

backward algorithm in some applications reflects the fact that for the purposes of these applications, the "true" model is of no particular interest as long as the fitted model performs well.

INITIAL STATISTICS

It follows that the initial values of (p, P, Q) can play two roles. They choose one of the values which maximize (perhaps only locally) the likelihood function L . As T tends to infinity, if the sequence of estimators obtained converges, then it must be to some point in the set M of correct values of (p, P, Q) . Thus, the initial values also choose, *a fortiori*, one of the many models consistent with perfect knowledge of the (unconditional) distribution of the observable string. Of course, this second choice, the choice of a single "true" value of (p, P, Q) , is not revealed to us with only a finite amount of data. In both roles, the initial values represent decisive "prior" information for the problem at hand. In the speech application, the simply fitted model produces poor speech recognition, yet an averaging of the initial and final estimates (with weights fitted by another use of the forward-backward algorithm similar to Jelinek *et al.* [4]) produces a successful model.

CONSEQUENCES FOR PRACTICAL SPEECH RECOGNITION

The singularity of the likelihood function in the case where the observables have a probability density function as described suggests either some Bayesian technique, as mentioned, or choosing a model without singularities such as a model having multinomial distributions for the conditional distribution of the observables. The multinomial distributions arise naturally in models for data encoded by vector quantization. For continuous speech recognition, we prefer the latter alternative, partly because it is simpler, but mainly because of an additional practical advantage of the multinomial model. It is nonparametric, and hence, ultimately can give a better fit to the complex high-dimensional distribution of speech data than is possible in practice by parametric density models such as mixtures of multivariate Gaussian distributions. It would be interesting to use smooth nonparametric density estimates (see, e.g., [8]), but this appears difficult in high dimensions.

A practical consequence of the lack of identifiability of the distribution may be seen in an application to the continuous speech recognition problem of training the hidden Markov model to a new talker. If the model were identifiable, then one might hope to obtain the value of (p, P, Q) for the new talker from a small speech sample of the new talker and the large set of (p', P', Q') available from having trained a large number of talkers previously. This could be done by using some metric or other to interpolate among the known (p', P', Q') points using the new speech data as a guide. However, if each (p', P', Q') is an arbitrary choice from equivalent points in a set M' , then any simple notion of distance is likely to be useless. It is not clear how much the use of a common (to all talkers) initial value (p'', P'', Q'') would help in this. Clearly, the use of such a common initial statistic is inconsistent with attempts to maximize the likelihood for any one talker by finding the largest of several local maxima.

REFERENCES

- [1] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model in ecology," *Bull. Amer. Math. Soc.*, vol. 73, pp. 360-363, 1966.
- [2] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, vol. 37, no. 6, pp. 1554-1563, 1966.
- [3] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic func-

tions of Markov chains," *Ann. Math. Statist.*, vol. 41, pp. 164-171, 1970.

- [4] F. Jelinek and R. L. Mercer, "Interpolated estimation of Markov source parameters from sparse data," in *Pattern Recognition in Practice*, E. L. Gelsema and L. N. Kanal, Ed. New York: North-Holland, 1980, pp. 381-387.
- [5] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, pp. 532-555, Apr. 1976.
- [6] L. A. Liporace, "Maximum-likelihood analysis for multivariate observations of Markov sources," in *Proc. Symp. Appl. of Hidden Markov Models to Text and Speech*, 1980, pp. 58-87, J. D. Ferguson, Ed., Inst. for Defense Analyses, Thanet Rd., Princeton, NJ 08540.
- [7] T. Petrie, "Probabilistic functions of finite Markov chains," *Ann. Math. Statist.*, vol. 40, no. 1, pp. 97-115, 1969.
- [8] A. R. Tapia and J. R. Thompson, *Nonparametric Probability Density Estimation*. Baltimore, MD: Johns Hopkins Univ. Press, 1978.

A Relationship Between the Levinson Algorithm and the Conjugate Direction Method

ARYE NEHORAI AND MARTIN MORF

Abstract—This note provides a connection between the (block) Levinson algorithm and the conjugate direction method (CDM). We first present an extension of the scalar CDM to the block matrix case. Then we show how the (block) Levinson recursions can be derived as a special case of the (extended) CDM. This gives a new derivation and interpretation for Levinson's algorithm.

I. INTRODUCTION

The conjugate direction method (CDM) [1] is a well-known class of recursive optimization algorithms that are used in the minimization of quadratic functions without the need for inversion of the Hessian matrix [2], [3]. Levinson's algorithm [4] was originally designed to efficiently compute one-step forward predictors of a stationary random signal, given the signal covariance. It can also be interpreted as an efficient method for inversion of Toeplitz matrices. Whittle [5] and Wiggins and Robinson [6] have extended the Levinson algorithm to the case of block Toeplitz matrices. Levinson's algorithm has been extensively used in the areas of statistics, geophysics, and speech analysis (see, e.g., [7]). The purpose of this note is to show that the (block) Levinson algorithm can be derived and viewed as a special case of the CDM, with a proper choice of conjugate directions.

