COMPUTATIONAL MATHEMATICS
BANACH CENTER PUBLICATIONS, VOLUME 13
PWN – POLISH SCIENTIFIC PUBLISHERS
WARSAW 1984

NUMERICAL METHODS IN SYSTEM DESIGN AND IDENTIFICATION WITH APPLICATION TO WAVE PROPAGATION IN LAYERED MEDIA

HANS KAISER

Sektion Mathematik/Physik, Pädagogische Hochschule "Karl Liebknecht", Am Neuen Palais, 15 Potsdam, DDR

and

HANS-CHRISTOPH KAISER

Institut für Mathematik, Akademie der Wissenschaften, Mohrenstr. 39, 108 Berlin, DDR

1. Introduction

We consider transfer systems Σ which convert input values x_i, $i = 1, \ldots, n$, into output values y_j, $j = 1, \ldots, l$. Each y_j is a function of $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ and certain quantities characterizing the structure of Σ. By *system identification* we mean determination of the structure of an actual system Σ from its behaviour, viz. by means of y_j-measurements for certain input vectors $\boldsymbol{x} \in X$. This will be effected within a system class $\mathfrak{S}$, the choice of which has the character of a recognition hypothesis. Within the same framework the problem of system design or synthesis is to be solved, which consists in determining a system which best performs a prescribed behaviour on an input set X. In this case the chosen system class $\mathfrak{S}$ represents the technical conditions for a device realizing the input-ouput relation. One is mainly interested in the three questions:

(i) Can the problem be solved within $\mathfrak{S}$?

(ii) Is the solution uniquely determined within $\mathfrak{S}$?

(iii) Can an efficient algorithm be found for constructing solutions?

A simple example shall be given for illustration. As input-ouput system Σ consider the plane three-bar linkage shown in Figure 1. The structure of Σ is constituted by the lengths of the bars and the distance between the fixed points A_1 and A_2, viz. by the values a, b, c (see Fig. 1). The rotation angle ϑ is to be taken as input; outputs be the Cartesian coordinates y_1 and y_2 of M.

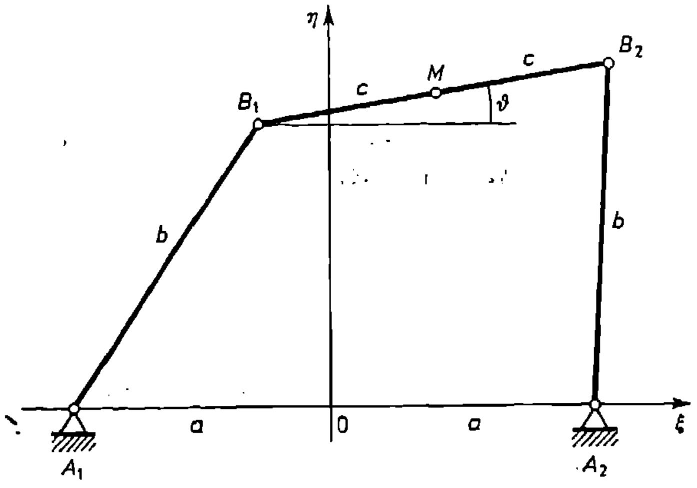

Fig. 1

Given ϑ and the structure parameters a, b, c, one derives

$$
(1)\qquad
\begin{aligned}
y_1 &= c \sin\vartheta \sqrt{\frac{b^2 - a^2 - c^2 + 2ac\cos\vartheta}{a^2 + c^2 - 2ac\cos\vartheta}},\\
y_2 &= (a - c\cos\vartheta)\sqrt{\frac{b^2 - a^2 - c^2 + 2ac\cos\vartheta}{a^2 + c^2 - 2ac\cos\vartheta}}.
\end{aligned}
$$

The system identification within the three-parameter class $\mathfrak{S}_{TB}$ of plane three-bars requires the determination of a, b, c using the information contained in a trajectorial segment of M, viz. the coordinate data of its points. A technically relevant synthesis problem might refer to the construction of a mechanism conducting M approximately along a straight line segment. If the identification problem within $\mathfrak{S}_{TB}$ can be solved, the solution is not uniquely determined. There are three mechanisms of the kind considered, all generating the same three-bar curve. This follows from a theorem due to Roberts concerning a more general type of three-bars, where the bar B_1B_2 is substituted by a triangle B_1B_2M (cf. Fig. 2). In Figure 2 one draws the parallelograms $A_1B_1MC_1$ and $A_2B_2MD_1$ and constructs over the sides C_1M and MD_1 two triangles similar to B_1B_2M. Finally the parallelogram $MD_2A_3C_2$ is drawn. It is easy to show that during any motion of the linkage corresponding to Figure 2 (which is a system with one degree of freedom) the point A_3 remains in the same position. Therefore this point can be fixed and one gets the three-bars

$A_1A_2B_2B_1$, $A_2A_3D_2D_1$, $A_1A_3C_2C_1$, all generating the same trajectory. This subject has been treated by Roberts [1] and Cayley [2].

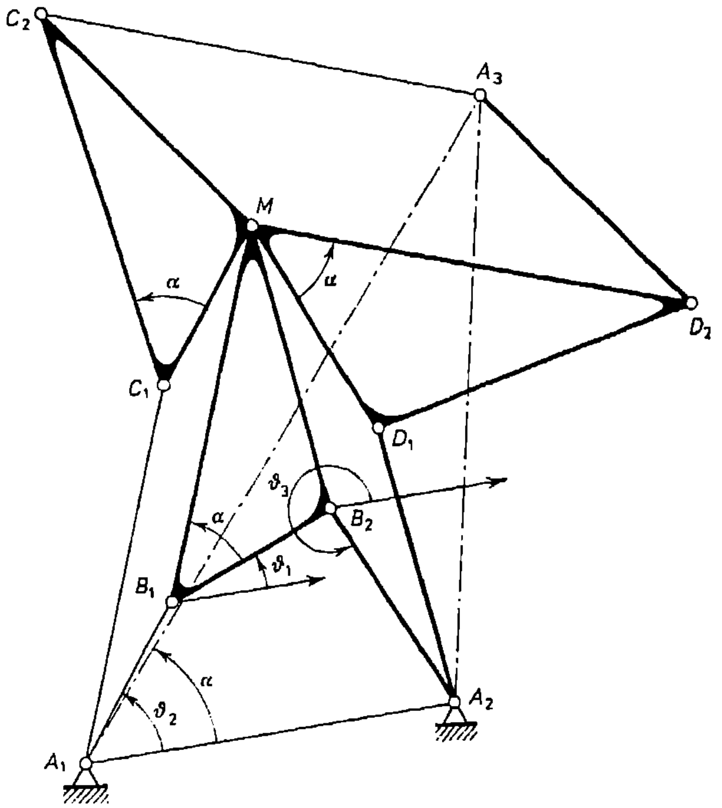

Fig. 2

From a general point of view one may observe the following methodological trends in dealing with system identification and design:

(i) The identification problem is reduced to an inverse problem for a differential equation (cf. Section 2).

