5) Every polynomial function $P(A)$ of a matrix A in G is itself a member of that same group. In particular, this is true of the exponential matrix $\exp(A)$. Moreover, if

$$\exp(A) = \sum_{i=0}^{N-1} \alpha_i K^i \tag{6}$$

then

$$\alpha_0 = \exp(a_0).$$

6) The determinant of A in G is given by

$$|A| = a_0^N. \tag{7}$$

A matrix A in G has N identical characteristic values

$$\lambda_i(A) = a_0, \qquad i = 1, \cdots, N. \tag{8}$$

7) All the preceding properties observed with respect to lower triangular matrices of the form (2) are equally valid when the matrix K is consistently replaced everywhere by its transpose K^T. In other words, there exists another similar (but different) group G_u of upper triangular matrices with precisely the same properties.

Uses

The preceding properties listed are useful in two ways.

1) A matrix $M \in G$ can be completely specified by a set of N numbers, m_i, $i = 0, 1, \ldots, N-1$. All the information contained in M is condensed in its first column. Thus matrix inversion as well as the evaluation of the transition matrix $\exp(At)$ can be reduced to the determination of a single first column of the inverse or of the transition matrix.

2) In time-varying systems with $A = A(t) \in G$ for all t, there is no difficulty in attaining an analytical solution. This is true by virtue of the commutativity property 2). It is indeed well known [2, p. 363] that when $A(t_1)\,A(t_2) = A(t_2)\,A(t_1)$ for all t_1, t_2, then the state transition matrix is given by

$$\phi(t, \tau) = \exp\left\{\int_\tau^t A(\lambda)\, d\lambda\right\}. \tag{9}$$

Conclusions

Matrices of the special triangular form (2) have some simple properties. These can be exploited in the simulation and control of processing plants such as chemical reactors that are structured as a train or sequence of similar processing steps with no recycle or other feedback paths. This is particularly useful in view of the time-varying situation often encountered in such plants.

References

[1] W. J. Cody, *SIAM Rev.*, vol. 12, p. 400, 1970.
[2] P. M. De Russo, R. J. Roy, and C. M. Close, *State Variables for Engineers*. New York: Wiley, 1967.
[3] C. E. Gall and R. Aris, *Can. J. Chem. Eng.*, vol. 43, p. 16, 1965.
[4] F. R. Gantmacher, *The Theory of Matrices*, vol. 1. New York: Chelsea Publ. Co., 1960.

Conditions for Minimax Approximation Obtained from the l_p Norm

J. W. BANDLER AND C. CHARALAMBOUS

***Abstract*—It is shown how the conditions for an optimal approximation in the minimax sense for general nonlinear functions can be obtained from the l_p norm. Applications in the optimal design of networks and systems are envisaged.**

Manuscript received July 20, 1971; revised October 6, 1971. This work was supported in part by the National Research Council of Canada under Grant A7239 and in part by a Frederick Gardner Cottrell grant from the Research Corporation.
The authors are with the Communications Research Laboratory, Department of Electrical Engineering, McMaster University, Hamilton, Ont., Canada.

Bandler [1] has recently derived necessary conditions for an optimal approximation in the minimax sense for general nonlinear approximating functions such as are encountered in network and system optimization problems. They were derived from the Kuhn–Tucker relations. It is well known that least pth approximation with a sufficiently large value of p can result in a solution very close to the minimax solution [2]–[4]. It is the purpose of this correspondence to show that necessary conditions for a minimax optimum can be fairly easily derived from the l_p norm. A note on sufficiency is included.

Consider an objective function of the form

$$U(\boldsymbol{\phi}) = \left\{\sum_{i\in I} |e_i(\boldsymbol{\phi})|^p\right\}^{1/p}, \qquad p \geq 1 \tag{1}$$

where $e_i(\boldsymbol{\phi})$, in general, represents a weighted error or deviation between a complex specified function (desired response) and a complex approximating function (actual response) at some sample point i of a finite set I, and $\boldsymbol{\phi}$ represents the k variable parameters. $U(\boldsymbol{\phi})$ is the norm $\|\boldsymbol{e}\|_p$. Minimization of $U(\boldsymbol{\phi})$ is called least pth approximation.

Theorem 1: At the optimum point $\check{\boldsymbol{\phi}}_\infty$ for a minimax approximation problem

$$\sum_{i\in J} u_i \frac{\operatorname{Re}\{e_i^*(\check{\boldsymbol{\phi}}_\infty)\nabla e_i(\check{\boldsymbol{\phi}}_\infty)\}}{|e_i(\check{\boldsymbol{\phi}}_\infty)|} = 0$$

$$\sum_{i\in J} u_i = 1, \qquad u_i \geq 0, \quad i \in J$$

where

$$\nabla \triangleq \begin{bmatrix} \dfrac{\partial}{\partial\phi_1} \\ \dfrac{\partial}{\partial\phi_2} \\ \cdot \\ \cdot \\ \cdot \\ \dfrac{\partial}{\partial\phi_k} \end{bmatrix} \tag{2}$$

