of the iterative process, satisfying the associated normal equations Eq. (4.11). I will show that the corresponding mean square errors are equal. Calling

$$\begin{aligned} \hat{L}'_0 &= \hat{L}_0 + \Delta \hat{L}_0 \\ &\vdots \\ \hat{L}'_k &= \hat{L}_k + \Delta \hat{L}_k \end{aligned} \tag{4.12}$$

it follows from (4.11) that

$$\begin{aligned} &\Sigma_i [(\Delta \hat{L}_0)_\alpha + (\Delta \hat{L}_1 X)_{i\alpha} + \cdots + (\Delta \hat{L}_k C_k)_{i\alpha}] = 0 \\ &\quad \vdots \\ &(\Delta \hat{L}_0 + \cdots + \Delta \hat{L}_k C_k) C_k^* = 0\,. \end{aligned} \tag{4.13}$$

It is then straightforward to verify that

$$\begin{aligned} \|E_{\hat{L}'_0, \ldots, \hat{L}'_k}\| &= \mathrm{tr}\{[Y - (\hat{L}'_0 + \hat{L}'_1 X + \cdots + \hat{L}'_k C_k)] \\ &\quad \cdot [Y - (\hat{L}'_0 + \hat{L}'_1 X + \cdots \hat{L}'_k C_k]^*\} \\ &= \|E_{L_0, \ldots, L_k}\|\,. \end{aligned} \tag{4.14}$$

5. Linear "Coding"

If the key-signals matrix, X, is such that $(X^* X)$ is diagonal, then (Theorem 2.2) the optimal polynomial estimation is identical with the optimal linear estimation. This simple observation suggests a possible "coding" scheme which has obvious analogies with the whitening approach used in deriving optimal estimators in gaussian noise.

"*Whitening*" *Scheme:* given Y and X, X can be always transformed into

$$X' = X S \qquad X \in \mathcal{M}(r, n) \qquad S \in \mathcal{M}(n, n) \tag{5.1}$$

with the matrix S. S is the unitary matrix which reduces $(X^* X)$ to a diagonal form (since $X^* X$ is symmetric S always exists) through $S^*(X^* X)S$. Then the optimal polynomial solution of $Y = 0(X')$ is the optimal linear one.

Note that the optimal estimate for $Y = 0(X)$ and the optimal one for $Y = 0(X')$ generally do not give the same error. The following theorem indicates a class of linear "coding" transformations on the key signals x^j, which do not affect the performance of the estimator.

Theorem 5.1

If X is transformed into

$$X' = T(X), \quad T \text{ linear mapping}, \quad X' \in \mathcal{M}(z, n) \tag{5.2}$$

the optimal (nonlinear) polynomial solution of $Y = 0(X')$ yields the same performance as the optimum (nonlinear) solution of $Y = 0(X)$, provided that T^{-1}, defined as

$$T^{-1}[T(X)] = X, \tag{5.3}$$

exists.

Proof: Suppose that

$$\hat{L}_0 + \hat{L}_1(X) + \cdots + \hat{L}_k(X, \ldots, X) \tag{5.4}$$

is the optimal k order polynomial estimate for $Y = 0(X)$, and

$$\hat{L}'_0 + \hat{L}'_1(X') + \cdots + \hat{L}'_k(X', \ldots, X') \tag{5.5}$$

it the optimal k order polynomial estimate for $Y = 0(X')$. I claim that the performances of the two estimations are identical. Suppose that (5.5) were a better estimation than (5.4): this is absurd since the k-th order

$$\hat{L}'_0 + \hat{L}'_1 T(X) + \cdots + \hat{L}'_k T \ldots T(X, \ldots, X) \tag{5.6}$$

would be better than the optimal estimation (5.4). Suppose that (5.5) were a worse estimation than (5.4): this is absurd since the k-th order

$$\hat{L}_0 + L_1 T^{-1}(X') + \cdots + \hat{L}_k T^{-1} \ldots T^{-1}(X', \ldots, X') \tag{5.7}$$

would be better than the optimal estimation (5.5).

Clearly the transformation $X' = X S$ by the matrix S [see (5.1)] cannot generally satisfy the requirement (5.3): note that $[X, S] = 0$ implies $(X^* X)$ being already diagonal!