(ii) When looking for a solution within a prescribed m-parametric system class

$$\mathfrak{S}_A = \{\Sigma(\boldsymbol{a}) : \boldsymbol{a} \in A\}, \quad A \subseteq \boldsymbol{R}^m, \tag{2}$$

one is concerned with a problem of classical approximation theory, as the following considerations will make evident.

Just like in the example above, $\boldsymbol{y} = (y_1, \ldots, y_l)^T$ is represented by a function

$$\boldsymbol{y} = \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{a}) = \boldsymbol{F}_{\boldsymbol{a}}(\boldsymbol{x}), \tag{3}$$

depending on a vector parameter $\boldsymbol{a}$ and the input vector $\boldsymbol{x}$. The function $\boldsymbol{F}$ is assumed to be known from a theory about the system class

$\mathfrak{S}_A$. The measurements or the desired behaviour shall be given as the function

$$f\colon X \to \boldsymbol{R}^l, \tag{4}$$

where X denotes a certain set of input vectors. Then the identification or synthesis requires the determination of a parameter vector $\boldsymbol{a}^* \in A$ such that the merit function

$$Z(\boldsymbol{a}) := \|\boldsymbol{f} - \boldsymbol{F}_{\boldsymbol{a}}\| \tag{5}$$

constructed with a suitably chosen norm $\|\cdot\|$ attains minimum at $\boldsymbol{a} = \boldsymbol{a}^*$:

$$Z(\boldsymbol{a}^*) = \min_{\boldsymbol{a} \in A} Z(\boldsymbol{a}); \tag{6}$$

in other words, one has to determine the best approximation of $\boldsymbol{f}$ within the family of functions

$$\{\boldsymbol{F}_{\boldsymbol{a}}\colon \boldsymbol{a} \in A\}. \tag{7}$$

In general, minimization of the function Z may be achieved by applying search procedures of nonlinear optimization. We refrain from a detailed description of these methods, which — as far as universally employable — often prove rather inefficient on account of their loose connection with the inherent properties of the problem.

An immediate approach to the best approximation is offered by the use of criteria characterizing $\boldsymbol{F}_{\boldsymbol{a}^*}$ for special problems (6). Many practical tasks require, for instance, uniform approximation; then Chebyshev's theorem can be viewed as such a criterion (cf. Section 3).

In the case where the system behaviour is described by differential equations containing the parameter vector $\boldsymbol{a}$, Σ can be detected within $\mathfrak{S}_A$ by means of quasilinearization. This is an iterative technique developed by Bellman, Kalaba and others (cf. Section 4).

In Section 5 we discuss methods of pattern recognition applied to the components of the output $\boldsymbol{y} = \boldsymbol{F}_{\boldsymbol{a}}(\boldsymbol{x})$ with regard to relevant shape phenomena. They are usually employed together with a discretization of the parameter space A by a finite set $\tilde{A}$. Recently effective algorithms have been developed, some of them based upon linguistic methods, for a complete scanning of the graphs $\boldsymbol{F}_{\boldsymbol{a}}$ corresponding to $\tilde{A}$.

We shall consider these methods mainly with regard to the propagation of electromagnetic waves in stratified media (cf. [3], [4], [5]). In this context, a medium will be called stratified, if the refractive index is constant on each plane orthogonal to a fixed spatial direction. Choosing this one as z-axis of the Cartesian coordinate system, the structure of such a layered medium is essentially determined by the refractive index profile $n = n(z)$. In order to influence the energy flow and polarization of electromagnetic waves, our interest concentrates on systems Σ as represented in Figure 3, where a layered medium extending in $0 \leqslant z \leqslant d$ is placed between two homogeneous media of refractive indices n_0 and n_g, respectively. Throughout this paper n_0, $n(z)$, n_g are assumed to be real valued, viz. the stratified medium non absorbing.

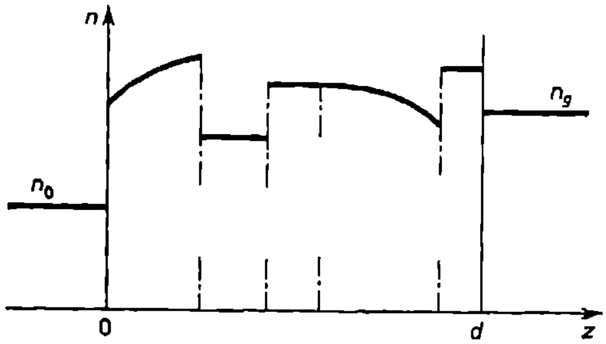

Fig. 3

Consider a plane electromagnetic wave incident upon the y, z-plane under the angle ϑ_0; due to the influence of the layered medium the wave will be split into a reflected one and a transmitted one.

The wavelength λ and/or the angle of incidence ϑ_0 are the input values of the transfer system Σ; one might take as outputs the reflectivity or transmittance, viz. the part of incoming energy transported by the reflected and the transmitted wave, respectively. The design or identification is then concerned with the determination of a refractive index profile.

