CONSTRUCTIVE APPROACHES TO APPROXIMATE SOLUTIONS OF OPERATOR EQUATIONS AND CONVEX PROGRAMMING

by

Henry Wolkowicz

A thesis submitted to the Faculty of Graduate Studies and Research, McGill University, in partial fulfillment of the requirements for the Degree of Doctor of Philosophy.

Department of Mathematics

McGill University

Montreal, Quebec

August, 1978.

To my mother

and

the memory of my father.

ABSTRACT

The research in this thesis lies in two related areas of applied mathematics: approximation and optimization. In the area of approximation, new classes of iterative methods are introduced to calculate best approximate solutions of operator equations in Banach spaces. Also, Kantorovich's approximation theory is extended to include, possibly inconsistent, operator equations. As a special case, the convergence of a Galerkin type method is established. In the second part of the thesis, the geometry of optimality conditions for nondifferentiable convex optimization problems is studied. Necessary and sufficient conditions, under which the Kuhn-Tucker theory is valid, are stated. The results are used to formulate a numerical algorithm and to calculate various objects which have recently appeared in the theory of optimization.

RÉSUMÉ

La recherche contenue dans cette thèse porte sur deux domaines voisins des mathématiques appliquées: l'approximation et l'optimisation. En théorie de l'approximation, on présente une nouvelle classe des méthodes itératives pour calculer les meilleures solutions approximatives d'équations d'opérateurs dans des espaces de Banach. On généralise aussie, la théorie d'approximation de Kantorovich au cas des équations d'opérateurs possiblement incohérentes. Dans ce contexte la preuve de convergence d'une méthode de type Galerkin apparaît comme un cas particulier. Dans la deuxième partie, on étudie la géométrie des conditions d'optimalité pour des problèmes non-différentiables d'optimisation convexe. On énonce également des conditions nécessaires et suffisantes pour que la théorie de Kuhn-Tucker soit valide. Ces résultats servent à formuler un algorithme numérique pour calculer certaines expressions qui on fait leur apparition récemment en théorie de l'optimisation.

ACKNOWLEDGEMENT

I would like to express my thanks to my supervisor, Professor S. Zlobec, whose personal interest in my work led to my studying mathematics. I would also like to thank Professor Zlobec for the help and advice received throughout my graduate career. His imaginative ideas have been the main influence in my thesis and in my research in general.

I would like to thank Professor G.P.H. Styan and my wife, Gail, for their continuous encouragement and support. In particular, I am grateful to Professor Styan for many enjoyable and helpful conversations and to my wife, Gail, for help in both discussing the contents and proofreading this manuscript.

Finally, I would like to thank Marlene Pyykko and Suzanna Clarkson for their careful and expert typing and Louis Paul Rivest for translating the abstract into French.

This research was supported by the National Research Council of Canada, le Gouvernement du Québec and McGill University.

TABLE OF CONTENTS

PART B. CONVEX PROGRAMMING

PART A

OPERATOR EQUATIONS

I. INTRODUCTION

1.

A solution of a consistent operator equation

$$Ax = b, \tag{1.1}$$

where A is a bounded linear operator from a Banach space X into itself and b is an element of X, can be calculated in two ways. One can use a simple iterative scheme set up in X, e.g. Krasnosel'skii et al. [33, Chapter 1], or an extension to Banach spaces of various well-known matrix iterative schemes, as suggested by e.g. Petryshyn [47], [49], Kammerer and Plemmons [29], and Gudder and Neumann [22]. The other way is to approximate the original equation (1.1) by a sequence of equations

$$\bar{A}\bar{x} = \bar{b}, \tag{1.2}$$

which are possibly easier to handle, and use appropriate error analysis. The latter approach is generally more successful. One of the first theories which studies the relationship between (1.1) and (1.2) was given by Kantorovich [30] and elaborated in the book by Kantorovich and Akilov [31]. Kantorovich's theory has been developed only for consistent equations. In particular, it is concerned with the following problems:

(i) Find conditions under which the consistency of (1.1) implies the consistency of (1.2).

(ii) If both (1.1) and (1.2) are consistent, estimate the distance between their solutions.

(iii) Find conditions under which the solutions of a sequence of approximate equations (1.2) converge to the solution of the equation (1.1).

(iv) Estimate the norm of A in terms of the norm of $\bar{A}$ and vice versa.

Kantorovich's approximation theory is rather general and, therefore, it is in principle applicable in many consistent situations, including the study and numerical treatment of infinite systems of linear equations, integral equations, ordinary differential equations and boundary value problems.

Various approximation theories have been recently developed and applied to particular problems by different authors, many of whom use the Kantorovich theory as a starting point. For instance, Thomas [59] refines some of Kantorovich's ideas and applies them to develop an approximation theory for the Nyström method of solving integral equations. Phillips [50] and Prenter [51] formulate approximation theories for the collocation method, while Ikebe [26] works with the Galerkin method. (For more details on these and other approaches for solving integral equations, see e.g. Houstis and Papatheodorou [25], Delves and Walsh [19] and Atkinson [7].)

Anselone [4] and Anselone and Moore [5] use the notion of collectively compact operators to formulate a different error analysis. Moore and Nashed [40] further developed the ideas of Anselone and Moore for possibly inconsistent operator equations in Banach spaces. They use the notions of generalized inverses of linear operators on Banach spaces and "best approximate" solutions of linear operator equations. Furthermore, they get, in special cases, some results in the perturbation theory of rectangular matrices obtained earlier by Ben-Israel [10] and Stewart [57].

An approximation theory for general, possibly inconsistent, linear equations in Hilbert spaces has been studied using the classical approach of Kantorovich (rather than the one of Moore and Nashed) by Zlobec [65]. One of the objectives of Part A is to continue the latter approach and formulate Kantorovich's theory for general, possibly inconsistent, linear equations in Banach spaces. The basic idea here is to establish and explore a relationship between approximate solutions of (1.1) and (1.2) and then use this relationship as a source for formulating various specific schemes for calculating approximate solutions of (1.1).

In the iterative computation of approximate solutions, as well as in Kantorovich's theory for singular equations, we will often use the concept of the generalized inverse of an operator. Some basic results on generalized inverses in Banach spaces are

summarized in Chapter II. In Chapter III, stationary and non-stationary iterative schemes are set up in Banach spaces for calculating both an approximate solution and the generalized inverse. This section extends from Hilbert to Banach spaces some results from the book by Ben-Israel and Greville [12, Chapter 8]. In Chapter IV, conditions for the consistency of $Ax = y$, for every y in a given subspace, are stated in terms of an approximate equation. Various error estimates are obtained as special cases. Kantorovich's theory for general linear equations is also formulated. The results are formulated in such a way that a comparison with the corresponding results for the nonsingular case from [31] is easily made. The most important results in this chapter are Theorem IV.3.1, which gives an error estimate, and Theorem IV.4.1, which gives conditions for convergence of approximate schemes. Using Kantorovich's theory, in Chapter V, a Galerkin-type method for calculating the best approximate solution is stated and its convergence is established for a class of operator equations in Banach spaces.

Situations where inconsistent linear operator equations arise are numerous and they include: integral equations in the theory of elasticity, potential theory and hydromechanics, e.g. Muskhelishvili [42], the integral formulation of the interior Neumann problem for the Laplacian, e.g. Kammerer and Nashed [27] and Atkinson [7], the eigenvalue problem in the case of a nonhomogeneous integral equation

when the associated homogeneous equation has a nontrivial solution, e.g. Kammerer and Nashed [27], and boundary value problems, e.g. Langford [34], Varga [62], Bramble and Shatz [16-18] and Serbin [54]. They also appear in the numerical solution of differential equations, for instance in the collocation method when the number of collocation points is bigger than the number of coefficients to be determined, e.g. Krasnosel'skii et al. [33] and in the numerical solution of nonlinear equations where the Fréchet derivative is singular, e.g. Boggs [15] and Gay [20]. If the number of collocation points is smaller than the number of coefficients, then, if consistent, the approximate equation (1.2) has infinitely many solutions and one may again be interested in calculating the best approximate solution. Under- and over-determined initial value problems have been studied by Lovass-Nagy and Powers [38]. In the finite dimensional case, the under- and over-determined systems appear frequently in statistics, e.g. Rao and Mitra [52], see also Ben-Israel and Greville [12], Abdelmalek [1], [2] and Anderson [3].

II. PRELIMINARIES

1. Notation

$X, Y, \overline{X}, \overline{Y}$	real or complex Banach spaces
$\ell(X,Y)$	the set of all linear operators from X into Y
$\ell_b(X,Y)$	the set of all bounded linear operators from X into Y
$\ell(X)$ and $\ell_b(X)$	the sets $\ell(X,X)$ and $\ell_b(X,X)$, respectively
ℓ_1	the Banach space of absolutely convergent sequences
$\|A\|$	the operator norm of A
$A_{\|S}$	A restricted to the set S
$\sigma(A)$	the spectrum of A
$\rho(A)$	the spectral radius of A
$R(A)$	the range space of A
$N(A)$	the null space of A
A^*	the adjoint of A.

For the above notions and their properties, see e.g. Taylor [58].

$R\{A,B\} = \{Z \in \ell_b(X,Y): Z = AUB$ for some $U \in \ell(X,Y)\}$, e.g. Ben-Israel [9]	
$\overline{M}$	the closure of the set M
P_M	the projection onto the set M, II.2
$M \oplus N$	the direct sum of M and N, II.2
M^c	the topological complement of M, II.2

x^*	the best approximate solution, II.2
y^*	an approximate solution, II.2
$\mathcal{D}(A)$	the domain of the operator A, II.2
A^+	the generalized inverse of A, II.2

2. Approximate and Best Approximate Solutions

In order to formulate iterative methods for calculating approximate solutions and develop Kantorovich's theory for general, possibly inconsistent, operator equations in Banach spaces, we employ the following notions.

A linear operator $P \in \mathcal{L}(X)$, is called a projection (of X) if $P^2 = P$. If $\mathcal{R}(P) = M$, then we denote P by P_M and call it the projection of X onto M. Every projection P_M decomposes X into two algebraic complements, $M = \mathcal{R}(P_M)$ and $N = \mathcal{R}(I - P_M)$. This implies $X = M + N$ and we write $N = M^c$. If M and N are both closed, then we say that M has a topological complement in X and write

$$X = M \oplus N. \tag{2.1}$$

For an example of decomposition (2.1), the reader is referred to Nashed's paper [43, p.327]. Recall that a closed subspace M of

X has a topological complement if and only if there exists a continuous projection P_M (of X), e.g. Taylor [58, p.241]. However, not every closed subspace has a topological complement, as shown by Murray [41] in 1937.

Consider $A \in \ell_b(X,Y)$. We shall assume that there exist continuous projections, $P_{N(A)} \in \ell_b(X)$ and $P_{R(A)} \in \ell_b(Y)$. (In particular, such an A must have a closed range.) $P_{N(A)}$ determines the complement $N(A)^c = (I - P_{N(A)})X$. Similarly, $P_{R(A)}$ determines the complement $R(A)^c = (I - P_{R(A)})Y$. Hence, $X = N(A) \oplus N(A)^c$ and $Y = R(A) \oplus R(A)^c$. When $A \in \ell_b(X,Y)$ and projections $P_{N(A)} \in \ell_b(X)$ and $P_{R(A)} \in \ell_b(Y)$ are given, then the system

$$A^+A = P_{N(A)^c} \tag{2.2}$$

$$AA^+ = P_{R(A)} \tag{2.3}$$

$$A^+P_{R(A)} = A^+ \tag{2.4}$$

always has a unique solution $A^+ \in \ell_b(Y,X)$, called the <u>generalized inverse</u> of A (relative to the projections $P_{N(A)}$ and $P_{R(A)}$). The operator A^+ then establishes a one-to-one correspondence between $R(A)$ and $N(A)^c$, i.e. $A^+|_{R(A)} = (A|_{N(A)^c})^{-1}$, e.g. Nashed [43], Kammerer and Plemmons [29]. Note that, by the closed graph theorem, A^+ is bounded when $R(A)$ is closed.

If z is any vector in $N(A)$, then $y^* = A^+b + z$ is called an <u>approximate solution of the equation</u> $Ax = b$ (relative to $P_{R(A)}$),

e.g. Gudder and Neumann [22], while $x^* = A^+b$ is called the best approximate solution of the equation $Ax = b$ (relative to $P_{N(A)}$ and $P_{R(A)}$), e.g. Moore and Nashed [40]. We see that $\{A^+b + z : z \in N(A)\}$ is the set of all solutions of the projectional equation $Ax = P_{R(A)}b$, while A^+b is the unique one which lies in $N(A)^c$.

Remark 2.1. If $R(A)$ is not closed but the complement $M = \overline{R(A)}^c$ exists, then A has a unique unbounded generalized inverse $A^+ \in \mathcal{L}(\mathcal{D}(A^+),X)$ relative to the complements M and $N(A)^c$, where $\mathcal{D}(A^+) = R(A) + M$. Many of the results herein can be extended to include this case as well as the converse case when A is a densely defined unbounded operator with closed range, see e.g. the approach in Nashed [44].

Remark 2.2. The term "best approximate" solution is used by Newman and Odell [46] under different circumstances. There $\hat{x}$ is a "best approximate" solution of $Ax = b$, where $A:X \to Y$, $b \in Y$ if, for every $x \in X$ with $x \neq \hat{x}$, either

$$\|A\hat{x} - b\| < \|Ax - b\|$$

or

$$\|A\hat{x} - b\| = \|Ax - b\| \quad \text{and} \quad \|\hat{x}\| < \|x\| .$$

(This corresponds to the notion of best least squares solution in the case of Hilbert spaces.) In order to avoid possible ambiguity, we shall refer to the above $\hat{x}$ as the "X, Y-best approximate"

solution of the equation $Ax = b$. If the norms on X and Y are strictly convex, then an "X, Y-best approximate" solution exists. If they are not strictly convex, then an "X, Y-best approximate" solution may not exist. In order to find $\hat{x}$, we need the notion of an X-projection (also called a "metric projection" by Blather, Morris and Wulbert in [14]). Suppose that S is a subspace of X. Then the mapping E_S is the X-projection onto S if, for every $x \in X$, $E_S x$ solves the minimization problem $\min_{y \in S} \|x - y\|$. In general, the mapping E_S is not linear. An instance in which E_S is linear is when S and S^c have a basis and the norm in X is a "TK norm" with respect to these bases, e.g. Singer [55]. In Hilbert spaces, E_S corresponds to the orthogonal projection P_S. When the "X, Y-best approximate" solution $\hat{x}$ exists, then $\hat{x} = Bb$, where $B = (I - E_{N(A)})A^+ E_{R(A)}$ and A^+ is any generalized inverse of A with respect to some $P_{N(A)^c}$ and $P_{R(A)}$. (Note that B need not be linear.) Thus, we see that when $E_{N(A)}$ and $E_{R(A)}$ are linear, one may choose $P_{N(A)^c} = I - E_{N(A)}$ and $P_{R(A)} = E_{R(A)}$ in which case the "X, Y-best approximate" solution $\hat{x} = Bb$ coincides with the "best approximate" solution $x^* = A^+ b$.

Suppose that Y (but not necessarily X) is a Hilbert space, $A \in \mathcal{L}_b(X,Y)$, A has closed range and $X = N(A) \oplus N(A)^c$. Then one may choose $P_{R(A)} = E_{R(A)}$, which is now the orthogonal projection on $R(A)$, i.e. $R(A)^c = R(A)^\perp$, and write $Y = R(A) \oplus R(A)^c$.

If A^+ is the generalized inverse of A with respect to $P_{N(A)^c}$ and $P_{R(A)}$, the best approximate solution $x^* = A^+ b$ is the unique least squares solution of $Ax = b$ in $N(A)^c$, i.e. x^* solves the problem

$$\min_{x \in X} \|Ax - b\| \tag{2.5}$$

and among all solutions of (2.5), it is the only one in $N(A)^c$, e.g. Kammerer and Plemmons [29]. The set of all least squares solutions corresponds to the set of all approximate solutions of $Ax = b$ with respect to $R(A)^c = R(A)^\perp$. If both X and Y are Hilbert spaces and $A \in \mathcal{L}_b(X,Y)$ has closed range, we may choose $P_{R(A)} = E_{R(A)}$ and $P_{N(A)^c} = I - E_{N(A)}$. These are now the orthogonal projections, i.e. $N(A)^c = R(A^*)$, $R(A)^c = N(A^*)$ and $\hat{x} = Bb = A^+ b = x^*$ is the best least squares solution of the equation $Ax = b$. This means that x^* is the only solution in $N(A)^c$ of the minimization problem (2.5) and among all solutions of (2.5) it is the unique one of smallest norm. For a detailed discussion of the generalized inverse and best least squares solution in Hilbert spaces, the reader is referred to the book by Ben-Israel and Greville [12].

III. ITERATIVE METHODS

1. Stationary Methods

In order to calculate an approximate solution y^* of the operator equation (I.1.1)

$$Ax = b,$$

where $A \in \ell_b(X,Y)$, $P_{N(A)^c} \in \ell_b(X)$ and $P_{R(A)} \in \ell_b(Y)$, one can use the following iterative scheme:

(1.1) $$x_{k+1} = x_k - BAx_k + Bb, \quad k = 0,1,\ldots,$$

where

$$B \in R\{P_{N(A)^c}, P_{R(A)}\}.$$

This scheme has been suggested for calculating the best least squares solution in Hilbert spaces in [65], see also [12, p.356].

Theorem 1.1. Let $A \in \ell_b(X,Y)$, $b \in Y$, $P_{N(A)^c} \in \ell_b(X)$, $P_{R(A)} \in \ell_b(Y)$ and $B \in R\{P_{N(A)^c}, P_{R(A)}\}$ be given. Then the sequence $\{x_k\}$, generated by (1.1), converges, for any $x_0 \in X$, to the approximate solution $y^* = x^* + P_{N(A)}x_0$ of $Ax = b$, if $\rho(P_{N(A)^c} - BA) < 1$. Moreover, if $x_0 \in N(A)^c$, then $y^* = x^*$, i.e. we obtain the best approximate solution.

Proof. We find that

$$x_{k+1} - x^* = (I - BA)x_k + Bb - x^*, \quad \text{by (1.1)}$$

$$= (I - BA)x_k + BP_{R(A)}b - x^* , \quad \text{since} \quad B \in R\{P_{N(A)^c}, P_{R(A)}\}$$

$$= (I - BA)(x_k - x^*) , \quad \text{since} \quad P_{R(A)}b = Ax^*$$

$$= (I - BA)^{k+1}(x_0 - x^*) , \quad \text{by iteration}$$

$$= (I - BA)^{k+1}(P_{N(A)}x_0 + P_{N(A)^c}x_0 - x^*)$$

$$= (P_{N(A)^c} - BA)^{k+1}(x_0 - x^*) + P_{N(A)}x_0 , \quad \text{since} \quad x^* \text{ is in } N(A)^c . \tag{1.2}$$

But $\rho(P_{N(A)^c} - BA) = \lim_{n\to\infty} \sup \|(P_{N(A)^c} - BA)^n\|^{1/n} < 1$, by the property of ρ (e.g. Taylor [58]) and the assumption. Therefore, there exists a real number s and a positive integer n_0 such that

$$\|(P_{N(A)^c} - BA)^n\|^{1/n} \le s < 1 , \quad \text{for all} \quad n \ge n_0 .$$

Hence, $\|(P_{N(A)^c} - BA)^n\| \le s^n \to 0$ as $n \to \infty$. This implies $(P_{N(A)^c} - BA)^n \to 0$ as $n \to \infty$. Thus x_k converges to $x^* + P_{N(A)}x_0$, by (1.2). □

Remark 1.1. Necessary conditions for convergence of x_k to $x^* + P_{N(A)}x_0$, for every $x_0 \in X$, are

$$\rho(P_{N(A)^c} - BA) \le 1 \quad [\text{not} \quad \rho(P_{N(A)^c} - BA) < 1 \ !]$$

and

$$P_{N(A)^c} - BA \quad \text{has no eignevalue} \quad \lambda \quad \text{such that} \quad |\lambda| = 1 .$$

Proof. When $x_k \to x^* + P_{N(A)}x_0$ for any $x_0 \in X$, then

$$(P_{N(A)^c} - BA)^k(x_0 - x^*) \to 0 \quad \text{as} \quad k \to \infty,$$

by (1.2). Hence, for every $x \in X$,

$$\sup_{k\geq 1} \|(P_{N(A)^c} - BA)^k x\| < \infty.$$

Now, by the Banach-Steinhaus theorem, there exists $M > 0$ such that $\|(P_{N(A)^c} - BA)^k\| \leq M$, $k = 1,2,\ldots$. But

$$[\rho(P_{N(A)^c} - BA)]^k = \rho[(P_{N(A)^c} - BA)^k], \quad \text{by the spectral mapping theorem}$$

$$\leq \|(P_{N(A)^c} - BA)^k\| \leq M.$$

Therefore, $\rho(P_{N(A)^c} - BA) \leq 1$. In order to prove the second necessary condition, we observe that if $0 \neq x \in X$ and $|\lambda| = 1$, such that $(P_{N(A)^c} - BA)x = \lambda x$, then $x \in N(A)^c$ and $(P_{N(A)^c} - BA)^k x = \lambda^k x \not\to 0$ as $k \to \infty$, contradicting $x_k \to x^*$, by (1.2). □

Example 1.1. The above remark is demonstrated by the operator $A \in \ell_b(\ell_1)$ defined on $x = (x_i)$ by

$$(Ax)_1 = 0 \quad \text{and} \quad (Ax)_i = (1 - 2^{-i})x_i, \quad i = 2,3,\ldots.$$

It is clear that

$$N(A) = \{x \in \ell_1 : x_i = 0, \ i = 2,3,\ldots\}$$

and

$$R(A) = \{x \in \ell_1 : x_1 = 0\} .$$

If we choose $N(A)^c = R(A)$, $R(A)^c = N(A)$ and $B = 2A$, then $P_{N(A)^c}$ is defined by

$$(P_{N(A)^c}x)_1 = 0 \quad \text{and} \quad (P_{N(A)^c}x)_i = x_i , \quad i = 2,3,\ldots ,$$

while $BA = 2A^2$ is defined by

$$(BAx)_1 = 0 \quad \text{and} \quad (BAx)_i = 2(1 - 2^{-i})^2 x_i , \quad i = 2,3,\ldots .$$

Therefore

$$\begin{aligned} \rho(P_{N(A)^c} - BA) &= \sup_{i\geq 2}\{|1 - 2(1 - 2^{-i})^2|\} \\ &= \sup_{i\geq 2}\{|-1 + 2^{2-i} - 2^{1-2i}|\} \\ &= 1 . \end{aligned}$$

But, for every $x \in X$,

$$((I - BA)^k x)_i = \begin{cases} x_1 , & \text{if } i = 1 \\ (-1 + 2^{2-i} - 2^{1-2i})^k x_i , & \text{if } i = 2,3,\ldots . \end{cases}$$

Now, let $\varepsilon > 0$ be given. Then, since

$$x \in \ell_1 \quad \text{and} \quad |-1 + 2^{2-i} - 2^{1-2i}| < 1 \quad \text{for} \quad i = 2,3,\ldots ,$$

there exists integers $N \geq 2$ and K such that

$$\sum_{i=N+1}^{\infty} |x_i| < \varepsilon \quad \text{and} \quad \sum_{i=2}^{N} |-1 + 2^{2-i} - 2^{1-2i}|^k |x_i| < \varepsilon ,$$

when $k \geq K$. Therefore

$$\|(I - BA)^k x - P_{N(A)} x\| = \sum_{i=2}^{\infty} |((I - BA)^k x)_i| < 2\varepsilon ,$$

when $k \geq K$. This implies that

$$(I - BA)^k x \to P_{N(A)} x , \quad \text{for any } x \in \ell_1 ,$$

which, in turn, proves convergence of the iterative scheme, by (1.2).

Remark 1.2. It is a consequence of Remark 1.1 that Theorem 1.1 gives not only a sufficient but also a necessary condition for convergence, if X is finite dimensional. In the case of a Hilbert space X and a normal operator $T = P_{N(A)^c} - BA$, one can show that $T^n x \to 0$, for all $x \in X$, if and only if $\rho(T) \leq 1$ and the spectrum of T has no mass on the unit circle $|z| = 1$. For, by the spectral theorem, e.g. Rudin [53],

$$T^n x = \int_{\sigma(T)} \lambda^n \, dE_\lambda x .$$

And since, on $\sigma(T)$, $|\lambda^n| \leq 1$ and $\lambda^n \to 0$ almost everywhere, the Lebesgue dominated convergence theorem implies that $T^n x \to 0$.

This proves sufficiency. Now suppose that the spectrum of T has mass on the unit circle. By Remark 1.1, to prove necessity we need only find $x \in X$ such that $T^n x \not\to 0$. By the assumption, there exists a Borel subset ω of the unit circle such that, if the subspace M is the range of the projection $E_\lambda(\omega)$, then $M \neq \{0\}$. But then T is invariant on M and $\sigma(T_{|M}) = \omega$ is a subset of the unit circle. This implies that T is unitary on M and thus

$$T^n x \not\to 0, \quad \text{for all} \quad 0 \neq x \in M.$$

Specifying

$$P_{N(A)^c} = P_{R(A^*)} \quad \text{and} \quad B = \alpha A^*,$$

one can show, see e.g. Petryshyn [47], that

$$\rho(T) = \| P_{R(A^*)} - \alpha A^* A \| \leq 1$$

if and only if

$$0 \leq \alpha \leq \frac{2}{\|A^*A\|}.$$

One now establishes the following characterization of convergence: For any $x_0 \in X$, the sequence $\{x_k\}$ converges to the least squares solution $y^* = x^* + P_{N(A)} x_0$ if and only if

(i) $$0 < \alpha \leq \frac{2}{\|A^*A\|}$$

and

(ii) $\|A^*A\|$ is not an eigenvalue of A^*A if $\alpha = \frac{2}{\|A^*A\|}$.

We recall that Petryshyn gives only the sufficient condition

$$0 < \alpha < \frac{2}{\|A^*A\|} .$$

Specifying B in (1.1), one obtains various iterative schemes for computing an approximate solution. In particular, if one splits $A = M + N$, chooses M^+ with respect to the continuous projections $P_{R(M)}$ and $P_{N(M)^c}$ such that $P_{R(M)} = P_{R(M)}P_{R(A)}$ and $P_{N(M)^c} = P_{N(A)^c}P_{N(M)^c}$ and specifies $B = \omega M^+$, $\omega \neq 0$, then (1.1) becomes

$$x_{k+1} = [(1 - \omega)I - \omega M^+ N]x_k + \omega M^+ b . \tag{1.3}$$

Further, for $\omega = 1$, (1.3) becomes

$$x_{k+1} = -M^+ N x_k + M^+ b . \tag{1.4}$$

If both A and M are invertible, then (1.3) and (1.4) become, respectively,

$$x_{k+1} = [(1 - \omega)I - \omega M^{-1} N]x_k + \omega M^{-1} b \tag{1.5}$$

and

$$x_{k+1} = -M^{-1} N x_k + M^{-1} b . \tag{1.6}$$

The scheme (1.5) has been studied by Pertryshyn [48], who calls it the "Extrapolated Jacobi Method". The scheme (1.6) is the well-known Jacobi method. Other methods can be obtained by the splitting

$A = D + S + Q$ with $B = \left(\frac{1}{\omega}D + S\right)^{+}$, $\omega \neq 0$ where $\frac{1}{\omega}D + S \in \ell_b(X,Y)$

and $P_{R\left(\frac{1}{\omega}D+S\right)} = P_{R\left(\frac{1}{\omega}D+S\right)} P_{R(A)}$, $P_{N\left(\frac{1}{\omega}D+S\right)^c} = P_{N(A)^c} P_{N\left(\frac{1}{\omega}D+S\right)^c}$.

Then (1.1) becomes

$$x_{k+1} = (D + \omega S)^{+}[(1 - \omega)D - \omega Q]x_k + \omega(D + \omega S)^{+}b . \tag{1.7}$$

If both A and $D + \omega S$ are invertible, the scheme (1.7) becomes

$$x_{k+1} = (D + \omega S)^{-1}[(1 - \omega)D - \omega Q]x_k + \omega(D + \omega S)^{-1}b , \tag{1.8}$$

which is known as the "Successive Over-Relaxation Method" (abbreviated SOR method). Specifying $\omega = 1$ in (1.8) one obtains

$$x_{k+1} = -(D + S)^{-1}Qx_k + (D + S)^{-1}b , \tag{1.9}$$

which is known as the Gauss-Seidel method. In the case of an $n \times n$ invertible matrix $A = (a_{ij})$, one frequently specifies

$$D = M = (a_{ii}) , \quad i = 1,\ldots,n$$

$$S = (a_{ij}) , \quad i > j , \quad i = 1,\ldots,n; \quad j = 1,\ldots,n-1$$

$$Q = A - D - S ,$$

$$N = A - M$$

in (1.5), (1.6), (1.8), and (1.9). Properties of these schemes for systems with invertible matrices (and linear operators) have been studied extensively, see e.g. Varga [60], [61] and Petryshyn [47], [48]. Scheme (1.4) has been studied by Berman and Plemmons [13]

for systems with singular matrices and by Gudder and Neumann [22] for singular operator equations in Hilbert space. For some other schemes using splittings of A, see e.g. Hadjidimos [23] and Meijerink and van der Vorst [39].

Remark 1.3. One calls the splitting $A = M + N$ a proper splitting, if $R(A) = R(M)$, $R(A)^c = R(M)^c$, $N(A) = N(M)$ and $N(A)^c = N(M)^c$. Note that to obtain a proper splitting in the case of Hilbert spaces and orthogonal complements, one need only check that $N(A) = N(M)$ and $R(A) = R(M)$, e.g. [13] and [22]. The proper splittings are not only useful in iterative calculation of least squares solutions, but they also play an important role in the Kantorovich approximation theory (see Chapter V).

One can slightly modify (1.1) in order to compute A^+, the generalized inverse of A relative to given projections, $P_{N(A)}$ and $P_{R(A)}$.

Theorem 1.2. Let $A \in \ell_b(X,Y)$, $P_{N(A)} \in \ell_b(X)$ and $P_{R(A)} \in \ell_b(Y)$. If $B \in R\{P_{N(A)^c}, P_{R(A)}\}$, then the sequence $\{X_k\}$, generated by

$$X_{k+1} = X_k - BAX_k + B, \quad k = 0,1,2,\ldots, \tag{1.10}$$

converges to $A^+ + P_{N(A)}X_0$ for all $X_0 \in \ell_b(Y,X)$, if $\rho(P_{N(A)^c} - BA) < 1$. Moreover, if $R(X_0) \subset N(A)^c$, then we obtain the generalized inverse A^+.

Proof. Here

$$X_{k+1} - A^+ = X_k - BAX_k + B - A^+$$

$$= (I - BA)X_k + BP_{R(A)} - P_{N(A)^c}A^+$$

$$= (I - BA)(X_k - A^+), \quad \text{by (II.2.3)}$$

$$= (P_{N(A)^c} - BA)^{k+1}(X_0 - A^+) + P_{N(A)}X_0 .$$

The rest of the proof is analogous to the proof of Theorem 1.1 □

Note that, for $b \in Y$,

$$(A^+ + P_{N(A)}X_0)b = A^+b + P_{N(A)}X_0b$$

is an approximate solution of $Ax = b$. Moreover, when X_0 satisfies the condition

$$X_0 = X_0P_{R(A)},$$

we see that $A^+ + P_{N(A)}X_0$ is the generalized inverse of A relative to (i) the same $P_{R(A)}$ as A^+, and, by (II.2.2), (ii) the new projection on $N(A)$,

$$P_{N(A)} + P_{N(A)}X_0A .$$

If (1.10) is modified as follows:

(1.11) $$X_{k+1} = X_k - BX_k + B, \quad k = 0,1,2,\ldots ,$$

and if one chooses

$$B \in R\{P_{R(A)}, P_{R(A)}\} \quad \text{and} \quad X_0 \in \ell_b(Y), \tag{1.12}$$

such that $R(X_0) \subset R(A)$, then the sequence generated has the property

$$X_{k+1} - P_{R(A)} = (P_{R(A)} - B)^{k+1}(X_0 - P_{R(A)}), \quad k = 0,1,2,\ldots .$$

Hence one concludes that whenever $\rho(P_{R(A)} - B) < 1$, the sequence $\{X_k\}$, generated by (1.11) and (1.12), converges to the projection $P_{R(A)}$. If X and Y are Hilbert spaces and $P_{R(A)}$ is the orthogonal projection on $R(A)$, then one can choose

$$X_0 = AZ_1 \quad \text{and} \quad B = AZ_2A^*,$$

where $Z_1 \in \ell_b(Y,X)$ and $Z_2 \in \ell_b(X,X)$. In particular, one can specify $Z_1 = A^*$ and $Z_2 = \alpha I$, where α is a real parameter with the property $\rho(P_{R(A)} - \alpha AA^*) < 1$.

2. Nonstationary Methods

We can further modify the scheme (1.1) by varying B at each step. Such methods include gradient methods of finding approximate solutions.

Theorem 2.1. Let $A \in \ell_b(X,Y)$, $P_{N(A)} \in \ell_b(X)$ and $P_{R(A)} \in \ell_b(Y)$. If $B_k \in R\{P_{N(A)^c}, P_{R(A)}\}$ for $k = 0,1,2,\ldots$, then the sequence x_k generated by

$$x_{k+1} = x_k - B_k A x_k + B_k b, \quad k = 0,1,2,\ldots \tag{2.1}$$

converges, for any $x_0 \in X$, to the approximate solution $y^* = x^* + P_{N(A)} x_0$ of $Ax = b$, if $\|P_{N(A)^c} - B_k A\| \leq 1 - \alpha_k$ for $k = 0,1,2,\ldots$, where $0 \leq \alpha_k \leq 1$ and $\sum \alpha_k = \infty$. Moreover, if $x_0 \in N(A)^c$, then $y^* = x^*$, i.e. we obtain the best approximate solution.

Proof. As in the proof of Theorem 1.1, we see that

$$x_{k+1} - x^* = \left[\prod_{i=0}^{k} (P_{N(A)^c} - B_i A)\right](x_0 - x^*) + P_{N(A)} x_0 .$$

But

$$\left\| \prod_{i=0}^{k} (P_{N(A)^c} - B_i A) \right\| \leq \prod_{i=0}^{k} (1 - \alpha_k) \to 0$$

if and only if the series $\sum \alpha_k$ diverges, e.g. Knopp [32, p.92]. □

Specifying B_k in (2.1) again leads to various iterative schemes for computing the best approximate solution. For example, if X and Y are Hilbert spaces and we choose $B_k = \alpha_k A^*$, for some scalars α_k, $k = 0,1,2,\ldots$, then we can rewrite (2.1) as

(2.2) $$x_{k+1} = x_k - \alpha_k A^*(Ax_k - b) .$$

Since $2A^*(Ax - b)$ is the gradient of the function $\|Ax - b\|^2$, we see that (2.2) defines a gradient method for minimizing $\|Ax - b\|$, with step size given by $\alpha_k/2$.

If we let $\Gamma_k = A^*(Ax_k - b)$ and $\alpha_k = \|\Gamma_k\|^2 \|A\Gamma_k\|^{-2}$, then (2.2) becomes the method of steepest descent. For a discussion on gradient methods, see e.g. Nashed [44, p.380] and Kammerer and Nashed [28].

3. Examples

The iterative scheme (1.1) can be used to calculate an approximate solution of the equation (I.1.1) in abstract spaces. However, in many situations, it is actually used to calculate an approximate solution of an approximate equation (I.1.2) which is frequently a more manageable finite system of linear algebraic equations. The first case will now be demonstrated.

Example 3.1. Let us calculate the best least squares solution x^* of the inconsistent equation $Ax(s) = (I - K)x(s) = b(s)$, where

$$Kx(s) = \frac{2}{\pi}\int_0^{\pi} (\sin s \sin \xi + \tfrac{1}{2}\cos s \cos \xi)x(\xi)\, d\xi$$

and $b(s) = s$. The operator K is chosen from Stakgold [56]. The problem will be solved in $X = Y = L_2[0,\pi]$ using the iterative scheme (1.1).

We choose $B = \alpha A^*$, where $0 < \alpha < 2 = \frac{2}{\|A^*A\|}$. This guarantees here that $\rho(P_{R(A)} - \alpha AA^*) < 1$ and that the scheme (1.1) converges to x^* for every choice of $x_0(s)$ in $N(A)^c = R(A^*)$. For $x_0(s) = 0$ one finds:

$$x_1(s) = \alpha(s - 2\sin s + \frac{2}{\pi}\cos s)$$

$$= \alpha_1^{(1)} s - \alpha_2^{(1)} \sin s + \alpha_3^{(1)} \cos s,$$

where $\alpha_1^{(1)} = \alpha$, $\alpha_2^{(1)} = 2\alpha$ and $\alpha_3^{(1)} = \frac{2\alpha}{\pi}$.

$$x_2(s) = [(1-\alpha)\alpha_1^{(1)} + \alpha]s - [(1-\alpha)\alpha_2^{(1)} + 2\alpha]\sin s$$

$$+ [(1 - \frac{\alpha}{4})\alpha_3^{(1)} - \frac{3\alpha}{\pi}\alpha_1^{(1)} + \frac{2}{\pi}\alpha]\cos s$$

$$= \alpha_1^{(2)} s - \alpha_2^{(2)} \sin s + \alpha_3^{(2)} \cos s,$$

where $\alpha_1^{(2)} = (1-\alpha)\alpha_1^{(1)} + \alpha$, $\alpha_2^{(2)} = (1-\alpha)\alpha_2^{(1)} + 2\alpha$, and $\alpha_3^{(2)} = (1 - \frac{\alpha}{4})\alpha_3^{(1)} - \frac{3\alpha}{\pi}\alpha_1^{(1)} + \frac{2}{\pi}\alpha$. In general, if

$$x_k(s) = \alpha_1^{(k)} s - \alpha_2^{(k)} \sin s + \alpha_3^{(k)} \cos s$$

then

$$x_{k+1}(s) = [(1-\alpha)\alpha_1^{(k)} + \alpha]s - [(1-\alpha)\alpha_2^{(k)} + 2\alpha]\sin s$$
$$+ [(1-\frac{\alpha}{4})\alpha_3^{(k)} - \frac{3\alpha}{\pi}\alpha_1^{(k)} + \frac{2}{\pi}\alpha]\cos s$$
$$= \alpha_1^{(k+1)}s - \alpha_2^{(k+1)}\sin s + \alpha_3^{(k+1)}\cos s .$$

Since the iterative schemes

$$\alpha_1^{(k+1)} = (1-\alpha)\alpha_1^{(k)} + \alpha$$
$$\alpha_2^{(k+1)} = (1-\alpha)\alpha_2^{(k)} + 2\alpha$$
$$\alpha_3^{(k+1)} = (1-\frac{\alpha}{4})\alpha_3^{(k)} - \frac{3\alpha}{\pi}\alpha_1^{(k)} + \frac{2}{\pi}\alpha$$

are convergent themselves with the solutions $\alpha_1 = 1$, $\alpha_2 = 2$ and $\alpha_3 = -\frac{4}{\pi}$, respectively, one concludes that

$$x^*(s) = s - 2\sin s - \frac{4}{\pi}\cos s$$

is the best least squares solution.

