10 JUL 1979

TRITA-NA 7407
G1

THE ROYAL INSTITUTE OF TECHNOLOGY
STOCKHOLM
SWEDEN

DEPARTMENT OF
NUMERICAL ANALYSIS
AND COMPUTING SCIENCE

CERN LIBRARIES, GENEVA

CM-P00069452

TRITA-NA-7907

Inst för Numerisk Analys och Datalogi
KTH
100 44 STOCKHOLM 70

Dept of Numerical Analysis
and Computing Science

The Royal Institute of Technology
S-100 44 STOCKHOLM 70, Sweden

ON NUMERICAL ANALYSIS IN
SEMI-INFINITE PROGRAMMING

by

Sven-Åke Gustafson

TRITA-NA-7907

ON NUMERICAL ANALYSIS IN SEMI-INFINITE PROGRAMMING

Sven-Åke Gustafson
Department of Numerical Analysis
Royal Institute of Technology
S-100 44 STOCKHOLM 70, Sweden

Abstract. In this paper which is a companion to [9] we shall discuss the theoretical questions which arise by the computational treatment of semi-infinite programs. Fairly strong regularity assumptions will be needed to insure satisfactory results of the computational schemes described here and which have proved effective in actual calculations.

1. Notations and preliminaries

We immediately introduce the concept of a dual pair of semi-infinite programs (SIP's). We will adhere to the notations in [3] and [9]. We will work with the entities:

S is a fixed but arbitrary index-set

$a_1, a_2, \dots, a_n$ and b are $n+1$ functions defined on S

$c \in \mathbb{R}^n$ a fix vector.

With these data we now define the two problems:

Program (P). Minimize the linear form

$$(1a) \quad c^T y$$

over all vectors $y \in \mathbb{R}^n$ subject to the constraints

$$(1b) \quad a(s)^T y \geq b(s), \quad s \in S$$

//

and

Program (D). Determine the integer q , the subset $\{s_1, s_2, \dots, s_q\} \subset S$ and the reals $x_1, x_2, \dots, x_q$ such that the expression

$$(2a) \quad \sum_{i=1}^q x_i b(s_i)$$

is maximized under the constraints

$$(2b) \quad \sum_{i=1}^q x_i a(s_i) = c$$

$$(2c) \quad x_i \geq 0, \quad i = 1, 2, \dots, q$$

//

Here $a(s) \in \mathbb{R}^n$ is the vector whose components are $a_r(s)$, $r = 1, 2, \dots, n$. In the applications it is very common that the index set S has the form

$$(3) \quad S = \bigcup_{j=0}^{\ell} S_j$$

where S_0 has finitely many elements only. We reformulate Programs (P) and (D) for this case and get

Program (P). Minimize the linear form

$$(4a) \quad c^T y$$

over all vectors $y \in R^n$ subject to the constraints

$$(4b) \quad a(s)^T y \geq b(s), \quad s \in S_j, \quad j = 0, 1, \dots, \ell \quad //$$

and

Program (D). Determine the integers q_j , $j = 0, 1, \dots, \ell$ the subsets $\{s_1^j, s_2^j, \dots, s_{q_j}^j\} \subset S_j$, $j = 0, 1, \dots, \ell$ and the reals $x_1^j, x_2^j, \dots, x_{q_j}^j$ such, that the expression

$$(5a) \quad \sum_{j=0}^{\ell} \sum_{i=1}^{q_j} x_i^j b(s_i^j)$$

is rendered a maximum under the constraints

$$(5b) \quad \sum_{j=0}^{\ell} \sum_{i=1}^{q_j} x_i^j a(s_i^j) = c$$

$$(5c) \quad x_i^j \geq 0, \quad i = 1, 2, \dots, q_j, \quad j = 0, 1, \dots, \ell. \quad //$$

Obviously the problem (5) is a special case of (2). Note that we use in this paper the convention that when referring to an entire group of formulas whose labels consist of a certain number followed by letters we use the number only. Thus by (5) we mean (5a), (5b) and (5c).

In a computer we can only store finitely many numbers with a limited precision. Further we can only carry out finitely many arithmetic operations. Thus it must be possible to determine whether an arbitrary s belongs to S by performing only finitely many such operations. We shall normally assume that $S \subset R^k$, $k < \infty$ although some of the results will be formulated for a slightly more general situation. The functions a_r and b will be defined through computer programs which must be of finite length and call for only finitely many arithmetic operations.

An optimal solution of Program (P) will be represented by an ordered n -tuple. For Program (D) the situation is not that easy, since q may be arbitrarily large. However, the following result holds.

Lemma 1. Let Program (D) assume its optimal value. Then it has an optimal solution specified by q , $\{s_1, s_2, \dots, s_q\}$, $x_1, x_2, \dots, x_q$ such that

$$(6a) \quad i) \quad q \leq n$$

$$(6b) \quad ii) \quad x_i > 0, \quad i = 1, 2, \dots, q$$

(6c) iii) $a(s_i), i=1,2,\dots,q$ are linearly independent.

The elementary *proof* can be carried out straightforwardly using properties of convexity and linear dependency. See e.g. [3] p 120-121.