6. A Restriction of "Stationarity"

Define the matrices $[X \in \mathcal{M}(r, n);\ Y \in \mathcal{M}(m, n)]$

$$\begin{aligned} S_{it} &= Y_{i\alpha'} X^*_{\alpha' t} \qquad S \in \mathcal{M}(m, r) \\ T_{jt} &= X_{j\alpha} X^*_{\alpha t} \qquad T \in \mathcal{M}(r, r) \end{aligned} \tag{6.1}$$

and suppose that the signals $y^\alpha_i = Y_{i\alpha}$, $x^{\alpha'}_j = X_{j\alpha'}$ satisfy the conditions

$$S_{it} = r u_q \delta_{i, t+q}, \tag{6.2}$$

$$T_{jt} = r v_p \delta_{j, t+p}; \tag{6.3}$$

the vectors u and v are defined as

$$u_q = \frac{1}{r} \Sigma_{\alpha'}\, y^{\alpha'}_{t+q} x^{\alpha'}_t, \tag{6.4}$$

$$v_p = \frac{1}{r} \Sigma_\alpha x^\alpha_{w+p} x^\alpha_w. \tag{6.5}$$

with $-(r-1) \leqq q \leqq (m-1)$ and $-(r-1) \leqq p \leqq (r-1)$; indices t and w can assume any arbitrary value for which the corresponding vector components exist. The optimal first order solution (without zero-order term) of $Y = 0(X)$ is given as usual by $Y = \hat{L}_1 X$ with $\hat{L}_1 = Y X^\dagger$. Property (A.15) gives

$$\hat{L}_1 = Y X^* (X X^*)^\dagger, \tag{6.6}$$

which becomes, through (6.1) and (A.7)

$$\hat{L}_1 = S T^\dagger. \tag{6.7}$$

Note that the (finite) matrices S and T have the typical structure of the matrices which belong to the multiplication space of the convolution algebra (see Borsellino *et al.*, 1973). Of course T is a real, symmetric matrix $(T = X X^*)$. Moreover, if the vectors x^α are sample sequences of a stationary (wide sense) stochastic process, the vector v_p is an estimate of the ensemble autocorrelation of the stochastic process. In this case standard theorems (see Doob, 1953) can be used to characterize the matrix T (Poggio, 1975b). When the inverse of $v_p \delta_{f, j+p}$ exists, Eq. (6.7) becomes completely equivalent to the standard result of the Wiener-Kolmogorov filtering theory.

It is interesting to consider the case in which the "ensemble autocorrelation" of x^α satisfies to

$$v_p = \delta_{p,0}, \tag{6.8}$$

that is

$$(X X^*)_{fj} = T_{fj} = r \delta_{f,j}. \tag{6.9}$$

Then the optimal linear estimation is

$$Y_{i\alpha} = (\hat{L}_1)_{ij} X_{j\alpha} = u_q \delta_{i,j+q} X_{q\alpha} = (u * x^\alpha)_i . \quad (6.10)$$

If the x^α vectors are "noiselike unities", in the sense of Borsellino *et al.* ($x_i^\alpha x_j^{\alpha'} = \delta_{ij}\delta_{\alpha\alpha'}$), then

$$X^* X = I . \quad (6.11)$$

This is a finite discrete scheme analoguous to the one usually considered in holographic-like associative memories [see for instance Eq. (3) in Borsellino *et al.*, 1973]; the introduction of (6.11) is similar to the hypothesis of ergodicity on the processes of which x^α and y^α are sample sequences. Of course (6.11) implies by itself that the optimal linear estimation (6.10) is also the (exact) optimal nonlinear estimation. This is a familiar result, since the optimality (in the Wiener sense) of the holographic coding scheme for "noiselike" key signals is well known in the theory of holography. Cowan (unpublished note) and more recently Pfaffelhuber (1975) have stressed a similar point which is now recovered in our nonlinear framework.

Extensions of the arguments of this section to nonlinear higher-order corrections are possible: for instance, the nonlinear holographic-like algorithm proposed by Poggio (1973) which extends the typical correlation-convolution sequence to a sequence of a generalized correlations and generalized convolutions [$\circledast^i$ and $*^i$ defined in Poggio (1973) and Geiger *et al.* (1975), respectively], can be thus recovered.

7. The Identification Problem

I shall briefly outline the system identification problem in the framework of this approach. One may distinguish two classes of problems in the identification of a system:

a) the input-output set of data is given (as the X and Y matrices, in this paper) and the problem is to find the (nonlinear) input-output mapping, which corresponds, in practical cases, to a physical system;

b) the (nonlinear) "black-box" is physically accessible, the set of (discrete) input vectors $\{x^\alpha\} = X$ can be chosen and the output set $\{y^\alpha\} = Y$ can be observed. The problem is to choose the input X in such a way that the input-output mapping performed by the system can be always determined.