Using (1.10) with B defined by $Bx(s) = x(s) - \frac{2}{\pi}\int_0^{\pi}\sin s \sin\xi\, x(\xi)\, d\xi$, one can show that X_k converges to A^+, which is here

$$A^+x(s) = x(s) - \frac{2}{\pi}\int_0^{\pi}(\sin s \sin\xi - \cos s \cos\xi)x(\xi)\, d\xi .$$

The best least squares solution of the equation introduced in Example 3.1 will be calculated in Example V.2.1 via Kantorovich's approximation theory. We conclude this section by demonstrating how the iterative scheme (1.1) can be applied to matrices.

Example 3.2. We now calculate the best least squares solution, using the iterative scheme (1.1), of the inconsistent system

$$\begin{aligned} x_1 \qquad + 3x_3 &= 1 \\ -x_1 + x_2 \qquad &= 1 \\ x_1 - x_2 \qquad &= 1 \\ x_2 + x_3 &= 0 . \end{aligned}$$

Here

$$A = \begin{pmatrix} 1 & 0 & 3 \\ -1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix} .$$

Specifying

$$B = \alpha A^t , \qquad \text{where} \quad \alpha = \frac{2}{\operatorname{trace} A^t A}$$

one obtains

$$B = \frac{1}{8}\begin{pmatrix} 1 & -1 & 1 & 0 \\ 0 & 1 & -1 & 1 \\ 3 & 0 & 0 & 1 \end{pmatrix} .$$

It is easy to verify that for the above choice of B and α, $\rho(P_{R(A^t)} - BA) < 1$, whenever rank $A > 1$. One can start iterating from $x^0 = 0$, in which case the following numerical results are obtained:

k	x_1^k	x_2^k	x_3^k
1	0.2500000	0.0000000	0.2500000
2	0.2500000	0.0625000	0.3125000
3	0.2656250	0.0625000	0.3281230
4	0.2656260	0.0664063	0.3320313
5	0.2666016	0.0664063	0.3330018
6	0.2666016	0.0666504	0.3332520
7	0.2666626	0.0666504	0.3333130
8	0.2666626	0.0666657	0.3333282
9	0.2666664	0.0666657	0.3333321
10	0.2666664	0.0666666	0.3333330
11	0.2666667	0.0666666	0.3333333

The eleventh approximation x^{11} gives the best least squares solution correct to six decimal places:

$$x_1^* = \frac{4}{15}, \quad x_2^* = \frac{1}{15}, \quad x_3^* = \frac{1}{3}.$$

Example 3.3. In this example we apply scheme (1.1) to solve the problem

$$\min_{x \in X} \|Ax - b\|$$

where X is a two-dimensional Banach space of scalars

$$x = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \quad A = \begin{pmatrix} 2 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

If we specify the norm

$$\|x\| = |\xi_1| + |\xi_2| + \max\{|\xi_1|, |\xi_2|, |\xi_1 + \xi_2|\}$$

(which is a TK norm with respect to the basis $\binom{1}{0}, \binom{0}{1}$), then $E_{R(A)}$ is linear and equal to

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} .$$

Thus, in order to solve the problem, we must choose $P_{R(A)} = E_{R(A)}$. Since $N(A)^c$ can be arbitrarily chosen, let it be $N(A)^c = R(A)$. Hence

$$P_{N(A)^c} = \begin{pmatrix} 1 & \frac{1}{2} \\ 0 & 0 \end{pmatrix} .$$

Further we choose

$$B = \tfrac{1}{4} P_{N(A)^c} P_{R(A)} = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad x_0 = 0 .$$

The sequence $x_1 = \frac{1}{4}\binom{1}{0}$, $x_2 = \frac{3}{8}\binom{1}{0}$, ... converging to $x^* = \frac{1}{2}\binom{1}{0}$ is obtained.

IV. KANTOROVICH'S THEORY FOR SINGULAR EQUATIONS

1. Condition for the Consistency of the Exact Equation

One of the most useful results in the formulation of the classical Kantorovich theory is a lemma which gives a condition for the consistency of the exact equation (I.1.1), e.g. [31, p.543]. This lemma will now be extended so that it also applies to singular equations.

Lemma 1.1. Let $V \in \ell_b(X,Y)$, E a closed subspace of X and F any subspace of Y containing $V(E)$. If there exists, for every $y \in F$, an $\hat{x} \in E$ such that

$$\|V\hat{x} - y\| \leq q\|y\| \quad \text{and} \quad \|\hat{x}\| \leq \alpha\|y\| \tag{1.1}$$

where $q < 1$ and α are constants, then the equation

$$Vx = y \tag{1.2}$$

has, for every $y \in F$, a solution $x \in E$ satisfying

$$\|x\| \leq \frac{\alpha}{1-q}\|y\| . \tag{1.3}$$

Proof. Similarly to the proof in [31], we will construct an exact solution of (1.2) by recursion. Take an arbitrary $y \in F$. Set $y_1 = y$. By hypothesis, an $\hat{x}_1 \in E$ exists such that

$$\|V\hat{x}_1 - y_1\| \leq q\|y_1\| , \quad \|\hat{x}_1\| \leq \alpha\|y_1\| . \tag{1.4}$$

Denote

(1.5) $$y_2 = y_1 - V\hat{x}_1 .$$

Clearly $y_2 \in F$, since $y_1 \in F$ and F is a subspace containing $V(E)$. We now apply the condition (1.1) to y_2. This implies the existence of $\hat{x}_2 \in E$ such that

$$\|V\hat{x}_2 - y_2\| \le q\|y_2\| = q\|y_1 - V\hat{x}_1\| , \quad \text{by (1.5)}$$

$$\le q^2\|y_1\| , \quad \text{by (1.4)}.$$

Also $\|\hat{x}_2\| \le \alpha\|y_2\| \le \alpha q\|y_1\|$. Continuing this process, sequences $\{y_k\}$ and $\{\hat{x}_k\}$ are obtained such that

(1.6) $$y_{k+1} = y_k - V\hat{x}_k , \quad k = 1,2,...$$

and

(1.7) $$\|y_k\| \le q^{k-1}\|y_1\| , \quad \|\hat{x}_k\| \le \alpha q^{k-1}\|y_1\| .$$

By iteration, (1.5) and (1.6) give

(1.8) $$y_{k+1} = y_1 - V(\hat{x}_1 + \hat{x}_2 + ... + \hat{x}_k) , \quad k = 1,2,... .$$

Using the second inequality in (1.7), and recalling that $y_1 = y$, one obtains

$$\Big\| \sum_{k=1}^{\infty} \hat{x}_k \Big\| \le \sum_{k=1}^{\infty} \|\hat{x}_k\| \le \alpha \sum_{k=1}^{\infty} q^{k-1}\|y\| .$$

Since $q < 1$, the series $\sum_{k=1}^{\infty} \hat{x}_k$ is convergent. Hence $x \triangleq \sum_{k=1}^{\infty} \hat{x}_k$

belongs to E, since E is a closed subspace. Furthermore

$$\|x\| \leq \alpha \sum_{k=1}^{\infty} q^{k-1}\|y\| = \frac{\alpha}{1-q}\|y\| .$$

So taking limits in (1.8) gives

$$\lim_{k\to\infty} y_k = \lim_{k\to\infty} [y_1 - V(\hat{x}_1 + \hat{x}_2 + \ldots + \hat{x}_k)]$$

$$= y_1 - Vx , \quad \text{by the continuity of } V .$$

Also, since $q < 1$, we see from (1.7) that $\lim_{k\to\infty} y_k = 0$. Thus $Vx = y$. We have shown that $x \in E$ is a solution of (1.2) and it satisfies (1.3). □

Lemma 1.1 has been proved in [31] in the special case when $E = X$ and $F = Y$. The above result will be used in the next section in the approximation theory. However, Lemma 1.1 is of an independent interest and in the remainder of this section we will show how, using the lemma, one can establish some new and some well-known estimates related to the equation (I.1.1)

Proposition 1.1. Let $M \in \mathfrak{L}_b(X,Y)$ and let $N(M)$ and $R(M)$ have topological complements. Denote by M^+ the generalized inverse of M with respect to these topological complements. Consider

(1.9) $$A = M + N$$

such that $A \in \mathcal{L}_b(X,Y)$ and $\|NM^+\| < 1$. If

$$R(N) \subset R(M), \tag{1.10}$$

then the equation

$$Ax = y \tag{1.11}$$

has, for every $y \in R(M)$, a solution $x^* \in N(M)^c$. Also

$$\|x^*\| \leq \frac{\|M^+\|}{1 - \|NM^+\|}\|y\| \tag{1.12}$$

and $R(A) = R(M)$. In addition to the above assumptions, if

$$N(M)^c \cap N(A) = \{0\}, \tag{1.13}$$

then

$$\|A^+\| \leq \frac{\|M^+\|}{1 - \|NM^+\|}, \tag{1.14}$$

where A^+ denotes the generalized inverse of A with respect to $P_{N(A)^c} = P_{N(M)^c}$ and $P_{R(A)} = P_{R(M)}$.

Proof. We will show that the assumptions of Lemma 1.1 are satisfied with $V = A$, $E = N(M)^c$, $F = R(M)$, $q = \|NM^+\|$ and $\alpha = \|M^+\|$. Choose an arbitrary $y \in R(M)$ and let $\hat{x} = M^+y$. Then

$$\begin{aligned}\|A\hat{x} - y\| &= \|(M+N)\hat{x} - y\| \\ &= \|(M+N)\hat{x} - MM^+y\|, \quad \text{since } y \in R(M) \\ &\leq \|NM^+\|\|y\|,\end{aligned}$$

and $\|\hat{x}\| = \|M^+y\| \leq \|M^+\|\|y\|$. Thus (1.1) holds. Now we apply Lemma 1.1 to the equation (1.11) to conclude that for every $y \in R(M)$, the equation $Ax = y$ has a solution $x^* \in N(M)^c$ satisfying (1.12). This implies that

(1.15) $$R(M) \subset R(A) .$$

However,

(1.16) $R(A) \subset R(M)$, by (1.9) and (1.10).

Now (1.15) and (1.16) imply

(1.17) $$R(A) = R(M) .$$

Since $N(M)^c$ is isomorphic to $R(M)$ via M, it is also isomorphic via A, by (1.17) and the assumption (1.13). Thus, one can choose $N(A)^c = N(M)^c$. This determines A^+ with respect to the topological complements $N(M)^c$ and $R(M)$. Now, since $x^* \in N(A)^c = N(M)^c$, we have $x^* = A^+y$, and (1.14) follows from (1.12). □

Corollary 1.1. Let $H \in \mathcal{L}_b(X)$ with $\|H\| < 1$ and $P \in \mathcal{L}_b(X)$ such that $P^2 = P$. If $R(H) \subset R(P)$, then the equation $(P + H)x = y$ has a solution $x^* \in R(P)$ for each $y \in R(P)$. Also $\|x^*\| \leq \frac{1}{1 - \|H\|}\|y\|$, $R(P + H) = R(P)$ and $\|(P + H)^+\| \leq \frac{1}{1 - \|H\|}$, where $(P + H)^+$ denotes the generalized inverse with respect to $P_{N(P+H)^c} = P$ and $P_{R(P+H)} = P$.

Proof. We will show that the hypotheses of Proposition 1.1 are satisfied with $M = P$ and $N = H$. Clearly, $P \in \ell_b(X)$ has topological complements $N(P)^c = R(P)$ and $R(P)^c = N(P)$. So P^+ $(=P)$ is the generalized inverse of P with respect to these complements. Also $P + H \in \ell_b(X)$, since $\|H\| < 1$. Take an arbitrary $y \in R(P)$. Then $\hat{x} \triangleq P^+ y$ (see the proof of Proposition 1.1, where $M = P$) is equal to y, i.e. $\hat{x} = y$, since $P = P^+$. Therefore, the assumption $\|HP^+\| < 1$, in Proposition 1.1, can be replaced by $\|H\| < 1$. Also, the assumption (1.13), which reads here

$$N(P)^c \cap N(P + H) = \{0\} \tag{1.18}$$

is satisfied. If (1.18) were not true, there would exist an $x \neq 0$ such that both

$$x \in N(P)^c = R(P) \quad \text{and} \quad x \in N(P + H).$$

Hence $(P + H)x = x + Hx = 0$, which contradicts the assumption $\|H\| < 1$. □

Corollary 1.2. (Ben-Israel [10].) Let H be an $n \times n$ real matrix, $\|H\| < 1$ and L be a subspace of R^m such that $R(H) \subset L$. Then

$$\|(P_L + H)^+\| \leq \frac{1}{1 - \|H\|}.$$

Proof. Specify $P = P_L$ in Corollary 1.1. □

The following classical result of Banach, e.g. [31], is also obtained.

Corollary 1.3. Let $H \in \ell_b(X)$. If $\|H\| < 1$, then $(I + H)^{-1}$ exists and

$$\|(I + H)^{-1}\| \leq \frac{1}{1 - \|H\|}.$$

Proof. Specify $P = I$ in Corollary 1.1. □

Corollary 1.4. (Kantorovich and Akilov [31, p.172.]) Let $M \in \ell_b(X)$ and suppose that $M^{-1} \in \ell_b(X)$ exists. If $N \in \ell_b(X)$ and $\|NM^{-1}\| < 1$, then $(M + N)^{-1}$ exists and

$$\|(M + N)^{-1}\| \leq \frac{\|M^{-1}\|}{1 - \|NM^{-1}\|}.$$

Proof. Apply Proposition 1.1 to the case when $X = Y$ and M^{-1} exists. □

2. Solvability of the Approximate Equation

Let $\tilde{X}$ and $\tilde{Y}$ be closed subspaces of the Banach spaces X and Y, respectively. Further, let $\tilde{X}$ and $\tilde{Y}$ be isomorphic via mappings J_0 and H_0 to the Banach spaces $\overline{X}$ and $\overline{Y}$, respectively. Suppose also that J and H are linear extensions of J_0 and H_0 to all of X and Y, respectively. Such extensions always exist, for we may take $J = J_0 P_{\tilde{X}}$ and $H = H_0 P_{\tilde{Y}}$. In many practical situations, $\overline{X}$ and $\overline{Y}$ are chosen to be finite.dimensional.

Consider again the equations

$$Ax = b, \tag{I.1.1}$$

where $A: X \to Y$, $b \in Y$, and

$$\overline{A}\overline{x} = \overline{b}, \tag{I.1.2}$$

where $\overline{A}: \overline{X} \to \overline{Y}$, $\overline{b} \in \overline{Y}$. We shall refer to (I.1.1) as the "exact" equation and to (I.1.2) as its "approximate" equation. We assume that $A \in \ell_b(X,Y)$, $\overline{A} \in \ell_b(\overline{X},\overline{Y})$ and that the following decompositions are possible:

$$X = N(A) \oplus N(A)^c, \quad \overline{X} = N(\overline{A}) \oplus N(\overline{A})^c$$

$$Y = R(A) \oplus R(A)^c, \quad \text{and} \quad \overline{Y} = R(\overline{A}) \oplus R(\overline{A})^c .$$

The symbol $\oplus$ is here used to indicate that all eight complements are necessarily closed. Denote by $A^+ \in \ell_b(Y,X)$ the generalized inverse of A relative to the continuous projections $P_{N(A)}$ and $P_{R(A)}$, and by $\overline{A}^+ \in \ell_b(\overline{Y},\overline{X})$ the generalized inverse of $\overline{A}$ relative

to the continuous projections $P_{N(\bar{A})}$ and $P_{R(\bar{A})}$. Let us denote by $x^* = A^+b$ and $\bar{x}^* = \bar{A}^+\bar{b}$, the best approximate solutions of the equations (I.1.1) and (I.1.2) respectively.

In the sequel we will state results relating the best approximate solutions of the exact and approximate equations when some or all of the following conditions are satisfied:

(I) The operator A is represented as $A = M+N$, where M is bounded and $X = N(M) \oplus N(M)^c$, $N(A)^c \subset N(M)^c$ and $Y = R(M) \oplus R(M)^c$. (In this situation M^+ denotes the generalized inverse of M, relative to the continuous projections $P_{N(M)^c}$ and $P_{R(M)}$.)

(II) J_0 maps $N(M)^c \cap \tilde{X}$ into $N(\bar{A})^c \cap \bar{X}$.

(III) H has the property $\bar{A}^+\bar{b} = \bar{A}^+HP_{R(A)}b$.

(IV) $\|AJ_0\tilde{x} - HA\tilde{x}\| \leq \varepsilon\|\tilde{x}\|$ for some constant $\varepsilon \geq 0$ and all $\tilde{x} \in N(M)^c \cap \tilde{X}$.

(V) For every $x \in N(A)^c$ there is a $u \in N(M)^c \cap \tilde{X}$ such that $\|Mu - P_{R(M)}Nx\| \leq \eta_1\|x\|$ for some constant $\eta_1 \geq 0$.

(VI) There exists a vector $v \in N(M)^c \cap \tilde{X}$ such that $\|Mv - P_{R(M)}P_{R(A)}b\| \leq \eta_2\|P_{R(A)}b\|$ for some constant $\eta_2 \geq 0$.

Theorem 2.1. (*Conditions for the solvability of the approximate equation.*) Let the conditions (I), (IV) and (V) be satisfied. In addition, suppose that

(A) $M^+ : \tilde{Y} \to \tilde{X}$,

(B) $H_0^{-1} H R(A) \subset R(A)$,

(C) $J(N(M)^c)$ is closed and

(D) $R(\bar{A}) \subset HR(A)$

If

$$q = [\varepsilon(1 + \eta_1 \|M^+\|) + \eta_1 \|HA\| \|M^+\|] \|A^+ H_0^{-1}\| < 1 , \tag{2.1}$$

then the equation

$$\bar{A}\bar{x} = HP_{R(A)}b \tag{2.2}$$

has a solution $\bar{x}^* \in J(N(M)^c)$ for every $b \in Y$. Furthermore

$$\|\bar{x}^*\| \leq \frac{\alpha}{1 - q} \|HP_{R(A)}b\| , \tag{2.3}$$

where

$$\alpha = \|J_0\| (1 + \eta_1 \|M^+\|) \|A^+ H_0^{-1}\| . \tag{2.4}$$

Proof. It is sufficient to show that the conditions (1.1) of Lemma 1.1 are satisfied for the equation (2.2) with $E = J(N(M)^c)$, $F = H R(A)$, $V = \bar{A}$, $y = HP_{R(A)}b$ and α and q as in (2.1) and (2.4). First, consider the equation $Ax = H_0^{-1} HP_{R(A)}b$ and its solution $x_0 = A^+ H_0^{-1} HP_{R(A)}b$. Denote $z = Mx_0 - H_0^{-1} HP_{R(A)}b$. Since $x_0 \in N(A)^c$ and $N(A)^c \subset N(M)^c$, by condition (I), we find that

(2.5) $$x_0 = M^+ z + M^+ H_0^{-1} H P_{R(A)} b$$

and also

(2.6) $$Ax_0 = H_0^{-1} H P_{R(A)} b ,$$

by definition of x_0 and condition (B). Therefore $z = Mx_0 - Ax_0 = -Nx_0$, since $A = M + N$.

Now, for $x = -x_0$, condition (V) implies that there exists $u \in N(M)^c \cap \tilde{X}$ such that

(2.7) $$\| Mu - P_{R(M)} N(-x_0) \| = \| Mu - P_{R(M)} z \| \leq \eta_1 \| x_0 \| ,$$

for some $\eta_1 \geq 0$. Denote $\tilde{x} = u + M^+ H_0^{-1} H P_{R(A)} b$. Note that $\tilde{x} \in N(M)^c \cap \tilde{X}$, by conditions (V) and (A). We will now show that $J_0 \tilde{x}$ is the required element $\hat{x}$ of E in Lemma 1.1. First,

$$\begin{aligned} \| \bar{A} J_0 \tilde{x} - H P_{R(A)} b \| &= \| \bar{A} J_0 \tilde{x} - H H_0^{-1} H P_{R(A)} b \| \\ &= \| \bar{A} J_0 \tilde{x} - H A x_0 \| , \quad \text{by condition (B) and definition of } x_0 \\ &\leq \| \bar{A} J_0 \tilde{x} - H A \tilde{x} \| + \| H A \tilde{x} - H A x_0 \| , \quad \text{by the triangular inequality} \\ &\leq \varepsilon \| \tilde{x} \| + \| H A \| \| \tilde{x} - x_0 \| , \quad \text{by condition (IV).} \end{aligned}$$

Since

$$\| \tilde{x} - x_0 \| = \| u + M^+ H_0^{-1} H P_{R(A)} b - M^+ z - M^+ H_0^{-1} H P_{R(A)} b \| ,$$

by definition of $\tilde{x}$ and x_0

$$= \|u - M^+ z\| = \|M^+ Mu - M^+ P_{R(M)} z\|$$

$$\leq \eta_1 \|M^+\| \|x_0\| , \quad \text{by (2.7)} \tag{2.8}$$

$$\leq \eta_1 \|M^+\| \|A^+ H_0^{-1}\| \|HP_{R(A)} b\| , \quad \text{by definition of } x_0$$

and

$$\|\tilde{x}\| \leq \|x_0\| + \|\tilde{x} - x_0\|$$

$$\leq (1 + \eta_1 \|M^+\|) \|x_0\| , \quad \text{by (2.8)}$$

$$\leq (1 + \eta_1 \|M^+\|) \|A^+ H_0^{-1}\| \|HP_{R(A)} b\| ,$$

the above inequality gives

$$\|\bar{A} J_0 \tilde{x} - HP_{R(A)} b\| \leq [\varepsilon(1 + \eta_1 \|M^+\|) \|A^+ H_0^{-1}\|$$

$$+ \eta_1 \|HA\| \|M^+\| \|A^+ H_0^{-1}\|] \|HP_{R(A)} b\|$$

$$= q \|HP_{R(A)} b\| , \tag{2.9}$$

where q is the constant defined by (2.1). Also

$$\|J_0 \tilde{x}\| \leq \|J_0\| \|\tilde{x}\| \leq \alpha \|HP_{R(A)} b\| , \tag{2.10}$$

where α is defined by (2.4). The inequalities (2.9) and (2.10) correspond to the assumptions (1.1) of Lemma 1.1 with $V = \bar{A}$, $\hat{x} = J_0 \tilde{x}$ and $y = HP_{R(A)} b$. Conditions (C) and (D) guarantee that $E = J(N(M)^c)$ be a closed subspace of X and $F = H\mathcal{R}(A)$ be a subspace of Y containing $V(E) = \bar{A}(J(N(M)^c))$. All conditions

of Lemma 1.1 are now satisfied, and the conclusions of Theorem 2.1 follow.

□

Corollary 2.1. Let $A \in \mathcal{L}_b(X,Y)$ and suppose that the conditions (I), (IV), (V), (A), (B), (C) and (D) are satisfied. If $q < 1$ and $\overline{A}$ satisfies the condition

$$\text{(E)} \qquad J(N(M)^c) \subset N(\overline{A})^c$$

then

$$\|\overline{A}^+\| \leq \frac{\alpha}{1-q}\|P_{R(\overline{A})}\| ,$$

where q and α are as in Theorem 2.1.

Proof. We need to show that

$$\|\overline{A}^+ \bar{y}\| \leq \frac{\alpha}{1-q}\|P_{R(\overline{A})}\|\|\bar{y}\| , \quad \text{for every } \bar{y} \in \overline{Y} .$$

Let $\bar{x}^*$ denote the solution in $J(N(M)^c)$ of the equation $\overline{A}\bar{x} = P_{R(\overline{A})}\bar{y}$. Such an $\bar{x}^*$ exists, by Theorem 1.1. In fact, by condition (E), $\bar{x}^* = \overline{A}^+\bar{y}$, i.e. $\bar{x}^*$ is a unique best approximate solution in $N(\overline{A})^c$ of the equation $\overline{A}\bar{x} = \bar{y}$. But $P_{R(\overline{A})}\bar{y} = HP_{R(A)}b$ for some $b \in Y$, by condition (D). So $\bar{x}^*$ is also a solution of the equation $\overline{A}\bar{x} = HP_{R(A)}b$ and it satisfies

$$\|\bar{x}^*\| \leq \frac{\alpha}{1-q}\|HP_{R(A)}b\| , \quad \text{by (2.3)}.$$

Therefore,

$$\|\overline{A}^{+}\bar{y}\| \le \frac{\alpha}{1-q}\|HP_{\mathcal{R}(A)}b\| = \frac{\alpha}{1-q}\|P_{\mathcal{R}(\overline{A})}\bar{y}\| \le \frac{\alpha}{1-q}\|P_{\mathcal{R}(\overline{A})}\|\|\bar{y}\| .$$

□

If $X = Y$, $\tilde{X} = \tilde{Y} = \overline{X} = \overline{Y}$, $M = I$ and A^{-1} exists then the conditions (A), (B), (C), (D), (I) and (II) are trivially satisfied, the requirement on H in (III) becomes $\overline{A}^{+}\overline{b} = \overline{A}^{+}Hb$ while (IV), (V) and (VI) reduce to:

(IV') $\|\overline{A}\tilde{x} - HA\tilde{x}\| \le \varepsilon\|\tilde{x}\|$ for some constant $\varepsilon \ge 0$ and all $\tilde{x} \in \tilde{X}$.

(V') For every $x \in X$ there is a $u \in \tilde{X}$ such that $\|u - Nx\| \le \eta_1\|x\|$ for some constant $\eta_1 \ge 0$.

(VI') There exists a vector $v \in \tilde{X}$ such that $\|v - b\| \le \eta_2\|b\|$ for some constant $\eta_2 \ge 0$.

In the nonsingular case, Theorem 1.1 reduces to the following result of Kantorovich and Akilov [31, p.545].

Corollary 2.2. Let $A \in \mathcal{L}_b(X)$ have an inverse and let the conditions (IV') and (V') be satisfied. If

$$q = [\varepsilon(1+\eta_1) + \eta_1\|HA\|]\|A^{-1}\| < 1 ,$$

then the equation $\overline{A}\bar{x} = \overline{b}$ has a solution $\bar{x}^* \in \overline{X}$ for every $\overline{b} \in \overline{X}$. Also

$$\|\bar{x}^*\| \leq \frac{\alpha}{1-q}\|\bar{b}\| ,$$

where $\alpha = (1+\eta_1)\|A^{-1}\|$.

An estimate for the norm of the generalized inverse $\bar{A}^+$ was obtained using Theorem 1.1. In the nonsingular case, Corollary 2.1 gives the following result of Kantorovich and Akilov [31, p.546].

Corollary 2.3. Let the hypotheses and the notation of Corollary 2.2 hold and let $\bar{A}$ satisfy the condition:

(E') "The existence of a solution of the equation $\bar{A}\bar{x} = \bar{b}$ for every $\bar{b} \in \bar{X}$ implies its uniqueness".

Then

$$\|\bar{A}^{-1}\| \leq \frac{\alpha}{1-q} .$$

Proof. Condition (E') and Corollary 2.2 imply the existence of $\bar{A}^{-1}$. The result now follows from Corollary 2.1 since the conditions (D) and (E) are satisfied when $\bar{A}^{-1}$ and A^{-1} exist. □

3. Error Estimates

The following theorem estimates the distance between the best approximate solution $\bar{x}^*$ of the approximate equation and the best approximate solution x^* of the exact equation.

Theorem 3.1. (*Estimate of the distance between best approximate solutions.*) Consider the equation $Ax = b$ and its approximate equation $\overline{A}\bar{x} = \overline{b}$. If the conditions (I) to (VI) are satisfied, then

$$\|x^* - J_0^{-1}\bar{x}^*\| \le p\|x^*\| , \tag{3.1}$$

where

$$p = \varepsilon(1 + c)\|J_0^{-1}\|\|\overline{A}^+\| + c(1 + \|J_0^{-1}\overline{A}^+HA\|) \tag{3.2}$$

and

$$c = \min\{1 , (\eta_1 + \eta_2\|A\|)\|M^+\|\} . \tag{3.3}$$

Proof. First we show that x^* can be approximated by an $\tilde{x} \in N(M)^c \cap \tilde{X}$ to the order of $\eta_1 + \eta_2$. We know, by conditions (V) and (VI), that there exist u and v in $N(M)^c \cap \tilde{X}$ such that

$$\|Mu - P_{R(M)}Nx^*\| \le \eta_1\|x^*\| \tag{3.4}$$

and

$$\|Mv - P_{R(M)}P_{R(A)}b\| \le \eta_2\|P_{R(A)}b\| . \tag{3.5}$$

Denote $\tilde{x} = M^+(Mv - Mu)$. Clearly $\tilde{x} \in N(M)^c \cap \tilde{X}$. We now show that x^* can be approximated by $\tilde{x}$ to the order of $\eta_1 + \eta_2$.

$$\|x^* - \tilde{x}\| = \|M^+Mx^* - \tilde{x}\|, \quad \text{by condition (I)}$$

$$= \|-M^+Nx^* + M^+(M+N)x^* - M^+(Mv - Mu)\|, \quad \text{by definition of } \tilde{x}$$

$$= \|-M^+Nx^* + M^+P_{R(A)}b - M^+(Mv - Mu)\|, \quad \text{since } A = M+N \text{ and } Ax^* = P_{R(A)}b$$

$$\le \|M^+\|(\|Mu - P_{R(M)}Nx^*\| + \|Mv - P_{R(M)}P_{R(A)}b\|), \quad \text{since } M^+ = M^+P_{R(M)}$$

$$\le \|M^+\|(\eta_1\|x^*\| + \eta_2\|P_{R(A)}b\|), \quad \text{by (3.4) and (3.5)}$$

$$\le \|M^+\|(\eta_1 + \eta_2\|A\|)\|x^*\|, \quad \text{since } \|P_{R(A)}b\| = \|Ax^*\| \le \|A\|\|x^*\|.$$

Hence we conclude that there exists an $\tilde{x} \in N(M)^c \cap \tilde{X}$ such that

$$\|x^* - \tilde{x}\| \le c\|x^*\|, \tag{3.6}$$

where $c = \min\{1, (\eta_1 + \eta_2\|A\|)\|M^+\|\}$. (Note that $c \le 1$, since we can always choose $\tilde{x} = 0$ in (3.6).)

Let us now prove (3.1). Denote $\bar{x}_0 = \bar{A}^+HA\tilde{x}$. Then

$$\|x^* - J_0^{-1}x^*\| \le \|x^* - \tilde{x}\| + \|\tilde{x} - J_0^{-1}\bar{x}_0\| + \|J_0^{-1}\bar{x}_0 - J_0^{-1}\bar{x}^*\|. \tag{3.7}$$

The first term on the right hand side $\|x^* - \tilde{x}\|$ is estimated by (3.6). The two remaining terms will now be estimated. First

$$\tilde{x} - J_0^{-1}\bar{x}_0 = \tilde{x} - J_0^{-1}\bar{A}^+HA\tilde{x}, \quad \text{by definition of } \bar{x}_0$$

$$= J_0^{-1}(J_0 - \bar{A}^+HA)\tilde{x}, \quad \text{since } \tilde{x} \in \tilde{X}$$

$$= J_0^{-1}\overline{A}^{+}(\overline{A}J_0 - HA)\tilde{x}\,, \quad \text{since } \overline{A}^{+}\overline{A}J_0\tilde{x} = P_{N(\overline{A})^c}J_0\tilde{x}$$
$$= J_0\tilde{x}\,, \quad \text{by condition (II).}$$

Hence

$$\|\tilde{x} - J_0^{-1}\bar{x}_0\| \le \|J_0^{-1}\overline{A}^{+}\|\,\|(\overline{A}J_0 - HA)\tilde{x}\|$$
$$\le \varepsilon\|J_0^{-1}\overline{A}^{+}\|\,\|\tilde{x}\|\,, \quad \text{by condition (IV)}$$
$$\le \varepsilon\|J_0^{-1}\overline{A}^{+}\|\,(\|x^*\| + \|x^* - \tilde{x}\|)\,, \quad \text{by the triangle inequality}$$
$$\le \varepsilon(1+c)\|J_0^{-1}\overline{A}^{+}\|\,\|x^*\|\,, \quad \text{by (3.6).}$$

The third term is estimated as follows:

$$\|J_0^{-1}\bar{x}_0 - J_0^{-1}\bar{x}^*\| = \|J_0^{-1}(\overline{A}^{+}HA\tilde{x} - \overline{A}^{+}\overline{A}\bar{x}^*)\|\,, \quad \text{by definition of } \bar{x}_0$$
$$= \|J_0^{-1}(\overline{A}^{+}HA\tilde{x} - \overline{A}^{+}P_{R(\overline{A})}\overline{b})\|$$
$$= \|J_0^{-1}(\overline{A}^{+}HA\tilde{x} - \overline{A}^{+}HP_{R(A)}b)\|\,, \quad \text{by condition (III)}$$
$$\le \|J_0^{-1}\overline{A}^{+}HA\|\,\|\tilde{x} - x^*\|\,, \quad \text{since } P_{R(A)}b = Ax^*$$
$$\le c\|J_0^{-1}\overline{A}^{+}HA\|\,\|x^*\|\,, \quad \text{by (3.6).}$$

After substituting the above estimates into (3.7) the conclusion follows. □

Remark 3.1. It may happen that we could approximate x^* by some $\tilde{x} \in N(M)^c \cap \tilde{X}$ directly. Then we no longer need the conditions (V) and (VI) and we can set $c = \min\{1\,, \|x^* - \tilde{x}\|\}$ in (3.2).

Remark 3.2. If $p < 1$, we can write the estimate (3.1) as follows

$$\|x^* - J_0^{-1}\bar{x}^*\| \leq \frac{p}{1-p}\|J_0^{-1}\bar{x}^*\| .$$

This is true, since

$$\|x^* - J_0^{-1}\bar{x}^*\| \leq p\|x^*\| \leq p(\|J_0^{-1}\bar{x}^*\| + \|x^* - J_0^{-1}\bar{x}^*\|) .$$

Remark 3.3. If X and Y are Hilbert spaces and $N(A)^c = R(A^*)$, $R(A)^c = N(A^*)$, then Theorem 3.1 reduces to a result obtained by Zlobec in [65]. However, the Hilbert space version of Theorem 3.1 is proved there under slightly different assumptions.

Remark 3.4. It may happen that one cannot satisfy condition (III) but that a constant η_3, such that

$$\|\overline{A^+b} - \overline{A^+}HP_{R(A)}b\| \leq \eta_3\|x^*\| ,$$

is found. In this case the constant p in (3.1) is different. Now

$$\|J_0^{-1}\bar{x}_0 - J_0^{-1}\bar{x}^*\| \leq (c\|J_0^{-1}\overline{A^+}HA\| + \eta_3\|J_0^{-1}\|)\|x^*\|$$

and hence

$$P = [\varepsilon(1+c)\|\overline{A^+}\| + \eta_3]\|J_0^{-1}\| + c(1 + \|J_0^{-1}\overline{A^+}HA\|) .$$

In the nonsingular case, we get the following result from [31, p.547]. (Recall that in our setting $x^* \in N(A)^c$. Also, as

in [31], we specify $\bar{b} = Hb$.)

Corollary 3.1. Let the conditions (IV'), (V') and (VI') be satisfied and let A^{-1} and $\bar{A}^{-1}$ exist. Then

$$\|x^* - \tilde{x}^*\| \le p\|x^*\| ,$$

where x^* is the solution of the exact equation (I.1.1), $\tilde{x}^*$ is the solution of the approximate equation (I.1.2) and

$$p = 2\varepsilon\|\bar{A}^{-1}\| + (\eta_1 + \eta_2\|A\|)(1 + \|\bar{A}^{-1}HA\|) .$$

Proof. Specify $M = I$, $X = Y$ and $\tilde{X} = \tilde{Y} = \bar{X} = \bar{Y}$ in Theorem 3.1. □

4. Convergence Criteria

Our next result gives conditions for convergence of approximation schemes. Suppose that the exact equation $Ax = b$ is approximated by a sequence of equations $\bar{A}_n\tilde{x} = \bar{b}_n$, $n = 1,2,\ldots$, rather than by a single equation. This determines a sequence of the spaces $\tilde{X}_n$, $\tilde{Y}_n$, $\bar{X}_n$, $\bar{Y}_n$, the operators $\bar{A}_n$, $(J_0)_n$, $(H_0)_n$, J_n, H_n and the constants ε_n, $(\eta_1)_n$, $(\eta_2)_n$, c_n, q_n, α_n, p_n, $n = 1,2,\ldots$. For the sake of notational simplicity these indices will generally

be omitted in the sequel. The following theorem gives conditions for the convergence of the sequence $\bar{x}_n^*$, the best approximate solution of $\bar{A}_n\bar{x}_n = \bar{b}_n$, $n = 1,2,\ldots$, to x^*, the best approximate solution of the exact equation.

Theorem 4.1. (*Convergence of the best approximate solutions.*) Consider the equation $Ax = b$ and a sequence of approximate equations $\bar{A}\bar{x} = \bar{b}$. Suppose that for each $n = 1,2,\ldots$ the conditions:

(i) (I) to (VI) and (A), (B), (C), (D) and (E)

(ii) $\sup_n \|J_0\| < \infty$, $\sup_n \|H_0^{-1}\| < \infty$ and $\sup_n \|P_{R(\bar{A})}\| < \infty$

(iii) $\lim_{n\to\infty} \varepsilon\|J_0^{-1}\| = 0$, $\lim_{n\to\infty} \eta_1\|J_0^{-1}\|\|H\| = 0$, $\lim_{n\to\infty} \eta_2\|J_0^{-1}\|\|H\| = 0$

are satisfied. Then $\lim_{n\to\infty} \eta_1 = 0$, $\lim_{n\to\infty} \eta_2 = 0$ and the sequence of best approximate solutions of $\bar{A}\bar{x} = \bar{b}$ converges to the best approximate solution x^* of $Ax = b$, i.e.

$$\lim_{n\to\infty} \|x^* - J_0^{-1}\bar{x}^*\| = 0 .$$

More precisely,

$$\|x^* - J_0^{-1}\bar{x}^*\| \le \varepsilon c_1\|J_0^{-1}\| + \eta_1(c_2 + c_3\|J_0^{-1}\|\|H\|) + \eta_2(c_4 + c_5\|J_0^{-1}\|\|H\|)\|x^*\| ,$$

where c_1 to c_5 are some constants independent of the index n.

