Prom. Nr. 3399

The Approximation Problem of Electrical Filters

Thesis presented to

THE SWISS FEDERAL INSTITUTE OF TECHNOLOGY, ZURICH

for the degree of

DOCTOR OF MATHEMATICS

by

ROSHDI ABDEL-RAHMAN AMER
B. Sc. Electrical Engineering, Dipl. Math. E.T.H.
Citizen of the U.A.R.

Accepted on the recommendation of
Prof. Dr. E. Stiefel
P.D. Dr. P. Läuchli

Birkhäuser Verlag Basel
und Stuttgart
1964

Separatabdruck aus:
Mitteilung Nr. 9 aus dem Institut für angewandte Mathematik
an der Eidgenössischen Technischen Hochschule in Zürich
Herausgegeben von Prof. Dr. E. Stiefel
Birkhäuser Verlag Basel und Stuttgart 1964

ACKNOWLEDGEMENTS

The idea of applying linear programming methods for the solution of the approximation problem of electrical filters was suggested by Professor Dr. E. STIEFEL in a colloquium concerning electrical filters which was held at the Institute of Applied Mathematics at the E.T.H. in Zurich during the year 1961. I wish to acknowledge my indeptedness to him for his continuing encouragement and interest with which he supervised the developement of this research.

Equally, I wish to express my appreciation to the participants in the colloquium for their many instructive discussions, specially Professor Dr. H. RUTISHAUSER, P.D. Dr. P. LÄUCHLI, Dr. H. R. SCHWARZ and Dipl. Ing. A. SCHAI (who suggested the applicability of linear programming methods by means of classifying the functions according to the signs of their numerators and denomenators).

Further, I express my sincere thanks to Dipl. Phys. A. GANZ and Dipl. Math. C. A. ZEHNDER for their generous assistance during compiling and testing of the ALGOL programs. Also, I should like to thank my wife for the preparation of the punch-cards for the ERMETH-code programs.

Zurich, March 1964. R. A.-R. AMER

Leer - Vide - Empty

CONTENTS

The Approximation Problem of Electrical Filters

R. A.-R. AMER

Leer - Vide - Empty

The Approximation Problem of Electrical Filters

R. A.-R. AMER

Chapter I. Introduction

This work deals with the so called:

'Approximation problem of electrical filters'.

This problem can be stated in the following form:

Given two sets D and S of intervals on the positive x-axis:

$$D = \{D_1, D_2, \ldots, D_\alpha, \ldots, D_{N_D}\}$$
$$\text{and} \qquad S = \{S_1, S_2, \ldots, S_\beta, \ldots, S_{N_S}\} \tag{1}$$

using the terminology of electrical filter theory, the intervals D_α and S_β will be called *'pass-bands'* and *'stop-bands'* respectively;

and a class F of functions $R(x)$ of the form:

$$R(x) = g(x) \cdot \frac{P_m(x)}{Q_n(x)}, \tag{2}$$

where: $P_m(x)$ and $Q_n(x)$ are *relatively prime* polynomials of *maximum degrees* m and n respectively,

and $g(x)$ is a given fixed positive weight function defined in the pass-bands D_α and in the stop-bands S_β.

It is then required to choose a function $R^*(x)$ out of the class F such that the expression:

$$\Delta^* = \frac{\max\limits_{x \in D} | R^*(x) |}{\min\limits_{x \in S} | R^*(x) |} \tag{3}$$

be a minimum, compared with the corresponding expressions for all other functions $R(x)$ in F.

To solve this problem the so called 'Exchange method' is used here in combination with linear programming methods (or—in some special cases—combined with solving an auxiliary eigenvalue problem).

The exchange method has been successfully used for approximation of continuous functions by polynomials [3]*); it has also been used (in combination with linear programming methods) for approximation by rational functions, and some examples were calculated by the author at the ERMETH in Zürich.

In the case of the approximation problem of electrical filters, the function to be approximated is zero in some intervals (the pass-bands) and infinity in some other intervals (the stop-bands); therefore some modifications are necessarily made in applying the exchange method.

Chapter II. The Exchange Method

In order to describe the exchange method which is used for solving the approximation problem of electrical filters the following definitions are introduced; most of the terms used are taken from [3].

A. Definitions

1. The 'maximum deviation' and the 'extreme points' of a function

The *maximum deviation* $\Delta(R(x))$ of a function $R(x)$ of the form (2) is defined by:

$$\Delta(R(x)) = \frac{\max\limits_{x \in D} | R(x) |}{\min\limits_{x \in S} | R(x) |} . \tag{4}$$

The points at which $| R(x) |$ assumes its maximum in D or its minimum in S are called *the extreme points* of the function $R(x)$.

2. The 'global optimal function' and the 'global T-deviation'

If the maximum deviation Δ^* of a function $R^*(x) \in F$ is such that:

$$\Delta^* = \Delta(R^*(x)) \leqq \Delta(R(x)) \tag{5}$$

for all functions $R(x) \in F$, then Δ^* is called *the global T-deviation***) and the function $R^*(x)$ is called *the global optimal function* for the given filter problem.

---

*) References used are indicated by numbers in square brackets.

**) The letter T will be used as an abbreviation for «TCHEBYCHEFF».

3. *The 'local classes'*

The class F is devided into a number of subclasses:

$$F_0, F_1, F_2, \ldots, F_f, \ldots, F_{N_F}.$$

The subclass F_0 consists of all functions in F which have either poles in D or zeros in S; it is obvious that all such functions have an infinite maximum deviation and therefore must be dropped out of consideration. The functions of the classes $F_1, F_2, \ldots, F_{N_F}$ have then the following properties:

a) $Q_n(x)$ has a constant sign σ_{D_α} in each pass-band D_α.

b) $P_m(x)$ has a constant sign σ_{S_β} in each stop-band S_β.

Either a positive or a negative sign may be chosen in each band; but since the functions

$$g(x)\frac{P_m(x)}{Q_n(x)}, \quad g(x)\frac{-P_m(x)}{Q_n(x)}, \quad g(x)\frac{P_m(x)}{-Q_n(x)} \quad \text{and} \quad g(x)\frac{-P_m(x)}{-Q_n(x)}$$

are equivalent, it is allowed to prescribe the sign of $P_m(x)$ in one of the stop-bands and the sign of $Q_n(x)$ in one of the pass-bands. Throughout this work the following choice is made:

$$\begin{aligned} P_m(x) &> 0 \quad \text{in } S_1, \\ Q_n(x) &> 0 \quad \text{in } D_1. \end{aligned} \tag{6}$$

It remains then to choose the signs σ_{D_α} and σ_{S_β} in the remaining $N_D + N_S - 2$ bands; this gives a number of

$$N_F = 2^{(N_D + N_S - 2)} \tag{7}$$

possible sign-combinations. Each sign-combination defines one of the classes $F_1, F_2, \ldots, F_{N_F}$.

These classes will be refered to as *the local classes.*

4. *The 'local optimal function' and the 'local T-deviation'*

If the maximum deviation Δ_f of a function $R_f(x) \in F_f$ is such that

$$\Delta_f = \Delta\big(R_f(x)\big) \leq \Delta\big(R(x)\big) \tag{8}$$

for all $R(x) \in F_f$, then Δ_f is called *the local T-deviation* for the local class F_f; and the function $R_f(x)$ is called *the local optimal function* of the local class F_f.

It is clear that *the global optimal function* for the filter problem *is one of the local optimal functions*; namely that which has the smallest local T-deviation.

5. A 'discrete set' and a 'reference'

a) *A discrete set* (d, s) is a set of points:

$$\begin{aligned} d &= xd_0, xd_1, \ldots, xd_r, \ldots, xd_\mu \quad xd_r \in D \\ \text{and} \quad s &= xs_0, xs_1, \ldots, xs_t, \ldots, xs_\nu \quad xs_t \in S \end{aligned} \tag{9}$$

having a total number of $\mu + \nu + 2 \geq m + n + 2$.

b) *A reference* $(\bar{d}, \bar{s})$ is a discrete set consisting of a number of exactly $m + n + 2$ points.

$$\mu + \nu + 2 = m + n + 2 . \tag{10}$$

6. The 'maximum discrete deviation' of a function

For a given discrete set, the expression

$$\delta_{(d,\,s)}(R(x)) = \frac{\max\limits_{x \in d} | R(x) |}{\min\limits_{x \in s} | R(x) |} \tag{11}$$

will be called *the maximum discrete deviation of the function* $R(x)$ *in the discrete set* (d, s).

7. The 'local discrete problem', the 'local discrete optimal function' and the 'local discrete T-deviation'

The *local discrete problem* is to construct a function $R_{(f)}(x) \in F_f$ such that its maximum discrete deviation, in a given discrete set (d, s), be a minimum compared with these of all functions in the local class F_f; this function $R_{(f)}(x)$ is then called the *local discrete optimal function*, and its maximum discrete deviation is called the *local discrete T-deviation* for the local class F_f in the discrete set (d, s):

$$\delta_{(f)}(d, s) = \delta_{(d,\,s)}(R_{(f)}(x)) . \tag{12}$$

8. A 'reference function', a 'leveled reference function' and the 'reference deviation'

Let the points of a reference $(\bar{d}, \bar{s})$ be numbered according to their order on the x-axis from left to right:

$$x_1, x_2, \ldots, x_j, \ldots, x_{m+n+2} ;$$

given a function $R(x)$, corresponding numbers θ_j are then defined by:

$$\theta_j = \begin{cases} R(x_j) & \text{for } x_j \in \bar{d} \\ -1/R(x_j) & \text{for } x_j \in \bar{s} . \end{cases} \tag{13}$$

A function $R(x)$ with the property that the numbers θ_j have alternating signs is called *a reference function with respect to the reference* $(\bar{d}, \bar{s})$. If these numbers θ_j are also equal in absolute value, then the function $R(x)$ is called *a leveled reference function with respect to the reference* $(\bar{d}, \bar{s})$ and its maximum discrete deviation in the set $(\bar{d}, \bar{s})$ is called *the reference deviation.*

The exchange method is based on the following theorems; these will be proved in the Appendix.

B. Theorems

Theorem 1: *If* $\bar{R}(x) \in F_f$ *is a reference function with respect to the reference* $(\bar{d}, \bar{s})$, *then:*

i) *The following relation is satisfied by all functions* $R(x) \in F_f$ *which are not proportional to* $\bar{R}(x)$*:*

$$\Delta(R(x)) > \frac{\min\limits_{x \in \bar{d}} |\bar{R}(x)|}{\max\limits_{x \in \bar{s}} |\bar{R}(x)|} \tag{14}$$

and

ii) *If* $\bar{R}(x)$ *is not a local optimal function in* F_f *then:*

$$\Delta(\bar{R}(x)) > \Delta_f > \frac{\min\limits_{x \in \bar{d}} |\bar{R}(x)|}{\max\limits_{x \in \bar{s}} |\bar{R}(x)|}. \tag{15}$$

The corresponding theorem for the case of the local discrete problem is:

Theorem 1′: *If* $\bar{R}(x) \in F_f$ *is a reference function with respect to the reference* $(\bar{d}, \bar{s})$ *which is included in the discrete set* (d, s), *then:*

i) *The following relation is satisfied by all functions* $R(x) \in F_f$, *which are not proportional to* $\bar{R}(x)$*:*

$$\delta_{(d,\,s)}(R(x)) > \frac{\min\limits_{x \in \bar{d}} |\bar{R}(x)|}{\max\limits_{x \in \bar{s}} |\bar{R}(x)|}. \tag{14′}$$

ii) *If* $\bar{R}(x)$ *is not a local discrete optimal function in* F_f *for the local discrete problem of the set* (d, s), *then:*

$$\delta_{(d,\,s)}(\bar{R}(x)) > \delta_f(d, s) > \frac{\min\limits_{x \in \bar{d}} |\bar{R}(x)|}{\max\limits_{x \in \bar{s}} |\bar{R}(x)|}. \tag{15′}$$

Theorem 2:

i) *If a function $R_f(x) \in F_f$ has a set of $m + n + 2$ extreme points such that they build a reference with respect to which $R_f(x)$ is a leveled reference function, then the function $R_f(x)$ is a local optimal function in F_f; and each other local optimal function for the same local class is proportional to $R_f(x)$.*

ii) *If a local optimal function $R_f(x) \in F_f$ satisfies the assumption:*

$$P_m(x) \text{ and } Q_n(x) \text{ are relatively prime polynomials of maximum degrees } m \text{ and } n \text{ respectively, such that at least one of them attains its maximum degree ,} \tag{16}$$

then the function $R_f(x)$ has at least $m + n + 2$ extreme points which contain a reference $(\bar{d}, \bar{s})$ with respect to which $R_f(x)$ is a leveled reference function.

The corresponding theorem for the case of the local discrete problem is:

Theorem 2′:

i) *A leveled reference function $R(x) \in F_f$ [with respect to a reference $(\bar{d}, \bar{s})$ included in the given discrete set (d, s)] which has a maximum discrete deviation equal to the reference deviation, is a local discrete optimal function in F_f.*

ii) *If a local discrete optimal function $\bar{R}_f(x)$ satisfies the assumption* (16), *then there exists a reference*

$$(\bar{d}, \bar{s}) \subset (d, s)$$

with respect to which the function $\bar{R}_f(x)$ is a leveled reference function with a reference deviation equal to its maximum discrete deviation.

In order to obtain the global optimal function $R^*(x)$, which is one of the N_F local optimal functions:

$$R_1(x), R_2(x), \ldots, R_f(x), \ldots, R_{N_F}(x) ,$$

each one of these functions is constructed and the best one among them is chosen (i.e. the local optimal function which has the smallest local T-deviation).

To construct one local optimal function $R_f(x)$ (for the local class F_f), the following exchange algorithm is used.

C. The Exchange Algorithm

The exchange algorithm is an iterative procedure in which the k-th iteration consists of the following steps:

1. For the discrete set $(d^{(k)}, s^{(k)})$ the local discrete problem is to be solved; the result is—according to Theorem 2′(ii)—a leveled reference function with respect to a reference included in $(d^{(k)}, s^{(k)})$.

Let this function be:

$$R^{(k)}(x) .$$

2. The local maximum and minimum points of the function $R^{(k)}(x)$ are determined, then out of these a reference

$$(\bar{d}^{(k+1)}, \bar{s}^{(k+1)})$$

is chosen such that:

i) $R^{(k)}(x)$ is a reference function with respect to the reference $(\bar{d}^{(k+1)}, \bar{s}^{(k+1)})$.

ii) For at least one point xd of $\bar{d}^{(k+1)}$: $\left| R^{(k)}(xd) \right| = \max\limits_{x \in D} \left| R^{(k)}(x) \right|$,
and at least one point xs of $\bar{s}^{(k+1)}$: $\left| R^{(k)}(xs) \right| = \min\limits_{x \in S} \left| R^{(k)}(x) \right|$.

iii) The expression:

$$\alpha^{(k)} = \frac{\min\limits_{x \in \bar{d}^{(k+1)}} \left| R^{(k)}(x) \right|}{\max\limits_{x \in \bar{s}^{(k+1)}} \left| R^{(k)}(x) \right|} \tag{17}$$

is to be made as large as possible.

According to theorem 1(ii) an upper and a lower bound for the local T-deviation Δ_f are given by:

$$\alpha^{(k)} < \Delta_f < \Delta\left(R^{(k)}(x)\right) , \tag{18}$$

if these bounds are close enough, the algorithm for the local class F_f can be stopped and the function $R^{(k)}(x)$ can be taken as a local optimal function for the local class F_f; this is justified by theorem 2(i).

3. The next discrete set $(d^{(k+1)}, s^{(k+1)})$ is constructed such that it includes at least the points of the reference $(\bar{d}^{(k+1)}, \bar{s}^{(k+1)})$.

The convergence of the exchange algorithm is based on the following theorem:

Theorem 3: *The lower bounds $\alpha^{(k)}$ for the local T-deviation Δ_f build a monotonicaly increasing sequence:*

$$\alpha^{(1)} < \alpha^{(2)} < \cdots < \alpha^{(k)} < \alpha^{(k+1)} < \cdots . \tag{19}$$

Proof:

The k-th iteration step of the exchange algorithm yields, for the local T-deviation Δ_f, the lower bound $\alpha^{(k)}$. At step 1 of the next iteration the local discrete optimal function $R^{(k+1)}(x)$ of F_f is constructed for the set $(d^{(k+1)}, s^{(k+1)})$. It follows from Theorem 1′(ii) and since the reference $(\bar{d}^{(k+1)}, \bar{s}^{(k+1)})$ is included in the set $(d^{(k+1)}, s^{(k+1)})$, that:

$$\delta_f(d^{(k+1)}, s^{(k+1)}) > \alpha^{(k)} . \tag{20}$$

The function $R^{(k+1)}(x)$ is—according to Theorem 2'(ii)—a leveled reference function with respect to a reference included in $(d^{(k+1)}, s^{(k+1)})$, and has a reference deviation equal to $\delta_f(d^{(k+1)}, s^{(k+1)})$.

At step 2 the reference $(d^{(k+2)}, \bar{s}^{(k+2)})$ is constructed and the number $\alpha^{(k+1)}$ is calculated. According to condition (iii) for the choice of $(\bar{d}^{(k+2)}, \bar{s}^{(k+2)})$ it must be that:

$$\alpha^{(k+1)} \geq \delta_f(d^{(k+1)}, s^{(k+1)}) . \tag{21}$$

From relations (20) and (21) it follows that:

$$\alpha^{(k+1)} > \alpha^{(k)} .$$

It is thus proved that the lower bound $\alpha^{(k)}$ for Δ_f is always raised after each iteration step of the exchange algorithm; but it is not proved that the upper bound $\Delta(R^{(k)}(x))$ is lowered; thus, a complete convergence proof is not established.

However, it has been practically found that a number of 3 to 5 iteration steps was sufficient to get the local optimal function within a permissible tolerance.

Chapter III. The Local Discrete Problem

The local discrete problem—obtained from the exchange algorithm which has been discussed in Chapter II—is to be solved in the present chapter.

Restatement of the problem:

Given a local class F_f of functions

$$R(x) = g(x) \frac{P_m(x)}{Q_n(x)} ,$$

which have the following properties:

a) $P_m(x)$ has a constant sign σ_{S_β} in the stop-band S_β, $\beta = 1, 2, \ldots, N_S$,

b) $Q_n(x)$ has a constant sign σ_{D_α} in the pass-pand D_α, $\alpha = 1, 2, \ldots, N_D$,

and a discrete set (d, s):

$$d = \{xd_0, xd_1, \ldots, xd_r, \ldots, xd_\mu\} ,$$
$$s = \{xs_0, xs_1, \ldots, xs_t, \ldots, xs_\nu\} .$$

with a total number of points

$$\mu + \nu + 2 \geq m + n + 2 .$$

It is then required to construct a function $\bar{R}_f(x) \in F_f$ such that its maximum discrete deviation in the set (d, s) satisfy the relation:

$$\delta_{(d,\,s)}(\bar{R}_f(x)) \leq \delta_{(d,\,s)}(R(x)) \quad \text{for all functions } R(x) \in F_f .$$

In order to solve this problem for the general case, where the number of pass- and stop-bands $N_D + N_S$ may be greater than three, linear programming methods will be used. For the special case of three bands ($N_D + N_S = 3$), another method can be used, in which the local discrete problem is expressed as an eigenvalue problem which is to be solved using a simple iterative process.