Case (a) is closely connected to the estimation problem considered in this paper. Theorem 2.1 and Theorem 4.2 provide a general solution to (a). Note that in some specific situations the mapping to be determined may have to be restricted *a priori* to specific terms of representation (1.2). In these instances Theorem 2.1 rather than the iterative method represents the basic tool to solve the problem.

Case (b) lies somewhat outside the scope of this paper. However, it has a practical interest and has obvious connections with the well known equivalent problem for operators or functionals (see Lee and Schetzen, 1965). I will here outline a solution to (b), adapting a method of Barrett (1963) to the formalism of this paper. A more complete discussion will be given in Poggio (1975b).

For a given set of vector inputs $\{x_i^j\} = X_{ij}, X \in \mathcal{M}(r, n)$ an orthogonalization procedure can always produce a set of symmetrical polynomials P_k in X with the property that any two polynomials of different degree are orthogonal; namely,

$$\Sigma_j [P_k(X)]_{\alpha_1 \ldots \alpha_k, j} [P_h^*(X)]_{j, \beta_1 \ldots \beta_h} = \delta_{k,h} R_k \quad (7.1)$$

with $R_k = c(k)$ if $\beta_1, \ldots \beta_k$ is a permutation of $\alpha_1, \ldots, \alpha_k$, $R_k = 0$ otherwise.

A simple case is when the vectors $\{x_i^j\}$ are chosen as sample sequences (one for each j) of independent and normally distributed variables with zero mean and unit variance. In this case the polynomials are multidimensional Hermite polynomials; summation over j gives an estimation (good for large n) of the ensemble average. Thus (7.1) holds approximately with $c(k) = k!$. The first polynomials are

$$\begin{aligned} H_0(X) &= 1 \\ [H_1(X)]_{\alpha_1, j} &= X_{\alpha_1}^j \\ [H_2(X)]_{\alpha_1 \alpha_2, j} &= X_{\alpha_1}^j X_{\alpha_2}^j - \delta_{\alpha_1 \alpha_2} . \end{aligned} \quad (7.2)$$

The "black-box" to be identified is assumed to have the representation $\{L_0', \ldots, L_k'\}$; the input-output mapping is

$$\begin{aligned} Y_{ij} = (L_0')_{i,j} + \sum_{\alpha_1} (L_1')_{i,\alpha_1} [H_1(X)]_{\alpha_1, j} + \cdots \\ + \sum_{\alpha_1 \ldots \alpha_k} (L_k')_{i,\alpha_1 \ldots \alpha_k} [H_k(X)]_{\alpha_1 \ldots \alpha_k, j} . \end{aligned} \quad (7.3)$$

The identification of the system is performed through the following operation on the output

$$(L_h')_{i,\alpha_1 \ldots \alpha_h} \simeq \frac{1}{h!} \sum_j Y_{ij} [H_h^*(X)]_{j,\alpha_1 \ldots \alpha_h} \qquad 1 \leqq h \leqq k . \quad (7.4)$$

Of course from (7.3) a representation of the form (1.2) can be recovered. Under restrictions of "stationarity" (in the sense of Section 6) the approach outlined above leads in a natural way to an identification scheme essentially equivalent to the "white-noise" method (see Lee and Schetzen, 1965).

8. Conclusion

More detailed theoretical extensions of the theory outlined here will be presented in a forthcoming paper (Poggio, 1975b). A generalization to signal vectors being infinite sequences of (complex) numbers should allow a more general reformulation of the theory. Connections with the theory of polynomial operators will also be examined.

The present paper suggests a number of interesting applications. They will be discussed and further developed in a future work (Poggio, 1975b). I will briefly outline some of them.

Regression analysis and, in general, nonlinear optimal estimation have a number of connections with the approach presented here which seems to provide in many cases a simple and constructive answer to the identification problem, the problem of characterizing a nonlinear "black-box" from a set of input-output data. The usual problem deals with the characterization of *functional* operators; however,

it is important to point out that for many practical purposes the available data have the discrete, finite structure assumed here.