Proof. Since $H_0 H_0^{-1} = I$, it follows that $\|H_0^{-1}\|\|H_0\| \geq 1$, $n = 1,2,\ldots$. Hence

$$\inf_n \|H_0\| > 0 \tag{4.1}$$

using the second assumption in (ii). Similarly, one concludes that

$$\inf_n \|J_0^{-1}\| > 0 \tag{4.2}$$

using the first assumption in (ii). Also $\|H\| \geq \|H_0\|$, since $H_0 = H|_{\tilde{Y}}$. Therefore, by (4.1),

$$\inf_n \|H\| > 0 . \tag{4.3}$$

From (4.2), (4.3) and condition (iii), we conclude that, in particular,

$$\lim_{n\to\infty} \varepsilon = 0 , \quad \lim_{n\to\infty} \eta_1 = 0 , \quad \lim_{n\to\infty} \eta_2 = 0 \quad \text{and} \quad \lim_{n\to\infty} \eta_1\|H\| = 0 . \tag{4.4}$$

Recall the constant q introduced in Theorem 2.1:

$$q = [\varepsilon(1 + \eta_1\|M^+\|) + \eta_1\|HA\|\|M^+\|]\|A^+H_0^{-1}\| .$$

For sufficiently large n, using (4.4), one has $q < \frac{1}{2}$ and for such values of n Theorem 2.1 is applicable. But we can also apply Corollary 2.1 to obtain

$$\|\overline{A}^+\| \leq \frac{\alpha}{1 - q}\|P_{R(A)}\| \leq 2\alpha\|P_{R(\overline{A})}\| .$$

Since $\alpha = \|J_0\|(1 + \eta_1\|M^+\|)\|A^+H_0^{-1}\|$, one concludes, using the third assumption in (ii), that $\|\bar{A}^+\|$ is bounded independently of the index n, i.e.

$$\sup_n \|\bar{A}^+\| = s < \infty. \tag{4.5}$$

The desired estimate now follows for sufficiently large n:

$$\|x^* - J_0^{-1}\bar{x}^*\| \le [2\varepsilon\|J_0^{-1}\|\|\bar{A}^+\| + (\eta_1 + \eta_2\|A\|)(1 + \|J_0^{-1}\bar{A}^+HA\|)\|M^+\|]\|x^*\|,$$

by (3.2) and (3.3)

$$\le [\varepsilon c_1\|J_0^{-1}\| + \eta_1(c_2 + c_3\|J_0^{-1}\|\|H\|) + \eta_2(c_4 + c_5\|J_0^{-1}\|\|H\|)]\|x^*\|,$$

where $c_1 = 2s$, $c_2 = \|M^+\|$, $c_3 = s\|A\|\|M^+\|$, $c_4 = \|A\|\|M^+\|$, $c_5 = s\|A\|^2\|M^+\|$ and s is defined by (4.5). The right-hand side in the above inequality tends to zero when $n \to \infty$, by (iii) and (4.4)

□

A corresponding result in the nonsingular case is given in [31, p.549] as follows.

Corollary 4.1. Consider the equation $Ax = b$ and a sequence of approximate equations $\bar{A}\bar{x} = \bar{b}$. Assume that A^{-1} exists and that $\bar{A}$ satisfies condition (E') for each $n = 1,2,\ldots$. Assume further that for each $n = 1,2,\ldots$ the conditions (IV'), (V'), (VI') are satisfied, and that

$$\lim_{n\to\infty} \varepsilon = 0 , \quad \lim_{n\to\infty} \eta_1\|H\| = 0 \quad \text{and} \quad \lim_{n\to\infty} \eta_2\|H\| = 0 .$$

Then the approximate equations are consistent for sufficiently large n and the sequence of approximate solutions converges to the exact solution x^* of $Ax = b$, i.e.

$$\lim_{n\to\infty} \|x^* - \bar{x}^*\| = 0 .$$

More precisely,

$$\|x^* - \bar{x}^*\| \le \varepsilon c_1 + \eta_1(c_2 + c_3\|H\|) + \eta_2(c_4 + c_5\|H\|) ,$$

where c_1 to c_5 are some constants indpendent of the index n.

Proof. Set $X = Y$, $\overline{X} = \overline{Y} = \widetilde{X} = \widetilde{Y}$ and $M = I$ in Theorem 2.1. Then $J_0 = H_0 = I$ and $\|H\| = \|J\| \ge 1$. Furthermore, from Corollary 1.3, we see that $R(\overline{A}) = \overline{X}$. So condition (E) is satisfied and $\|P_{R(\overline{A})}\| = 1$. Conditions (i) to (iii) of Theorem 4.1 hold and the result follows. □

Remark 4.1. Constants ε, η_1, η_2 and c_1 to c_5 in Theorem 4.1 reduce, in the nonsingular case, to the corresponding constants in Corollary 4.1.

Let us recall that Corollary 2.1 gave us an estimate for $\|\overline{A}^+\|$ in terms of $\|A^+\|$ via constants α and q. Our last result gives us the reverse estimate.

Theorem 4.2. (*Estimate for the norm of the generalized inverse.*) Let $A \in \ell_b(X,Y)$ and $\overline{A} \in \ell_b(\overline{X},\overline{Y})$. If the conditions (I) to (V) are satisfied and

$$r = [\varepsilon(1 + \eta_1\|M^+\|)\|J_0^{-1}\|\|\overline{A}^+\| + \eta_1\|M^+\|(1 + \|J_0^{-1}\overline{A}^+HA\|)] < 1$$

then

$$\|A^+\| \le (1 - r)^{-1}[\|J_0^{-1}\overline{A}^+H\| + \|M^+\|(1 + \varepsilon\|J_0^{-1}\|\|\overline{A}^+\| + \|J_0^{-1}\overline{A}^+HA\|)] .$$

Proof. Take $\hat{x} \in X$, $\hat{x} \notin N(A)$. Then $x^* = P_{N(A)^c}\hat{x}$ is clearly the best approximate solution of the equation

$$Ax = A\hat{x} . \tag{4.6}$$

By condition (V), there exists a $u \in N(M)^c \cap \tilde{X}$ such that

$$\|Mu - P_{R(M)}N(-P_{N(A)^c}\hat{x})\| \le \eta_1\|P_{N(A)^c}\hat{x}\| . \tag{4.7}$$

Now

$$\|x^* - u\| = \|P_{N(A)^c}\hat{x} - u\| = \|M^+MP_{N(A)^c}\hat{x} - M^+Mu\| , \quad \text{by condition (I)}$$

$$\le (\|MP_{N(A)^c}\hat{x} + NP_{N(A)^c}\hat{x}\| + \|P_{R(M)}NP_{N(A)^c}\hat{x} + Mu\|)\|M^+\|$$

$$\le \left(\eta_1 + \frac{\|Ax^*\|}{\|P_{N(A)^c}\hat{x}\|}\right)\|P_{N(A)^c}\hat{x}\|\|M^+\| , \quad \text{by (4.7)} .$$

Apply Remark 3.1 to the exact equation (4.6) and its approximate equation $\overline{A}\overline{x} = HA\hat{x}$, with

$$c = \min\{1 , \|M^+\| \left(\eta_1 + \frac{\|A\hat{x}\|}{\|P_{(A)^c}\hat{x}\|}\right)\} \quad \text{in (3.2)} .$$

Then

$$\|x^* - J_0^{-1}\bar{x}^*\| = \|P_{N(A)^c}\hat{x} - J_0^{-1}\bar{A}^+HA\hat{x}\|$$

$$\leq \left\{\varepsilon\left[1 + \|M^+\|\left(\eta_1 + \frac{\|A\hat{x}\|}{\|P_{N(A)^c}\hat{x}\|}\right)\right]\|J_0^{-1}\|\|A^+\| \right.$$

$$\left. + \|M^+\|\left(\eta_1 + \frac{\|A\hat{x}\|}{\|P_{N(A)^c}\hat{x}\|}\right)(1 + \|J_0^{-1}\bar{A}^+HA\|)\right\}\|P_{N(A)^c}\hat{x}\| ,$$

by (3.1) and (3.2)

$$\leq [\varepsilon(1 + \eta_1\|M^+\|)\|J_0^{-1}\|\|\bar{A}^+\| + \eta_1\|M^+\|(1 + \|J_0^{-1}\bar{A}^+HA\|)]\|P_{N(A)^c}\hat{x}\|$$

$$+ \|M^+\|(1 + \varepsilon\|J_0^{-1}\|\|\bar{A}^+\| + \|J_0^{-1}\bar{A}^+HA\|)\|A\hat{x}\| .$$

Now

$$\|P_{N(A)^c}\hat{x}\| \leq \|P_{N(A)^c}\hat{x} - J_0^{-1}\bar{A}^+HA\hat{x}\| + \|J_0^{-1}\bar{A}^+H\|\|A\hat{x}\| ,$$

by the triangle inequality

$$\leq r\|P_{N(A)^c}\hat{x}\| + [\|J_0^{-1}\bar{A}^+H\| + \|M^+\|(1 + \varepsilon\|J_0^{-1}\|\|\bar{A}^+\| + \|J_0^{-1}\bar{A}^+HA\|)]\|A\hat{x}\| ,$$

by (4.8) and the definition of r.

Hence

$$\|A\hat{x}\| \geq \frac{1 - r}{\|J_0^{-1}\bar{A}^+H\| + \|M^+\|(1 + \varepsilon\|J_0^{-1}\|\|\bar{A}^+\| + \|J_0^{-1}\bar{A}^+HA\|)}\|P_{N(A)^c}\hat{x}\|$$

$$= t\|P_{N(A)^c}\hat{x}\| ,$$

where t denotes the coefficient of $\|P_{N(A)^c}\hat{x}\|$. Take an arbitrary $0 \neq y \in R(A)$. Then there exists an $\hat{x} \in X$, $\hat{x} \notin N(A)$ such that $y = A\hat{x}$. Hence $A^+y = P_{N(A)^c}\hat{x}$. Furthermore, by the above inequality

$$\|A^+y\| = \|P_{N(A)^c}\hat{x}\| \leq \frac{1}{t}\|A\hat{x}\| = \frac{1}{t}\|y\|,$$

which gives the desired estimate for A^+. Note that here $t > 0$, since $r < 1$. If $y = 0$, the above inequality is trivially satisfied. □

Remark 4.2. If $A \in \ell_b(X)$ and in addition $X = Y$, $\overline{X} = \overline{Y} = \tilde{X} = \tilde{Y}$, then the above result reduces to the bound for a left inverse of A given in [31, p.550].

V. GALERKIN'S METHOD FOR BEST APPROXIMATE SOLUTIONS

1. Description of the Method

In this section we will use Kantorovich's theory to prove that a Galerkin type method, when applied to a certain kind of, possibly inconsistent, operator equation, produces the best approximate solution. This solution is obtained as the limit of a sequence of best approximate solutions of, possibly inconsistent, systems of linear algebraic equations. In the case of Hilbert space, another method is suggested by Nashed [44]. Unlike our approach he finds the best least squares solution by applying Galerkin's method, with a suitably chosen basis, to the consistent equations $A^*Ax = A^*b$ and $Ax = P_{R(A)}b$, rather than to $Ax = b$.

Consider an equation $Ax = b$ in a separable Banach space X, where $A \in \mathcal{L}_b(X)$ and $b \in X$, not necessarily $b \in R(A)$, are given. We assume that $A = I + N$, where N is compact (which implies that $N(A)$ is finite dimensional) and $R(A)$ is closed. Further we assume that $R(A) = N(A)^c$. Denote by $\{\phi_i : i = 1,\ldots,m\}$ a basis of $N(A)$, $m = \dim N(A)$, and by $\{\psi_i : i = 1,2,\ldots\}$ a basis of $R(A)$. It is assumed that $R(A)$ has a countable basis. Then every $x \in X$ can be written as

$$x = \sum_{i=1}^{m} c_i(x)\phi_i + \sum_{i=1}^{\infty} d_i(x)\psi_i ,$$

where $c_i(x)$, $i = 1,\ldots,m$ and $d_i(x)$, $i = 1,2,\ldots$ are some

coefficients which depend on x. The above situation occurs, for instance, in $X = C[0,1]$ with A a Fredholm integral operator of the second kind with a continuous and symmetric kernel. The best approximate solution of such an equation can be calculated by Galerkin's method as follows: For sufficiently large n solve the system of n linear algebraic equations in n unknowns

$$\sum_{j=1}^{n} d_i(A\psi_j)\xi_j = d_i(b), \quad i = 1,\ldots,n. \tag{1.1}$$

We will show that, for sufficiently large n, the system (1.1) is consistent and that the sequence of solutions $\bar{x} = (\xi_j)$ converges to the best approximate solution of $Ax = b$, with respect to $P_{N(A)^c} = P_{R(A)}$ and $P_{R(A)}$, when $n \to \infty$. (Note that in this situation both A and A^+ leave $R(A)$ invariant.) In order to prove the consistency of (1.1), for large n, we will use a result from Krasnosel'skii et al. [33, p.212] which is stated here as the following lemma.

Lemma 1.1. Let $T \in \mathcal{L}_b(X)$ be compact and let $\{P_n: n = 1,2,\ldots\}$ be a sequence of projections in $\mathcal{L}_b(X)$, where X is a Banach space. If $P_n \to I$ strongly, i.e., for every $x \in X$,

$$\|P_n x - x\| \to 0 \quad \text{as} \quad n \to \infty,$$

then $\|(I - P_n)T\| \to 0$ as $n \to \infty$.

In our situation we specify

$$Y = X, \quad \tilde{Y} = \tilde{X} = \text{span}\{\psi_1, \ldots, \psi_n\} \quad \text{and} \quad P_{\tilde{Y}} = P_{\tilde{X}},$$

which is defined by

$$P_{\tilde{X}}x = P_{\tilde{X}}\left(\sum_{i=1}^{m} c_i(x)\phi_i + \sum_{i=1}^{\infty} d_i(x)\psi_i\right) = \sum_{i=1}^{n} d_i(x)\psi_i .$$

Further, J and H are defined by

$$Jx = Hx = \begin{pmatrix} d_1(x) \\ \vdots \\ d_n(x) \end{pmatrix}$$

while

$$J_0 = H_0 = J|_{\tilde{X}}, \quad \bar{b} = Hb \quad \text{and} \quad \bar{A} = HAJ_0^{-1} .$$

Note that $P_{\tilde{X}} = H_0^{-1}H$ and $\bar{X} = \bar{Y}$ is the space of all n-tuples. The norm in $\bar{X}$ and $\bar{Y}$ is defined by

$$\|\bar{x}\| = \|Jx\| = \sup_{k=1,\ldots,n} \left\| \sum_{i=1}^{k} d_i(x)\psi_i \right\| . \tag{1.2}$$

We will first show that the system (1.1) is consistent for large n and then that all conditions of Theorem IV.4.1 are satisfied.

Matrix $\bar{A} = (\bar{a}_{ij})$, $\bar{a}_{ij} = d_i(A\psi_j)$, $i, j = 1,\ldots,n$ has an inverse if and only if $P_{\tilde{X}}A|_{\tilde{X}}$ has an inverse. By Lemma 1.1, where $X = R(A)$, $I = I|_{R(A)}$, $P_n = P_{\tilde{X}}|_{R(A)}$ and $T = N|_{R(A)}$,

$$\| \left(I\big|_{R(A)} - P_{\tilde{X}}\big|_{R(A)}\right) N\big|_{R(A)} \| = \| (P_{\tilde{X}} - I) P_{R(A)} N\big|_{R(A)} \| \to 0$$

as $n \to \infty$. Since $\left(A\big|_{R(A)}\right)^{-1}$ is bounded (by the assumptions on A), this further implies that

$$\| (P_{\tilde{X}} - I) P_{R(A)} N\big|_{R(A)} \| \, \| \left(A\big|_{R(A)}\right)^{-1} \| < 1$$

for sufficiently large n. Now, by specifying $X = Y = R(A)$,

$$M = A\big|_{R(A)} \quad \text{and} \quad N = (P_{\tilde{X}} - I) P_{R(A)} N\big|_{R(A)}$$

in Proposition IV.1.1 we conclude that $M + N$ is invertible, which is here

$$A\big|_{R(A)} + (P_{\tilde{X}} - I) P_{R(A)} N\big|_{R(A)} = I\big|_{R(A)} + P_{\tilde{X}} N\big|_{R(A)}$$

$$= P_{\tilde{X}} A\big|_{\tilde{X}}, \text{ when restricted to } \tilde{X}.$$

Therefore $\overline{A}$ is invertible, which implies that the system (1.1) is consistent for large n.

Let us now show that all assumptions of Theorem IV.4.1 are satisfied for sufficiently large n.

Condition I: Since $M = I$, this condition is obviously satisfied.

Condition II: We know that $\overline{A}$ is invertible, so this condition, for large n, reduces to $J_0 : \tilde{X} \to \overline{X}$, which is always satisfied.

Condition III: Since $\overline{A}$ is invertible, the condition becomes $\overline{b} = H P_{R(A)} b$, which is satisfied by our construction

of H and $\overline{b}$.

Condition IV: One can specify $\varepsilon = 0$, because $\overline{A} = HAJ_0^{-1}$.

Condition V: For an arbitrary $x \in N(A)^c$, take $u = P_{\tilde{X}}Nx$. Then

$$\|Mu - P_{R(M)}Nx\| = \|(P_{\tilde{X}} - P_{R(A)})Nx\|, \quad \text{since}$$

$$M = I \quad \text{and} \quad N(A)^c = R(A)$$

$$\leq \|(P_{\tilde{X}} - P_{R(A)})N\big|_{R(A)}\| \|x\| .$$

So, one can specify

$$\eta_1 = \|(P_{\tilde{X}} - P_{R(A)})N\big|_{R(A)}\| .$$

Condition VI: Take $v = P_{\tilde{X}}P_{R(A)}b$. Then

$$\|Mv - P_{R(M)}P_{R(A)}b\| = \|(P_{\tilde{X}} - P_{R(A)})P_{R(A)}b\| .$$

Therefore, one can choose

$$\eta_2 = \frac{1}{\|P_{R(A)}b\|}\|(P_{\tilde{X}} - P_{R(A)})P_{R(A)}b\| .$$

If $P_{R(A)}b = 0$, then set $\eta_2 = 0$.

Condition A: Since $M = I$ and $\tilde{X} = \tilde{Y}$, the condition is satisfied.

Condition B: By our construction of H and H_0, $H_0^{-1}H = P_{\tilde{X}}$. Since $\tilde{X} \subset R(A)$, one concludes that $H_0^{-1}HR(A) \subset R(A)$.

Condition C: $J(N(M)^c) = \overline{X}$, which is closed.

Condition D: Since $R(\overline{A}) = \overline{X}$, this condition is satisfied by the construction of H.

Condition E: For large n, $\overline{A}$ is invertible and $N(\overline{A})^c = \overline{X}$, so the condition is satisfied.

In order to prove conditions (ii) of Theorem IV.4.1 we proceed as follows: Define a linear mapping T from $R(A)$ into the space of sequences $Tx = (d_i(x))$, $i = 1,2,\ldots$ such that $\sum_{i=1}^{\infty} d_i(x)\psi_i$ is an element in $R(A)$. It is shown in [37, p.135] that T is a linear bijection and that T is bounded and has the bounded inverse T^{-1}, if the norm in the sequence space is defined by

$$\|Tx\| = \sup_k \left\| \sum_{i=1}^{k} d_i(x)\psi_i \right\| . \tag{1.3}$$

Since $\|J_0 x\| \le \|Tx\|$ for every $x \in \tilde{X} \subset R(A)$, where the norms are taken as in (1.2) and (1.3), respectively, one concludes that $\|J_0\| \le \|T\| < \infty$, regardless of n. Hence $\sup_n \|J_0\| < \infty$. Space $\overline{X}$ is homeomorphic with the subspace of the above sequence space consisting of all sequences with zero components from $(n+1)$-st on. Therefore $H_0^{-1}\bar{x} = T^{-1}\bar{x}$ for all $\bar{x} \in \overline{X}$. Hence $\|H_0^{-1}\| = \|T^{-1}|_{\overline{X}}\| \le \|T^{-1}\| < \infty$, regardless of n, and one concludes that $\sup_n \|H_0^{-1}\| < \infty$. Since $\overline{A}$ is invertible for large n, $P_{R(\overline{A})} = I$ and thus $\sup_n \|P_{R(\overline{A})}\| < \infty$.

Finally, the conditions (iii) are satisfied, since $\varepsilon = 0$, $\eta_1 \to 0$ and $\eta_2 \to 0$ as $n \to \infty$, by Lemma 1.1, while $J_0 = H_0$, and thus $\sup_n \|J_0^{-1}\| = \sup_n \|H_0^{-1}\| < \infty$, and $\|Hx\| \le \|Tx\|$ for every

$x \in R(A)$, $Hx = 0$ for $x \in R(A)^c$, by the construction of H and T, which implies $\sup_n \|H\| \le \|T\| < \infty$, regardless of n.

All the conditions of Theorem IV.4.1 are satisfied and one concludes that $\lim_{n\to\infty} \|x^* - J_0^{-1}\bar{x}^*\| = 0$, where x^* is the best approximate solution of $Ax = b$ and $\bar{x}^*$ is the exact solution (for large n) of the approximate equation (1.1).

The best approximate solution of $Ax = b$ can also be calculated by solving systems of linear algebraic equations (1.3) in the case of a proper splitting $A = M + N$ if, in addition to the proper splitting, $M^+ : \tilde{X} \to \tilde{X}$ for sufficiently large n. All the conditions of Theorem IV.4.1 are still satisfied. The only modification is that u and v in Conditions V and VI are taken as follows: $u = M^+ P_{\tilde{X}} N x$ and $v = M^+ P_{\tilde{X}} P_{R(A)} b$. Here $\tilde{X}$ is still $\text{span}\{\psi_1, \ldots, \psi_n\}$ in $R(A)$. In fact, this requirement on $\tilde{X}$ can be relaxed. One can choose $\tilde{X} = \text{span}\{\tau_1, \ldots, \tau_n\}$, where $\{\tau_1, \ldots, \tau_n\}$ is an arbitrary set of linearly independent vectors in X, provided that $P_{\tilde{X}} P_{R(A)} = P_{R(A)} P_{\tilde{X}}$ for sufficiently large n and $\tau_1, \ldots, \tau_n, \tau_{n+1}, \ldots$ is a basis of X. However, with this arbitrary construction of $\tilde{X}$, the system (1.1) may be inconsistent for sufficiently large n, in which case the best approximate solution $\bar{x}^* = \bar{A}^+ \bar{b}$ is obtained. Now one can show, using Lemma 1.1, that

$$J_0(N(A)^c \cap \tilde{X}) = N(\bar{A})^c \cap \bar{X} \quad \text{and} \quad H_0(N(A) \cap \tilde{X}) = N(\bar{A}) \cap \bar{X}$$

for sufficiently large n. These relations imply that the only

conditions which need verification, i.e. Conditions II and III are also satisfied.

A Galerkin method for calculating the best approximate solution of $Ax = b$ can be formulated as follows:

(i) Find a proper splitting $A = M + N$ with N compact.

(ii) Find a basis $\{\tau_1, \tau_2, \ldots\}$ of X such that

$$M^+ : \tilde{X} \to \tilde{X} \quad \text{and} \quad P_{\tilde{X}} P_{R(A)} = P_{R(A)} P_{\tilde{X}} \tag{1.4}$$

for sufficiently large n, where $\tilde{X} = \text{span}\{\tau_1, \ldots, \tau_n\}$.

(iii) Calculate $\bar{A} = HAJ_0^{-1}$ and $\bar{b} = Hb$. The elements of $\bar{A} = (a_{ij})$ are determined by $a_{ij} = e_i(A\tau_j)$, where $e_i(x)$ is the i-th coefficient of x in the expansion $x = \sum_{i=1}^{\infty} e_i(x)\tau_i$, while the elements of $\bar{b} = (b_i)$ are determined by $b_i = e_i(b)$.

(iv) Calculate the best approximate solution of $\bar{A}\bar{x} = \bar{b}$, i.e. $\bar{x}^* = \bar{A}^+\bar{b}$.

If A is written as $A = I + N$, in which case we may not have a proper splitting, then the basis $\{\tau_1, \tau_2, \ldots\}$ must be chosen as a basis of $R(A)$. The conditions (1.4) are then redundant and $\bar{A}$ is invertible for sufficiently large n.

2. An Example

Example 2.1. The best least squares solution of the equation $Ax = b$ from Example III.3.1 will now be calculated using Galerkin's method.

First, the operator A can be written as $A = M + N$, where

$$Mx(s) = x(s) - \left(x(s), \sqrt{\frac{2}{\pi}}\sin s\right)\sqrt{\frac{2}{\pi}}\sin s$$

$$Nx(s) = -\tfrac{1}{2}\left(x(s), \sqrt{\frac{2}{\pi}}\cos s\right)\sqrt{\frac{2}{\pi}}\cos s\,.$$

Here $(\cdot\,,\cdot)$ denotes the inner product in $L_2[0,\pi]$. Since M is the orthogonal projection on $(\text{span}\{\sin s\})^{\perp}$ and $[Nx(s), \sin s] = 0$ for every $x \in L_2[0,\pi]$, one concludes that $R(N) \subset (\text{span}\{\sin s\})^{\perp} = R(M)$. Furthermore

$$\|N\| \le \left(\int_0^{\pi}\int_0^{\pi}\left(\frac{1}{\pi}\cos t\,\cos\xi\right)^2 dt\,d\xi\right)^{\frac{1}{2}} = \tfrac{1}{2} < 1\,,$$

which implies that $A = M + N$ is a proper splitting, by Corollary IV.1.1.

Second, we choose the following basis of X:

$$\sqrt{\frac{2}{\pi}}\cos s\,,\quad \sqrt{\frac{2}{\pi}}\sin s\,,\quad \sqrt{\frac{2}{\pi}}\sin 3s\,,\quad \sqrt{\frac{2}{\pi}}\sin 5s\,,\ \ldots\ . \tag{2.1}$$

The conditions (1.4) are now satisfied for every n.

Third, we calculate $\overline{A}$ and $\overline{b}$ for $n = 1,2,\ldots$.

$$a_{11} = \left(\sqrt{\tfrac{2}{\pi}}\cos s\,,\, A\left(\sqrt{\tfrac{2}{\pi}}\cos s\right)\right) = \tfrac{1}{2}\,, \quad \text{since} \quad A(\cos s) = \tfrac{1}{2}\cos s$$

$$b_1 = \left(s\,,\sqrt{\tfrac{2}{\pi}}\cos s\right) = -2\sqrt{\tfrac{2}{\pi}}\ .$$

Thus $\bar{A}\bar{x} = \bar{b}$ for $n = 1$ is given by $\bar{x} = -2\sqrt{\frac{2}{\pi}}$, which gives $\bar{x}^* = -4\sqrt{\frac{2}{\pi}}$. For $n = 2$, the system (1.1) is

$$\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{pmatrix} \bar{x} = \begin{pmatrix} -2\sqrt{\frac{2}{\pi}} \\ \pi\sqrt{\frac{2}{\pi}} \end{pmatrix}$$

and its best least squares solution is

$$\bar{x}^* = \begin{pmatrix} -4\sqrt{\frac{2}{\pi}} \\ 0 \end{pmatrix}.$$

For $n = 3$, the system (1.1) becomes

$$\begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \bar{x} = \begin{pmatrix} -2\sqrt{\frac{2}{\pi}} \\ \pi\sqrt{\frac{2}{\pi}} \\ \frac{\pi}{3}\sqrt{\frac{2}{\pi}} \end{pmatrix}$$

with the best least squares solution

$$\bar{x}^* = \begin{pmatrix} -4\sqrt{\frac{2}{\pi}} \\ 0 \\ \frac{\pi}{3}\sqrt{\frac{2}{\pi}} \end{pmatrix}.$$

At the n-th step $(n \geq 3)$ we obtain

$$\bar{A} = \begin{pmatrix} \frac{1}{2} & & & & 0 \\ & 0 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & \ddots \\ 0 & & & & 1 \end{pmatrix}, \quad \bar{b} = \sqrt{\frac{2}{\pi}} \begin{pmatrix} -2 \\ \pi \\ \frac{\pi}{3} \\ \frac{\pi}{5} \\ \vdots \\ \frac{\pi}{2n-3} \end{pmatrix}$$

and the best least squares solution is

$$\bar{x}^* = \sqrt{\frac{2}{\pi}} \begin{pmatrix} -4 \\ 0 \\ \frac{\pi}{3} \\ \frac{\pi}{5} \\ \vdots \\ \frac{\pi}{2n-3} \end{pmatrix}.$$

Hence

$$J_0^{-1}\bar{x}^* = \frac{2}{\pi}\left[-4\cos s + \frac{\pi}{3}\sin 3s + \frac{\pi}{5}\sin 5s + \ldots + \frac{\pi}{2n-3}\sin(2n-3)\right].$$

Since the coefficients $(x^*(s), \tau_i)$, $i = 1,\ldots,n$ of the function $x^*(s) = s - 2\sin s - \frac{4}{\pi}\cos s$, in the basis (2.1), are $-4\sqrt{\frac{2}{\pi}}$, 0, $\frac{\pi}{3}\sqrt{\frac{2}{\pi}}$, $\frac{\pi}{5}\sqrt{\frac{2}{\pi}}$, $\ldots$, $\frac{\pi}{2n-3}\sqrt{\frac{2}{\pi}}$ for every n, we conclude that $J_0^{-1}\bar{x}^* \to x^*(s)$, i.e. $x^*(s)$ is the best least squares solution of $Ax = b$. The same result has been obtained in Example III.3.1 using iteration in $L_2[0,\pi]$.

REFERENCES

[1] N.N. Abdelmalek, "Computing the strict Chebyshev solution of overdetermined linear equations", Mathematics of Computation, 31 (1977), 974-983.

[2] N.N Abdelmalek, "An efficient method for the discrete linear L_1 approximation problem", Mathematics of Computation, 29 (1975), 844-850.

[3] N. Anderson, "A generalization of the method of averages for overdetermined linear systems", Mathematics of Computation, 29 (1975), 607-614.

[4] P.M. Anselone, Collectively Compact Approximation Theory, Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[5] P.M. Anselone and R.H. Moore, "Approximate solutions of integral and operator equations", Journal of Mathematical Analysis and Applications, 9 (1964), 268-277.

[6] K.E. Atkinson, "The solution of non-unique linear integral equations", Numerische Mathematik, 10 (1967), 117-124.

[7] K.E. Atkinson, A Survey of Numerical Methods for the Solution of Fredholm Integral Equations of the Second Kind, SIAM, Philadelphia, Pennsylvania, 1976.

[8] A. Ben-Israel, "A note on an iterative method for generalized inversion of matrices", Mathematics of Computation, 20 (1966), 439-440.

[9] A. Ben-Israel, "On direct sum decompositions of Hestenes algebras", Israel Journal of Mathematics, 2 (1964), 50-54.

[10] A. Ben-Israel, "On error bounds for generalized inverses", SIAM Journal of Numerical Analysis, 3 (1966), 585-592.

[11] A. Ben-Israel and Dan Cohen, "On iterative computation of generalized inverses and associated projections", SIAM Journal of Numerical Analysis, 3 (1966), 410-419.

[12] A. Ben-Israel and T.N.E. Greville, Generalized Inverses, Theory and Applications, Wiley-Interscience, New York, 1974.

[13] A. Berman and R.J. Plemmons, "Cones and iterative methods for best least squares solutions of linear systems", SIAM Journal of Numerical Analysis, 11 (1974), 145-154.

[14] J. Blather, P.D. Morris and D.E. Wulbert, "Continuity of set-valued metric projection", Mathematishe Annalen, 178 (1968), 12-24.

[15] Paul T. Boggs, "The convergence of the Ben-Israel iteration for nonlinear least squares problems", Mathematics of Computation, 30 (1976), 512-522.

[16] J.H. Bramble and A.H. Shatz, "On the numerical solution of elliptic boundary value problems by least squares approximation of the data", Numerical Solutions of Partial Differential Equations, II, B.E. Hubbard, ed., Academic Press, New York (1971), 107-131.

[17] J.H. Bramble and A.H. Shatz, "Least squares method for 2m-th order elliptic boundary value problems", Mathematics of Computation, 25 (1971), 1-32.

[18] J.H. Bramble and A.H. Shatz, "Rayleigh-Ritz-Galerkin methods for Dirichlet's problem using subspaces without boundary conditions", Communications on Pure and Applied Mathematics, 23 (1970), 653-676.

[19] L.M. Delves and J. Walsh, Numerical Solution of Integral Equations, Clarendon Press, Oxford, 1974.

[20] David Gay, "Modifying singular values: existence of solutions to systems of nonlinear equations having a possibly singular Jacobian matrix", Mathematics of Computation, 31 (1977), 962-973.

[21] C.W. Groetsch, Computational Theory of Generalized Inverses of Bounded Linear Operators, Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio, 1975.

[22] S.P. Gudder and M. Neumann, "Splittings and iterative methods for approximate solutions to singular operator equations in Hilbert spaces", Journal of Mathematical Analysis and Applications, 62 (1978), 272-294.

[23] Apostolos Hadjidimos, "Accelerated overrelaxation method", Mathematics of Computation, 32 (1978), 149-157.

[24] Michael P. Hanna, "Generalized overrelaxation and Gauss-Seidel convergence on Hilbert space", Proceedings of the American Mathematical Society, 35 (1972), 524-530.

[25] E.N. Houstis and T.S. Papatheodorou, "A collocation method for Fredholm integral equations of the second kind", Mathematics of Computation, 32 (1978), 159-173.

[26] Y. Ikebe, "The Galerkin method for the numerical solution of Fredholm integral equations of the second kind", Rept. CNA-S, University of Texas, Austin, Texas, (1970).

[27] W.J. Kammerer and M.Z. Nashed, "On the convergence of the conjugate gradient method for singular linear operator equations", SIAM Journal of Numerical Analysis, 9 (1972), 165-181.

[28] W.J. Kammerer and M.Z. Nashed, "Steepest descent for singular linear operators with nonclosed range", Applicable Analysis, 1 (1971), 143-159.

[29] W.J. Kammerer and R.H. Plemmons, "Direct iterative methods for least squares solutions to singular operator equations", Journal of Mathematical Analysis and Applications, 49 (1975), 512-526.

[30] L.V. Kantorovich, "Functional analysis and applied mathematics", Uspehi Matematičeskih Nauk, 3 (1948), 89-195 (Russian).

[31] L.V. Kantorovich and G.P. Akilov, Functional Analysis in Normed Spaces, Translated from Russian by D.E. Brown, Pergamon Press, Oxford, 1964.

[32] Konrad Knopp, Infinite Sequences and Series, Translated from German by F. Bagemihl, Dover Publications, New York, 1956.

[33] M.A. Krasnosel'skii et al., Approximate Solution of Operator Equations, Walters-Noordhoff Publ., Groningen, Holland, 1972.

[34] W.F. Langford, "The generalized inverse of an unbounded linear operator with unbounded constraints", SIAM Journal of Mathematical Analysis (to appear).

[35] Charles Lawson and Richard Hanson, Solving Least Squares Problems, Prentice-Hall, Englewood Cliffs, New Jersey, 1974.

[36] L. Marie Lawson, "Computational methods for generalized inverse matrices arising from proper splittings", Linear Algebra and its Applications, 12 (1975), 111-126.

[37] L.A. Liusternik and V.I. Sobolev, Elements of Functional Analysis, Frederic Ungar Publ. Comp., New York, 1961.

[38] V. Lovass-Nagy and D.L. Powers, "On under- and over-determined initial value problems", International Journal on Control, 19 (1974), 653-656.

[39] J.A. Meijerink and H.A. van der Vorst, "An iterative method for linear systems of which the coefficient matrix is a symmetric M-matrix", Mathematics of Computation, 31 (1977), 148-162.

[40] R.H. Moore and M.Z. Nashed, "Approximations to generalized inverses of linear operators", SIAM Journal of Applied Mathematics, 27 (1974), 1-16.

[41] F.J. Murray, "On complementary manifolds and projections in spaces L_p and ℓ_p", Transactions of the American Mathematical Society, 41 (1937), 138-152.

[42] N.I. Muskhelishvili, Singular Integral Equations, Noordhoff, Groningen, Holland, 1946.

[43] M.Z. Nashed, "Generalized inverses, normal solvability, and iteration for singular operator equations", in Nonlinear Functional Analysis and Applications (L.B. Rall, editor), Academic Press, New York (1971), 311-359.

[44] M.Z. Nashed, "Perturbations and approximations for generalized inverses and linear operator equations", in Generalized Inverses and Applications (M.Z. Nashed, editor), Academic Press, New York (1976), 325-396.

[45] M.Z. Nashed and Grace Wahba, "Convergence rates of approximate least squares solutions of linear integral and operator equations of the first kind", Mathematics of Computation, 28 (1974), 69-80.

[46] T.G. Newman and P.L. Odell, "On the concept of a p-q generalized inverse of a matrix", SIAM Journal of Applied Mathematics, 17 (1969), 520-525.

[47] W.V. Petryshyn, "On generalized inverses and on the uniform convergence of $(I - \beta K)^n$ with application to iterative methods", Journal of Mathematical Analysis and Applications, 18 (1967), 417-439.

[48] W.V. Petryshyn, "On the generalized overrelaxation method for operation equations", Proceedings of the American Mathematical Society, 14 (1963), 917-924.

[49] W.V. Petryshyn, "On the extrapolated Jacobi or simultaneous displacements method in the solution of matrix and operator equations", Mathematics of Computations, 19 (1965), 37-56.

[50] J.L. Phillips, "Collocation as a projection method for solving integral and other operator equations", Thesis, Purdue University (1969).

[51] P.M. Prenter, "Collocation method for the numerical solution of integral equations", SIAM Journal of Numerical Analysis, 10 (1973), 570-581.

[52] C.R. Rao and S.K. Mitra, Generalized Inverse of Matrices and its Applications, Wiley, New York, 1971.

[53] W. Rudin, Functional Analysis, McGraw-Hill, New York, 1973.

[54] S.M. Serbin, "Computational investigations of least-squares type methods for the approximate solutions of boundary value problems", Mathematics of Computation, 29 (1975), 777-793.

[55] I. Singer, Bases in Banach Spaces I, Springer-Verlag, New York, 1970.

[56] I. Stakgold, Boundary Value Problems of Mathematical Physics, Volume I, Macmillan Series in Advanced Mathematics and Theoretical Physics, Macmillan, New York, 1969.

[57] G.W. Stewart, "On the continuity of the generalized inverse", SIAM Journal of Applied Mathematics, 17 (1969), 33-45.

[58] A.E. Taylor, Introduction to Functional Analysis, Wiley, New York, 1958.

[59] K.S. Thomas, "On the approximate solution of operator equations", Numerische Mathematik, 23 (1975), 231-239.

[60] R.S. Varga, Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.

[61] R.S. Varga, "Extensions of the successive overrelaxation theory with applications to finite element approximations", Department of Mathematics, Kent University, Kent, Ohio.

[62] R.S. Varga, Functional Analysis and Approximation Theory in Numerical Analysis, SIAM, Regional Conference Series in Applied Mathematics, No.3, Chapter 6, 1971.

[63] H. Wolkowicz and S. Zlobec, "Calculating the best approximate solution of an operator equation", Mathematics of Computation, October (1978), (to appear).

[64] Gerhard Zielke, Beiträge: Zur Theorie Und Berechnung Von Verallgemeinerten Inversen Matrizen, Martin-Luther-Universität, Halle-Wittenberg, 1977.

[65] S. Zlobec, "On computing the best least squares solutions in Hilbert spaces", Rendiconti del Circolo Matematico di Palermo, 26 (1977), 1-15.

ADDITIONAL REFERENCES

[66] M.B. Gagua, "On the approximate solution of linear equations", Trudy Vycisl. Centra Akad. Nauk Gruzin. SSSR, 3 (1963), 27-47.

[67] V.V. Ivanov, The Theory of Approximate Methods and Their Application to the Numerical Solution of Singular Integral Equations, Noordhoff International Publishing, Leyden, 1976.

