Correspondence

Proof of a Special Case of Shanks' Conjecture

B. D. O. ANDERSON AND E. I. JURY

Abstract—In 1972 Shanks conjectured that the least squares inverse of a two-dimensional polynomial is stable, and verified the conjecture numerically for certain low-degree two-dimensional polynomials. Recently the conjecture was proved false. However, in this note we prove the conjecture for all polynomials of a restricted and low degree. The key to the verification lies in utilizing the centrosymmetric properties of the Toeplitz matrix which arises in an equation yielding the coefficients of the approximate inverse.

Introduction

It is well known [1] that the least squares inverse of a one-dimensional polynomial is always stable. This fact can be utilized for effective design of one dimensional digital filters. Based on this work, Shanks [2] in 1972 conjectured the same result for two-dimensional digital filters, and he verified the conjecture numerically for several low-degree polynomials. Recently the conjecture was shown in general to be false [3]. By restricting the degree of the polynomials involved, we can, however, demonstrate the truth of the conjectiure in a limited situation.

Main Result

We shall suppose there is prescribed a two-dimensional polynomial

$$A(z_1, z_2) = a_{00} + a_{10}z_1 + a_{01}z_2 + a_{11}z_1z_2 \qquad (1)$$

where, without loss of generality, $a_{00} > 0$. We shall establish the stability of an approximate inverse of the form

$$B(z_1, z_2) = b_{00} + b_{10}z_1 + b_{01}z_2 + b_{11}z_1z_2. \qquad (2)$$

Using two-dimensional convolution, we form

$$A(z_1, z_2)B(z_1, z_2) = C(z_1, z_2) = c_{00} + c_{10}z_1 + c_{01}z_2 + c_{11}z_1z_2 + c_{20}z_1^2 + c_{21}z_2^2 + c_{21}z_1^2z_2 + c_{12}z_1z_2^2 + c_{22}z_1^2z_2^2.$$

The c_{ij} are easily computed from the a_{ij}, b_{ij}. To obtain the approximate inverse, we form the quantity

$$1 - c_{00}^2 + \sum c_{ij}^2 = 1 - a_{00}^2 b_{00}^2 + \sum c_{ij}^2$$

and seek to minimize it with respect to the b_{ij}. The result is that the b_{ij} must satisfy the equation

$$\begin{bmatrix} \Gamma_{00} & \Gamma_{10} & \Gamma_{01} & \Gamma_{11} \\ \Gamma_{10} & \Gamma_{00} & \Gamma_{11}^1 & \Gamma_{01} \\ \Gamma_{01} & \Gamma_{11}^1 & \Gamma_{00} & \Gamma_{10} \\ \Gamma_{11} & \Gamma_{01} & \Gamma_{10} & \Gamma_{00} \end{bmatrix} \begin{bmatrix} b_{00} \\ b_{10} \\ b_{01} \\ b_{11} \end{bmatrix} = \begin{bmatrix} a_{00} \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad (3)$$

or in vector form

$$\underline{\Gamma b} = \underline{a}.$$

The Γ_{ij} are given by a simple calculation as follows:

$$\begin{aligned} \Gamma_{00} &= a_{00}^2 + a_{10}^2 + a_{01}^2 + a_{11}^2 \\ \Gamma_{11} &= a_{11}a_{00}, \Gamma_{11}^1 = a_{11}a_{10} \\ \Gamma_{01} &= a_{01}a_{00} + a_{11}a_{10} \\ \Gamma_{10} &= a_{10}a_{00} + a_{11}a_{01}. \end{aligned} \qquad (4)$$

The matrix Γ is block Toeplitz and centrosymmetric. Utilizing the centrosymmetry, the above equation yields:

$$\begin{bmatrix} \Gamma_{00} + \Gamma_{11} & \Gamma_{01} + \Gamma_{10} & 0 & 0 \\ \Gamma_{10} + \Gamma_{01} & \Gamma_{00} + \Gamma_{11}^1 & 0 & 0 \\ 0 & 0 & \Gamma_{00} - \Gamma_{11} & \Gamma_{10} - \Gamma_{10} \\ 0 & 0 & \Gamma_{10} - \Gamma_{01} & \Gamma_{00} - \Gamma_{11}^1 \end{bmatrix} \begin{bmatrix} b_{00} + b_{11} \\ b_{10} + b_{01} \\ -b_{00} + b_{11} \\ -b_{10} + b_{01} \end{bmatrix} = \begin{bmatrix} a_{00} \\ 0 \\ -a_{00} \\ 0 \end{bmatrix}.$$