Another area in which the results of this paper may bear some interest has to do with the theories of associative recall, in connection either with the field of parallel computation or with the problem of how the brain stores and retrieves information. A variety of models of associative recall have been recently proposed. They have usually taken the form of a *specific* network, which stores information in a distributed and content-addressable fashion. Most of these networks have a similar formal structure but quite different "physical" implementations. Some, like the Associative Net of Willshaw (1972) may be realized in neural tissue (compare Marr, 1969). However the physiological evidence is, at present, far from providing useful constraints for the mechanisms actually involved. Therefore it seems important to characterize in a general way the common underlying logic of these models of associative distributed memory. So far no general formalism was available. A solution to the problem may be now provided by the formal scheme outlined in this paper. In fact a very large class of nonlinear associative memory algorithms can be described by a representation of the type of (1.2). Providing the optimum algorithm (in the mean-square sense) of this form, this paper may also suggest how specific models can be classified and compared. The issue of nonlinearity, embedded in a natural way in the present scheme, may prove to be especially important (see comment in Section 7 and: Longuet-Higgins *et al.*, 1970; Poggio, 1973, 1974; Cooper, 1974).

Finally, this paper is hoped to provide a theoretical background for developing a general approach to the learning of classifications as well as to inductive generalization, somewhat in the directions implied by the work of Marr (1969, 1970) and Willshaw (1972).

Appendix

The generalized inverse exists for any (possibly rectangular) matrix whatsoever with complex elements. Here the conjugate transpose of A is indicated with A^*.

The generalized inverse of A is defined (Penrose, 1955) to be the unique matrix $A^\dagger$ satisfying to

$$A A^\dagger A = A\,, \tag{A-1}$$

$$A^\dagger A A^\dagger = A^\dagger\,, \tag{A-2}$$

$$(A A^\dagger)^* = A A^\dagger\,, \tag{A-3}$$

$$(A^\dagger A)^* = A^\dagger A\,. \tag{A-4}$$

If A is real, so also is $A^\dagger$; if A is nonsingular, then $A^\dagger = A^{-1}$. Essentially three theorems, due to Penrose (1955) are needed in this paper. They are given here for convenience.

Theorem A-1

The four Eqs. (A-1)–(A-4) have a unique solution $A^\dagger$ for any given A.

Theorem A-2

A necessary and sufficient condition for the equation $AXB = C$ to have a solution is

$$A A^\dagger C B^\dagger B = C$$

in which case the general solution is

$$X = A^\dagger C B^\dagger + Y - A^\dagger A Y B B^\dagger\,,$$

where Y is arbitrary.

Theorem A-3

$B A^\dagger$ is the unique best approximate solution of the equation $XA = B$.

According to Penrose X_0 is the best approximate solution of

$$G = f(X) \text{ if for all } X \text{ either}$$

$$\|f(X) - G\| > \|f(X_0) - G\| \quad \text{or}$$

$$\|f(X) - G\| = \|f(X_0) - G\| \quad \text{and} \quad \|X\| \geqq \|X_0\|\,,$$

where $\|A\| = \text{trace}\,(A^* A)$.

It is straightforward to check that $A^\dagger B$ is the solution of the normal equation associated to the optimal (least square) solution of $XA = B$ (see Kohonen *et al.*, 1973).

The following relations are also useful

$$A^{\dagger\dagger} = A\,, \tag{A-5}$$

$$A^{\dagger *} = A^{* \dagger}\,, \tag{A-6}$$

$$(\lambda A)^\dagger = \lambda^\dagger A^\dagger\,, \tag{A-7}$$

$$(U A V)^\dagger = V^* A^\dagger U^* \text{ if } U, V \text{ unitary}\,, \tag{A-8}$$

$$A^* = A^* A A^\dagger\,, \tag{A-9}$$

$$A^\dagger = A^* A^{\dagger *} A^\dagger\,, \tag{A-10}$$

$$A^\dagger = (A^* A)^\dagger A^*\,, \tag{A-11}$$

$$(A^* A)^\dagger = A^\dagger A^{\dagger *}\,, \tag{A-12}$$

$$(A^\dagger A)^\dagger = A^\dagger A\,, \tag{A-13}$$

$$(A A^*)^\dagger = A^{\dagger *} A^\dagger\,, \tag{A-14}$$

$$A^\dagger = A^* (A A^*)^\dagger\,. \tag{A-15}$$

Relations (A-5)–(A-12) are given by Penrose as consequences of the definitions. (A-13) and (A-14) can be easily verified by substituting the right hand side into the corresponding definitions of the generalized inverse, since the latter is always unique. (A-15) is obtained from (A-10) and (A-14). Relations (A-8) and (A-11) provide a method of calculating the generalized inverse of a matrix A since $A^* A$, being hermitian, can be reduced to diagonal form by a unitary transformation. Thus

$A^* A = U D U^*$ with $D = \text{diag}(d_1 \ldots d_n)$ and

$A^\dagger = (A^* A)^\dagger A^* = U D^\dagger U^* A^*$ with $D^\dagger = \text{diag}(d_1^\dagger \ldots d_n^\dagger)$.