In Section II we present an extension of the scalar CDM to the block matrix case and then review the block Levinson algorithm. In Section III we use the CDM to derive the Levinson recursions as a special case. This gives a new derivation and interpretation for Levinson and related algorithms, as is discussed in Section IV.

II. THE BLOCK CDM AND REVIEW OF LEVINSON'S ALGORITHM

The Block CDM

We consider here the block extension of the CDM which solves the problem

Manuscript received October 20, 1981; revised September 30, 1982. This work was supported by the Defense Advanced Research Projects Agency under Contract MDA903-80-C-0331.

The authors are with the Information Systems Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA 94305.

$$\text{minimize } \text{tr} \left\{ \frac{1}{2} x^T Q x - c^T x \right\} \quad (1)$$

where the superscript T denotes transpose, Q is an $np \times np$ symmetric positive definite matrix, and x and c are $np \times m$ block vectors. The corresponding unique solution is

$$x^* = Q^{-1} c. \quad (2)$$

The method assumes that we have available n Q -orthogonal full column rank $np \times p$ block vectors $\{d_j\}_{j=0}^{n-1}$, i.e.,

$$d_j^T Q d_l = 0 \quad \text{if } j \neq l. \quad (3)$$

The following theorem is the block extension of the scalar case.

Block Conjugate Direction Theorem: Let $\{d_j\}_{j=0}^{n-1}$ be a set of full column rank Q -orthogonal block vectors. For any $x_0 \in R^{np \times m}$, the sequence $\{x_i\}$ generated according to

$$x_{i+1} = x_i + d_i \alpha_i \quad (4)$$

with

$$\alpha_i = -[d_i^T Q d_i]^{-1} d_i^T g_i \quad (5)$$

and

$$g_i = Q x_i - c \quad (6)$$

converges to the unique solution $x^* = Q^{-1} c$ after n steps, i.e., $x_n = x^*$.

To prove the above theorem, one can directly extend the corresponding scalar version in [2] or [3].

The Block Levinson Algorithm

In order to see the similarity of the block Levinson algorithm to the CDM, we first write it as a minimization of the form

$$\text{minimize } \text{tr} \{ E e_{n,t} e_{n,t}^T \} \quad (7)$$

where $e_{n,t}$ is the n th-order forward linear prediction error of the observation signal y_t based on its n past values, namely,

$$e_{n,t} = y_t - \sum_{j=1}^n a_{n,j}^T y_{t-j}, \quad t \geq n \quad (8)$$

where y_t is a $p \times 1$ stationary random signal vector, and the coefficients $\{a_{n,j}\}$ are $p \times p$ matrices. This minimization problem can be equivalently written as

$$\text{minimize } \text{tr} \left\{ \frac{1}{2} a^T R_{n-1} a - R_{1:n}^T a \right\} \quad (9)$$

where a is the n th-order prediction filter of y_t , given by

$$a = [a_{n,1}^T, \dots, a_{n,n}^T]^T \quad np \times p, \quad (10)$$

the covariance matrix R_{n-1} has (with the stationarity assumption) the following block Toeplitz structure

$$R_{n-1} = \begin{bmatrix} R_0 & R_1 & \cdots & R_{n-1} \\ R_{-1} & & & \\ \vdots & & & \\ R_{-n+1} & R_{-1} & & R_0 \end{bmatrix} \quad np \times np \quad (11)$$

where

$$R_j = E y_t y_{t-j}^T \quad p \times p, \quad (12)$$

and the block correlation vector $R_{1:n}$ in (9) is

$$R_{1:n} = [R_1, \dots, R_n]^T \quad np \times p. \quad (13)$$

The definition (12) implies that

$$R_j = R_{-j}^T. \quad (14)$$

In order to solve the forward prediction problem (7), Levinson's algorithm also solves the backward estimation problem, namely,

$$\text{minimize } \text{tr} \{ E r_{n,t} r_{n,t}^T \} \quad (15)$$

where

$$r_{n,t} = y_{t-n} - \sum_{j=1}^n b_{n,j}^T y_{t-n+j}, \quad (16)$$

i.e., $r_{n,t}$ is the linear estimation error of y_{t-n} based on its n future values. The corresponding recursive solution of the block Levinson algorithm is

$$a_{i+1} = \begin{bmatrix} a_i \\ 0 \end{bmatrix} + \begin{bmatrix} -b_i \\ I_p \end{bmatrix} R_{r,i}^{-1} \Delta_{i+1} \quad (17a)$$