Denote the top left 2×2 submatrix by A, the lower right 2×2 submatrix by C. Let $\Delta_1 = \det A$, $\Delta_2 = \det C$. Then one easily finds that

$$b_{00} = \frac{1}{2} a_{00} \left[\frac{A_{22}}{\Delta_2} + \frac{C_{22}}{\Delta_2}\right] \quad b_{11} = \frac{1}{2} a_{00} \left[\frac{A_{22}}{\Delta_1} - \frac{C_{22}}{\Delta_1}\right]$$

$$b_{01} = \frac{1}{2} a_{00} \left[\frac{C_{21}}{\Delta_2} - \frac{A_{21}}{\Delta_1}\right] \quad b_{10} = \frac{1}{2} a_{00} \left[-\frac{C_{21}}{\Delta_2} - \frac{A_{21}}{\Delta_1}\right]. \qquad (6)$$

In the appendix, it is shown that

$$|A_{21}| < A_{22}, \quad |C_{21}| < C_{22}, \quad \Delta_1 > 0, \quad \Delta_2 > 0. \qquad (7)$$

Now the following conditions on the b_{ij} are necessary and sufficient for $B(z_1, z_2)$ to be stable [2]:

$$\left|\frac{b_{01}}{b_{00}}\right| < 1 \qquad (8)$$

$$\left|1 \pm \frac{b_{10}}{b_{00}}\right| > \left|\frac{b_{11} \pm b_{01}}{b_{00}}\right|. \qquad (9)$$

We can check these are satisfied. Using (6) we have

$$\left|\frac{b_{01}}{b_{00}}\right| = \left|\frac{C_{21}\Delta_1 - A_{21}\Delta_2}{C_{22}\Delta_1 + A_{22}\Delta_2}\right| \leqslant \frac{|C_{21}|\Delta_1 + |A_{21}|\Delta_2}{C_{22}\Delta_1 + A_{22}\Delta_2} < 1$$

by (7). Next, (9) follows if and only if

$$|b_{00} \pm b_{10}| > |b_{11} \pm b_{01}|$$

since b_{00} is easily checked to be nonzero. In turn, this follows if and only if

Manuscript received December 17, 1975; revised June 18, 1976. This work was supported by the Australian Research Grants Committee and the Joint Services Electronics Program under Contract F44620-71-C-0087 (to the University of California, Berkeley).

B. D. O. Anderson is with the Department of Electrical Engineering, University of Newcastle, New South Wales, 2308, Australia.

E. I. Jury is with the Department of Electrical Engineering, University of California, Berkeley, CA 94720.

$$|(A_{22}\Delta_2 + C_{22}\Delta_1) \mp (A_{21}\Delta_2 + C_{21}\Delta_1)| > |(A_{22}\Delta_2 - C_{22}\Delta_1) \pm (C_{21}\Delta_1 - A_{21}\Delta_2)|.$$

With

$$x = A_{22}\Delta_2 \mp A_{21}\Delta_2 \qquad y = C_{22}\Delta_1 \mp C_{21}\Delta_1$$

these inequalities are equivalent to

$$|x + y| > |x - y|. \tag{10}$$

However, by (7), x and y are positive, so that (10) is trivially true. Thus (9) is verified.

We remark that minor adjustments can be made to cover the case when the roles of z_1 and z_2 are interchanged.

Conclusion

We have verified Shanks conjecture for polynomials of very restricted degree. The properties of the Toeplitz matrix arising in the formulations are of importance in the proof, and it may be that further restricted degree results could be obtained (despite the falsity of the conjecture in general) by using such Toeplitz matrix properties.

Appendix

Verification of (7)

$$\begin{aligned}|A_{21}| &= |\Gamma_{10} + \Gamma_{01}| \\ &= |(a_{00} + a_{11})(a_{01} + a_{10})| \\ &\leq \tfrac{1}{2}(a_{00} + a_{11})^2 + \tfrac{1}{2}(a_{01} + a_{10})^2 \\ &= \tfrac{1}{2}(a_{00}^2 + a_{11}^2 + a_{01}^2 + a_{10}^2) + a_{00}a_{11} + a_{10}a_{01}.\end{aligned}$$