and where it is assumed that $e_i(\boldsymbol{\phi})$ is continuous with continuous partial derivatives for all i, at least in the feasible region of $\boldsymbol{\phi}$. The asterisk denotes complex conjugate. Note that

$$\max |e_i(\boldsymbol{\phi})| = \lim_{p\to\infty} \left\{\sum_i |e_i(\boldsymbol{\phi})|^p\right\}^{1/p}, \qquad i \in I \tag{3}$$

and

$$J \triangleq \{i \,|\, |e_i(\check{\boldsymbol{\phi}}_\infty)| = \max |e_i(\check{\boldsymbol{\phi}}_\infty)|, \; i \in I\} \tag{4}$$

that is, J is the finite (nonempty) set of indices corresponding to the equal maxima of $|e_i(\check{\boldsymbol{\phi}}_\infty)|$.

Proof: Differentiating (1), for $p > 1$ and $U(\boldsymbol{\phi}) > 0$,

$$\nabla U(\boldsymbol{\phi}) = \|\boldsymbol{e}(\boldsymbol{\phi})\|_p^{1-p} \sum_{i\in I} |e_i(\boldsymbol{\phi})|^{p-2} \operatorname{Re}\{e_i^*(\boldsymbol{\phi})\nabla e_i(\boldsymbol{\phi})\}. \tag{5}$$

The necessary conditions for an optimum of $U(\boldsymbol{\phi})$ are that

$$\nabla U(\check{\boldsymbol{\phi}}_p) = \boldsymbol{0} \tag{6}$$

where $\check{\boldsymbol{\phi}}_p$ denotes the optimum parameter vector for particular values of p. Therefore, assuming $e_i(\check{\boldsymbol{\phi}}_p) \neq 0$,

$$\|\boldsymbol{e}(\check{\boldsymbol{\phi}}_p)\|_p \sum_{i\in I} \frac{|e_i(\check{\boldsymbol{\phi}}_p)|^p}{\|\boldsymbol{e}(\check{\boldsymbol{\phi}}_p)\|_p^p} \frac{\operatorname{Re}\{e_i^*(\check{\boldsymbol{\phi}}_p)\nabla e_i(\check{\boldsymbol{\phi}}_p)\}}{|e_i(\check{\boldsymbol{\phi}}_p)|^2} = 0. \tag{7}$$

Let

$$u_i = \lim_{p\to\infty} \left\{\frac{|e_i(\check{\boldsymbol{\phi}}_p)|}{\|\boldsymbol{e}(\check{\boldsymbol{\phi}}_p)\|_p}\right\}^p. \tag{8}$$

Then, it is clear that

$$u_i \begin{cases} = 0, & i \notin J \\ \geq 0, & i \in J \end{cases} \tag{9}$$

and, from the definition of $||e||_p$,

$$\sum_{i \in J} u_i = 1. \tag{10}$$

Therefore, (7) can be written, for $p \to \infty$,

$$\sum_{i \in J} u_i \frac{\operatorname{Re}\{e_i^*(\check{\phi}_\infty)\nabla e_i(\check{\phi}_\infty)\}}{|e_i(\check{\phi}_\infty)|} = 0 \tag{11}$$

where it is noted that the $|e_i(\check{\phi}_\infty)|$ for $i \in J$ are all equal to $||e(\check{\phi}_\infty)||_\infty$.

Theorem 2: If the relations in Theorem 1 are satisfied at a point $\check{\phi}_\infty$ and the $|e_i(\phi)|$ for $i \in I$ are convex, then $\check{\phi}_\infty$ is optimal.

If the appropriate $|e_i(\phi)|$ are convex, then it is relatively straightforward to prove that $||e(\phi)||_p$ is convex for $p \geq 1$, so that $\check{\phi}_p$ locates a minimum.

It is felt that the ideas discussed here could lead to a better understanding of the relationship between least pth and minimax approximations for network and system design problems. Note that no assumptions concerning the number of equal maxima of $|e_i(\check{\phi}_\infty)|$ compared with the number of parameters has been made. Furthermore, (8) relating u_i to $|e_i(\check{\phi}_\infty)|$ is not immediately obvious from a$_n$ application of the Kuhn–Tucker relations [1].

ACKNOWLEDGMENT

The authors wish to thank R. E. Seviora of the Department of Electrical Engineering, University of Toronto, Toronto, Ont., Canada, whose early suggestions proved to be a turning point in this work.

REFERENCES

[1] J. W. Bandler, "Conditions for a minimax optimum," *IEEE Trans. Circuit Theory* (Corresp.), vol. CT-18, pp. 476–479, July 1971.
[2] G. C. Temes and D. Y. F. Zai, "Least pth approximation," *IEEE Trans. Circuit Theory* (Corresp.), vol. CT-16, pp. 235–237, May 1969.
[3] J. W. Bandler, "Optimization methods for computer-aided design," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-17, pp. 533–552, Aug. 1969.
[4] R. Fletcher, J. A. Grant, and M. D. Hebden, "The calculation of linear best L_p approximations," *Comput. J.*, vol. 14, pp. 276–279, Aug. 1971.