We will only discuss the representation of an optimal solution of Program (D) satisfying (6) when $S \subset R^k$ and with S given by (3). The solution

$$q, \{s_1, s_2, \dots, s_q\} \quad x_1, x_2, \dots, x_q$$

is specified by among other things q vectors $s_1, s_2, \dots, s_q$ which can be ordered in $q!$ different ways. In order to remove this ambiguity in the computer representation we introduce a canonical ordering in two phases as follows. First we order the vectors $s_1, s_2, \dots, s_q$ such that first come the elements out of $\{s_1, s_2, \dots, s_q\}$ belonging to S_0 (if any) then those in S_1 , etc. Next we order the elements in each S_j in lexicographic order. (If $u \in R^k$ and $v \in R^k$, then u comes before v in the lexicographic order, if *either* $u_1 < v_1$ or there is an $i, 2 \leq i \leq n$ such that $u_j = v_j, j=1,2,\dots,i-1$ and $u_i < v_i$). If we order $s_1, s_2, \dots, s_q$ as described above then an optimal solution of Program (D) is uniquely specified by the integers $q_0, q_1, \dots, q_q$, the vectors $s_1, s_2, \dots, s_q$ and the reals $x_1, x_2, \dots, x_q$, i.e. in total $1 + l + q(k+1)$ numbers.

Definition 1. Two dual pairs (P), (D) and $(\tilde{P}), (\tilde{D})$ specified by respectively S, a, b, c and $S, \tilde{a}, \tilde{b}, \tilde{c}$ are said to be *computationally equivalent*, if the computer representations coincide for the following pairs of numbers:

$$a_r(s) \text{ and } \tilde{a}_r(s), r=1,2,\dots,n, s \in S$$

$$b(s) \text{ and } \tilde{b}(s), s \in S$$

$$c_r \text{ and } \tilde{c}_r, r=1,2,\dots,n$$

//

Example 1. Put

$$b(s) = e^s \quad \tilde{b}(s) = \sum_{k=0}^{20} \frac{s^k}{k!} \quad -1 \leq s \leq 1$$

The representations of b and $\tilde{b}$ coincide for a computer working with the relative accuracy 10^{-8} .

//

We now want to prove that if S is compact and $a_r, r=1,2,\dots,n$ and b continuous on S , then Programs (P) and (D) are computationally equivalent to linear programs.

We first need

Definition 2. Let $T = \{t_1, t_2, \dots, t_N\}$ be a finite subset of a set $S \subset R^k$. Then T will be called a *grid*. Denote by $k(T)$ its convex hull. Next determine N continuous functions $\rho_1, \rho_2, \dots, \rho_N$ such that

i) each $s \in k(T)$ has a representation

$$s = \sum_{j=1}^N \rho_j(s) t_j$$

- ii) $\sum_{j=1}^N \rho_j(s) = 1$
- iii) $\rho_j(s) \geq 0, \quad j = 1, 2, \dots, N$
- iv) for each $s \in k(T)$ at most $k+1$ $\rho_j(s)$ are strictly positive.

(Note that ρ_j are not uniquely determined by i) through iv)).

Let now a function f be defined on T . Define Lf through

$$(Lf)(s) = \sum_{j=1}^N \rho_j(s) f(t_j)$$

Then L will be called the *positive linear interpolator* induced by T and

$\rho_1, \rho_2, \dots, \rho_N$.

//

We next state

Lemma 2. Let $S, T, k(T)$ and L be defined as in Definition 2. Then $y \in R^n$ satisfies the inequalities

$$a(t_j)^T y \geq b(t_j), \quad j = 1, 2, \dots, N$$

if and only if y satisfies the inequalities

$$La^T(s)y \geq Lb(s), \quad s \in k(T)$$

The proof is immediate. See e.g. [3] p 148.

//

We now arrive at the result announced earlier.

Theorem 1. Let $S \subset R^k$ be a compact set, $a_r, r = 1, 2, \dots, n$ and b continuous functions on S . Then there is a finite subset $T \subset S, T = \{t_1, t_2, \dots, t_N\}$ such that Program (P) is computationally equivalent to the task: Minimize the linear form

$$(7a) \quad c^T y,$$

over all vectors $y \in R^n$ subject to the constraints

$$(7b) \quad La(s)^T y \geq b(s), \quad s \in S$$

The problem (7) can be solved by means of linear programming. L is defined as in Definition 2.

Proof. Let $\epsilon > 0$ be given. Since S is compact and $a_r, r = 1, 2, \dots, n$ and b are continuous on S it is possible to find a subset $T = \{t_1, t_2, \dots, t_N\}$ and a corresponding positive linear interpolator L such that

- i) $T \subset S \subset k(T)$
- ii) $|La_r(s) - a_r(s)| < \epsilon, \quad s \in S, \quad r = 1, 2, \dots, n$
- iii) $|Lb(s) - b(s)| < \epsilon, \quad s \in S$

L may be constructed e.g. through interpolating linearly in each of the k dimen-

sions. By Lemma 2 (7b) is equivalent to

$$a(t_j)^T y \geq b(t_j), \quad j = 1, 2, \dots, N$$

and hence Problem (7) is equivalent to a linear program with n variables and N constraints. By choosing ϵ small enough we achieve the desired result. //