The next part A of this chapter deals with the eigenvalue problem; the linear programming method is discussed in the second part B.

A. The Eigenvalue Problem

In this part the local discrete problem is solved by directly constructing the local discrete optimal function $R_f(x)$. In this case the given discrete set (d^k, s^k) must be a reference $(\bar{d}^k, \bar{s}^k)$ having a number of $\mu + \nu + 2 = m + n + 2$ points.

The function $R_f(x)$ is expressed in the form:

$$R_f(x) = \frac{\sum_{i=0}^{m} \xi_i\, p_i(x)}{\sum_{k=0}^{n} \eta_k\, q_k(x)} = \frac{(\vec{\xi},\, \vec{p}(x))}{(\vec{\eta},\, \vec{q}(x))} \tag{22}$$

where: $p_0(x), p_1(x), \ldots, p_i(x), \ldots, p_m(x)$
and $q_0(x), q_1(x), \ldots, q_k(x), \ldots, q_n(x)$

are suitably chosen basis functions, such that:

$$\left.\begin{aligned} p_i(x) &= \text{a polynomial of degree } i, \\ q_k(x) &= (1/g(x)) \times \text{a polynomial of degree } k. \end{aligned}\right\} \tag{23}$$

According to theorem (2'ii), the function $\bar{R}_f(x)$ must be a leveled reference function with respect to the reference $(\bar{d}^k, \bar{s}^k)$; let the reference deviation be equal to $1/\lambda^2$. The following equations are then satisfied by $\bar{R}_f(x)$ at the points of the reference:

$$\left.\begin{aligned} \bar{R}_f(xd_r) &= \quad (-1)^j \cdot 1/\lambda\,, \quad r = 0, 1, \ldots, \mu \\ \text{and} \qquad \bar{R}_f(xs_t) &= -(-1)^j \cdot \lambda\,, \quad t = 0, 1, \ldots, \nu \end{aligned}\right\} \tag{24}$$

where j is the order of the point xd_r (or xs_t) among the points of the reference on the x-axis. Or, using the representation (22) for $\bar{R}_f(x)$:

$$\left.\begin{aligned} (-1)^j \cdot (\vec{\eta}, \vec{q}(xd_r)) &= \lambda \cdot (\vec{\xi}, \vec{p}(xd_r)) \\ \text{and} \qquad -(-1)^j \cdot (\vec{\xi}, \vec{p}(xs_t)) &= \lambda \cdot (\vec{\eta}, \vec{q}(xs_t)) \end{aligned}\right\} \tag{25}$$

which can be written in the following matrix form:

$$\begin{array}{c|cc|c|c} & 0 \quad i \quad m & 0 \quad k \quad n & & \\ \hline 0 & & \vdots & \xi_0 & \\ r & 0 & \cdots\ qd_{rk}\ \cdots & \xi_1 \ \vdots\ \xi_i & \\ \mu & & \vdots & \xi_m & \\ \hline 0 & \vdots & & \eta_0 \ \vdots & \\ t & \cdots\ ps_{ti}\ \cdots & 0 & \eta_k & = \\ \nu & \vdots & & \eta_n & \end{array} \tag{26}$$

$$\lambda \cdot \begin{array}{c|cc|c|c} & 0 \quad i \quad m & 0 \quad k \quad n & & \\ \hline 0 & \vdots & & & \\ r & \cdots\ pd_{ri}\ \cdots & 0 & \xi_i & \\ \mu & \vdots & & & \\ \hline 0 & & \vdots & & \\ t & 0 & \cdots\ qs_{tk}\ \cdots & \eta_k & \\ \nu & & \vdots & & \end{array}$$

where

$$\left.\begin{aligned} pd_{ri} &= p_i(xd_r)\,, & ps_{ti} &= -(-1)^j\, p_i(xs_t) \\ qd_{rk} &= (-1)^j\, q_k(xd_r)\,, & qs_{tk} &= q_k(xs_t) \end{aligned}\right\} \tag{27}$$

or in partitioned-matrix form:

$$\left(\begin{array}{c|c} 0 & (qd) \\ \hline (ps) & 0 \end{array}\right)\left(\begin{array}{c} \vec{\xi} \\ \hline \vec{\eta} \end{array}\right) = \lambda \cdot \left(\begin{array}{c|c} (pd) & 0 \\ \hline 0 & (qs) \end{array}\right)\left(\begin{array}{c} \vec{\xi} \\ \hline \vec{\eta} \end{array}\right) \tag{28}$$

where the matrices: (pd), (ps), (qd), (qs) have the elements given in (27).

The local discrete problem is thus reduced to the eigenvalue problem in (28). A solution of the first problem must also be a solution of the second one; but in order that a solution of the eigenvalue problem may be a correct solution for the local discrete problem, it must satisfy the following two conditions:

i) The eigenvalue λ must be as large as possible, and

ii) the corresponding eigenvector $(\vec{\xi}, \vec{\eta})$ must yield a function $R(x)$, according to (22), such that:

$$\frac{(\vec{\xi}, \vec{p}(x))}{(\vec{\eta}, \vec{q}(x))} = R(x) \in F_f \,. \tag{29}$$

In the general case of more than 3 bands, it may be necessary to go through the eigenvectors one after the other, each time checking if the corresponding function $R(x)$ lies in the required local class or not.

However, due to the following theorem, in the case of 3 bands the situation is much more simple. A filter consisting of 3 bands may be either a band-pass or a band-stop filter; the following discussion applies to a band-pass filter (a band-stop filter can be discussed similarly).

Theorem 4 (proof in the Appendix): *For a band-pass filter the global optimal function has the following two properties:*

a) *All its zeros lie in the pass-band.*

b) *It has $m + 1$ extreme points in the pass-band and $n + 1$ extreme points in the stop-bands, such that these $m + n + 2$ extreme points build a reference with respect to which the function is a leveled reference function.*

This theorem simplifies the problem in the following way:

1. Since we can choose the positive sign for $P_m(x)$ in the first stop-band and for $Q_n(x)$ in the pass-band, it follows from the first part (a) of the theorem that $P_m(x)$ has a constant sign $(-1)^m$ in the second stop-band. Therefore, it is sufficient to go through the exchange algorithm—as explained in Chapter II with the only exceptions that the discrete set (d^k, s^k) must be a reference—for only one local class, namely that defined by:

$$\left.\begin{array}{ll} P_m(x) & > 0 \quad \text{for } x \in S_1 \,, \\ Q_n(x) & > 0 \quad \text{for } x \in D \,, \\ (-1)^m \cdot P_m(x) & > 0 \quad \text{for } x \in S_2 \,. \end{array}\right\} \tag{30}$$

2. The second part (b) of the theorem allows the choice of the reference (d^k, s^k) for each step of the exchange algorithm, such that the number of points in d^k and s^k are:

$$\mu + 1 = m + 1 \quad \text{and} \quad \nu + 1 = n + 1 \,. \tag{31}$$

In the following it will be shown how this simplifies the problem.

According to (31), the matrices (pd) and (qs) have a square form; by proper choice of the basis functions $p_i(x)$ and $q_k(x)$, these matrices have inverses:

$$(pd)^{-1} \quad \text{and} \quad (qs)^{-1} \quad \text{respectively.}$$

Equations (28) can thus be written in the form:

$$\left(\begin{array}{c|c} (pd)^{-1} & 0 \\ \hline 0 & (qs)^{-1} \end{array}\right) \left(\begin{array}{c|c} 0 & (qd) \\ \hline (ps) & 0 \end{array}\right) \left(\begin{array}{c} \vec{\xi} \\ \hline \vec{\eta} \end{array}\right) = \lambda \cdot \left(\begin{array}{c} \vec{\xi} \\ \hline \vec{\eta} \end{array}\right)$$

or:

$$\left(\begin{array}{c|c} 0 & (pd)^{-1}\,(qd) \\ \hline (qs)^{-1}\,(ps) & 0 \end{array}\right) \left(\begin{array}{c} \vec{\xi} \\ \hline \vec{\eta} \end{array}\right) = \lambda \cdot \left(\begin{array}{c} \vec{\xi} \\ \hline \vec{\eta} \end{array}\right)$$

which can be written:

$$\left.\begin{array}{l} \left(\begin{array}{c|c} 0 & (D) \\ \hline (S) & 0 \end{array}\right) \left(\begin{array}{c} \vec{\xi} \\ \hline \vec{\eta} \end{array}\right) = \lambda \cdot \left(\begin{array}{c} \vec{\xi} \\ \hline \vec{\eta} \end{array}\right) \\ \text{where} \quad (D) = (pd)^{-1}\,(qd) \\ \text{and} \quad (S) = (qs)^{-1}\,(ps) \,. \end{array}\right\} \tag{32}$$

This is equivalent to:

$$\left.\begin{array}{rl} & (D)\,\vec{\eta} = \lambda \cdot \vec{\xi} \\ \text{and} & (S)\,\vec{\xi} = \lambda \cdot \vec{\eta} \end{array}\right\} \tag{33}$$

which leads to the following two eigenvalue problems:

$$(DS)\,\vec{\xi} = \lambda^2 \cdot \vec{\xi} \tag{34a}$$

$$\text{and} \quad (SD)\,\vec{\eta} = \lambda^2 \cdot \vec{\eta} \,. \tag{34b}$$

Let the matrix (DS) be iterated on a suitable vector $\vec{\xi}^{(0)}$ (each time normalizing the last component to the value 1); this defines the following sequence of vectors:

$$\vec{\xi}^{(0)}, \vec{\xi}^{(1)}, \vec{\xi}^{(2)}, \ldots, \vec{\xi}^{(k)}, \ldots \,.$$

For this iteration process the following theorem will be proved:

Theorem 5:

a) *The iteration process:*

$$\vec{\xi}^{(k)} = 1/(\lambda^{(k)})^2 \cdot (DS)\, \vec{\xi}^{(k-1)} \tag{35}$$

where $(\lambda^{(k)})^2$ *is determined such that the last component of* $\vec{\xi}^{(k)}$ *be equal to* 1, *yields the numbers* $(\lambda^{(k)})^2$ *which converge to the largest eigenvalue* λ_1^2 *of the matrix* (DS) *and the vectors* $\vec{\xi}^{(k)}$ *which converge to the corresponding eigenvector* $\hat{\vec{\xi}}$.

b) *The function* $\hat{R}(x)$ *defined by:*

$$\hat{R}(x) = \frac{(\hat{\vec{\xi}}, \vec{p}(x))}{(\hat{\vec{\eta}}, \vec{q}(x))} \tag{36}$$

where

$$\hat{\vec{\eta}} = (1/\lambda_1) \cdot (S)\, \hat{\vec{\xi}} \tag{37}$$

is the required solution for the given local discrete problem.

Proof: The following two points are at first proved:

α) *If the problem* (32) *has an eigenvalue* λ *(with a corresponding eigenvector* $(\vec{\xi}, \vec{\eta})$*), then* λ^2 *is an eigenvalue for the matrix* (DS); *the corresponding eigenvector is* $\vec{\xi}$.

This can be proved in the same way in which (34a) has been deduced from (32).

β) *If the matrix* (DS) *has an eigenvector* $\vec{\xi}$ *whose eigenvalue is* λ^2, *then the vector* $(\vec{\xi}, \vec{\eta})$ *with* $\vec{\eta} = (1/\lambda) \cdot (S)\, \vec{\xi}$ *is an eigenvector, having an eigenvalue* λ, *for the problem* (32). *Equations* (24) *are then satisfied by the function* $R(x)$ *which is defined by the eigenvector* $(\vec{\xi}, \vec{\eta})$ *according to* (22).

This can be proved by substitution in (32); the function $R(x)$ satisfies then the equations (24) because these are equivalent to (32).

Proof of a): The convergence proof of the iteration process (35) is established by proving:

i) *The eigenvalues of the matrix* (DS) *are all real; and the largest eigenvalue* λ_1^2 *is positive**).

ii) *The other eigenvalues* λ_j^2 $(j \neq 1)$ *satisfy the inequality:*

$$-\lambda_1^2 < \lambda_j^2 < \lambda_1^2 . \tag{38}$$

---

*) The numbers λ can assume imaginary values.

To prove i) examine (32) in its original form (25); it is then clear that if λ is an eigenvalue, then also $-\lambda$; namely if λ corresponds to the eigenvector $(\vec{\xi}, \vec{\eta})$, then $-\lambda$ corresponds to the eigenvector $(\vec{\xi}, -\vec{\eta})$; and since the matrix in (32) is real, therefore its eigenvalues are either real or pure imaginary numbers.

From this fact and (β) above, it follows that the eigenvalues of the matrix (DS) must be all real.

Further, if the local discrete problem has a solution, then (32) has at least one real eigenvalue; and from (α) it follows that the matrix (DS) has at least one positive eigenvalue; i.e. its largest eigenvalue is positive.

To prove ii) the cases where (38) is not satisfied are shown to be impossible; these are:

ii a) *An eigenvalue* $\lambda_j^2 = \lambda_1^2$ (a larger eigenvalue $\lambda_j^2 > \lambda_1^2$ can not exist since λ_1^2 is assumed to be the largest).

In this case it follows from (β) that the corresponding functions $R^{(j)}(x)$ and $R^{(1)}(x)$ are both leveled reference functions having the same discrete T-deviation. From theorem 1'(i) it follows that these functions must be proportional (in this case they are identical). This means that the eigenvectors corresponding to the eigenvalues λ_j^2 and λ_1^2 are the same; i.e. these eigenvalues are actually one and the same eigenvalue.

ii b) *An eigenvalue* $-a^2 \leq -\lambda_1^2$ (a real).

If the corresponding eigenvector is $\vec{\xi}$ (which must be real), then according to (β), the function:

$$R(x) = \frac{(\vec{\xi}, \vec{p}(x))}{i(\vec{\eta}, \vec{q}(x))}, \quad \text{where} \quad \vec{\eta} = (1/a) \cdot (S)\, \vec{\xi}$$

satisfies the equations:

$$\frac{(\vec{\xi}, \vec{p}(xd_r))}{i(\vec{\eta}, \vec{q}(xd_r))} = \pm 1/i\, a, \quad r = 0, 1, \ldots, m$$

and

$$\frac{(\vec{\xi}, \vec{p}(xs_t))}{i(\vec{\eta}, \vec{q}(xs_t))} = \pm i\, a, \quad t = 0, 1, \ldots, n.$$

It follows that the real function:

$$R_{real}(x) = \frac{(\vec{\xi}, \vec{p}(x))}{(\vec{\eta}, \vec{q}(x))}$$

has a maximum discrete T-deviation $\delta(R_{real}(x))$ which satisfies the inequality

$$\delta(R_{real}(x)) = 1/a^2 \leq 1/\lambda_1^2 = \delta(R^{(1)}(x)) \tag{39}$$

and since $R^{(1)}(x)$ is a leveled reference function it follows that (39) can be satisfied only if the functions are proportional, which means that the eigenvectors corresponding to $-a^2$ and λ_1^2 are proportional; i.e. these are one and the same eigenvalue; but this is impossible.

Proof of b): The function $\hat{R}(x)$ defined in (36) is, according to (β), a leveled reference function with respect to the given reference. Its reference deviation is $1/\lambda^2$; and no other leveled reference function with respect to the same reference have a smaller or equal reference deviation, since then and from (α) there would be for the matrix (DS) an eigenvalue that does not satisfy the inequality (38).

It remains only to prove that the function $\hat{R}(x)$ actually lies in the required local class as defined in (30):

i) At first we prove that no poles of $\hat{R}(x)$ can lie in the pass-band.

Assume that the function $\hat{R}(x)$ has a pole at the point $xp \in D$. A new function $R'(x)$ is then constructed by shifting the pole in xp to a new position xp':

$$R'(x) = \varrho(x) \cdot \hat{R}(x) \tag{40}$$

where

$$\varrho(x) = b \cdot \frac{x - xp}{x - xp'} .$$

By proper choice of b and xp' it can be reached that:

$$\left.\begin{aligned} |\varrho(x)| &\leq 1 \quad \text{for} \quad x \in D \\ \text{and} \quad |\varrho(x)| &> 1 \quad \text{for} \quad x \in S . \end{aligned}\right\} \tag{41}$$

The function $R'(x)$ has thus a maximum discrete deviation which is smaller than that, $1/\lambda_1^2$, of the original function $\hat{R}(x)$. There must exist therefore a local optimal function which has a still smaller maximum discrete deviation; and since this must be a leveled reference function with respect to the given reference (Theorem 2'ii), equations (24) [and (32)] are thus satisfied with a larger value for $\lambda = \overline{\lambda_1}$. From ($\alpha$) it follows that the matrix (DS) has an eigenvalue $\overline{\lambda_1^2} > \lambda_1^2$ which is impossible, since λ_1^2 is the largest eigenvalue. Our assumption that a pole of $\hat{R}(x)$ lies in the pass-band is thus proved to be false.

ii) Since the function $\hat{R}(x)$, which has been proved in i) to be a continuous function in D, has $m + 1$ alternating signs at the points $xd_r \in D$, it follows that its m zeros must be all real and lie inside the pass-band. Conditions (30) are thus satisfied by this function (after multiplication by a suitable factor $= \pm 1$), which means that $\hat{R}(x)$ lies in the required local class.

The Algorithm: The basis functions $p_i(x)$ and $q_k(x)$ defined in (23) are chosen such that the polynomials in these functions have their zeros in the pass-bands and in the stop-bands respectively, and such that the polynomials $p_m(x)$ and that one in $q_n(x)$ have their zeros in roughly estimated positions for the zeros and poles of the function $\hat{R}(x)$.

The exchange algorithm is then started with a reference $(d^{(0)}, s^{(0)})$ consisting of a number of $m + 1$ points in D and a number of $n + 1$ points in S. As explained in Chapter II, each iteration step of the exchange algorithm leads to a local discrete problem, which is to be solved using the following procedure:

1. The matrices (D) and (S) are calculated*) out of the matrices (pd), (ps), (qd) and (qs) defined in (27).

$$(D) = (pd)^{-1} \cdot (qd) , \quad (S) = (qs)^{-1} \cdot (ps) .$$

The product-matrix is then constructed: $(DS) = (D) \cdot (S)$.

2. A suitable vector $\vec{\xi}^{(0)}$ is then chosen; this can be either the vector $\hat{\vec{\xi}}$ from the previous exchange step, or (at the beginning) the vector $\vec{\xi}^{(0)} = (0, 0, \ldots, 0, 1)$ is chosen which is justified by the special choice of $p_m(x)$. The eigenvector $\hat{\vec{\xi}}$ belonging to the largest eigenvalue λ_1^2 of the matrix (DS) is calculated using the iteration:

$$\vec{\xi}^{(k)} = (1/\lambda^{(k)})^2 \cdot (DS)\, \vec{\xi}^{(k-1)} \tag{42}$$

where $\lambda^{(k)}$ is chosen such that the last component of $\vec{\xi}^{(k)}$ be equal to 1. According to Theorem 5 this process must converge such that:

$$\vec{\xi}^{(k)} \to \hat{\vec{\xi}} \quad \text{and} \quad \lambda^{(k)} \to \lambda_1 .$$

3. The local discrete optimal function $\hat{R}(x)$ is constructed according to equations (36) and (37).

The exchange algorithm is then continued by calculating the local maximum and minimum points of the function $\hat{R}(x)$ and choosing a new reference as explained in Chapter II. It can be easily shown that the new reference contains $m + 1$ points in D and $n + 1$ points in S, so that the resulting local discrete problem can be solved using the above process.