[68] J.J. Koliha, "On the iterative solution of linear operator equations with selfadjoint operators", J. Australian Math. Soc., 13 (1972), 241-255.

[69] J.J. Koliha, "Convergent and stable operators and their generalizations", J. Math. Anal. Appl., 43 (1973), 778-794.

[70] J.J. Koliha, "Series representation of pseudoinverses and partial inverses of operators", Bull. Amer. Math. Soc., 80 (1974), 325-328.

[71] J.J. Koliha, "Power convergence and pseudoinverses of operators between Banach spaces", J. Math. Anal. Appl., 48 (1974) 446-469.

[72] M.Z. Nashed and G.F. Votruba, "A unified approach to generalized inverses of linear operators: I. Algebraic topological and projectional properties", Bull. Amer. Math. Soc., 80 (1974), 831-835.

[73] M.Z. Nashed and G.F. Votruba, "A unified operator theory of generalized inverses", in Generalized Inverses and Application, M.Z. Nashed, ed., Academic Press, New York, 1976, 1-109.

[74] L. Rakovshchik, "Approximate solutions of equations with normally solvable operators", U.S.S.R. Computational Math. and Math. Phys., 6 (1966), 3-11.

[75] G.F. Votruba, "Generalized Inverses and Singular Equations in Functional Analysis", Doctoral Dissertation, University of Michigan, Ann Arbor, Michigan, 1963.

PART B

CONVEX PROGRAMMING

I. INTRODUCTION

1.

Consider the convex program

$$(P) \qquad \begin{aligned} & f^0(x) \to \min \\ & \text{s.t.} \\ & f^k(x) \le 0, \ k \in P = \{1,\ldots,m\}, \end{aligned}$$

where $f^k : X \to R$, $k \in \{0\} \cup P$, are continuous convex functions and X is a locally convex linear Hausdorff space. The convexity assumption represents the natural framework for the treatment of optimization problems. One can now develop an elegant theory that allows a wide range of applications. The optimality conditions and algorithms developed, using this theory, usually assume that Slater's condition or some other constraint qualification is satisfied. However, recently Ben-Israel, Ben-Tal and Zlobec [10], [11] have used the cones of directions of constancy of the constraints, to characterize optimality without assuming a constraint qualification. Algorithms for solving (P) were then given in [14]. In this part of the thesis we continue this approach and study the role played by the cones of directions of constancy when deriving optimality conditions and algorithms. In particular, we use the approach of Gould and Tolle [30] and study the relationship between the geometry of the feasible set and the analytic properties of the constraints. We will see that the cones of directions of constancy provide exactly

the missing analytic information needed in order to describe the feasible set and thus characterize optimality.

In Chapter II we summarize some basic results in convex analysis and optimization. We also introduce the set of 'badly behaved' constraints at x, $P^b(x)$, and present a closedness criterion for the sum of two, not necessarily convex, cones. In Chapter III we first give a lemma which shows the different relationships between the cones of directions of constancy, the tangent cone of the feasible set and the cone of subgradients. Using this lemma, we derive some old and some new characterizations of optimality. Chapter IV examines the notions of constraint qualifications and regularization. The most important results here are the regularization technique in Theorem IV.4.2 and the weakest constraint qualifications in Theorem IV.3.1. Finally, in Chapter V we present the Method of Reduction, which solves program (P) with faithfully convex constraints. The method includes an algorithm for finding the cone of directions of constancy of a faithfully convex function and also a generalization of the algorithm for finding the equality set, $P^=$, given by Abrams and Kerzner [3]. We conclude with several applications and examples.

II. PRELIMINARIES

1. Notation

We will be dealing with real locally convex linear spaces $X, Y, \ldots$ and various mappings defined on them. We list below some of the abbreviations and symbols used and the sections where their meanings are explained.

tvs	a topological linear space
lcs	a locally convex (Hausdorff) space
$\mathcal{L}(X,Y)$	the set of all linear operators from X into Y
X'	the topological dual of X, with the weak* topology
w^*-closed	weak* closed
R^n	the real Euclidean n-space
$R(B)$	the range space of the operator B
$N(B)$	the null space of the operator B
$K^\perp$	the annihilator (in X') of a set K (in X)
span K	the vector subspace spanned by K
$\overline{K}$	topological closure
bdry K	the topological boundary of K
int K	the topological interior of K
cone	a set closed under nonnegative scalar multiplication
blunt cone	a set closed under positive scalar multiplication

cone K	the convex cone generated by K
conv K	the convex hull of K
$\emptyset$	the empty set
$\phi \cdot x$	the value of $\phi(x)$ or $x(\phi)$, where $x \in X$ and $\phi \in X'$
(P)	the convex program, II.2
(P_r)	the regularized program, IV.4
P	the indexing set of the constraints, II.2
S	the feasible set, II.2
$P(x)$	the binding constraints, II.2
$P^=$	the equality set, II.2
$P^<(x) = P(x) \setminus P^=$,	II.2
$P^b(x)$	the 'badly behaved' set, II.5
dist(x,K)	the distance from x to the set K, II.5
$E_k(x)$	the directions of vanishing directional derivative, II.5
$D_f^{<,\leq,=,>}$	the cone of directions of decrease, nonincrease, constancy and increase, II.3
$\nabla f(x;d)$	the directional derivative, II.4
$\partial f(x)$	the subdifferential, II.4
$\nabla f(x)$	the gradient, II.4
M^*	the polar of M, II.6
$F^\circ(x)$	the set of continuous, convex functions which achieve their minimum in S at x, II.7

$DF^{\circ}(x) = \{\phi \in \partial f^{\circ}(x) : f^{\circ} \in F^{\circ}(x)\}$, II.7

$C_{P(x)}(x)$	the linearizing cone, II.7
$B_{P(x)}(x)$	the cone of subgradients, II.7
$T(S,x)$	the cone of tangents, II.8
K-T point	Kuhn-Tucker point, IV.1
K-T conditions	Kuhn-Tucker conditions, IV.1
regular point	IV.2
CQ	constraint qualification, IV.2
WCQ	weakest constraint qualification, IV.3
(PL)	the lexicographic problem, V.5
(PP)	the Pareto optimal problem, V.5
(PS)	the semi-infinite problem, V.5

2. The Convex Program

We consider the convex programming problem

$$\text{(P)} \qquad \begin{aligned} & f^0(x) \to \min \\ & \text{s.t.} \\ & f^k(x) \le 0, \quad k \in P = \{1,\dots,m\}, \end{aligned}$$

where $f^k: X \to R$ are continuous convex functions, defined on the *lcs* X, for all $k \in \{0\} \cup P$. (Without loss of generality, we assume that none of the functions is constant.) Unless otherwise specified, we assume that the feasible set

$$S = \{x \in X: f^k(x) \le 0, \text{ for all } k \in P\}$$

is not empty. The set of binding constraints, at $x \in S$, is

$$P(x) = \{k \in P: f^k(x) = 0\}.$$

An important subset of P, independent of x, is the equality set

$$P^= = \{k \in P: f^k(x) = 0, \text{ for all } x \in S\}.$$

(See e.g. Zoutendijk [53], Rockafellar [43] and Abrams and Kerzner [3].) This is the set of indices k for which the constraint f^k vanishes on the entire feasible set. We then denote

$$P^<(x) = P(x) \setminus P^=.$$

Note that unlike $P^=$, $P^<(x)$ depends on x.

3. Cones of Directions and Faithfully Convex Functions

Following Ben-Israel, Ben-Tal and Zlobec [11], we define the relations

$$\text{"relation" is "=", "<", "}\leq\text{" or ">",}$$

by

$$D_f^{\text{"relation"}}(x) = \{d \in X: \text{there exists } \bar{\alpha} > 0 \text{ with } f(x+\alpha d) \text{ "relation" } f(x), \text{ for all } 0 < \alpha \leq \bar{\alpha}\}.$$

These are the cones of directions of constancy, descent, nonincrease and increase respectively. For simplicity of notation, we let

$$D_k^{\text{"relation"}}(x) = D_{f^k}^{\text{"relation"}}(x)$$

and

$$D_\Omega^{\text{"relation"}}(x) = \bigcap_{k\in\Omega} D_k^{\text{"relation"}}(x), \quad \text{for } \Omega \subset P.$$

Remark 3.1. For a function f in the class of faithfully convex functions, the cone $D_f^{=}(x)$ is a subspace independent of x. Following Rockafellar [44], we say that a convex function f is faithfully convex if: f is affine on a line segment only if it is affine on the whole line containing that segment. If $X = R^n$, then Rockafellar has shown that f is faithfully convex if and only if it is of the form

$$f(x) = h(Ax+b) + a\cdot x + \alpha, \tag{3.1}$$

where $A \in R^{m\times n}$, $b \in R^m$, $a^t \in R^n$, $\alpha \in R$ and the function $h:R^m \to R$ is strictly convex. It is easy to see that $D_f^{=}(x) = N\left(\begin{bmatrix} A \\ a^t \end{bmatrix}\right)$ and is a subspace independent of x.

In the following lemma we collect some properties of the directions. We also show directly that the cone of directions of constancy of a continuous faithfully convex function on X, a *lcs*, is a subspace independent of $x \in X$.

Lemma 3.1. Suppose that $f:X \to R$ is a convex function and $x \in S$. Then:

a) $D_f^{\leq}(x)$ is a convex cone, $D_f^{<}(x)$ is a convex blunt cone.

b) $conv\ D_f^{=}(x) \subset D_f^{\leq}(x)$.

c) $D^{\leq}_{P(x)}(x) = D^{\leq}_{P^{<}(x)}(x) \cap D^{=}_{P^{=}}(x)$

$= D^{\leq}_{P^{<}(x)}(x) \cap conv\ D^{=}_{P^{=}}(x)$.

d) $D^{=}_{P^{=}}(x) \cap D^{<}_{P^{<}(x)}(x) \neq \emptyset$.

e) If f is both faithfully convex and continuous, then $D_f^{=}(x) = D_f^{=}$ is a subspace of X, independent of x.

Proof. For (a)-(d), see e.g. [11], [12].

e) First, let us show that $D_f^=(x)$ is a subspace. Suppose that $d_1, d_2 \in D_f^=(x)$ and let $d = d_1 + d_2$. If $\alpha \in R$, then

$$\begin{aligned} f(x+\alpha d) &= f(\tfrac{1}{2}(x+2\alpha d_1) + \tfrac{1}{2}(x+2\alpha d_2)) \\ &\le \tfrac{1}{2}f(x+2\alpha d_1) + \tfrac{1}{2}f(x+2\alpha d_2), \text{ since } f \text{ is convex} \\ &= f(x), \text{ since } d_1, d_2 \in D_f^=(x) \text{ and } f \text{ is faithfully convex.} \end{aligned}$$

Therefore f is bounded above on the whole line $x+\alpha d$, $\alpha \in R$, which implies that f is constant on this line (see e.g. Rockafellar [41, p.69]. Thus, $d \in D_f^=(x)$. This shows that $D_f^=(x)$ is closed under addition. That $D_f^=(x)$ is closed under scalar multiplication is clear, from the definition of a faithfully convex function.

We have left to show that $D_f^=(x) = D_f^=$, i.e. it is independent of x. Suppose that $x, y \in X$ and $d \in D_f^=(x)$. We will show that $d \in D_f^=(y)$.

Case (i): Suppose that $f(y) \le f(x)$. We will first show that

$$f(y+\alpha d) \le f(x), \quad \text{for all } \alpha \in R.$$

Let $\alpha \in R$ and $1 > t_k > 0$ with $t_k \to 0$ as $k \to \infty$. Consider the directions $z^k = \alpha d + t_k(x-y)$ and let $\gamma_k = 1/t_k$. Then

$$\begin{aligned} f(y) &\le f(x) \\ &= f(x+\gamma_k \alpha d), \text{ since } d \in D_f^=(x) \\ &= f(y+\gamma_k z^k). \end{aligned}$$

By convexity of f and since $\gamma_k > 1$, we conclude that

$$f(y + z^k) \leq f(x)$$

and thus, by continuity of f, we see that

$$f(y + \alpha d) = \lim_{k\to\infty} f(y + z^k) \leq f(x) .$$

This shows that f is bounded on the line $y + \alpha d$, $\alpha \in R$, and therefore, f is constant on this line, i.e. $d \in D_f^{=}(y)$.

Case (ii): Suppose that $f(x) < f(y)$. By a similar argument to case (i), we see that

$$f(y + \alpha d) = \lim_{k\to\infty} f(y + \alpha d + t_k(x - y)) \leq f(y) ,$$

for all $\alpha \in R$, i.e. $d \in D_f^{=}(y)$.

□

Remark 3.2. For faithfully convex functions f, on R^n, one can calculate $D_f^{=}$ explicitly, see [49] and Section V.2. The class of faithfully convex functions is quite large. It includes all analytic convex, as well as strictly convex, functions. The algorithms in Chapter V will deal mainly with these functions. In general, however, the cone of directions of constancy may not be a subspace. In fact it may be neither convex nor closed. (See [10] for examples.)

4. Subdifferentiability

We have assumed that our functions are convex, but not necessarily differentiable. Nonsmooth, or nondifferentiable, functions occur quite often in convex analysis. Applications for these functions arise in approximation theory, e.g. Dem'yanov and Malozemov [23], duality theory, e.g. Rockafellar [42] and semi-infinite programming, e.g. Ben-Tal, Kerzner and Zlobec [13]. (See also Clarke [19] and Pshenichnyi [40].) For convex functions, it is possible to develop a complete calculus without assuming differentiability, e.g. Rockafellar [41], Pshenichnyi [40] and Holmes [34]. We now recall some concepts dealing with directional derivatives and subgradients of a convex function f, defined on the *lcs* X.

The directional derivative of f at x, in the direction d, is defined as

$$\nabla f(x;d) = \lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t} .$$

Convex functions have the useful property that the directional derivatives exist universally, e.g. [41, Theorem 23.1].

A vector $\phi \in X'$ is said to be a subgradient of a convex function f, at the point x, if

$$f(z) \geq f(x) + \phi \cdot (z - x) , \quad \text{for all } z \in X .$$

The set of all subgradients of f at x is then called the subdifferential of f at x and is denoted by $\partial f(x)$.

If the directional derivative of f at x is a continuous linear functional, i.e. if $\nabla f(x;\cdot) = \phi \in X'$, then

$$\phi \cdot d = \lim_{t \to 0} \frac{f(x + td) - f(x)}{t}$$

and ϕ is called the gradient of f at x and denoted $\nabla f(x)$. Note that in this case

$$\partial f(x) = \{\nabla f(x)\}.$$

We collect some useful properties in the following lemma. For more details and proofs, see e.g. [34], [41].

Lemma 4.1. Suppose that $f: X \to R$ is convex. Then

a) $\nabla f(x;\cdot)$ is a finite, sublinear functional on X, for all $x \in X$.

If, in addition, f is continuous at x, then:

b) $\nabla f(x;d) = \max\{\phi \cdot d : \phi \in \partial f(x)\}$

and

c) $\partial f(x)$ is a non-empty, w^*-compact convex subset of X'.

The next lemma presents some of the relations that exist between the subgradients and the directions introduced in Section 3. For the proofs see Ben-Tal and Ben-Israel [12].

Lemma 4.2. Suppose that $f:X \to R$ is convex. Then

a) $D_f^<(x) = \{d \in X:\nabla f(x;d) < 0\}$.

If $\nabla f(x)$ exists, then:

b) $D_f^<(x) = \{d \in X:\nabla f(x)\cdot d < 0\}$

and

c) $conv\ D_f^=(x) = D_f^=(x) \subset \{d \in X:\nabla f(x)\cdot d = 0\}$.

5. The 'Badly Behaved' Constraints

For $x \in S$, let

$$P^b(x) \triangleq \{k \in P^=:(E_k(x) \cap D_k^>(x) \cap C_{P(x)}(x)) \setminus \overline{D^=_{P^=}(x)} \neq \emptyset\},$$

where

$$E_k(x) = \{d \in X:\nabla f^k(x;d) = 0\}$$

and

$$C_{P(x)}(x) = \{d \in X:\nabla f^k(x;d) \leq 0, \quad \text{for all} \quad k \in P(x)\}.$$

(See Section 7 below for further properties of the 'linearizing cone', $C_{P(x)}(x)$.) We call $P^b(x)$ the set of 'badly behaved' constraints at $x \in S$, for program (P).

The set $P^b(x)$ is the set of constraints that creates problems in the Kuhn-Tucker theory. We can see that these are the constraints in $P^=$, whose analytic properties (given by the directional derivatives) do not fully describe the geometry of the feasible set (given by the feasible directions). It will be shown in IV.3 that

$$P^b(x) = \emptyset$$

is an essential condition for the Kuhn-Tucker theory to hold at x, independent of f^0. The set $P^b(x)$ will also be used in the characterizations of optimality in III.4.

Abrams and Kerzner [3] have given an algorithm that finds the set $P^=$. (See V.3 for a modified version of their algorithm.) Once $P^=$ is found, then, for any given index $k_o \in P^=$, we see that $k_o \in P^b(x)$ if and only if the system

$$\begin{cases} \nabla f^{k_o}(x;d) = 0 \\ \nabla f^{k}(x;d) \leq 0, \quad \text{for all} \quad k \in P(x)\backslash k_o, \\ \qquad d \notin D^=_{k_o}(x) \cup \overline{D^=_{P^=}(x)} \end{cases}$$

is consistent. (Note that when $D^=_{k_o}(x)$ is closed, then $\overline{D^=_{P^=}(x)} \subset D^=_{k_o}(x)$. This simplifies the above system and thus, the corresponding definition for the 'badly behaved' set.)

The set $P^b(x)$ is not equal to $P^=$ in general. In fact, if

(5.1) $$E_k(x) = D^=_k(x),$$

then f^k is 'never badly behaved' at x, i.e. $k \notin P^b(x)$ independent of the other constraints. This class of functions which are 'never badly behaved' at x includes all continuous linear functionals on X. Furthermore, if $X = R^n$, $\nabla f(x) \neq 0$ and f is a strictly convex function of one variable, considered as a function on R^n (i.e. if the restriction of f to R^1 is strictly convex), then f is a nonlinear function which is 'never badly behaved' at x. (See Ben-Israel, Ben-Tal and Zlobec [10] for definitions and properties of functions whose restrictions are strictly convex.)

The class of functions which are 'never badly behaved' at x also includes the 'distance' functions defined below. We will see, in IV.4, that every program (P) can be 'regularized' by the addition of one such 'distance' function.

Lemma 5.1. Suppose that X is a Hilbert space, K is a nonempty closed convex cone in X, $x \in S$ and $k \in P$. If, for $y \in X$,

$$f^k(y) = \operatorname{dist}(y - x, K)$$

$$\triangleq \min_{z \in K} \| (y - x) - z \| , \tag{5.2}$$

then f^k is a convex function on X, which is 'never badly behaved' at x. Furthermore,

$$\nabla f^k(x;d) = \begin{cases} 0 & \text{if } d \in K \\ \text{positive} & \text{otherwise.} \end{cases} \tag{5.3}$$

Proof. First, let us show that the function f^k is convex. Suppose that $y^1, y^2 \in X$. Since X is a Hilbert space and K is a nonempty, closed convex set, there exist unique points z^1 and z^2, in K, which are closest to $y^1 - x$ and $y^2 - x$ resp., see e.g. Rudin [46, p.78]. Thus

$$f^k(y^1) = \|y^1 - x - z^1\|, \quad f^k(y^2) = \|y^2 - x - z^2\|.$$

Let $0 \le \lambda \le 1$. Then

$$\lambda z^1 + (1 - \lambda) z^2 \in K$$

and

$$\begin{aligned} f^k(\lambda y^1 + (1 - \lambda) y^2) &= \operatorname{dist}(\lambda y^1 + (1 - \lambda) y^2 - x, K) \\ &\le \|(\lambda y^1 + (1 - \lambda) y^2 - x) - (\lambda z^1 + (1 - \lambda) z^2)\| \\ &\le \lambda \|(y^1 - x) - z^1\| + (1 - \lambda) \|(y^2 - x) - z^2\| \\ &= \lambda f^k(y^1) + (1 - \lambda) f^k(y^2). \end{aligned}$$

Therefore f^k is a convex function on X. Now let $d \in X$. Then

$$\begin{aligned} \nabla f^k(x;d) &= \lim_{t \downarrow 0} \frac{f^k(x + td) - f^k(x)}{t} \\ &= \lim_{t \downarrow 0} \frac{\operatorname{dist}(td, K)}{t} \\ &= \operatorname{dist}(d, K), \quad \text{since } K \text{ is a cone.} \end{aligned}$$

This yields (5.3) and further implies that (5.1) holds. Therefore f^k is 'never badly behaved' at x. □

Example 5.1. Consider the program (P) with the single constraint in one variable, $f^1(x) \leq 0$, where

$$f^1(x) = \begin{cases} x^2 & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$P^b(x) = \begin{cases} \{1\} & \text{if } x = 0, \\ \emptyset & \text{otherwise.} \end{cases}$$

However, if

$$f^1(x) = \begin{cases} x^2 + x & \text{if } x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

then $P^b(x) = \emptyset$ for all x, i.e. f^1 is not 'badly behaved' at $\bar{x}$, though $1 \in P^=$.

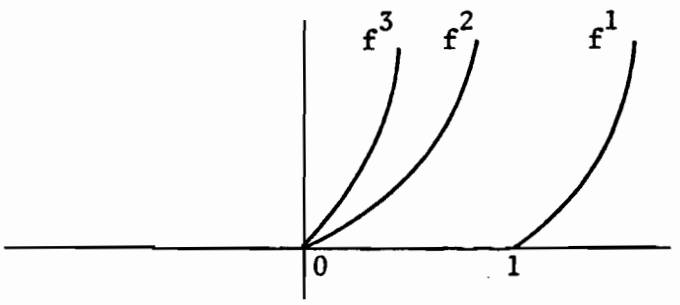
Example 5.2. Now consider the three functions

$$f^1(x) = \begin{cases} (x-1)^2 & \text{if } x \geq 1, \\ 0 & \text{otherwise,} \end{cases}$$

$$f^2(x) = \begin{cases} x^2 & \text{if } x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$f^3(x) = \begin{cases} x^2 + x & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$



If the program (P) has just the two constraints f^1 and f^2, then

$$P^b(x) = \begin{cases} \{2\} & \text{if } x = 0, \\ \emptyset & x \neq 0 \text{ and } x \in S. \end{cases}$$

If, however, the program (P) has all three constraints, then

(5.4) $$P^b(x) = \emptyset \quad \text{for all } x \in S.$$

As mentioned above, we shall see that, (5.4) is essential for the Kuhn-Tucker theory to hold, independent of the choice of the objective function f^0.

6. Polar Sets and Closedness Criteria

In this section we collect some useful results on polar sets. These results can be found in e.g. Girsanov [28] and Holmes [34]. See also Borwein [16]. We also present a closedness criterion for the sum of two cones.

Recall that for $M \subset X$ and X a *lcs*, the polar of M is

$$M^* = \{\phi \in X' : \phi \cdot x \geq 0 \quad \text{for all} \quad x \in M\}.$$

M^* is then a w^*-closed convex cone in X'. However, if $M \subset X'$, then we define its polar to be

$$M^* = \{x \in X : \phi \cdot x \geq 0 \quad \text{for all} \quad \phi \in M\}.$$

M^* is now a w-closed convex cone in X.

Lemma 6.1. Suppose that K and L are subsets of X and C is a subset of X'. Then:

a) $K \subset L$ implies $L^* \subset K^*$.

b) $K^* = (\overline{conv}\ K)^*$, $C^* = (\overline{conv}\ C)^*$, $K^{**} = \overline{cone}\ K$ and $C^{**} = \overline{cone}\ C$.

If, in addition, K and L are closed convex cones, then

c) $(K \cap L)^* = \overline{K^* + L^*}$.

The following closedness criteria will be used in deriving the optimality conditions in Chapter III.

Lemma 6.2. Suppose that L is a closed cone in X, C is a compact subset of X not containing the origin, and K is the cone generated by C, i.e.

$$K = \bigcup_{\lambda \geq 0} \lambda C .$$

If

$$C \cap (-L) = \emptyset , \tag{6.1}$$

then

$K + L$ is closed.

Proof. Suppose that the net $k^n + \ell^n \to p$, where $k^n \in K$ and $\ell^n \in L$. We need to show that $p \in K + L$. For $0 \neq \phi \in K$, let

$$\|\phi\| \triangleq \inf\{t > 0 : t^{-1}\phi \in C\} .$$

Note that $\|\phi\| < \infty$, for all $0 \neq \phi \in K$, by the definition of K. (In fact, $\|\cdot\|$ is the Minkowski functional of C and is a seminorm on X, when C is a balanced convex absorbing set.)

Case (i): Suppose that $\sup \|k^n\| < \infty$. Then

$$k^n = t_n c^n , \quad \text{for some } c^n \in C , \ t_n > 0 \ \text{ and } \ \sup t_n < \infty .$$

Since C is a compact set, we can assume that $k^n \to k \in K$. Therefore

$$\lim \ell^n = \lim (p - k^n) = p - k .$$

Let $\ell \triangleq p - k$. Then $\ell \in L$, since L is closed, and

$$p = k + \ell \in K + L .$$

Case (ii): Suppose that $\sup \|k^n\| = \infty$. We can assume that $\|k^n\| \to \infty$, or else we can extract a bounded subnet and use case (i). Let $\alpha_n = 1/\|k^n\|$. Then $\alpha_n \to 0$ and $\|\alpha_n k^n\| = 1$ for all n. As in case (i), we can assume that

$$\alpha_n k^n \to k \in K .$$

Furthermore, since $\|\alpha_n k^n\| = 1$, we see that

$$\alpha_n k^n \in C, \quad \text{for all } n .$$

This implies that $k \in C$. Now, since

$$k^n + \ell^n \to p \quad \text{and} \quad \alpha_n \to 0 ,$$

we see that

$$\alpha_n k^n + \alpha_n \ell^n \to \lim \alpha_n p = 0 ,$$

i.e.

$$\alpha_n \ell^n \to \lim (-\alpha_n k^n) = -k ,$$

and $-k \in L$, since L is closed. Thus

(6.2) $$k \in C \cap (-L) ,$$

contradicting the hypothesis. This implies that $\sup \|k^n\| < \infty$ and we can apply case (i). □

Note that since X' with the w^*-topology is a *lcs*, the lemma holds when X is replaced with X'.

Remark 6.1. The condition

$$K \cap (-L) = \{0\}$$

alone, does not necessarily imply that

$$K + L \text{ is closed.}$$

Halmos [33, p.28] has given an example of two subspaces, in a Hilbert space, with zero intersection and a non-closed sum.†

Closedness of a sum is related to closedness of the linear image of a set. Suppose that $U \subset X$ and $T: X \to Y$ is a bounded linear operator, where X and Y are Banach spaces. If T has closed range, then the linear image TU is closed if and only if $U + N(T)$ is closed, see e.g. Atteia [5] and Holmes [35, p.142]. Therefore Lemma 6.2 implies that

$$TU \text{ is closed}$$

if, T has closed range, U is a cone generated by a compact set C which does not contain the origin and

$$C \cap N(T) = \emptyset .$$

---

† I would like to thank Professor Rockafellar for pointing out this example to me.

The above criteria is used to prove the existence of solutions in optimal control and spline approximation problems, see e.g. [35, Section 21], [21] and [27].

Remark 6.2. Lemma 6.2 holds if (6.1) is weakened to read

$$C \cap (-L) \cap (-bdry\ L) = \emptyset .$$

Proof. Following the proof of Lemma 6.2, we see that we fail to obtain a contradiction if

$$k \in C \cap (-int\ L)$$

in (6.2). Therefore, in this case, we need to show that $p \in K + L$. We accomplish this by showing that

$$p \in int\ (K + L) .$$

Since $k \in -int\ L$, we can find a convex neighborhood of the origin, η, such that

$$-k + \eta \subset L .$$

(Recall that every *lcs* X has a convex local base, see e.g. [45].) It is therefore sufficient to show that the open set

$$p + \eta \subset \overline{K + int\ L}.$$

Let

$$\overline{p} \in p + \eta \quad \text{and} \quad q \triangleq \overline{p} - p .$$

Then $q \in \eta$ and since $\alpha_n \ell^n \to -k$, $\alpha_n \to 0$ and η is convex, we can assume that

$$\alpha_n \ell^n + \alpha_n q \in -k + \eta \subset \mathit{int}\ L, \quad \text{for all} \quad n.$$

Therefore

$$\ell^n + q = \frac{1}{\alpha_n}(\alpha_n \ell^n + \alpha_n q) \in \mathit{int}\ L, \quad \text{for all} \quad n,$$

since L is a cone. This implies that

$$k^n + \ell^n + q \to p + q = \bar{\bar{p}},$$

where $k^n \in K$ and $\ell^n + q \in \mathit{int}\ L$, i.e. $\bar{\bar{p}} \in \overline{K + \mathit{int}\ L}$. □

7. Some Special Cones

We now present some well-known definitions of cones used in mathematical programming, see e.g. Gould and Tolle [30] and Abadie [1]. However, the definitions are stated here in terms of subgradients.

By $F^\circ(x)$, we denote the cone of all continuous convex objective functions f° with the property that x minimizes f° over S. Then the cone

$$DF^{\circ}(x) \triangleq \{\phi \in X' : \phi \in \partial f^{\circ}(x), \text{ for some } f^{\circ} \in F^{\circ}(x)\}.$$

For every subset Ω of $P(x)$, the linearizing cone at $x \in S$, with respect to Ω, is

$$C_{\Omega}(x) = \{d \in X : \phi \cdot d \leq 0 \text{ for all } \phi \in \partial f^{k}(x) \text{ and all } k \in \Omega\}.$$

By Lemma 4.1(b), we see that

$$C_{\Omega}(x) = \{d \in X : \nabla f^{k}(x;d) \leq 0 \text{ for all } k \in \Omega\}.$$

This formulation corresponds with the definition of the linearizing cone given in Section 5.

The cone of subgradients at x is

$$B_{\Omega}(x) = \{\phi \in X' : \phi = \sum_{k \in \Omega} \lambda_k \phi^k \text{ for some } \lambda_k \geq 0 \text{ and } \phi^k \in \partial f^{k}(x)\}.$$

This cone is convex. It is also w^*-closed, if $0 \notin \mathit{conv} \bigcup_{k \in \Omega} \partial f^{k}(x)$, by Lemma 6.2. We now set

$$B_{\emptyset}(x) = \{0\}.$$

The linearizing cone and the cone of subgradients have the following dual property.

Lemma 7.1. Suppose that $x \in S$ and $\Omega \subset P(x)$. Then

$$\overline{B_\Omega(x)} = - C^*_\Omega(x).$$

Proof. Since

$$B_k(x) = \partial f^k(x)^{**}, \quad \text{by Lemma 6.1(b)}$$

$$= - C^*_k(x), \quad \text{by definition,} \tag{7.1}$$

we conclude that

$$- C^*_\Omega(x) = - \overline{\sum_{k\in\Omega} C^*_k(x)}, \quad \text{by Lemma 6.1(c)}$$

$$= \overline{B_\Omega(x)}, \quad \text{by (7.1).}$$

□

Gould and Tolle [31] used Farkas' Lemma to prove the above result, for differentiable functions on R^n. Note that in the differentiable case, $B_\Omega(x)$ is closed. This may fail in the nondifferentiable case. (Consider $B_1(0)$, where f^1 is the support function of the set $\{x \in R^2 : \|x - (0,1)\| \le 1\}$.[†]) The closure of $B_{P(x)}(x)$ will play an essential role in the characterization of regularity in IV.3.

[†] I would like to thank Professor J. Borwein for pointing this out to me.

8. The Cone of Tangents

For $x \in M$, where M is an arbitrary set in X, the cone of tangents to M at x is defined by

$$T(M,x) = \{d \in X : d = \lim \lambda_k(x^k - x), \text{ where } x^k \in M, \lambda_k \geq 0 \text{ and } x^k \to x.\}$$

This cone is closed and it is convex if M is. In fact. when M is convex, it is exactly the $\overline{cone}\ (M - x)$, the support cone of M at x. For further properties, see e.g. Guignard [32] and Holmes [34]. For a relationship of the cone of tangents with derivatives in mathematical programming, see e.g. Massam [38] and Massam and Zlobec [39].

The cone of tangents is used in optimization theory to describe the geometry of the feasible set. For example, one gets the following characterization of optimality.

Theorem 8.1. [34, p.30] $x \in S$ is optimal for (P) if and only if

$$\partial f^\circ(x) \cap T^*(S,x) \neq \emptyset.$$

This result will be the starting point for our characterizations of optimality. Note that the characterization is in terms of the feasible set, rather than the constraints.

III. CHARACTERIZATIONS OF OPTIMALITY

1. Introduction

Optimality conditions of the Kuhn-Tucker type usually use the analytic properties of the constraints, given by the cone of subgradients, $B_{P(x)}(x)$, or the linearizing cone, $C_{P(x)}(x)$. However, these cones may not provide all the required information needed to characterize optimality. As seen in Theorem II.8.1, the cone of tangents of the feasible set does provide enough information to characterize optimality. However, this cone is hard to use computationally. In [10], it was shown that the cones of directions of constancy can be used to characterize optimality. We will see how these cones provide the missing information. Using Theorem II.8.1, we first present optimality conditions of the type given by Guignard [32] and Gould and Tolle [30]. (See Theorem 3.1.) From this result we deduce several new, as well as known, optimality conditions.

2. A Basic Lemma

The following lemma presents several relationships between the types of cones mentioned in Section 1. This lemma is of key importance in the proofs in the rest of this chapter and in Chapter IV.

Lemma 2.1. Suppose that $x \in S$, Ω satisfies $P^b(x) \subset \Omega \subset P^=$ and either $conv\ D^=_\Omega(x)$ is closed or $\Omega = P^=$.

Then:

a) $T(S,x) = \overline{D^{\leq}_{P(x)}(x)}$.

b) $\overline{conv\ D^=_\Omega(x)} \cap C_{P(x)}(x) = \overline{D^=_\Omega(x)} \cap C_{P(x)}(x) = \overline{D^=_{P^=}(x)} \cap C_{P(x)}(x)$.

c) $T(S,x) = \overline{conv\ D^=_\Omega(x)} \cap C_{P(x)}(x)$.

d) $conv \left\{ \bigcup_{k \in P^<(x)} \partial f^k(x) \right\} \cap \left(D^=_\Omega(x) \right)^* = \emptyset$.

e) $T^*(S,x) = \overline{\left(D^=_\Omega(x)\right)^* + C^*_{P(x)}(x)} = \overline{\left(D^=_\Omega(x)\right)^* - B_{P(x)}(x)}$.

f) $T^*(S,x) = \left(D^=_{P^=}(x)\right)^* + C^*_{P(x)}(x) = \left(D^=_{P^=}(x)\right)^* - B_{P(x)}(x)$.

(Recall that X' is given the w^*-topology and thus is a *lcs*.)

Proof. Since the point $x \in S$ is fixed throughout, we will omit it in this proof when the intended meaning is clear, e.g. $D^{\leq}_P$ will denote $D^{\leq}_{P(x)}(x)$, $P^<$ will denote $P^<(x)$, etc.

a) Suppose that $d \in T(S,x)$, with associated nets $\{x^k\}$ and $\lambda_k \geq 0$. Let $d^k = x^k - x$. Since $x^k \in S$, we see that $x + d^k \in S$, i.e. $d^k \in D_p^{\leq}$. Furthermore, since $D_p^{\leq}$ is a cone, $\lambda_k d^k \in D_p^{\leq}$. But

$$d = \lim \lambda_k (x^k - x) = \lim \lambda_k d^k,$$

which implies that $d \in \overline{D_p^{\leq}}$. Thus

$$T(S,x) \subset \overline{D_p^{\leq}}.$$

Conversely, suppose that $d \in D_p^{\leq}$. Then, there exists $\bar{\alpha} > 0$ such that

$$x + \alpha d \in S, \quad \text{for all} \quad 0 < \alpha \leq \bar{\alpha}.$$

Let

$$x^k = x + \frac{\bar{\alpha}}{k} d, \quad \text{for} \quad k \geq 1.$$

Then $x^k \in S$ and, since X is a *tvs*, $x^k \to x$ as $k \to \infty$. By choosing $\lambda_k = \frac{k}{\bar{\alpha}}$, we see that

$$\lambda_k (x^k - x) = d, \quad \text{for all} \quad k \geq 1.$$

Therefore $d \in T(S,x)$, i.e. we have shown that $D_p^{\leq} \subset T(S,x)$. The result now follows since $T(S,x)$ is closed.

b) (i) First, let us show that

$$\overline{conv}\ D_\Omega^{=} \cap C_p \subset \overline{\overline{conv}\ D_\Omega^{=} \cap C_p}. \tag{2.1}$$

By hypothesis and the fact that $\overline{conv}\ D^{=}_{P^{=}} \subset C_{P^{=}}$ and $C_P = C_{P^{=}} \cap C_{P^{<}}$, it is sufficient to show that

$$\overline{conv}\ D^{=}_{P^{=}} \cap C_{P^{<}} \subset conv\ D^{=}_{P^{=}} \cap C_{P^{<}}. \tag{2.2}$$

But this follows since

$$\hat{d} \in D^{=}_{P^{=}} \cap D^{<}_{P^{<}} \subset conv\ D^{=}_{P^{=}} \cap int\ C_{P^{<}} \neq \emptyset, \tag{2.3}$$

by Lemma II.3.1(d).

Let us define

$$d_\lambda \triangleq \lambda\hat{d} + (1 - \lambda)d, \tag{2.4}$$

for scalars λ and vectors d.

(ii) Next, let us show that

$$\overline{D^{=}_{\Omega} \cap C_P} \subset \overline{D^{=}_{P^{=}}} \cap C_P\ . \tag{2.5}$$

Suppose that

$$d \in (D^{=}_{\Omega} \cap C_P) \setminus (\overline{D^{=}_{P^{=}}} \cap C_P)\ .$$

We will find a set $I \subset P^{=}$ and feasible directions $d_\lambda \in D_P^{\leq}$, which are directions of decrease for f^k, $k \in I$. This will contradict the definition of $P^{=}$.

