PROCEEDINGS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 103, Number 2, June 1988

A LEARNING THEOREM FOR LINEAR OPERATORS

JAN MYCIELSKI

(Communicated by R. Daniel Mauldin)

ABSTRACT. We prove a convergence theorem in linear dynamic approximation theory which yields estimates of certain series of errors.

In [**5** and **6**] we have developed a model for the function of the cerebral cortex, which is based upon a theorem of linear dynamic approximation theory. Here we extend that theorem from the case of functionals to the case of operators. This allows new variants of the models developed in [**6**]. Instead of looking at the errors of each learning line separately, we can use now the errors of bunches of n learning lines. But the theorem which is proved below has a straightforward motivation which does not require the knowledge of [**5**, **6**]. A sequence of points $\xi_0, \xi_1, \ldots \in \mathbf{R}^k$ represent the consecutive inputs to a system S. Other inputs $y_0, y_1, \ldots \in \mathbf{R}^n$ represent the desired outputs of S but they reach S a little later so that S must compute its guess $\hat{y}_t \in \mathbf{R}^n$ on the basis of $\xi_0, \ldots, \xi_t$ and $y_0, \ldots, y_{t-1}$, before getting y_t. Of course $y_t - \hat{y}_t$ is the error vector and this error is known to S at the time when $\hat{y}_{t+1}$ is to be computed. We want to design S such that the errors will be minimized in some sense, although this will require of course certain assumptions about the sequences (ξ_t) and (y_t). (In our model in [**6**] the neocortex is represented as a system of millions of overlapping thin columns perpendicular to the cortical layers. Each such column contains a system S. At the bottom of each of them the inputs ξ_t and y_t are received and in that particular model y_t is one of the coordinates of ξ_{t+1}.)

We specify S as follows. First S has a fixed (nonlinear) preparatory map $\phi\colon \mathbf{R}^k \to \mathbf{R}^m - \{0\}$. For example, ϕ could be the system of all monomials of degree $\leq d$ in the coordinates of $\mathbf{R}^k$. (In the applications in [**6**] there is a brief discussion of the ϕ's which appear natural for the model of the neocortex considered there. In that case surely $d \leq 3$ but still k is too large so that not all $\binom{k+3}{3}$ monomials of degree ≤ 3 could appear in ϕ. On the other hand functions of a different nature, such as characteristic functions of intervals, could appear among the coordinates of ϕ.) This map ϕ does not depend on t. Then S seeks a linear map $M_t\colon \mathbf{R}^m \to \mathbf{R}^n$ such that the error vector $y_t - M_t(\phi(\xi_t))$ be small in a sense. So in our system S, we have $\hat{y}_t = M_t(\phi(\xi_t))$. The linear operator M_t is called the long term memory of S while ϕ could be called the instinctive memory of S. M_t is computed from $\phi(\xi_{t-1})$, M_{t-1} and y_{t-1} only, and this computation is very simple so that it can be accomplished in fractions of a second by the tissues of the central nervous system.

Received by the editors December 20, 1986.
1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 41A25; Secondary 92A08, 92A25.

In the present paper ϕ will play no role. So we put $x_t = \phi(\xi_t)$ and we study only a subsystem A of S which is defined as follows.

$t = 0, 1, 2, \ldots$ is the time variable.

$x_t \in \mathbf{R}^m - \{0\}$ is the input to A at time t.

$y_t \in \mathbf{R}^n$ is the desired output of A at time t.

Both x_t and y_t are viewed as column vectors.

L_{mn} is the space of linear maps $M\colon \mathbf{R}^m \to \mathbf{R}^n$, i.e., of real $(m \times n)$-matrices.

$M_t \in L_{mn}$ is the state of the memory of A at time t.

$\hat{y}_t = M_t x_t$ is the prediction of y_t which A makes at time t.

The error of this prediction is defined by the formula

$$e_t = (y_t - \hat{y}_t)/||x_t||.$$

So $e_t \in \mathbf{R}^n$ and e_t is a vector of relative errors.

Given any $M_0 \in L_{mn}$, the memory of A at time 0, A updates its memory according to the following formula

$$M_{t+1} = M_t + \left[1 - \frac{\theta}{||e_t||}\right]_+ \frac{e_t x_t^T}{||x_t||}, \tag{1}$$

where $\theta \geq 0$ is a certain constant, $[\alpha]_+ = \max(0, \alpha)$ and x_t^T is the transpose of x_t.

It is easy to check that the formula (1) is equivalent to the following statement. M_{t+1} is the operator which minimizes the norm $||M_{t+1} - M_t||$ while satisfying the inequality

$$||y_t - M_{t+1}x_t||/||x_t|| \leq \theta.$$

(Recall the definition: for all $M = (a_{ij}) \in L_{mn}$,

$$||M|| = \left(\sum_{ij} a_{ij}^2\right)^{1/2} = (\mathrm{Tr}(MM^T))^{1/2}.)$$

The rule (1) is motivated by its computational simplicity and the upper bounds given in the theorem below.