An ALGOL program**) for solving the approximation problem of a band-filter using the above method has been constructed and tested on the ERMETH by calculating the following example.

---

*) These matrices can be calculated directly without inverting (pd) and (qs).

**) The ALGOL programs are available on request at the Institute of Applied Mathematics at the ETH in Zürich.

An example (band-pass filter):

The filter requirements:

$$\text{stop-band } S_1\colon\ 0 \leq x \leq 1.5\,,$$
$$\text{pass-band } D\colon\ 2 \leq x \leq 4\,,$$
$$\text{stop-band } S_2\colon\ 5 \leq x\,.$$

The class F of functions:

$$R(x) = \frac{P_3(x)}{\sqrt{x}\,Q_2(x)} \qquad (m = 3,\ n = 2)\,.$$

Basis functions:

$$p_0(x) = 1 \qquad q_0(x) = \sqrt{x}\,,$$
$$p_1(x) = (x - 3) \qquad q_1(x) = \sqrt{x}\,(x - 6)\,,$$
$$p_2(x) = (x - 3)^2 - 1/2 \qquad q_2(x) = \sqrt{x}\,(x - 6)\,(x - 1)\,,$$
$$p_3(x) = (x - 3)^3 - 3\,(x - 3)/4$$

The reference $(\bar{d}^{(0)}, \bar{s}^{(0)})$:

$$\bar{d}^{(0)}\colon\ xd_0 = 2 \quad xd_1 = 2.5 \quad xd_2 = 3.5 \quad xd_3 = 4\,,$$
$$\bar{s}^{(0)}\colon\ xs_0 = 1.5 \quad xs_1 = 5 \quad xs_2 = 7$$

The solution of the first local discrete problem defined by the reference $(\bar{d}^{(0)}, \bar{s}^{(0)})$ was obtained after 5 iteration steps (42). The following local discrete optimal function was obtained (Figure 1):

$$\hat{R}(x) = \frac{(\hat{\xi}, \vec{p}(x))}{(\hat{\eta}, \vec{q}(x))}$$

where $\hat{\xi} = (-0.003\,122\,29,\ -0.033\,8401,\ 0.175\,690,\ 1)$

and $\hat{\eta} = (-1.015\,29\ ,\ -0.105\,459,\ -0.328\,843)\,.$

The maximum discrete deviation $\delta(\hat{R}(x)) = 1/\lambda_1^2 = 0.016\,5720$.

The first exchange step yielded the new reference $(\bar{d}^{(1)}, \bar{s}^{(1)})$:

$$\bar{d}^{(1)}\colon\ xd_0 = 2 \quad xd_1 = 2.376\,66 \quad xd_2 = 3.440\,66 \quad xd_3 = 4\,,$$
$$\bar{s}^{(1)}\colon\ xs_0 = 1.5 \quad xs_1 = 5 \quad xs_2 = 7.786\,42\,.$$

The upper and lower bounds for the T-deviation Δ^* were given:

$$\alpha^{(1)} = 0.016\,5720 < \Delta^* < 0.019\,0032 = \Delta(\hat{R}(x))\,.$$

N. B. The number $1/\lambda_1^2$ is generally smaller than the lower bound $\alpha^{(1)}$.

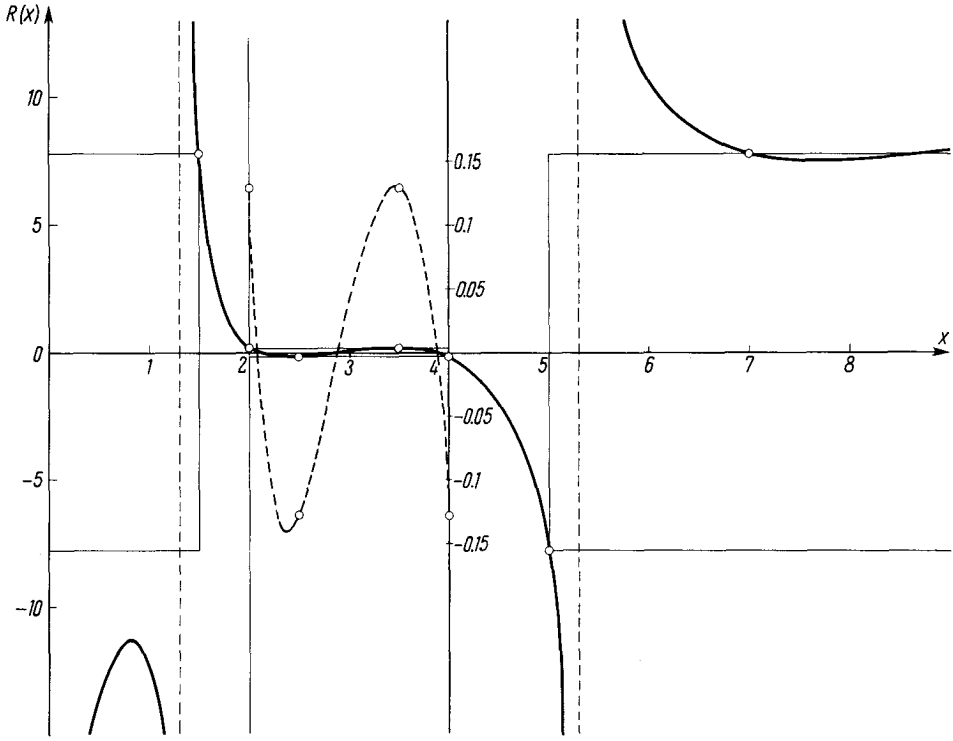

Figure 1
The local discrete optimal funktion $\hat{R}(x)$
The dashed curve in the pass-band represents the function $R(x)$ to the scale marked on the vertical at $x = 4$
o: Reference points

The second local discrete problem (for the reference $(\bar{d}^{(1)}, \bar{s}^{(1)})$) yielded the local discrete optimal function:

$$\hat{R}'(x) = \frac{(\hat{\vec{\xi}}', \vec{p}(x))}{(\hat{\vec{\eta}}', \vec{q}(x))}$$

where $\hat{\vec{\xi}}' = (-0.010\,0617,\ -0.024\,1942,\ 0.197\,815,\ 1)$

and $\hat{\vec{\eta}}' = (-1.011\,85,\ -0.103\,647,\ -0.332\,477)$.

The maximum discrete deviation $\delta(\hat{R}'(x)) = 1/\lambda_1'^2 = 0.017\,4711$.

The second exchange step gave for Δ^* the following upper and lower bounds:

$$\alpha^{(2)} = 0.017\,4711 < \Delta^* < 0.017\,4753 = \Delta(\hat{R}'(x)) .$$

The function $\hat{R}'(x)$ was thus accepted as the required solution for the approximation problem of the given filter. This function has the following zeros and poles:

Zeros: 2.085 74, 2.860 71 and 3.855 71;

Poles: 1.341 56 and 5.346 69.

B. The Linear Programming Method

This part deals with the general case of a filter having more than 3 bands ($N_D + N_S \geq 3$). In Chapter II it has been found that for the construction of the global optimal function, each one of the $2^{(N_D+N_S-2)}$ local optimal functions are to be calculated using the exchange algorithm which decomposes the problem into a number of local *discrete* problems. In Part A of this chapter the local discrete problem has been expressed as an eigenvalue problem, which could be easily solved in the case of a band-pass or a band-stop filter. However, in the case of more than 3 bands it is much more difficult to obtain the solution of the eigenvalue problem which gives a function $R(x)$ that lies in the required local class given in the local discrete problem. In this part B this algebraic problem is decomposed into a number of linear problems; namely linear programming problems.

In the local discrete problem there is given:

a) A discrete set (d, s)

$$d = \{xd_0, xd_1, \ldots, xd_r, \ldots, xd_\mu\}$$

and

$$s = \{xs_0, xs_1, \ldots, xs_t, \ldots, xs_\nu\}$$

such that the total number of points in (d, s) be:

$$\mu + \nu + 2 \geq m + n + 2\,.$$

b) A local class F_f of functions $R(x) = g(x) \cdot P_m(x)/Q_n(x)$ such that:

$$\left.\begin{aligned} &\operatorname{sign}\big(P_m(x)\big) = \sigma_{S_\beta} \text{ for } x \in S_\beta \\ \text{and} \quad &\operatorname{sign}\big(Q_n(x)\big) = \sigma_{D_\alpha} \text{ for } x \in D_\alpha\,. \end{aligned}\right\} \tag{43}$$

It is then required to construct the local discrete optimal function $R(x) \in F_f$ which has the smallest maximum discrete deviation in (d, s).

In order to solve this problem a sequence of linear programs is constructed in the following way:

1. Construction of a linear program:

For a given number l it is required to construct a function $R(x) \in F_f$ which satisfies the following conditions:

$$\left.\begin{array}{ll} & |R(xd_r)| \leq l\,, \quad \operatorname{sign}\big(Q_n(xd_r)\big) = \sigma_{D_\alpha} \ (xd_r \in D_\alpha) \\ \text{and} & |R(xs_t)| \geq 1/l\,, \ \operatorname{sign}\big(P_m(xs_t)\big) = \sigma_{S_\beta} \ (xs_t \in S_\beta) \\ \text{for} & r = 0, 1, \ldots, \mu \quad \text{and} \quad t = 0, 1, \ldots, \nu\,. \end{array}\right\} \tag{44}$$

(i.e. the maximum discrete deviation of $R(x)$ is to be $\leq l^2$).

Using for $R(x)$ the representation:

$$R(x) = \frac{(\vec{\xi},\ \vec{p}(x))}{(\vec{\eta},\ \vec{q}(x))}\,,$$

the conditions (44) can be written:

$$\left.\begin{array}{ll} -l \leq \dfrac{(\vec{\xi},\ \vec{p}(xd_r))}{(\vec{\eta},\ \vec{q}(xd_r))} \leq l & \\ \sigma_{D_\alpha} \cdot \big(\vec{\eta},\ \vec{q}(xd_r)\big) > 0 & r = 0, 1, \ldots, \mu \\ \text{and:} & \\ -l \leq \dfrac{(\vec{\eta},\ \vec{q}(xs_t))}{(\vec{\xi},\ \vec{p}(xs_t))} \leq l & \\ \sigma_{S_\beta} \cdot \big(\vec{\xi},\ \vec{p}(xs_t)\big) > 0 & t = 0, 1, \ldots, \nu \\ \text{where } xd_r \in D_\alpha \quad \text{and} \quad xs_t \in S_\beta & \end{array}\right\} \tag{45}$$

which is equivalent to the system of inequalities:

$$\left.\begin{array}{l} \sigma_{D_\alpha} \cdot \big[-\big(\vec{\xi},\ \vec{p}(xd_r)\big) + l \cdot \big(\vec{\eta},\ \vec{q}(xd_r)\big)\big] \geq 0 \\ \sigma_{D_\alpha} \cdot \big[\big(\vec{\xi},\ \vec{p}(xd_r)\big) + l \cdot \big(\vec{\eta},\ \vec{q}(xd_r)\big)\big] \geq 0 \qquad r = 0, 1, \ldots, \mu \\ \text{where } xd_r \in D_\alpha \\ \text{and} \\ \sigma_{S_\beta} \cdot \big[l \cdot \big(\vec{\xi},\ \vec{p}(xs_t)\big) - \big(\vec{\eta},\ \vec{q}(xs_t)\big)\big] \geq 0 \\ \sigma_{S_\beta} \cdot \big[l \cdot \big(\vec{\xi},\ \vec{p}(xs_t)\big) + \big(\vec{\eta},\ \vec{q}(xs_t)\big)\big] \geq 0 \qquad t = 0, 1, \ldots, \nu \\ \text{where } xs_t \in S_\beta\,. \end{array}\right\} \tag{46a}$$

Since the problem is homogeneous in $\vec{\xi}$ and $\vec{\eta}$ the solution of the above inequalities can be always chosen such that it satisfies the additional inequalities:

$$\left.\begin{array}{ll} -1 \leq \xi_i \leq 1 & i = 0, 1, \ldots, m\,, \\ \text{and} & \\ -1 \leq \eta_k \leq 1 & k = 0, 1, \ldots, n\,. \end{array}\right\} \quad (46\text{b}), (47\text{b})$$

The system of inequalities (46a) and (46b) is said to be *consistent* if there exists a vector $(\vec{\xi}, \vec{\eta})$ *whose components do not all vanish,* such that it satisfies these inequalities; otherwise the system is said to be *inconsistent.*

To examine the consistency of the above system and to construct a solution vector (if consistent), a new variable ζ is introduced and the system (46) is extended to:

$$\left.\begin{aligned}
Y(xd_r) &= \sigma_{D_\alpha} \cdot \left[-(\vec{\xi}, \vec{p}(xd_r)) + l \cdot (\vec{\eta}, \vec{q}(xd_r)) \right] + G(xd_r) \cdot \zeta \geq 0 \\
Y^*(xd_r) &= \sigma_{D_\alpha} \cdot \left[(\vec{\xi}, \vec{p}(xd_r)) + l \cdot (\vec{\eta}, \vec{q}(xd_r)) \right] + G(xd_r) \cdot \zeta \geq 0 \\
Y(xs_t) &= \sigma_{S_\beta} \cdot \left[l \cdot (\vec{\xi}, \vec{p}(xs_t)) - (\vec{\eta}, \vec{q}(xs_t)) \right] + G(xs_t) \cdot \zeta \geq 0 \\
Y^*(xs_t) &= \sigma_{S_\beta} \cdot \left[l \cdot (\vec{\xi}, \vec{p}(xs_t)) + (\vec{\eta}, \vec{q}(xs_t)) \right] + G(xs_t) \cdot \zeta \geq 0 \\
&(xd_r \in D_\alpha \quad \text{and} \quad xs_t \in S_\beta) .
\end{aligned}\right\} \tag{47a}$$

The constants $G(xd_r)$ and $G(xs_t)$ are positive weight-numbers for which the following values are chosen:

$$\left.\begin{aligned}
G(xd_r) &= \sqrt{\vec{p}^2(xd_r) + l^2 \cdot \vec{q}^2(xd_r)} \\
\text{and} \qquad G(xs_t) &= \sqrt{\vec{q}^2(xs_t) + l^2 \cdot \vec{p}^2(xs_t)} .
\end{aligned}\right\} \tag{48}$$

This choice has the following geometrical meaning:

If the hyperplanes

$$Y(xd_r) = 0, \; Y^*(xd_r) = 0 \quad \text{and} \quad Y(xs_t) = 0, \; Y^*(xs_t) = 0$$

are represented in the cartesian space $\xi_0, \xi_1, \ldots, \xi_m, \eta_0, \ldots, \eta_n, \zeta$ then these will all have an inclination of 45° to the ζ-axis.

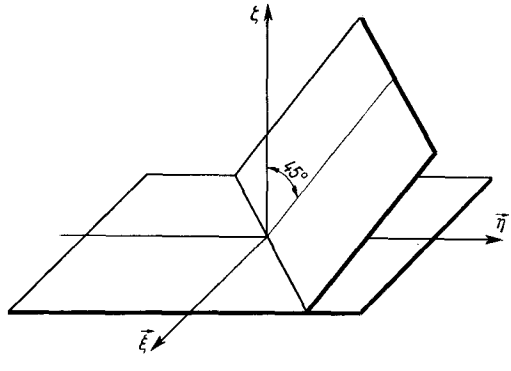

Figure 2

In this geometrical model the points $(\vec{\xi}, \vec{\eta}, \zeta)$ which satisfy the inequalities (47a) lie in the convex cone $C_l^{(m+n+3)}$ built by the intersection of the half-spaces (47a); its vertix is the origin. The points which satisfy the inequalities (46a) [and thus correspond to functions $R(x)$ which satisfy conditions (44)] lie in the convex cone $C_l^{(m+n+2)}$ built by the intersection of the cone $C_l^{(m+n+3)}$ with the hyperplane $\zeta = 0$. The cone $C_l^{(m+n+2)}$ will be refered to as *the feasible cone*; its size depends on the value of l and for different values of l the corresponding feasible cones obey the relation:

$$C_{l_2}^{(m+n+2)} \subset C_{l_1}^{(m+n+2)} \quad \text{for} \quad l_2 \leq l_1 \,. \tag{49}$$

The feasible cone and the relation (49) are illustrated by the Figures (3a) and (3b) respectively:

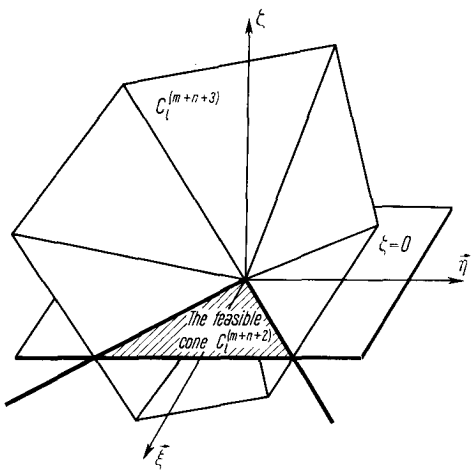

Figure 3a
The feasible cone

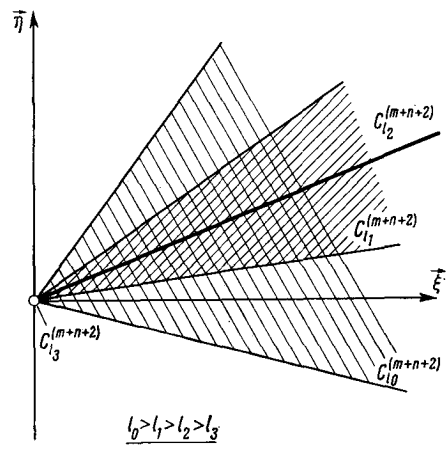

Figure 3b

In order to investigate the consistency of the system (46a) the following linear program is solved:

$$\left.\begin{array}{l} \textit{To determine values for } \xi_0, \xi_1, \ldots, \xi_m, \eta_0, \eta_1, \ldots, \eta_n, \zeta \\ \textit{maximizing the objective function: } Z = -\zeta \\ \textit{subject to the constraints: } (47a) \text{ and } (47b). \end{array}\right\} \tag{50}$$

The relation between the consistency of the system (46a) and the solution of the linear program (50) is given by the following theorem:

2. Theorem 6:

The system of inequalities (46a) *is consistent* (*and has a solution* $(\hat{\vec{\xi}}, \hat{\vec{\eta}})$), *if and only if the linear program* (50) *has an optimal solution* $(\hat{\vec{\xi}}, \hat{\vec{\eta}}, \hat{\zeta}) \neq \vec{0}$.

Proof:

i) If the linear program (50) has an optimal solution $(\hat{\vec{\xi}}, \hat{\vec{\eta}}, \hat{\zeta})$, then since $(\vec{\xi}, \vec{\eta}, \zeta) = \vec{0}$ is a feasible solution, the value of $\hat{\zeta}$ must be either zero or negative. From this and since $(\hat{\vec{\xi}}, \hat{\vec{\eta}}, \hat{\zeta})$ satisfies the inequalities (47a) it follows that the system of inequalities (46a) has a solution $(\hat{\vec{\xi}}, \hat{\vec{\eta}})$.

ii) If the system (46a) is consistent having a solution $(\overset{*}{\vec{\xi}}, \overset{*}{\vec{\eta}})$ then $(\overset{*}{\vec{\xi}}, \overset{*}{\vec{\eta}}, 0)$ is a feasible solution for the linear program (50). The optimal solution of (50) must be, therefore, such that $\hat{\zeta} \leq 0$.