2. An inverse problem for the wave equation and the Schrödinger equation

The problem stated at the end of Section 1 is regarded for the special case of a normally incident TE-wave (electric vector perpendicular to the plane of incidence), where the inhomogeneous medium described by $n(z)$ extends over the half space $z > 0$. Then, in $z > 0$, the x-component of the electric vector may be written as

$$(1) \qquad E_x = U(z, k)e^{i\omega t}, \qquad k = 2\pi/\lambda = \omega/c,$$

(λ wavelength, c light speed in the vacuum), where U is a solution of the wave equation

$$(2) \qquad \frac{d^2U}{dz^2} + k^2 n^2(z)\, U = 0.$$

Usually this equation is to be solved for a given function $n(z)$. An inverse problem related to (2) requires the determination of $n(z)$ from the spectral characteristic of a functional of U corresponding to an observable physical quantity. In most cases this is the reflection coefficient $r = r(k)$, as e.g. in the following considerations. r is defined as the ratio of the x-components of the electric vector of the reflected and incident wave, respectively, at the interface $z = 0$. An approximate solution for the inverse problem will be constructed, assuming n to be a continuously differentiable function for $z \geqslant 0$, $n_0 = n(0)$ being the refractive index in $z < 0$. The solution of (2), which represents the field induced by the incident wave, is given in $z > 0$ by the Bremmer series

(3) $$U(z, k) = \sum_{j=0}^{\infty} U_j(z, k)$$

(cf. [6], [7]). One arrives at this result by a decomposition of the field in $z > 0$ into partial waves; $U_j(z, k)$ denotes the amplitude composed of all contributions from waves undergoing exactly j reflections at interfaces lying in $z > 0$. Introducing the optical thickness

(4) $$x = x(z) = \int_0^z n(s)\, ds$$

one gets

(5a) $$U_0(x, k) = \sqrt{\frac{n(0)}{n(x)}}\, e^{-ikx},$$

(5b) $$U_{2j}(x, k) = \frac{1}{\sqrt{n(x)}} \int_0^x U_{2j-1}(s, k)\, e^{-ik(x-s)} \frac{d\sqrt{n}}{ds}\, ds,$$

(5c) $$U_{2j+1}(x, k) = \frac{-1}{\sqrt{n(x)}} \int_x^{\infty} U_{2j}(s, k)\, e^{-ik(s-x)} \frac{d\sqrt{n}}{ds}\, ds.$$

It can be shown that the Bremmer series converges towards a solution of the wave equation (2), if

(6) $$|n(z)| \geqslant a > 0 \quad \text{for all } z \geqslant 0, \qquad \int_0^{\infty} |n'(z)|\, dz = b, \quad b < 2a.$$

One obtains for the first derivative of the Bremmer series the formula

$$\frac{dU}{dz} = ikn(z)(W-V), \tag{7}$$

where

$$V = \sum_{j=0}^{\infty} U_{2j}, \quad W = \sum_{j=0}^{\infty} U_{2j+1}. \tag{8}$$

The series V and W satisfy the integral equations

$$V(x, k) = U_0(x, k) + \frac{1}{\sqrt{n(x)}} \int_0^x \frac{d\sqrt{n}}{ds} W(s, k) e^{-ik(x-s)} ds; \tag{9}$$

$$W(x, k) = \frac{-1}{\sqrt{n(x)}} \int_x^{\infty} \frac{d\sqrt{n}}{ds} V(s, k) e^{-ik(s-x)} ds. \tag{10}$$

From (7) one derives for the y-component of the magnetic vector (magnetic permeability $\mu \equiv 1$) the formula

$$H_y = n(V-W) e^{i\omega t}. \tag{11}$$

Let A and B be the amplitude quantities of the incident and reflected wave, respectively. Then, by the continuity of the tangent component of the electric and magnetic vector at the interface $z = 0$, one gets

$$A+B = U(0, k) = 1+W(0, k), \tag{12}$$

$$A-B = 1-W(0, k). \tag{13}$$

From this follows

$$r(k) = W(0, k) = -\frac{1}{\sqrt{n(0)}} \int_0^{\infty} \frac{d\sqrt{n}}{dx} V(x, k) e^{-ikx} dx. \tag{14}$$

In particular, (14) implies

$$r(-k) = r^*(k). \tag{15}$$

Taking the first term in the series defining V one gets from (14) the approximate equality

$$r(k) = -\frac{1}{2} \int_0^{\infty} \frac{d}{dx} (\ln n) e^{-2ikx} dx. \tag{16}$$

Applying Fourier transformation to the real and imaginary part of this equation we get

$$\frac{2}{\pi}\int_{-\infty}^{\infty} r(k)e^{2ikx}\,dk = \begin{cases} \dfrac{d}{dx}\ln n(-x) & \text{for } x<0,\\ 0 & \text{for } x \geqslant 0; \end{cases} \tag{17}$$

that means,

$$\sqrt{\frac{n(x)}{n(0)}} = \exp\Big(-\int_{-\infty}^{2x} F(s)\,ds\Big), \qquad x \geqslant 0, \tag{18}$$

where

$$F(s) = \frac{1}{2\pi}\int_{-\infty}^{\infty} r(k)e^{-iks}\,dk. \tag{19}$$

There exists a well known correspondence between the wave equation and the Schrödinger equation. With a solution U of (2) we define the function

$$\Psi(x,k) = \sqrt{n(x)}\,U(x,k), \tag{20}$$

depending on the optical thickness (4) and the wave number k. Then Ψ satisfies the Schrödinger equation

$$-\frac{d^2\Psi}{dx^2} + V(x)\Psi = k^2\Psi \tag{21}$$

with the scattering potential

$$V(x) = \frac{1}{\sqrt{n(x)}}\frac{d^2}{dx^2}\sqrt{n(x)}. \tag{22}$$

Thus one obtains from the approximate solution (18) of the inverse problem associated with the wave equation the formula

$$V(x) = -2\frac{d}{dx}F(2x) + 4F^2(2x). \tag{23}$$

This approximate solution of the inverse problem for the Schrödinger equation has been determined by Moses [8] via the Gelfand–Levitan integral equation. In [9] (18) has been deduced using the relation (23).