By the assumption, we can find a nonempty set $I \subset P^{=}\backslash\Omega$ such that

$$d \in C_P = \bigcap_{k\in P} (-\partial f^k)^*, \quad d \in D_{P^{=}\backslash I}^{=} \quad \text{but}$$

$$d \notin D_k^{=} \cup \overline{D_{P^{=}}^{=}}, \quad \text{for each } k \in I.$$

Recall that when $k_\circ \in P^{=}$, then $f^{k_\circ}$ is 'badly behaved' at x if the system

$$\begin{cases} \nabla f^{k_\circ}(x;d) = 0 \\ \nabla f^{k}(x;d) \leq 0, \ k \in P(x)\backslash k_\circ \\ d \notin D_{k_\circ}^{=} \cup \overline{D_{P^{=}}^{=}} \end{cases}$$

is consistent. Therefore, since

$$I \subset P^{=}\backslash\Omega \subset P^{=}\backslash P^{b},$$

we see that

$$\nabla f^k(x;d) < 0, \quad \text{for all } k \in I,$$

i.e.

$$d \in D_I^{<} \cap D_{P^{=}\backslash I}^{=}. \tag{2.6}$$

Let $\hat{d}$ and d_λ be defined as in (2.3) and (2.4) respectively. Then, by (2.3), (2.4) and (2.6),

(2.7) $$d_\lambda \in D_I^<, \quad \text{for all} \quad 0 \le \lambda < 1.$$

Furthermore, Lemma II.3.1(b) implies that

(2.8) $$d_\lambda \in D_{P^=\setminus I}^{\le}, \quad \text{for all} \quad 0 < \lambda < 1.$$

Now, by continuity and (2.3), there exists $0 < \beta < 1$ such that

(2.9) $$d_\lambda \in D_P^<, \quad \text{for all} \quad \beta \le \lambda < 1.$$

From (2.7), (2.8) and (2.9), we conclude that

$$d_\lambda \in D_P^{\le} \cap D_I^<, \quad \text{for all} \quad \beta \le \lambda < 1,$$

contradiction. Thus we have shown that

$$D_\Omega^= \cap C_P \subset \overline{D_{P^=}^=} \cap C_P.$$

The inclusion (2.5) follows, since both C_P and $\overline{D_{P^=}^=}$ are closed.

(iii) By a similar argument, in particular employing Lemma II.3.1(b) again, we see that

(2.10) $$\mathit{conv}\ D_\Omega^= \cap C_P \subset D_\Omega^= \cap C_P.$$

The desired result now follows from (2.1), (2.5), (2.10) and

$$\overline{D_{P^=}^=} \cap C_P \subset \overline{D_\Omega^=} \cap C_P \subset \overline{\mathit{conv}}\ D_\Omega^= \cap C_P.$$

c) By (a), (b) and (2.1), it is sufficient to show that

$$\overline{D_P^{\leq}} = \overline{conv\ D_P^{=} \cap C_P}.$$

That

$$\overline{D_P^{\leq}} \subset \overline{conv\ D_P^{=} \cap C_P} \tag{2.11}$$

is clear from the definitions and Lemma II.3.1(b). To prove the converse, we first show that

$$conv\ D_P^{=} \cap C_P \subset \overline{D_P^{\leq}}. \tag{2.12}$$

Suppose that we are given

$$d \in conv\ D_P^{=} \cap C_P$$

and the neighbourhood of the origin, η. We need to show that

$$D_P^{\leq} \cap (\eta + d) \neq \emptyset. \tag{2.13}$$

Let $\hat{d}$ and d_λ be defined as in (2.3) and (2.4) respectively. Then

$$d_\lambda \cdot \phi = \lambda \hat{d} \cdot \phi + (1 - \lambda) d \cdot \phi < 0,$$

for all $0 < \lambda \leq 1$ and all $\phi \in \bigcup_{k \in P^<} \partial f^k$. Therefore

$$d_\lambda \in D_P^{<} \cap conv\ D_P^{=} \subset D_P^{\leq}, \quad \text{for all } 0 < \lambda \leq 1.$$

Furthermore, $d_\lambda \in \eta + d$ for sufficiently small λ. This proves (2.13) and thus (2.12).

The desired result now follows since $\overline{D^{\leq}_{P}}$ is closed.

d) Let

$$C \triangleq conv \left\{ \bigcup_{k \in P^<} \partial f^k \right\} . \tag{2.14}$$

Suppose that the intersection is not empty. Then $P^< \neq \emptyset$ and there exists

$$\phi \in C \cap \left(D^{=}_{\Omega}\right)^*$$

where $\phi = \sum_{k \in P^<} \lambda_k \phi^k$ for some $\lambda_k \geq 0$, $\Sigma \lambda_k = 1$ and $\phi^k \in \partial f^k$.

By Lemma II.6.1(a) and (b),

$$D^{=}_{\Omega} \subset \{\phi\}^* \quad \text{and} \quad -C^* \subset -\{\phi\}^*.$$

Therefore

$$D^{=}_{\Omega} \cap -C^* \subset \{\phi\}^{\perp}.$$

Let $\hat{d}$ be as in (2.3). Then

$$\hat{d} \in D^{=}_{P^=} \cap D^{<}_{P^<} \subset D^{=}_{\Omega} \cap -C^* \subset \{\phi\}^{\perp},$$

i.e. $\hat{d} \cdot \phi = 0$. But

$$\hat{d} \cdot \phi = \hat{d} \cdot \sum_{k \in P^<} \lambda_k \phi^k < 0,$$

since $\hat{d} \in D_P^{<}$, $\lambda_k \geq 0$ and $\Sigma\lambda_k = 1$. Contradiction.

e) The result follows from (c) and Lemmas II.6.1(c) and II.7.1.

f) First, note that

$$D_P^{=*} + C_P^* = D_P^{=*} + C_P^{*<}, \tag{2.15}$$

since $conv\ D_P^{=} \subset C_P^{=}$. Now, let C be as in (2.14). Then C is w^*-compact, since the subdifferentials ∂f^k are w^*-compact and $P^{<}$ is finite. Furthermore, since $0 \notin C$ by (d), we get that $B_{P<}$ is closed by Lemma II.6.2 and thus

$$-C_{P<}^* = B_{P<} = cone\ C, \quad \text{by Lemma II.7.1.}$$

The result now follows from (d), (e), (2.15) and Lemma II.6.2. □

3. Gould and Tolle Optimality Criteria

We will be interested in optimality criteria of the following type:

$x \in S$ is optimal if and only if the system

$$(G - T) \begin{cases} \phi^{\circ} + \sum\limits_{k \in P(x)} \lambda_k \phi^k \in G \\ \phi^{\circ} \in \partial f^{\circ}(x) \\ \phi^k \in \partial f^k(x),\ \lambda_k \geq 0 \end{cases}$$

is consistent,

where G is a nonempty cone in X'. Gould and Tolle have considered such optimality criteria for differentiable, not necessarily convex, functionals on R^n. They have shown that if

$$T^*(S,x) = C^*_{P(x)}(x) + G, \tag{3.1}$$

then the $(G - T)$ conditions are necessary for $x \in S$ to be optimal. (Note that the condition (3.1) depends only on the constraints and not on the objective function f°.) One obvious candidate for G is

$$T^*(S,x) \setminus C^*_{P(x)}(x) \cup \{0\}.$$

Moreover, if $G = \{0\}$ satisfies (3.1), then we get the well-known Kuhn-Tucker conditions, e.g. [36], [37]. In our setting, i.e. for the convex program (P), we can say:

Theorem 3.1. Suppose that $x \in S$ and $G \subset X'$. Then the statement:

"x is optimal for (P) if and only if the system

$$\left\{ \begin{array}{l} \phi^\circ + \sum\limits_{k \in P(x)} \lambda_k \phi^k \in G \\ \phi^\circ \in \partial f^\circ(x) \\ \phi^k \in \partial f^k(x),\ \lambda_k \geq 0 \end{array} \right. \tag{3.2}$$

is consistent",

holds, for any fixed objective function f°, if and only if

G satisfies (3.1).

Proof. We need to show that we can choose G in (3.2) if and only if

$$T^*(S,x) = - B_{P(x)}(x) + G, \tag{3.3}$$

Sufficiency: Suppose that G satisfies (3.3). By Theorem II.8.1, we know that x is optimal if and only if $\partial f^\circ(x) \cap T^*(S,x) \neq \emptyset$. By (3.3), this implies that x is optimal if and only if $\partial f^\circ(x) \cap (-B_{P(x)}(x) + G) \neq \emptyset$, i.e. if and only if (3.2) is consistent.

Necessity: We need to show that (3.3) holds. Suppose that $\phi \in T^*(S,x)$ and f° is defined by the linear functional $\phi(\cdot)$ on X. Then $\phi \in \partial f^\circ(x) \cap T^*(S,x)$ and we can conclude that x is optimal for (P), i.e. $\phi = f^\circ \in F^\circ(x)$. Therefore, by the conditions (3.2), we see that $\phi \in -B_{P(x)} + G$. Thus

$$T^*(S,x) \subset -B_{P(x)}(x) + G.$$

Conversely, let $\phi \in -B_{P(x)}(x) + G$. Then we can find $\lambda_k \geq 0$ and $\phi^k \in \partial f^k(x)$ such that

$$\phi + \sum_{k \in P(x)} \lambda_k \phi^k \in G. \tag{3.4}$$

Again we let f° be the linear functional ϕ. Then $\phi = f^\circ \in F^\circ(x)$, by (3.2). Since $\partial f^\circ(x) = \{\phi\}$, Theorem II.8.1 implies that $\phi \in T^*(S,x)$. Thus

$$-B_{P(x)}(x) + G \subset T^*(S,x).$$

□

The condition (3.1) is frequently referred to as a necessary and sufficient constraint qualification, or a weakest constraint qualification.

4. Some Choices for the Cone G

By specifying G in (3.3), we get necessary and sufficient conditions for optimality which hold without a constraint qualification. For example,

Theorem 4.1. The point $x \in S$ is optimal for (P) if and only if the system

$$\begin{cases} \phi^\circ + \sum\limits_{k \in P(x)} \lambda_k \phi^k \in \left(D^{\leq}_{P(x)}(x)\right)^* \\ \phi^\circ \in \partial f^\circ(x) \\ \phi^k \in \partial f^k(x), \quad \lambda_k \geq 0 \end{cases}$$

is consistent.

Proof. By Theorem 3.1, we need only show that $G = \left(D^{\leq}_{P(x)}(x)\right)^*$ satisfies (3.3). Now

$$T^*(S,x) = \left(D^{\leq}_{P(x)}(x)\right)^*, \quad \text{by Lemmas 2.1(a) and II.6.1(a)}$$

$$= -B_{P(x)}(x) + \left(D^{\leq}_{P(x)}(x)\right)^*, \quad \text{since } D^{\leq}_{P(x)}(x) \subset C_{P(x)}(x) \text{ and}$$

$$-B_{P(x)}(x) \subset C^*_{P(x)}(x). \quad \square$$

Other, possibly more useful candidates for G are given in the next theorem.

Theorem 4.2. Suppose that $x \in S$, the set Ω satisfies $P^b(x) \subset \Omega \subset P^=$ and both *conv* $D_\Omega^=(x)$ and $-B_{P(x)}(x) + \left(D_\Omega^=(x)\right)^*$ are closed. Then, x is optimal for (P) if and only if the system

$$(4.1) \qquad \begin{cases} \phi^\circ + \sum\limits_{k \in P(x)} \lambda_k \phi^k \in \left(D_\Omega^=(x)\right)^* \\ \phi^\circ \in \partial f^\circ(x) \\ \phi^k \in \partial f^k(x), \quad \lambda_k \geq 0 \end{cases}$$

is consistent.

Proof. The result follows immediately from Theorem 3.1 and Lemma 2.1(e). □

We have assumed that *conv* $D_\Omega^=$ and $-B_{P(x)}(x) + \left(D_\Omega^=(x)\right)^*$ are closed. (This can be considered as a kind of constraint qualification.) The sets are closed, for example, when the constraints are faithfully convex and differentiable. For then both cones in the sum are polyhedral. The following two examples show that the closure assumptions are necessary.

Example 4.1. Consider the program

$$f^\circ(x) \to \min$$

s.t.

$$f^k(x) \leq 0, \quad k \in P = \{1, 2, 3\},$$

where $x = (x_i) \in R^3$, $f^1(x) = x_1$, $f^2(x) = -x_1$, $f^3(x) = \big(\mathrm{dist}(x,K)\big)^2$ (see (II.5.2)) and K is the self-polar, 'ice-cream' cone

$$K \triangleq \{x \in R^3 : x_1 \geq 0, \quad x_1 x_2 \geq x_3^2\}.$$

Note that now

$$\nabla f^3(0,d) = \lim_{t \downarrow 0} \frac{\min_{z \in K} \|td - z\|^2}{t}$$

$$= 0, \quad \text{for all } d \in R^3.$$

Let $\bar{x} = 0$. Then $\bar{x} \in S$, $P^= = P$ while $P^b(\bar{x}) = \{3\}$. Furthermore,

$$C^*_{P(0)}(0) = \mathrm{span}\left\{\begin{pmatrix}1\\0\\0\end{pmatrix}\right\} \quad \text{and} \quad \big(D^=_{P^b(0)}(0)\big)^* = K.$$

Let us show that

$$C^*_{P(0)}(0) + \big(D^=_{P^b(0)}(0)\big)^* \quad \text{is not closed.}$$

Choose

$$k^i = \begin{pmatrix} i \\ \frac{1}{i} \\ 1 \end{pmatrix} \in K \quad \text{and} \quad \ell^i = \begin{pmatrix} -i \\ 0 \\ 0 \end{pmatrix} \in C^*_{P(0)}(0), \quad i = 1, 2, \ldots .$$

Then

$$k^i + \ell^i \to \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \notin C^*_{P(0)}(0) + \big(D^=_{P^b(0)}(0)\big)^*.$$

Example 4.2. Consider the program

$$\begin{aligned} & f^{\circ}(x) \to \min \\ \text{s.t.} \quad & f^{k}(x) \le 0, \quad k \in P = \{1,2\}, \end{aligned}$$

where $x = (x_i) \in R^2$,

$$f^1(x) = \begin{cases} (x_1^2 + x_2^2 - 1)^2 & \text{if } x_1^2 + x_2^2 - 1 \ge 0 \\ 0 & \text{otherwise,} \end{cases}$$

$f^2(x) = \operatorname{dist}(x - \bar{x}, K)$, $K = \{x \in R^2 : x_1 \ge 0,\ x_2 \ge 0\}$ and $\bar{x} = (1,0)^t$. Then $S = \{\bar{x}\}$, $P^{=} = P$ while $P^{b}(\bar{x}) = \{1\}$. Let $\Omega = \{1\}$. Then

$$D_{\Omega}^{=}(\bar{x}) = \{d \in R^2 : d_1 < 0\} \cup \{0\}$$

is not closed. Furthermore

$$T^{*}(S,\bar{x}) = R^2 \quad \text{and} \quad C^{*}_{P(\bar{x})}(\bar{x}) = K.$$

Therefore

$$\left(D_{\Omega}^{=}(\bar{x})\right)^{*} + C^{*}_{P(\bar{x})}(\bar{x}) = \{x \in R^2 : x_2 \ge 0\} \neq T^{*}(S,x).$$

This implies that (3.1) fails and that we cannot choose $\Omega = \{1\}$ in Theorem 4.2.

Remark 4.1. When the sum $C^*_{P(x)}(x) + D^{=*}_{\Omega}(x)$ is not closed, we can, however, get the following asymptotic conditions: If $P^b(x) \subset \Omega \subset P^=$ and *conv* $D^=_\Omega$ is closed, then $x \in S$ is optimal if and only if there exists nets

$$\{b^n\} \subset -B_{P(x)}(x) \quad \text{and} \quad \{d^n\} \subset \left(D^=_\Omega(x)\right)^*$$

such that

$$\lim(b^n + d^n) \in \partial f^\circ(x).$$

For more details on asymptotic conditions see e.g. Zlobec [50], [51].

Using the fact that the sum $-B_{P(x)}(x) + D^*_{P^=}(x)$ is always closed (see Lemma 2.1(f)), we get the following theorem.

Theorem 4.2. Let $x \in S$. Then x is optimal for (P) if and only if the system

$$\begin{cases} \phi^\circ + \sum\limits_{k \in P(x)} \lambda_k \phi^k \in \left(D^=_{P^=}(x)\right)^* \\ \phi^\circ \in \partial f^\circ(x) \\ \phi^k \in \partial f^k(x), \quad \lambda_k \geq 0 \end{cases}$$

is consistent.

Proof. The result follows from Theorem 3.1 and Lemma 2.1(f).

□

This result is equivalent to the following characterization of optimality given in [12]. (See also [3] and [11] for the differentiable case.)

Corollary 4.1. Let $x \in S$. Then x is optimal for (P) if and only if the system

$$\begin{cases} \phi^\circ + \sum_{k \in P^<(x)} \lambda_k \phi^k \in \left(D_P^=(x)\right)^* \\ \phi^\circ \in \partial f^\circ(x) \\ \phi^k \in \partial f^k(x), \ \lambda_k \geq 0 \end{cases}$$

is consistent.

Proof. The result follows from (2.15) and Lemma II.7.1.

□

The above optimality criteria holds without a constraint qualification. However, when Theorem 4.2 is applicable, it may prove useful to choose $G = \left(D_P^= b_{(x)}(x)\right)^*$, since this is a smaller G. One would therefore, have simpler necessary conditions to check. This question will be examined in more detail in Section IV.5.

The case when $G = \{0\}$ deserves special attention because then, the system (3.2) reduces to the Kuhn-Tucker conditions. We will study this case in the next chapter.

IV. CONSTRAINT QUALIFICATIONS AND REGULARIZATION TECHNIQUES

1. Kuhn-Tucker Points

A point $x \in S$ is called a Kuhn-Tucker (*K-T*) point for (P) if the Kuhn-Tucker (*K-T*) conditions are satisfied at x, i.e. if the system

$$(1.1) \qquad \begin{cases} \phi^\circ + \sum\limits_{k \in P(x)} \lambda_k \phi^k = 0 \\ \phi^\circ \in \partial f^\circ(x) \\ \phi^k \in \partial f^k(x), \quad \lambda_k \geq 0 \end{cases}$$

is consistent. It is well-known that if $x \in S$ is a *K-T* point, then x solves program (P). However, the converse does not always hold.

2. Regular Points and Slater's Condition

We call $x \in S$ a regular point (Lagrange regular point), if the *K-T* conditions (1.1) hold for every $f^\circ \in F^\circ(x)$, i.e. we can choose $G = \{0\}$ in (III.3.2), see e.g. [31]. A constraint qualification (CQ) is then a condition on the set of constraints which

guarantees that x is a regular point.

Slater's condition, i.e. the requirement that

" there exists $\hat{x} \in X$ with $f^k(\hat{x}) < 0$, for all $k \in P$ "

is a well-known CQ which guarantees that each point $x \in S$ is a regular point. Slater's condition is equivalent to the fact that

$$P^= = \emptyset, \tag{2.1}$$

see e.g. [11]. (This follows from Lemma II.3.1(d).) Note that Slater's condition is not equivalent to

$$int\ S \neq \emptyset,$$

see e.g. [11]. However, we can say the following.

Theorem 2.1. Suppose that, when $P^= \neq \emptyset$ there exists $k \in P^=$ such that f^k is faithfully convex. Then

(i) Slater's condition holds

if and only if

(ii) $int\ S \neq \emptyset$.

(In particular, the equivalence holds when all the constraints are faithfully convex.)

Proof. That (i) implies (ii) follows by continuity. It is now sufficient to show that (ii) implies (2.1). Suppose that (2.1) fails, i.e. there exists $k \in P^=$ and, by hypothesis, f^k is

faithfully convex. Then, $D_k^{=}(x) = D_k^{=}$ is a proper subspace of X independent of $x \in X$, see Lemma II.3.1(e). This implies that $S \subset x + D_k^{=}$, which implies that $int\ S = \emptyset$. □

For a discussion on constraint qualifications, see e.g. [4], [7], [8].

3. Weakest Constraint Qualifications

A weakest constraint qualification (WCQ) is a constraint qualification, that holds if and only if x is a regular point. In other words, it is a condition that holds if and only if x is a *K-T* point for all $f^\circ \in F^\circ(x)$. Gould and Tolle [30], [31] have shown that, in their setting (see e.g. Section III.3)

$$T^*(S,x) = C^*_{P(x)}(x) \quad \text{is a} \quad \text{WCQ}.$$

By Theorem III.3.1, we see that in our more general setting,

(3.1) $$T^*(S,x) = -B_{P(x)}(x) \quad \text{is a WCQ}.$$

Note that this requires $B_{P(x)}(x)$ be closed. In this section, we present two different WCQ's.

Theorem 3.1. Suppose that $x \in S$, $P^b(x) \subset \Omega \subset P^=$ and either *conv* $D^=_\Omega(x)$ is closed or $\Omega = P^=$. Then:

a) $P^b(x) = \emptyset$ and $B_{P(x)}(x)$ is closed is a WCQ.

b) $C_{P(x)}(x) \subset$ *conv* $D^=_\Omega(x)$ and $B_{P(x)}(x)$ is closed is a WCQ.

Proof. Suppose that (a) holds. Then, Lemma III.2.1(c) implies that

$$T(S,x) = C_{P(x)}(x) \quad \text{and} \quad B_{P(x)}(x) \quad \text{is closed.}$$

By (3.1) and Lemma II.6.1(a), the above is a WCQ. This implies that x is a regular point.

Conversely, suppose that $P^b(x) \neq \emptyset$. Recall that $k \in P^b(x)$ if $k \in P^=$ and there exists

$$d \in \left(D^>_k(x) \cap E_k(x) \cap C_{P(x)}(x)\right) \setminus \overline{D^=_{P^=}(x)}.$$

But, this implies that

$$d \notin \overline{D^=_{P^=}(x)} \cap C_{P(x)}(x)$$

$$= T(S,x), \quad \text{by Lemma III.2.1 (b) and (c).}$$

Therefore,

$$T(S,x) \neq C_{P(x)}(x),$$

which, as above, implies that x is not a regular point. That $B_{P(x)}(x)$ is not closed implies that x is not a regular point, follows from (3.1). This proves (a).

To prove (b) note that, as in the proof of (a), we need only show that: $C_{P(x)}(x) \subset \overline{conv}\ D^{=}_{\Omega}(x)$ if and only if $T(S,x) = C_{P(x)}(x)$. But this follows directly from Lemma III.2.1(b) and (c). □

Remark 3.1. Suppose that $B_{P(x)}(x)$ is closed and we can find $\hat{x} \in S$ and $\Omega \subset P$ such that $f^k(\hat{x}) < 0$, for all $k \in P\backslash\Omega$, and f^k is 'never badly behaved', for all $k \in \Omega$, i.e.

$$E_k(x) = D^{=}_k(x), \quad \text{for all } x \in S \text{ and } k \in \Omega.$$

Then, since $P^b(x) \subset P^{=}$, this implies that $P^b(x) = \emptyset$, for all $x \in S$, i.e. $x \in S$ is a regular point. Thus we see that, when checking if Slater's condition holds, we need not worry about the functions which are 'never badly behaved'. In particular, we can ignore all linear functionals.

Remark 3.2. The condition given in (b) may be easier to check computationally than the one in (a). For example, when $X = R^n$, the constraints f^k, $k \in P(x)$, are differentiable and the constraints f^k, $k \in \Omega$, are faithfully convex, then $C_{P(x)}(x)$ is a polyhedral cone while $D^{=}_{\Omega}(x) = D^{=}_{\Omega}$ is a subspace, independent of x. Furthermore, $D^{=}_{\Omega}$ can be calculated explicitly, see Section V.2.

Remark 3.3. Suppose that S contains two distinct points. Then Slater's condition is a WCQ with respect to the Fritz John optimality conditions.

Proof. The Fritz John optimality conditions state that:

$x \in S$ is optimal if and only if the system

$$\begin{cases} \sum_{k\in P(x)\cup\{0\}} \lambda_k \phi^k = 0 \\ \phi^k \in \partial f^k(x), \quad \lambda_k \geq 0, \quad \sum_{k\in P(x)\cup\{0\}} \lambda_k = 1 \end{cases}$$

is consistent. Necessity always holds. We need to show that, if S contains two distinct points (Note that when $S = \{x\}$, then x is optimal for any f° chosen, the Fritz John optimality conditions hold, but Slater's condition fails.) then the Fritz John conditions are sufficient for optimality, independent of the objective function f°, if and only if Slater's condition holds. This follows from the fact that the system

$$\begin{cases} \sum_{k\in P(x)} \lambda_k \phi^k = 0 \\ \phi^k \in \partial f^k(x), \ \lambda_k \geq 0, \quad \sum_{k\in P(x)} \lambda_k = 1 \end{cases}$$

is consistent if and only if Slater's condition fails (i.e. if and only if $P^= \neq \emptyset$), which in turn follows from Motzkin's Theorem of the alternative, see e.g. [52].

□

4. Regularization

Gould and Tolle have posed the question: "Can the program (P) be regularized by the addition of a finite number of constraints?" Augunwamba [6] has considered the nonconvex, differentiable case and has shown that one can always regularize with the addition of an infinite number of constraints. He has also given necessary and sufficient conditions to insure the number of constraints added may be finite. In this section, we show that one can always regularize (P) at $\bar{x}$, by the addition of one (possibly nondifferentiable) constraint. Furthermore, in the case of faithfully convex constraints, we can regularize (P) by the addition of a finite number of linear constraints. (In the following theorem, we assume that $B_{P(x)}(x)$ is closed.)

Theorem 4.1. Suppose that $\bar{x} \in S$, X is a Hilbert space $P^b(\bar{x}) \subset \Omega \subset P^=$ and either $conv\ D^=_\Omega(\bar{x})$ is closed or $\Omega = P^=$. Consider program (P) with the additional constraint

$$f^{m+1}(x) \stackrel{\Delta}{=} \text{dist}\left((x - \bar{x}),\ \overline{conv}\ D^=_\Omega(\bar{x})\right).$$

Then $\bar{x}$ is a regular point.

Proof. By Lemma II.5.1, f^{m+1} is not 'badly behaved' at $\bar{x}$ and therefore, $P^b(\bar{x})$ is not increased by the addition of f^{m+1}. Now, by Theorem 3.1, we need only show that

$$C_{P(\bar{x})\cup\{m+1\}}(\bar{x}) \subset \overline{conv}\ D^=_\Omega(\bar{x}).$$

But

$$C_{m+1}(\bar{x}) = \{d \in X : \nabla f^{m+1}(\bar{x};d) \leq 0\}$$
$$= \overline{conv}\ D^{=}_{\Omega}(\bar{x}), \quad \text{by (II.5.3).}$$

□

Note that the feasible set remains unchanged after the addition of f^{m+1}. For, let $\bar{S}$ denote the feasible set after the addition. Then

$$x \in \bar{S} \Leftrightarrow x \in S \text{ and } x - \bar{x} \in \overline{conv}\ D^{=}_{\Omega}(\bar{x})$$
$$\Leftrightarrow x \in S, \text{ since } \Omega \subset P^{=} \text{ and } D^{=}_{\Omega}(\bar{x}) \subset \overline{conv}\ D^{=}_{\Omega}(\bar{x}).$$

We have, therefore, regularized the point $\bar{x}$, by the addition of a 'redundant' constraint.

Theorem 4.2. Let $\bar{x} \in S$ and f^k, $k \in P^{=}$, be faithfully convex. Suppose that $B : Y \to X$ is the linear operator satisfying

$$D^{=}_{P^{=}} = R(B),$$

where Y is a *lcs*. Consider the program, in the variable $y \in Y$,

$$(P_r) \qquad \text{s.t.} \quad \begin{array}{l} f^{\circ}(\bar{x} + By) \to \min \\ f^k(\bar{x} + By) \leq 0, \ k \in P \backslash P^{=}. \end{array}$$

Then Slater's condition is satisfied for (P_r), and $y = 0$ is a feasible point of (P_r). Moreover, if y^* solves (P_r), then $\bar{x} + By^*$ solves (P).

Proof. The result follows from the characterization of optimality in [3], [11] (see also Corollary III.5.1) and the fact that $P^{=} = \emptyset$ if and only if Slater's condition holds. □

In the next chapter we will see how to calculate $D^{=}_{P^{=}}$ and how to apply the above theorem to find a feasible point and solve program (P). Note that, after the substitution, (P_r) has fewer constraints (and as we shall see, fewer variables) than (P).

5. Strongest and Weakest Optimality Conditions

In Chapter III, we presented several optimality criteria of the type:

$$(5.1)\quad \begin{cases} x \in S \text{ is optimal if and only if the system} \\ \qquad \phi^{\circ} + \sum_{k \in P(x)} \lambda_k \phi^k \in G \\ \qquad \phi^{\circ} \in \partial f^{\circ}(x) \\ \qquad \lambda_k \geq 0,\ \phi^k \in \partial f^k(x),\ k \in P(x), \\ \text{is consistent,} \end{cases}$$

where G is a closed convex cone satisfying

(5.2) $$T^*(S,x) = -B_{P(x)}(x) + G.$$

We have seen that we can always choose

(5.3) $$G = \left(D^{=}_{P^{=}}(x)\right)^* \quad \text{or} \quad \left(D^{\leq}_{P(x)}(x)\right)^*.$$

Furthermore, when

$$P^b(x) \subset \Omega \subset P^{=}$$

and the sets

$$conv\ D^{=}_{\Omega}(x) \quad \text{and} \quad -B_{P(x)}(x) + \left(D^{=}_{\Omega}(x)\right)^*$$

are closed, we can choose

(5.4) $$G = \left(D^{=}_{\Omega}(x)\right)^*.$$

Among the choices of G in (5.3) and (5.4), clearly

$$G = \left(D^{=}_{P^b(x)}(x)\right)^*$$

is the smallest.

Gould and Tolle [30] have posed the following question: "When does there exist a 'smallest' G for (5.1) ?" By a 'smallest' G we shall mean a nonempty, closed convex cone that satisfied (5.2) but which contains no proper convex subset which also satisfied (5.2). It is of interest to use the 'smallest' G, because then algorithms which use (5.1) have fewer necessary conditions to check, i.e. we have a 'tighter' theory.

From Lemma III.2.1(a), we know that

$$T^*(S,x) = \left(D^{\leq}_{P(x)}(x)\right)^*.$$

Therefore $G = \left(D^{\leq}_{P(x)}(x)\right)^*$ will always satisfy (5.2) and this G will then be the 'largest' possible. On the other hand, we have seen that, when $B_{P(x)}(x)$ is closed, $G = \{0\}$ satisfies (5.2) if and only if $P^b(x) = \emptyset$. In this case, $G = \{0\}$ is clearly a 'smallest' G. However, when $P^b(x) \neq \emptyset$, the cone

$$G = \left(D^{=}_{P^b(x)}(x)\right)^*$$

need not be a 'smallest' G. In fact a 'smallest' G need not exist (See Example 5.1 below.). The following theorem gives conditions for finding a 'smallest' G.

Theorem 5.1. Suppose that $x \in S$. If there exists a closed convex cone H such that

$$C_{P(x)} \cap H = T(S,x)$$

and

$$-B_{P(x)}(x) + H^* \text{ is closed,}$$

then

$$G = H^*$$

satisfies (5.2). Furthermore, if there exists a largest, by inclusion,

such H, then H^* is a 'smallest' G for (5.2).

Proof. Follows from Lemma II.6.1(c). □

Corollary 5.1. If $P^b(x) \subset \Omega \subset P^=$ and

$$C_{P(x)}(x) = X,$$

then

$$\left(D^=_{P=}(x)\right)^* = \left(D^=_{\Omega}(x)\right)^* = \left(D^=_{P^b(x)}(x)\right)^*,$$

and, furthermore, this cone is the unique and so 'smallest' G satisfying (5.2).

Proof. Follows from Lemma III.2.1(c). □

Corollary 5.2. If there exists a halfspace H such that

$$C_{P(x)}(x) \cap H = T(S,x)$$

and

$$-B_{P(x)}(x) + H^* \text{ is closed,}$$

then

$$G = \begin{cases} \{0\} & \text{if } P^b(x) = \emptyset \\ H^* & \text{otherwise} \end{cases}$$

is a 'smallest' G.

Example 5.1. A smallest G need not exist. Consider the program (P) with the two constraints

$$f^1(x) = x_1^2 + x_3^2 \leq 0$$

$$f^2(x) = x_1 \leq 0.$$

Then $S = \{x = (x_i) \in R^3 : x_1 = x_3 = 0\}$, $\bar{x} = (0,0,0)^t \in S$, $P^= = \{1,2\}$, $D^{\leq}_{P(\bar{x})}(\bar{x}) = D^{=}_{P^=}(\bar{x}) = S$, $T^*(S,x) = \left(D^{\leq}_{P(\bar{x})}(\bar{x})\right)^* = \{x \in R^3 : x_2 = 0\}$ and $C_{P(\bar{x})}(\bar{x}) = \{x \in R^3 : x_1 \leq 0\}$. We can now set

$$G = cone\{(1,0,\varepsilon)^t, (1,0,-\varepsilon)^t\}$$

where $\varepsilon > 0$, since

$$C^*_{P(\bar{x})}(\bar{x}) + G = \{x \in R^3 : x_2 = x_3 = 0,\ x_1 \leq 0\} + G$$

$$= T^*(S,\bar{x}).$$

However, we cannot set $\varepsilon = 0$ in G.

Example 5.2. Consider the program (P) with the single constraint

$$f^1(x) = x_1^2 + x_3^2 \leq 0.$$

Then $\bar{x} = (0,0,0)^t \in S$, $P^= = P^b(\bar{x}) = \{1\}$, $C_{P(\bar{x})}(\bar{x}) = R^3$ and

$D^{=}_{P^{=}}(\bar{x}) = \{x \in R^3 : x_1 = x_3 = 0\}$. By Corollary 5.1,

$$G = \left(D^{=}_{P^{=}}(\bar{x})\right)^* = \left(D^{=}_{P^{\leq}(\bar{x})}(\bar{x})\right)^* = \{x \in R^3 : x_2 = 0\}$$

is the unique and so the 'smallest' G.

V. THE METHOD OF REDUCTION

1. Introduction

The purpose of this chapter is to introduce a numerical method for solving finite dimensional convex programs (P), regardless of whether Slater's condition is satisfied or not. Ben-Tal and Zlobec [14] have presented a feasible directions algorithm that solves program (P) without assuming a constraint qualification. They find feasible directions d, by solving the system

$$d \cdot \nabla f^k(x) < 0, \quad k \in \{0\} \cup \Omega$$
$$d \in D_k^{=}(x), \quad k \in P(x) \setminus \Omega,$$

where Ω is some subset (possibly empty) of $P(x)$. If no solution exists, for any $\Omega \subset P(x)$, then x is an optimal point for (P). Abrams and Kerzner [3] have shown that one need only consider the single system, $\Omega = P^{=}$. They have also presented an algorithm that finds $P^{=}$. (Note that it may be computationally better to use other subsets Ω, rather than just $P^{=}$, since this allows more feasible directions to choose from.) Zoutendijk [53] has suggested that, in the absence of Slater's condition, one should solve the perturbed program

$$(P_\varepsilon) \qquad \begin{aligned} & f^0(x) \to \min \\ \text{s.t.} \quad & f^k(x) \le \varepsilon, \quad k \in \Omega \\ & f^k(x) \le 0, \quad k \in P \setminus \Omega, \end{aligned}$$

where $\varepsilon > 0$ and $\Omega = P$ or $P^{=}$.

If the feasible set is compact, this is a 'stable' perturbation of (P), i.e. if $x(\varepsilon)$ denotes a solution of (P_ε), then every cluster point of the net $\{x^*(\varepsilon)\}_{\varepsilon\downarrow 0}$ is a solution of (P). Moreover, $f^\circ(x^*(\varepsilon)) \to f^\circ(x^*)$ as $\varepsilon \to 0$, where x^* is a solution of (P), see e.g. G. Wolkowicz [48].

In this chapter, we combine the approaches in [3] and [14] with the regularization technique in Theorem IV.4.2, to formulate the method, which we call the "Method of Reduction". The method first finds a feasible point x° and, in the process of finding x°, it reduces program (P) to an equivalent program, in fewer variables and fewer constraints, for which Slater's condition is satisfied. The solution of (P) is now calculated by any method that works when Slater's condition is satisfied.

An integral part of the algorithm is finding the cone $D^=_{P^=}(x)$. This is done in Sections 2 and 3. The method of reduction is then presented in Section 4. Applications and examples follow in Sections 5 and 6.

2. Calculating the Cone of Directions of Constancy†

Recall that the cone of directions of constancy $D_f^=$ of a faithfully convex function $f : R^n \to R$ is a subspace of R^n independent of the choice of x, see Lemma II.3.1(e). We now formulate an algorithm that finds $D_f^= \cap R(A_0)$, where f is a faithfully convex function, A_0 is any specified $n \times p$ matrix and $R(A_0)$ denotes the range space of A_0. Calculation of the intersection $D_f^= \cap R(A_0)$ is useful in the situation when the intersection of two or more cones of directions of constancy is needed. If $A_0 = I$, the identity matrix, then the algorithm calculates the cone of directions of constancy of f.

The algorithm is based on the fact that D_f lies in the orthogonal complement of ϕ, for any $\phi \in \partial f(x)$. By repeatedly considering the restriction of f to this orthogonal complement, we calculate $D_f^=$.

First we present the following two lemmas.

Lemma 2.1. Suppose that $0 \neq d \in R^k$ and i_0 is the smallest positive integer such that the i_0-th component of d is nonzero, i.e. $d_{i_0} \neq 0$. Let

---

† This algorithm has been published, in the case of differentiable faithfully convex functions, in [49].

$$A = \left[\begin{array}{c|ccc} I_{(i_0-1)x(i_0-1)} & & 0 & \\ \hline & \frac{d_{i_0+1}}{d_{i_0}} & \cdots & \frac{d_k}{d_{i_o}} \\ 0 & & & \\ & -I_{(k-i_0)x(k-i_0)} & & \end{array}\right] .$$

Then $R(A) = N(d)$, where $N(d)$ denotes the null space of d.

Proof. Obvious. □

Lemma 2.2. Let $\phi \in \partial f(x)$, where $f : R^n \to R$ is a faithfully convex function and $x \in R^n$. Then

$$D_f^{=} \subset N(\phi).$$

Proof. Let $d \in D_f^{=}$. Then $\nabla f(x;d) = 0$ and Lemma II.4.1(b) implies that $\phi \cdot d \leq 0$. But $-d \in D_f^{=}$, since $D_f^{=}$ is a subspace when f is faithfully convex. Thus $\phi \cdot d = 0$.

□

Let $E_k = \{e^i : i = 1,\ldots,k\}$ denote the set of unit vectors in R^k and $A_0 \in R^{nxp}$ be given.

Algorithm A:

Initialization: Set $P_0 = A_0$ and $i = 1$.

i-th step $(1 \le i \le p)$: Find a point x in the set of $p - i + 2$ vectors $\{0\} \cup E_{p-i+1}$ such that

$$\phi P_{i-1} \neq 0, \quad \text{for some } \phi \in \partial f(P_{i-1}x). \tag{2.1}$$

Case (i): If such an x exists and $i < p$, then using Lemma 2.1, determine $A_i \in R^{(p-i+1)x(p-i)}$ such that

$$R(A_i) = N(\phi P_{i-1}). \tag{2.2}$$

Set $P_i = P_{i-1}A_i$ and proceed to step $i + 1$.