If $\hat{\zeta} = 0$, then $(\vec{\xi}, \vec{\eta}, \zeta) = (\overset{*}{\vec{\xi}}, \overset{*}{\vec{\eta}}, 0)$ is an optimal solution of the linear program.

If $\hat{\zeta} < 0$, then it follows from (47a) that not all components of $\hat{\vec{\xi}}$ and $\hat{\vec{\eta}}$ can vanish.

From the above discussion it is seen that the solution of the linear program (50) defines a function $\hat{R}(x)$ which satisfies the conditions (44) for a given value of l if such a function exists; if not then this is indicated by the result that the linear program has the only optimal solution $(\hat{\vec{\xi}}, \hat{\vec{\eta}}, \hat{\zeta}) = \vec{0}$.

The aim of the algorithm outlined below is to get a value for l such that the system (46a) be just consistent. The solution of the corresponding linear program defines a function:

$$\hat{R}(x) = \frac{(\hat{\vec{\xi}}, \vec{p}(x))}{(\hat{\vec{\eta}}, \vec{q}(x))} \tag{51}$$

where $(\hat{\vec{\xi}}, \hat{\vec{\eta}}, \hat{\zeta})$, $(\hat{\zeta} \cong 0)$ is the optimal solution of the linear program (50).

The function $\hat{R}(x)$ is then the required local discrete optimal function.

3. Outlines of the algorithm:

The algorithm that uses the linear programming method for solving one local discrete problem, obtained from the exchange algorithm as described in Chapter II, consists of the following steps:

0. A value l_0 is chosen for l at the beginning such that l_0^2 be a roughly estimated value for the local T-deviation of the given local class.

1. For the chosen value of l the corresponding linear program as defined in (50) is solved. This leads to one of the following situations:

a) The optimal solution is $(\hat{\xi}, \hat{\eta}, \hat{\zeta}) = \vec{0}$; in this case the inequalities (46a) are inconsistent (Theorem 6)—the feasible cone $C_l^{(m+n+2)}$ consists then of only one point namely the origin—a greater value for l is thus to be chosen and the corresponding linear program is solved.

b) An optimal solution is obtained such that the corresponding function $\hat{R}(x)$ gives for the local discrete T-deviation:

$$\left.\begin{aligned} &\text{an upper bound } \delta = \max_r \left| R(xd_r) \right| / \min_t \left| R(xs_t) \right| \\ &\text{and a lower bound } \beta = \max_{(\bar{d}, \bar{s}) \subset (d, s)} \left(\min_{x \in \bar{d}} \left| R(x) \right| / \max_{x \in \bar{s}} \left| R(x) \right| \right). \end{aligned}\right\} \tag{52}$$

If the bounds δ and β are not close enough to one another then the algorithm goes to step 2 for the choice of a smaller value for l. In the situation b) the feasible cone $C_l^{(m+n+2)}$ has a considerable size.

c) An optimal solution is obtained such that the corresponding function $\hat{R}(x)$ gives two bounds δ and β for the local discrete T-deviation [according to (52)], such that they are practically equal. The function $\hat{R}(x)$ is then a practically leveled reference function with respect to the reference $(\bar{d}, \bar{s})$ and can thus be accepted as a solution for the local discrete problem. Situation c) corresponds to a narrow feasible cone $C_l^{(m+n+2)}$.

2. A new value for l is chosen as follows:

A possible choice is such that $l^2 = \delta$ [given by (52)], this has the advantage that the linear program of the next step has always an optimal solution different from $\vec{0}$, but then the convergence to the required solution of the local discrete problem is too slow. However, it was practically found that the relation between the values of l_i and the values of $\hat{\zeta}_i$ given by the optimal solutions of the corresponding linear programs is almost a linear relation, so that for the linear program i the following choice can be made for l_i:

$$\left.\begin{aligned} &i = 1: \quad \beta < l_1^2 < \delta\,, \\ &i \geq 2: \quad l_i = \frac{l_{i-1} \cdot \hat{\zeta}_{i-2} - l_{i-2} \cdot \hat{\zeta}_{i-1}}{\hat{\zeta}_{i-2} - \hat{\zeta}_{i-1}}\,. \end{aligned}\right\} \tag{53}$$

With this new value for l step 1. above is then repeated.

It was found, by numerical experiments, that a total number of 3 to 6 linear programs was sufficient to obtain a solution for the local discrete problem.

N. B. For the calculation of the bounds δ and β the values of the function $\hat{R}(x)$ at the points of the given discrete set are needed; these can be directly

calculated from the values of

$$\hat{Y}(xd_r),\ \hat{Y}^*(xd_r) \quad r = 0, 1, \ldots, \mu$$
$$\hat{Y}(xs_t),\ \hat{Y}^*(xs_t) \quad t = 0, 1, \ldots, \nu \text{ and } \hat{\zeta}$$

given by the optimal solution of the linear program:

$$\left.\begin{aligned} \hat{R}(xd_r) &= \frac{\hat{Y}^*(xd_r) - \hat{Y}(xd_r)}{\hat{Y}^*(xd_r) + \hat{Y}(xd_r) - 2 \cdot G(xd_r) \cdot \hat{\zeta}} \cdot l\,, \\ 1/\hat{R}(xs_t) &= \frac{\hat{Y}^*(xs_t) - \hat{Y}(xs_t)}{\hat{Y}^*(xs_t) + \hat{Y}(xs_t) - 2 \cdot G(xs_t) \cdot \hat{\zeta}} \cdot l\,. \end{aligned}\right\} \tag{54}$$

4. The Simplex Algorithm:

In order to solve the linear program (50), the simplex algorithm is used in its direct form as described in [4].

The inequalities (47a) are written in the form of a table:

	$Z =$	$Y(x_j) =$	$Y^*(x_j) =$
$-\xi_0$	0	$p_{0,2j-1}$	$p_{0,2j}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$-\xi_i$	0	$p_{i,2j-1}$	$p_{i,2j}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$-\xi_m$	0	$p_{m,2j-1}$	$p_{m,2j}$
$-\eta_0$	0	$q_{0,2j-1}$	$q_{0,2j}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$-\eta_k$	0	$q_{k,2j-1}$	$q_{k,2j}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$-\eta_n$	0	$q_{n,2j-1}$	$q_{n,2j}$
$-\zeta$	1	$-G(x_j)$	$-G(x_j)$
1	0	0	0

(55)

where: for $x_j \in d$:
$$p_{i,2j-1} = -p_{i,2j} = \sigma_{D_\alpha} \cdot p_i(x_j)$$
$$q_{k,2j-1} = q_{k,2j} = -\sigma_{D_\alpha} \cdot l \cdot q_k(x_j)$$

and for $x_j \in s$:
$$p_{i,2j-1} = p_{i,2j} = -\sigma_{S_\beta} \cdot l \cdot p_i(x_j)$$
$$q_{k,2j-1} = -q_{k,2j} = \sigma_{S_\beta} \cdot q_k(x_j)\,.$$

(55′)

If an initial solution $(\vec{\xi}^0, \vec{\eta}^0)$ is known from a previous linear program, then the above table is transformed in the following way:

1. The variables ξ_i and η_k are expressed in the form:

$$\left.\begin{aligned} \xi_i &= \xi_i^0 + \delta\xi_i \quad i = 0, 1, \ldots, m \\ \text{and} \qquad \eta_k &= \eta_k^0 + \delta\eta_k \quad k = 0, 1, \ldots, n\,. \end{aligned}\right\} \tag{56a}$$

By substitution in the table (55) this is modified as follows:

The zeros in the last row, with the exception of that in the column Z, are replaced by the constant elements:

$$b'_v = -\sum_{i=0}^{m} \xi_i^0 \cdot a_{i,v} - \sum_{k=0}^{n} \eta_k^0 \cdot a_{m+1+k,v} \tag{56b}$$

($a_{u,v}$ is the general element in the table which lies in the row u—the rows being numbered from 0—and in the column v—the column 0 is the column of the variable Z). The row variables $\xi_0, \xi_1, \ldots, \xi_m, \eta_0, \eta_1, \ldots, \eta_n$ are then replaced by the new variables

$$\delta\xi_0, \delta\xi_1, \ldots, \delta\xi_m, \delta\eta_0, \delta\eta_1, \ldots, \delta\eta_n\,.$$

2. If any of the elements b'_v $(= a_{m+n+3,v})$ are found to be negative then the following transformation must be also made, since it is required to have positive (or zero) elements in the last row before the simplex algorithm can be started:

$$\zeta = \zeta^0 + \delta\zeta\,. \tag{56c}$$

This replaces the elements b'_v of the last row by the new elements

$$b_v = b'_v - \zeta^0 \cdot a_{m+n+2,v}\,. \tag{56d}$$

The value of ζ^0 is chosen according to:

$$\zeta^0 = \max_v (b'_v / a_{m+n+2,v})\,. \tag{56e}$$

From the fact that the elements $a_{m+n+2,v}$ are all negative it follows that the new elements b_v in the last row will be all ≥ 0. At the end of the transformations (56a–e) the table has the variables

$$\delta\xi_0, \delta\xi_1, \ldots, \delta\xi_m, \delta\eta_0, \delta\eta_1, \ldots, \delta\eta_n, \delta\zeta$$

written on the left. These transformations are done for the following purpose:

i) To remove the zeros in the last row, which cause the problem to be seriously degenerated; in fact if no initial solution $(\vec{\xi}^0, \vec{\eta}^0)$ is given, then to remove the degeneracy, any initial solution may be assumed.

ii) To make use of the information obtained from the previous linear program by reducing the number of A.T.-steps (these will be defined later*) necessary to attain the optimal solution of the present linear program; in fact, if the initial solution is close enough to the optimal solution, then the number of A.T.-steps is reduced to its minimum $m + n + 3$.

It remains only to take the inequalities (47b) into consideration; these can be written in the form

$$\left.\begin{aligned} X_i &= 1 - \xi_i = (1 - \xi_i^0) - \delta\xi_i \geq 0 \\ X_i^* &= 1 + \xi_i = (1 + \xi_i^0) + \delta\xi_i \geq 0 \end{aligned}\right\} \quad i = 0, 1, \ldots, m$$

$$\left.\begin{aligned} \text{and } X_{m+1+k} &= 1 - \eta_k = (1 - \eta_k^0) - \delta\eta_k \geq 0 \\ X_{m+1+k}^* &= 1 + \eta_k = (1 + \eta_k^0) + \delta\eta_k \geq 0 \end{aligned}\right\} \quad k = 0, 1, \ldots, n. \tag{57}$$

These could be written in an additional part of the table:

	$X_i =$	$X_i^* =$	$X_{m+1+k} =$	$X_{m+1+k}^* =$
	0	0	0	0
	.	.	.	.
	0	0	0	0
$-\delta\xi_i$	1	−1	0	0
	0	0	0	0
	.	.	.	.
	.	.	.	.
	0	0	0	0
$-\delta\eta_k$	0	0	1	−1
	0	0	0	0
	.	.	.	.
$-\delta\zeta$	0	0	0	0
1	$1 - \xi_i^0$	$1 + \xi_i^0$	$1 - \eta_k^0$	$1 + \eta_k^0$

(58)

This has the disadvantage that the table (55) will be almost doubled. However, this can be avoided by modifying the standard simplex algorithm using the fact that any of the columns in (58) can be calculated from the elements of the original table (55) after any number of A.T.-steps; the rules for the choice of the pivot element will thus be modified, and from time to time a column of type (58) will be added to or taken away from the table as will be described in detail later.

The simplex algorithm used here consists of two stages each consisting of a number of A.T.-steps in which the pivot element is chosen according to the rules detailed below. The A.T.-step itself is defined by the following:

*) See also the book [4].

i) The A.T.-step

Let the pivot element be in the row *upiv* and in the column *vpiv*, i.e. the element $a_{upiv,\,vpiv}$ of the table which is assumed to have the following form:

	$Z =$	$V_{vpiv} =$	$V_v =$	
$-\ U_{upiv}$	.	. . . $a_{upiv,\,vpiv}$	. $a_{upiv,\,v}$	. . .
$-\ U_u$	.	. . . $a_{u,\,vpiv}$	. $a_{u,\,v}$	. . .
1		.	.	

(59)

The variable U_{upiv} is expressed in terms of the variable V_{vpiv} and the other variables U_u ($u \neq upiv$) by solving the equation in the column *vpiv* for U_{upiv}. The variables V_v ($v \neq vpiv$) are then also expressed in terms of V_{vpiv} and U_u ($u \neq upiv$) so that the table will have the new form:

	$Z =$	$U_{upiv} =$	$V_v =$	
$-\ V_{vpiv}$	.	. . . $a'_{upiv,\,vpiv}$	. $a'_{upiv,\,v}$	. . .
$-\ U_u$	.	. . . $a'_{u,\,vpiv}$	. $a'_{u,\,v}$	. . .
1		.	.	

(59′)

where:
$$a'_{upiv,\,vpiv} = 1/a_{upiv,\,vpiv}$$
$$a'_{upiv,\,v} = -\,a_{upiv,\,v}/a_{upiv,\,vpiv} \quad (v \neq vpiv)$$
$$a'_{u,\,vpiv} = a_{u,\,vpiv}/a_{upiv,\,vpiv} \quad (u \neq upiv)$$
and
$$a'_{u.\,v} = a_{u,\,v} - a_{u,\,vpiv} \cdot a_{upiv,\,v}/a_{upiv,\,vpiv}$$
$$(u \neq upiv \text{ and } v \neq vpiv).$$

(59″)

ii A) The first stage of the simplex algorithm
(The Elimination of the free variables):

Since the variables $\delta\xi_i$ $(i = 0, 1, \ldots, m)$, $\delta\eta_k$ $(k = 0, 1, \ldots, n)$ and $\delta\zeta$ are allowed to assume either positive or negative values (they are called the free variables) they must be removed from the left side of the table by a number of $m + n + 3$ A.T.-steps, in each of which the pivot element is chosen as follows*):

1. *Pivot row:*

Among the rows which still have a free variable the one having the largest absolute value for the element $a_{u,0}$ is taken as the pivot row *upiv*.

2. *Pivot column:*

The following characteristic quotients Q are to be considered:

(i) For variables of type Y or Y^* written at the top of the table; let $Y(x_j)$ be at the top of the column v, then the corresponding characteristic quotient is

$$Q(Y(x_j)) = a_{m+n+3,\,v}/a_{upiv,\,v}\,. \tag{60i}$$

(ii) For variables of type X or X^* corresponding to variables $\delta\xi_i$ or $\delta\eta_k$ written at the top of the table; let $\delta\xi_i$ be at the top of the column v, then the two characteristic quotients are to be considered

$$\left.\begin{aligned} Q(X_i) &= (1 - \xi_i^0 - a_{m+n+3,\,v})/(-a_{upiv,\,v}) \\ \text{and}\quad Q(X_i^*) &= (1 + \xi_i^0 + a_{m+n+3,\,v})/a_{upiv,\,v}\,. \end{aligned}\right\} \tag{60ii}$$

(iii) For the two variables X and X^* corresponding to a variable $\delta\xi_i$ (or $\delta\eta_k$) written in the pivot row

$$\left.\begin{aligned} Q(X_i) &= 1 - \xi_i^0 \\ \text{and}\quad Q(X_i^*) &= -1 - \xi_i^0\,. \end{aligned}\right\} \tag{60iii}$$

Out of these characteristic quotients only those which have an opposite sign to $a_{upiv,\,0}$ are calculated and the one having the smallest absolute value is chosen; this belongs either to one of the variables Y (or Y^*) or to one of the variables X (or X^*):

(a) If the characteristic quotient $Q(Y(x_j))$, belonging to the variable $Y(x_j)$ written at the top of the column v, is chosen then the column v is taken as the pivot column *vpiv*.

(b) If $Q(X_i)$ (or $Q(X_i^*)$) is chosen, then the column corresponding to the variable X_i (or X_i^*) is added to the table and taken as the pivot column *vpiv*.

---

*) The general term in the table is assumed to be $a_{u,\,v}$, where u is the number of the row $u = 0, 1, \ldots, m + n + 3$ and v is the number of the column $v = 0, 1, \ldots, N_{col}$.

This column is calculated from the rest of the table in the following way:

(b1) If the variable $\delta\xi_i$ (corresponding to X_i (or X_i^*)) is at the top of the column v, then:

$$\left.\begin{aligned} &\textit{Column } X_i\text{:} \quad a_{u,\,vpiv} = -a_{u,\,v} \quad u = 0, 1, \ldots, m+n+2 \\ &\qquad\qquad a_{m+n+3,\,vpiv} = 1 - \xi_i^0 - a_{m+n+3,\,v} \\ &\textit{Column } X_i^*\text{:} \quad a_{u,\,vpiv} = a_{u,\,v} \quad u = 0, 1, \ldots, m+n+2 \\ &\qquad\qquad a_{m+n+3,\,vpiv} = 1 + \xi_i^0 + a_{m+n+3,\,v}\,. \end{aligned}\right\} \tag{61 b1}$$

(b2) If $\delta\xi_i$ is written in the pivot row, then:

$$\left.\begin{aligned} &\textit{Column } X_i\text{:} \quad a_{u,\,vpiv} = 0 \quad \text{for all } u \neq upiv \\ &\qquad\qquad a_{upiv,\,vpiv} = 1\,, \quad a_{m+n+3,\,vpiv} = 1 - \xi_i^0 \\ &\textit{Column } X_i^*\text{:} \quad a_{u,\,vpiv} = 0 \quad \text{for all } u \neq upiv \\ &\qquad\qquad a_{upiv,\,vpiv} = -1, \quad a_{m+n+3,\,vpiv} = 1 + \xi_i^0\,. \end{aligned}\right\} \tag{61 b2}$$

iiB) The second stage of the simplex algorithm (The actual simplex stage):

The free variables are now at the top of the table and the actual simplex algorithm can begin with the following rules for the choice of the pivot element:

1. Pivot row:

The row with the most negative element $a_{u,0}$ is chosen; if none of the elements $a_{u,0}$ $(u = 0, 1, \ldots, m+n+2)$ is negative then the simplex algorithm is at its end.

2. Pivot column:

The same characteristic quotients as given by (60i) and (60ii) in the first stage are considered (with the only exception, that if a variable X_i (or X_i^*) is in the pivot row then the corresponding characteristic quotient is not considered). Out of these only the positive characteristic quotients are calculated and the smallest is chosen. The pivot column is determined, and if necessary added to the table, in the same way as in the first stage.

In order to avoid an unnecessary increase in the number of columns of the table, the following rules are to be followed: After each A.T.-step, if the variable that newly comes to the top of the table is an X_i (or X_i^*), then:

a) If this comes to the top of the last column, then this last column is to be cancelled.

b) If it comes to the top of another column, then this is replaced by the last column in the table with the variable at its top.