3. Chebyshev methods in system design

From its beginning the development of approximation theory has been associated with the application to system design. The fundamentals of uniform approximation have been laid by P. L. Chebyshev in the second half of the 19^{th} century when he was concerned with the design of bar-linkages realizing certain prescribed trajectories (cf. [10], [11]). The basic idea shall be illustrated by the previously regarded three-bar linkage of Figure 1. Eliminating the input ϑ in (1.1) one derives for the trajectory of the point M the algebraic equation

$$\zeta^3+(2r+y_2^2)\zeta^2+\big(r^2+2y_2^2(r+2a^2)\big)\zeta+y_2^2\big(r^2+4a^2(y_2^2-b^2)\big) = 0, \tag{1}$$

where

$$\zeta = y_1^2, \qquad r = y_2^2-a^2+c^2-b^2. \tag{2}$$

The synthesis aims at the construction of a mechanism that forces the point M to move along a path parallel to the ξ-axis. Let $[\alpha, \beta]$ be the corresponding interval for ζ and H the ordinate of the points of the parallel segment. According to the desired behaviour of the three-bar, the polynomial

$$\begin{aligned} P_3(\zeta) = \zeta^3+(2r \quad +H^2)\zeta^2+\big(2H^2(r_H+2a^2)+r_H^2\big)\zeta+ \\ +H^2\big(r_H^2+4a^2(H^2-b^2)\big), \\ r_H = H^2-a^2+c^2-b^2, \end{aligned} \tag{3}$$

is required to realize the least possible deviation from zero on the interval $[\alpha,\beta]$, i.e. P_3 is to coincide with the transformed Chebyshev polynomial

$$\tilde{T}_3(\zeta) = \frac{(\beta-\alpha)^3}{2^5}T_3\left(\frac{2\zeta-(\beta+\alpha)}{\beta-\alpha}\right) = \zeta^3+p_2\zeta^2+p_1\zeta+p_0 \tag{4}$$

with the coefficients

$$\begin{aligned} p_2(\alpha,\beta) &= -\tfrac{3}{2}(\alpha+\beta), \\ p_1(\alpha,\beta) &= \tfrac{3}{16}(\alpha+3\beta)(\beta+3\alpha), \\ p_0(\alpha,\beta) &= -\tfrac{1}{32}(\alpha+\beta)(\alpha^2+\beta^2+14\alpha\beta), \end{aligned} \tag{5}$$

derived from the Chebyshev polynomial of degree three

$$T_3(t) = 4t^3-3t. \tag{6}$$

The maximum deviation from zero of the polynomial (4) on $[\alpha, \beta]$ is given by

$$L = (\beta - \alpha)^3/2^5. \tag{7}$$

Comparing the coefficients in (3) and (4) one obtains

$$\begin{aligned} 2r_H + H^2 &= -\tfrac{3}{2}(\alpha+\beta), \\ r_H^2 + 2H^2(r_H + 2a^2) &= \tfrac{3}{16}(\alpha+3\beta)(\beta+3\alpha), \\ \big(r_H^2 + 4a^2(H^2 - b^2)\big)H^2 &= -\tfrac{1}{32}(\alpha+\beta)(\alpha^2+\beta^2+14\alpha\beta). \end{aligned} \tag{8}$$

Elimination of α, β gives the so-called basic equation

$$27p_0 = p_2(9p_1 - 2p_2^2), \tag{9}$$

from which restrictions for the construction of the required system can be derived. Substituting the left-hand sides of (8) for the p_j into (9) one gets

$$H^2 = \frac{(a^2+b^2-c^2)^3}{18a^2(2c^2+b^2-2a^2)}. \tag{10}$$

Thus H is characterized as a function of the system parameters a, b, c, and it becomes evident that the problem has no solution if

$$\frac{a^2+b^2-c^2}{2c^2+b^2-2a^2} < 0. \tag{11}$$

On account of symmetry, the case $\alpha = 0$ is of particular interest; there (8) takes the form

$$\begin{aligned} 2r_H + H^2 &= -\tfrac{3}{2}\beta, \\ r_H^2 + 2H^2(r_H + 2a^2) &= \tfrac{9}{16}\beta^2, \\ H^2\big(r_H^2 + 4a^2(H^2 - b^2)\big) &= -\tfrac{1}{32}\beta^3. \end{aligned} \tag{12}$$

Approximation of the straight line $\eta = H$ is to be effected on an interval of the ξ-axis of length $l = 2\sqrt{\beta}$, and, considering (7), the maximum deviation from zero of the polynomial P_3 is

$$L = \frac{\beta^3}{32} = \frac{l^6}{2048}. \tag{13}$$

Let L and H be prescribed; then one has to compute β from (13), and the synthesis problem will be solved by successively determining r_H, a, b from (12) and finally c from (3b).

The above exemplified method using Chebyshev's polynomials can be applied likewise to various other system classes. The application may be extended to linear combinations of Chebyshev systems (cf. [10]).

For further illustration of the method we consider the design of achromatic systems, consisting of a stack of S homogeneous dielectric layers with equal optical thickness

$$\Delta = n_j h_j, \quad j = 1, \ldots, S \tag{14}$$

(h_j geometrical thickness, n_j refractive index of the jth layer); we assume normal incidence and given refractive indices n_0 and $n_{S+1} = n_g$ of the enclosing media, i.e. we suppose that the refractive index profile represented in Figure 3 is a special step function. Then in (1.2) we have $m = S$, and the parameters a_j are given by the refractive indices n_j of the layers. The output be the reciprocal transmittance T^{-1} depending from the so-called phase angle φ which is related to the input λ according to

$$\varphi = \frac{2\pi}{\lambda} \Delta . \tag{15}$$

One can show that $T^{-1}(\varphi)$ is the trigonometric polynomial

$$T^{-1}(\varphi) = \sum_{j=0}^{S} B_j \cos 2j\varphi \tag{16}$$

whose coefficients are rational functions in the refractive indices $n_1, \ldots, n_S$ or Fresnel's interface coefficients

$$r_j = \frac{n_{j-1} - n_j}{n_{j-1} + n_j}, \quad j = 1, \ldots, S+1, \tag{17}$$

respectively. Numerous algorithms have been developed for the computation of the B_j (cf. [12], [13]). (16) can be transformed into the algebraic polynomial