As a consequence of Theorem 1 a SIP could be replaced by a linear program although the latter generally will be very large. This is the rationale behind the methods based on discretization where a SIP is approximated by a linear program. This can always be solved in a finite time. The iterative methods to be treated in Section 4 are often more rapid provided an initial approximation of the optimal solution y together with some other information are given. These methods have been discussed extensively in [12], [13]. However, for these methods convergence is not guaranteed in many practical situations. An initial approximation of the optimal solution may be constructed by means of linear programming and then refined by means of iterative schemes. In case of convergence failure the initial approximation is refined before iteration is recommenced. Such a combined approach is described in [3], [5], [10] and implemented in the computer codes in [2].

2. Wellposedness in SIP

If the optimal value $v(P)$ of Program (P) varies very strongly by small perturbations in input data a, b, c then $v(P)$ is not well determined by the computer representation of Program (P). We illustrate this by

Example 2. Minimize the linear form

$$y_1 + \frac{1}{2}y_2 + \left(\frac{1}{2} + \frac{\epsilon}{3}\right)y_3$$

over all $y \in \mathbb{R}^3$ subject to the constraints

$$y_1 + y_2 s + y_3(s + \epsilon s^2) \geq e^s, \quad s \in [0, 1]$$

We find: $v(p) = \frac{1}{2}(1 + e) \quad \epsilon = 0, \quad v(P) = \frac{1}{4}(3e^{1/3} + e), \quad \epsilon \neq 0$ //

Example 3. Minimize the linear form

$$y_1 + (1 + \epsilon)y_2$$

over all $y \in \mathbb{R}^2$ subject to the constraints

$$y_1 + y_2 s \geq -\frac{1}{1+s}, \quad s \in [0, 1].$$

We get $v(P) = -\infty, \quad \epsilon > 0 \quad v(P) = -\frac{1}{2}, \quad \epsilon = 0 \quad v(P) = -\frac{1}{2+\epsilon}, \quad \epsilon < 0$ //

In the two examples $v(P)$ is poorly determined, if $|\epsilon|$ is a small number in comparison to the computational errors.

We will now give conditions which insure that the problem to determine the value

$v(P)$ is well-posed. The corresponding optimal solution can, under the same conditions be shown to vary continuously with input data, provided it is unique. (In case of nonuniqueness we could take the optimal solution with the smallest Euclidean norm. Unfortunately this imposed uniqueness does not imply that the so selected solution varies continuously with input data, as easily seen from simple examples).

Lemma 3. Let (1b) be consistent and assume that c has a representation

$$\sum_{i=1}^n x_i a(s_i) = c \quad x_i > 0, \quad s_i \in S, \quad i = 1, 2, \dots, n$$

and where $a(s_1), a(s_2), \dots, a(s_n)$ are linearly independent. Then there is an $F > 0$ such that all optimal solutions y of Program (P) satisfy

$$|y_r| \leq F, \quad r = 1, 2, \dots, n$$

Proof. By (7) on p 106 in [3] Program (P) has an optimal solution y . It must also met (1b). Hence there are numbers d_i such that

$$(8) \quad y^T a(s_i) = b(s_i) + d_i \quad \text{with } d_i \geq 0, \quad i = 1, 2, \dots, n$$

We get also

$$y^T c = \sum_{i=1}^n x_i b(s_i) + \sum_{i=1}^n x_i d_i$$

Since $x_i > 0$ by assumption there is a $B > 0$ such that

$$(9) \quad 0 \leq d_i \leq B, \quad i = 1, 2, \dots, n$$

We may look upon (8) as a linear system with $y \in R^n$ as the unknown. Since $a(s_1), a(s_2), \dots, a(s_n)$ are assumed to be linearly independent this system has a nonsingular matrix. Due to (9) its right hand side is bounded by $B + \max_i |b(s_i)|$ and the desired result follows. //

We next derive an expression for the change in the value $v(P)$ which is caused by perturbations in input data a , b and c .

Lemma 4. Let S be a compact set, a_r and $\tilde{a}_r$, $r = 1, 2, \dots, n$ and $b, \tilde{b}$ be continuous functions on S . c and $\tilde{c}$ are given vectors. With these data we form the two tasks:

Program (P): Minimize $c^T y$

subject to $a(s)^T y \geq b(s), \quad s \in S$

Program ($\tilde{P}$): Minimize $\tilde{c}^T y$

subject to $\tilde{a}(s)^T y \geq \tilde{b}(s), \quad s \in S$

We make the assumptions:

- i) (P) and ($\tilde{P}$) have optimal solutions y and $\tilde{y}$ respectively
- ii) There is a vector y^S and a number $\omega > 0$ such that

$$a(s)^T y^S - b(s) \geq \omega, \quad s \in S \quad \tilde{a}(s)^T y^S - \tilde{b}(s) \geq \omega, \quad s \in S$$

Introduce the notations:

F is the largest of the $2n$ numbers $|y_r|$ and $|\tilde{y}_r|$, $r = 1, 2, \dots, n$,

$$\delta_r = \max_{s \in S} |a_r(s) - \tilde{a}_r(s)|, \quad r = 1, 2, \dots, n, \quad \delta_{n+1} = \max_{s \in S} |b(s) - \tilde{b}(s)|$$

$$\Delta = \delta_{n+1} + F \sum_{r=1}^n \delta_r$$

$$C_r = \max(|c_r|, |\tilde{c}_r|), \quad r = 1, 2, \dots, n$$

$$Y_r = \max(|y_r^S - y_r|, |y_r^S - \tilde{y}_r|), \quad r = 1, 2, \dots, n$$

Then we may state

$$|v(P) - v(\tilde{P})| \leq F \sum_{r=1}^n |\tilde{c}_r - c_r| + \frac{\Delta}{\omega + \Delta} \sum_{r=1}^n C_r Y_r$$

Proof. We want to construct a vector $\bar{y}$ of the form $\bar{y} = (1-\lambda)y + \lambda y^S = y + \lambda(y^S - y)$ which is feasible for $\tilde{P}$. We get first

$$\tilde{a}(s)^T \bar{y} - \tilde{b}(s) = a(s)^T y - b(s) + \{\tilde{a}(s)^T - a(s)^T\} y + b(s) - \tilde{b}(s)$$

Since y is feasible for (P) , $a(s)^T y - b(s) \geq 0$.

Thus

$$\tilde{a}(s)^T \bar{y} - \tilde{b}(s) \geq -(\delta_{n+1} + F \sum_{r=1}^n \delta_r) = -\Delta$$

Next

$$\tilde{a}(s)^T \bar{y} - \tilde{b}(s) = (1-\lambda)[\tilde{a}(s)^T \bar{y} - \tilde{b}(s)] + \lambda[\tilde{a}(s)^T y^S - \tilde{b}(s)] \geq -(1-\lambda)\Delta + \lambda\omega = 0$$

for $\lambda = \Delta/(\omega + \Delta)^{-1}$

With this choice of λ , $\bar{y}$ is hence feasible for $\tilde{P}$ and we get therefore

$$v(\tilde{P}) \leq \tilde{c}^T \bar{y} = \tilde{c}^T \{y + \lambda(y^S - y)\} = c^T y + (\tilde{c} - c)^T y + \lambda \tilde{c}^T (y^S - y)$$

Since $v(P) = c^T y$ we obtain

$$v(\tilde{P}) - v(P) \leq (\tilde{c} - c)^T y + \lambda \tilde{c}^T (y^S - y)$$

Using the notations introduced above we arrive at

$$(10) \quad v(\tilde{P}) - v(P) \leq F \sum_{r=1}^n |\tilde{c}_r - c_r| + \frac{\Delta}{\omega + \Delta} \sum_{r=1}^n C_r Y_r$$

If we interchange the roles of P and $\tilde{P}$ in the calculation above we establish that $v(P) - v(\tilde{P})$ is also bounded by the right hand side of (10) and hence the assertion of the lemma is valid. //

Definition 3. Let S be a set, a_r , $r = 1, 2, \dots, n$ and b functions defined on S . If there is a vector y^S and a number $\omega > 0$ such that

$$(11) \quad a(s)^T y^S - b(s) \geq \omega, \quad s \in S,$$

then $a_1, a_2, \dots, a_n$ and b are said to meet Slater's condition over S . //

Definition 4. Let S be a compact set and let a_r , $r=1,2,\dots,n$ and b belong to $C(S)$, the linear space of continuous functions defined on S and equipped with the maximum norm. Let $c \in \mathbb{R}^n$ and define the linear space U of ordered triples $u = (a,b,c)$. Define $\|u\|$ through $\|u\| = \max(\|a_1\|, \|a_2\|, \dots, \|a_n\|, \|b\|, |c_1|, |c_2|, \dots, |c_n|)$. Then U is called the *dataspace* of Program (P). //

We now establish the main result of this Section:

Theorem 2. Let $u = (a,b,c)$ be as in Definition 4. Make the following assumptions

- i) S is a compact subset of $\mathbb{R}^k$
- ii) $a_1, a_2, \dots, a_n$ and b meet Slater's condition over S
- iii) c has the representation required in Lemma 3.

On the dataspace U we introduce the norm given in Definition 4. Then we assert: There is a neighbourhood $N(u)$ of u with the properties

- a) The optimal value $v(\tilde{P})$ of each Program $\tilde{P}$ formed with respect to $\tilde{u} = (\tilde{a}, \tilde{b}, \tilde{c})$ is a continuous function on $N(u)$.
- b) Provided each Program $\tilde{P}$ has a unique optimal solution $\tilde{y}$ it varies continuously as a function of input data $\tilde{a}, \tilde{b}, \tilde{c}$.