We still need the notations

$$e_t(M) = (y_t - Mx_t)/||x_t||, \tag{2}$$

in particular $e_t = e_t(M_t)$,

$$E_s(M) = \operatorname*{Sup}_{t \geq s} ||e_t(M)||, \tag{3}$$

and, for $\alpha = 1, 2$, we define the following series of errors

$$S_\alpha = \sum_{t=0}^{\infty} ([||e_t|| - \theta]_+)^\alpha.$$

With these notations our result takes the following form

THEOREM. (i) $||M_{t+1} - M_t|| = [||e_t|| - \theta]_+$.

(ii) $S_2 \leq \mathrm{Inf}\{||M - M_0||^2 | E_0(M) \leq \theta\}$.

(iii) $S_1 \leq \mathrm{Inf}\{(||M - M_0||^2 - S_2)/2(\theta - E_0(M)) | E_0(M) < \theta\}$.

(iv) *If* $\operatorname{Inf}\{E_s(M) | M \in L_{mn},\ s = 0, 1, \dots\} < \theta$, *then* $S_1 < \infty$.
(*We have stipulated here that* $\operatorname{Inf} \varnothing = \infty$.)

PROOF. (i) is obvious from (1).

To show (ii) we define $x_t^0 = x_t/||x_t||$, $\sigma_t = ||M - M_t||^2$ and $\alpha_t = [1 - \theta/||e_t||]_+$. By (2) we have

$$(M - M_t)x_t^0 = e_t - e_t(M), \tag{4}$$

and, by (3),

$$||e_t(M)|| \leq E_0(M). \tag{5}$$

A routine calculation using (4), Schwarz's inequality and (5) yields

$$\begin{aligned}
\sigma_{t+1} &= ||(M - M_t) - \alpha_t e_t x_t^{0T}||^2 \\
&= \sigma_t - 2\alpha_t ||e_t||^2 + 2\alpha_t \operatorname{Tr}(e_t(M)e_t^T) + \alpha_t^2 ||e_t||^2 \\
&\leq \sigma_t - 2\alpha_t ||e_t||^2 + 2\alpha_t E_0(M)||e_t|| + \alpha_t^2 ||e_t||^2 \\
&= \sigma_t - [||e_t|| - \theta]_+^2 - 2(\theta - E_0(M))[||e_t|| - \theta]_+.
\end{aligned}$$

Hence, since $\sigma_{t+1} \geq 0$ for all t, we get

$$0 \leq \sigma_0 - S_2 - 2(\theta - E_0(M))S_1.$$

So if $\exists M[\theta \geq E_0(M)]$ we get (ii) and if $\exists M[\theta > E_0(M)]$ we get (iii).

(iv) follows of course from (3) and (iii).

REMARKS. 1. The work of Gersho and Lucky [**1**, **2**, **3**, **4**] (see also [**7**, **8**] and references therein) suggests that if the inputs x_t are loaded with random noise, then the rule (1) should be replaced by the rule

$$M_{t+1} = M_t + c\left[1 - \frac{\theta}{||e_t||}\right]_+ \frac{e_t x_t^T}{||x_t||}$$

where c is a suitable constant in the interval $[0, 1]$. But we do not know any clear-cut theorem elucidating this situation.

2. One can replace the domain space $\mathbf{R}^m$ by any Hilbert space and it is easy to generalize the theorem accordingly. One can replace both spaces $\mathbf{R}^m$ and $\mathbf{R}^n$ by the Hilbert space l_2, but then our norm $||(a_{ij})|| = (\sum a_{ij}^2)^{1/2}$ may be infinite. So our Theorem may be significant only for appropriate choices of M_0. The unit matrix is a natural choice for M_0 in such applications in which $y_t = x_{t+1}$, for $t = 0, 1, \dots$.

3. One can also modify the theorem to a continuous time variable in the style of §3 in [**6**].

REFERENCES

1. A. Gersho, *Automatic equalization technique for multilevel pulse transmissions*, Bell Telephone Laboratories Technical Memorandum, MM 65-1381-13, December 1965.
2. ——, *Adaptive equalization of highly dispersive channels for data transmission.* I, Bell Telephone Laboratories Technical Memorandum, MM 68-1386-3, April 1968.
3. R. W. Lucky, *Automatic equalization for digital communication*, Bell System Tech. J. **44** (1965), 547–588.
4. ——, *Techniques for adaptive equalizations for digital communication systems*, Bell System Tech. J. **45** (1966), 255–286.

5. J. Mycielski, *Can mathematics explain natural intelligence?*, Physica **22D** (1986), 366–375. Also appeared as a Los Alamos National Laboratory Report LA-10492-MS, UC32 July 1985.
6. J. Mycielski and S. Swierczkowski, *A model of the neocortex*, Los Alamos National Laboratory Report LA-11140-MS, UC32 November 1987.
7. R. S. Sutton, *Learning to predict by the method of temporal differences*, Technical Report TR87-509.1, GTE Laboratories Inc., Waltham, Mass., 1987.
8. B. Widrow and S. D. Stearns, *Adaptive signal processing*, Prentice-Hall, Englewood Cliffs, N. J., 1985.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF COLORADO, BOULDER, COLORADO 80309-0426