$$T^{-1}(\zeta) = \sum_{j=0}^{S} A_j \zeta^j \tag{18}$$

in

$$\zeta = \cos^2 \varphi . \tag{19}$$

The construction of an achromatic system then implies the determination of $n_1, \ldots, n_S$ such that T^{-1} is approximately constant over a certain interval $[\alpha, \beta]$ of the ζ-axis. The approximation problem (1.6) is to be solved with regard to

$$F = T^{-1}, \tag{20}$$

$$f(\zeta) = H, \quad \alpha \leqslant \zeta \leqslant \beta. \tag{21}$$

Practical considerations require uniform approximation. Following the same way as in three-bar design, one has to compare the polynomial

$$\frac{T^{-1}(\zeta)-H}{A_S} = \zeta^S + \frac{A_{S-1}}{A_S}\zeta^{S-1} + \dots + \frac{A_1}{A_S}\zeta + \frac{A_0-H}{A_S} \tag{22}$$

with the transformed Chebyshev polynomial

$$\begin{aligned}\tilde{T}_S(\zeta) &= \frac{(\beta-\alpha)^S}{2^{2S-1}}\, T_S\left(\frac{2\zeta-(\alpha+\beta)}{\beta-\alpha}\right) \\ &= \zeta^S + p_{S-1}(\alpha, \beta)\zeta^{S-1} + \dots + p_0(\alpha, \beta).\end{aligned} \tag{23}$$

By means of the differential equation for the Chebyshev polynomials one gets a recursion formula for the p_j; explicitly, one derives the following expressions:

$$\begin{aligned} p_j &= (-1)^{S-j}\binom{S}{j}\left(\frac{\alpha+\beta}{2}\right)^{S-j} * \\ &* \sum_{i=0}^{\left[\frac{S-j}{2}\right]} (-1)^i 2^{-2i}\left(\frac{\beta-\alpha}{\alpha+\beta}\right)^{2i} \frac{\binom{S-j}{2i}}{\binom{S-1}{i}}\binom{2i}{i}, \quad j = 0, 1, \dots, S-1;\end{aligned} \tag{24a}$$

in particular, for $\alpha = 0$ one gets

$$p_j = (-1)^{S-j}\binom{S}{j}\left(\frac{\beta}{2}\right)^{S-j} \frac{(2S-3)(2S-5)\dots(2j+1)}{(2S-2)(2S-3)\dots(S+j)}. \tag{24b}$$

Comparing the coefficients in (22) and (23) provides as equations for thin-film synthesis

$$\begin{aligned} \frac{A_0-H}{A_S} &= p_0(\alpha, \beta), \\ \frac{A_1}{A_S} &= p_1(\alpha, \beta), \\ &\dots\dots\dots\dots \\ \frac{A_{S-1}}{A_S} &= p_{S-1}(\alpha, \beta). \end{aligned} \tag{25}$$

Systems with refractive indices satisfying the equations (25) for certain values H, α, β shall be called "Chebyshev stacks". They are characterized by their realizing best approximation of the H-plateau on a frequency interval corresponding to $\alpha \leqslant \zeta \leqslant \beta$, as compared with any other system of the class described. The maximum deviation L from H is given by

$$(26)\qquad L = \frac{(\beta-\alpha)^S}{2^{2S-1}}\,|A_S| = (\beta-\alpha)^S|B_S| = 2(\beta-\alpha)^S|r_1 r_{S+1}|\prod_{j=1}^{S+1}\frac{1}{1-r_j^2}.$$

As regards elimination of the unknown parameters α, β on the right-hand side of (25), the following should be noted: Using two of the equations (25), one can eliminate α and β in the others Thus one obtains from (25) a system of $S-2$ basic equations

$$(27)\qquad \begin{aligned} g_1(n_1,\dots,n_S) &= 0,\\ &\cdots\cdots\cdots\\ g_{S-2}(n_1,\dots,n_S) &= 0. \end{aligned}$$

As in the case of the three-bar design, the assumption $\alpha = 0$ is of particular interest on account of symmetry. Eliminating β in (25) one obtains $S-1$ basic equations. β is related to the maximal deviation (26) over the achromaticity interval according to

$$(28)\qquad \beta = \left(\frac{L}{|B_S|}\right)^{1/S}.$$

As an example we consider the synthesis of a two-layer antireflection system. For this purpose, let $H = 1.01$ and $L = 0.01$ be prescribed. A numerical solution of (25), observing the relation (28), yields

$$(29)\qquad n_1 = 1.18878, \quad n_2 = 1.26450.$$

Figure 4 shows the transmittance of this system in the φ-plot.

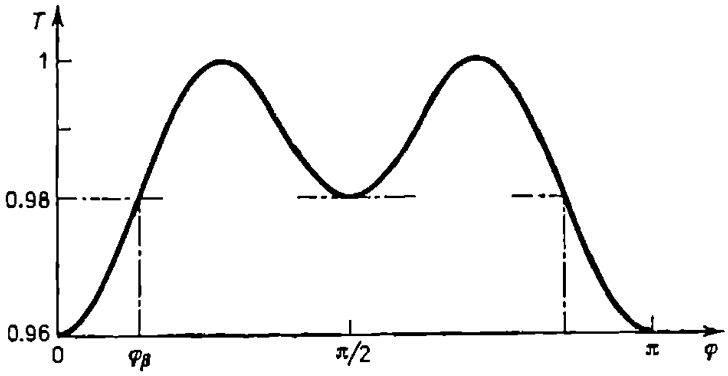

Fig. 4

Further investigations are directed towards general theorems on Chebyshev stacks providing restrictions for the construction and the characteristics of the systems' performance.