Proof. It is clear that there is a neighbourhood of u such that the assumptions ii) and iii) hold. Hence by the results on p 115 in [3] the corresponding Programs have optimal solutions. By a slight modification of the argument in the proof of Lemma 3 we established that these optimal solutions must be confined to a bounded subset of $\mathbb{R}^n$. The statement a) is now an easy consequence of Lemma 4. Assume now that each Program $\tilde{P}$ has a unique optimal solution $\tilde{y}$. We want to establish that $\tilde{y}$ is a continuous function of $\tilde{u}$ at $\tilde{u} = u$. Assume the contrary. Let y^0 be the optimal solution corresponding to u . Thus there is a sequence $u^1, u^2, \dots$ such that $u^k \rightarrow u$ but the corresponding optimal solutions y^k do not converge to y^0 . However, $\{y^k\}_0^\infty$ is confined to a bounded subset of $\mathbb{R}^n$ and hence has an accumulation point z . We thus select a subsequence $y^{k_k} \rightarrow z$. Since $u^{k_k} \rightarrow u$ we find that z is the unique optimal solution corresponding to u , thus $z = y^0$. We find that all convergent subsequences of $\{y^k\}$ have y^0 as limitpoint and this fact establishes the contradiction sought. //

3. Discrete approximations of semi-infinite programs

Let the assumptions of Theorem 2 prevail! Form a sequence of grids $\{T^\ell\}_1^\infty$ with

$$T^1 \subset T^2 \subset \dots \subset S$$

and such

$$\lim_{\ell \rightarrow \infty} \max_{s \in S} \min_{t \in T^\ell} \|s - t\| \rightarrow 0$$

where $\| \cdot \|$ is a norm on R^k . Then there is a ℓ_0 such that for $\ell \geq \ell_0$ the linear program obtained when T^ℓ replaces S has an optimal solution. Assume now that the optimal solution y^ℓ is unique. By Theorem 2 the sequence $\{y^\ell\}_{\ell_0}^\infty$ must converge to an optimal solution of Program (P).

A similar result cannot be established for Program (D), however. We illustrate with

Example 4. Determine the integer q , the subset $\{s_1, s_2, \dots, s_q\} \subset [0, 1]$ and the reals $x_1, x_2, \dots, x_q$ such that the expression

$$-\sum_{i=1}^q x_i / (1+s_i)$$

is maximized under the constraints

$$\sum_{i=1}^q x_i s_i^{r-1} = 1/r, \quad r=1, 2$$

$$\text{and } x_i \geq 0, \quad i=1, 2, \dots, q$$

The optimal solution is $q=1$ $x_1=1$ $s_1=1/2$ and the corresponding value is $-2/3$. Let us now discretize this problem and replace $[0, 1]$ with the grids $T^\ell = \{s_1^\ell, s_2^\ell, \dots, s_\ell^\ell\}$ where

$$s_i^\ell = \frac{i-1}{\ell-1}, \quad i=1, 2, \dots, \ell, \quad \ell=2, 3, \dots$$

In the cases ℓ odd and ℓ even we find the unique optimal solutions:

$$\ell = 2m+1: \quad q=1, \quad s_1^\ell = \frac{1}{2}, \quad x_1^\ell = 1 \quad \text{value} = -2/3$$

$$\ell = 2m: \quad q=2, \quad s_1^\ell = \frac{m-1}{2m-1}, \quad s_2^\ell = \frac{m}{2m-1}, \quad x_1^\ell = x_2^\ell = \frac{1}{2}, \quad \text{value} = -\frac{3}{2} \frac{(2m-1)^2}{(3m-2)(3m-1)}$$

//

Large values of ℓ correspond to small perturbations in input data. We may even by selecting ℓ odd get an arbitrarily close approximation where the optimal solution of the discretized problem has $q=2$ but the originally SIP has $q=1$. For further illustrations of this phenomena see e.g. [2], [3], [8]. Thus the problem to calculate q is not well posed. However, we can derive the following results.

Lemma 5. Let Program D have an optimal solution $q, \{s_1, s_2, \dots, s_q\}, x_1, x_2, \dots, x_q$ corresponding to the value $v(D)$. Let further $a_1, a_2, \dots, a_n$ and b meet Slater's condition over S . Then we can state, using the notations of Definition 3.

$$(12) \quad \sum_{i=1}^q x_i \leq \omega^{-1} [c^T y^S - v(D)]$$

Proof. Since $q, \{s_1, s_2, \dots, s_q\}$ and $x_1, x_2, \dots, x_q$ is also feasible for Program (D) we have

$$c^T y^S - v(D) = \sum_{i=1}^q x_i \{a(s_i)^T y^S - b(s_i)\} \geq \omega \sum_{i=1}^q x_i \quad \text{giving (12).}$$

//

We next introduce

Definition 5. The functions $a_1, a_2, \dots, a_n$ which are defined on a set S are said to meet *Krein's condition* over S , if there is a vector $y^K \in R^n$ and a number $\lambda > 0$ such that

$$(13) \quad a(s)^T y^K \geq \lambda, \quad s \in S \quad //$$

We note that if Krein's condition is met by the continuous functions $a_1, a_2, \dots, a_n$ over a compact set S , then the $n+1$ functions $a_1, a_2, \dots, a_n$ and b meet Slater's condition over S for all continuous b . In particular, Krein's condition is met if $a_1(s) = 1$, a case often occurring in practice.