In this way the variables X and X^* are allowed to stay only at the left of the table which keeps the total number of columns less than $2(\mu + \nu + 2) + m + n + 3$; actually, the total number of columns in most of the cases is equal to $2(\mu + \nu + 2) + 1$.

At the end of the simplex algorithm, the optimal solution of the linear program (50) is read from the table in the following way:

a) *The variables that are written at the left side of the table are put equal to zero.*

b) *The variables written at the top of the table are given the values of the last elements in their columns.*

The values of the variables

$$\begin{aligned} &\delta\xi_i && i = 0, 1, \ldots, m \\ &\delta\eta_k && k = 0, 1, \ldots, n \\ &Y(x_j),\ Y^*(x_j) && j = 1, 2, \ldots, \mu + \nu + 2 \end{aligned}$$

and Z

for the optimal solution of the linear program (50) are thus all defined.

5. The problem of degeneracy:

The type of degeneracy encountered here is that two or more of the calculated characteristic quotients become equal to zero, so that the pivot column be always chosen in their columns and the value of the objective function Z remains constant which means that the simplex algorithm never arrives to an end. This type of degeneracy can occur in one of the following situations:

1. At the beginning of the first simplex-algorithm in the first exchange step, if the initial solution

$$(\vec{\xi}^0, \vec{\eta}^0) = \vec{0}$$

is chosen. This situation can be avoided if any roughly estimated initial solution is taken.

2. During the simplex algorithm, if the $m + n + 3$ variables at the left side of the table become all of the type Y or Y^*, then the characteristic quotients calculated for the other variables of the same type must be all equal to zero [because the current solution given by the table in this situation must be $(\vec{\xi}, \vec{\eta}, \zeta) = \vec{0}$].

However, it was practically found in all such situations that the characteristic quotients have very small values instead of zeros (due to the round-off errors made during the previous A.T.-steps). Thus, it was possible to continue the simplex algorithm as usual as if no degeneracy had occured.

If, however, it is not possible to choose an initial solution different from $\vec{0}$ (to avoid the first situation), or if the round-off errors do not save the second situation, the following method can be used:

1. The elements of the last row are stored to be used later.

2. The columns for which the characteristic quotients are zero are determined (let these be called the degenerated columns); their zero elements in the last row are then replaced by positive numbers (these are thought to be multiplied by a common factor ε which will be later put equal to zero).

3. The simplex algorithm is continued as before with the exception that the pivot column is to be chosen only among the degenerated columns; as long as such a pivot column can be found.

4. If no pivot can be found among the degenerated columns (this must happen after a finite number of A.T.-steps, since the first element in the last row is always increased and a maximum value must be reached), the following is then done:

a) The elements of the last row are replaced by their original values stored by step 1. (This is equivalent to putting ε equal to zero.)

b) The pivot column is then chosen by the usual simplex algorithm (this must be one different from the degenerated columns), and the simplex algorithm is continued.

6. An example:

The results obtained for a filter with 2 pass-bands and 3 stop-bands are discussed here. The 8 local optimal functions were calculated. The final results are given by curves for these functions together with their zeros and poles. The exchange steps required to obtain the best local optimal function (i.e. *the global optimal function*) are also discussed.

The filter requirements:

stop-bands	pass-bands
S_1: $0 \leq x \leq 1$	D_1: $1.3 \leq x \leq 2$
S_2: $2.5 \leq x \leq 4$	D_2: $4.5 \leq x \leq 6$
S_3: $6.5 \leq x$	

The class F of functions:

$$R(x) = \frac{P_5(x)}{\sqrt{x}\,Q_4(x)}$$

$$(m = 5,\ n = 4,\ g(x) = 1/\sqrt{x})\ .$$

The following choice was made for the basis functions: [The zeros of $p_i(x)$, with the exception of $p_1(x)$, lie in the pass-bands; and the zeros of $q_k(x)$ lie in the stop-bands]:

$$p_0(x) = 1$$

$$p_1(x) = (x - 3.65)$$

$$p_2(x) = (x - 1.65)\,(x - 5.25)$$

$$p_3(x) = (x - 1.65)\,(x - 4.719\,66)\,(x - 5.780\,33)$$

$$p_4(x) = (x - 1.402\,65)\,(x - 1.897\,45)\,(x - 4.719\,66)\,(x - 5.780\,33)$$

$$p_5(x) = (x - 1.402\,65)\,(x - 1.897\,45)\,(x - 4.600\,48)\,(x - 5.899\,52)\,(x - 5.25)$$

$$q_0(x) = \sqrt{x}$$

$$q_1(x) = \sqrt{x}\,(x - 3.25)$$

$$q_2(x) = \sqrt{x}\,(x - 3.25)\,(x - 8.034\,44)$$

$$q_3(x) = \sqrt{x}\,(x - 0.809\,02)\,(x - 3.25)\,(x - 8.034\,44)$$

$$q_4(x) = \sqrt{x}\,(x - 0.809\,02)\,(x - 2.719\,66)\,(x - 3.780\,33)\,(x - 8.034\,44)\,.$$

The local classes: The number of local classes for the given number of pass- and stop-bands is given by

$$N_F = 2^{(N_D + N_S - 2)} = 8\,;$$

these are defined by the following sign-distributions (the local T-deviation for each local class is also given in the table):

local class	σ_{S_1}	σ_{D_1}	σ_{S_2}	σ_{D_2}	σ_{S_3}	the local T-deviation
F_1	+ 1	+ 1	+ 1	+ 1	+ 1	0.0901
F_2	+ 1	+ 1	+ 1	+ 1	− 1	0.0690
F_3	+ 1	+ 1	+ 1	− 1	+ 1	0.1389
F_4	+ 1	+ 1	+ 1	− 1	− 1	0.1098
F_5	+ 1	+ 1	− 1	+ 1	+ 1	0.2597
F_6	+ 1	+ 1	− 1	+ 1	− 1	0.0806
F_7	+ 1	+ 1	− 1	− 1	+ 1	0.3802
F_8	+ 1	+ 1	− 1	− 1	− 1	0.1587

In order to obtain *the global optimal function* (the local optimal function for the local class F_2) the following exchange steps were used:

The first exchange step: A discrete set with the following points was chosen (their number exceeds by 3 the number required for a reference, i.e. $m + n + 2 = 11$):

in D_1: $xd_0 = 1.3$ $xd_1 = 1.65$ $xd_2 = 2$	*in* S_1: $xs_0 = 0.5$ $xs_1 = 1$
in D_2: $xd_3 = 4.5$ $xd_4 = 4.875$	*in* S_2: $xs_2 = 2.5$ $xs_3 = 3.25$
	$xs_4 = 4$
$xd_5 = 5.625$ $xd_6 = 6$	*in* S_3: $xs_5 = 6.5$ $xs_6 = 13$

For the first simplex algorithm the following choice for l was made

$$l_0 = 0.25 .$$

The initial solution $(\vec{\xi}^0, \vec{\eta}^0) = \vec{0}$ was chosen; the resulting simplex algorithm was thus *degenerated* at the beginning and the method described above for the treatment of the degeneracy was used. At the end of the simplex algorithm the function in Figure 4 was obtained. This is an almost leveled

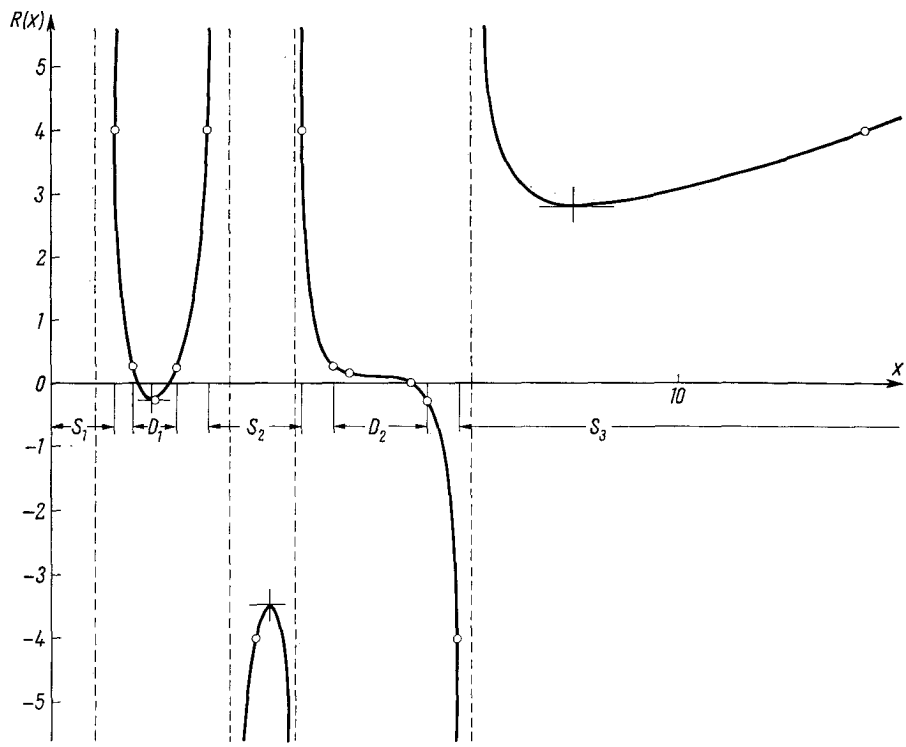

Figure 4
The first local discrete optimal function
o: Points of the first discrete set

reference function with respect to a reference consisting of 11 points contained in the discrete set as seen in Figure 4.

The second exchange step: The maximum and minimum points (including those at the ends of the intervals) of the function obtained from the first exchange step were calculated; then out of these the following discrete set was chosen (which is actually a reference)

$$xs_0 = 1 \quad xs_1 = 2.5 \quad xs_2 = 3.475 \quad xs_3 = 4 \quad xs_4 = 6.5 \quad xs_5 = 8.25$$
$$xd_0 = 1.3 \quad xd_1 = 1.6 \quad xd_2 = 2 \quad xd_3 = 4.5 \quad xd_4 = 6 .$$

For this discrete set *the local discrete problem* was again solved, which needed a number of 4 simplex algorithms to obtain the function shown in Figure 5.2. This function was taken as the local optimal function of the local class F_2. The local optimal functions for the other local classes were also calculated in the same way; the results are shown in Figures 5.1–5.8 and prove clearly that the local optimal function of F_2 is *the global optimal function* for the given filter.

Discussion of the results:

The local optimal functions obtained have the following features which have no similar in the classical case of a low-pass (or a high-pass) filter or in the case of a band-pass (or a band-stop) filter:

1. *Complex zeros or poles.* Actually, the global optimal function has a pair of complex zeros.

2. *Zeros or poles that lie between pass- and stop-bands.* The third best local optimal function $R_1(x)$ has a pole between the first pass-band and the second stop-band.

3. *Negative zeros or poles.* For example the function $R_1(x)$ has a negative zero, and the function $R_4(x)$ has a negative pole.

4. *The ends of the intervals are not necessarily extreme points of the local optimal functions.* (For each local optimal function there are *exactly* 11 extreme points as shown in Figure 5.) The end of the first pass-band and the beginning of the second stop-band are not extreme points for the local optimal function $R_1(x)$; the first pass-band may be extended to $x = 2.045$ and the second stop-band may begin in $x = 2.16$ and the same local optimal function will be obtained.

It may be a serious loss of time to calculate all the local optimal functions whose number N_F may be very large ($N_F = 2^{(N_D+N_S-2)}$ increases *exponentially* with the total number of pass- and stop-bands). But on the other hand, if the best function has complex zeros or poles the problem of realisation of the electrical filter becomes more complicated and *it may be found more convenient to realise a second or a third best filter with no complex zeros or poles.*

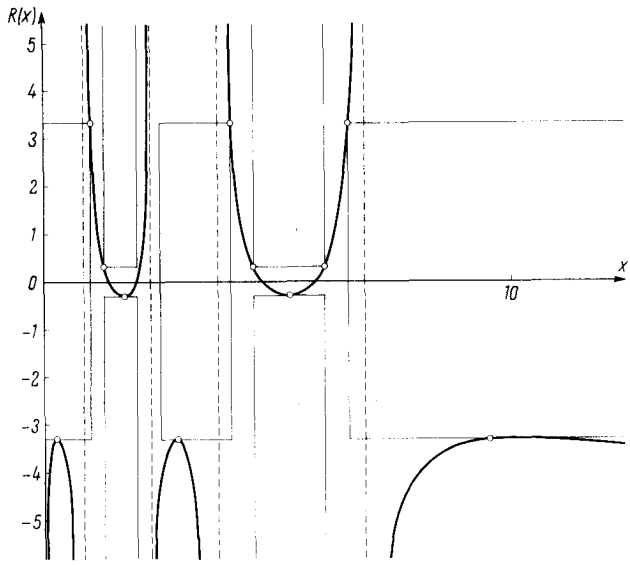

Figure 5.1
The local optimal function $R_1(x)$
Zeros: $-0.767\,202$; $1.409\,23$; $1.972\,79$; $4.691\,27$; $5.819\,47$ – *Poles:* $0.834\,102$; $2.282\,79$; $3.723\,50$; $6.829\,08$ – *Maximum deviation:* $\Delta_1 = 0.0901$ – ∘: Extreme points

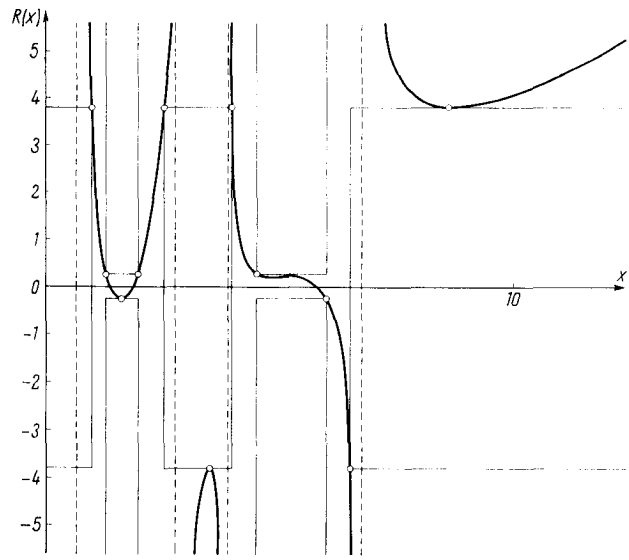

Figure 5.2
The local optimal function $R_2(x)$ (= the global optimal function)
Zeros: $1.376\,43$; $1.880\,65$; $5.814\,67$; $(4.479\,14 \pm 0.491\,382\,i)$ – *Poles:* $0.663\,276$; $2.782\,75$; $3.892\,37$; $6.745\,84$ – *Maximum deviation:* $\Delta_2 = 0.0690 = \Delta^*$ – ∘: Extreme points

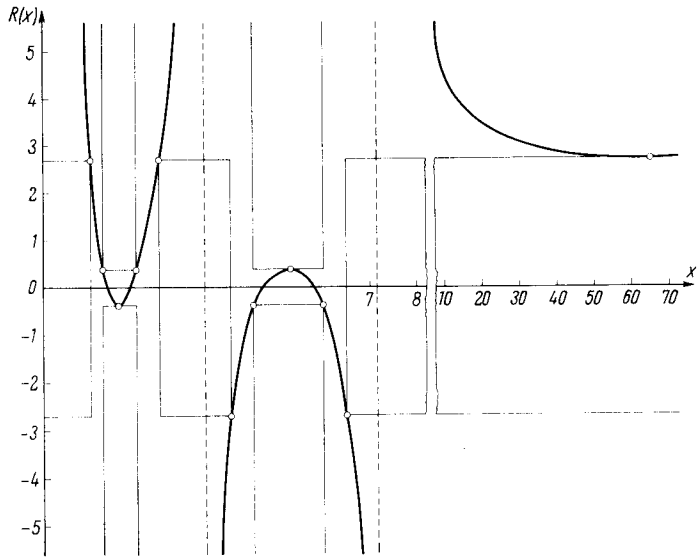

Figure 5.3
The local optimal function $R_3(x)$
Zeros: -92.6175; 1.389 27; 1.878 49; 4.699 04; 5.803 41 – *Poles:* 3.427 32; 7.154 67; $(1.188\ 04 \pm 1.407\ 36\ i)$ – *Maximum deviation:* $\Delta_3 = 0.1389$ – o: Extreme points

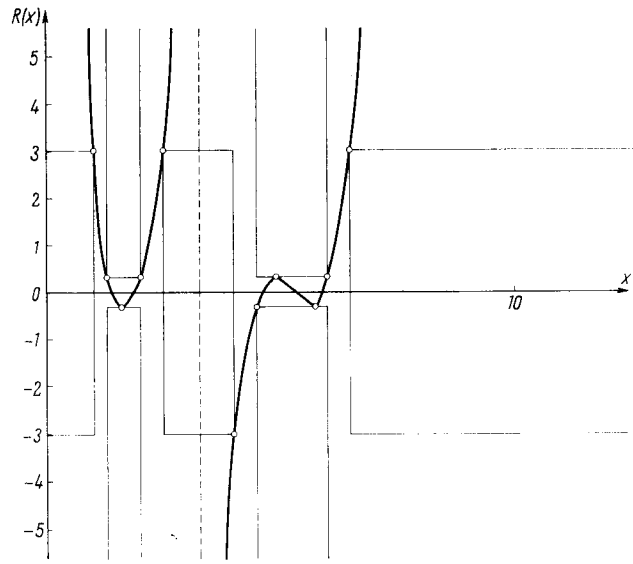

Figure 5.4
The local optimal function $R_4(x)$
Zeros: 1.385 10; 1.881 67; 4.629 82; 5.326 63; 5.900 12 – *Poles:* $-0.440\ 082$; 3.277 47; $(5.736\ 03 \pm 1.009\ 19\ i)$ – *Maximum deviation:* $\Delta_4 = 0.1098$ – o: Extreme points

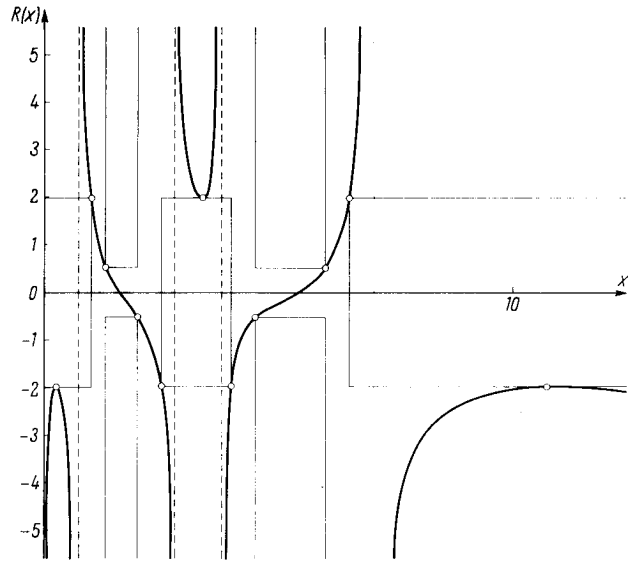

Figure 5.5
The local optimal function $R_5(x)$
Zeros: $-0.459\,010$; $1.592\,40$; $5.466\,57$; $(3.666\,47 \pm 1.072\,98\,i)$ – *Poles:* $0.743\,846$; $2.798\,79$; $3.796\,23$; $7.037\,98$ – *Maximum deviation:* $\Delta_5 = 0.2597$ – ∘: Extreme points