4. Quasilinearization

In this section it is assumed that the behaviour of the system is described by an initial value problem which contains as parameters the structure quantities $a_1, \ldots, a_m$:

$$\text{(1a)} \qquad \boldsymbol{w}' = \boldsymbol{g}(z, \boldsymbol{w}, \boldsymbol{a}), \qquad 0 \leqslant z \leqslant d,$$

$$\text{(1b)} \qquad \boldsymbol{w}(d) = \boldsymbol{c}(\boldsymbol{a}),$$

$$\text{(1c)} \qquad \boldsymbol{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_L \end{bmatrix}, \quad \boldsymbol{g} = \begin{bmatrix} g_1 \\ \vdots \\ g_L \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_L \end{bmatrix}.$$

The variables $w_1, \ldots, w_l$, $l \leqslant L$, are related to observable outputs provided by measurements (or prescribed values)

$$\text{(2)} \qquad \boldsymbol{Q} = (Q_1, \ldots, Q_l)^T,$$

which constitute the function $\boldsymbol{f}$ in (1.4). Quasilinearization is an iterative technique for the identification of the parameter vector $\boldsymbol{a}$ corresponding to the measurements (2). In each step one has to solve an approximation problem within the solution space of a system of linear differential equations related to (1). The procedure presents a certain analogy to Newton's method for the solution of nonlinear equations. It is based on the following transformation of the initial value problem (1): The parameter vector $\boldsymbol{a}$ which realizes best approximation to the measurements (2) is preliminarily regarded as a solution of the initial value problem

$$\text{(3)} \qquad \boldsymbol{a}' = \boldsymbol{0}, \qquad \boldsymbol{a}(d) = \boldsymbol{a}.$$

Together with (1) one obtains the enlarged initial value problem

$$\text{(4)} \qquad \boldsymbol{W}' = \boldsymbol{G}(z, \boldsymbol{W}), \qquad \boldsymbol{W}(d) = \boldsymbol{C}(\boldsymbol{a}),$$

where

$$\text{(5)} \qquad \boldsymbol{W}(z) = \begin{bmatrix} \boldsymbol{w}(z) \\ \boldsymbol{a}(z) \end{bmatrix}, \quad \boldsymbol{G}(z, \boldsymbol{W}) = \begin{bmatrix} \boldsymbol{g}(z, \boldsymbol{w}, \boldsymbol{a}) \\ \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{C}(\boldsymbol{a}) = \begin{bmatrix} \boldsymbol{c}(\boldsymbol{a}) \\ \boldsymbol{a} \end{bmatrix}.$$

The parameter adjustment is effected by constructing a sequence $\boldsymbol{W}^k(z)$, $k = 1, 2, \ldots$, for which $\boldsymbol{W}^k(0)$ converges, tending towards $\boldsymbol{Q}$ at the first l positions.

The first step is the choice of an initial parameter vector $\boldsymbol{a}^0$ and the determination of the solution of (4) subject to the initial condition $\boldsymbol{W}(d) = \boldsymbol{C}(\boldsymbol{a}^0)$ associated with $\boldsymbol{a}^0$. Starting with $\boldsymbol{W}^k$ one obtains $\boldsymbol{W}^{k+1}$ by solving the following problems:

(a) Linearization of the differential equation (4) at the point $\boldsymbol{W}^k$,

$$\boldsymbol{W}' = \boldsymbol{G}(z, \boldsymbol{W}^k) + \boldsymbol{J}(z, \boldsymbol{W}^k)(\boldsymbol{W} - \boldsymbol{W}^k), \tag{6}$$

where

$$\boldsymbol{J} = (J_{ij}) = \left(\frac{\partial G_i}{\partial W_j}\right) \tag{7}$$

is the Jacobian of $\boldsymbol{G}$.

(b) Determination of the general solution

$$\boldsymbol{W}(z) = \boldsymbol{P}(z) + \boldsymbol{H}(z)\boldsymbol{b} \tag{8}$$

of (6), where $\boldsymbol{P}$ is a solution of the inhomogeneous equation with the initial condition

$$\boldsymbol{P}(d) = \boldsymbol{0}, \tag{9}$$

and $\boldsymbol{H}$ is an $(L+m)\times(L+m)$ matrix whose columns $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_{L+m}$ form a fundamental system for the homogeneous equation corresponding to the initial condition

$$\boldsymbol{H}(d) = \text{unit matrix}. \tag{10}$$

Then holds

$$\boldsymbol{W}(d) = \boldsymbol{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_{L+m} \end{bmatrix} \tag{11}$$

(c) Adjustment of the b_i contained in the general solution (8) of (6) to the measurements $\boldsymbol{Q}$. Let $\boldsymbol{v}, \boldsymbol{p}, \boldsymbol{h}_i$ be the vectors composed of the first l components of $\boldsymbol{W}, \boldsymbol{P}, \boldsymbol{H}_i$, respectively. Then the term

$$\left\| \boldsymbol{Q} - \boldsymbol{p}(0) - \sum_{i=1}^{L+m} b_i \boldsymbol{h}_i(0) \right\| \tag{12}$$

is to be minimized. Because of

$$\mathbf{a} = \begin{bmatrix} b_{L+1} \\ \vdots \\ b_{L+m} \end{bmatrix} \tag{13}$$

one has to take into account the constraint (cf. (11), (5))

$$\begin{bmatrix} b_1 \\ \vdots \\ b_L \end{bmatrix} = \mathbf{w}(d) = \mathbf{c}(b_{L+1}, \ldots, b_{L+m}). \tag{14}$$

Thus the problem is reduced to minimizing the objective function

$$(15) \quad Z(\mathbf{a}) = Z(b_{L+1}, \ldots, b_{L+m})$$

$$:= \left\| \mathbf{Q} - \mathbf{p}(0) - \sum_{i=1}^{L} c_i(b_{L+1}, \ldots, b_{L+m})\, \mathbf{h}_i(0) - \sum_{i=1}^{m} b_{L+i} \mathbf{h}_{L+i}(0) \right\|.$$

If $\mathbf{a}^*$ is a solution of the stated minimum problem, i.e.

$$Z(\mathbf{a}^*) \leqslant Z(\mathbf{a}) \quad \text{for all } \mathbf{a} \in A, \tag{16}$$

then $\mathbf{W}^{k+1}$ is determined within (8) by

$$\mathbf{b}^* = \begin{bmatrix} \mathbf{c}(\mathbf{a}^*) \\ \mathbf{a}^* \end{bmatrix}. \tag{17}$$

Assuming the initial condition (1b) to be independent from $\mathbf{a}$, one is confronted with a problem of linear approximation. In this case the vector

$$\mathbf{S} := \mathbf{Q} - \mathbf{p}(0) - \sum_{i=1}^{L} c_i \mathbf{h}_i(0) \tag{18}$$

is to be approximated with respect to a norm $\|\cdot\|$, chosen in the space $\mathbf{R}^l$, by a linear combination

$$\sum_{i=1}^{m} a_i \mathbf{h}_{L+i}(0). \tag{19}$$

When dealing with least square approximation, a^* satisfies the normal equations. For their solution it is advisable to orthonormalize the vectors $h_{L+1}(0), \ldots, h_{L+m}(0)$; a factorization algorithm due to Rutishauser [14] proves effective here (cf. [15]). Regarding the convergence of the process, numerical experiments show that the situation is analogous to that occurring in application of Newton's method, namely, either convergence is achieved rapidly or divergence soon becomes evident. A theoretical investigation of the convergence can be found in [16].