Lemma 6. Let $q, \{s_1, s_2, \dots, s_q\} \subset S$ and $x_1, x_2, \dots, x_q$ be feasible for Program (D). If $a_1, a_2, \dots, a_n$ meet Krein's condition over S then, using the notations of Definition 5

$$(14) \quad \sum_{i=1}^q x_i \leq \lambda^{-1} c^T y^K$$

Proof. Since $q, \{s_1, s_2, \dots, s_q\}$ and $x_1, x_2, \dots, x_q$ is feasible for Program (D),

$$c^T y^K = \sum_{i=1}^q x_i a(s_i)^T y^K \geq \lambda \sum_{i=1}^q x_i$$

and hence (14) follows. //

One could say that Lemmas 5 and 6 are the counterpart of Lemma 3. If all of these Lemmas apply, they should guarantee that numbers of moderate size only occur during the numerical treatment of an SIP. If very large numbers appear nevertheless this fact signals numerical difficulties. One source of these troubles could be that $a_1, a_2, \dots, a_n, b$ are almost linearly dependent, i.e. one or more of the a_i 's can be approximated accurately with a linear combination of the remaining ones: Then the corresponding component in (2b) should be dropped and n decreased by 1. In particular, if b is approximated accurately by a linear combination of $a_1, a_2, \dots, a_n$, then $v(D)$ is correspondingly well determined by the conditions (2b), (2c), and the task to evaluate $v(D)$ should not be treated as a SIP. In the general situation one could contemplate replacing $a_1, a_2, \dots, a_n, b$ by linear combinations of the same functions selected to be "more linearly independent". This could be achieved using a scheme similar to the modified Gram-Schmidt orthogonalization procedure.

We next discuss the solution of a discretized SIP by means of the simplex method. It is essential that stable updating of the inverse is performed e.g. according to [3] Kap IV.

Let $\{s_1, s_2, \dots, s_n\} \subset S$ specify a basic solution obtained by the simplex method. Then $a(s_1), a(s_2), \dots, a(s_n)$ are linearly independent and

$$(15a) \quad \sum_{i=1}^n x_i a(s_i) = c \quad \text{with} \quad x_i \geq 0, \quad i = 1, 2, \dots, n$$

To test for optimality we determine $y \in R^n$ from the system

$$(15b) \quad y^T a(s_i) = b(s_i) \quad i = 1, 2, \dots, n$$

Next we determine s^* as the point minimizing $y^T a(s) - b(s)$ on the grid used for discretization. If the corresponding value is nonnegative, y is an optimal solution. Otherwise another simplex step must be performed and s^* must replace one of the s_i 's in (15a). To determine *which* we need to solve the system

$$(15c) \quad \sum_{i=1}^n \rho_i a(s_i) = a(s^*)$$

All of the systems (15) have a regular matrix and hence unique solutions. But their condition can be poor, if the grid is fine and some of the s_i lie closely together since then the vectors $a(s_i)$ are almost linearly dependent. Since the analysis of the general situation is complicated but does not give more insight we discuss only the special case $n = 2\ell$ and s_{2i}, s_{2i-1} lying closely together for $i = 1, 2, \dots, \ell$. We assume also that S is an interval. $a_r, r = 1, 2, \dots, n$ and b are also assumed to be continuously differentiable. (15a) takes the form

$$\sum_{i=1}^{\ell} \{x_{2i-1} a(s_{2i-1}) + x_{2i} a(s_{2i})\} = c$$

or

$$(16a) \quad \sum_{i=1}^{\ell} \left\{ u_i a(s_{2i-1}) + v_i \frac{a(s_{2i}) - a(s_{2i-1})}{s_{2i} - s_{2i-1}} \right\} = c$$

where $u_i = x_{2i-1} + x_{2i}$ $v_i = x_{2i}(s_{2i} - s_{2i-1})$, $i = 1, 2, \dots, \ell$. We see from (16a) that u_i and v_i are the solution of a system whose condition number remains bounded when $s_{2i} - s_{2i-1}$ becomes small. This means that the sums $x_{2i-1} + x_{2i}$ remain well determined provided that the quotients $(a(s_{2i}) - a(s_{2i-1})) / (s_{2i} - s_{2i-1})$ can be evaluated with full accuracy, e.g. if the derivative of a_r is numerically available. But the individual values x_{2i}, x_{2i-1} are obtained from the solution of a system whose inverse is roughly proportional to $\max_i (s_{2i} - s_{2i-1})^{-1}$. If only the functional values $a_r(s)$ are available, then the condition of the problem to determine u_i and v_i is also proportional to $(s_{2i} - s_{2i-1})^{-1}$ as apparent from (16a). These results are independent of ℓ , the number of groups s_{2i-1}, s_{2i} in (15). For (15c) analogous statements can be made. We turn now to (15b) which may be written

$$(16b) \quad \sum_{r=1}^n y_r a_r(s_{2i}) = b(s_{2i}), \quad \sum_{r=1}^n y_r \frac{a_r(s_{2i}) - a_r(s_{2i-1})}{s_{2i} - s_{2i-1}} = \frac{b(s_{2i}) - b(s_{2i-1})}{s_{2i} - s_{2i-1}}$$

Thus the condition of the problem to determine y is roughly independent of $s_1, s_2, \dots, s_{2\ell}$ provided that the quotients of the form $(a(s_{2i}) - a(s_{2i-1})) / (s_{2i} - s_{2i-1})$ can be evaluated with full accuracy e.g. by means of derivatives of a_r . Otherwise y is solution of a system of equations whose condition number is approximately proportional to $\max_i (s_{2i} - s_{2i-1})^{-1}$. In this discussion we have tacitly assumed that the systems, which arise from (16a), (16b) when the quotients are replaced by the

corresponding derivatives have nonzero determinants. The analysis carried out here may be extended to the case when $S \subset \mathbb{R}^k$ $k > 1$. See also [8].