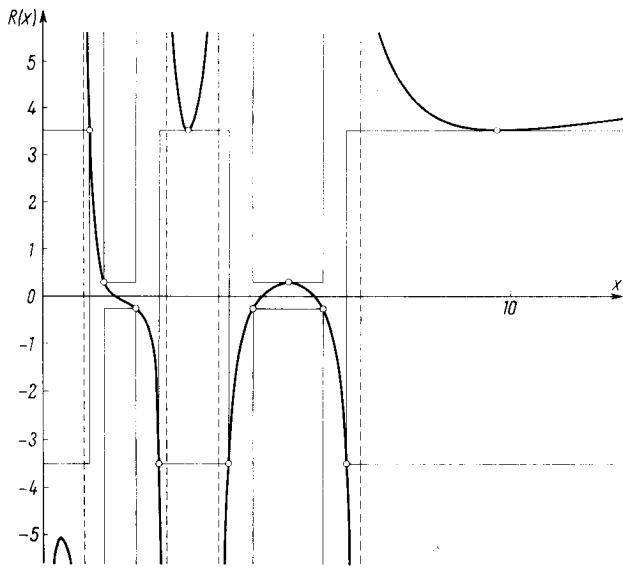

Figure 5.6
The local optimal function $R_6(x)$
Zeros: $1.498\,27$; $4.694\,31$; $5.826\,21$; $(1.861\,46 \pm 0.617\,012\,i)$ – *Poles:* $0.871\,519$; $2.638\,61$; $3.776\,05$; $6.800\,00$ – *Maximum deviation:* $\Delta_6 = 0.0806$ – ∘: Extreme points

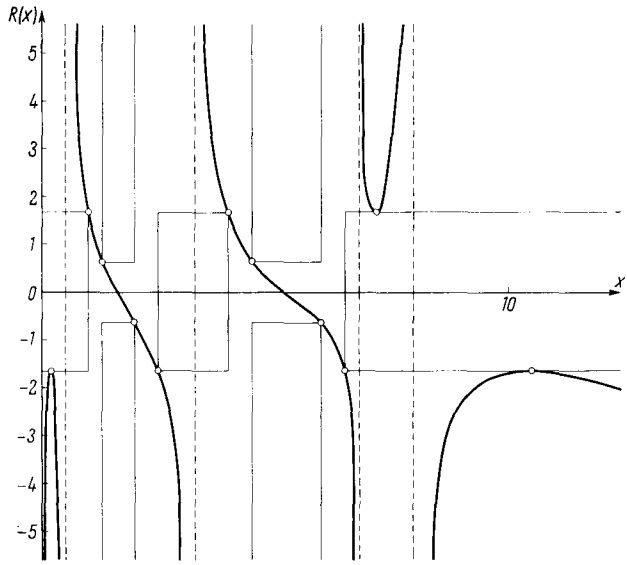

Figure 5.7
The local optimal function $R_7(x)$
Zeros: $-0.038\,517$; $1.614\,91$; $5.159\,71$; $(7.331\,12 \pm 0.782\,56\,i)$ – *Poles:* $0.520\,06$; $3.303\,15$; $6.814\,87$; $7.973\,17$ – *Maximum deviation:* $\Delta_7 = 0.3802$ – ∘: Extreme points

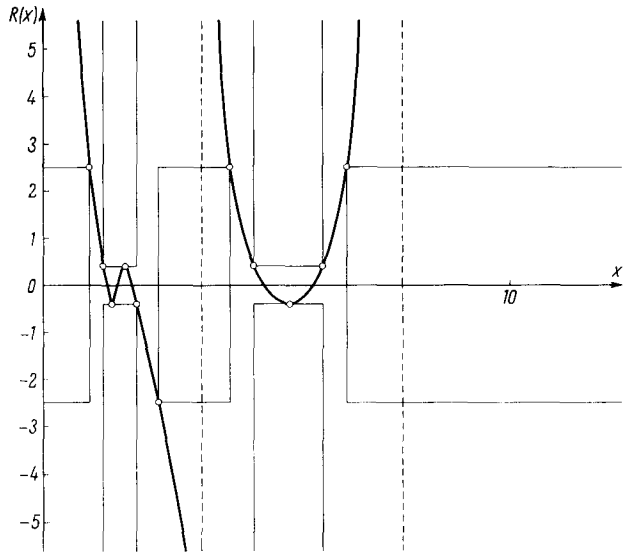

Figure 5.8
The local optimal function $R_8(x)$
Zeros: $1.348\,97$; $1.588\,87$; $1.894\,33$; $4.711\,72$; $5.803\,56$ – *Poles:* $3.394\,63$; $7.673\,39$; $(1.522\,69 \pm 0.216\,137\,i)$ – *Maximum deviation:* $\Delta_8 = 0.1587$ – ∘: Extreme points

Chapter IV. Mixed-Integer Programming Methods

In the preceding chapters the approximation problem of an electrical filter with more than 3 pass- and stop-bands has been solved by calculating *each one* of the N_F local optimal functions *seperately*; and then among these the global optimal function has been chosen. In this chapter the possibility of *directly determining the global optimal function* with the help of mixed-integer programming is studied. The exchange algorithm, as discussed in Chapter II, can not be expected to converge here; in fact, it has to be modified in the following way in order that it delivers the required result.

1. The modified exchange algorithm

The k-th iteration of *the modified exchange algorithm* consists of the following steps:

1. For the discrete set:

$$(d, s)^{(k)} = (d^{(k)}, s^{(k)})_{f_1} \cup (d^{(k)}, s^{(k)})_{f_2} \cup \cdots \cup (d^{(k)}, s^{(k)})_{f_{(\theta^{(k)})}} \tag{62}$$

corresponding to the local classes:

$$F_{f_1}, F_{f_2}, \ldots, F_{f_{(\theta^{(k)})}} \tag{63}$$

the following *global discrete approximation problem* is to be solved:

$$\left.\begin{array}{l} \textit{To choose a function } R^{(k)}(x) \in F_1 \cup F_2 \cup \cdots \cup F_{N_F} \\ \textit{such that its maximum discrete deviation in the set} \\ (d, s)^{(k)} \textit{ be a minimum.} \end{array}\right\} \tag{64}$$

The resulting *global discrete optimal function* must be one of the *local discrete optimal functions*; let it be that of the local class F_{f_j} [this must not necessarily be one of the local classes (63)].

2. The local maximum and minimum points of the function $R^{(k)}(x)$ are determined; then the new reference $(\bar{d}^{(k+1)}, \bar{s}^{(k+1)})_{f_j}$ and the number $\alpha_{f_j}^{(k)}$ are determined in the same way as in step 2 of the exchange algorithm discussed in Chapter II. It is obvious that *the global T-deviation* Δ^* satisfies the inequality:

$$\alpha_{f_j}^{(k)} \leq \Delta^* < \Delta\big(R^{(k)}(x)\big) . \tag{65}$$

If these bounds are close enough to each other, then the function $R^{(k)}(x)$ is *practically the global optimal function.*

3. a) If the local class F_{f_j} in which the function $R^{(k)}(x)$ lies is one of the local classes (63), then replace the set $(d^{(k)}, s^{(k)})_{f_j}$ by the new set $(d^{(k+1)}, s^{(k+1)})_{f_j}$ obtained in step 2; the remaining sets are left as they are. This defines *the new discrete set:*

$$\left.\begin{aligned}
&(d, s)^{(k+1)} = (d^{(k+1)}, s^{(k+1)})_{f_1} \cup (d^{(k+1)}, s^{(k+1)})_{f_2} \cup \cdots \\
&\qquad (d^{(k+1)}, s^{(k+1)})_{f_j} \cup \cdots \cup (d^{(k+1)}, s^{(k+1)})_{f_{(\theta^{(k+1)})}} \\
&\textit{where } \theta^{(k+1)} = \theta^{(k)} \\
&\textit{and } (d^{(k+1)}), s^{(k+1)})_{f_i} = (d^{(k)}, s^{(k)})_{f_i} \\
&\quad \text{also } \alpha_{f_i}^{(k+1)} = \alpha_{f_i}^{(k)}
\end{aligned}\right\} \text{for } f_i \neq f_j . \tag{66a}$$

b) If the local class F_{f_j} is not one of the local classes (63), then let this be the new local class $F_{f_{(\theta^{(k+1)})}} (= F_{f_j})$. The set $(d^{(k+1)}, s^{(k+1)})_{f_j}$ is then combined to the set $(d, s)^{(k)}$ to get *the new discrete set:*

$$\left.\begin{aligned}
&(d, s)^{(k+1)} = (d^{(k+1)}, s^{(k+1)})_{f_1} \cup (d^{(k+1)}, s^{(k+1)})_{f_2} \cup \cdots \\
&\qquad \cup (d^{(k+1)}, s^{(k+1)})_{f_i} \cup \cdots \cup (d^{(k+1)}, s^{(k+1)})_{f_{(\theta^{(k+1)})}} \\
&\textit{where} \quad \theta^{(k+1)} = \theta^{(k)} + 1 \\
&\qquad (d^{(k+1)}, s^{(k+1)})_{f_{(\theta^{(k+1)})}} = (d^{(k+1)}, s^{(k+1)})_{f_j} \\
&\textit{and} \quad (d^{(k+1)}, s^{(k+1)})_{f_i} = (d^{(k)}, s^{(k)})_{f_i} \\
&\qquad \text{also } \alpha_{f_i}^{(k+1)} = \alpha_{f_i}^{(k)}
\end{aligned}\right\} i = 1, 2, \ldots, \theta^{(k)} . \tag{66b}$$

With this new discrete set (66a) or (66b) step 1 above is then repeated.

At the beginning of the algorithm, the numbers $\alpha_1^{(0)}, \alpha_2^{(0)}, \ldots, \alpha_{N_F}^{(0)}$ are given all the value zero; a discrete set $(d^{(0)}, s^{(0)})$ is chosen for a start and then left out in the following exchange steps.

The total number $\theta^{(k)}$ of sets required can be one or more sets but a too large number near the maximum number $N_F = 2^{(N_D + N_S - 2)}$ is not to be expected (in the example mentioned at the end of this chapter it was found that one set was sufficient; the maximum possible number is $N_F = 4$).

It is easily seen that this exchange algorithm leads to the required result; in fact, using the same considerations as in Chapter II, it is obvious that the

numbers $\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(k)}, \ldots$ defined by:

$$\alpha^{(k)} = \max_{f} (\alpha_1^{(k)}, \alpha_2^{(k)}, \ldots, \alpha_f^{(k)}, \ldots, \alpha_{N_F}^{(k)}) \tag{67}$$

build *an increasing sequences of lower bounds for the global T-deviation.* The total number of *exchange steps,* required to obtain *the global optimal function,* is in any case not greater than the sum of the numbers of exchange steps required to get the local optimal functions of the $\theta^{(k)}$ local classes in (63), each function being obtained seperately using the exchange algorithm of Chapter II.

2. The global discrete approximation problem

In order to solve *the global discrete approximation problem* (64) as defined in step 1 of the modified exchange algorithm, a number of *mixed-integer programs* are used in a way similar to that discussed in Chapter III-B for the solution of *the local discrete problem* using linear programming*).

The mixed-integer program for a given discrete set $(d, s)^{(k)}$ and a given number l is constructed in the following way:

The given discrete set $(d, s)^{(k)}$ is devided into a number of subsets

$$\left.\begin{aligned} & xD_1, xD_2, \ldots, xD_\alpha, \ldots, xD_{N_D} \\ \text{and} \quad & xS_1, xS_2, \ldots, xS_\beta, \ldots, xS_{N_S} \end{aligned}\right\} \tag{68}$$

such that the points of $(d, s)^{(k)}$ which lie in D_α build the subset xD_α and those which lie in S_β build the subset xS_β. With the given initial value for l (which is determined in a similar way as in Chapter III-B) the following inequalities are constructed:

$$\left.\begin{aligned} & -l \leq \frac{(\vec{\xi}, \vec{p}(x))}{(\vec{\eta}, \vec{q}(x))} \leq l \quad \text{for all } x \in xD_1 \cup xD_2 \cup \cdots \cup xD_{N_D} \\ \text{and} \quad & -l \leq \frac{(\vec{\eta}, \vec{q}(x))}{(\vec{\xi}, \vec{p}(x))} \leq l \quad \text{for all } x \in xS_1 \cup xS_2 \cup \cdots \cup xS_{N_S}. \end{aligned}\right\} \tag{69}$$

*) It may be a faster method (but more complicated), to use *at the beginning one mixed-integer program* with the given initial value for l in order to determine a local class F_j in which the global optimal function $R^{(k)}(x)$ is expected to lie; then for this constant local class F_j a number of *linear programs* as explained in Chapter III-B are used to get an exact value for l; then *at the end a second mixed-integer program* is used to check if for this new value of l the function $R^{(k)}(x)$ actually lies in the local class F_j.

These are equivalent to the following system, if the function $R(x)$ defined by the vector $(\vec{\xi}, \vec{\eta})$ is to lie in one of the local classes $F_1, F_2, \ldots, F_{N_F}$:

$$
\left.
\begin{array}{l}
\textit{for all } x \in xD_1\textit{:} \\
\quad -(\vec{\xi}, \vec{p}(x)) + l(\vec{\eta}, \vec{q}(x)) \geq 0 \\
\text{and} \quad (\vec{\xi}, \vec{p}(x)) + l(\vec{\eta}, \vec{q}(x)) \geq 0 \\
\\
\textit{for all } x \in xS_1\textit{:} \\
\quad l(\vec{\xi}, \vec{p}(x)) - (\vec{\eta}, \vec{q}(x)) \geq 0 \\
\text{and} \quad l(\vec{\xi}, \vec{p}(x)) + (\vec{\eta}, \vec{q}(x)) \geq 0 \\
\\
\textit{for all } x \in xD_\alpha\textit{:} \\
\begin{array}{l|l}
\textit{either} & \textit{or} \\
-(\vec{\xi}, \vec{p}(x)) + l(\vec{\eta}, \vec{q}(x)) \geq 0 & (\vec{\xi}, \vec{p}(x)) - l(\vec{\eta}, \vec{q}(x)) \geq 0 \\
\text{and} \quad (\vec{\xi}, \vec{p}(x)) + l(\vec{\eta}, \vec{q}(x)) \geq 0 & \text{and} \quad -(\vec{\xi}, \vec{p}(x)) - l(\vec{\eta}, \vec{q}(x)) \geq 0
\end{array} \\
\\
\textit{for all } x \in xS_\beta\textit{:} \\
\begin{array}{l|l}
\textit{either} & \textit{or} \\
l(\vec{\xi}, \vec{p}(x)) - (\vec{\eta}, \vec{q}(x)) \geq 0 & -l(\vec{\xi}, \vec{p}(x)) + (\vec{\eta}, \vec{q}(x)) \geq 0 \\
\text{and} \quad l(\vec{\xi}, \vec{p}(x)) + (\vec{\eta}, \vec{q}(x)) \geq 0 & \text{and} \quad -l(\vec{\xi}, \vec{p}(x)) - (\vec{\eta}, \vec{q}(x)) \geq 0
\end{array} \\
\\
\text{where } \alpha = 2, 3, \ldots, N_D \\
\text{and} \quad \beta = 2, 3, \ldots, N_S\,.
\end{array}
\right\} \quad (70)
$$

In order to choose the proper inequalities, for the points in the pass-bands D_α ($\alpha \geq 2$) and for the points in the stop-bands S_β ($\beta \geq 2$), out of the above system (70) and to construct a solution for these, a number of $N_D + N_S - 2$ *integer variables* are introduced

$$\varrho_{D_2}, \varrho_{D_3}, \ldots, \varrho_{DN_D} \quad \text{and} \quad \varrho_{S_2}, \varrho_{S_3}, \ldots, \varrho_{SN_S}$$

and a variable ζ is introduced in the same way as in Chapter III-B. The following *mixed-integer program* is then solved:

To determine values for: $\vec{\xi}, \vec{\eta}, \zeta$

$\varrho_{D_\alpha}\ (\alpha = 2, 3, \ldots, N_D)$

and $\varrho_{S_\beta}\ (\beta = 2, 3, \ldots, N_S)$

maximizing the objective-function: $Z = -\zeta$,

subject to the following restrictions:

a1) *for all* $x \in xD_1$:

$$Y = -(\vec{\xi}, \vec{p}(x)) + l(\vec{\eta}, \vec{q}(x)) + G(x) \cdot \zeta \geq 0$$

and $$Y^* = (\vec{\xi}, \vec{p}(x)) + l(\vec{\eta}, \vec{q}(x)) + G(x) \cdot \zeta \geq 0$$

where $G(x) = \sqrt{\vec{p}^2(x) + l^2\, \vec{q}^2(x)}$.

a2) *for all* $x \in xS_1$:

$$Y = l(\vec{\xi}, \vec{p}(x)) - (\vec{\eta}, \vec{q}(x)) + G(x) \cdot \zeta \geq 0$$

and $$Y^* = l(\vec{\xi}, \vec{p}(x)) + (\vec{\eta}, \vec{q}(x)) + G(x) \cdot \zeta \geq 0$$

where $G(x) = \sqrt{l^2\, \vec{p}^2(x) + \vec{q}^2(x)}$.

a3) *for all* $x \in xD_\alpha\ (\alpha = 2, 3, \ldots, N_D)$: (71)

$$Y = -(\vec{\xi}, \vec{p}(x)) + l(\vec{\eta}, \vec{q}(x)) + G(x) \cdot \zeta + K(x) \cdot \varrho_{D_\alpha} \geq 0$$

and $$Y^* = (\vec{\xi}, \vec{p}(x)) + l(\vec{\eta}, \vec{q}(x)) + G(x) \cdot \zeta + K(x) \cdot \varrho_{D_\alpha} \geq 0$$

$$Y = (\vec{\xi}, \vec{p}(x)) - l(\vec{\eta}, \vec{q}(x)) + G(x) \cdot \zeta + K(x) \cdot (1 - \varrho_{D_\alpha}) \geq 0$$

and $$Y^* = -(\vec{\xi}, \vec{p}(x)) - l(\vec{\eta}, \vec{q}(x)) + G(x) \cdot \zeta + K(x) \cdot (1 - \varrho_{D_\alpha}) \geq 0$$

where $G(x) = \sqrt{\vec{p}^2(x) + l^2\, \vec{q}^2(x)}$

and $$K(x) = 2\, l \cdot \sum_{k=0}^{n} | q_k(x) \quad .$$

a4) *for all* $x \in xS_\beta\ (\beta = 2, 3, \ldots, N_S)$:

$$Y = l(\vec{\xi}, \vec{p}(x)) - (\vec{\eta}, \vec{q}(x)) + G(x) \cdot \zeta + K(x) \cdot \varrho_{S_\beta} \geq 0$$

$$Y^* = l(\vec{\xi}, \vec{p}(x)) + (\vec{\eta}, \vec{q}(x)) + G(x) \cdot \zeta + K(x) \cdot \varrho_{S_\beta} \geq 0$$

$$Y = -l(\vec{\xi}, \vec{p}(x)) + (\vec{\eta}, \vec{q}(x)) + G(x) \cdot \zeta + K(x) \cdot (1 - \varrho_{S_\beta}) \geq 0$$

and $$Y^* = -l(\vec{\xi}, \vec{p}(x)) - (\vec{\eta}, \vec{q}(x)) + G(x) \cdot \zeta + K(x) \cdot (1 - \varrho_{S_\beta}) \geq 0$$

where $G(x) = \sqrt{l^2\, \vec{p}^2(x) + \vec{q}^2(x)}$

and $$K(x) = 2\, l \cdot \sum_{i=0}^{m} | p_i(x)$$

$$\left.\begin{array}{ll} \text{b)} & \varrho_{D_\alpha} = 0 \quad \text{or} \quad 1 \quad \alpha = 2, 3, \ldots, N_D \\ \text{and} & \varrho_{S_\beta} = 0 \quad \text{or} \quad 1 \quad \beta = 2, 3, \ldots, N_S \\ \text{and} & \\ \text{c)} & |\xi_i| \leq 1 \quad i = 0, 1, 2, \ldots, m \\ \text{and} & |\eta_k| \leq 1 \quad k = 0, 1, 2, \ldots, n\,. \end{array}\right\} \tag{71}$$

The following Theorem 7 defines the relation between the optimal solution of the mixed-integer programming problem (71) and the optimal solutions of the N_F possible linear programs (50).