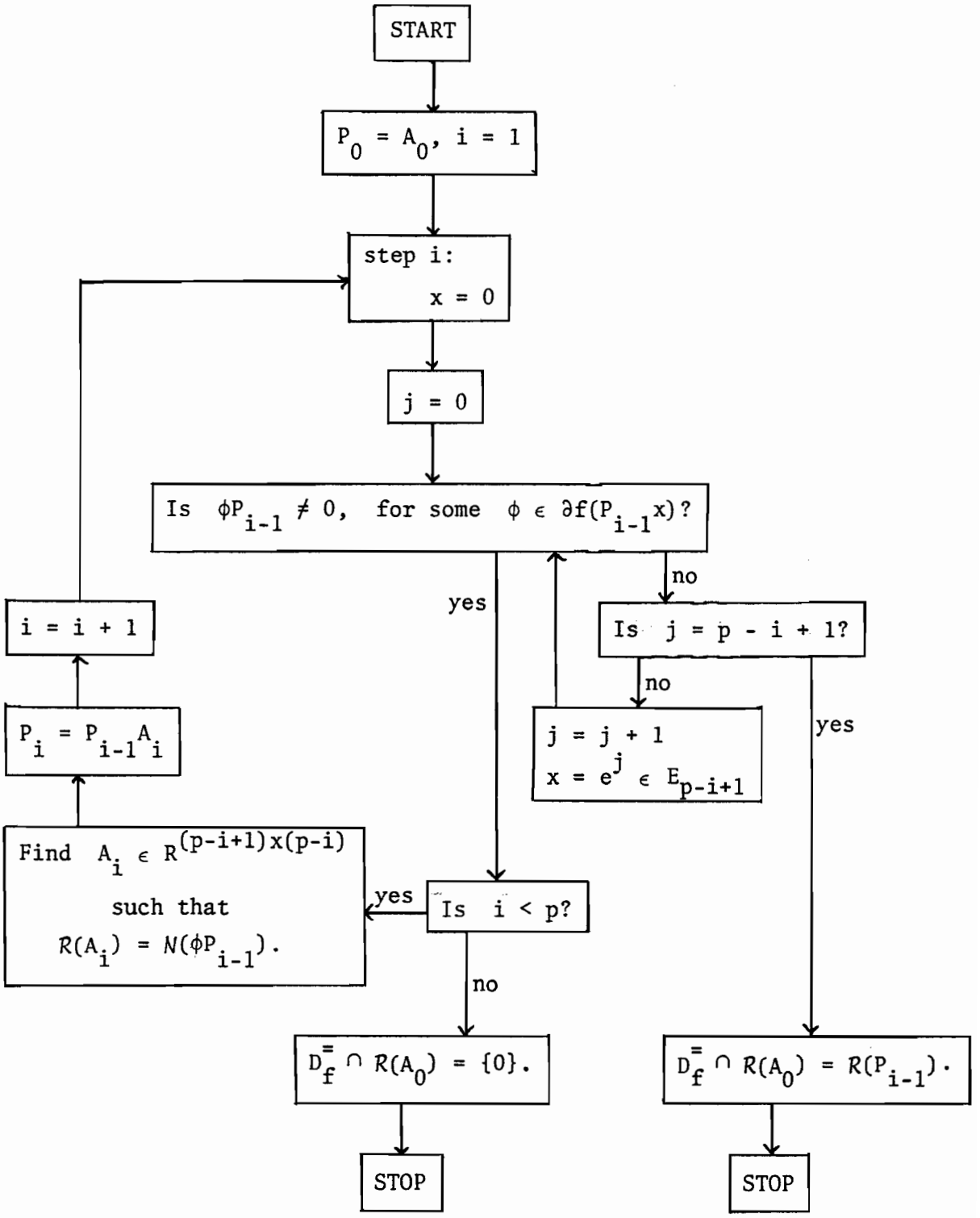
Case (ii): If such an x exists but $i = p$, then STOP.

Conclusion: $D_f^{=} \cap R(A_0) = \{0\}$.

Case (iii): If such an x does not exist, then STOP.

Conclusion: $D_f^{=} \cap R(A_0) = R(P_{i-1})$.

Theorem 2.1. Suppose that $f : R^n \to R$ is a faithfully convex function and A_0 is some given nxp matrix. Then the above algorithm finds $D_f^{=} \cap R(A_0)$ in at most $p - s + 1$ steps, where $s = \dim\left(D_f^{=} \cap R(A_0)\right)$.



Flowchart to find $D_f^{=} \cap R(A_0)$

Proof. Let x^i denote the point x which satisfies (2.1) at the i-th step and for $i \geq 0$ let $f_i = f \circ P_i$ denote the composite function formed by applying first P_i and then f. By the linearity of P_i, f_i is a faithfully convex function and so $D^{=}_{f_i}$ is a fixed subspace of R^{p-i}. Furthermore, $\partial f_i(x) = \partial f(P_i x) P_i$.

Now suppose that case (i) has occurred, i.e. $x^i \in \{0\} \cup E_{p-i+1}$, $\phi \in \partial f(P_{i-1} x^i)$, $\phi P_{i-1} \neq 0$ and $i < p$. Let us show that

$$D^{=}_{f} \cap R(A_0) = P_i D^{=}_{f_i}. \tag{2.3}$$

First, let us show that

$$D^{=}_{f} \cap R(A_0) = A_0 D^{=}_{f_0}. \tag{2.4}$$

Suppose that $d \in D^{=}_{f_0}$. This means that $f_0(\alpha d) = f_0(0)$ for all $\alpha \in R$. By definition of f_0 and the linearity of A_0, this gives $f(\alpha A_0 d) = f(0)$ for all $\alpha \in R$, i.e. $A_0 d \in D^{=}_{f}$. Furthermore, since $A_0 d \in R(A_0)$, $A_0 d \in D^{=}_{f} \cap R(A_0)$.

Conversely, suppose that $d \in D^{=}_{f} \cap R(A_0)$. Then there exists a $\bar{d} \in R^p$ such that $d = A_0 \bar{d}$ and $f(\alpha A_0 \bar{d}) = f(0)$ for all $\alpha \in R$. Again, by definition of f_0 and the linearity of A_0, we get that $f_0(\alpha \bar{d}) = f_0(0)$ for all $\alpha \in R$, i.e. $\bar{d} \in D^{=}_{f_0}$ where $d = A_0 \bar{d}$. This proves (2.4).

Next, let us show that

$$D^{=}_{f_{i-1}} = A_i D^{=}_{f_i} \quad \text{for} \quad i \geq 1. \tag{2.5}$$

Suppose that $d \in D^{=}_{f_i}$. This means that $f_i(\alpha d) = f_i(0)$ for all $\alpha \in R$. Since $f_i = f_{i-1} \circ A_i$, we get that $f_{i-1}(\alpha A_i d) = f_{i-1}(0)$ for all $\alpha \in R$, i.e. $A_i d \in D^{=}_{f_{i-1}}$.

Conversely, suppose that $d \in D^{=}_{f_{i-1}}$, i.e. $f_{i-1}(\alpha d) = f_{i-1}(0)$ for all $\alpha \in R$. By Lemma 2.2, $D^{=}_{f_{i-1}} \subset \partial f^{\perp}_{i-1}(x^i) \subset N(\phi P_{i-1})$ and $N(\phi P_{i-1}) = R(A_i)$ by (2.2). Therefore there exists a $\bar{d} \in R^{p-i}$ such that $d = A_i \bar{d}$. So $f_i(\alpha \bar{d}) = f_{i-1}(\alpha A_i \bar{d}) = f_{i-1}(\alpha d) = f_{i-1}(0) = f_i(0)$ for all $\alpha \in R$, i.e. $\bar{d} \in D^{=}_{f_i}$ and $d = A_i \bar{d}$. This proves (2.5).

By repeated substitution of (2.5) into (2.4), one gets that $D^{=}_{f} \cap R(A_0) = A_0 D^{=}_{f_0} = A_0 A_1 D^{=}_{f_1} = \ldots = P_i D^{=}_{f_i}$, which proves (2.3).

Now suppose that case (ii) has occurred, i.e. $x^i \in \{0\} \cup E_{p-i+1}$, $\phi P_{i-1} \neq 0$ but $i = p$. Since $f_{p-1} : R \to R$ is faithfully convex, we get that $D^{=}_{f_{p-1}} = \{0\}$. But, by (2.3), the $(p-1)$-st step implies that $D^{=}_{f} \cap R(A_0) = P_{p-1} D^{=}_{f_{p-1}}$. Substituting for $D^{=}_{f_{p-1}}$ yields the desired result that $D^{=}_{f} \cap R(A_0) = \{0\}$.

Finally, suppose that case (iii) has occurred, i.e. $\partial f_{i-1}(y) = \{0\}$ for all $y \in \{0\} \cup E_{p-i+1}$. Then, by the convexity of f_{i-1}, the conplete set E_{p-i+1} lies in $D^{=}_{f_{i-1}}$. But $D^{=}_{f_{i-1}}$ is a subspace

of R^{p-i+1} and so we conclude that $D_{f_{i-1}}^{=} = R^{p-i+1}$. Substituting into (2.3) yields $D_f^{=} \cap R(A_0) = R(P_{i-1})$. □

The algorithm will be illustrated by two examples.

Example 2.1. Consider the function

$$f(x) = -\left(4 + (x_1 + x_2)^2\right)^{\frac{1}{2}} + x_1 + x_2 + x_3^2.$$

This function is convex and analytic and so is faithfully convex. Let us determine its cone of directions of constancy $D_f^{=}$.

Initialization: $P_0 = A_0 = I_{3x3}$ and $i = 1$.

Step 1: Since $\nabla f(x) =$

$$\left(1 - \frac{x_1 + x_2}{\left(4 + (x_1 + x_2)^2\right)^{\frac{1}{2}}},\ 1 - \frac{x_1 + x_2}{\left(4 + (x_1 + x_2)^2\right)^{\frac{1}{2}}},\ 2x_3\right),$$

we see that $0 \in \{0\} \cup E_3$ and $\nabla f(P_0 0)P_0 = \nabla f(0) = (1, 1, 0) \neq 0$. Furthermore, since $i = 1 < p = 3$, we are in case (i). Using Lemma 2.1, we find that

$$P_1 = A_0 A_1 = A_1 = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Step 2: For $x \in R^2$, $\nabla f(P_1 x)P_1 = \nabla f\left(\begin{pmatrix} x_1 \\ -x_1 \\ -x_2 \end{pmatrix}\right)\begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} = (0, 2x_2)$.

Therefore, $\nabla f(P_1 e^2)P_1 = (0,2) \neq 0$, where $e^2 \in E_2$. Furthermore, since $i = 2 < p$ we are in case (i) again and so we find that

$$A_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad P_2 = P_1 A_2 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

Step 3: The finite point set $\{0\} \cup E_{p-i+1}$ is $\{0,1\}$ and

$\nabla f(P_2 0)P_2 = \nabla f(P_2 1)P_2 = (1,1,0)\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = 0$. Therefore, we are in case (iii) and

$$D_f^{=} = R(P_2) = \left\{ \begin{pmatrix} d \\ -d \\ 0 \end{pmatrix} \in R^3 : d \in R \right\}.$$

Example 2.2. Now consider the faithfully convex function $g(x) = -x_1 - x_2 + x_3^2$ and suppose that we wish to find $D_g \cap D_f = D_g \cap R\left(\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}\right)$, where f is the function in the previous example.

Initialization: $P_0 = A_0 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$ and $i = 1$.

Step 1: Since $p = 1$ and $\nabla g(x) = (-1,-1,2x_3)$ we see that $\{0\} \cup E_1 = \{0,1\}$ and that $\nabla g(P_0 0)P_0 = \nabla g(P_0 1)P_0 = 0$. Therefore, we are in case (iii) and

$$D_f^{=} \cap D_g^{=} = R(P_0) = \left\{ \begin{pmatrix} d \\ -d \\ 0 \end{pmatrix} \in R^3 : d \in R \right\}.$$

A computer program for the above algorithm appears in the appendix.

3. Calculating the Sets $P^{=}$ and $D^{=}_{P^{=}}$

In [3], an algorithm for calculating $P^{=}$ is given, for the program (P). We now present a modified version of this algorithm, in the case that the constraints f^k, $k \in P^{=}$, are faithfully convex. In actual fact, the algorithm finds

$$\overline{P^{=}} \quad \text{and} \quad D^{=}_{\overline{P^{=}}} \cap R(A_0),$$

in at most $t = \min\{\text{card}P(\bar{x}),\ n + 1 - \dim[S \cap (\bar{x} + R(A_0))]\}$ steps, where

$$\overline{P^{=}} = \{k \in P : f^k(x) = 0 \quad \text{for all} \quad x \in S \cap (\bar{x} + R(A_0))\}$$

and A_0 is any specified $n \times n_0$ matrix. (Recall that $D^{=}_k$ is independent of x, when the function f^k is faithfully convex.) If A_0 is specified to be the identity, then $P^{=}$ and $D^{=}_{P^{=}}$ are found. (The generalization, to find $\overline{P^{=}}$ and $D^{=}_{\overline{P^{=}}}$, will be needed in the sequel.)

The algorithm is a (finite) iterative method. We start with $P^{=}_0 = \emptyset$ and find the sets $P^{=}_{i+1} = P^{=}_i \cup J_i$ at each iteration. (The sets J_i are defined below.) The algorithm terminates when $P^{=}_i = \overline{P^{=}}$ is reached. The difference between this algorithm and the one in [3], is that, at each iteration, we discard the constraints f^k, $k \in J_i$, and, by a substitution technique, we then consider the remaining constraints as being restricted to the subspace $D^{=}_{J_i}$. In addition, when finding the set J_i, we first check if $\phi = 0$ is in the subdifferential of any of the (remaining) binding constraints. (Recall that, if $0 \in \partial f(x)$ and f is convex, then f achieves a global minimum at x.) The algorithm is demonstrated in Example 3.1 below.

Algorithm B:

Initialization: Let $\bar{x} \in S$, $\mathcal{P}_0 = \mathcal{P}(\bar{x})$, $\mathcal{P}_0^{=} = \emptyset$, $P_0 = A_0$ and $i = 0$.

i-th step $(0 \le i \le t)$: Find $k \in \mathcal{P}_i$ such that

$$0 \in \partial f^k(\bar{x}) P_i .$$

Case (i): If such a k exists, use Algorithm A to find the $n_i \times n_{i+1}$ matrix A_{i+1} satisfying

$$R(A_{i+1}) = \bigcap_{k \in J_i} D^{=}_{f^k \circ P_i} , \tag{3.1}$$

where

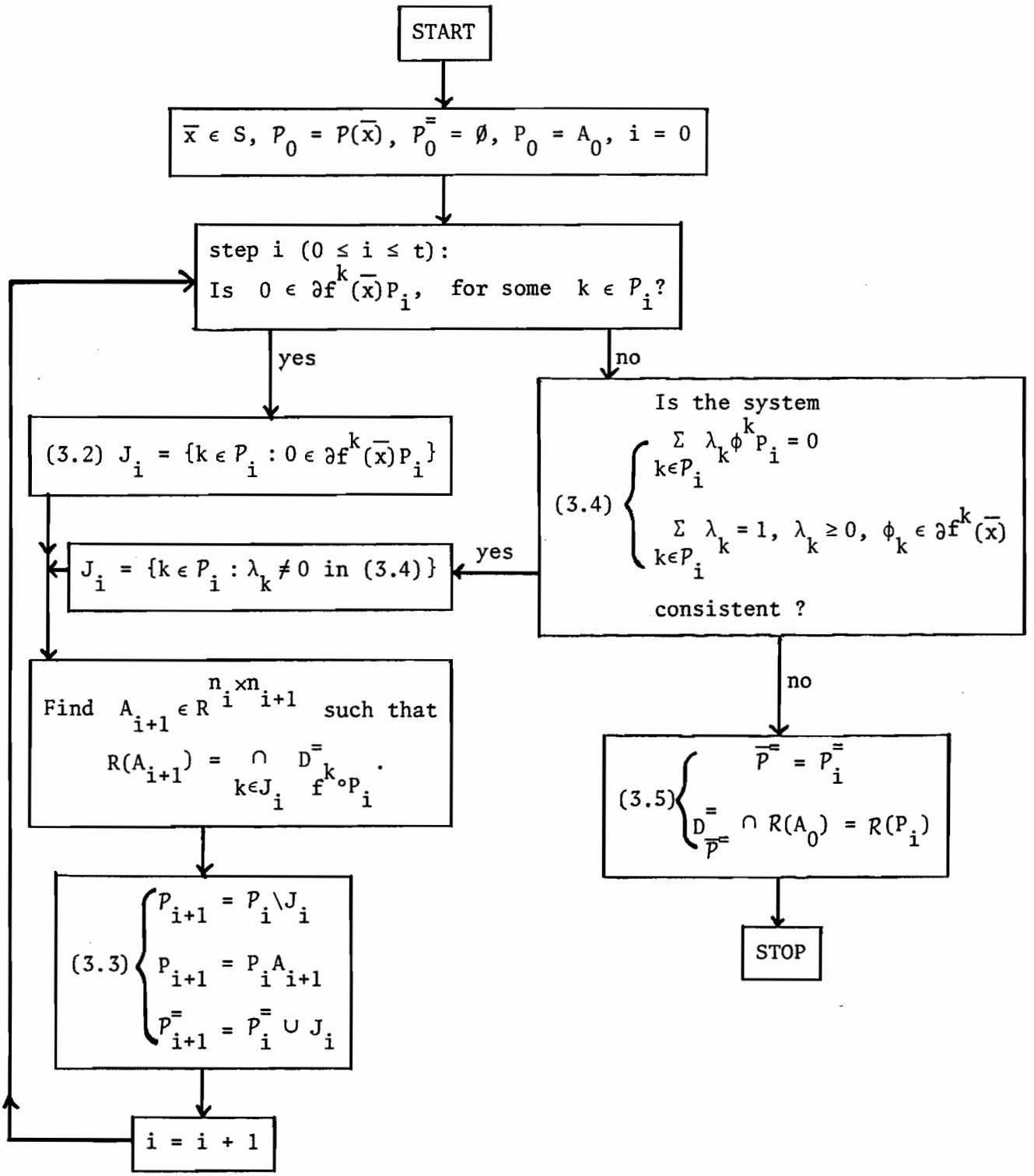
$$J_i = \{k \in \mathcal{P}_i : 0 \in \partial f^k(\bar{x}) P_i\}. \tag{3.2}$$

Then set

$$\begin{cases} \mathcal{P}_{i+1} = \mathcal{P}_i / J_i \\ P_{i+1} = P_i A_{i+1} \\ \mathcal{P}^{=}_{i+1} = \mathcal{P}^{=}_i \cup J_i \end{cases} \tag{3.3}$$

and proceed to step $i + 1$.

Case (ii): If such a k does not exist but the system



Flowchart to find $\overline{P^{=}}$ and $D^{=}_{\overline{P^{=}}} \cap R(A_0)$

$$(3.4)\qquad \begin{cases} \sum_{k\in P_i} \lambda_k \phi^k P_i = 0 \\ \sum_{k\in P_i} \lambda_k = 1,\ \lambda_k \geq 0,\ \phi^k \in \partial f^k(\bar{x}) \end{cases}$$

is consistent, then find A_{i+1}, P_{i+1}, $\dot{P}_{i+1}$, and $\bar{P}^{=}_{i+1}$ satisfying (3.1) and (3.3), where

$$J_i = \{k \in P_i : \lambda_k \neq 0 \text{ in (3.4)}\}.$$

Now proceed to step $i + 1$.

Case (iii): If such a k does not exist, but the system (3.4) is inconsistent, then STOP.

Conclusion:

$$(3.5)\qquad \begin{cases} \bar{P}^{=} = P^{=}_i \\ D^{=}_{\bar{P}^{=}} \cap R(A_0) = R(P_i)\ . \end{cases}$$

Before proving the convergence of the algorithm, let us first prove the following rather technical lemma.

Lemma 3.1. Denote

$$f^k_i(y) \triangleq f^k(\bar{x} + P_i y) \quad \text{and} \quad S_i \triangleq \{x \in R^{n_i} : f^k_i(x) \leq 0 \text{ for all } k \in P_i\}.$$

Then

$$(3.6)\qquad f^k_i(y) = f^k_{i-1}(A_i y).$$

$$D^{=}_{f_i^k} = D^{=}_{f^k \circ P_i}, \quad \text{for all } k \in P^{=}. \tag{3.7}$$

$$D^{=}_{P_i^{=}} \cap R(A_0) = R(P_i). \tag{3.8}$$

$$S_i \subset R(A_{i+1}). \tag{3.9}$$

$$S \cap (\bar{x} + R(A_0)) \subset \bar{x} + R(P_i). \tag{3.10}$$

Proof. Since

$$f_i^k(y) = f^k(\bar{x} + P_i y) = f^k(\bar{x} + P_{i-1}A_i y) = f_{i-1}^k(A_i y),$$

relation (3.6) is proved.

Now, when f^k is faithfully convex, there exists a strictly convex function g, a matrix B, vectors a and b and a constant c such that $f^k(x) = g(Ax + b) + a^t x + c$, with $D^{=}_{f^k} = N\begin{pmatrix} A \\ a^t \end{pmatrix}$, (see Remark II.3.1). Therefore

$$\begin{aligned} f_i^k(y) &= f^k(\bar{x} + P_i y) \\ &= g(A(\bar{x} + P_i y) + b) + a^t(\bar{x} + P_i y) + c \\ &= g(AP_i y + A\bar{x} + b) + a^t P_i y + a^t \bar{x} + c \end{aligned}$$

with

$$D^{=}_{f_i^k} = N\begin{pmatrix} AP_i \\ a^t P_i \end{pmatrix},$$

which equals the cone of directions of constancy of

$$f^k \circ P_i(y) = g(AP_i y + b) + a^t P_i y + c.$$

This proves (3.7).

Let us prove (3.8) by finite induction on i. The result holds for $i = 0$, since $P_0^= = \emptyset$ and $P_0 = A_0$. So, let us assume that $i \geq 1$ and that

$$D^=_{P^=_{i-1}} \cap R(A_0) = R(P_{i-1}). \tag{3.11}$$

(Note that we will consider $D^=_k$ as a subset of R^n and as a subset of R^{n_i} depending on the context, i.e. depending on whether we are considering the function f^k or f_i^k.) First, suppose that $d \in R(P_i)$, i.e.

$$d = A_0 A_1 \ldots A_i \bar{y}, \quad \text{for some } \bar{y} \in R^{n_i}.$$

This implies that $d \in R(P_{i-1}) = D^=_{P^=_{i-1}} \cap R(A_0)$, by (3.11). Now, to show that $d \in D^=_{P^=_i} \cap R(A_0)$, it is sufficient to show that $d \in D^=_{P^=_i \setminus P^=_{i-1}} = D^=_{J_{i-1}}$, by (3.3). So, let $k \in J_{i-1}$ and $\alpha \in R$. Then

$$f^k(\bar{x} + \alpha d) = f^k(\bar{x} + \alpha A_0 \ldots A_i \bar{y})$$

$$= f^k_{i-1}(\alpha A_i \bar{y}), \quad \text{by (3.6)}$$

$$= f_{i-1}^{k}(0), \quad \text{since} \quad k \in J_{i-1} \quad \text{and} \quad R(A_i) \subset D^{=}_{f_{i-1}^{k}},$$

by (3.1) and (3.7)

$$= f^{k}(\bar{x}), \quad \text{by (3.6)}.$$

This implies that $d \in D^{=}_{J_{i-1}}$. Thus we have shown that $R(P_i) \subset D^{=}_{P_i^{=}}$.

Conversely suppose that

$$d \in D^{=}_{P_i^{=}} \cap R(A_0).$$

Since $P_i^{=} \supset P_{i-1}^{=}$, (3.11) implies that $d = A_0 A_1 \cdots A_{i-1}\bar{y}$, for some $\bar{y} \in R^{n_{i-1}}$. To show that $D^{=}_{P_i^{=}} \cap R(A_0) \subset R(P_i)$, it is now sufficient to show that

$$\bar{y} = A_i \bar{z}, \quad \text{for some} \quad \bar{z} \in R^{n_i}.$$

Suppose that $k \in P_i^{=} \backslash P_{i-1}^{=} = J_{i-1}$ and $\alpha \in R$. Then

$$\begin{aligned} f_{i-1}^{k}(0) &= f^{k}(\bar{x}), \quad \text{by (3.6)} \\ &= f^{k}(\bar{x} + \alpha d), \quad \text{since} \quad d \in D^{=}_{P_i^{=}} \quad \text{and} \quad k \in P_i^{=} \\ &= f^{k}(\bar{x} + A_0 \cdots A_{i-1}(\alpha\bar{y})) \\ &= f_{i-1}^{k}(\alpha\bar{y}), \quad \text{by (3.6)}. \end{aligned}$$

This implies that $\bar{y} \in D^{=}_{J_{i-1}} = R(A_i)$, by (3.1). Thus, $\bar{y} = A_i \bar{z}$

for some $\bar{z} \in R^{n_i}$. This completes the proof of (3.8).

To prove (3.9), we consider two separate cases.

Case (a): Suppose that $0 \in \partial f_i^k(0)$, for some $k \in P_i$. (Note that $\partial f_i^k(0) = \partial f^k(\bar{x}) P_i$.) By (3.2), $0 \in \partial f_i^k(0)$ for all $k \in J_i$. Therefore, $y = 0$ is a global minimum for the convex functions f_i^k, $k \in J_i$. Now, suppose that $\bar{y} \in S_i$, i.e. $f_i^k(\bar{y}) \leq 0$ for all $k \in P_i$. Then $f_i^k(\bar{y}) = 0$ for all $k \in J_i$, since $y = 0$ is a global minimum for these functions and $f_i^k(0) = 0$. Since S_i is convex and $0 \in S_i$, we conclude that $\alpha\bar{y} \in S_i$ for all $0 \leq \alpha \leq 1$. This further implies that $f_i^k(\alpha\bar{y}) = 0$, for all $k \in J_i$ and $0 \leq \alpha \leq 1$, i.e. $\bar{y} \in D^{=}_{J_i} = R(A_{i+1})$. This proves (3.9), in case (a).

Case (b): Suppose that $0 \notin \partial f_i^k(0)$ for all $k \in P_i$. Also, assume that the system (3.4) is consistent, i.e. there exist $\lambda_k > 0$ such that

$$\sum_{k \in J_i} \lambda_k \phi^k = 0, \quad \phi^k \in \partial f_i^k(0). \tag{3.12}$$

(Note that if no such λ_k's exist, then the algorithm stops and (3.9) does not require proof.) As in case (a), we need only show that

$$\text{if } \bar{y} \in S_i \text{ and } k \in J_i, \text{ then } f_i^k(\bar{y}) = 0.$$

Suppose not. Then there exists $y \in S_i$ and $k_o \in J_i$ such that $f_i^k(y) \leq 0$, for all $k \in P_i$, and $f_i^{k_o}(y) < 0$. This implies that

$$\phi^k \cdot y \leq 0, \quad \text{for all } k \in J_i, \quad \text{and } \phi^{k_o} \cdot y < 0,$$

for all $\phi^k \in \partial f_i^k(0)$, $k \in J_i$ and $\phi^{k_o} \in \partial f_i^{k_o}(0)$. By Motzkin's theorem of the alternative [52], this contradicts (3.12). Therefore, (3.9) is proved.

Let us now prove (3.10), by finite induction on i. The result holds for $i = 0$, since $P_0 = A_0$. So, let us assume that $i \geq 0$ and

$$S \cap (\bar{x} + R(A_0)) \subset \bar{x} + R(P_{i-1}).$$

Let $x \in S \cap (\bar{x} + R(A_0))$. Then the above implies that $x = \bar{x} + P_{i-1}\bar{y}$ for some $\bar{y} \in R^{n_{i-1}}$. Thus

$$\begin{aligned} f_{i-1}^k(\bar{y}) &= f^k(x), \quad \text{by (3.6)} \\ &\leq 0, \quad \text{since } x \in S. \end{aligned}$$

Therefore, $\bar{y} \in S_{i-1}$. Now, by (3.9), $\bar{y} = A_i\bar{z}$ for some $\bar{z} \in R^{n_i}$. Substituting for $\bar{y}$ in the expression for x, implies that $x = \bar{x} + A_0 \ldots A_i\bar{z}$, which proves (3.10). (Note that the sets $\bar{x} + R(P_i)$ are decreasing linear manifolds containing the set $S \cap (\bar{x} + R(A_0))$. The algorithm essentially stops when $\bar{x} + R(P_i)$ is the smallest such linear manifold.)

□

We are now ready to prove the convergence of the algorithm. Recall that

$$\overline{P}^{=} = \{k \in P : f^k(x) = 0 \quad \text{for all} \quad x \in S \cap (\overline{x} + R(A_0))\}.$$

Theorem 3.1. Suppose that $\overline{x} \in S$, A_0 is an arbitrary $n \times n_0$ matrix and the constraints f^k, $k \in \overline{P}^{=}$ are faithfully convex. Then the above algorithm finds

$$\overline{P}^{=} \quad \text{and} \quad D^{=}_{\overline{P}^{=}} \cap R(A_0)$$

in at most $t = \min\{\text{card}P(\overline{x}),\ n_0 + 1 - \dim(S \cap (\overline{x} + R(A_0)))\}$ steps.

Proof. We need to prove that (3.5) holds when case (iii) occurs. So, suppose that $0 \notin \partial f_i^k(0)$, for all $k \in P_i$, and the system (3.4) is inconsistent. This implies that the system

$$\sum_{k \in P_i} y^k = 0,\ y^k \in \textit{cone}\ \partial f^k(\overline{x}) P_i,$$

is inconsistent. We now conclude, by the Dubovitski-Milyutin theorem of the alternative [52] and by Lemma II.7.1, that

$$\bigcap_{k \in P_i} \{y \in R^{n_i} : \phi^k P_i \cdot y < 0, \quad \text{for all} \quad \phi^k \in \partial f^k(\overline{x})\} \neq \emptyset.$$

This yields $\hat{y} \in R^{n_i}$ such that

(3.13) $$\phi^k P_i \cdot \hat{y} < 0, \quad \text{for all} \quad k \in P_i \quad \text{and} \quad \phi^k \in \partial f^k(\overline{x}).$$

Let

$$x(\alpha) = \bar{x} + P_i \alpha \hat{y}. \tag{3.14}$$

Then, (3.13) and (3.14) imply that

$$\begin{cases} f^k(x(\alpha)) < 0, & \text{for all } k \in P \backslash P(\bar{x}), \\ f^k(x(\alpha)) < 0, & \text{for all } k \in P_i, \end{cases} \tag{3.15}$$

for all $0 \le \alpha \le \bar{\alpha}$, for some $\bar{\alpha} > 0$. Furthermore, if $0 \le \alpha \le \bar{\alpha}$, then

$$\begin{aligned} f^k(x(\alpha)) &= f^k(\bar{x} + P_i \alpha \hat{y}), \quad \text{by (3.14)} \\ &= f^k(\bar{x}), \quad \text{for all } k \in P_i^=, \quad \text{by (3.8)} \\ &= 0, \quad \text{for all } k \in P_i^=, \quad \text{since } P_i^= \subset P(\bar{x}). \end{aligned} \tag{3.16}$$

Therefore, (3.15) and (3.16) imply that $x(\alpha) \in S \cap (\bar{x} + R(A_0))$ and, moreover,

$$\overline{P^=} \subset (P(\bar{x}) \backslash P_i) = P_i^=.$$

Since $D^=_{P_i^=} \cap R(A_0) = R(P_i)$ by (3.8), to prove (3.5) we have only left to show that

$$P_i^= \subset \overline{P^=}. \tag{3.17}$$

Let us prove this by finite induction on i. Now, (3.17) holds for $i = 0$, since $P_0^= = \emptyset$. Therefore, let us assume that $i \ge 1$ and

$P^{=}_{i-1} \subset \overline{P^{=}}$. Since $P^{=}_{i} = P^{=}_{i-1} \cup J_{i-1}$, by iteration, it is sufficient to show that $J_{i-1} \subset \overline{P^{=}}$. Suppose not. Then, there exists $x \in S \cap (\overline{x} + R(A_0))$ and $k_\circ \in J_{i-1}$ such that

(3.18) $\quad f^k(x) \leq 0$ for all $k \in P$ and $f^{k_\circ}(x) < 0$.

But $x = \overline{x} + A_0 \cdots A_i y$ for some $y \in R^{n_i}$, by (3.10), and

$$\begin{aligned} f^{k_\circ}(x) &= f^{k_\circ}(\overline{x} + A_0 \cdots A_i y) \\ &= f^{k_\circ}_{i-1}(A_i y), \text{ by (3.6)} \\ &= 0, \text{ since } D^{=}_{J_{i-1}} = R(A_i), \text{ by (3.1) and (3.7).} \end{aligned}$$

This contradicts (3.18). □

Example 3.1. Suppose $S \subset R^5$ is defined by the constraints

$$\begin{aligned}
f^1(x) &= e^{x_1} + x_2^2 - 1 \leq 0 \\
f^2(x) &= x_1^2 + x_2^2 + e^{-x_3} - 1 \leq 0 \\
f^3(x) &= x_1 + x_4^2 + x_5^2 - 1 \leq 0 \\
f^4(x) &= e^{-x_2} - 1 \leq 0 \\
f^5(x) &= (x_1 - 1)^2 + x_2^2 - 1 \leq 0 \\
f^6(x) &= x_1 + e^{-x_4} - 1 \leq 0 \\
f^7(x) &= x_2 + e^{-x_5} - 1 \leq 0.
\end{aligned}$$

Let us find $P^{=}$ and $D^{=}_{P^{=}}$.

Initialization: Let $\bar{x} = (0,0,1,\frac{1}{2}\sqrt{2},\frac{1}{2}\sqrt{2})$ be the chosen feasible point. Then $P_0 = A_0 = I_{5\times 5}$, $P_0 = P(\bar{x}) = \{1,3,4,5\}$ and $P_0^{=} = \emptyset$. The corresponding gradients are

$$\nabla f^1(\bar{x}) = (1,0,0,0,0)$$

$$\nabla f^3(\bar{x}) = (1,0,0,\sqrt{2},\sqrt{2})$$

$$\nabla f^4(\bar{x}) = (0,-1,0,0,0)$$

$$\nabla f^5(\bar{x}) = (-2,0,0,0,0) .$$

Step 0: Since $\nabla f^k(\bar{x})A_0 = \nabla f^k(\bar{x}) \neq 0$ for all $k \in P_0$, we solve the system given by (3.4), i.e.

$$\lambda_1 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \lambda_3 \begin{bmatrix} 1 \\ 0 \\ 0 \\ \sqrt{2} \\ \sqrt{2} \end{bmatrix} + \lambda_4 \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \lambda_5 \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

$$\lambda_1 + \lambda_3 + \lambda_4 + \lambda_5 = 1 \quad \text{and} \quad \lambda_k \geq 0.$$

A solution is $\lambda_1 = \frac{2}{3}$, $\lambda_3 = \lambda_4 = 0$ and $\lambda_5 = \frac{1}{3}$. Therefore,

$$J_0 = \{1,5\},\ P_1 = \{3,4\},\ P_1^{=} = \{1,5\},$$

$$A_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{with} \quad R(A_1) = \bigcap_{k\in J_0} D^{=}_{f^k \circ P_0} \quad \text{and} \quad P_1 = P_0 A_1 = A_1.$$

Step 1: Since $\nabla f^4(\bar{x})P_1 = 0$ while $\nabla f^3(\bar{x})P_1 \neq 0$ we get that $J_1 = \{4\}$, $P_2 = \{3\}$, $P_2^{=} = \{1,4,5\}$, $A_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ with $R(A_2) = D^{=}_{f^4 \circ P_1}$ and $P_2 = P_1 A_2 = A_1$.

Step 2: Since $P_2 = \{3\}$ and $\nabla f^3(\bar{x})P_2 \neq 0$ case (iii) occurs. STOP.

Conclusion:

$$P^{=} = P_2^{=} = \{1,4,5\}$$

and

$$D^{=}_{P^{=}} = R(P_2) = \{\begin{pmatrix} 0 \\ 0 \\ 0 \\ d_3 \\ d_4 \\ d_5 \end{pmatrix} \in R^5 : d_3, d_4, d_5 \in R\}.$$

Remark 3.1. Using the substitution technique and checking whether $\nabla f^k(\bar{x})P_i = 0$, reduces the number of computations required to find $P^{=}$ and $D^{=}_{P^{=}}$.

4. The Method of Reduction

We now collect the machinery presented in the previous two sections and formulate the method of reduction. This algorithm first finds a feasible point and then solves the general convex program (P) with faithfully convex constraints. No constraint qualification need be assumed. Let us denote by

$$S(P),$$

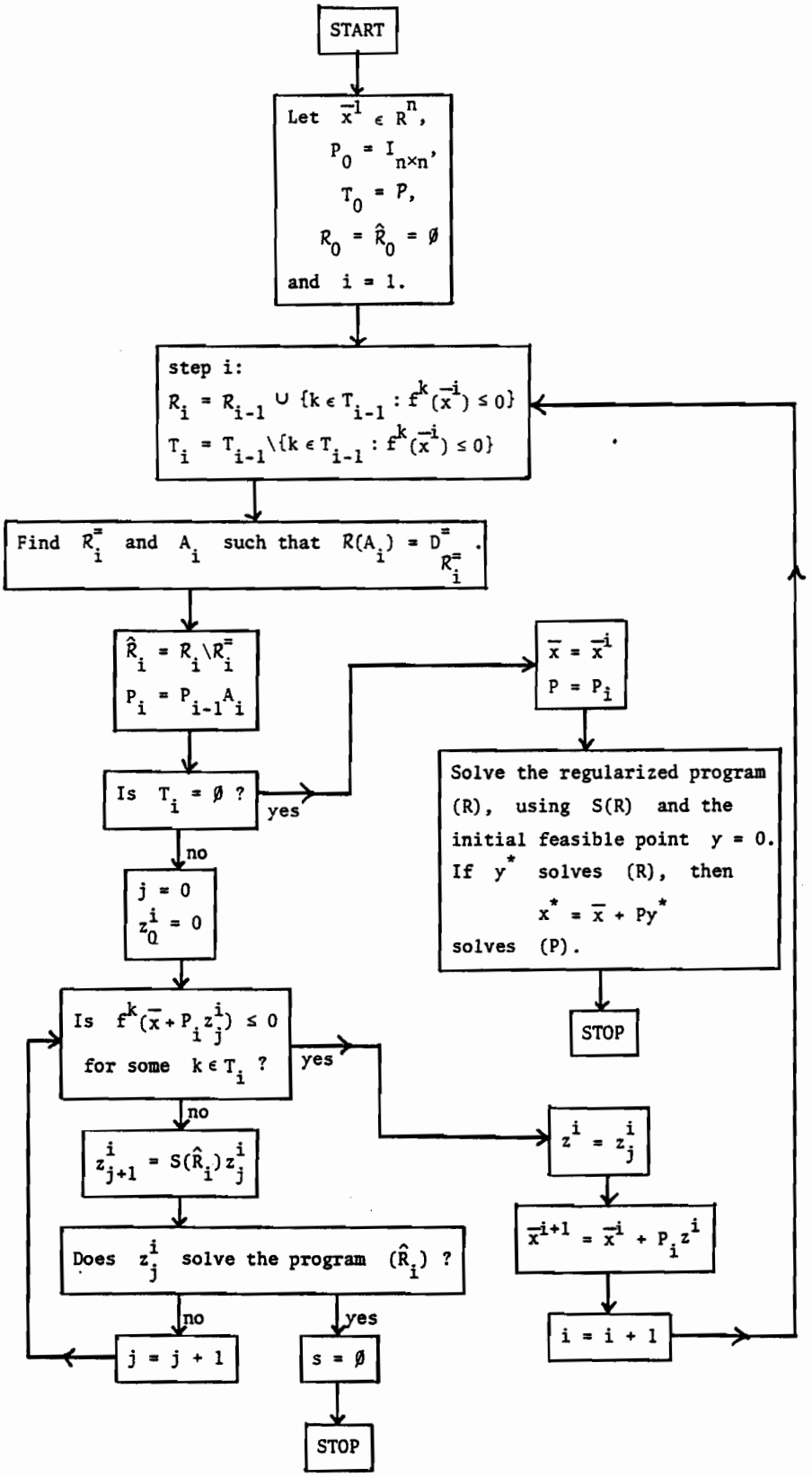
a method that solves program (P) under the assumption that Slater's condition is satisfied (e.g. Zoutendijk's feasible directions method [53], Robinson's method [52] or Powell's method [52]). The method of reduction finds the regularized program (P_r) of Theorem IV.4.2, in the process of finding a feasible point. It then solves (P_r), using $S(P_r)$. Furthermore, if Slater's condition was not satisfied for the original program (P), the regularized program (P_r) will always have fewer constraints and fewer variables than (P).