$$b_{i+1} = \begin{bmatrix} 0 \\ b_i \end{bmatrix} + \begin{bmatrix} I_p \\ -a_i \end{bmatrix} R_{e,i}^{-1} \Delta_{i+1}^T \quad (17b)$$

$$R_{e,i+1} = R_{e,i} - \Delta_{i+1}^T R_{r,i}^{-1} \Delta_{i+1} \quad (17c)$$

$$R_{r,i+1} = R_{r,i} - \Delta_{i+1} R_{e,i}^{-1} \Delta_{i+1}^T \quad (17d)$$

where I_p is the $p \times p$ identity matrix and

$$a_i = [a_{i,1}^T, \dots, a_{i,i}^T]^T \quad ip \times p \quad (17e)$$

$$b_i = [b_{i,i}^T, \dots, b_{i,1}^T]^T \quad ip \times p \quad (17f)$$

$$R_{r,i} = E r_{i,t} r_{i,t}^T \quad R_{e,i} = E e_{i,t} e_{i,t}^T \quad (17g)$$

$$\Delta_{i+1} = [0 \quad -b_i^T \quad I_p] R_{i+1} \begin{bmatrix} I_p \\ -a_i \\ 0 \end{bmatrix} \quad (17h)$$

$$a_0 = b_0 = 0 \quad R_{e,0} = R_{r,0} = R_0. \quad (17i)$$

The block vectors $\{a_j\}$ and $\{b_j\}$ have the interpretation of the least squares solutions to the minimization problems (7) and (15), respectively, with $n = j$ (see, e.g., [8]).

III. DERIVATION OF LEVINSON'S ALGORITHM AS A SPECIAL CDM

We shall now show that the (block) Levinson recursions can be derived and interpreted as a particular (block) CDM, with suitable substitution of conjugate directions and initial conditions in (4)–(6).

Consider the minimization problem (7). Here $Q = R_{n-1}$ and $c = R_{1:n}$. To obtain the Levinson solution (17a) as a CDM, assume by induction that we have available $\{b_j\}_{j=1}^i$. Now introduce the block vectors

$$d_{a,j} = [-b_j^T \quad I_p \quad 0, \dots, 0]^T \quad np \times p. \quad (18)$$

From the least squares interpretation of $\{b_j\}$, we have the relation

$$d_{a,j}^T R_{n-1} d_{a,l} = \begin{cases} 0, & \text{if } j \neq l \\ R_{r,j} & \text{if } j = l. \end{cases} \quad (19)$$

Note that this result is actually equivalent to the fact that the backward residuals of different orders are uncorrelated.

Equation (19) implies that we can use $\{d_{a,j}\}$ as conjugate directions to solve (7). With this substitution and with zero initial conditions, we shall later assume that the corresponding partial CDM solution x_i has, by induction, the form

$$x_{a,i} = [a_i^T \quad 0, \dots, 0]^T \quad np \times p. \quad (20)$$

Now we show that for the Levinson algorithm, $d_i^T g_i = -\Delta_{i+1}$. Substituting the general expression (6) for g_i and then using

TABLE I
SUMMARY OF THE RELATIONSHIPS BETWEEN THE LEVINSON ALGORITHM
AND THE CDM

CDM	Levinson	
Q	R_{n-1}	R_{n-1}
c	$R_{1:n}$	$R_{-n:-1}$
x_i	$x_{a,i}$	$x_{b,i}$
d_i	$d_{a,i}$	$d_{b,i}$
$d_i^T Q d_i$	$R_{r,i}$	$R_{e,i}$
$d_i^T g_i$	$-\Delta_{i+1}$	$-\Delta_{i+1}^T$

the previous expressions, we find that

$$\begin{aligned}
 d_i^T g_i &= d_i^T (Q x_i - c) \\
 &= d_{a,i}^T [R_{n-1} [d_i^T \ 0 \ \cdots \ 0]^T - R_{1:n}] \\
 &= -[-b_i^T I_p \ 0 \ \cdots \ 0] \begin{bmatrix} R_{-1} & R_0 & \cdots & R_{n-1} \\ & & & R_{n-1} \\ R_{-n} & R_{-n+1} & & R_0 \end{bmatrix} \begin{bmatrix} I_p \\ -a_i \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\
 &= -[0 \ -b_i^T \ I_p] R_{i+1} \begin{bmatrix} I_p \\ -a_i \\ 0 \end{bmatrix} \\
 &= -\Delta_{i+1}. \tag{21}
 \end{aligned}$$

Substituting all the above expressions in the general CDM solution (4)–(6), we obtain

$$[a_{i+1}^T \ 0 \ \cdots \ 0]^T = [a_i^T \ 0 \ \cdots \ 0]^T + [-b_i^T \ I_p \ 0 \ \cdots \ 0]^T R_{r,i}^{-1} \Delta_{i+1}. \tag{22}$$

The last recursion (22) is equivalent to the Levinson recursion (17a), as can be seen by referring to its nonzero upper $(i+1)p$ rows. Note that we were allowed to write the LHS of (22) in the above form since the last $(n-i-1)p$ rows of its RHS are identically zero. This verifies our assumption on the structure of the partial solutions by induction.