Theorem 7: *The mixed-integer programming problem* (71) *has an optimal solution:*

$$\hat{\xi}, \hat{\eta}, \hat{\zeta}; \quad \hat{\varrho}_{D_2}, \hat{\varrho}_{D_3}, \ldots, \hat{\varrho}_{D_{N_D}}; \quad \hat{\varrho}_{S_2}, \hat{\varrho}_{S_3}, \ldots, \hat{\varrho}_{S_{N_S}}$$

if and only if the linear program (50) *defined by the sign-distribution:*

$$\left.\begin{array}{l} \sigma_{D_\alpha} = \left\{\begin{array}{ll} +1 & \text{if} \quad \hat{\varrho}_{D_\alpha} = 0 \\ -1 & \text{if} \quad \hat{\varrho}_{D_\alpha} = 1 \end{array}\right\} \alpha = 2, 3, \ldots, N_D \\ \sigma_{S_\beta} = \left\{\begin{array}{ll} +1 & \text{if} \quad \hat{\varrho}_{S_\beta} = 0 \\ -1 & \text{if} \quad \hat{\varrho}_{S_\beta} = 1 \end{array}\right\} \beta = 2, 3, \ldots, N_S \end{array}\right\} \tag{72}$$

has an optimal solution: $\hat{\xi}, \hat{\eta}, \hat{\zeta}$, *such that* $-\hat{\zeta}$ *is the largest among the* N_F *values for* $-\zeta$ *of the optimal solutions for the* N_F *possible linear programs* (50).

The proof of this theorem follows directly from the following Lemma:

Lemma: *A choice of the integer variables in* (71) *subject to the conditions* b) *in* (71) *reduces the inequalities* a) *in* (71) *to the inequalities* (47a) *of a linear program* (50) *for which the sign-distribution is defined according to* (72) *by the choice of the integer variables.*

Proof of the Lemma:

1. *The inequalities* a1) *and* a2) *in* (71) *are the same as those in* (47a) *for the same points* x.

2. *For the inequalities* a3):

(i) *If* $\varrho_{D_\alpha} = 0$, then the first two inequalities in a3) become:

$$\left.\begin{aligned} &-(\vec{\xi}, \vec{p}(x)) + l(\vec{\eta}, \vec{q}(x)) + G(x)\cdot\zeta \geq 0 \\ \text{and}\quad &(\vec{\xi}, \vec{p}(x)) + l(\vec{\eta}, \vec{q}(x)) + G(x)\cdot\zeta \geq 0 \end{aligned}\right\} \tag{73i}$$

which are the corresponding inequalities in (47a) for $\sigma_{D_\alpha} = +1$. The third and fourth inequalities: $\bar{Y} \geq 0$ and $\bar{Y}^* \geq 0$ can be cancelled since they follow directly from the inequalities (73i) in the following manner:

From $-(\vec{\xi}, \vec{p}(x)) + l(\vec{\eta}, \vec{q}(x)) + G(x)\cdot\zeta \geq 0$ it follows that:

$$\begin{aligned} \bar{Y}^* &= -(\vec{\xi}, \vec{p}(x)) - l(\vec{\eta}, \vec{q}(x)) + G(x)\cdot\zeta + K(x) \\ &\geq -2\,l\cdot(\vec{\eta}, \vec{q}(x)) + K(x)\,. \end{aligned}$$

Substituting for $K(x)$ it follows:

$$\bar{Y}^* \geq K(x) - 2\,l\cdot(\vec{\eta}, \vec{q}(x)) = 2\,l\cdot\left(\sum_{k=0}^{n} |\,q_k(x)\,| - \sum_{k=0}^{n} \eta_k\cdot q_k(x)\right) \geq 0$$

since $\vec{\eta}$ satisfies the conditions c) in (71) [or (47b)].

Similarly, from $(\vec{\xi}, \vec{p}(x)) + l(\vec{\eta}, \vec{q}(x)) + G(x)\cdot\zeta \geq 0$ it follows that $\bar{Y} \geq 0$.

(ii) *If* $\varrho_{D_\alpha} = 1$, then the last two inequalities in a3) become:

$$\left.\begin{aligned} &(\vec{\xi}, \vec{p}(x)) - l(\vec{\eta}, \vec{q}(x)) + G(x)\cdot\zeta \geq 0 \\ \text{and}\quad &-(\vec{\xi}, \vec{p}(x)) - l(\vec{\eta}, \vec{q}(x)) + G(x)\cdot\zeta \geq 0 \end{aligned}\right\} \tag{73ii}$$

which are the corresponding inequalities in (47a) for $\sigma_{D_\alpha} = -1$. The first and second inequalities: $Y \geq 0$ and $Y^* \geq 0$ can be cancelled since they follow directly from the inequalities (73ii); namely $Y \geq 0$ follows from the first and $Y^* \geq 0$ follows from the second inequality.

3. *For the inequalities* a4):

It can be similarly shown that:

(i) *If* $\varrho_{S_\beta} = 0$, then they are equivalent to the two inequalities:

$$\left.\begin{aligned} &l(\vec{\xi}, \vec{p}(x)) - (\vec{\eta}, \vec{q}(x)) + G(x)\cdot\zeta \geq 0 \\ \text{and}\quad &l(\vec{\xi}, \vec{p}(x)) + (\vec{\eta}, \vec{q}(x)) + G(x)\cdot\zeta \geq 0 \end{aligned}\right\} \tag{74i}$$

which are the corresponding inequalities in (47a) for $\sigma_{S_\beta} = +1$.

(ii) *If* $\varrho_{S\beta} = 1$, then the inequalities a4) are equivalent to:

$$\left.\begin{aligned} &-l(\vec{\xi}, \vec{p}(x)) + (\vec{\eta}, \vec{q}(x)) + G(x) \cdot \zeta \geq 0 \\ \text{and} \quad &-l(\vec{\xi}, \vec{p}(x)) - (\vec{\eta}, \vec{q}(x)) + G(x) \cdot \zeta \geq 0 \end{aligned}\right\} \tag{74ii}$$

which are the corresponding inequalities in (47a) for $\sigma_{S\beta} = -1$.

Proof of Theorem 7:

Since each possible choice for the values of the integer-variables [subject to conditions (c) in (71)] reduces the inequalities (a) in (71) to the inequalities (47a) of one of the linear programs (50), and since the inequalities (c) in (71) are the same as those of (47b) of the linear program (50), and the objective function is the same in both problems [the linear program (50) and the mixed-integer program (71)], it follows that the optimal solution of the mixed-integer program must have the same values for $\hat{\vec{\xi}}, \hat{\vec{\eta}}, \hat{\zeta}$ as those of the optimal solution for one of the linear programs (50), namely for the linear program (50) whose optimal solution has the largest value for the objective function $-\zeta$.

From Theorem 7, and Theorem 6 of Chapter III-B, it follows that:

a) If there exists in any of the local classes $F_1, F_2, \ldots, F_{N_F}$ a function

$$R(x) = \frac{(\vec{\xi}, \vec{p}(x))}{(\vec{\eta}, \vec{q}(x))}$$

which satisfies the conditions (69), then such a function is given by the vectors $\hat{\vec{\xi}}$ and $\hat{\vec{\eta}}$ of the optimal solution of the mixed-integer program; at the same time this function is chosen out of the local class which is *most liable* to contain a solution for the global discrete approximation problem (since the solution with the largest value for $-\zeta$ is chosen by the mixed-integer program).

b) If it is not possible to construct a function $R(x)$ which satisfies the conditions (69), then the mixed-integer program has the optimal solution $(\hat{\vec{\xi}}, \hat{\vec{\eta}}, \hat{\zeta}) = \vec{0}$. In this case, the mixed-integer program is to be newly solved with a larger value for l.

In order to solve the mixed-integer program (71), an algorithm given by R. GOMORY [6] was used. *However, it was found that the time needed to solve one mixed-integer program was too long to make this method of any practical advantage over that given in Chapter III-B.*

The method (described in this chapter) is hoped to become useful if another mixed-integer programming method is developed, which consumes less calculating time (perhaps by making use of the fact that the integer variables are, in this special case, restricted to the special values 0 and 1).

3. An example

The approximation problem of the following filter was solved using the mixed-integer programming method.

Filter requirements:

pass-bands	stop-bands
D_1: $0.5 \leq x \leq 1.5$	S_1: $2 \leq x \leq 3.5$
D_2: $4 \leq x \leq 4.5$	S_2: $5.5 \leq x$

The class F of functions:

$$R(x) = \frac{P_4(x)}{(\sqrt{x} \cdot Q_3(x))}$$

$$(m = 4,\ n = 3 \text{ and } g(x) = 1/\sqrt{x}) .$$

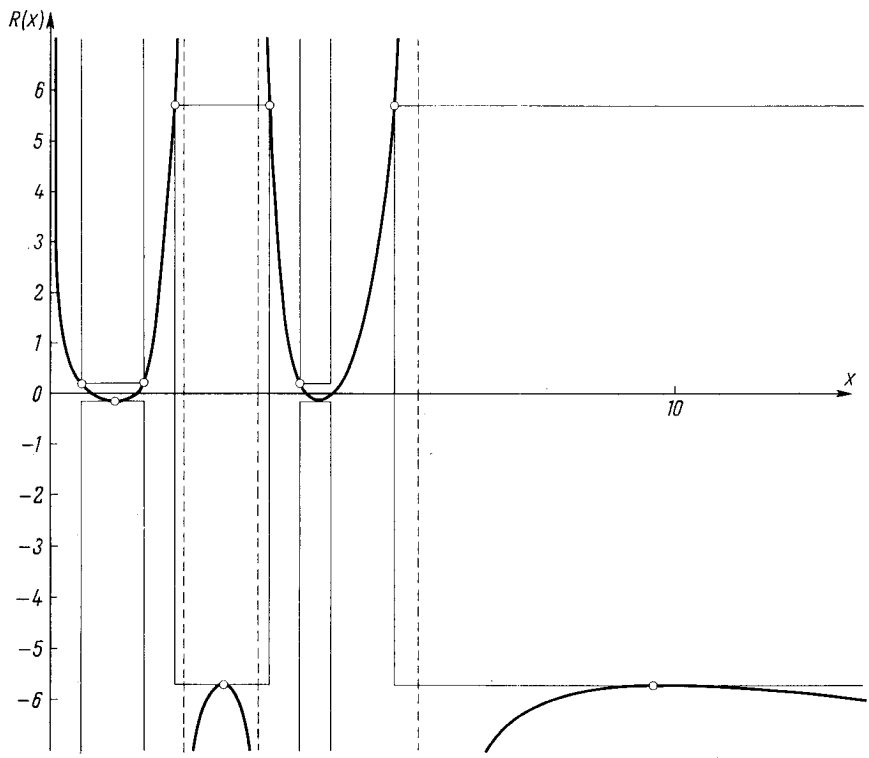

Figure 6
The global optimal function $R^*(x)$
o: Extreme points

The global optimal function:

After two exchange steps, in each of which one mixed-integer program consumed from 4 to 6 times the calculating time for one equivalent linear program, *the global optimal function* $R^*(x)$ shown in Figure 6 was obtained.

The global T-deviation Δ^* (the maximum deviation of the global optimal function):

$$\Delta^* = 0.0308 .$$

From Figure 6 it is seen that the global optimal function has a number of $m + n + 2 = 9$ extreme points; these build a reference with respect to which the function $R^*(x)$ is a leveled reference function. The second pass-band contains only one extreme point and it can be extended to $x = 4.65$ without changing the global optimal function $R^*(x)$.

The zeros and poles of $R^*(x)$ are:

Zeros: 0.658 436, 1.393 76, 4.099 27 and 4.596 08 ,

Poles: 2.146 33, 3.341 99, and 5.888 86 .

Appendix. Proofs of Theorems 1, 2 and 4

Theorem 1: *If* $\bar{R}(x) \in F_f$ *is a reference function with respect to a reference* $(\bar{d}, \bar{s})$, *then:*

i) *The following relation is satisfied by all functions* $R(x) \in F_f$, *which are not proportional to* $\bar{R}(x)$:

$$\Delta(R(x)) > \frac{\min\limits_{x \in \bar{d}} |\bar{R}(x)|}{\max\limits_{x \in \bar{s}} |\bar{R}(x)|} \tag{14}$$

and

ii) *If* $R(x)$ *is not a local optimal function in* F_f, *then:*

$$\Delta(\bar{R}(x)) > \Delta_f > \frac{\min\limits_{x \in \bar{d}} |\bar{R}(x)|}{\max\limits_{x \in \bar{s}} |\bar{R}(x)|} . \tag{15}$$

Proof:

Let a function $R(x) \in F_f$ *be such that:*

$$\Delta(R(x)) \leq \frac{\min\limits_{x \in \bar{d}} |\bar{R}(x)|}{\max\limits_{x \in \bar{s}} |\bar{R}(x)|} \tag{75}$$

then by multiplication of $R(x)$ by a suitable factor a, it can be reached that for $\tilde{R}(x) = a \cdot R(x)$:

$$\left.\begin{array}{ll} \textit{for all } x \in \bar{d}: & |\tilde{R}(x)| \leq |R(x)| \\ \textit{and for all } x \in \bar{s}: & |\tilde{R}(x)| \geq |\bar{R}(x)| . \end{array}\right\} \tag{76}$$

Using the relations (76), *it will be proved that in a number of* $m + n + 1$ *points* $\mathring{x}_i$ (the meaning of m and n is defined in Chapter I) *the following is satisfied:*

$$\tilde{R}(\mathring{x}_i) = \bar{R}(\mathring{x}_i) \quad i = 1, 2, \ldots, m + n + 1 \tag{77}$$

or:
$$g(\mathring{x}_i)\,\frac{\tilde{P}_m(\mathring{x}_i)}{\tilde{Q}_n(\mathring{x}_i)} = g(\mathring{x}_i)\,\frac{\bar{P}_m(\mathring{x}_i)}{\bar{Q}_n(\mathring{x}_i)}$$

or:
$$\frac{\tilde{P}_m(\mathring{x}_i)\cdot\bar{Q}_n(\mathring{x}_i) - \bar{P}_m(\mathring{x}_i)\cdot\tilde{Q}_n(\mathring{x}_i)}{\tilde{Q}_n(\mathring{x}_i)\cdot\bar{Q}_n(\mathring{x}_i)} = 0$$

i.e. the polynomial $A_{m+n}(x) = \tilde{P}_m(x)\cdot\bar{Q}_n(x) - \bar{P}_m(x)\cdot\tilde{Q}_n(x)$ (78)

whose maximum degree is $m + n$ has a number of $m + n + 1$*) zeros; it follows that $A_{m+n}(x)$ must *identically vanish*, which means that *the two functions* $\tilde{R}(x)$ *and* $\bar{R}(x)$ *are identical.*

The functions $R(x)$ and $R(x)$ are thus proportional if relation (75) is satisfied. This proves the first part i) of the theorem. *The second part* ii) *is proved by replacing* $R(x)$ *by a local optimal function for the local class* F_f *in relation* (14).

It remains only to prove (77). For this purpose, relation (75) is used to prove that the two curves:

$$\tilde{r}(x) = \frac{\tilde{P}_m(x)}{\tilde{Q}_n(x)} \quad \text{and} \quad \bar{r}(x) = \frac{\bar{P}_m(x)}{\bar{Q}_n(x)} \tag{79}$$

have a number of $m + n + 1$*) points of intersection $\mathring{x}_i$ ($i = 1, \ldots, m + n + 1$). This can not be proved directly since the curves $\tilde{r}(x)$ and $\bar{r}(x)$ are *not continuous*; the proof consists of the following steps:

---

*) The points $\mathring{x}_i$ are either different, or there can be among them some points which are to be *doubly counted* [this can occur only in the case of equality in (75)] such that their total number remains at least $m + n + 1$. For such a double point $\mathring{x}_j$ it will be proved (footnote page 59) that the two curves $\tilde{r}(x)$ and $\bar{r}(x)$ have a common tangent; from this it is easily seen that in a double point $\mathring{x}_j$:

$$A(\mathring{x}_j) = 0 \quad \text{and} \quad A'(\mathring{x}_j) = 0 .$$

1. *The two curves* $\tilde{r}(x)$ *and* $\bar{r}(x)$ *are transformed into the curves* $\tilde{\phi}(x)$ *and* $\bar{\phi}(x)$ *on the surface of the cylinder whose axis is the x-axis* [a curve $\bar{\phi}(x)$ on the developed surface of the cylinder is shown in the lower part of Figure 7]. This transformation is defined by:

$$\left.\begin{aligned} \sin\phi(x) &= \frac{P_m(x)}{\sqrt{P_m^2(x) + Q_n^2(x)}} \\ \text{and} \qquad & \\ \cos\phi(x) &= \frac{Q_n(x)}{\sqrt{P_m^2(x) + Q_n^2(x)}} . \end{aligned}\right\} \tag{80}$$

This defines the *continuous* curves $\tilde{\phi}(x)$ and $\bar{\phi}(x)$.

2. Since the functions $\tilde{R}(x)$ and $\bar{R}(x)$ *belong to the same local class* F_f [i.e. $\tilde{P}_m(x)$ and $\bar{P}_m(x)$ have the same sign in the stop-bands, and $\tilde{Q}_n(x)$ and $\bar{Q}_n(x)$ have the same sign in the pass-bands], therefore:

$$\left.\begin{aligned} &\textit{in } D_\alpha\text{: if } \sigma_{D_\alpha} = +1, \text{ then both curves } \tilde{\phi}(x) \text{ and } \bar{\phi}(x) \text{ lie in the} \\ &\qquad\qquad \text{neighbourhood of the line } \phi = 0; \\ &\text{and if } \sigma_{D_\alpha} = -1, \text{ then both curves } \tilde{\phi}(x) \text{ and } \bar{\phi}(x) \text{ lie in the} \\ &\qquad\qquad \text{neighbourhood of the line } \phi = \pi; \\ &\textit{in } S_\beta\text{: if } \sigma_{S_\beta} = +1, \text{ then both curves } \tilde{\phi}(x) \text{ and } \bar{\phi}(x) \text{ lie in the} \\ &\qquad\qquad \text{neighbourhood of the line } \phi = \pi/2; \\ &\text{and if } \sigma_{S_\beta} = -1, \text{ then both curves } \tilde{\phi}(x) \text{ and } \bar{\phi}(x) \text{ lie in the} \\ &\qquad\qquad \text{neighbourhood of the line } \phi = 3\pi/2. \end{aligned}\right\} \tag{81}$$

The curve $\bar{r}(x)$ and the corresponding curve $\bar{\phi}(x)$ on the developed surface of the cylinder are shown in Figure 7; the $m + n + 2$ reference points are marked by small circles, and the corresponding points $(x_k, \bar{\phi}(x_k))$ are also marked in the lower part of the figure.