Algorithm C:

Initialization: Let $P_0 = I_{n\times n}$, $T_0 = P$, $\hat{R}_0 = \emptyset$, $R_0 = \emptyset$, $n_0 = n$ and $\bar{x}^1 \in R^n$.

i-th step $(1 \le i \le \text{card}P)$: Set

$$R_i = \hat{R}_{i-1} \cup \{k \in T_{i-1} : f^k(\bar{x}^i) \le 0\} \tag{4.1}$$



Flowchart for the Method of Reduction

and

$$T_i = T_{i-1}/R_i. \tag{4.2}$$

Now consider the program

$$(R_i) \quad \text{s.t.} \quad \begin{array}{l} \sum_{k \in T_i} f^k(\bar{x}^i + P_{i-1}y) \to \min \\ f^k(\bar{x}^i + P_{i-1}y) \le 0, \ k \in R_i, \ y \in R^{n_{i-1}}. \end{array}$$

Using the feasible point 0 and Theorem IV.4.2, regularize this program, i.e. find $R_i^=$ (Note that $R_i^=$ is the equality set for (R_i).) and the $n_{i-1} \times n_i$ matrix A_i satisfying $D^=_{R_i^=} = R(A_i)$ (Use algorithm B). Now set

$$\hat{R}_i = R_i \backslash R_i^= \tag{4.3}$$

and

$$P_i = P_{i-1}A_i \tag{4.4}$$

to get the 'reduced' program

$$(\hat{R}_i) \quad \text{s.t.} \quad \begin{array}{l} \sum_{k \in T_i} f^k(\bar{x}^i + P_i y) \to \min \\ f^k(\bar{x}^i + P_i y) \le 0, \ k \in \hat{R}_i \ \text{ and } \ y \in R^{n_i}. \end{array}$$

(Note that Slater's condition is satisfied and 0 is a feasible point, by Theorem IV.4.2.)

Case (i): Suppose that $T_i = \emptyset$. Set $\bar{x} = \bar{x}^i$ and $P = P_i$ and solve the 'reduced' program,

$$(R) \qquad \begin{aligned} & f^\circ(\bar{x} + Py) \to \min \\ & \text{s.t.} \\ & f^k(\bar{x} + Py) \le 0, \ k \in \hat{R}_i \ \text{ and } \ y \in R^{n_i}, \end{aligned}$$

using the initial feasible point $y = 0$ and $S(R)$.

Conclusion: If y^* is a solution of R, then $x^* = \bar{x} + Py^*$ is a solution of the original program (P).

Case (ii): Suppose that $T_i \neq \emptyset$. Then, set $z_0^i = 0$ and

$$z_{j+1}^i = S(\hat{R}_i) z_j^i, \quad j = 0, 1, \ldots$$

i.e. z_{j+1}^i is the point obtained after one iteration of $S(\hat{R}_i)$, applied to the point z_j^i.

Case (ii) (a): Suppose that after j iterations of $S(\hat{R}_i)$, we find $k \in T_i$ such that $f^k(\bar{x}^i + P_i z_j^i) \le 0$. Then set

$$(4.5) \qquad \begin{cases} z^i = z_j^i \\ \bar{x}^{i+1} = \bar{x}^i + P_i z^i \end{cases}$$

and proceed to step $i + 1$.

Case (ii) (b): Suppose that after j iterations of $S(\hat{R}_i)$, we have not found $k \in T_i$ such that $f^k(\bar{x}^i + P_i z_j^i) \le 0$ but z_j^i solves the program $(\hat{R}_i)$.

Conclusion: $S = \emptyset$.

Before proving convergence, we prove the following lemma.

Lemma 4.1.

$$\bigcap_{j=1}^{i} D^{=}_{R^{=}_j} = R(P_i). \tag{4.6}$$

$$f^k(\overline{x^i}) \leq 0, \quad \text{for every } k \in P \backslash T_i. \tag{4.7}$$

$$S \subset \overline{x^i} + R(P_{i-1}). \tag{4.8}$$

$$R^{=}_i \subset P^{=}. \tag{4.9}$$

Proof. Let us prove (4.6) by finite induction on i. The result holds for $i = 1$, by (4.4). So, let us assume that $i \geq 2$ and that

$$\bigcap_{j=1}^{i-1} D^{=}_{R^{=}_j} = R(P_{i-1}). \tag{4.10}$$

(Recall that we consider $D^{=}_{R^{=}_j}$ as a subset of R^n and as a subset of $R^{n_{j-1}}$, depending on the context.) Suppose that $d \in R(P_i)$, i.e. $d = P_i y$ for some $y \in R^{n_i}$. Then $d \in R(P_{i-1}) = \bigcap_{j=1}^{i-1} D^{=}_{R^{=}_j}$. But, if $k \in R^{=}_i$ and $\alpha \in R$, then

$$f^k(\overline{x^i} + \alpha d) = f^k(\overline{x^i} + P_{i-1}(\alpha A_i y))$$

$$= f^k(\overline{x^i}), \quad \text{since} \quad D^{=}_{R_i^{=}} = R(A_i).$$

Therefore, $d \in D^{=}_{R_i^{=}}$. This shows that

$$R(P_i) \subset \bigcap_{j=1}^{i} D^{=}_{R_j^{=}}.$$

Conversely, suppose that $d \in \bigcap_{j=1}^{i} D^{=}_{R_j^{=}}$. By (4.10), $d = P_{i-1}y$ for some $y \in R^{n_{i-1}}$. To show $\bigcap_{j=1}^{i} D^{=}_{R_j^{=}} \subset R(P_i)$ it is now sufficient to show that $y = A_i z$ for some $z \in R^{n_i}$. Suppose that $k \in R_i^{=}$ and $\alpha \in R$. Then

$$f^k(\overline{x^i}) = f^k(\overline{x^i} + \alpha d), \quad \text{since} \quad d \in D^{=}_{R_i^{=}}$$

$$= f^k(\overline{x^i} + P_{i-1}\alpha y).$$

This implies that $y \in D^{=}_{R_i^{=}}$. Now, since $D^{=}_{R_i^{=}} = R(A_i)$, we see that $y = A_i z$ for some $z \in R^{n_i}$. This proves (4.6).

Let us also prove (4.7) by finite induction on i. The result holds for $i = 1$ by the initialization and the definition of T_1. So, let us assume that $i \geq 2$ and

(4.11) $\qquad f^k(\overline{x^{i-1}}) \leq 0, \quad \text{for every } k \in P \backslash T_{i-1}.$

Suppose that $k \in R_i$. Then

$$f^k(\bar{x}^i) = f^k(\bar{x}^{i-1} + P_{i-1}z^{i-1})$$

$$\leq 0, \quad \text{by (4.1) and the fact that } z^{i-1} \text{ is a feasible point of the problem } \hat{R}_{i-1}.$$

On the other hand, suppose that $k \in \bigcup_{j=1}^{i-1} \bar{R}_j = (P\backslash T_i)\backslash R_i$. Then

$$f^k(\bar{x}^i) = f^k(\bar{x}^{i-1} + P_{i-1}z^{i-1})$$

$$= f^k(\bar{x}^{i-1}), \quad \text{by (4.6)}$$

$$\leq 0, \quad \text{by (4.11)}.$$

This proves (4.7).

As above, let us prove (4.8) by finite induction on i. The result holds for $i = 1$, since $P_0 = I_{n\times n}$. So, let us assume that $i \geq 2$ and

$$S \subset \bar{x}^{i-1} + R(P_{i-2}). \tag{4.12}$$

Suppose that $x \in S$. Then, $x = \bar{x}^{i-1} + P_{i-2}y$ for some $y \in R^{n_{i-2}}$, by (4.12). So,

$$f^k(x) = f^k(\bar{x}^{i-1} + P_{i-2}y)$$

$$\leq 0, \quad \text{for every } k \in P, \quad \text{since } x \in S.$$

Therefore, y is a feasible point for the program (R_{i-1}). Since

$A_{i-1}z^{i-1}$ is a feasible point for (R_{i-1}) and $D^{=}_{R^{=}_{i}} = R(A_{i-1})$, we get that $y = A_{i-1}z^{i-1} + A_{i-1}z$ for some $z \in R^{n_{i-1}}$. Substituting for y gives

$$x = \bar{x}^{i-1} + P_{i-2}(A_{i-1}z^{i-1} + A_{i-1}z)$$
$$= \bar{x}^{i} + P_{i-1}z,$$

i.e. $x \in \bar{x}^{i} + R(P_{i-1})$. This proves (4.8).

Suppose that (4.9) fails to hold. Then, there exists $k_o \in R^{=}_{i}$ such that $k_o \in P^{=}$. This implies that there exists an $\hat{x} \in R^{n}$ such that $f^{k}(\hat{x}) \leq 0$, for $k \in P$, and $f^{k_o}(\hat{x}) < 0$. But, by (4.8), we see that $\hat{x} = \bar{x}^{i-1} + P_{i-1}y$ for some $y \in R^{n_{i-1}}$. Therefore, $f^{k}(\bar{x}^{i-1} + P_{i-1}y) \leq 0$, for $k \in P$, but $f^{k_o}(\bar{x}^{i-1} + P_{i-1}y) < 0$. This contradicts the assumption that $k_o \in R^{=}_{i}$. □

Let us now prove convergence of the algorithm.

Theorem 4.1. Assume that (P) and f^{k}, $k \in P$, are such that S(P) is a convergent method when Slater's condition is satisfied. Furthermore, suppose that the constraints f^{k}, $k \in P^{=}$, are faithfully convex. Then, the method of reduction first finds a feasible point $\bar{x}$ and then solves program (P) by solving the 'reduced' program (R). (If y^{*} solves (R), then $x^{*} = \bar{x} + Py^{*}$ solves (P).)

Proof. From (4.7), we get that $y = 0$ is a feasible point for (R_i). Therefore, from Theorem IV.4.2, we get that Slater's condition is satisfied and $y = 0$ is a feasible point for the regularized program $(\hat{R}_i)$. Now, let us treat each of the cases separately.

Case (i): From (4.6), (4.7), (4.9) and the fact that Slater's condition is satisfied for (R), we see that program (R) is actually the program obtained in Theorem IV.4.2, after regularizing our original program (P). We can thus solve (R) using S(R). That $x^* = \bar{x}^i + P_i y^*$ solves (P), if y^* solves (R), follows from Theorem IV.4.2.

Case (ii) (a): In this case we just proceed to step $i + 1$.

Case (ii) (b): Suppose that $S \neq \emptyset$ and $x \in S$. Then, $x = \bar{x}^i + P_{i-1} z$ for some $z \in R^{n_{i-1}}$, by (4.8). Moreover,

$$f^k(\bar{x}^i + P_{i-1} z) \leq 0, \quad k \in P.$$

Therefore, z is a feasible point for the program (R_i) and $\sum_{k \in T_i} f^k(\bar{x}^i + P_{i-1} z) \leq 0$. But then, $z = A_i y$ for some $y \in R^{n_i}$. Substituting for z, we get that $f^k(\bar{x}^i + P_i y) \leq 0$, $k \in P$, and $\sum_{k \in T_i} f^k(\bar{x}^i + P_i y) \leq 0$. Since this implies that y is a feasible point for the program $(\hat{R}_i)$, we have contradicted the optimality of z^i_j. □

Remark 4.1. We have assumed that S(P) solves programs of type (P), when Slater's condition is satisfied. However, it may happen that the objective function and the feasible set (for one of the regularized programs $(\hat{R}_i)$) may have a common direction of recession. Such programs are called degenerate, see e.g. Abrams [2]. The infimum for degenerate programs may not be achieved or may be achieved on an unbounded set. Abrams [2] has shown how to reduce such a program to a nondegenerate program in a finite number of steps. Another possible way of handling this situation is to find a $K > 0$, large enough, and add the constraint

$$f^{m+1}(x) = \|x\|^2 - K \leq 0.$$

Such a K can be found, if a solution x^* exists, for our original program (P), and if we can approximate $\|x^*\|$. Adding the constraint f^{m+1} and choosing $\bar{x}^1$ such that $\|\bar{x}^1\| < K$, will ensure that the programs $(\hat{R}_i)$ and the final regularized program (R) are nondegenerate.

5. Applications

The method of reduction is in particular applicable to convex programming problems for which Slater's condition is not satisfied. As mentioned in [14], one class of problems for which Slater's condition is never satisfied is the class of multicriteria problems. This includes the lexicographic problem and the Pareto optimal problem. Following [15], let us first define the lexicographic problem as follows: Suppose that $f^1, \ldots, f^m$ is an ordered set of objectives. The corresponding lexicographic problem (PL) consists in choosing decisions successively subject to $x \in R^n$. The set of all optimal solutions of (PL) can then be obtained by solving the following sequence of programming problems:

$$(PL_1) \quad \begin{array}{l} \text{Determine} \\ a_1 = \min\{f^1(x) : x \in R^n\}. \end{array}$$

$$(PL_2) \quad \begin{array}{l} \text{Determine} \\ a_2 = \min\{f^2(x) : f^1(x) - a_1 \leq 0\}. \end{array}$$

$$\ldots\ldots$$

$$(PL_m) \quad \begin{array}{l} \text{Solve} \\ \min\{f^m(x) : f^k(x) - a_k \leq 0,\ k = 1,\ldots,m-1\}. \end{array}$$

Note that Slater's condition is never satisfied for the programs (PL_j). In fact,

$$\{1,\ldots,j-1\} \subset PL_j^{=}, \quad j = 2,\ldots,m,$$

where $PL_j^=$ is the equality set of (PL_j).

Now suppose that the functions f^k, $k = 1,\ldots, m - 1$, are faithfully convex. Then, if we use the method of reduction, solving (PL) reduces to solving m unconstrained problems. After finding the matrix A such that $R(A) = \bigcap_{i=1}^{j-1} D_i^=$ we see that, to solve (PL_j), we need only find

$$\min_y f^j(x^{j-1} + Ay),$$

where x^{j-1} is a solution of (PL_{j-1}). The lexicographic problem is treated in greater detail by Ben-Tal and Zlobec [15]. They provide two different methods for finding solutions.

Let us now define the Pareto optimum problem:

(PP) Find $\bar{x} \in R^n$ such that there is no other point x satisfying

$$f^k(x) \leq f^k(\bar{x}), \quad k \in P,$$

with at least one strict inequality.

The point $\bar{x}$ is then called Pareto optimal or efficient, see e.g. [9]. Note that any lexicographic solution of f^k, $k \in P$, in any order, is a Pareto optimal solution. In fact, one can say even more.

Theorem 5.1. Suppose that $\Omega_1, \Omega_2, \ldots, \Omega_r$ is any disjoint partition of the index set P and

$$g^k(\cdot) \triangleq \sum_{i \in \Omega_k} f^i(\cdot), \quad \text{for } k = 1, \ldots, r.$$

If x^* is the lexicographic solution with the ordered set of objectives $g^1, \ldots, g^r$, then x^* is a Pareto optimal solution of (PP).

Proof. Suppose not. Then there exists $x \neq x^*$, k_o and ℓ such that $k_o \in \Omega_\ell$, $f^{k_o}(x) < f^{k_o}(x^*)$ and $f^k(x) \leq f^k(x^*)$ for all $k \in P$. But then $g^k(x) \leq g^k(x^*)$ for $k = 1, \ldots, r$ and $g^\ell(x) < g^\ell(x^*)$, which contradicts the fact that x^* solves the lexicographic problem. Thus x^* is a Pareto optimal point. □

Using this theorem, we can find efficient points by solving a finite number of unconstrained optimization problems. However, not all the efficient points of (PP) can be found in this way. Charnes and Cooper [18] have given the following characterization of efficiency.

Theorem 5.2. $\bar{x}$ is a Pareto optimum for (PP) if and only if $\bar{x}$ solves the program

$$(P_{\bar{x}}) \qquad \begin{aligned} &\sum_{k \in P} f^k(x) \to \min \\ &\text{s.t.} \\ &f^k(x) - f^k(\bar{x}) \leq 0, \quad k \in P. \end{aligned}$$

This result leads to the following characterization of efficiency in terms of the equality set of $(P_{\bar{x}})$.

Theorem 5.3. $\bar{x}$ is a Pareto optimum for (PP) if and only if

$$P = P^{=}_{\bar{x}},$$

where $P^{=}_{\bar{x}}$ is the equality set of $(P_{\bar{x}})$.

Proof. The proof is immediate from the definition of Pareto optimum. □

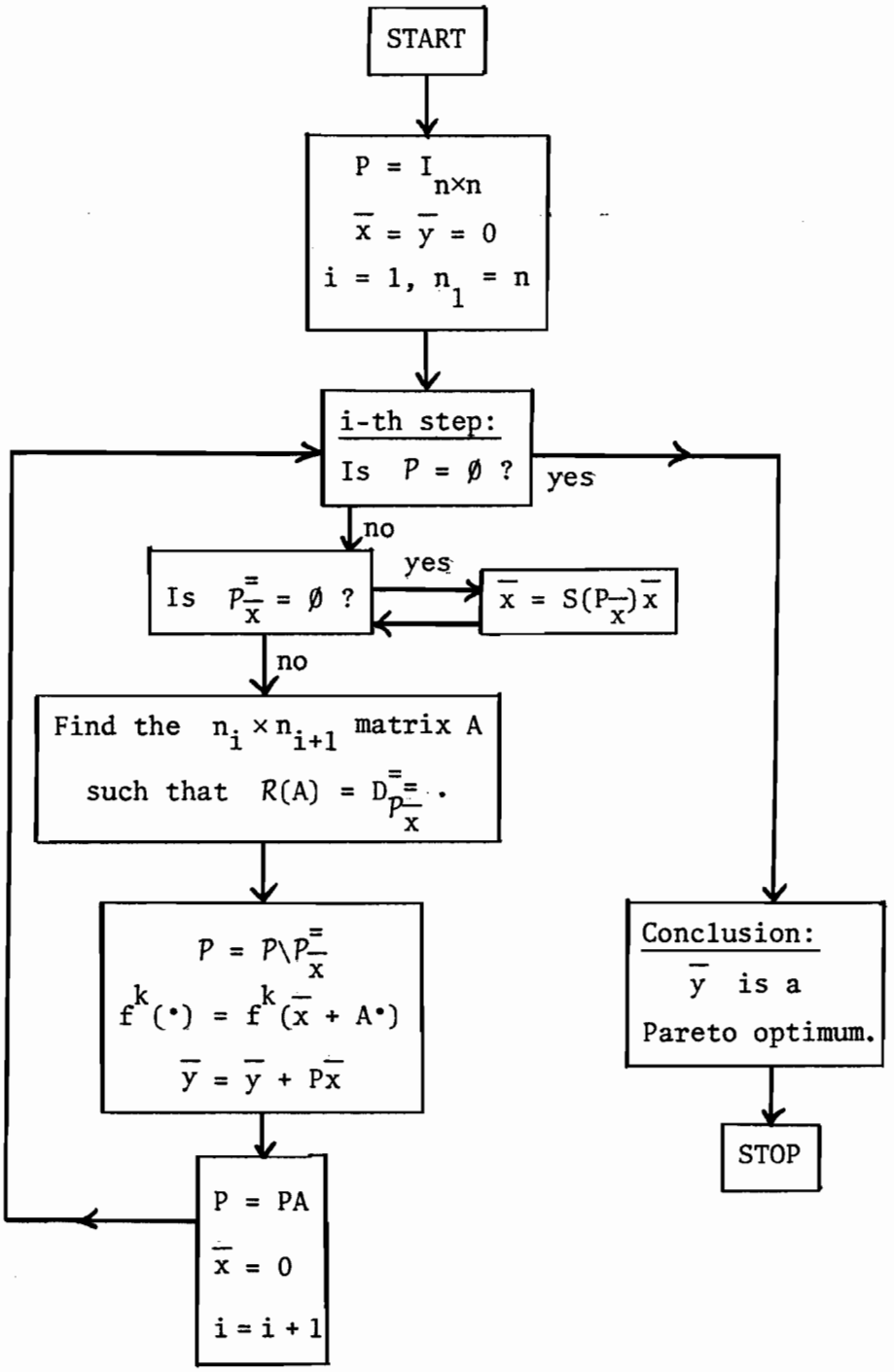
The above characterization suggests the following algorithm, when the functions f^k, $k \in P$, are faithfully convex.

Algorithm D:

Initialization: Let $P = I_{n\times n}$ and $\bar{x} = \bar{y} = 0$.

i-th step $(1 \leq i \leq \text{card}P)$: Use the feasible point $\bar{x}$ and $S(P_{\bar{x}})$ to perform one iteration for $(P_{\bar{x}})$. Redefine $\bar{x}$ to be the new point obtained, i.e. $\bar{x} \overset{\Delta}{=} S(P_{\bar{x}})(\bar{x})$. Continue until $P^{=}_{\bar{x}} \neq \emptyset$. Now, using Theorem 2.1, find the matrix A such that $R(A) = D^{=}_{P^{=}_{\bar{x}}}$ and let

$$P = P\backslash P^{=}_{\bar{x}},\quad f^k(\cdot) = f^k(\bar{x} + A\cdot)$$
$$\bar{y} = \bar{y} + P\bar{x},\ P = PA \text{ and } \bar{x} = 0.$$



Flowchart to find a Pareto optimum

Case (i): If $P \neq \emptyset$. Proceed to step $i + 1$.

Case (ii): If $P = \emptyset$, STOP.

Conclusion: $\bar{y}$ is a Pareto optimum.

Let us now consider the semi-infinite programming problem,

$$\text{(PS)} \qquad \begin{aligned} & g^{\circ}(x) \to \min \\ & \text{s.t.} \\ & g^{k}(x,t) \leq 0, \ t \in T^{k}, \ k \in P = \{1,\ldots,m\}, \ x \in R^{n}, \end{aligned}$$

where g° is convex, $g^{k}(x;t)$ is convex in x for each $t \in T^{k}$ and continuous in t for each x and T^{k} is compact in R^{ℓ}. Ben-Tal, Kerzner and Zlobec [13] have presented a characterization of optimality for (PS) which does not require a constraint qualification. Another way of treating program (PS) is by considering the convex functions

$$f^{k}(x) \overset{\Delta}{=} \sup_{t \in T^{k}} g^{k}(x,t), \quad k \in P.$$

This reduces (PS) to the form (P). We can now apply the results in the previous chapters. A third way of treating (PS) is by discretization, i.e. let T_{i}^{k}, $i = 1, 2, \ldots$ be finite subsets of T^{k} such that $\lim T_{i}^{k}$ is dense in T^{k} for each $k \in P$. Now solve the sequence of programs (of type (P))

$$(P_{i}) \qquad \begin{aligned} & g^{\circ}(x) \to \min \\ & \text{s.t.} \\ & g^{k}(x,t) \leq 0, \ t \in T_{i}^{k}, \ k \in P. \end{aligned}$$

Let X_i^* denote the set of optimal solutions (P_i) and let $x_i^* \in X_i^*$. Then every cluster point of the sequence $\{x_i^*\}$ is an optimal point of (PS). Conversely, every optimal point of (PS) is the limit of a subsequence of $\{x_i^*\}$ for some $x_i^* \in X_i^*$. Let us outline the proof: First, using the fact that a compact set-valued map is upper semi-continuous (u.s.c.) if and only if it's graph is closed (see e.g. Debreu [22]) we can show that the maps $\Omega^k : T^k \to \{x \in R^n : g^k(x,t) \leq 0\}$ are u.s.c. for each $k \in P$. This then implies that

$$\bigcap_{t \in T_i^k} \Omega^k(t) = \bigcap_{t \in T^k} \Omega^k(t), \quad k \in P,$$

and therefore $S_i \to S$ where S_i is the feasible set of (P_i) and S is the feasible set of (PS). The result now follows from Fiacco [26, Theorem 2.1].

6. Examples

To illustrate the method of reduction we consider the following three examples, which were solved using the computer program given in the appendix.

Example 6.1. Consider the program

$$
\begin{aligned}
(P) \qquad & f^0(x) = x_1 + x_2 + x_3 \to \min \\
& \text{s.t.} \\
& f^1(x) = x_1^2 + x_2^2 - 2 \le 0 \\
& f^2(x) = (x_1-2)^2 + (x_2-2)^2 - 2 \le 0 \\
& f^3(x) = e^{-x_3} - 1 \le 0.
\end{aligned}
$$

For the initial estimate, let us choose .

$$x^0 = (0,0,0)^t.$$

Since

$$f^1(x^0) < 0,\ f^3(x^0) = 0 \quad \text{while} \quad f^2(x^0) > 0,$$

the algorithm begins by considering the program

$$
\begin{aligned}
(P_1) \qquad & f^2(x) \to \min \\
& \text{s.t.} \\
& f^1(x) \le 0 \\
& f^3(x) \le 0.
\end{aligned}
$$

Slater's condition is satisfied and x^0 is a feasible starting point for (P_1). Applying Zoutendijk's method, yields the solution

$$\bar{x} = (1,1,5.65)^t.$$

Since $f^2(\bar{x}) = 0$, we can eliminate the last constraint from the objective function. We now consider the original program (P) with the feasible point $\bar{x}$. The algorithm now finds that

$$P^{=} = \{1,2\} \quad \text{and} \quad D^{=}_{P^{=}} = R(P) = R\left(\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right) .$$

After substituting P, we get the equivalent reduced program with one variable y and one constraint:

$$\begin{aligned} & f^{\circ}(\bar{x} + Py) = \bar{x} + y \to \min \\ \text{s.t.} \quad & \\ & f^{3}(\bar{x} + Py) = e^{-(5.65+y)} - 1 \leq 0 \\ & y \in R. \end{aligned}$$

Zoutendijk's method yields the solution

$$y^{*} = -5.65$$

and thus

$$x^{*} = \bar{x} + Py^{*} = (1,1,0)^{t}$$

is the solution of our original program (P). The above problem, when solved by the computer program in the appendix, gave the solution correct to 8 decimal places.

Example 6.2. Consider the program

$$
\begin{aligned}
(P)\quad \text{s.t.}\quad & f^0(x) = x_1 - x_2 + (x_3-1)^2 + (x_4-2)^2 + (x_5-2)^2 \to \min \\
& f^1(x) = e^{x_1} + x_2^2 - 1 \le 0 \\
& f^2(x) = x_1^2 + x_2^2 + e^{-x_3} - 1 \le 0 \\
& f^3(x) = x_1 + x_4^2 + x_5^2 - 1 \le 0 \\
& f^4(x) = e^{-x_2} - 1 \le 0 \\
& f^5(x) = (x_1-1)^2 + x_2^2 - 1 \le 0 \\
& f^6(x) = x_1 + e^{-x_4} - 1 \le 0 \\
& f^7(x) = x_2 + e^{-x_5} - 1 \le 0 .
\end{aligned}
$$

Starting with the initial (not feasible) point

$$x^0 = (2,2,2,2,2)^t,$$

the algorithm finds the feasible point

$$\bar{x} = (0,0,1.95,.8095,.5871)^t,$$

and

$$P^= = \{1,4,5\}; \quad D^=_{P^=} = R(P) = R\left(\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right) .$$

Using Zoutendijk's method, applied to the reduced program

$$f^{\circ}(\bar{x} + Py) = (.95 + y_1)^2 + (y_2 - 1.1905)^2 + (y_3 - 1.4129)^2 \to \min$$

s.t.

$$f^{2}(\bar{x} + Py) = e^{-1.95+y_1} - 1 \le 0$$

$$f^{3}(\bar{x} + Py) = (.8095 + y_2)^2 + (.5871 + y_3)^2 - 1 \le 0$$

$$f^{6}(\bar{x} + Py) = e^{-(.8095+y_2)} - 1 \le 0$$

$$f^{7}(\bar{x} + Py) = e^{-(.5871+y_3)} - 1 \le 0$$

we get that

$$x^* = (0,0,1,.707,.707)^t; \quad \text{and} \quad f(x^*) = 3.343146$$

is the solution of program (P). The computer program found the solution correct to 7 decimal places. (The above problem was also solved in [14] using the feasible direction methods MELP1 and MELP2.)

Example 6.3. Our last example is the program

$$f^{\circ}(x) = 10\sqrt{2}x_1 + (12 - 2\sqrt{2})x_2 + 10x_3 - 2x_4 - (12 + 2\sqrt{2})x_5 + 2x_1^2 + 2x_2^2 + 3x_3^2 + 3x_4^2 + 2x_5^2 + 2\sqrt{2}x_1x_3 - 2\sqrt{2}x_1x_4 - 4x_2x_5 + 2x_3x_4 \to \min$$

(P) s.t.

$$f^{1}(x) = 2e^{\frac{1}{2}(\sqrt{2}x_1 - x_3 + x_4)} + x_2^2 + x_5^2 + 2x_2x_5 - 2 \le 0$$

$$f^{2}(x) = 2x_1^2 + 2x_2^2 + x_3^2 + x_4^2 + 2x_5^2 - 2\sqrt{2}x_1x_3 + 2\sqrt{2}x_1x_4 + 4x_2x_5 - 2x_3x_4 + 4e^{\frac{1}{2}(x_2 - x_3 - x_4 - x_5)} - 4 \le 0$$

$$f^3(x) = 2\sqrt{2}x_1 - 2x_3 + 2x_4 + 2x_1^2 + x_2^2 + 2x_3^2 + 2x_4^2 + x_5^2$$
$$+ 2\sqrt{2}x_1x_3 - 2\sqrt{2}x_1x_4 + 2x_2x_3 + 2x_2x_4 - 2x_2x_5$$
$$- 2x_3x_5 - 2x_4x_5 - 4 \le 0$$

$$f^4(x) = x^{-\frac{1}{2}(\sqrt{2}x_2+\sqrt{2}x_5)} - 1 \le 0$$

$$f^5(x) = -4\sqrt{2}x_1 + 4x_3 - 4x_4 + 2x_1^2 + 2x_2^2 + x_3^2 + x_4^2 + 2x_5^2$$
$$- 2\sqrt{2}x_1x_3 + 2\sqrt{2}x_1x_4 + 4x_2x_5 - 2x_3x_4 \le 0$$

$$f^6(x) = \sqrt{2}x_1 - x_3 + x_4 + 2e^{\frac{1}{2}(\sqrt{2}x_1+x_3-x_4)} - 2 \le 0$$

$$f^7(x) = x_2 + x_5 + \sqrt{2}e^{\frac{1}{2}(x_2+x_3+x_4-x_5)} - \sqrt{2} \le 0.$$

Using the initial (not feasible) starting point

$$x^\circ = (0,0,2,1,1)^t,$$

the computer program in the appendix finds the feasible point

$$\bar{x} = (-.0547,-1.4604,1.1583,1.2357,1.4605)^t,$$

and

$$P^= = \{1,4,5\};\ D^=_{P^=} = R(P) = R\left(\begin{bmatrix} .707 & -.707 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right).$$

It then finds the solution

$$x^* = (-.5,-.855,-.206,.501,.845)^t;\quad f^\circ(x^*) = -22.627.$$

The output for this program is given in the appendix.

Remark 6.1. Example 6.3 was constructed from Example 6.2 by substituting the unitary matrix

$$T = \tfrac{1}{2}\begin{bmatrix} \sqrt{2} & 0 & -1 & 1 & 0 \\ 0 & \sqrt{2} & 0 & 0 & \sqrt{2} \\ 0 & -1 & 1 & 1 & 1 \\ -\sqrt{2} & 0 & -1 & 1 & 0 \\ 0 & -1 & -1 & -1 & 1 \end{bmatrix},$$

i.e. we replaced the functions $f^k(x)$ in Example 6.2 by the functions $f^k(Tx)$, for all $k \in P$. The objective function $f^\circ(x)$ was replaced by $4f^\circ(Tx) - 36$. (Multiplying by 4 and subtracting 36 were done in order to eliminate fractions and constants.) Example 6.3 is therefore the equivalent problem obtained, from Example 6.2, after a rotation of the axis. Since an exact solution of Example 6.2 is (see[14])

$$x^* = (0,0,1,\sqrt{2}/2,\sqrt{2}/2);\ f^\circ(x^*) = 9 - 4\sqrt{2},$$

we see that an exact solution of Example 6.3 is

$$T^t x^* = \tfrac{1}{2}(-1,-1-\sqrt{2}/2,1-\sqrt{2},1,1+\sqrt{2}/2)^t.$$

The exact value of the objective function is

$$4(9 - 4\sqrt{2}) - 36.$$

Furthermore, $D^{=}_{P^{=}}$ for Example 6.3 is T^tP, where P is the matrix in Example 6.2. Example 6.3 was constructed to illustrate how the

computer program works when the cones of directions of constancy of the constraints are difficult to calculate by inspection.

Remark 6.2. The above three examples illustrate that the computer program in the appendix can be used to solve convex programs, with faithfully convex constraints, independent of Slater's condition. We have not as yet compared this algorithm with any other existing techniques. Nor have we studied the stability of the algorithm, though we have had to account for round-off error in several instances. The author hopes to study these questions in the near future.