The optimal solutions of the discretized versions of a SIP generally have $q = n$ even if the optimal solutions of the SIP itself have $q < n$. From the discussion above of the condition of the systems (15) we conclude that in this case there is an optimal roughness of the grid when the discretized version gives the best estimate of the optimal solution of Program (P). The decrease of the discretization errors which is obtained must namely be balanced off against the deterioration of the condition of the systems (15), which results from making the grid finer. But if $q = n$ the condition of the system should be largely independent of the fineness of the grid. Lastly if derivatives of a_r , $r = 1, 2, \dots, n$ and b are available this information could be used to replace (15) with (16) and hence achieve that the condition of the systems becomes largely unaffected by the fineness of the grid even if $q < n$. However, the derivatives could profitably be used in the iterative schemes to be treated in the next Section.

4. Iterative schemes

Let q , $\{s_1, s_2, \dots, s_q\}$ and $x_1, x_2, \dots, x_q$ be an optimal solution of Program (D), $y \in \mathbb{R}^n$ an optimal solution of Program (P). Then these entities must simultaneously satisfy (1b), (2b) and (2c). We have also the complementary slackness equation (see e.g. [3], [5], [10]).

$$(17) \quad x_i (y^T a(s_i) - b(s_i)) = 0, \quad i = 1, 2, \dots, q$$

Put

$$(18) \quad f(s) = y^T a(s) - b(s)$$

By (1b) we have $f(s) \geq 0$, $s \in S$ and by (17) $f(s_i) = 0$, $i = 1, 2, \dots, n$. Therefore we conclude:

$$(19) \quad f \text{ has a global minimum at } s_i, \text{ if } x_i > 0, i = 1, 2, \dots, q$$

This fact is used to generate further relations. Thus if f is differentiable and s_i is in the interior of S , then we get

$$\nabla f(s_i) = 0$$

since the *global* minimum at s_i is also a *local* minimum. But if s_i is at the boundary of S , then we combine the equations arising from the fact that s_i is at the boundary with the circumstance that a local minimum for f occurs there. In [12] this is expressed in a uniform manner by means of so-called Kuhn-Tucker conditions. Then S is defined as the requirement that certain inequalities must be satisfied. In all problems studied so far (19) generates k independent equations, if $S \subset \mathbb{R}^k$.

To summarize, the optimality conditions may be written as follows

$$(20a) \quad \sum_{i=1}^q x_i a(s_i) = c$$

$$(20b) \quad x_i f(s_i) = 0, \quad i = 1, 2, \dots, q$$

$$(20c) \quad f \text{ has a local minimum at } x_i, \quad i = 1, 2, \dots, q, \text{ where } f \text{ is defined by (18)}$$

$$(21a) \quad x_i \geq 0, \quad i = 1, 2, \dots, q$$

$$(21b) \quad f(s) \geq 0 \quad s \in S$$

The relations (20) are combined into a nonlinear system of equations from which optimal solutions of Programs (P) and (D) are calculated. It must then be verified that the proposed solutions meet (21). Note that (21b) generally defines infinitely many inequalities and hence cannot be checked computationally by means of finitely many operations. Sometimes analytical methods can be used to confirm that (21b) is met for a calculated optimal solutions. In many practical applications one may only be able to verify the statement that

$$f(s) \geq -\delta, \quad s \in S$$

where δ is a known positive number. Then it is not certain that the calculated vector $\bar{y}$ is optimal for Program (D). Bounds for the errors in $v(P)$ caused by substituting y for the (unknown) optimal solution are given in [10].

We next describe how to construct the system (20) and an approximate solution to it from an optimal solution $\bar{q}$, $\{\tilde{s}_1, \tilde{s}_2, \dots, \tilde{s}_q\}$, $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_q$, $\bar{y}$ of a discretized version of Programs (P) and (D) corresponding to a certain grid T . Define $\tilde{f}$ through

$$\tilde{f}(s) = a(s)^T \bar{y} - b(s), \quad s \in S$$

Then $\tilde{f}(s) \geq 0$, $s \in T$ but $\tilde{f}(s) < 0$ is possible for $s \notin T$.

We take now q equal to the number of local minima of f which are such that the corresponding values of f are nonpositive. Denote the positions of these minima by $\sigma_1, \sigma_2, \dots, \sigma_q$. With each σ_i we associate the $\tilde{s}_j$:s lying closest to σ_i . Let G_i be the set of the corresponding j -values. The vectors $\tilde{s}_j$ with j belonging to a certain G_i generally lie closely together. As approximation for x_i we take the sum of those $\tilde{x}_j$:s with j in G_i . As approximation for s_i we could take either the center of gravity of the $\tilde{s}_j$:s with $j \in G_i$ and weighted with the x_j :s or we could use σ_i itself. The former strategy tends to give a slightly smaller residuals in (20a) while the latter on the other hand should give zero residuals in the equations obtained from (20c). The calculated vector $\tilde{y}$ is taken as an initial estimate for y .