Acknowledgement. I wish to thank E. Pfaffelhuber, W. Reichardt, N. J. Strausfeld, V. Torre, and especially D. J. Willshaw for many useful comments and suggestions and for reading the manuscript. I would also like to thank I. Geiss for typing it.

References

Anderson, J. A.: A memory storage model utilizing spatial correlation functions. Kybernetik **5**, 113 (1968)

Anderson, J. A.: A simple neural network generating an interactive memory. Math. Biosci. **14**, 197 (1972)

Balakrishnan, A. V.: Identification of control systems from input-output data. International Federation of Automatic Control, Congress 1967

Barrett, J. F.: The use of functionals in the analysis of non-linear physical systems. J. Electr. Contr. **15**, 567 (1963)

Borsellino, A., Poggio, T.: Holographic aspects of temporal memory and optomotor response. Kybernetik **10**, 58 (1971)

Borsellino, A., Poggio, T.: Convolution and correlation algebras. Kybernetik **13**, 113 (1973)

Cooper, L. N.: A possible organization of animal memory and learning. Proc. of the Nobel Symposium on Collective Properties of Physical Systems. Lundquist, B. (ed). New York: Academic Press 1974

Dieudonné, J.: Foundations of modern analysis. New York: Academic Press 1969

Doob, J. L.: Stochastic processes. London: J. Wiley 1963

Gabor, D.: Associative holographic memories. IBM J. Res. Devel. **13**, 2 (1969)

Geiger, G., Poggio, T.: The orientation of flies towards visual pattern: on the search for the underlying functional interactions. Biol. Cybernetics (in press)

Heerden, P. J., van: Theory of optical information storage in solids. Appl. Optics **2**, 387 (1963)

Katzenelson, J., Gould, L. A.: The design of nonlinear filters and control systems. Part I. Information and Control **5**, 108 (1962)

Kohonen, T.: Correlation matrix memories. IEEE Trans. Comput. C-**21**, 353—359 (1972)

Kohonen, T., Ruohonen, M.: Representation of associated data by matrix operators. IEEE Trans. on Comp. (1973)

Lee, Y. W., Schetzen, M.: Measurement of the Wiener kernels of a nonlinear system by cross-correlation. Int. J. Control **2**, 237 (1965)

Longuet-Higgins, H. C.: Holographic model of temporal recall. Nature (Lond.) **217**, 104 (1968)

Longuet-Higgins, H. C., Willshaw, D. J., Bunemann, O. P.: Theories of associative recall. Rev. Biophys. **3**, 223 (1970)

Marr, D.: A theory of cerebellar cortex. J. Phys. (Lond.) **202**, 437 (1969)

Marr, D.: A theory for cerebral neocortex. Proc. roy. Soc. B **176**, 161 (1970)

Penrose, R.: A generalized inverse for matrices. Proc. Cambridge Phil. Soc. **52**, 406 (1955)

Penrose, R.: On best approximate solutions of linear matrix equations. Proc. Cambridge Phil. Soc., **52**, 17 (1956)

Pfaffelhuber, E.: Correlation memory models – a first approximation in a general learning scheme. Biol. Cybernetics (in press)

Poggio, T.: On holographic models of memory. Kybernetik **12**, 237 (1973)

Poggio, T.: In: Vecli, A. (Ed.): Processing of visual information in flies. Proc. I. Symp. It. Soc. Biophys. Parma: Tipo-Lito 1974

Poggio, T.: A theory of nonlinear interactions in many inputs systems. In preparation (1975a)

Poggio, T.: On optimal associative black-boxes: theory and applications. In preparation (1975b)

Willshaw, D. J.: Models of distributed associative memory. Ph. D. Thesis. University of Edinburgh 1971

Willshaw, D. J.: A simple network capable of inductive generalization. Proc. roy. Soc. B **182**, 233 (1972)

Dr. T. Poggio
Max-Planck-Institut
für biolog. Kybernetik
Spemannstr. 38
D-7400 Tübingen
Federal Republic of Germany