Iterative Solution of the Riccati Equation

KAREL VIT

***Abstract*—A new convergence proof of an iterative method for the steady solution of the Riccati equation is presented and its geometric nature and close resemblance to the Newton method are emphasized. Uniqueness, rate of convergence and initialization of the iterative process are discussed.**

Following the paper by Kleinman [1], the process of computing the steady solution of the Riccati equation reduces to one of allocating the positive definite solution, if it exists, among all symmetric K satisfying the system of nonlinear algebraic equations

$$0 = KA + A^TK + C^TC - KNK \tag{1}$$

where $N = BR^{-1}B^T$. We assume that R is positive definite, (A, B) and (A, C) being completely controllable and observable pairs, respectively, A being a stable matrix. All matrices are real, square, and of equal order. Under these conditions the positive definite solution $K > 0$ of (1) is known to exist and to be unique.

Manuscript received August 13, 1971.
The author is with the Department of Electrical Engineering, University of Salford, Salford, England.

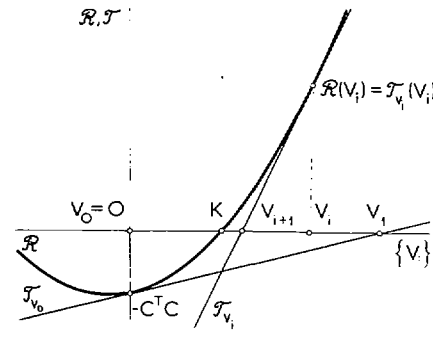

Fig. 1. Scalar case.

In order to avoid the use of a concept of a cost matrix and of matrix exponentials in the time domain as in [1], we introduce two symmetric operators mapping the space of all real symmetric matrices into and onto itself respectively. These are

$$\mathcal{R}(V) = -A^TV - VA + VNV - C^TC \tag{2}$$

$$\mathcal{J}_F(V) = (NF - A)^TV + V(NF - A) - FNF - C^TC \tag{3}$$

where F is a symmetric matrix. The solution of (1) is clearly equivalent to the solution of $\mathcal{R}(K) = 0$. $\mathcal{J}_F$ is easily shown to be a support of $\mathcal{R}$ at F. In such a way a simple and interesting geometric interpretation of (2) and (3) is introduced with its strong analogy to a scalar case in which $\mathcal{J}$ becomes a tangent to $\mathcal{R}$. The solution of a nonlinear system $\mathcal{R}(\cdot) = 0$ is replaced by a sequence of linear problems $\mathcal{J}(\cdot) = 0$ which are much easier to solve numerically.

For an easier understanding of the iterative process to be derived, a graphical interpretation of a scalar case presented in Fig. 1 may be referred to. The Lyapunov theorem will be used throughout without being explicitly mentioned.

Lemma 1: Let Y be a symmetric matrix such that $A - NY$ and $NY - A$ have no eigenvalues in common and let V be the solution of $\mathcal{J}_Y(V) = 0$. Then $\mathcal{R}(V) \geq 0$. If for the above condition also $V > 0$, then $A - NV$ is a stable matrix.

Proof: Due to the assumption, the solution of $\mathcal{J}_Y(V) = 0$ exists and is unique. Use of (2) and (3) together with $\mathcal{J}_Y(V) = 0$ gives

$$\mathcal{R}(V) = (Y - V)N(Y - V) \geq 0. \tag{4}$$

Directly from the definition (2)

$$\mathcal{R}(V) = (NV - A)^TV + V(NV - A) - VNV - C^TC. \tag{5}$$

From (4) and (5) one obtains

$$(A - NV)^TV + V(A - NV) = -(Y - V)N(Y - V) - VNV - C^TC \tag{6}$$

and the assumption $V > 0$ implies the stability of $A - NV$.

Lemma 2: Let $\{V_i\}$, $i = 0, 1, 2 \cdots$, be a sequence of symmetric matrices generated by $\mathcal{J}_{V_i}(V_{i+1}) = 0$, $V_0 = 0$. Then $V_{i+1} \leq V_i$, $i = 1, 2, \cdots$, and $V_i > 0$ for each $V_i \in \{V_i\}$.

Proof: Let us proceed by induction. For $i = 0$ we have $V_0 = 0$ and (3) reduces to $A^TV_1 + V_1A = -C^TC$. Since $V_1 > 0$ the matrix $A - NV_1$ is stable by Lemma 1. Assume now V_i to be the solution of $\mathcal{J}_{V_{i-1}}(V_i) = 0$ and $A - NV_i$ to be stable. Then $V_{i+1} > 0$ is the solution of

$$(A - NV_i)^TV_{i+1} + V_{i+1}(A - NV_i) = -V_iNV_i - C^TC \tag{7}$$

and, by Lemma 1, $A - NV_{i+1}$ is a stable matrix. In order to show that $\{V_i\}$ is decreasing we introduce $D_i = V_i - V_{i+1}$. From (2) and (3) follows