A general approach is now to linearize the system formed from (20) and improve the approximate solution by means of Newton-Raphson's method. Thus in each step we must solve a linear system with $n + (k+1)q$ equations. This is the method implemented in the computer codes of [2] where suitable stopping rules are given. Alternative methods are presented in [12], where rates of convergence are discussed in a general context. Sometimes (20) has a special structure which may be exploited to save labour. We mention two such instances:

- a) $q = n$. Then the equations generated by (20b), (20c) can be solved independently of those in (20a). Furthermore, in each iteration step one may first use (20b) to improve the estimate of y then (20c) to correct the estimates of s_i .
- b) (20a) can be solved independently. This occurs e.g. when S is a subset of the real line and $a_1, a_2, \dots, a_n$ and b form a Čebyšev system over S . Then the optimal solutions of Program (D) and the problem obtained when we *minimize* (2a) subject to (2b), (2c) correspond to generalized quadrature rules of the Gaussian type. A numerical method working for general Čebyšev systems is given in [6] but the very important special case when $a_r(s) = s^{r-1}$ is more efficiently solved by means of the method described in [4].

However, even in the general case a certain simplification is possible. Upon linearizing (20) we may use (20c) to express the corrections of s_i in terms of those of y . This relation is substituted into the linear equations obtained from (20a), (20b) whereupon the corrections in y and x_i are calculated. Then the improvements in s_i are found from (20c). Thus a decomposition of the linear system which must be solved in each Newton-Raphson iteration is achieved. Related ideas are discussed in [13].

In [11] it is shown that provided certain regularity conditions are met, Newton-Raphson's method converges to an approximate solution with small residuals even if the error in the solution itself may be relatively large. From Lemma 4 it transpires, that then the associated error in the optimal value is small since the residuals in (20a) correspond to $\tilde{c}_r - c_r$ and those of (20b) govern the size of Δ . Remaining numbers in the formula of the assertion of Lemma 4 are of moderate size when the assumptions of Lemmas 3 and 5 are valid.

The favourable theoretical results what regards the stability of computational schemes for solving SIP's agree with the experiences gained from actual calculations.

REFERENCES

- [1] Charnes, A., W.W. Cooper and K.O. Kortanek, *Duality, Haar programs and finite sequence spaces*, Proc. Nat. Acad. Sci., U.S. 48 (1962), 783-786
- [2] Fahlander, K., *Computer programs for semi-infinite optimization*, TRITA-NA-7312, Dept of Numerical Analysis, Royal Institute of Technology, S-10044 Stockholm 70, Sweden, 1973
- [3] Glashoff, K. and S.-Å. Gustafson, *Einführung in die Lineare Optimierung*, Wissenschaftliche Buchgesellschaft, Darmstadt, 1978
- [4] Golub, G.H. and J.H. Welsch, *Calculation of Gauss quadrature rules*, Math. Comp. 23 (1969), 221-230
- [5] Gustafson, S.-Å., *On the computational solution of a class of generalized moment problems*, SIAM J Numer. Anal. 7 (1970), 343-357
- [6] Gustafson, S.-Å., *Die Berechnung von verallgemeinerten Quadraturformeln vom Gausschen Typus, eine Optimierungsaufgabe*, in L. Collatz, W. Wetterling (Eds): *Numerische Methoden bei Optimierungsaufgaben*, ISNM 17 (1973), 59-71, Birkhäuser, Basel
- [7] Gustafson, S.-Å., *On computational applications of the theory of moment problems*, Rocky Mountain J on Math 4 (1974), 227-240
- [8] Gustafson, S.-Å., *Stability aspects on the numerical treatment of linear semi-infinite programs*, TRITA-NA-7604, Dept of Numerical Analysis, Royal Institute of Technology, S-10044 Stockholm 70, Sweden, 1976
- [9] Gustafson, S.-Å., *On semi-infinite programming in numerical analysis*, this volume
- [10] Gustafson, S.-Å. and K.O. Kortanek, *Numerical treatment of a class of semi-infinite programming problems*, NRLQ 20 (1973), 477-504
- [11] Gustafson, S.-Å., G. Dahlquist and L. Reichel, *Remarks on the convergence of Newton-Raphson's method in the presence of round-offs*, TRITA-NA-7615, Dept of Numerical Analysis, Royal Institute of Technology, S-10044 Stockholm 70, Sweden
- [12] Hettich, R., *Kriterien zweiter Ordnung für lokal beste Approximation*, Numer. Math. 22 (1974), 409-417
- [13] Hettich, R., *A Newton-method for nonlinear Chebyshev approximation* in R. Schaback, K. Scherer (Eds): *Approximation Theory, Lecture Notes in Math* 556 (1976), 222-236, Springer, Berlin-Heidelberg, New York