Applying quasilinearization to the design and identification of interference coatings we assume that the refractive index profile is imbedded in an m-parameter function family

$$N_A = \{n(\cdot, \boldsymbol{a}) \colon \boldsymbol{a} \in A,\ n(\cdot, \boldsymbol{a}) \in ([0, d] \to \boldsymbol{R})\}. \tag{20}$$

The observable quantities are related to the reflection coefficient. Differential equations for the latter can be derived in a classical way [5] or by means of Bellman's invariant imbedding principle (cf. [7]). Figure 5 shows the refractive index profile of the layered medium under consideration. The quantity

$$r(x) = \varrho(x) e^{i\Phi(x)}$$

is the reflection coefficient at the interface $z = x$ with respect to the refractive index profile shown in Figure 6, obtained from the original one in Figure 5 by truncation and continuation over $z < x$ with a constant refractive index $\tilde{n}(x)$. We assume $\tilde{n}$ to be a continuously differentiable function.

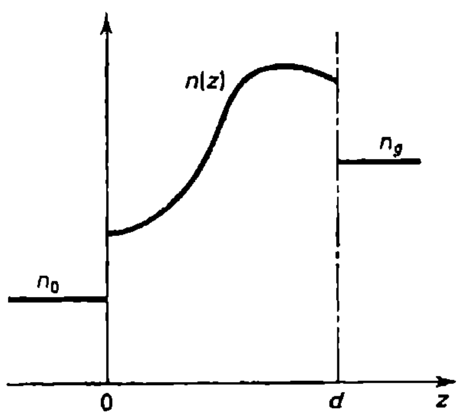

Fig. 5

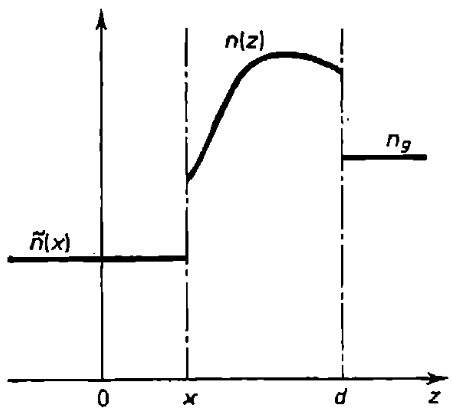

Fig. 6

The function $r = r(x)$ is a solution of the differential equation

$$(21a)\qquad \frac{d}{dx}r(x) = \frac{\frac{d\tilde{n}}{dx}(x)}{2\tilde{n}(x)}\left(1-r^2(x)\right)+ik\,n(x)*$$
$$*\left\{\left(\frac{n(x)}{\tilde{n}(x)}+\frac{\tilde{n}(x)}{n(x)}\right)r(x)+\frac{1}{2}\left(\frac{n(x)}{\tilde{n}(x)}-\frac{\tilde{n}(x)}{n(x)}\right)\left(1+r^2(x)\right)\right\},$$

subject to the initial condition

$$(21b)\qquad r(d) = \frac{\tilde{n}(d)-n_g}{\tilde{n}(d)+n_g}.$$

Using a recursion formula of Vlassov (cf. [3], [4]) one can show that the reflection coefficient r of the system in Figure 5 is given by

$$(22)\qquad \hat{r} = \frac{r_1+r(0)}{1+r_1r(0)},\qquad r_1 = \frac{n_0-\tilde{n}(0)}{n_0+\tilde{n}(0)}.$$

In most cases the values which are given as data for identification are the values $R_1, \ldots, R_l$ of the reflectivity $R = |\hat{r}|^2$ at wavelengths $\lambda_1, \ldots, \lambda_l$. From (21) we derive an initial value problem concerning the functions

$$(23)\qquad \begin{aligned}\varrho_j(x) &= \varrho(\lambda_j, x),\\ \Phi_j(x) &= \Phi(\lambda_j, x),\end{aligned}\qquad j = 1, \ldots, l,$$

thus arriving at the problem which is to be handled with quasilinearization.

The following choices of the function $\tilde{n}$ prove to be of particular interest:

$$\begin{aligned}(24)\qquad \tilde{n}(x) &= n_0,\\ (25)\qquad \tilde{n}(x) &= n(x),\end{aligned}\qquad 0 \leqslant x \leqslant d.$$

In [17] numerical experiments concerning the assumption (25) are reported. In case of (24) the initial condition (21b) is independent from the parameters a_j; that means, the approximation problem (15), (16) is a linear one. A more detailed discussion including numerical experiments is given in [15].

5. Pattern recognition methods

For the sake of simplicity we consider a one-dimensional output $y = F_a(x)$. Methods in system design, dealt with in this section presuppose an appropriate discretization of the parameter set A by a finite set $\tilde{A}$, sufficiently "dense" in A, and they are based upon a direct investigation of the graphs of F_a, $a \in \tilde{A}$, with regard to relevant shape phenomena. Depending from the number of elements in $\tilde{A}$ the computation time for a complete scanning may increase considerably; and therefore, in digitized waveform analysis effective methods of pattern recognition are urgently required. Recently techniques from formal language theory have been successfully incorporated; they are based upon a decomposition of the waveform into so-called pattern primitives. Substituting symbols for the latter, one gets the features of the waveform represented as a string of characters, i.e. a word of a certain pattern language L. Thus the investigation of the shape of the waveform can be effected by parsing a word of the pattern language in connection with a grammar G generating L (cf. [18], [19]).