APPENDIX

```
C
C     METHOD OF REDUCTION
C     SOLVES CONVEX PROGRAMS (P).
C     USES ZOUTENDIJK'S METHOD WITH
C     SIMPLEX METHOD AND FIBONNACCI SEARCH.
C     TAKES JAMMING INTO ACCOUNT.
C
C     THE USER NEEDS TO ADD:
C     ANALYTIC REPRESENTATIONS OF ALL THE FUNCTIONS AND GRADIENTS
C     THIS IS ADDED BETWEEN THE LINES **************
C     THE USER ADDS FOR DATA:
C        THE DIMENSION OF THE X VARIABLE - N
C        THE NUMBER OF CONSTRAINTS - PCNSTR
C        INITIAL VECTOR ESTIMATE - XBAR
C
      IMPLICIT REAL*8(A-H,O-Z)
      INTEGER   COLS,PCNSTR
      COMMON P(10,10),E(10,10),GRD(10),GRADS(10,10),P0(10),XBAR(10)
      COMMON F0,SMETH,FNK(10),REDUC,EPS1,EPS2,F0SAVE
      COMMON EPS3,EPS4,EPS5,EPS6,EPS7,EPS8,EPS9,EPS10
      COMMON N,KP,PCNSTR,JI(10),COLS,ITERS
C
C     APPROXIMATIONS OF 0 DUE TO ROUNDOFF:
C     EPS1  - CONSTRAINT SATISFIED
C     EPS2  - BINDING (ANTIZIG.)
C     EPS3  - GRADIENT IN SUBR. CONE
C     EPS4  - GRADIENT IN SUBR. PEQUAL
C     EPS5  - CONSISTENCY OF SYSTEM IN PEQUAL
C     EPS6  - BASIS ELEMENTS IN SIMPLEX
C     EPS7  - NORM OF GRADIENT OF OBJ. FUNCTION
C     EPS8  - SOLUTION FOUND
C     EPS9  - ENSURE PIVOT>0 IN SIMPLEX
C
      EPS1=EPS2=EPS3=EPS5=EPS7=EPS9=1.D-5
      EPS4=1.D-4
      EPS8=1.D-6
      EPS6=1.D-8
      IT=0
      READ,N,PCNSTR
      KP=N
      READ,(XBAR(I),I=1,N)
      PRINT 1003
 1003 FORMAT('1','OUTPUT FOR EXAMPLE 6.3')
      PRINT 1004
 1004 FORMAT('-','INITIAL ESTIMATE')
      PRINT 1005,(XBAR(I),I=1,N)
 1005 FORMAT(' '                                ,10F10.4)
      DO 1010 K=1,PCNSTR
 1010 P0(K)=0
      DO 1020 I=1,N
      DO 1020 J=1,N
      E(I,J)=0
      P(I,J)=0
 1020 IF(I.EQ.J)E(I,J)=P(I,J)=1
      DO 1030 K=1,PCNSTR
      CALL GRDF(K,XBAR)
      DO 1030 I=1,N
 1030  GRADS(I,K)=GRD(I)
```

```
      SMETH=1
      ITERS=0
      IFEAS=0
      CALL FNEVAL(XBAR,IFEAS)
      FOSAVE =F0+1
      CALL SMTHD1
      CALL PEQUAL
 1040 CALL SMTHD
      IF(REDUC.EQ.1)GO TO 1045
      IF(SMETH.EQ.0)GO TO 1050
      IT=IT+1
      SMETH=0
      DO 1047 K=1,NCONSTR
      IF(P0(K).EQ.0.AND.FNK(K).GE.EPS1 )SMETH=1
 1047 IF(P0(K).EQ.0.).AND.FNK(K).LT.EPS1 )P0(K)=1
      IF(IT.LT.NCONSTR)GO TO 1045
      PRINT,'FEASIBLE SET IS EMPTY'
      STOP
 1045 CONTINUE
      IF(SMETH.EQ.1)GO TO 1051
      PRINT 1044
 1044 FORMAT('-','FEASIBLE POINT FOUND')
      PRINT 1046,(XBAR(I),I=1,N)
 1046 FORMAT(' '                                    ,10F10.4)
      CALL PEQUAL
      PRINT 1048, N,KP
 1048 FORMAT('-',I1,'X',I1,'MATRIX FOR THE CONE OF CONSTANCY')
      DO 1049 I=1,N
 1049 PRINT 1054,(P(I,J),J=1,KP)
 1054 FORMAT(' ',10F12.4)
      GO TO 1040
 1051 CALL PEQUAL
      GO TO 1040
 1050 PRINT 1053
 1053 FORMAT('-','SOLUTION FOUND')
      PRINT 1052 ,(XBAR(I),I=1,N)
 1052 FORMAT(' ','X IS  ',10F10.4)
      PRINT,'VALUE OF OBJECTIVE FUNCTION IS    ',F0
      PRINT 1055
 1055 FORMAT('1')
      STOP
      END


      DOUBLE PRECISION FUNCTION FN0(X)
C
C     THE STATEMENT FUNCTION
C     EVALUATES THE OBJECTIVE FUNCTION
C     THE USER MUST ADD THE LINE CONTAINING FN0
C
      IMPLICIT REAL*8(A-H,O-Z)
      DIMENSION X(10)
      REAL*4 SQRT
C     **********************************************
C     **********************************************
      FN0=10* SQRT(2.)*X(1) +(12.-2.* SQRT(2.))*X(2)+10.*X(3)-2*X(4)-(
     1+2*SQRT(2.))*X(5)-4*X(2)*X(5)+2*X(3)*X(4)+2*X(2)**2+3*X(3)**2+
     1  3*X(4)**2+2*X(5)**2 +2*X(1)**2+2*SQRT(2.)*X(1)*X(3)-2*
     1SQRT(2.)*X(1)*X(4)
C     **********************************************
```

```
C       ********************************************************
        RETURN
        END


        SUBROUTINE GRDEVO(X)
C
C       SUBROUTINE EVALUATES THE GRADIENT OF FO
C       THE USER MUST ADD THE VECTOR OF PARTIAL DERIVATIVES
C
        IMPLICIT REAL*8(A-H,O-Z)
        INTEGER   COLS,PCNSTR
        COMMON P(10,10),E(10,10),GRD(10),GRADS(10,10),PO(10),XBAR(10)
        COMMON FO,SMETH,FNK(10),REDUC,EPS1,EPS2,FOSAVE
        COMMON EPS3,EPS4,EPS5,EPS6,EPS7,EPS8,EPS9,EPS10
        COMMON N,KP,PCNSTR,JI(10),COLS,ITERS
        DIMENSION X(10)
        REAL*4 SQRT
C       ********************************************************
C       ********************************************************
        GRD(1)=10*SQRT(2.)+4*X(1)+2*SQRT(2.)*X(3)-2*SQRT(2.)*X(4)
        GRD(2)=4*X(2)-4*X(5)+(12-2*SQRT(2.))
        GRD(3)=6*X(3)+2*X(4)+10    +2*SQRT(2.)*X(1)
        GRD(4)=6*X(4)+2*X(3)-2-2*SQRT(2.)*X(1)
        GRD(5)=4*X(5)-4*X(2)-2*SQRT(2.)   -12
C       ********************************************************
C       ********************************************************
        RETURN
        END


        SUBROUTINE GRDF(K,X)
C
C       THE SUBROUTINE FINDS THE GRADIENT OF THE CONSTRAINTS.
C       THE USER MUST ADD THE VECTORS OF PARTIAL DERIVATIVES
C
        IMPLICIT REAL*8(A-H,O-Z)
        INTEGER   COLS,PCNSTR
        COMMON P(10,10),E(10,10),GRD(10),GRADS(10,10),PO(10),XBAR(10)
        COMMON FO,SMETH,FNK(10),REDUC,EPS1,EPS2,FOSAVE
        COMMON EPS3,EPS4,EPS5,EPS6,EPS7,EPS8,EPS9,EPS10
        COMMON N,KP,PCNSTR,JI(10),COLS,ITERS
        DIMENSION X(10)
        REAL*4 SQRT
        EXP(T)=DEXP(T)
        GO TO(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,
       1 23,24,25,26,27,28,29,30),K
C       ********************************************************
C       ********************************************************
    1   CONTINUE
        GRD(1)=  EXP(.5*(SQRT(2.)*X(1)-X(3)+X(4)))*SQRT(2.)
        GRD(2)=2*X(2)+2          *X(5)
        GRD(3)=EXP(.5*(SQRT(2.)*X(1)-X(3)+X(4)))   *(-1)
        GRD(4)=EXP(.5*(SQRT(2.)*X(1)-X(3)+X(4)))
        GRD(5)=2          *X(2)          +2*X(5)
        RETURN
    2   CONTINUE
        GRD(1)=4*X(1)-2*SQRT(2.)*X(3)-2*SQRT(2.)*X(4)
        GRD(2)=4*X(2)+4*X(5)                      +2*EXP(.5*(X(2)-X(3)-X(4)
       1 -X(5)))
```

```
      GRD(3)=2*X(3)-2*SQRT(2.)*X(1)-2*X(4)    -2*EXP(.5*(X(2)-X(3)-X(4)
     1 -X(5)))
      GRD(4)=2*X(4)+2*SQRT(2.)*X(1)-2*X(3)    -2*EXP(.5*(X(2)-X(3)-X(4)
     1 -X(5)))
      GRD(5)=4*X(5)+4*X(2)                    -2*EXP(.5*(X(2)-X(3)-X(4)
     1 -X(5)))
      RETURN
3     CONTINUE
      GRD(1)=2*SQRT(2.)+4*X(1)+2*SQRT(2.)*X(3)-2*SQRT(2.)*X(4)
      GRD(2)=2*X(2)+2*X(3)+2*X(4)-2*X(5)
      GRD(3)=-2+4*X(3)+2*SQRT(2.)*X(1)+2*X(2)-2*X(5)
      GRD(4)=2+4*X(4)-2*SQRT(2.)*X(1)+2*X(2)-2*X(5)
      GRD(5)=2*X(5)-2*X(2)-2*X(3)-2*X(4)
      RETURN
4     CONTINUE
      GRD(1)=0
      GRD(2)=-EXP(-.5*(SQRT(2.)*X(2)+SQRT(2.)*X(5)))/SQRT(2.)
      GRD(3)=0
      GRD(4)=0
      GRD(5)=-EXP(-.5*(SQRT(2.)*X(2)+SQRT(2.)*X(5)))/SQRT(2.)
      RETURN
5     CONTINUE
      GRD(1)=4*X(1)-2*SQRT(2.)*X(3)+2*SQRT(2.)*X(4)-4*SQRT(2.)
      GRD(2)=4*X(2)+4*X(5)
      GRD(3)=2*X(3)-2*SQRT(2.)*X(1)+4-2*X(4)
      GRD(4)=2*X(4)+2*SQRT(2.)*X(1)-2*X(3)-4
      GRD(5)=4*X(5)+4*X(2)
      RETURN
6     CONTINUE
      GRD(1)=SQRT(2.)+SQRT(2.)*EXP(.5*(SQRT(2.)*X(1)+X(3)-X(4)))
      GRD(2)=0
      GRD(3)=-1       +          EXP(.5*(SQRT(2.)*X(1)+X(3)-X(4)))
      GRD(4)= 1       -          EXP(.5*(SQRT(2.)*X(1)+X(3)-X(4)))
      GRD(5)=0
      RETURN
7     CONTINUE
      GRD(1)=0
      GRD(2)=  EXP(.5*(X(2)+X(3)+X(4)-X(5)))/SQRT(2.)    +1
      GRD(3)=  EXP(.5*(X(2)+X(3)+X(4)-X(5)))/SQRT(2.)
      GRD(4)= EXP(.5*(X(2)+X(3)+X(4)-X(5)))/SQRT(2.)
      GRD(5)=1-EXP(.5*(X(2)+X(3)+X(4)-X(5)))/SQRT(2.)
      RETURN
8     CONTINUE
      RETURN
9     CONTINUE
      RETURN
10    CONTINUE
      RETURN
11    CONTINUE
      RETURN
12    CONTINUE
      RETURN
13    CONTINUE
      RETURN
14    CONTINUE
      RETURN
15    CONTINUE
      RETURN
16    CONTINUE
      RETURN
```

```
   17   CONTINUE
        RETURN
   18   CONTINUE
        RETURN
   19   CONTINUE
        RETURN
   20   CONTINUE
        RETURN
   21   CONTINUE
        RETURN
   22   CONTINUE
        RETURN
   23   CONTINUE
        RETURN
   24   CONTINUE
        RETURN
   25   CONTINUE
        RETURN
   26   CONTINUE
        RETURN
   27   CONTINUE
        RETURN
   28   CONTINUE
        RETURN
   29   CONTINUE
        RETURN
   30   CONTINUE
        RETURN
C       ************************************************
C       ************************************************
        END


        SUBROUTINE FNEVAL(X,IFEAS)
C
C       THIS SUBROUTINE EVALUATES THE FUNCTIONS AT THE POINT X
C       THE USER MUST SUPPLY THE ANALYTIC REPRESENTATIONS OF THE FUNCTION
C
        IMPLICIT REAL*8(A-H,O-Z)
        INTEGER   COLS,PCNSTR
        COMMON P(10,10),F(10,10),GRD(10),GRADS(10,10),F0(10),XBAR(10)
        COMMON F0,SMETH,FNK(10),REDUC,EPS1,EPS2,FOSAVE
        COMMON EPS3,EPS4,EPS5,EPS6,EPS7,EPS8,EPS9,EPS10
        COMMON N,KP,PCNSTR,JI(10),COLS,ITERS
        DIMENSION X(10)
        REAL*4 SQRT
        FXP(T)=DEXP(T)
        IF(SMETH.EQ.1.OR.IFEAS.NE.0)GO TO 3001
        F0=FN0(X)
        RETURN
 3001   IF=1
        F0=0
        DO 80 K=1,PCNSTR
        IF(P0(K).LT.0)GO TO 80
        IF(IFEAS.EQ.0.AND.P0(K).NE.0)GO TO 80
        IF(IFEAS.EQ.1.AND.P0(K).NE.1.AND.P0(K).NE.2)GO TO 80
        GO TO (31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49
       1,50),K
C       ************************************************
C       ************************************************
```

```
31    CONTINUE
      F=2*EXP(.5*(SQRT(2.)*X(1)-X(3)+X(4)))+2*          X(2)*X(5)+
     1  X(2)**2+X(5)**2-2
      FNK(K)=F
      GO TO 70
32    CONTINUE
      F=4*EXP(.5*(X(2)-X(3)-X(4)-X(5)))-2*SQRT(2.)*X(1)*X(3)+2*SQRT(2.
     1  *X(1)*X(4)-2*X(3)*X(4)+4*X(2)*X(5)+2*X(1)**2+2*X(2)**2+X(3)**2+
     1  X(4)**2+2*X(5)**2-4
      FNK(K)=F
      GO TO 70
33    CONTINUE
      F=2*SQRT(2.)*X(1)-2*X(3)+2*X(4)+2*SQRT(2.)*X(1)*X(3)-2*SQRT(2.)*
     1  X(1)*X(4)+2*X(2)*X(3)+2*X(2)*X(4)-2*X(2)*X(5)-2*X(3)*X(5)-2*X(4
     1  *X(5)+2*X(1)**2+X(2)**2+2*X(3)**2+2*X(4)**2+X(5)**2-4
      FNK(K)=F
      GO TO 70
34    CONTINUE
      F=     EXP(    -.5*(SQRT(2.)*X(2)+SQRT(2.)*X(5)))-1
      FNK(K)=F
      GO TO 70
35    CONTINUE
      F=-4*SQRT(2.)*X(1)+4*X(3)-4*X(4)-2*SQRT(2.)*X(1)*X(3)+2*SQRT(2.)*
     1  X(1)*X(4)-2*X(3)*X(4)+4*X(2)*X(5)+2*X(1)**2+2*X(2)**2+X(3)
     1  **2+X(4)**2+2*X(5)**2
      FNK(K)=F
      GO TO 70
36    CONTINUE
      F=2*EXP(.5*(SQRT(2.)*X(1)+          X(3)-X(4)))+SQRT(2.)*X(1)
     1  -   X(3)+   X(4)-2
      FNK(K)=F
      GO TO 70
37    CONTINUE
      F=SQRT(2.)*EXP(.5*(X(2)+X(3)+X(4)-X(5)))+X(2)+X(5)-SQRT(2.)
      FNK(K)=F
      GO TO 70
38    CONTINUE
      FNK(K)=F
      GO TO 70
39    CONTINUE
      FNK(K)=F
      GO TO 70
40    CONTINUE
      FNK(K)=F
      GO TO 70
41    CONTINUE
      GO TO 70
42    CONTINUE
      FNK(K)=F
      GO TO 70
43    CONTINUE
      FNK(K)=F
      GO TO 70
44    CONTINUE
      FNK(K)=F
      GO TO 70
45    CONTINUE
      FNK(K)=F
      GO TO 70
46    CONTINUE
```

```
      FNK(K)=F
      GO TO 70
   47 CONTINUE
      FNK(K)=F
      GO TO 70
   48 CONTINUE
      FNK(K)=F
      GO TO 70
   49 CONTINUE
      FNK(K)=F
      GO TO 70
   50 CONTINUE
      FNK(K)=F
      GO TO 70
C     ********** ********** ********** ********** ********** *****
C     ********** ********** ********** ********** ********** *****
   70 CONTINUE
      IF(IFFEAS.EQ.0)GO TO 75
      IF(F.GE.EPS1 )IF=0
      IF(F.LT.EPS1 .AND.F.GT.-EPS2 )P0(K)=1
      IF(F.LT.-EPS2 )P0(K)=2
   75 IF(IFFEAS.EQ.1)GO TO 80
      F0=F0+F
   80 CONTINUE
      IFFEAS=IF
      RETURN
      END

      SUBROUTINE CONE
C
C     WE FIND THE RANGE SPACE OF P INTERSECT THE CONE OF CONSTANCY.
C     THE RESULT IS PUT BACK INTO P
C
      IMPLICIT REAL*8(A-H,O-Z)
      INTEGER   COLS,PCNSTR
      COMMON P(10,10),E(10,10),GRD(10),GRADS(10,10),P0(10),XBAR(10)
      COMMON F0,SMETH,FNK(10),REDUC,EPS1,EPS2,F0SAVE
      COMMON EPS3,EPS4,EPS5,EPS6,EPS7,EPS8,EPS9,EPS10
      COMMON N,KP,PCNSTR,JI(10),COLS,ITERS
      REAL*8 B(10,10),Y(10),A(10,10),X(10),TEMP(10)
C
C     IF JI(K)=0 WE SKIP THE K-TH FUNCTION
C
      DO 250 NCNSTR=1,PCNSTR
      IF(JI(NCNSTR).EQ.0)GO TO 240
  100 J=0
      DO 110 I=1,N
  110 X(I)=0
      GO TO 140
  115 DO 120 I=1,N
      X(I)=0
      DO 120 K=1,KP
  120 X(I)=X(I)+P(I,K)*E(K,J)
  140 CALL GRDF(NCNSTR,X)
      DO 150 I=1,KP
      Y(I)=0
      DO 150 K=1,N
  150 Y(I)=Y(I)+GRD(K)*P(K,I)
      DO 180 I=1,KP
```

```
180   IF(DABS(Y(I)).GT.EPS3 )GO TO 190
      IF(J.EQ.KP)GO TO 240
      J=J+1
      GO TO 115
190   CONTINUE
      IF(KP.EQ.1)GO TO 220
      KPM1=KP-1
      DO 170 K1=1,KP
      DO 170 K2=1,KPM1
      K2PL=K2+1
      A(K1,K2)=0
      IF(K1.EQ.K2.AND.K1.LT.I)A(K1,K2)=-1
      IF(K1-1.EQ.K2.AND.K1.GT.I)A(K1,K2)=-1
170   IF(K1.EQ.I.AND.K2.GE.I)A(K1,K2)=Y(K2PL)/Y(I)
      DO 200 I1=1,N
      DO 200 I2=1,KPM1
      B(I1,I2)=0
      DO 200 K=1,KP
200   B(I1,I2)=P(I1,K)*A(K,I2)+B(I1,I2)
      DO 210 I1=1,N
      DO 210 I2=1,KPM1
210   P(I1,I2)=B(I1,I2)
      DO 214 K=1,PCNSTR
      IF(PO(K).EQ.-1)GO TO 214
      DO 212 I=1,KPM1
      TEMP(I)=0
      DO 212 J=1,KP
212   TEMP(I)=TEMP(I)+GRADS(J,K)*A(J,I)
      DO 213 I=1,KPM1
213   GRADS(I,K)=TEMP(I)
214   CONTINUE
      KP=KPM1
      GO TO 100
220   CONTINUE
      DO 230 I=1,N
230   P(I,1)=0
240   CONTINUE
250   CONTINUE
      RETURN
      END


      SUBROUTINE  SMPLEX(A,C,ARTIF1,ROWS,BASIS,COLS)
C
C     THE SUBROUTINE SOLVES PROBLEMS AX=B BY THE SIMPLEX METHOD
C
      IMPLICIT REAL*8(A-H,O-Z)
      INTEGER TOLS
      INTEGER   PCNSTR,COLS,ARTIF1,COLSM1,ROWS,PIVOT(2),BASIS(11)
      COMMON P(10,10),B(10,10),GRD(10),GRADS(10,10),PO(10),XBAR(10)
      COMMON FO,SMETH,FNK(10),REDUC,EPS1,EPS2,FOSAVE
      COMMON EPS3,EPS4,EPS5,EPS6,EPS7,EPS8,EPS9,EPS10
      COMMON N,KP,PCNSTR,JI(10),TOLS,ITERS
      REAL*8 A(11,22),C(21)
      ABS(T)=DABS(T)
      COLSM1=COLS-1
      NOITER=0
      DO 406 I=1,ROWS
      K=ARTIF1+I
      BASIS(I)=K
```

```
      D  405 J=1,ROWS
405   A(J,K)=0
406   A(I,K)=1
407   ZJMIN=0
      DO 420 J=1,COLSM1
      IF(C(J).EQ.-1000.)GO TO 420
      ZJ=0
      DO 410 I=1,ROWS
410   ZJ   =ZJ   +A(I,J)*C(BASIS(I))
      IF(ZJ   -C(J).GE.ZJMIN)GO TO 420
      ZJMIN=ZJ   -C(J)
      PIVOT(2)=J
420   CONTINUE
      IF(ZJMIN.EQ.0)RETURN
      THETA0=-1
      DO 430 I=1,ROWS
      IF(A(I,PIVOT(2)).LE.0)GO TO 430
      T=A(I,COLS)/A(I,PIVOT(2))
      IF(T.LT.0.OR.(T.GT.THETA0.AND.THETA0.GE.0))GO TO 430
      PIVOT(1)=I
      THETA0=T
430   CONTINUE
      IF(THETA0.GE.0)GO TO 432
      PRINT,'ERROR1'
      STOP
432   BASIS(PIVOT(1))=PIVOT(2)
      DO 435 I=1,ROWS
      IF(I.EQ.PIVOT(1))GO TO 435
      A(I,COLS)=A(I,COLS)-THETA0*A(I,PIVOT(2))
      IF(A(I,COLS).GT.0.AND.A(I,COLS).LT.EPS9 )A(I,COLS)=0
435   CONTINUE
      A(PIVOT(1),COLS)=THETA0
      IF(A(PIVOT(1),COLS).GT.0.AND.A(PIVOT(1),COLS).LT.EPS9 )
     1 A(PIVOT(1),COLS)=0
      DO 455 J=1,COLSM1
      IF(C(J).EQ.-1000)GO TO 455
      IF(J.EQ.PIVOT(2))GO TO 455
      DO 450 I=1,ROWS
      IF(I.EQ.PIVOT(1))GO TO 450
      A(I,J)=A(I,J)-A(PIVOT(1),J)*A(I,PIVOT(2))/A(PIVOT(1),PIVOT(2))
      IF(A(I,J).GT.0.AND.A(I,J).LT.EPS9 )A(I,J)=0
450   CONTINUE
      A(PIVOT(1),J)=A(PIVOT(1),J)/A(PIVOT(1),PIVOT(2))
455   CONTINUE
      DO 460 I=1,ROWS
460   A(I,PIVOT(2))=0
      A(PIVOT(1),PIVOT(2))=1
      NOITER=NOITER +1
      IF(NOITER.LE.50)GO TO 407
      PRINT,'ERROR2'
      STOP
      END


      SUBROUTINE PEQUAL
C
C     SUBROUTINE FINDS THE EQUALITY SET
C
      IMPLICIT REAL*8(A-H,O-Z)
      INTEGER   PCNSTR,ROWS,COLS,ARTIF1,BASIS(11),ILAMDA(10)
```

```
      COMMON P(10,10),B(10,10),GRD(10),GRADS(10,10),F0(10),XBAR(10)
      COMMON F0,SMETH,ENK(10),REDUC,EPS1,EPS2,F0SAVE
      COMMON EPS3,EPS4,EPS5,EPS6,EPS7,EPS8,EPS9,EPS10
      COMMON N,KP,PCNSTR,JI(10),COLS,ITERS
      REAL*8 A(11,22),C(21)
      ABS(T)=DABS(T)
500   DO 510 K=1,PCNSTR
510   JI(K)=0
      DO 550 K=1,PCNSTR
      IF(P0(K).NE.1)GO TO 550
      DO 512 I=1,KP
512   IF(DABS(GRADS(I,K)).GT.EPS4 )GO TO 550
      JI(K)=1
      P0(K)=-1
      IF(K.EQ.PCNSTR)GO TO 630
      K1=K+1
      DO 520 I=K1,PCNSTR
      IF(P0(K).NE.1)GO TO 520
      DO 515 J=1,KP
515   IF(DABS(GRADS(J,K)).GT.EPS4 )GO TO 520
      JI(K)=1
      P0(K)=-1
520   CONTINUE
      GO TO 630
550   CONTINUE
      ARTIF1=0
      ROWS=KP+1
      DO 570 K=1,PCNSTR
      IF(P0(K).NE.1)GO TO 570
      ARTIF1=ARTIF1+1
      ILAMDA(ARTIF1)=K
      DO 560 I=1,KP
560   A(I,ARTIF1)=GRADS(I,K)
      A(ROWS,ARTIF1)=1
570   CONTINUE
      IF(ARTIF1.LE.1)RETURN
      COLS=ARTIF1+ROWS+1
      DO 580 I=1,KP
580   A(I,COLS)=0
      A(ROWS,COLS)=1
      DO 590 I=1,ARTIF1
590   C(I)=0
      IPL=ARTIF1+1
      IMINUS=COLS-1
      DO 600 I=IPL,IMINUS
600   C(I)=-1000
      CALL SMPLEX(A,C,ARTIF1,ROWS,BASIS,COLS)
      SUM=0
      DO 610 I=1,ROWS
610   SUM=SUM+A(I,COLS)*C(BASIS(I))
      IF(SUM.LT.-EPS5)RETURN
      DO 620 I=1,ROWS
      IF(ABS(A(I,COLS)).LT.EPS6    )GO TO 620
      P0(ILAMDA(BASIS(I)))=-1
      JI(ILAMDA(BASIS(I)))=1
620   CONTINUE
630   CONTINUE
      CALL CONE
      GO TO 500
      END
```

```
      SUBROUTINE SMTHD
C
C     THIS SUBROUTINE PERFORMS ITERATIONS ASSUMING SLATER'S CONDITIO
C     THE CONSTRAINTS ARE THOSE FOR WHICH PO(K).GT.0
C     THE OBJECTIVE FUNCTION IS THE SUM OF FO AND ANY CONSTRAINTS
C     ZOUTENDIJK'S METHOD IS USED
C
      IMPLICIT REAL*8(A-H,O-Z)
      INTEGER   ROWS,PCNSTR,COLS,ARTIF1,BASIS(11)
      COMMON P(10,10),E(10,10),GRD(10),GRADS(10,10),PO(10),XBAR(10)
      COMMON FO,SMETH,FNK(10),REDUC,EPS1,EPS2,FOSAVE
      COMMON EPS3,EPS4,EPS5,EPS6,EPS7,EPS8,EPS9,EPS10
      COMMON N,KP,PCNSTR,JI(10),COLS,ITERS
      DIMENSION GRDF0(10)
      DIMENSION D(10),DKP(10)
      DIMENSION XY(10),XZ(10)
      DIMENSION TEMP1(10)
      REAL*8    A(11,22),C(21)
      ABS(T)=DABS(T)
      ARTIF1=2*KP+1
      ITERS=0
  700 ROWS=0
      DSUM=0
      ITERS=ITERS+1
      REDUC=0
      DO 730 K=1,PCNSTR
      IF(PO(K).NE.1)GO TO 730
      ROWS=ROWS+1
      DO 720 J=1,KP
      J2=2*J
      J2M=2*J-1
      A(ROWS,J2M)=GRADS(J,K)
  720 A(ROWS,J2)=-A(ROWS,J2M)
  730 CONTINUE
      ROWS=ROWS+1
      DO 7301 J=1,ARTIF1
 7301 A(ROWS,J)=0
      IF(SMETH.EQ.1)GO TO 732
      FOSAVE=FNO(XBAR)
      CALL GRDEVO(XBAR)
      DO 7302 I=1,KP
      TEMP1(I)=0
      DO 7302 J=1,N
 7302 TEMP1(I)=TEMP1(I)+GRD(J)*P(J,I)
      DO 731 I=1,KP
      I2M=2*I-1
  731 A(ROWS,I2M)=TEMP1(I)
      GO TO 736
  732 IFEAS=0
      CALL FNEVAL(XBAR,IFEAS)
      FOSAVE=FO
      DO 735 K=1,PCNSTR
      IF(PO(K).NE.0)GO TO 735
      DO 733 I=1,KP
      I2M=2*I-1
  733 A(ROWS,I2M)=A(ROWS,I2M)+GRADS(I,K)
  735 CONTINUE
  736 CONTINUE
      DO 737 I=1,KP
      I2=2*I
```

```
      I2M=I2-1
      DSUM=DSUM+ ABS(A(ROWS,I2M))
737   A(ROWS,I2)=DKP(I)=-A(ROWS,I2M)
      IF(DSUM.LT.EPS7)RETURN
      IF(ROWS.GT.1)GO TO 739
      DO 738 I=1,KP
738   DKP(I)=DKP(I)/DSUM
      GO TO 845
739   COLS=ROWS+ARTIF1+2
      DO 740 I=1,ROWS
      A(I,COLS)=0
740   A(I,ARTIF1)=1
      ROWS=ROWS+1
      I2=ARTIF1-1
      DO 750 J=1,I2
      C(J)=0
750   A(ROWS,J)=1
      A(ROWS,ARTIF1)=0
      A(ROWS,COLS)=1
      DO 760 J=ARTIF1,COLS
760   C(J)=0
      C(ARTIF1)=1
      CALL SMPLEX(A,C,ARTIF1,ROWS,BASIS,COLS)
      DO 790 I=1,ROWS
790   IF(BASIS(I).EQ.ARTIF1.AND.ABS(A(I,COLS)).GT.EPS6)GO TO 810
      RETURN
810   CONTINUE
      DO 820 I=1,KP
820   DKP(I)=0
      DO 840 I=1,ROWS
      IF(BASIS(I).GT.2*KP)GO TO 840
      IDIV=BASIS(I)/2
      IF(BASIS(I).EQ.IDIV*2)GO TO 830
      IDIV=IDIV+1
      DKP(IDIV)=DKP(IDIV)+A(I,COLS)
      GO TO 840
830   DKP(IDIV)=DKP(IDIV)-A(I,COLS)
840   CONTINUE
845   DO 850 I=1,N
      D(I)=0
      DO 850 J=1,KP
850   D(I)=P(I,J)*DKP(J)+D(I)
      SUM=20
      ALPHA=20
      DO 860 I=1,N
860   XBAR(I)=XBAR(I)+ALPHA*D(I)
      DO 880 ITER=1,30
      SIGN=-1
      IFEAS=1
      CALL FNEVAL(XBAR,IFEAS)
      IF(IFEAS.EQ.1.AND.ITER.EQ.1)GO TO 885
      IF(IFEAS.EQ.1)SIGN=1
      ALPHA=ALPHA/2
      SUM=SUM+SIGN*ALPHA
      DO 880 I=1,N
880   XBAR(I)=XBAR(I)+ALPHA*D(I)*SIGN
885   CONTINUE
      IFEAS=1
      CALL FNEVAL(XBAR,IFEAS)
      IF(IFEAS.EQ.1)GO TO 900
```

```
      ALPHA=ALPHA/2
      SUM=SUM-ALPHA
      DO 890 I=1,N
  890 XBAR(I)=XBAR(I)-ALPHA*D(I)
      IFEAS=1
      CALL FNEVAL(XBAR,IFEAS)
  900 CONTINUE
C
C     NOW FIND THE MINIMUM OF F0 IN THE ONE DIRECTION
C
C
C     FIBONACCI SEARCH
C
      A0=0
      TOL=.10000
      NN=1
      N0=1
      N1=1
  960 N2=N0+N1
      IF(N2.GE.TOL)GO TO 970
      NN=NN+1
      N0=N1
      N1=N2
      GO TO 960
  970 Y=SUM*N0/N2
      Z=SUM-Y
      DO 975 I=1,N
      XY(I)=XBAR(I)-Y*D(I)
  975 XZ(I)=XBAR(I)-Z*D(I)
      IFEAS=0
      CALL FNEVAL(XY,IFEAS)
      FY=-F0
      IFEAS=0
      CALL FNEVAL(XZ,IFEAS)
      FZ=-F0
      DO 990 I=1,NN
      DIST=SUM-A0
      IF(FY.LT.FZ)XN=Z
      IF(FY.GE.FZ)XN=Y
      IF(I.EQ.NN)GO TO 995
      IF(FY.LT.FZ)GO TO 980
      SUM=Z
      Z=Y
      FZ=FY
      Y=A0+SUM-Y
      DO 977 J=1,N
  977 XY(J)=XBAR(J)-Y*D(J)
      IFEAS=0
      CALL FNEVAL(XY,IFEAS)
      FY=-F0
      GO TO 990
  980 A0=Y
      Y=Z
      FY=FZ
      Z=A0+SUM-Z
      DO 982 J=1,N
  982 XZ(J)=XBAR(J)-Z*D(J)
      IFEAS=0
      CALL FNEVAL(XZ,IFEAS)
      FZ=-F0
```

```
990   CONTINUE
995   CONTINUE
      DO 996 I=1,N
996   XBAR(I)=XBAR(I)-XM*D(I)
      IFEAS=1
      CALL FNEVAL(XBAR,IFEAS)
      IFEAS=0
      CALL FNEVAL(XBAR,IFEAS)
      IF(SMETH.EQ.0)F0=FN0(XBAR)
      DO 998 K=1,PCNSTR
      IF(P0(K).EQ.-1)GO TO 998
      CALL GRDF(K,XBAR)
      DO 997 I=1,KP
      GRADS(I,K)=0
      DO 997 J=1,N
997   GRADS(I,K)=GRADS(I,K)+GRD(J)*P(J,I)
998   CONTINUE
      IF( DABS(F0-F0SAVE).LT.EPS8)GO TO 918
      IF(ITERS.GT.50)RETURN
      IF(SMETH.EQ.0)GO TO 918
      ENTRY SMTHD1
      ARTIF1=2*KP+1
918   REDUC=0
      SMETH=0
      DO 950 K=1,PCNSTR
      IF(P0(K).LT.0)GO TO 950
      IF(ENK(K).LE.  -EPS2 )GO TO 920
      IF(ENK(K).GE.EPS1.AND.P0(K).EQ.0)GO TO 940
      IF(P0(K).EQ.0)REDUC=1
      P0(K)=1
      GO TO 950
920   IF(P0(K).EQ.0)REDUC=1
      P0(K)=2
      GO TO 950
940   IF(P0(K).EQ.0)SMETH=1
950   CONTINUE
      IF(REDUC.EQ.0.AND. DABS(F0-F0SAVE).GT.EPS8 )GO TO 700
      RETURN
      END

$DATA
```

```
OUTPUT FOR EXAMPLE 6.3

INITIAL ESTIMATE
    0.0000     0.0000     2.0000     1.0000     1.0000

FEASIBLE POINT FOUND
   -0.0547    -1.4604     1.1583     1.2357     1.4605

5X3MATRIX FOR THE CONE OF CONSTANCY
       0.7071      -0.7071       0.0000
      -0.0000       0.0000      -1.0000
       1.0000       0.0000       0.0000
       0.0000       1.0000       0.0000
       0.0000       0.0000       1.0000

SOLUTION FOUND
X IS      -0.4999    -0.8549    -0.2059     0.5011     0.8549
VALUE OF OBJECTIVE FUNCTION IS                      -22.627421378454
```

REFERENCES

[1] J. Abadie, "On the Kuhn-Tucker theorem", Nonlinear Programming, J. Abadie, ed., North-Holland Publishing Co., Amsterdam, 1967.

[2] R.A. Abrams, "Projections of convex programs with unattained infima", SIAM Journal of Control, 13 (1975), 706-718.

[3] R.A. Abrams and L. Kerzner, "A simplified test for optimality", Journal of Optimization Theory and Applications (forthcoming).

[4] K. Arrow, L. Hurwicz and H. Uzawa, "Constraint qualifications in maximization problems", Naval Research Logistics Quarterly, 8 (1961), 175-191.

[5] M. Atteia, "Fonctions spline definies sur un ensemble convexe", Numerische Mathematik, 12 (1968), 192-210.

[6] C.C. Augunwamba, "Optimality condition: constraint regularization", Mathematical Programming, 13 (1977), 38-48.

[7] M.S. Bazaraa, J.J. Goode and C.M. Shetty, "Constraint qualifications revisited", Management Science, 18 (1972), 567-573.

[8] M.S. Bazaraa, C.M. Shetty, J.J. Goode and M.Z. Nashed, "Nonlinear programming without differentiability in Banach spaces: necessary and sufficient constraint qualification", Applicable Analysis, 5 (1976), 165-173.

[9] A. Ben-Israel, A. Ben-Tal and A. Charnes, "Necessary and sufficient conditions for a Pareto optimum in convex programming", Econometrica, 45 (1977), 811-820.

[10] A. Ben-Tal, A. Ben-Israel and S. Zlobec, "Characterization of optimality in convex programming without a constraint qualification", Journal of Optimization Theory and Applications, 20 (1976), 417-437.

[11] A. Ben-Israel, A. Ben-Tal and S. Zlobec, "Optimality conditions in convex programming", The IX International Symposium on Mathematical Programming, Budapest, Hungary, August, 1976.

[12] A. Ben-Tal and A. Ben-Israel, "Characterizations of optimality in convex programming: the nondifferentiable case", Report (1976).

[13] A. Ben-Tal, L. Kerzner and S. Zlobec, "Optimality conditions for convex semi-infinite programming problems", Report (1976).

[14] A. Ben-Tal and S. Zlobec, "A new class of feasible direction methods", Report (1977).

[15] A. Ben-Tal and S. Zlobec, "Convex programming and the lexicographic multicriteria problem", Math. Operation-forsch. Statist. Set. Optimization, 8 (1977), 61-73.

[16] J. Borwein, "Proper efficient points for maximizations with respect to cones", SIAM Journal of Control and Optimization, 15 (1977), 57-63.

[17] F. Burns, M. Fiedler and E. Haynsworth, "Polyhedral cones and positive operators", Linear Algebra and its Applications, 8 (1974), 547-559.

[18] A. Charnes and W.W. Cooper, Management Models and Industrial Applications of Linear Programming, Vol.1, John Wiley & Sons, New York, 1961.

[19] F.H. Clarke, "Nondifferentiable functions in optimization", Applied Math. Notes, 2, Dec., (1976).

[20] B.D. Craven and J.J. Koliha, "Generalizations of Farkas' Theorem", SIAM Journal of Math. Analysis, 8, Nov., (1977), 983-997.

[21] J.W. Daniel and L.L. Schumaker, "On the closedness of the linear image of a set, with applications to generalized spline functions", Applicable Analysis, 4 (1974), 191-205.

[22] G. Debreu, Theory of Value. An Axiomatic Study of Economic Equilibrium, John Wiley & Sons, 1959.

[23] V.F. Dem'yanov and V.N. Malozemov, Introduction to Minimax, Translated from Russian by D. Louvish, John Wiley & Sons, New York, 1974.

[24] J. Dieudonné, "Sur la séparation des ensembles convexes", Mathematische Annalen., 163 (1966), 1-3.

[25] A. Ya Dubovitskii and A.A. Milyutin, "Extremum problems in the presence of restrictions", Zh. vychish. Mat. mat. Fiz., 5 (1965), 395-453.

[26] A.V. Fiacco, "Convergence properties of local solutions of sequences of mathematical problems in general spaces", Journal of Optimization Theory and Applications, 13 (1974), 1-12.

[27] S.D. Fisher and J.W. Jerome, Minimum Norm Extremals in Function Spaces, Lecture Notes in Mathematics, 479, Springer-Verlag, New York, 1975.

[28] I.V. Girsanov, Lectures on Mathematical Theory of Extremum Problems, Lecture Notes in Economics and Mathematical Systems, No.67, Springer-Verlag (1972).

[29] J.L. Goffin, "On convergence rates of subgradient optimization methods", Mathematical Programming, 13 (1977), 329-347.

[30] F.J. Gould and J.W. Tolle, "Geometry of optimality conditions and constraint qualifications", Mathematical Programming, 2 (1972), 1-18.

[31] F.J. Gould and J.W. Tolle, "A necessary and sufficient qualification for constrained optimization", SIAM Journal of Applied Mathematics, 20, 2 (1971), 164-171.

[32] M. Guignard, "Generalized Kuhn-Tucker conditions for mathematical programming problems in a Banach space", SIAM Journal of Control, 7 (1969), 232-241.

[33] P.R. Halmos, Introduction to Hilbert space, Chelsea Publishing Co., New York, 1957.

[34] R.B. Holmes, A Course on Optimization and Best Approximation, Lecture Notes in Mathematics, No.257, Springer-Verlag (1972).

[35] R.B. Holmes, Geometric Functional Analysis and its Applications, Graduate Tests in Mathematics, 24, Springer-Verlag (1975).

[36] H.W. Kuhn, "Nonlinear programming: a historic view", SIAM-AMS Proceedings, 9 (1969), 1-26.

[37] O.L. Mangasarian, Nonlinear Programming, McGraw-Hill (1969).

[38] H. Massam, "Mathematical programming with cones", Ph.D. Thesis, McGill University, 1977.

[39] H. Massam and S. Zlobec, "Various definitions of the derivative in mathematical programming", Mathematical Programming, 7 (1974), 144-161.

[40] B.N. Pshenichnyi, Necessary Conditions for an Extremum, Marcel Dekker, INC., New York (1971).

[41] R.T. Rockafellar, Convex Analysis, Princeton University Press (1972).

[42] R.T. Rockafellar, Conjugate Duality and Optimization, SIAM, Regional Conference Series No.16.

[43] R.T. Rockafellar, "Ordinary convex programs without a duality gap", Journal of Optimization Theory and Applications, 7, 3 (1971) 143-148.

[44] R.T. Rockafellar, "Some convex programs whose duals are linearly constrained", Nonlinear Programming, J.B. Rosen, O.L. Mangasarian and K. Ritter, eds., Academic Press, New York, 1970, 293-322.

[45] W. Rudin, Functional Analysis, McGraw-Hill, 1973.

[46] W. Rudin, Real and Complex Analysis, McGraw-Hill, 1966.

[47] M. Slater, "Lagrange multipliers revisited. A contribution to nonlinear programming", Cowles Commission Paper, Math., 403 (1950).

[48] G. Wolkowicz, "Some aspects of stability in nonlinear programming", M.Sc. Thesis, McGill University, 1978.

[49] H. Wolkowicz, "Calculating the cone of directions of constancy", Journal of Optimization Theory and Applications, 25 (1978), 451-457.

[50] S. Zlobec, "Asymptotic Kuhn-Tucker conditions for mathematical programming problems in a Banach space", SIAM Journal of Control, 8, 4 (1970), 505-512.

[51] S. Zlobec, "Extensions of Asymptotic Kuhn-Tucker conditions in mathematical programming", SIAM Journal of Applied Mathematics, 21, 3 (1971), 448-460.

[52] S. Zlobec, Mathematical Programming, Lecture Notes, McGill University, 1977.

[53] G. Zoutendijk, Methods of Feasible Directions, A Study in Linear and Non-linear Programming, Elsevier, New York, 1960.