3. From the fact that $\bar{R}(x)$ is *a reference function with respect to the reference* $(\bar{d}, \bar{s})$, it follows that the points with $x = x_k$ (the abscissas of the reference points) on the lines defined by (81) must be situated successively on *opposite* sides of the curve $\bar{\phi}(x)$; as shown by the arrows in Figure 7.

4. From the relations (75) it follows that the points $(x_k, \tilde{\phi}(x_k))$ on the curve $\tilde{\phi}(x)$ are situated *closer* to the lines defined by (81) than the corres-

ponding points $(x_k, \bar{\phi}(x_k))$ on the curve $\bar{\phi}(x)$; i.e. these $m + n + 2$ points of the curve $\tilde{\phi}(x)$ lie relative to the curve $\bar{\phi}(x)$ in the directions of the arrows (Figure 7).

5. From 4. and since $\bar{\phi}(x)$ and $\tilde{\phi}(x)$ are both *continuous curves on the surface of the cylinder*, it follows that between each two successive reference points x_k and x_{k+1} there must be *either* a point of intersection of the two curves: $(\mathring{x}_i, \phi_i)$ *or* two points: $(\mathring{x}_i, \phi_i)$ on the curve $\tilde{\phi}(x)$, and $(\mathring{x}_i, \phi_i + \pi)$ on the curve $\bar{\phi}(x)$.

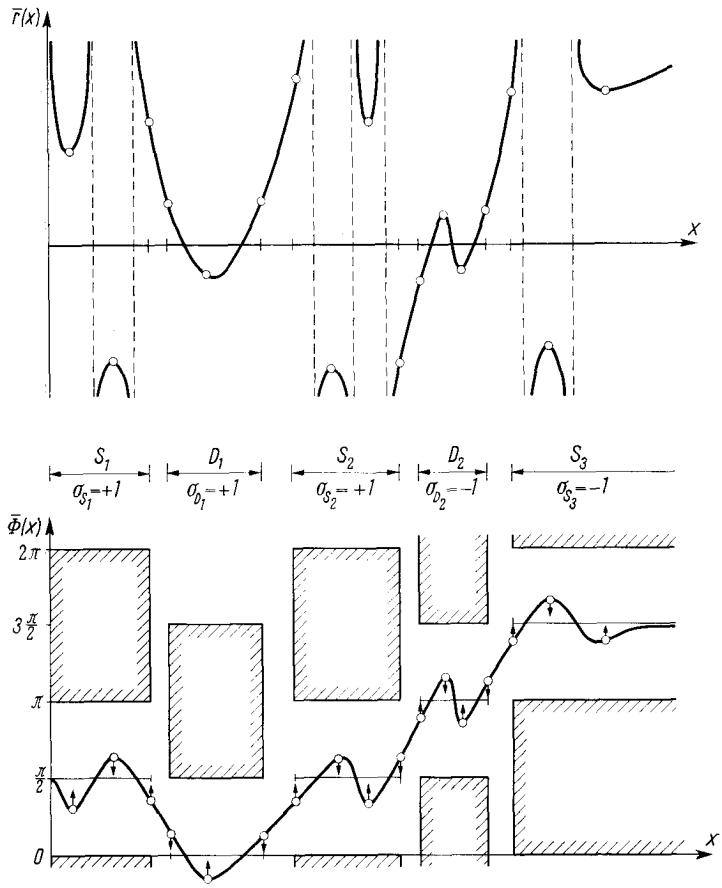

Figure 7
The curves $\bar{r}(x)$ and $\bar{\phi}(x)$
o: reference points; forbidden regions hatched

In both cases the functions $\tilde{r}(x)$ *and* $\bar{r}(x)$ *have the same value at the point* $\mathring{x}_i$. Relation (77) is thus satisfied in the $m + n + 1$*) points $\mathring{x}_i$, this completes the proof of Theorem 1.

Theorem 1′: is the *discrete case* of Theorem 1 and can be proved in exactly the same way.

Theorem 2:

i) *If a function* $R_f(x) \in F_f$ *has a set of* $m + n + 2$ *extreme points, such that they build a reference* $(\bar{d}, \bar{s})$ *with respect to which* $R_f(x)$ *is a leveled reference function, then the function* $R_f(x)$ *is a local optimal function in* F_f*; and each other local optimal function for the same local class is proportional to* $R_f(x)$.

ii) *If a local optimal function* $R_f(x) \in F_f$ *satisfies the assumption:*

$$\left.\begin{array}{l} P_m(x) \text{ and } Q_n(x) \text{ are relatively prime polynomials of} \\ \text{maximum degrees } m \text{ and } n \text{ respectively, such that at least} \\ \text{one of them attains its maximum degree,} \end{array}\right\} \qquad (16)$$

then the function $R_f(x)$ *has at least* $m + n + 2$ *extreme points which contain a reference* $(\bar{d}, \bar{s})$ *with respect to which* $R_f(x)$ *is a leveled reference function.*

Proof:

(i) Since $R_f(x)$ is a leveled reference function with respect to the reference $(\bar{d}, \bar{s})$ and the points of this reference are extreme points of $R_f(x)$, therefore:

$$\Delta(R(x)) = \frac{\max\limits_{x \in \bar{d}} |R_f(x)|}{\min\limits_{x \in \bar{s}} |R_f(x)|} = \frac{\min\limits_{x \in \bar{d}} |R_f(x)|}{\max\limits_{x \in \bar{s}} |R_f(x)|}. \qquad (82)$$

From part ii) of Theorem 1 it follows then, since the inequality (15) can not be satisfied, that $R_f(x)$ *must be a local optimal function in* F_f. From part i) of Theorem 1 it follows directly that *all other local optimal functions in* F_f *must be proportional to* $R_f(x)$.

---

*) In the case of equality in (75), the curve $\tilde{\phi}(x)$ can coincide with $\bar{\phi}(x)$ at some reference points. For each point $\mathring{x}_j$ of this sort (with the exception of the first and the last points of the reference), there is:

either a) at least one point of intersection for $\tilde{\phi}(x)$ with $\bar{\phi}(x)$, different from $\mathring{x}_j$, which lie between the preceding and the succeding reference points.

or b) a common tangent for the two curves $\tilde{\phi}(x)$ and $\bar{\phi}(x)$ at the point $\mathring{x}_j$. In such a double point $\mathring{x}_j$ it is easily proved that the curves $\tilde{r}(x)$ and $\bar{r}(x)$ have a common tangent.

(ii) *Let $R_f(x)$ be a local optimal function which satisfies assumption* (16) *but its extreme points do not build a reference with respect to which $R_f(x)$ is a leveled reference function*; or in other words, there is *at most* a number of $m + n + 1$ points $\bar{x}_k$ (numbered according to their order on the x-axis from left to right) such that:

$$\left.\begin{aligned} a \cdot R_f(\bar{x}_k) &= \begin{cases} (-1)^k \cdot l & \text{for } \bar{x}_k \in D \\ (-1)^k \cdot -\frac{1}{l} & \text{for } \bar{x}_k \in S \end{cases} \\ \text{where } l^2 &= \Delta\big(R_f(x)\big) \\ \text{and } \quad a &= \text{a suitable factor .} \end{aligned}\right\} \tag{83}$$

It can be shown, by constructing a function $R_\lambda(x) \in F_f$ which has a maximum deviation smaller than l^2, that the function $R_f(x)$ is not a local optimal function in F_f.

The function $R_\lambda(x)$ is constructed in the following way:

1) A set of points $\mathring{x}_k$ is chosen such that:

$$\bar{x}_k < \mathring{x}_k < \bar{x}_{k+1} . \tag{84}$$

The number $\mathring{N}$ of these points is less by one than the number of the points $\bar{x}_k$; i.e. $\mathring{N} \leq m + n$.

2) A polynomial $B_{m+n}(x)$ of *maximum degree* $m + n$ is then defined by:

$$B_{m+n}(x) = \pm \prod_{k=1}^{\mathring{N}} (x - \mathring{x}_k) \tag{85}$$

(the sign will be later chosen). Since the polynomials $P_m(x)$ and $Q_n(x)$ satisfy the assumption (16), two polynomials $U_n(x)$ and $V_m(x)$, of *maximum degrees* n and m respectively, can be determined such that:

$$B_{m+n}(x) = U_n(x) \cdot P_m(x) - V_m(x) \cdot Q_n(x) . \tag{86}$$

3) The function $R_\lambda(x)$ is then defined by:

$$R_\lambda(x) = \frac{P_m(x) + \lambda \cdot V_m(x)}{Q_n(x) + \lambda \cdot U_n(x)} \cdot g(x) . \tag{87}$$

To prove that $R_\lambda(x)$ can be constructed such that its maximum deviation is smaller than l^2, consider the difference

$$R_\lambda(x) - R_f(x) = -\, g(x) \, \frac{\lambda \cdot B_{m+n}(x)}{Q_n^2(x) + \lambda \cdot Q_n(x) \cdot U_n(x)} . \tag{88}$$

By the choice of a suitable sign in (85) it can be reached that:

$$\left.\begin{array}{ll} |R_\lambda(x)| < |R_f(x)| & \text{for } x \text{ in the neighbourhood*) of any of the points } \bar{x}_k \in D \\ \text{and } |R_\lambda(x)| > |R_f(x)| & \text{for } x \text{ in the neighbourhood*) of any of the points } \bar{x}_k \in S \end{array}\right\} \quad (89)$$

and by taking λ small enough, the points at which $|R_\lambda(x)|$ assumes the value of its maximum in D or the value of its minimum in S, can be kept in the neighbourhoods*) of the points $\bar{x}_k$. From (89) it follows then that:

$$\Delta(R_\lambda(x)) < \Delta(R_f(x)) = l^2 . \quad (90)$$

Theorem 2′: is the *discrete case* of Theorem 2, and can be proved in ecaxtly the same way.

Theorem 4: For a *band-pass* filter *the global optimal function* $\bar{R}(x)$ has the following properties:

a) *All its zeros are real and lie in the pass-band.*

b) *It has* $m + 1$ *extreme points in the pass-band and* $n + 1$ *extreme points in the stop-band, such that these* $m + n + 2$ *extreme points build a reference* $(\bar{d}, \bar{s})$ *with respect to which* $\bar{R}(x)$ *is a leveled reference function.*

Proof:

a1) *The assumption* (16) *of Theorem 2 is satisfied by the polynomials* $P(x)$ *and* $Q(x)$ *of the global optimal function* $\bar{R}(x)$.

If assumption (16) were not satisfied then the polynomial $P(x)$ must have a degree less than m, so that a new function $R(x) = a(x - x_0) \cdot \bar{R}(x)$ which lies in the class F could be constructed such that [by proper choice of a and x_0**)] it has a smaller maximum deviation than that of the function $\bar{R}(x)$; which is impossible since $\bar{R}(x)$ is *a global optimal function.*

a2) *The function* $\bar{R}(x)$ *has no complex zeros or poles:*

From point 1 above it follows that $\bar{R}(x)$ *is a leveled reference function with respect to a reference* $(\bar{d}, \bar{s})$ *consisting of* $m + n + 2$ *extreme points*

---

*) Or more exactly in the interval:

$\mathring{x}_{k-1} < x < \mathring{x}_k$ (for $k \neq 1, \mathring{N}$)
or $x < \mathring{x}_1$ ($k = 1$)
or $\mathring{x}_{\mathring{N}} < x$ ($k = \mathring{N}$) .

**) Take vor x_0 the middle point of the interval.

(Theorem 2ii). From the definition of a reference function (in Chapter II) it follows that a reference function, *in the case of a band-filter*, must have a total number of *at least* $m + n + 1$ *real zeros and poles* inside the bands. The function $\bar{R}(x)$ can thus have no pairs of complex zeros or poles.

a3) *The function* $\bar{R}(x)$ *has a number of* m *real zeros which lie all in the pass-band:*

From points 1 and 2 above it is seen that $\bar{R}(x)$ must have a number of m *real zeros*. In order to prove that all these zeros lie in the pass-band D, assume that a zero lies in x_0 outside D. Consider then the new function:

$$\tilde{R}(x) = a \cdot \frac{x - \tilde{x}_0}{x - x_0} \cdot \tilde{R}(x) \tag{91}$$

this is also a function that belongs to the class F; and by proper choice of a and the new position of the zero: $\tilde{x}_0$*) it can be reached that

$$\Delta(\tilde{R}(x)) < \Delta(\bar{R}(x)) \tag{92}$$

but this is impossible since $\bar{R}(x)$ is *a global optimal function.*

b) From point 1 of proof a) and Theorem 2, it has been proved that $R(x)$ is a leveled reference function with respect to a reference $(\bar{d}, \bar{s})$ consisting of $m + n + 2$ extreme points of $R(x)$. *It remains now to prove that exactly* $m + 1$ *extreme points lie in the pass-band and* $n + 1$ *extreme points lie in the stop-band:*

b1) *Let a function* $\theta(x)$ *be defined by:*

$$\theta(x) = \begin{cases} R(x) & \text{for } x \in D \\ -1/R(x) & \text{for } x \in S\,. \end{cases} \tag{93}$$

Consider the variation of the sign of $\theta(x)$ as x goes through the bands S_1, D and S_2 from left to right on the x-axis; *the intervals* I_k *(numbered from left to right* $k = 1, 2, \ldots, n_I$*) are defined such that:*

$$\left.\begin{array}{l} \text{i) } I_1 \cup I_2 \cup \cdots \cup I_{n_I} \equiv S_1 \cup D \cup S_2\,. \\ \text{ii) } \theta(x) \text{ has a constant sign in any interval } I_k\,. \\ \text{iii) } \theta(x) \text{ has opposite signs in any two successive intervals } I_k \text{ and } I_{k+1}. \end{array}\right\} \tag{94}$$

---

*) If the ends of the pass-band are the points $x = b$ and $x = c$, then choose $\tilde{x}_0$ such that the four points x_0, $\tilde{x}_0$, b and c build a harmonic range [i.e. the double ratio $(x_0, \tilde{x}_0, b, c) = -1$]. The factor a can be then chosen such that the rational function $\varrho(x) = a(x - \tilde{x}_0)/(x - x_0)$ has an absolute value:

$$|\varrho(x)| \begin{cases} \leq 1 & \text{for all } x \in D \\ > 1 & \text{for all } x \in S\,. \end{cases}$$

The number n_I of these intervals must be such that:

$$m + n + 2 \leq n_I \leq m + n + 3 . \tag{95}$$

The first inequality follows from the existance of the reference $(\bar{d}, \bar{s})$; the second inequality must be satisfied since the number of zeros of the function $\theta(x)$ can not be greater than $m + n$ [because it is equal to the total number of zeros and poles of $R(x)$ which lie in the pass- and stop-bands].

b2) *The pass-band consists of exactly $m + 1$ intervals I_k:*

Since no zeros of $\bar{R}(x)$ can lie between the pass-band and either of the stop-bands, and using a similar construction to that of foot-note page 62 it is easily shown that *no poles of $\bar{R}(x)$ can lie between the stop-bands, therefore $\bar{R}(x)$ must have the same sign* [*i.e. $\theta(x)$ must change its sign*] as x goes from S_1 to D and from D to S_2. From this and since $\theta(x)$ has m zeros in D, it follows that *the pass-band consists of exactly $m + 1$ intervals I_k.*

b3) *From each interval I_k choose a point x_k such that:*

$$\theta(x_k) = \max_{x \in I_k} \left(\theta(x) \right) \quad k = 1, 2, \ldots, n_I . \tag{96}$$

This defines a number of n_I points x_k, at which $\theta(x)$ has alternating signs, and from which a number of $m + 1$ points lie in the pass-band. From (95) it is seen that *the number n_I can be either equal to $m + n + 2$ or $m + n + 3$.* In the first case the reference $(\bar{d}, \bar{s})$ consists of exactly the $m + n + 2$ points x_k; in the second case either the first or the last one of the points x_k is excluded to obtain the $m + n + 2$ points of the reference [since the function $\theta(x)$ must have alternating signs and $|\theta(x)|$ must assume its maximum at the points of $(\bar{d}, \bar{s})$]. *The reference $(\bar{d}, \bar{s})$ consists in both cases of exactly $m + 1$ points in the pass-band and $n + 1$ points in the stop-bands.*

REFERENCES

[1] E. STIEFEL, *Le problème d'approximation dans la théorie des filtres électriques*, Colloque sur l'analyse numérique à Mons (1961).
[2] W. CAUER, *Theorie der linearen Wechselstromschaltungen*, Akademie-Verlag, (Berlin 1954).
[3] E. STIEFEL, *Numerical methods of Tchebycheff Approximation*, Proceedings of a symposium conducted by the Mathematics Research Center (University of Wisconsin, April 1958).
[4] E. STIEFEL, *Einführung in die numerische Mathematik* (Teubner Verlagsgesellschaft 1961).
[5] N. J. ACHIESER, *Vorlesungen über Approximationstheorie* (Akademie-Verlag, Berlin 1953).
[6] R. GOMORY, *An Algorithm for mixed-integer problems* (The RAND corporation, July 1960).
[7] G. B. DANTZIG, *On the significance of solving linear programming problems with some integer variables*, Econometrica, Vol. *28*, 1 (January 1960).

ZUSAMMENFASSUNG

Das Approximationsproblem der elektrischen Filter wird mit Hilfe eines modifizierten Austauschverfahrens auf eine Folge von diskreten Problemen zurückgeführt. Das diskrete Problem im Falle eines Bandfilters wird durch ein dazu äquivalentes Eigenwertproblem gelöst. Im allgemeinen Fall von mehreren Sperr- und Durchlassbereichen werden die Methoden der linearen Programmierung verwendet, um die lokal optimalen Funktionen zu bestimmen. Unter diesen wird dann die Lösungsfunktion herausgegriffen. Die direkte Bestimmung der Lösungsfunktion kann auf ein ganzzahliges lineares Programm zurückgeführt werden.

CURRICULUM VITAE

Name: Roshdi Abdel-Rahman Amer. Citizen of the U.A.R. Born in Cairo on the 23 March 1934.

Education:

1941 to 1950: Primary and secondary schools in Cairo; passed the Maturity examination and obtained a scholarship from Cairo University on July 1950.

1950 to June 1955: Study at the Faculty of Engineering in Cairo University; received the B.Sc. degree in Electrical Engineering on June 1955.

July 1955 to January 1956: Served as an electrical engineer in the egyptian government military factories.

January 1956 to September 1958: Assistant at the Electrical Engineering Department of the Faculty of Engineering, Cairo University.

October 1958 to November 1961: Study at the Department of Mathematics and Physics of the Swiss Federal Institute of Technology; received the degree of Dipl. Math. on November 1961.

November 1961 to February 1964: Research work at the Institute of Applied Mathematics at the E.T.H. Zurich, where the present thesis has been developed.