Also,

$$|a_{00}a_{11}| \leq \tfrac{1}{2}(a_{00}^2 + a_{11}^2).$$

So,

$$\begin{aligned}|A_{21}| &\leq \tfrac{1}{2}(a_{00}^2 + a_{11}^2 + a_{01}^2 + a_{10}^2) + a_{00}a_{11} + a_{01}a_{10} \\ &\leq \tfrac{1}{2}(a_{00}^2 + a_{11}^2 + a_{01}^2 + a_{10}^2) + \tfrac{1}{2}(a_{00}^2 + a_{11}^2) + a_{01}a_{10} \\ &\leq (a_{00}^2 + a_{11}^2 + a_{01}^2 + a_{10}^2) + a_{01}a_{10} \\ &= \Gamma_{00} + \Gamma_{11}^1 \\ &= A_{22}.\end{aligned}$$

An examination of the last inequality above shows that if $|A_{21}| = A_{22}$, it is necessary that $a_{01} = a_{10} = 0$. But then $|A_{21}| = 0$, and to have $|A_{21}| = A_{22}$ would imply $A_{22} = 0$, in turn requiring $a_{00} = a_{11} = 0$, which is impossible.

Similar calculations yield $|C_{21}| < C_{22}$, $|A_{21}| < A_{11}$, and $|C_{21}| < C_{11}$. Since $\Delta_1 = A_{11}A_{22} - A_{21}^2$, $\Delta_2 = C_{11}C_{22} - C_{21}^2$, $\Delta_1 > 0$, $\Delta_2 > 0$.

References

[1] E. A. Robinson, *Statistical Communication and Detection*. New York: Hafner, 1967, pp. 173-174.

[2] J. L. Shanks, S. Treitel, and J. H. Justice, "Stability of synthesis of two-dimensional recursive filters," *IEEE Trans. Audio Electroacous.*, vol. AU-20, pp. 115-128, June 1972.

[3] Y. Genin and Y. Kamp, "Co-interexample in the least-squares inverse stabilization of 2 D recursive filters," *Electron. Lett.*, vol. 11, pp. 330-331, July 1975.

Comments on "On the Approximation Problem for Recursive Digital Filters with Arbitrary Attenuation Curve in the Pass-Band and the Stop-Band"

M. T. DOLAN

Abstract—**This correspondence contains a demonstration of the effectiveness and validity of the Fletcher-Powell optimization technique as opposed to the algorithm presented in the above paper by Dubois and Leich.**[1]

Dubois and Leich state in the above paper that "Methods based on the classical optimization procedures (example Fletcher-Powell) are very heavy and give results only for low degree filters. . ." (p. 206).[1] In contrast, our experience indicates that the Fletcher-Powell optimization technique is an extremely powerful tool; it has been used successfully by us on filters of order as high as 24 without difficulty.

An apparent drawback of the linear programming approach they describe is in the storage requirements for slack variables, artificial variables, and a double-precision scratch matrix necessary for inversion. Note that a Fletcher-Powell solution has no direct evaluation of an inverse matrix. Another weakness of their scheme is that the linear programming problem ". . .has feasible solutions if the specifications for phase 1 are well chosen." The Fletcher-Powell method is far more robust in this regard since it is capable of converging even if only a poor initial solution is known.

Summary of the Fletcher-Powell Optimization

The application to filter design is essentially that of Lasdon and Waren [1]. The procedure requires an initial design usually obtained through standard approximation techniques. The coefficients of the initial transfer function are to be varied such that certain performance requirements are satisfied. Specifications on the insertion-loss response in the frequency domain are described as a system of inequalities. These constraints are to be satisfied at a finite number of discrete frequencies, the philosophy being that a solution (not necessarily a global optimum) satisfies the specifications. The optimization is performed in the space of an augmented vector containing the coefficients, gain factor, and an additional variable which, when positive, represents the maximum amount by which the constraints are not satisfied, and when negative, represents the minimum amount by which the constraints are satisfied.

The actual solution of this nonlinear program is accomplished by a sequence of unconstrained minimizations through the use of a penalty function. A detailed treatment of this technique is given by Fiacco and McCormick [2]. A carefully chosen variation in step size and a quadratic interpolation produce a succession of minima in a direction of search which is determined by the Fletcher-Powell algorithm [3]. These minor iterations continue until two consecutive penalty function evaluations are within 0.01 percent of each other to complete a major iteration.

Note that there are no constraints on stability. Since we are dealing with insertion loss only, if a final design is unstable, it is made stable by inversion of the poles which lie inside the unit circle in the Z^{-1} plane.

Examples

The printed coefficients in the tables below correspond to the cascade realization:

Manuscript received September 9, 1975; revised April 2, 1976 and August 4, 1976.

The author is with Bell Laboratories, Murray Hill, NJ 07974.

[1] H. Dubois and H. Leich, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 202-207, Apr. 1975.