In a similar way, we can also derive the dual Levinson recursion (17b) as a specialized CDM where we use the block conjugate directions

$$d_{b,j} = [0, \cdots, 0 \ I_p \ -a_j^T]^T \quad np \times p; \tag{23}$$

the corresponding partial solution x_i is

$$x_{b,i} = [0, \cdots, 0 \ b_i^T]^T \quad np \times p. \tag{24}$$

This completes the derivation of the Levinson algorithm as a specialized CDM by induction, the results of which are summarized in Table I.

IV. DISCUSSION

In the above sections we have shown how the Levinson algorithm can be derived and interpreted as a special case of the CDM. In view of this relationship, we note the following properties of Levinson's algorithm that indicate how it exploits the Toeplitz structure of the covariance matrix.

1) Levinson's algorithm generates the conjugate directions required for each recursion independently in the following special manner. In the computation of the new partial solution $x_{a,i+1}$, it uses the previous result of $x_{b,i}$ to construct the required conjugate direction $d_{a,i}$. In the computation of $x_{b,i+1}$, it uses $x_{a,i}$ for a similar generation of $d_{b,i}$. This means that

Levinson's algorithm is a particular combination of two inter-coupled CDM solutions.

2) While in the general CDM the partial solutions $\{x_j\}$ and the matrix Q have maximum dimension at each stage of the recursion, the dimensions of the corresponding filters $\{a_j\}$ and the matrices $\{R_j\}$ grow at each step of the recursion.

The connections found in this paper, together with the deterministic geometrical interpretation of the CDM (cf. [2] and [3]), can now be used to provide a new geometrical interpretation for the Levinson and related algorithms. Note that this description deals with the optimal prediction filters, while the previously well-known explanations are usually concerned with the stochastic interpretation of the innovation signals. However, note that these two descriptions are analogous by the isomorphism between the two corresponding spaces.

Other algorithms for which the CDM explanation also holds include the time update of the fast recursive least squares algorithm in [9], and the pure order recursions of the ladder (or lattice) filter algorithm in, see, e.g., [10], developed by Morf *et al.* The CDM can be shown to be a common basis for these methods and others; see [11].

REFERENCES

- [1] M. R. Hestenes and E. L. Stiefel, "Methods of conjugate gradients for solving linear systems," *J. Res. Nat. Bur. Standards*, vol. 49, sect. B, pp. 409–436, 1952.
- [2] M. R. Hestenes, *Conjugate Direction Methods in Optimization*. New York: Springer-Verlag, 1980.
- [3] D. G. Luenberger, *Introduction to Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley, 1973.
- [4] N. Levinson, "The Wiener RMS (root-mean-square) error criterion in filter design and prediction," *J. Math Phys.*, vol. 25, pp. 261–278, Jan. 1947.
- [5] P. Whittle, *Prediction and Regulation*. New York: Van Nostrand-Reinhold, 1963.
- [6] R. Wiggins and E. A. Robinson, "Recursive solution to the multi-channel filtering problem," *J. Geophys. Res.*, vol. 70, no. 8, pp. 1885–1891, 1965.
- [7] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 501–580, 1975.
- [8] T. Kailath, "A view of three decades of linear filtering theory," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 146–181, 1974.
- [9] M. Morf, T. Kailath, and L. Ljung, "Fast algorithms for recursive identification," in *Proc. 15th IEEE Conf. Decision Contr.*, Clearwater Beach, FL, Dec. 1976, pp. 916–921.
- [10] M. Morf, B. Dickinson, T. Kailath, and A. Vieira, "Efficient solutions of covariance equations for linear prediction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 429–435, Oct. 1977.
- [11] A. Nehorai and M. Morf, "A unified derivation for fast estimation algorithms by the conjugate direction method," in preparation.

Error Analysis of Good-Winograd Algorithm Assuming Correlated Truncation Errors

G. PANDA, R. N. PAL, AND B. CHATTERJEE

Abstract—This paper investigates the error performance of the Good-Winograd algorithm (GWA) when implemented in fixed-point mode us-

Manuscript received October 26, 1981; revised August 23, 1982 and November 5, 1982.

G. Panda is with the Department of Electronics and Telecommunications Engineering, University College of Engineering, Burla, Orissa, India 768018.

R. N. Pal and B. Chatterjee are with the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur-721302, India.