Regardless of a subsequent use of linguistic methods, many concepts of digitized waveform analysis include feature generation, most usually effected by segmented best approximation, and frequently by piecewise straight line approximation. It may be useful to admit discontinuities at the breakpoints (cf. [20], [21], [22], [23]).

Generally, segmented best approximation is constructed respecting one of the following constraints:

(i) Given the number M of segments. Determination of an optimal segmentation such that the maximum of the approximation errors – relative to the best approximation in each segment – is as small as possible.

(ii) Given a tolerance ε. Determination of a segmentation with the fewest segments such that the error – relative to the best approximation – over each of them does not exceed ε.

When dealing with the detection of special phenomena, e.g. peaks or plateaus, ad-hoc algorithms prove efficient (cf. [24], [25]).

In application to thin-film synthesis we consider the following system class: owing to technological demands, multilayer systems with a given number S of layers and a fixed sequence

$$n_0//n_1/n_2/\ \dots\ /n_S//n_0 \tag{1}$$

of refractive indices are regarded; system parameters are the optical thicknesses $\Delta_1, \dots, \Delta_S$ of the layers. The considered discretization $\mathfrak{S}_{\tilde{A}}$

contains all systems with

$$\Delta_j = a_j \Delta, \quad j = 1, \ldots, S, \tag{2a}$$

$$a_1, \ldots, a_S \text{ integers}, \quad a_1 + a_2 + \ldots + a_S = P, \tag{2b}$$

where P is a fixed integer measuring the fineness of the discretization and $P\Delta$ is the total optical thickness of the stack. $\mathfrak{S}_\Delta^S$ comprises $\binom{P-1}{S-1}$ systems. Output is to be the reciprocal transmittance $T^{-1}(\varphi)$, which is given by a trigonometric polynomial of degree P, the coefficients being rational expressions in $n_1, \ldots, n_S$ (cf. Section 3). In order to determine achromatic systems, a continuous version of the algorithm [25] was employed. Algorithms for a complete scanning of $\mathfrak{S}_\Delta^S$ have been developed in [12]. A detailed description and results can be found in [26].

The above described thickness optimization can be related to the Chebyshev methods (cf. Section 3), originally conceived for an optimal adjustment of the refractive indices, thereby permitting a considerably faster detection of achromatic plateaus. This increase in efficiency is caused by the intensive utilization of the theoretical insight into the analytic structure of the output quantity. A comprehensive report including comparative complexity considerations is in preparation.

References

[1] S. Roberts, *On three-bar motion in plane space*, Proc. London Math. Soc. 7 (1875/76), 14–23.

[2] A. Cayley, *Three-bar motion*, ibid. 7 (1875/76), 142.

[3] Z. Knittl, *Optics of Thin Films*, London–New York–Sydney–Toronto 1976.

[4] П. Г. Кард, *Анализ и синтез многослойных интерференционных пленок*, Таллин 1971.

[5] H. Kaiser and H. C. Kaiser, *Elektromagnetische Wellen in geschichteten Medien*, Potsdamer Forschungen, Reihe B, Heft 16 (1979).

[6] H. Bremmer, *The WBK approximation as the first term of a geometrical-optical series*, *The Theory of Electromagnetic Waves, A Symposium*, New York 1951, 169–179.

[7] R. Bellman and R. Kalaba, *Functional equations, wave propagation and invariant imbedding*, J. Math. Mech. 8 (1959), 683–704.

[8] H. E. Moses, *Calculation of the scattering potential from reflection coefficients*, Phys. Review 102 (1956), 559–567.

[9] J. Hirsch, *An analytic solution to the synthesis problem for dielectric thin-film layers*, Optica Acta 26 (1979), 1273–1279.

[10] З. Ш. Блох, *Приближенный синтез механизмов*.

[11] P. L. Chebyshev, *Oeuvres*, St. Petersburg 1899 (vol. I), 1907 (vol. II).

[12] H. Kaiser, *Algorithmen zur Analyse und Synthese von Interferenzschichtsystemen*, Wiss. Zeitschrift PH Karl Liebknecht Potsdam, Jg. 17 (1973), Heft 1, 117–132.

[13] —, *Extremwerte des Reflexionsgrades*, ibid. Jg. 21 (1977), Heft 1, 189–203.

[14] H. Rutishauser, *Algorithmus 2: Orthonormierung von Vektoren nach E. Schmidt*, Computing 1/2 (1966), 159–161.

[15] H. Kaiser and H. C. Kaiser, *Synthesis and identificaton of thin-film systems by means of quasilinearization*, Wiss. Zeitschrift PH Karl Liebknecht Potsdam 25.1 (1981), 111–120.

[16] R. Kalaba, *On nonlinear differential equations, the maximum operation and monotone convergence*, J. Math. Mech. 8 (1959), 519–574.

[17] H. H. Kagiwada, *System Identification: Methods and Applications*, Reading (Mass.) 1974.

[18] A. V. Aho and J. D. Ullman, *The Theory of Parsing*, Translation and Compiling, Englewood 1973.

[19] K. S. Fu (Ed.), *Syntactic Pattern Recognition, Applications*, Berlin–Heidelberg–New York 1977.

[20] T. Pavlidis, *Waveform segmentation through functional approximation*, IEEE Trans. Computers, C-22 (1973), 689–697.

[21] T. Pavlidis and S. L. Horowitz, *Segmentation of plane curves*, ibid. C-23 (1974), 860–870.

[22] C. L. Lawson, *Characteristical properties of the segmented rational minmax approximation problem*, Numer. Math. 6 (1964), 293–301.

[23] G. M. Phillips, *Algorithms for piecewise straight line approximation*, The Computer Journal II (1968), 211–212.

[24] S. L. Horowitz, *A syntactic algorithm for peak detection in waveforms with applications to cardiography*, Comm. ACM 18 (1975), 281–285.

[25] A. F. Blumenthal, L. S. Davis, A. Rosenfeld, *Detecting natural "plateaus" in one-dimensional patterns*, IEEE Trans. Computers, C-26 (1977), 178–179.

[26] H. Kaiser and H. C. Kaiser, *Methods of pattern recognition in the synthesis and identification of thin-film systems*, Thin Solid Films 74 (1980), 173–178.

Presented to the Semester
Computational Mathematics
February 20–May 30, 1980